Using Classification Trees to Link Serial Crimes

by

Rebecca Mugford

A thesis submitted to
the Faculty of Graduate and Postdoctoral Affairs
in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

in

Psychology

Carleton University
Ottawa, Canada

Abstract

In the investigative setting, police must often decide whether multiple crimes have been committed by a single offender. Using a variety of statistical techniques, studies have shown that it is possible to link serial crimes in a relatively accurate fashion using behavioural information (i.e., a process often referred to as behavioural linkage analysis; BLA). Despite this, practitioners often resist using these techniques, in a similar fashion to how clinical psychologists often resist actuarial techniques. In an attempt to develop an approach to BLA that may be better received by end users, this dissertation explored how classification trees (CTs) can be used to link serial crimes. Specifically, three variations of a CT approach were explored: a standard, single CT, an iterative CT (ICT), and the combination of multiple standard CTs and/or ICTs (i.e., a multiple model approach). Using separate samples of serial break and enters from Saint John, New Brunswick (N = 170) and serial sexual assaults from Quebec (N = 260), the ability of these approaches to link serial crimes were compared to one of the most commonly employed statistical approaches to BLA: main-effects logistic regression analysis. Generally, results revealed that all statistical approaches achieved high (and similar) levels of predictive accuracy; however, a number of potential advantages of a simple, standard CT approach were identified (e.g., transparency and ease-of-use). The findings reported in this dissertation have implications for BLA researchers (e.g., how behavioural domains are defined, how crime samples are selected, etc.) and police practitioners (e.g., the availability of a user-friendly statistical linking tool, the need for better data collection protocols, etc.). However, before a CT-based approach to BLA is implemented in practice, future research is required to address some of the limitations of the current research.

Acknowledgements

First, I would like to thank my supervisor, Dr. Craig Bennell, for all of his help and encouragement throughout my entire post-secondary journey. I know for a fact that none of what I have accomplished or the opportunities that I have been provided over the years would have been possible without your help and support. I am truly grateful for everything you have done and don't think I will ever be able to thank you enough.

I would also like to thank all of my committee members – Drs. Shelley Brown, Kevin Nunes, George Pollard, and Matthew Tonkin – for reading my dissertation and providing me with helpful feedback. I know you are all busy, particularly around this time of year, so I am genuinely appreciative of the time and effort you have invested in this. I would also like to thank Etelle and the rest of the Psychology Department's administrative staff – you have been extremely helpful over the years whenever I have had questions about the graduate student process. I would also like to express my gratitude to Angela Totten at the Saint John Police Force and Dr. Eric Beauregard for providing me with samples of Canadian crime for this dissertation.

I would also like to thank all members of the Police Research Lab – both past and present – for all of their support over the years. You have all made the entire graduate student experience much more enjoyable. I would like to give a special shout out to Karla Emeno, Holly Ellingwood, and Brittany Blaskovits – your support has been crucial, particularly over the last couple of years! Finally, I would like to thank my parents, my sister, Sara, and my friends (especially Kailey, Laura, and Celine) for your support and encouragement over the years.

Table of Contents

List of Tables

List of Figures

List of Appendices

CHAPTER 1

Introduction

In the investigative setting, police must often decide whether multiple crimes have been committed by a single offender (Grubin, Kelly, & Brunsdon, 2001). When forensic evidence is unavailable or compromised, practitioners must depend on crime scene behaviours to make connections between crimes, a process commonly referred to as behavioural linkage analysis (BLA; Woodhams, Hollin, & Bull, 2007).[1] To be a valid investigative technique, two assumptions underlying BLA must be supported: (1) intra-offender behavioural consistency and (2) inter-offender behavioural distinctiveness (Canter, 1995). Intra-offender behavioural consistency refers to the presence of similar crime scene behaviours across crimes in a given offender's series of crimes, whereas inter-offender behavioural distinctiveness refers to the presence of dissimilar crime scene behaviours across crimes committed by different offenders.

There are both practical and theoretical benefits to studying BLA. From a practical standpoint, BLA has the potential to be a valuable investigative tool that can allow for a more effective and efficient investigation of serial crimes (Grubin et al., 2001; Labuschagne, 2012; Woodhams et al., 2007). This is largely because BLA can facilitate the consolidation of evidence across the linked crime scenes, which ultimately provides a more complete picture of the crimes under investigation (Grubin et al., 2001). Moreover, in certain countries, BLA evidence is being admitted into court when questions arise about whether the accused is responsible for committing multiple crimes (see Labuschagne, 2012, 2014 and Charron & Woodhams, 2010 for some examples of court

[1] Throughout this dissertation, the term "practitioners" will be used to collectively refer to those who link crimes in practice, which can include (but is not limited to): police officers, crime analysts, and Behavioural Investigative Advisors in the UK.

cases). Consequently, it is important that research tests the assumptions of BLA to determine if the technique does (or does not) meet the evidentiary standards of the jurisdictions accepting it as evidence (e.g., the Daubert criteria; *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 1993).

Theoretically, studying BLA allows for a better understanding of the patterns of behaviour exhibited by serial offenders which can, in turn, suggest explanations for why certain patterns might emerge (e.g., the degree to which criminal behaviour is offender- vs. situation-driven, or dependent on offender-situation interactions; Funder & Colvin, 1991). Moreover, studying BLA can also advance theoretical models that attempt to explain consistency and distinctiveness in human behaviour (e.g., aspects of the cognitive-affective personality system [CAPS]; Mischel & Shoda, 1995). For instance, determining if, when, and how serial offenders are consistent and distinctive in their crime scene behaviours may provide extensions to existing theoretical models, or conversely, may more precisely define the boundary conditions under which a particular model may not apply.

Using a variety of statistical techniques, such as logistic regression analysis, a number of studies over the last few decades have shown that serial offenders do exhibit a reasonable degree of intra-offender consistency and inter-offender distinctiveness, which makes it possible to link serial crimes in a relatively accurate fashion (e.g., Bennell & Canter, 2002; Canter et al., 1991; Santtila, Junkkila, & Sandnabba, 2005; Slater, Woodhams, & Hamilton-Giachritsis, 2015; Tonkin, Grant, & Bond, 2008; Winter et al., 2013). However, despite the demonstrated value of these techniques, linking decisions in the investigative setting are still typically made in an unsystematic (and potentially

inferior) way, with different practitioners relying on different behavioural features when making linking decisions, sometimes even using linking cues that are not supported by empirical research (e.g., see Bennell, Bloomfield, Snook, Taylor, & Barnes, 2010; Burrell & Bull, 2011; Canter et al., 1991; Hazelwood & Warren, 2003; Keppel & Weis, 2006; Santtila, Korpela, & Häkkänen, 2004). In fact, practitioners often respond with resistance to the use of such techniques (e.g., Bennell et al., 2010) in a similar fashion to how clinicians have often resisted actuarial techniques in the field of clinical psychology (Dawes, Faust, & Meehl, 1989; Grove & Meehl, 1996).

There are many potential explanations for this resistance, but two obvious ones have been raised anecdotally and are supported by research in other domains (e.g., Swets, Dawes, & Monahan, 2000). First, some practitioners argue that the statistical techniques that currently exist for conducting BLA are potentially too complex and cumbersome for them to be of use in practice. Second, existing statistical approaches to BLA ultimately lead the practitioner to rely on certain nomothetic, one-size-fits-all strategies for linking crimes, which practitioners find dubious (i.e., general patterns of behaviour that are assumed to point to linked sets of crimes, such as short inter-crime distances (ICDs); e.g., Bennell & Canter, 2002; see Rainbow & Gregory, 2011, for a very similar argument in relation to criminal profiling, where practitioners use crime scene behaviours to predict offender characteristics). Both of these potential problems (i.e., the complexity and one-size-fits-all nature of many of the statistical approaches used to explore to BLA to date) may cause practitioners to question the value of actuarial tools, regardless of how effective they are at linking crimes.

The goal of this dissertation is to present a relatively novel approach to BLA, which relies on classification trees (CTs; Monahan et al., 2001; Steadman et al., 2000). Although CTs have been used to develop predictive models in a variety of fields for a number of years (e.g., marketing and medicine; Magidson, 1993; Podgorelec, Kokol, Stiglic, & Rozman, 2002, respectively), only recently has this approach been explored within the BLA field (e.g., Tonkin, Woodhams, Bull, Bond, & Santtila, 2012). The argument will be presented in this dissertation that this statistical approach provides a viable alternative to other statistical techniques that are commonly used to conduct BLA, most notably logistic regression analysis.

Not only is the CT approach predicted to link crimes with the same (or greater) degree of accuracy as other statistical approaches, this approach is arguably less computationally complex and cumbersome than other approaches, and more intuitively appealing, both in terms of how it can be used by practitioners and in terms of the linking strategies that it produces (Monahan et al., 2001; Steadman et al., 2000). That is, a CT approach may produce linking strategies that are more idiographic in nature while maintaining simplicity in terms of the computational procedures involved in the statistical analyses. As such, a CT approach to BLA may be more readily accepted into investigative practice. Equally valuable, a CT approach may uncover important patterns of serial offending behaviour that have previously gone unnoticed using the traditional statistical approaches to BLA because these approaches have relied on a main-effects, one-size-fits-all approach to examining behavioural consistency and distinctiveness.

To determine the value of the CT approach to BLA, two Canadian samples of crime (break and enter, and sexual assault) will be used to explore how different

variations of a CT-based approach to crime linkage compare to one of the most commonly employed statistical approaches to BLA: main-effects logistic regression analysis. Throughout this dissertation, various CT and logistic regression linking models will be developed and compared (sometimes only qualitatively) across a range of features, including: predictive accuracy, generalizability, usability, and their ability to capture unique patterns in behavioural consistency and distinctiveness. Ultimately, the goal of this dissertation is to determine whether CTs offer any practical or theoretical advantages to the BLA field.

This dissertation is structured as follows. In Chapter 2, an overview of the findings of BLA research to date is provided and the limitations of the statistical approaches used in this research are explored in relation to the purpose of this dissertation. In Chapter 3, a description of CTs is first provided, followed by an overview of the three approaches for constructing CT-based prediction tools that were developed in a large-scale violence risk assessment study (Monahan et al., 2001), and ultimately explored in the current dissertation. An overview of the possible advantages of adopting these different CT approaches is also provided. In Chapter 4, a brief overview of the goals and hypotheses of the current dissertation is provided. Chapter 5 outlines the various methodological steps employed to produce the linking models. The results of the analyses are then presented using samples of serial break and enters (Chapter 6) and serial sexual assaults (Chapter 7). Finally, in Chapter 8, the findings are summarized and the results from the two datasets are compared. This chapter ends with a discussion of the implications and limitations of this research, as well as suggestions for future research.

CHAPTER 2

Literature Review

In this chapter, an overview of the research exploring our ability to link crimes using statistical approaches is provided, followed by a discussion of some of the potential disadvantages of the statistical approach adopted in most of this research to date.

Empirical Support for Behavioural Linkage Analysis

In their review of BLA research, Woodhams et al. (2007) identified a number of studies that have examined whether behavioural information can be used to accurately link crimes using a range of crime types. For instance, many studies have supported the use of BLA with a variety of property crime types, including residential burglary (e.g., Tonkin, Santtila, & Bull, 2012), commercial burglary (e.g., Bennell & Canter, 2002), commercial robbery (e.g., Woodhams & Toye, 2007), arson (e.g., Santtila, Fritzon, & Tamelander, 2004), and car theft (e.g., Tonkin et al., 2008). Other studies have found that it is possible to use behavioural information to link crimes of an interpersonal nature, including sexual assault (e.g., Slater et al., 2015), personal robbery (e.g., Burrell, Bull, & Bond, 2012), and homicide (e.g., Santtila et al., 2008). Finally, some evidence is emerging that suggests it is possible to use behavioural information to link different types of crime committed by the same offender (e.g., a commercial burglary and commercial robbery committed by the same offender; Tonkin & Woodhams, 2015; Tonkin, Woodhams, Bull, Bond, & Palmer, 2011).

A variety of statistical techniques have been employed in these studies, including: cluster analysis (e.g., Green, Booth, & Biderman, 1976), discriminant function analysis (e.g., Santtila, Fritzon, et al., 2004; Santtila et al., 2008), multidimensional scaling

analysis (e.g., Santtila et al., 2005), Bayesian modelling (e.g., de Zoete, Sjerps, Lagnado, & Fenton, 2015; Winter et al., 2013), and logistic regression analysis (e.g., Bennell & Canter, 2002; Slater et al., 2015; Tonkin et al., 2008; Woodhams & Labuschagne, 2012). Despite the use of different statistical approaches, some underlying methodological commonalities are evident across many of these studies. For example, many researchers measure across-crime behavioural similarity in some manner, use a statistical procedure to compare the behavioural similarity observed between crimes committed by the same offender versus different offenders for the purpose of classifying crimes as linked or unlinked, and then, based on some evaluative criteria, make inferences as to how that particular procedure performs in terms of predictive accuracy.

However, it is important to note that two different types of linking tasks are also examined across these studies: (1) a pairwise task of deciding whether two crimes are linked to one another (i.e., are these two crimes linked?; e.g., Bennell & Jones, 2005, Slater et al., 2015), or (2) a series-level task of deciding whether a single crime belongs to a particular series of crimes that can vary in length (i.e., does crime A belong to series Z?; e.g., de Zoete et al., 2015; Santtila et al. 2008). Moreover, the statistical techniques adopted by researchers tend to be applied to one task or the other (e.g., logistic regression analysis has always been applied to the pairwise task, whereas Bayesian modelling has always been applied to the series-level task; this is not to say, however, that these statistical techniques could not be used to examine both tasks).

In one of the first studies to examine BLA, Green and colleagues (1976) used cluster analysis to determine whether linking was possible with a sample of 15 solved residential burglaries committed by three offenders in Tennessee. Green et al. first

calculated a composite across-crime similarity score for each pair of crimes, reflecting whether the behaviours exhibited within the crimes (e.g., type of material taken, location on block, etc.) were similar or different from one another. These similarity scores were then used as input for a cluster analysis that provided a graph displaying the individual crimes, with the space between each crime representing their degree of behavioural similarity (i.e., the higher the similarity score for two burglaries, the closer they would be in space on the graph). In an attempt to isolate burglaries committed by the three offenders, the authors then subjectively separated the crimes into three distinct clusters on the graph. Examining the crimes contained within each of these clusters, Green et al. found that 93% of the burglaries were attributed to the correct offender.

In a more recent study, Santtila, Fritzon, et al. (2004) employed principal component analysis (PCA) and discriminant function analysis (DFA) to examine BLA with serial arsons. Using behavioural information from 248 arsons committed by 42 offenders in Finland, each case was first coded for the presence or absence of behaviours found across the offences (e.g., forced entry, gasoline used, victim known, etc.). A PCA was then conducted to identify groups of behaviours that represented the same core theme. To determine the extent to which the identified behavioural themes could distinguish among series of arsons, DFA was conducted using the summary scores for each theme as the predictor variables, and the series that each offence belonged to as the grouping variable. Using the resulting discriminant functions, 32% of the sample could be classified as belonging to the correct crime series, well beyond that expected by chance (3%). Moreover, for 52% of the cases, the correct crime series (i.e., the series that

the case belonged to in reality) was listed among the top ten most probable series identified by the discriminant functions.

Perhaps the most common statistical approach adopted in the field of BLA research to date, and one that will be focused on in this dissertation, is logistic regression analysis. First used by Bennell and Canter (2002) and Bennell and Jones (2005) to examine the feasibility of BLA for commercial and residential burglaries, respectively, this statistical approach has now been used to test BLA for a wide range of property and interpersonal crimes (e.g., Markson, Woodhams, & Bond, 2010; Slater et al., 2015; Tonkin et al., 2008; Tonkin et al., 2011; Woodhams & Labuschagne, 2012; Woodhams & Toye, 2007). For instance, Tonkin et al. (2008) coded 386 solved car thefts committed by 193 offenders in the UK against a checklist consisting of behaviours characterizing three domains: target selection choices, target acquisition behaviours, and disposal behaviours. Across-crime similarity scores were calculated for every pair of crimes for each of these three domains.[2] Tonkin et al. also measured spatial similarity for every pair of crimes in two ways: by calculating the inter-crime distance (ICD; i.e., the distance in kilometres between the locations where the cars were stolen) and the inter-dump distance (i.e., the distance in kilometres between the locations where the stolen cars were eventually located). A stepwise (optimal) regression model and separate simple logistic regression models were developed to classify crime pairs as linked or unlinked on the basis of their across-crime similarity. To determine whether these models could generalize well to a

[2] Like most linking studies adopting a similar methodological approach, across-crime similarity in this study was measured using Jaccard's coefficient (*J*) (Jaccard, 1908), which can range in value from 0 (indicating that no similar behaviours are observed across both crimes in a given pair) to 1 (indicating that the behaviours observed across both crimes are completely similar). Jaccard's is calculated as $J = a/(a + b + c)$, with *a* equal to the number of behaviours that are the same across crime 1 and crime 2, *b* equal to the number of behaviours present in crime 1 but not crime 2, and *c* equal to the number of behaviours present in crime 2 but not crime 1.

new sample of crimes, the models were then applied to a test sample of crime pairs (i.e., a process referred to as cross-validation). Subsequent Receiver Operating Characteristic (ROC) analyses revealed that the highest levels of predictive accuracy were achieved by ICD (test AUC = .81) and inter-dump distance (test AUC = .77).[3] All other domains were associated with relatively low levels of predictive accuracy (test AUCs ranging from .56 to .57). Tonkin et al. attributed the lower predictive accuracy associated with some domains to the fact that the behaviours making up those domains are more likely to be context-specific, whereas the decision of where to take and leave a stolen car is likely to be more under the control of the offender.

In a review of the published linking studies that used ROC analysis to assess linking accuracy, Bennell, Mugford, Ellingwood, and Woodhams (2014) found that the AUCs reported across the studies ranged from just below chance-level accuracy (AUC = .45; Burrell et al., 2012) to extremely high levels of accuracy (AUC = .96; Melnyk, Bennell, Gauthier, & Gauthier, 2011). Overall, approximately one-third (36%) of the AUCs reported in these studies were in the low range, half (49%) were in the moderate range, and one-fourth (14%) were in the high range, according to Swets' (1988) guidelines.

[3] ROC analysis is commonly used to measure the overall accuracy of decision-making approaches (e.g., linking models) when the decision outcome is binary (i.e., yes [linked crime pair] versus no [unlinked crime pair]). The analysis provides an ROC graph as output, which plots the hit rate (i.e., correct "linked" decisions) on the *y*-axis and the false alarm rate (i.e., incorrect "linked" decisions) on the *x*-axis across all possible decision thresholds for the particular decision criterion being used (e.g., all possible values of Jaccard's coefficient). When those points are connected, this produces an ROC curve. The area of the graph that falls below this curve, called the Area Under the Curve (AUC), is used as an overall index of accuracy. The AUC can range from 0 (perfect inaccuracy) to 1.00 (perfect accuracy), with .50 indicating chance levels of accuracy. According to Swets' (1988) guidelines, AUCs ranging from .50 to .70 indicate low levels of predictive accuracy, AUCs ranging from .70 to .90 indicate moderate accuracy, and AUCs ranging from .90 to 1.00 indicate high accuracy.

Based on their review, Bennell et al. (2014) highlighted some identifiable patterns of linking accuracy that emerged across these studies. For instance, they noted that AUCs for studies using interpersonal crime types (e.g., serial homicide, serial sexual assault) tended to be higher than those reported for property crime types (e.g., residential burglary, car theft), with the exception of the AUCs associated with certain features of property crime (e.g., features related to the distance in space and/or time between the crimes). They also noted that geographic differences appear to account for some of the variation in the AUCs reported, with differences even observed between boroughs (divisions) within the same police force (e.g., Bennell & Jones, 2005). Finally, they noted that the highest AUCs tend to be found for measures of geographical and temporal distance between crimes rather than modus operandi behaviours, such as entry behaviours in burglary, presumably because geography and time are able to be controlled largely by the offender (it is important to note that these space-time features have only been examined within property crime types to date).

Taken together, the results of the studies reviewed above, including the review by Bennell et al. (2014), suggest that it is possible to link serial crimes to a single offender using behavioural information, and that this task can be accomplished in a relatively accurate fashion. However, the degree of linking accuracy achieved in research studies may be moderated by a number of factors (e.g., crime type, geographical region, type of crime scene behaviours relied on, etc.), although it is not yet fully understood why these factors moderate accuracy.

Benefits and Pitfalls of Traditional Approaches to Behavioural Linkage Analysis

As illustrated in the above review, BLA has the potential to be a worthwhile investigative technique. From a practical standpoint, one of the goals of BLA might be to arrive at the most accurate and reliable decision concerning whether a crime pair is linked or unlinked given the available behavioural information.[4] In the investigative setting, linkage decisions are often made by relying on linking cues that, based on training or prior experience, a practitioner deems relevant to the BLA task (e.g., see Bennell, Snook, MacDonald, House, & Taylor, 2012; Burrell & Bull, 2011; Hazelwood & Warren, 2003). Decades of research in other areas of decision-making, however, have illustrated that decisions based on empirically established relationships between variables (i.e., actuarial decision-making) often outperform these clinical, or experienced-based, decisions on a variety of tasks (e.g., Æisdóttir et al., 2006; Dawes & Corrigan, 1974; Dawes et al., 1989; Grove & Meehl, 1996; Grove, Zald, Lebow, Snitz, & Nelson, 2000; Meehl, 1954).

As such, a general advantage of the approaches to BLA described above is that they are actuarial, rather than clinical, in nature. Given the consistent support for actuarial decisions over clinical judgment in other areas, it seems logical to assume that linking predictions made by these statistical approaches are likely to be more accurate than those made by an individual practitioner who is basing their decision solely on experience. This is because actuarial linking decisions are formed on the basis of systematic relationships between across-crime similarity scores and linkage status rather than an

[4] It is important to note that there are a number of linking tasks that are undertaken by police practitioners in the real-world setting, including: (1) searching for pairs of linked crimes in a large database of crimes, (2) being provided with an "index" crime (e.g., by a police investigator) and searching a large database for possible linkages to that index crime, and (3) being provided with a pre-set number of crimes (e.g., 5 or 6) and determining whether any of those crimes are likely to be linked or not. The current dissertation deals directly with the first task. As such, the results and conclusions are applicable to the first task only.

individual practitioner's (potentially biased) view of what behaviours might be most useful for linking crimes. Research that has compared the performance of human judges to actuarial methods on linking tasks generally supports this notion.

For instance, building on previous research (i.e., Canter et al., 1991; Santtila, Korpela, & Häkkänen, 2004), Bennell et al. (2010) directly compared the accuracy of linking decisions made by individuals varying in investigative experience to the predictions made using a logistic regression model. Students (n = 40) and police professionals (n = 31) were provided with information on 38 pairs of commercial burglaries committed in the UK and were asked to determine whether each pair was linked or unlinked. The information included the location of each offence on a map, the ICD for each pair, entry methods, target selection characteristics, and the property stolen. Prior to examining the information, all participants were told that only 20% of the pairs they were provided with were linked. Approximately half of the students and half of the police professionals (i.e., the 'trained' groups) were advised that BLA research has indicated that the single most relevant linking cue for determining whether the same offender has committed a pair of commercial burglaries is the ICD. The logistic regression equation containing only the ICD from Bennell and Canter's (2002) study on commercial burglaries was also applied to the same 38 crime pairs to allow for a comparison between linking decisions.

Using ROC analysis to assess linking accuracy, results from this study revealed that all groups (including the logistic regression model) performed significantly better than chance ($p < .001$). However, the linking decisions made by students (AUC = .76) were significantly more accurate than the decisions made by police professionals (AUC =

.67). Similarly, the linking decisions of trained participants (AUC = .77) were significantly more accurate than the decisions of untrained participants (AUC = .67). Overall, however, the logistic regression model made significantly more accurate linking predictions (AUC = .87) than those made by all other groups combined (AUC = .72).

These results suggest that, even when individuals have some amount of investigative experience, or have been provided with additional training about what cues should be attended to when making linking decisions, statistical models can still outperform them. This is potentially because human judges rely on available information regardless of whether it is relevant or not, whereas the statistical model only takes into account relevant linking cues. Indeed, when asked what information they used to form their linking decisions, a number of the trained police professionals in Bennell et al.'s (2010) study continued to prioritize behaviours they believed to be important to the linking task (e.g., similarity in property stolen), despite research that has found these behaviours to be largely unrelated to successful linking decisions (e.g., Bennell & Canter, 2002; Bennell & Jones, 2005). These illusory correlations – or beliefs that a relationship exists between two things (e.g., linkage status and similarity in the property stolen) when it does not truly exist – are common in many fields of decision-making (e.g., Arkes, 1981; Dawes, 1989).

Although there are potential advantages to adopting a statistical approach to BLA, there are also some drawbacks to the statistical methods currently used. The results of Bennell et al.'s (2010) study (i.e., the fact that trained participants still relied on ineffective linking cues) provide an indirect illustration of one of the central problems surrounding current statistical approaches to BLA – there may be resistance to the idea of

using empirical research (and the tools that emerge from this research) to make linking decisions in the investigative setting.

Although a decision support tool can easily be constructed based on the statistical approaches to BLA presented above, actuarial linking methods are rarely, if ever, used in practice (see Hazelwood & Warren, 2003 and Martineau & Corey, 2008 for a description of how linkage decisions are typically made in practice).[5] This dilemma is not exclusive to the investigative domain – researchers in other areas have also found that actuarial instruments, although promising, are often not adopted into the environment for which they have been developed (e.g., clinical psychology and HIV diagnoses; Dawes et al., 1989; Swets et al., 2000). A number of reasons have been proposed for why this is the case, with many of them directly relevant to the investigative setting.

First, as mentioned above, a common criticism of actuarial tools is that they are too complex for individuals lacking a background in statistics to use and understand (Dawes et al., 1989; Grove & Meehl, 1996; Swets et al., 2000). While actuarial tools could be implemented as easy-to-use, straightforward computer programs that require little understanding of the underlying statistical model it represents, practitioners should still understand the rationale behind the tool. Given the fact that practitioners may lack extensive training in advanced statistics, it would likely be difficult for them to fully grasp the processes guiding some statistical techniques, such as how the features of a logistic regression equation (e.g., regression coefficients, etc.) are estimated. Indeed, for

[5] In fact, some statistical tools that serve to support investigative decision-making are already in existence, but they have not been accepted into practice. For example, the Computer Aided Tracking and Characterization of Homicides (CATCH; Kangas, 2001) software was developed over a decade ago to help the police more systematically assess massive amounts of crime data, yet its use is not widespread across law enforcement agencies. Likewise, despite the fact that approaches to linking based on MDS have been shown to improve the accuracy of human judgment on the linking task (Canter et al., 1991), these methods have also not been incorporated into widespread practice.

BLA practitioners to see any value in adopting a certain linking tool into their everyday practice, it seems important that they accept and understand the reasoning behind why a given decision is reached.

Second, even if the statistical approach itself is understood, it may be difficult for practitioners to communicate the linking results to others (Grove & Meehl, 1996; Swets et al., 2000). This is primarily because the actual decision-making process of a particular tool is often not made explicit to the user. For instance, consider a computer program developed for linking commercial burglaries that uses the logistic regression equation developed by Bennell & Canter (2002):

$$\text{Log}\left(\frac{p}{1-p}\right) = -2.82 - 0.88(\text{inter-crime distance})$$

The practitioner would first input the crime locations into the program, and the program would calculate the distance between the two crimes in kilometres. Using the above equation, the program would then provide as output the estimated probability that the crime pair in question is linked (as well as some potential threshold the practitioner could adopt to decide whether that probability is high enough to consider the crime pair linked). Even if a practitioner understands the logic used to obtain this probability, it may be difficult for them to articulate the process to others when all that is provided to them are numerical values.

Since practitioners may be asked to justify their decisions in court or provide a rationale to their superiors for why investigative resources were allocated in a certain way on the basis of their decisions, it is crucial that they are able to effectively explain the steps taken to arrive at their decisions (e.g., *State v. Fortin*, 2004). Moreover, even

though the linking decision they made (on the basis of the actuarial tool) will potentially be more accurate than a decision made without the tool, the credibility of the decision may be questioned if the practitioner cannot explain how they arrived at their decision. As such, any advantage associated with adopting an actuarial approach to linking may be lost if the decision-making process of the tool is not made easily accessible to the practitioner.

A final issue with current BLA methods concerns the fact that they typically take a one-size-fits-all approach to the linking task (Monahan et al., 2001; Steadman et al., 2000; Swets et al., 2000). Consider a regression model containing ICD as the sole predictor of linkage status (e.g., Bennell & Canter, 2002; Bennell & Jones, 2005; Markson et al., 2010; Tonkin et al., 2008; Tonkin, Santtila, et al., 2012). Using this model to make linkage predictions, the exact same information would be applied in the same way to each crime pair. That is, regardless of the ICD calculated for each crime pair, all scores would be multiplied by the exact same weight. This one-size-fits-all issue is a common criticism of regression-based approaches because it may not always be appropriate to treat all cases in the same manner (Monahan et al., 2001).

This concern has certainly been raised by linkage practitioners, who point out the problems of applying the same linking cues to all crime pairs to determine linkage status. For example, with respect to ICD, while it is true that offenders are likely to commit their crimes in close proximity to one another (Bennell & Canter, 2002; Bennell & Jones, 2005; Rossmo, 2000; Tonkin, Santtila, et al., 2012), practitioners are well aware of the fact that there are many instances where offenders commute to their crimes and exhibit large ICDs (Canter & Larkin, 1993; Kocsis & Irwin, 1997). As a result, there may be

crimes that cannot be linked based on ICD; however, these crimes could potentially be linked using other behaviours. Although interactions between linking variables can, in principle, be modeled within a logistic regression analysis, it can be difficult to interpret the interaction terms (Menard, 2002). This would be especially true if the individual attempting to comprehend the decision does not have a strong understanding of statistics.

CHAPTER 3

A Classification Tree Approach to Behavioural Linkage Analysis

Given the potential advantage of actuarial over clinical BLA decisions, completely abandoning all efforts to introduce a statistical approach to BLA into the investigative setting (because of the aforementioned concerns) would be unwise. Instead, it seems sensible to explore whether statistical approaches to BLA can be adopted that may increase the chance that an empirically-based tool is accepted by linkage practitioners. One promising alternative to logistic regression modelling is to employ a classification tree (CT) approach to BLA.

This chapter begins with a general description of CTs, including their structure and terminology. An overview of the methodology, rationale, and results of the MacArthur Violence Risk Assessment Study (Monahan et al., 2001; Steadman et al., 2000) is then provided, as this study developed a number of apparent improvements to a CT-based modelling approach that are explored within in the remainder of this dissertation. The chapter ends with a discussion of the possible advantages of adopting a CT-based approach to linking crimes, followed by a brief overview of the results of the only published study to date that has examined this approach within the BLA context (Tonkin, Woodhams, et al., 2012).

Classification Trees

Classification trees have been widely used (and shown to be effective) in research pertaining to a range of decision-making settings, including: marketing (e.g., Lee, Harp, Horridge, & Russ, 2003; Magidson, 1993), medicine (e.g., Bosson & Labarere, 2006; Podgorelec et al., 2002), and the prediction of violence and recidivism (e.g., Berk,

Sherman, Barnes, Kurtz, Ahlman, 2009; Gardner, Lidz, Mulvey, & Shaw, 1996; Neuilly, Zgoba, Tita, & Lee, 2011; Steadman et al., 2000). Despite decades of use in other domains, a CT approach to linking has only been explored in the last few years (e.g., Tonkin, Woodhams, et al., 2012).

CT structure and terminology. Generally speaking, a CT embodies the many relationships that can be found between different combinations of predictors and the outcome of interest, displaying these complex relationships in an easy-to-follow hierarchical format (Rokach & Maimon, 2008; Steadman et al., 2000). The organization of a CT is determined by applying a statistical algorithm to the data used to construct the CT. In instances of categorical classification (e.g., linked vs. unlinked crimes), the Chi-squared Automatic Interaction Detection (CHAID) algorithm is commonly used which, as the name suggests, relies on the chi-square test of independence to split the CT using the various predictors (Kass, 1980; Magidson, 1993). The terminology used to describe the various components of a CT, as well as a hypothetical example of a standard CT resulting from the application of this methodology to burglary crime pairs, is provided in Figure 1.

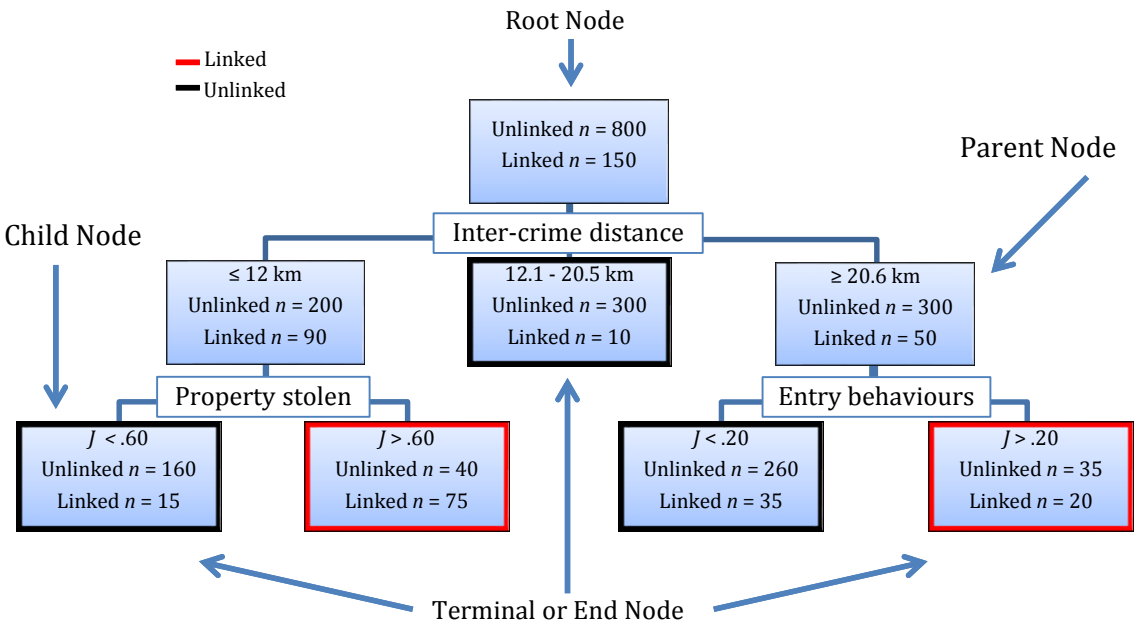


Figure 1. A hypothetical example of the structure of a standard burglary CT with "linked" versus "unlinked" as the decision outcome using a single decision threshold.

As shown in Figure 1, the CT first begins with all cases, regardless of group membership, at what is commonly called the 'root node'. The algorithm then runs a chi-square test of independence for all predictors and the outcome, and selects the predictor holding the most significant chi-square value to split the data into sub-groups (i.e., that best distinguishes linked from unlinked crime pairs). Each resulting sub-group is comprised of the cases that fall into the different categories of this predictor variable. These sub-groups are referred to as 'child nodes' because they arise from an earlier node, or what is called a 'parent node'. In Figure 1, for instance, ICD was the predictor chosen to initially partition the data into sub-groups.

Next, for each child node, this splitting process continues until the cases can no longer be further distinguished from one another on the basis of the predictor variables. That is, for each child node, the chi-square test of independence is again run for all possible predictors and the outcome variable, and the best predictor is again selected to further partition the data. If a variable is identified that can further differentiate the cases in the resulting child nodes, then this child node is split into new sub-groups based on the levels of this second predictor. This situation can be seen in Figure 1, where two levels of ICD are further partitioned by property stolen or entry behaviours. In this instance, the initial child node is also called a parent node since additional sub-groups (i.e., additional child nodes) can emerge from it.

Alternatively, if no additional predictors can differentiate between the cases in a given child node, the search for group differences ends, and that child node is then labelled a 'terminal' or 'end node.' In Figure 1, this scenario is illustrated in the middle category of ICD (crime pairs with distances between 12.1 and 20.5 kilometres), as well as for all categories of across-crime similarity (as represented by categories of *J*-scores) for either property stolen or entry behaviours.

Overall, the purpose of the CT is to create sub-groups that share certain qualities (e.g., values of different predictor variables) that are also the same on the outcome of interest (e.g., whether the crime pairs are linked or unlinked) (Monahan et al., 2001). As seen in Figure 1, each child node will, however, contain a certain proportion of cases from each decision alternative (i.e., from both the linked and unlinked groups). Analogous to the process of selecting and applying a decision threshold (i.e., a predicted probability value) to make linking decisions based on a logistic regression model, some

classification criteria must be established and applied to the end nodes in order to make a linked versus unlinked decision based on the resulting CT structure. It is common practice to make these decisions in reference to the base rate of the target decision outcome (e.g., Banks et al., 2004; Monahan et al., 2001; Silver & Chow-Martin, 2002). In BLA research, this would refer to the base rate of linked crime pairs in the entire sample. More specifically, a particular end node is commonly designated as representing the target decision outcome (i.e., linked) if the proportion of target cases comprising that node is more than twice the amount of the overall base rate (Monahan et al., 2001; Steadman et al., 2000).

For instance, consider the hypothetical burglary CT provided in Figure 1. The base rate of linked crimes in this situation is 15.78% (i.e., 15.78% of the crime pairs that were used to construct this particular CT are linked). As such, the threshold for classifying a node as linked would be 31.56% (i.e., more than 31.56% of the crime pairs in a given node must be linked crime pairs for that node to be classified as linked). Applying this threshold to the five end nodes in Figure 1, the two end nodes outlined in red would be classified as linked. Consequently, all remaining end nodes would be classified as unlinked (outlined in black).

By adopting the single threshold approach outlined above, it is clear that a number of crime pairs would be classified as unlinked even if, in reality, they are linked (e.g., all linked cases contained in the end nodes in Figure 1 outlined in black). In fact, this would correspond to 40% of the total number of linked crime pairs contained in the sample comprising the CT in Figure 1. Researchers employing a CT approach in other settings have recognized that it is impractical (and erroneous) to assume that any single prediction

model can correctly classify all cases into two mutually exclusive groups (Monahan et al., 2001; Steadman et al., 2000). For this reason, attempts to improve the decisions made using this standard CT approach have been proposed. These attempts are discussed next.

Enhancing the Standard CT Approach

One of the most comprehensive undertakings in the study of violence risk assessment to date, the MacArthur Violence Risk Assessment Study, involved the development of a CT-based actuarial tool for assessing violence risk in patients diagnosed with mental illness (Monahan et al., 2001; Steadman et al., 2000). Participants (N = 939) recruited from mental health hospitals in the United States were assessed on a total of 134 potential risk factors for violence (e.g., child abuse, gender, social preferences, delusions, etc.) prior to being discharged from the facility. Approximately 20-weeks after discharge, violence in the community was assessed to determine which patients did and did not engage in violent behaviour. Using actual violence as the outcome measure (violent versus not violent) and the various potential risk factors as the predictor variables, a forward stepwise logistic regression and a CHAID analysis were conducted and the level of predictive accuracy achieved by each model was compared. Using ROC analysis, the logistic regression and standard CT models were found to have relatively good (and comparable) levels of predictive accuracy (AUCs = .81 and .79, respectively).

Hypothesizing that predictions based on the CT approach could be enhanced, researchers involved in the MacArthur Study introduced three novel methodological refinements to CT modelling, including: (1) adopting two decision thresholds, (2) iterating the standard CT, and (3) constructing multiple iterative CT models in an attempt

to further improve the accuracy of predictions. Each of these modifications is discussed next.

Adopting two decision thresholds. As briefly mentioned, the MacArthur Study researchers acknowledged that risk assessment predictions (or any predictions for that matter) are not necessarily as simple and straightforward as a single decision threshold approach implies. In fact, there is likely to be a certain proportion of cases that one cannot unequivocally classify into high or low violence risk groups because they are not markedly low or high in violence risk (Monahan et al., 2001; Steadman et al., 2000). In other words, the risk of violence displayed by these ambiguous cases cannot be differentiated from the base rate of violence using the sequence of predictors initially identified by the standard CT model (Monahan et al., 2001; Steadman et al., 2000).

The same argument can be made concerning the BLA task. Although a certain prediction model may be successful at classifying some crime pairs, for other crime pairs, the similarity observed in the behaviours may not be extreme enough (in the direction of either high or low similarity) for accurate linked or unlinked decisions to be made using the initial set of predictor variables. In other words, these ambiguous crime pairs exhibit average levels of similarity for the particular behaviours included in the initial CT model (i.e., certain ranges of ICD and certain ranges of *J*-scores for a given sequence of behavioural domains).

For this reason, researchers involved in the MacArthur Study opted to use the violence base rate to construct *two* decision thresholds – one for each of the possible risk outcomes (Monahan et al., 2001; Steadman et al., 2000). In addition to labelling high violence risk groups as those exhibiting twice the violence base rate, they also used a

second threshold identifying the low violence risk groups as those nodes that contained less than half the base rate of violence. Using this two-threshold approach, they found that the logistic regression model could classify 57.1% of the cases into either the high or low risk group, whereas the standard CT model could classify 50.8% of the cases into either group. Thus, approximately half of the sample remained unclassified because they represented an average, or indistinguishable, level of violence risk when using either of these statistical approaches.

To illustrate how this approach can be applied to the BLA context, consider the previous hypothetical burglary example (now depicted below in Figure 2 using two thresholds). Since the base rate of linked crimes in this instance is 15.78%, the threshold for classifying a node as unlinked would be 7.89%. Whereas three end nodes were labelled unlinked when using the single decision threshold (Figure 1), only one end node (outlined in black) would be classified as containing unlinked crime pairs when using the two-threshold approach (Figure 2). Moreover, since the two remaining nodes in Figure 2 (outlined in yellow) fall between the linked and unlinked thresholds, they are labelled as unclassified end nodes. That is, relative to the number of unlinked crime pairs contained in these end nodes, the number of linked crime pairs remains too high to confidently classify them as unlinked.

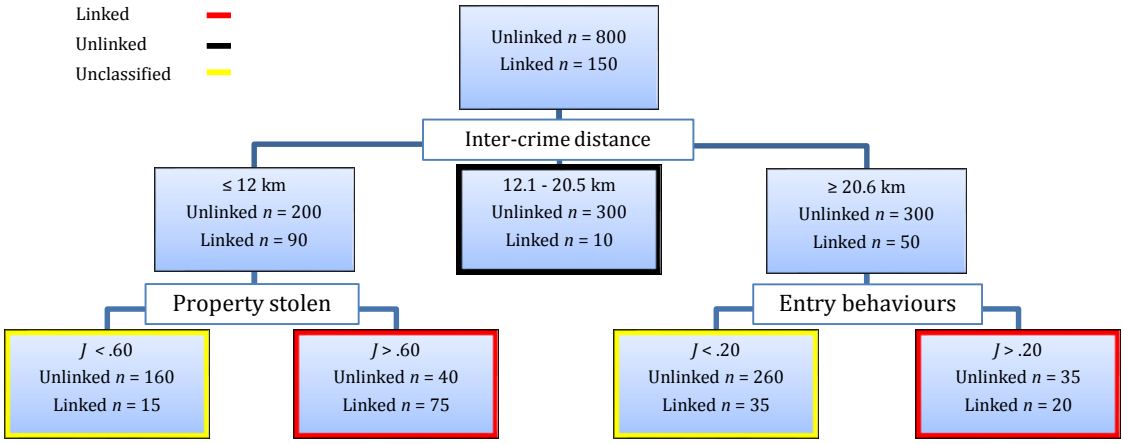


Figure 2. The hypothetical burglary classification tree using two decision thresholds, resulting in "linked", "unlinked", and "unclassified" decisions.

Iterating the standard CT. Although relatively good levels of predictive accuracy could be achieved using a standard CT approach to violence risk assessment, researchers involved in the MacArthur Study were concerned with the fact that only approximately half of all cases could be classified into a high and low violence risk group by applying the two decision thresholds to the standard CT (Monahan et al., 2001; Steadman et al., 2000). To resolve this issue, they devised an extension of the standard CT approach known as the iterative classification tree (ICT; Monahan et al., 2001; Steadman et al., 2000).

Essentially, the ICT approach involves the re-analysis of cases that are deemed unclassifiable by the standard CT when the two-threshold approach is used (e.g., the cases contained in the end nodes outlined in yellow in Figure 2). To increase the number of cases that could be classified into a high or low violence risk group, researchers

involved in the MacArthur Study combined the sub-groups containing unclassifiable cases and subjected them to another iteration (or re-analysis) using the CHAID algorithm. What resulted from this second analysis was a CT similar in structure to the first one, but containing different combinations of predictor variables to differentiate between the included cases. They then applied the two thresholds to this second CT, and again, re-analyzed the cases remaining unclassified after this second analysis. They repeated this process until the unclassified cases could no longer be distinguished from one another using different combinations of risk factors.

The ICT procedure undertaken in the MacArthur Study was terminated after four iterations. In other words, a total of four CTs were created. The first CT (or first iteration) represented the standard CT mentioned above. As explained, this CT classified 50.8% of the total sample into the high or low risk group based on the two thresholds. Using the same two thresholds, repeating the CHAID algorithm 3 more times classified an additional 25.8% of all cases above those classified by the standard CT alone. This resulted in a total of 76.6% of cases being classified into a high or low violence risk group. Moreover, the predictive accuracy achieved by the ICT (AUC = .82) was slightly better than that achieved using the standard CT (AUC = .79) or logistic regression model (AUC = .81). Thus, the ICT approach could still maintain similar levels of predictive accuracy as both previously examined approaches (logistic regression and the standard CT), yet the iterative process made it possible to classify a considerably higher number of cases into the definitive groups using the two decision thresholds.

Constructing multiple iterative CTs. Finally, researchers involved in the MacArthur Study also hypothesized that the accuracy achieved with an ICT model could

be further enhanced by constructing many ICT models and combining the risk predictions based on these models to provide a more robust 'combined' estimate of violence risk for each case (Banks et al., 2004; Monahan et al., 2001). As explained by the MacArthur Study researchers, improved predictions may be observed when combining multiple ICT models because the approach "may capture a different but important facet of the interactive relationship between the measured risk factors and violence" (Banks et al., 2004, p. 324). Using a different predictor variable to initially split the cases for each CT, 10 different ICT models were developed (applying the two thresholds to classify the cases as low risk, high risk, or unclassifiable). Next, all cases were scored on each of the 10 models, with each score reflecting whether the case was classified as low, high, or unclassifiable on the corresponding ICT model. For each case, these scores were then summed to create an overall risk estimate of violence.[6] Conducting an ROC analysis on the predictions made by the multiple ICT model, predictive accuracy was found to be higher (AUC = .88) than that achieved by the single ICT model (AUC = .82).

Promoting Acceptance of Actuarial Tools for Behavioural Linkage Analysis

Considering the collective results of the MacArthur Study, it is possible that a CT approach may be a promising alternative to current BLA methods. Although a CT approach would likely arrive at decisions in a noticeably different manner than a logistic regression approach, existing studies have shown that CTs are able to achieve a level of predictive accuracy that is comparable to logistic regression models (Monahan et al., 2001; Steadman et al., 2000), which is the current method of choice for carrying out BLA

[6] For each of the 10 models, a score of -1 was provided to a participant if they were in the low violence risk category for that model, a score of 0 was provided if they were unclassified, and a score of +1 was provided if they were in the high violence risk category. For each participant, these 10 model scores were then summed to provide the combined risk score. A more detailed description of this approach is present in Chapter 5.

in an actuarial manner (Bennell et al., 2014). The CT approach also holds a number of appealing qualities that can serve to bridge the gap between BLA research and practice by facilitating the adoption of actuarial tools into the investigative setting (Gardner et al., 1996; Monahan et al., 2001; Steadman et al., 2000). For example, recall that one issue with the traditional regression approach to BLA is that it is based on relatively complex statistical procedures. A CT approach to BLA can resolve this issue because, as explained above, the CTs are produced using relatively easy and comprehensible statistical procedures (i.e., the chi-square test; Kass, 1980). As such, a decision support tool resulting from a CT approach should be much easier for practitioners to grasp, even if they possess little prior statistical knowledge.

Similarly, the CT approach can overcome the common criticism that the decision-making process of a regression-based tool is not made sufficiently transparent to the end-user (Gardner et al., 1996; Monahan et al., 2001). The CTs formed from the CHAID analyses can easily be transformed into a visual decision support tool that guides the user, step-by-step, through the decision-making process. To illustrate this transparency, consider how the hypothetical CT presented in Figure 2 may be used in practice. Essentially, a practitioner can arrive at a decision by asking, in reference to the burglary pair of interest, a number of questions by referring to the CT. The order of these questions follows the progression of the CT beginning from the root node.

For instance, based on Figure 2, a practitioner would first examine the distance between the two burglaries. If this was less than or equal to 12 kilometres, the practitioner would then assess the *J*-score associated with the property stolen in the two burglaries. If this *J*-score is greater than .60, then the practitioner can classify that

burglary pair as linked. If it is less than .60, then that crime pair is unclassifiable and the practitioner would then proceed through the second CT (i.e., a CT that would be produced in iteration 2 of the CHAID analysis when the tool was developed) in the same manner, until he or she has either classified the crime pair as linked, unlinked, or unclassifiable. In fact, this decision-making process can be made far more accessible to the user in that a printout of the CTs that resulted in the linking decision can be provided. That is, the practitioner could be provided with a tangible copy of the sequence of questions they proceeded through to determine whether a crime pair is linked or not. This would undoubtedly improve their ability to communicate their decision-making process to others and would make it easier for them to justify their decisions to anyone who questions them.

Finally, as mentioned, a common criticism of traditional logistic regression approaches is that they take a very oversimplified approach to the decision-making process (Monahan et al., 2001; Steadman et al., 2000). In contrast, the CT approach has the capacity to easily deal with many complex interactions, which can potentially improve linking accuracy and provide new insights into serial offender behaviour. Although these interactions can become very complex, the CT approach illustrates the nature of the interactions between predictors in a straightforward manner.

For instance, consider the decision-making sequences for a practitioner when the crime pair in question is greater than or equal to 20.6 kilometres apart (as opposed to less than or equal to 12 kilometres apart as provided in the previous example). In this case, instead of referring to across-crime similarity in property stolen, the practitioner would assess the across-crime similarity score for the entry behaviours exhibited by that pair of

burglaries. If the similarity score is greater than .20, the practitioner would classify the crime pair as linked. If it is less than .20, the crime pair would be deemed unclassifiable, and the practitioner would again proceed to the second CT to ask a new sequence of questions until the crime pair can be placed into one of the sub-groups. Thus, the same ultimate decision (i.e., that a crime pair is 'linked') can be made for two different crime pairs by relying on the similarity evident in two different behavioural domains. Indeed, the fact that these interactions are taken into account will likely foster acceptance of an actuarial tool based on the CT approach, since practitioners may no longer perceive the tool as too generic or oversimplified to account for the complexities of offender behaviour.

Previous Research Exploring CTs in the Linking Context

As mentioned, only one published study to date has explored the use of CTs within the crime linkage context. Tonkin, Woodhams, et al. (2012) compared the linking models produced by logistic regression and CHAID analysis using two samples of serial property crimes: car thefts from the UK ($N = 376$) and residential burglaries from Finland ($N = 160$). Following the same *single* ICT procedure outlined in Monahan et al. (2001), Tonkin and colleagues' study reported a number of important findings.

First, only standard CTs were produced for both datasets: no second iterations were produced when the unclassifiable cases were pooled and re-analyzed (see Figure 3 for the CT produced using the serial burglary data). Second, a statistically comparable level of discrimination accuracy was found for the stepwise logistic regression and CT-based models across both crime types. For the residential burglary data, for instance, both the stepwise logistic regression model and CT model (as shown in Figure 3 for the test

sample) were associated with moderate levels of predictive accuracy and the differences in predictive accuracy observed were not statistically significant (test sample AUCs = .87 and .80, respectively). Third, although both approaches took into account the same predictors for the residential burglary data (i.e., ICD, similarity in entry behaviours, and similarity in internal behaviours), the CT approach resulted in *two* distinct pathways to a linked decision. As shown in Figure 3, when the model was applied to the test sample, a linked decision could be made by relying on either: (1) shorter ICDs (i.e., under 2.73 km) or (2) moderate ICDs (i.e., between 3.71 and 6.19 km) in combination with a higher degree of similarity in internal behaviours (i.e., similarity scores above .45).[7]

Arguably one of the most meaningful findings reported in this study was that the residential burglary CT did not generalize well from development to test (this was not the case for car theft). That is, the level of predictive accuracy achieved by the CT model was significantly lower in the test sample (AUC = .80, 95% CI [.71, .88]) than the development sample (AUC = .96, 95% CI [.94, .98]); however, this problem did not arise with the logistic regression model. Based on this finding, Tonkin and colleagues concluded that, although CT-based models may have a number of practical advantages as outlined above (e.g., increased transparency, ease of use, etc.), their reliability might be questionable. If the ultimate goal of conducting research of this variety is to develop linking models that can be used on new samples of crime pairs in practice, this issue is a legitimate concern.

[7] It is important to note that the serial car theft CT and stepwise logistic regression model both included ICD and disposal behaviours; however, the logistic regression model also included target selection behaviours (although there was no statistically significant increase in accuracy when this domain was included in the model).

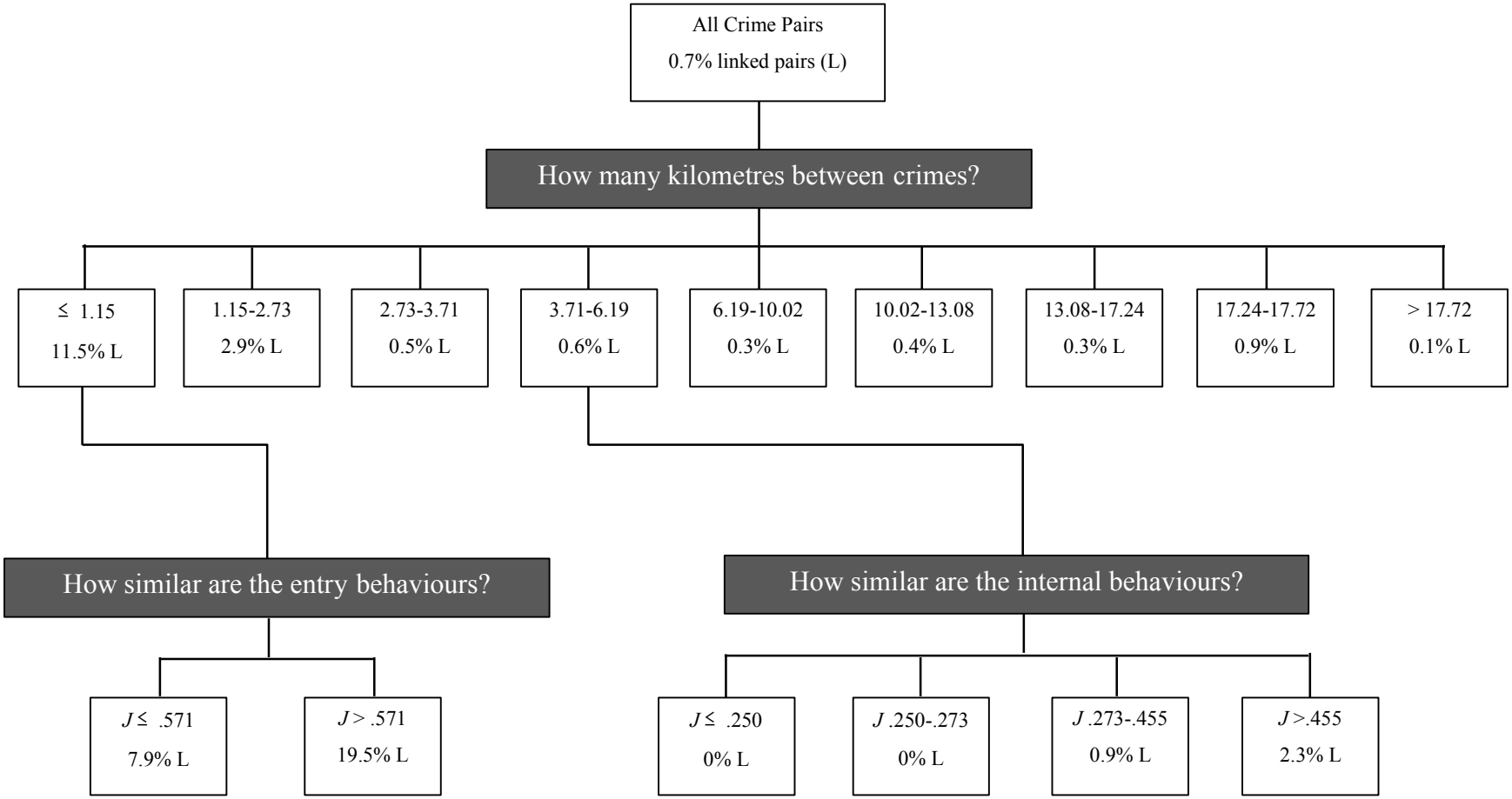


Figure 3. Standard classification tree produced by Tonkin, Woodhams, et al. (2012) using serial residential burglaries from the UK. Percentages refer to the proportion of linked ("L") crime pairs in each node when the tree was applied to the test sample of crime pairs. Adapted and reprinted with permission from the first author.

CHAPTER 4

The Current Study

In addition to focusing on Canadian crime data, which is something that has rarely been done in the BLA literature to date, the central goal of this study was to provide further insight into the performance of a CT-based approach to BLA relative to the traditional logistic regression approach. This study extended the research conducted by Tonkin, Woodhams, et al. (2012) in a number of ways. First, this study examined the value of adopting a multiple CT/ICT approach[8] to linking serial crimes (in addition to the standard CT and single ICT approaches that were tested in Tonkin, Woodhams, et al.'s study). Second, given that Tonkin, Woodhams, et al.'s study focused on a CT-based approach to linking property crimes only, the current study examined the performance of CTs when linking property crime (serial break and enters) *and* interpersonal crime (serial sexual assaults). Third, this study also incorporated temporal proximity (i.e., the number of days between two crimes) into the property crime linking models that were developed. Tonkin, Woodhams, et al. (2012) acknowledged that future research should examine how temporal proximity contributes to a CT-based approach to linking since it has consistently been found to be important to the prediction of linkage status in previous studies (e.g., Burrell et al., 2012; Markson et al., 2010; Tonkin, Santtila, et al., 2012).

Although the results of Tonkin, Woodhams, et al.'s (2012) study suggest that a simple, standard CT-based approach may be most appropriate for the BLA field, it is possible that the value of an ICT approach (and/or multiple CT/ICT approach) may be realized with crimes of an interpersonal nature. Indeed, the increased complexities

[8] This approach was ultimately dubbed the "multiple CT/ICT approach" because it may be the case that only standard CTs (or a mixture of standard CTs and ICTs) are produced prior to being combined to form the "multiple model approach" to CT-based linking.

involved in interpersonal crimes versus property crimes (e.g., due to the presence of a victim) may be better suited for the intricacies of an ICT (and/or multiple CT/ICT) approach.[9] With that said, Tonkin, Woodhams, et al.'s results do suggest that the generalizability of CT-based models is a valid concern. As such, it is worthwhile to explore whether generalizability concerns exist when using more complex CT-based approaches to BLA (i.e., ICTs and multiple CTs/ICTs), when exploring both property crime data and interpersonal crime data, and when using data from a geographical area (i.e., Canada) that has yet to be adequately explored in BLA research.

Finally, considering the potential theoretical advantages of adopting a CT-based approach to BLA, another goal of this study was to compare and contrast the behavioural patterns that emerged from a CT-based approach to BLA to patterns that emerged when using regression modelling. Although Tonkin, Woodhams, et al.'s (2012) study did not reveal any markedly different behavioural patterns between the two approaches in their serial burglary sample (i.e., the same predictors arose in both approaches and the CT produced was relatively low in complexity), this may be because the analyses were restricted to property crime types. Again, it may be that a CT-based approach is well suited to capture the increased behavioural complexity inherent in interpersonal crime types (versus property crime types). As such, more complex CTs (even standard CTs) may arise for interpersonal crime types, such as serial sexual assault.

[9] Indeed, research has found that sexual crimes are highly complex (e.g., Leclerc, Smallbone, & Wortley, 2014; Winter, 2014; Woodhams & Komarzynska, 2014). For instance, Leclerc et al. found that the ultimate sexual behaviours engaged in by child sexual offenders depends on the reaction of the victim to the offenders' initial actions. Given that victims are not integral to the commission of property offences, the behavioural patterns observed during interpersonal offences are expected to be more complex in nature.

Hypotheses

Based on previously conducted BLA research (e.g., Bennell & Canter, 2002) it was expected that crimes committed by the same offender would be characterized by a higher degree of behavioural similarity than crimes committed by different offenders. As a result, it was also expected that it would be possible to distinguish, or separate, linked crimes from unlinked crimes in a relatively accurate fashion (i.e., at levels greater than chance). This was expected to be true regardless of what model was used to carry out the linking task (i.e., logistic regression model, standard CT model, ICT model, or multiple CT/ICT model).

Based on previous research from other domains (e.g., Monahan et al., 2001) and from Tonkin, Woodhams, et al.'s (2012) study, the standard CT model was expected to achieve similar levels of predictive accuracy to the stepwise logistic regression model, whereas the ICT model was expected to result in a slightly higher level of accuracy than either the standard CT or stepwise logistic regression model. Relying on multiple ICTs, however, was expected to lead to the greatest level of linking accuracy when compared to any of the other models alone (stepwise logistic regression model, standard CT model, or single ICT model). Purely on the basis of the increased complexity of the behaviours observed in interpersonal crimes, it was also hypothesized that these differences in model performance would be more pronounced for the serial sexual assault dataset.

Based on the results of Tonkin, Woodhams, et al.'s (2012) study, it was also expected that all CT-based models would be less robust (i.e., have more issues with generalizability) than the stepwise logistic regression models, regardless of the crime type examined.

Finally, due to the interactive nature of CTs, patterns in behavioural consistency and distinctiveness that remain hidden when using the logistic regression approach were expected to emerge when using the various CT approaches, particularly when linking interpersonal crimes. Given that no research has examined a CT-based approach to linking crimes of an interpersonal nature, however, hypotheses were not made about the expected nature of these patterns.

CHAPTER 5

Methodology

Methodologically, this study involved four broad phases. Phase 1 involved conducting a descriptive analysis of similarity scores, Phase 2 involved developing and evaluating the main effects logistic regression models, Phase 3 involved developing and evaluating the standard CT and ICT models, and Phase 4 involved developing and evaluating the multiple CT/ICT models. A more detailed description of the steps involved in each phase is provided in the 'data analysis procedures' section below.

Data

Two samples of Canadian serial crime were analyzed in the current study: residential break and enters and sexual assaults. For the purpose of this study, serial offenders were defined as any offender who committed two or more offences of the same type (i.e., break and enters or sexual assaults) during the timeframe specified for each dataset (Federal Bureau of Investigation, 2008). In Canada, an individual can be charged with breaking and entering when they break and enter a dwelling for the purpose of committing (or intending to commit) an indictable offence, such as stealing property (Criminal Code, 1985). Sexual assault is defined as intentionally applying (or attempting to apply) force of a sexual nature to another person without their consent (Criminal Code, 1985). The same data analysis procedures were followed for each sample of crime.

Serial break and enter sample. Information from a total of 175 serial residential break and enters committed by 66 offenders between January 2001 and December 2013 was extracted from the Saint John Police Force's internal computer incident database. This sample represented all identified serial break and enters committed within the Saint

John Police Force's jurisdiction during this time period. Saint John is the largest city in the province of New Brunswick, with a population of 70,063 (Statistics Canada, 2012a). The city area covers 315.82 km^2 (population density 221.80 per km^2). The only criterion for inclusion in the current study was that the offender had been charged with a minimum of two residential break and enters that occurred between January 2001 and December 2013. This particular date range was selected to ensure the sample size was adequate for all analyses.[10] Each offender's crime series ranged in length from 2 to 9 crimes (M = 2.65, SD = 1.21; Median and Mode = 2.00). Offenders ranged in age from 12 to 52 (M = 24.13; SD = 9.38) and the vast majority were male (93.94%).

Coding incident files. Each electronic file contained several narratives, including the investigating officers' field notes, witness statements, and forensic reports (all typed verbatim into the database).[11] Behavioural checklists from previous linking research on residential burglaries were consulted and merged to form a comprehensive checklist for the current study (e.g., Bennell, 2002; Tonkin, 2012). Each crime was given a unique identifier prior to coding (e.g., 1-1 for offender 1, crime 1). Behaviours were then coded dichotomously (1 = behaviour/characteristic present; 0 = behaviour/characteristic absent) using the behavioural checklist presented in Appendix A. As shown in Appendix A, behaviours were grouped into the following domains to facilitate the calculation of similarity scores across linking features sharing the same function: (1) target selection (e.g., apartment, detached dwelling, daytime, etc.; 24 variables), (2) entry behaviours

[10] Although, in practice, crime linkages can still be made over such wide timeframes, linking models were explored when temporal proximity was both included and excluded from the analyses using the Saint John data (as discussed in more detail on pp. 85-86).

[11] The number of narratives available varied from crime to crime, with some crimes having more narratives to code than others. For example, when two officers attended the crime scene, there were two separate narratives containing officer field notes. Likewise, some crimes had forensic reports whereas others did not.

(e.g., access using door, access from rear, etc.; 18 variables), (3) internal behaviours (e.g., tidy search, intrusive search, etc.; 22 variables), and property stolen (e.g., small electronics, cash, etc.; 43 variables). A second researcher coded approximately 10% (n = 23) of the crimes in order to provide an index of inter-rater reliability for the data. Percent agreement was calculated across all behaviours and a satisfactory level of overall agreement was achieved (95.22%). A satisfactory level was also achieved when examining each behavioural domain separately (target selection = 86.96%; entry behaviours = 86.96%; internal behaviours = 93.67%; property stolen = 95.53%).[12]

Each break and enter also had a date and address associated with it in the incident database. Using geographic information system software (ArcGIS), addresses for each crime were converted to geo-coded *x*-*y* coordinates to the nearest metre to facilitate the calculation of spatial similarity scores. Specifically, the offence locations were converted to the Universal Transverse Mercator (UTM) geographic coordinate system, zone 19 (representing the location of Saint John within this particular coordinate system). Dates that the offences were committed were also extracted for each crime in order to facilitate the calculation of temporal proximity scores. An exact date of occurrence was not available for a minority of the break and enters because the occupants of the home were absent over an extended period of time when the crime was committed (n = 16; 9.10%). Given that this did not represent a large proportion of the crimes in the dataset, the

[12] Although percent agreement is not ideal given that it does not correct for chance agreement (whereas Cohen's Kappa does), percent agreement was used in this instance because there were a number of variables where Kappa could not be calculated given the nature of the data (e.g., one rater had all zeros – no variability – for all cases, whereas the second rater indicated a 1 for a single case). Since it is relatively rare in linking research to have a measure of inter-rater agreement, percent agreement was deemed an acceptable compromise given the characteristics of the data.

incident report date was used in these cases to avoid missing data when calculating temporal proximity.

Controlling the impact of prolific offenders. A common, and legitimate, concern in linking research is that prolific offenders may disproportionately influence the linking models developed (that is, that a small subset of offenders whose identified crime series is considerably larger than that of the typical offender in the data contribute a disproportionately high number of linked crime pairs to the dataset; e.g., Bennell & Canter, 2002; Burrell et al., 2012, Tonkin et al., 2008). For instance, in the serial break and enter dataset just described, the majority of offenders' crime series consist of two break and enters (65.20% of offenders); therefore, most offenders would only contribute one pair of linked crimes to the final dataset. However, one offender in the serial break and enter dataset committed 9 crimes, meaning they would ultimately contribute 36 linked crime pairs to the final dataset (since all possible pairs of crimes are used to develop the linking models). This is problematic if the goal is to develop linking models that are generalizable to offenders outside of the sample used to construct the linking models.

Crime linkage researchers typically control for the impact of prolific offenders by selecting a constant number of crimes per offender (e.g., Bennell & Canter, 2002; Burrell, Bull, Bond, & Harrington, 2015; Markson et al., 2010; Tonkin, Woodhams, et al., 2012). More recently, however, some researchers have started recommending that all identified crimes committed by the offenders during the timeframe examined be included in the study sample, suggesting that the decision to use a constant number of crimes per offender decreases the ecological validity of the resulting linking models (Tonkin &

Woodhams, 2015; Woodhams & Labuschagne, 2012). Given that both of these issues are legitimate concerns (i.e., biased linking models as a result of prolific offenders vs. biased linking models as a result of an artificially reduced dataset), a different approach to selecting the final sample was used in the current study that attempts to satisfy both arguments simultaneously (or at least offers a compromise).

More specifically, prolific offenders were first identified by detecting outliers on the variable "number of crimes in series" using a boxplot approach.[13] A variation on Winsorizing (Field, 2013) was then used to bring the crime series of prolific offenders within a "normal" range of crimes. That is, for each offender/crime series that was identified as an outlier, they were provided with a new "number of crimes per series" value that was one unit higher than the highest non-outlying value (i.e., non-prolific offender). A random sample of each identified prolific offender's crimes that corresponded to this new number of crimes per series was then selected for inclusion in the final dataset. This procedure is a reasonable compromise between the two previous approaches taken in the literature given that it reduces the impact of prolific offenders while simultaneously maintaining a larger dataset than would be the case if one restricted crime series to a constant number per offender. As such, it still maintains a greater degree of variability in the number of crimes committed per offender, which is more reflective of what may be encountered in operational settings.

[13]A common approach to identifying outliers is to transform raw values to *z*-scores and then classify cases with a *z*-score exceeding a cut-off *z*-value (e.g., typically 3.29; Field, 2013) as an outlier. This approach, however, was not used in this case given that the presence of some relatively extreme outliers resulted in an inflated standard deviation for the serial sexual assault data (*M* crimes per series = 5.00, *SD* = 6.04). Using the *z*-score approach, the inflated standard deviation masked many of the less extreme outlying scores (Wilcox, 2001). As such, the boxplot method (which relies on percentiles, not the standard deviation) was used instead (Wilcox, 2001). The boxplot method to identifying outliers was also adopted for both datasets for sake of consistency.

Using this approach, four crime series/offenders were identified as outliers (i.e., prolific offenders) in the serial break and enter data (crimes per series = 5, 5, 6, and 9). The highest non-outlying value was 4. As such, a random sample of 5 crimes was selected from the original series for the two offenders with 6 and 9 crimes. This resulted in a reduced final dataset of 170 break and enters that was then used in all subsequent analyses (series ranging in length from 2 to 5 crimes, $M = 2.58$, $SD = 0.91$).[14]

Serial sexual assault sample. Data collected as part of a large-scale Canadian study examining the characteristics of sex offenders and their offending patterns (i.e., Beauregard, 2005) were used to facilitate the comparison of logistic regression and CT-based linking approaches for interpersonal crime. This dataset has been used in several published studies to date (e.g., Beauregard, Leclerc, & Lussier, 2012; Hewitt, Beauregard, & Davies, 2012; Reid, Beauregard, Fedina, & Frith, 2014). Although a portion of these studies have examined behavioural consistency across the offenders' crimes, these studies typically focus on a subset of the variables examined in the current research (e.g., variables reflecting environmental or crime site selection consistency; Deslauriers-Varin & Beauregard, 2014a, 2014b). Likewise, none of the studies examining behavioural consistency have employed the methodological approach typically used in BLA research as described previously (i.e., development of stepwise logistic regression

[14] In fact, the analyses conducted in the current dissertation were carried out using data from the three different sampling approaches described above (i.e., using a constant (2) number of crimes per offender, using all crimes per offender, and using the outlier approach to controlling prolific offenders). Depending on the sampling approach employed, the multivariable models produced (i.e., stepwise logistic regression and all CT-based models) changed considerably in terms of the sequence of predictors included in the models and the predictive accuracy of the statistical models developed. As expected, predictive accuracy was higher when all crimes were included relative to when the outlier approach or the 2-crimes-per-offender sampling approach was used. The implications of employing these different sampling methods on future linking research are explored in Chapter 8.

models followed by ROC analysis) and none have examined behavioural consistency using CT-based approaches.

Unlike most linking research, which typically collects data from internal police computer databases, this dataset was primarily collected through semi-structured interviews with federally incarcerated offenders, supplemented by crosschecks of police investigative reports included in each offender's correctional file (Beauregard, 2005). To facilitate data collection, Beauregard (2005) identified all offenders who had committed a minimum of two stranger sexual offences and were serving a sentence of two or more years in a Quebec penitentiary between 1995 and 2004.[15] A total of 92 sex offenders met these sample criteria, with 72 offenders agreeing to participate in the research. However, only 69 offenders provided sufficient qualitative information to be included in the current analyses (Beauregard, 2005).[16] According to Statistics Canada (2012b), Quebec is the second largest province/territory in Canada both geographically (land area: 1,365,128 km^2) and population-wise (population 7,903,001; population density: 5.8 per km^2).

The original dataset contained a total of 347 stranger serial sexual assaults committed by 69 offenders between 1975 and 2003.[17] Crime series ranged in length from 2 to 37 crimes (M = 5.00, SD = 6.04; Median = 3.00; Mode = 2.00), and offenders ranged in age from 18 to 55 at the start of their crime series (M = 30.73; SD = 9.40). As

[15] Similar to studies examining our ability to link interpersonal crimes (e.g., Grubin et al., 2001; Santtila et al., 2005; Slater et al., 2015; Woodhams, Hollin, & Bull, 2008), "stranger" crimes were used in the current study and were operationalized as situations where the offender had "no personal relationship with the victim prior to the day the offense was committed" (Beauregard, 2005, p. 90).

[16] Three offenders did not provide sufficient qualitative information because they refused to answer many of the questions during the semi-structured interview and were therefore excluded from this particular dataset (E. Beauregard, personal communication, December 2, 2015).

[17] Although this is the best estimate of the date range possible, it is difficult to ascertain whether this represents the true date range for this sample of sexual assaults due to an abundance of missing dates - a total of 164 crimes (47%) were missing an exact date of crime and 26 crimes (8%) were missing a date range between which the offence reportedly occurred.

expected, the vast majority of sexual assaults were committed in the province of Quebec (n = 342); however, a small number were committed in: Ottawa, Ontario (n = 1), Ottawa, Ontario and Hull, Quebec (n = 1), Winnipeg, Manitoba (n = 1), and Calgary, Alberta (n = 2).

Decisions concerning the exclusion, inclusion, and recoding of variables. The original dataset contained over 500 variables pertaining to various aspects of the offence, offender, and victim. Generally speaking, the variables included in the original dataset could be subdivided into five broad domains: (1) pre-crime conditions (e.g., activities that the offender or victim were engaging in prior to the crime, the offender's mood prior to the crime, etc.), (2) crime scene behaviours (e.g., offender used restraints during the crime, offender brought a rape kit, sexual behaviours engaged in by the offender, verbal strategies used by the offender during the assault, etc.), (3) environmental and geographic conditions (e.g., location of the crime, type of area where the crime was committed, how often the offender changed locations during the commission of the assault, offender routine activity locations, etc.), (4) demographic characteristics (e.g., age of victim, sex of victim, race of victim, offender height/weight, offender marital status, etc.), and (5) other offender-related factors (e.g., scores on various psychological tests, whether the offender showed evidence of distorted thinking about the assault, whether the victim was selected based on various factors, etc.).

Given that the goal of the current study was to develop models that could potentially be used for BLA in practice, variables reflecting information that would likely be unknown to police at the time that the crime occurred were omitted from the current research. For instance, all offender-related variables that were not incident-specific were

deleted (e.g., offender demographic/physical characteristics and scores on the various psychological tests). Likewise, offender-related variables that may have been incident-specific but were unlikely to be known to police were deleted (e.g., the offender's pre-crime activities, whether the offender selected the victim randomly or non-randomly, whether the offender selected the victim based on their appearance, etc.). Finally, some behaviours that were specific to the offender's modus operandi behaviour were also omitted if it seemed unlikely that the victim or police would know the information at the time of the offence (e.g., the offender lied to the victim about his name).

Once each original variable was assessed against this inclusion criterion, all multilevel variables that were retained for subsequent analyses were transformed into several binary coded variables to facilitate the calculation of behavioural similarity scores. For instance, the variable "strategy to approach the victim" had the following eight levels: seduction/persuasion, money/gift, games, trick/false identity, using drugs/alcohol, act directly on victim, threat, and physical violence. Separate dichotomously coded variables were created for seven of these levels (1 = strategy present, 0 = strategy absent).[18] All variables were then separated into the following six domains prior to calculating similarity scores: (1) control behaviours (e.g., offender used false identity/trick to approach the victim, offender used physical violence to commit the crime; knife was used during the crime, offender retrieved bindings from the scene; 32 variables), (2) environmental behaviours (e.g., crime occurred on a weekday, offender changed locations once, offender encountered the victim in a park, victim was released in a residential area, etc.; 52 variables), (3) escape behaviours (e.g., offender used a

[18] The level "act directly on victim" was excluded from analyses because it was unclear how this differed from other levels for this variable (e.g., physical violence).

disguise, offender wore gloves, offender used a condom, etc.; 5 variables), (4) sexual behaviours (e.g., vaginal intercourse with fingers, exhibitionism, offender forced victim to engage in fellatio, etc.; 17 variables), (5) style behaviours (e.g., offender complimented victim, non-sexual part of victim's body was mutilated, offender used more force than necessary to commit the crime; offender stole something from the victim, etc.; 23 variables), and (6) victim selection (e.g., victim was male, victim was a child aged 0-12, victim was at a bar or nightclub before crime occurred, etc.; 19 variables). Please see Appendix B for a complete list of the variables included in each behavioural domain used in the serial sexual assault analyses.[19]

Finally, the dataset also had a date of crime variable and a number of location-related variables. More specifically, there was a location variable for four temporal phases across the assault: the encounter site (i.e., the location where the offender first encounters the victim), the attack site (i.e., the location where the offender first attacks the victim), the crime site (i.e., the location where the actual sexual assault occurs), and the victim release site (i.e., the location where the offender left their victim, signalling the end of the crime commission process; Rossmo, 2000). Large amounts of this data were missing across all of these variables. For instance, only 53% (n = 183) of the crimes in the dataset had a date of crime that could be used to calculate temporal proximity. Likewise, although a location was listed for every crime for each temporal phase, it was often not a useable address from which x-y coordinates could be produced to calculate spatial similarity scores. For instance, frequently the location was simply listed as "the woods", "the street", or "victim" (presumably referring to the victim's house) in a particular city or town in Quebec. In total, only 34% of crimes (n = 119) had a useable

[19] Unfortunately, no index of inter-rater reliability was provided for the serial sexual assault dataset.

encounter address, 40% (n = 140) had a useable attack address, 41% (n = 143) had a useable crime address, and 37% (n = 129) had a useable victim release address. The options for dealing with missing data were to: (1) delete cases that were missing the values and run the analyses on a subset of cases, (2) substitute the missing values with some replacement strategy (e.g., multiple imputation), or (3) omit the problematic variables from the analyses (Tabachnick & Fidell, 2007).

From a statistical perspective, deleting crimes where useable dates and addresses were not available would be problematic for a number of reasons. First, eliminating a large proportion of crimes might cause difficulties when estimating the linkage potential of the behavioural domains mentioned above. Second, a substantial decrease in sample size would potentially jeopardize the stability of the models developed. Third, a closer inspection of the pattern of missing data suggested that data were not missing at random and thus, deleting cases with missing data may provide biased estimates for the variables examined in this research (e.g., deleting the cases without useable addresses would mean that a large proportion of the behaviours engaged in by a specific subset of serial sexual assaulters would be excluded from the models [i.e., those who primarily commit their crimes in more public areas]).

Using multiple imputation to estimate the missing values was also deemed an inappropriate strategy for a number of reasons. First, imputing such a large number of dates and addresses seemed improper and the nature of the variables themselves makes this situation relatively unusual (i.e., estimating dates and addresses rather than scores on psychological measures or scales, for example). Second, from a practical perspective, it is very difficult to argue that linkage practitioners would find a model useful where the

majority of the data for some of the variables has been statistically estimated. Moreover, practitioners have argued that spatial and temporal similarity, although empirically powerful, may not be useful linkage features in practice, particularly when it comes to linking interpersonal crimes. For instance, a police investigator may seek out the expertise of a linkage practitioner for a certain cluster of crimes because they are in fact close in time, space, or both. Conversely, practitioners may also be asked to make linkage decisions for crimes that have been committed far apart (and, as previously discussed, there are a certain subset of offenders who travel far distances to their crimes). As such, models that emphasize linking based on short distances may not be the most practical. It is important for research to uncover other reliable indicators of behavioural consistency and distinctiveness that can be more useful to practitioners (e.g., control behaviours).

Given all of these issues, it was decided that the most appropriate course of action was to exclude the temporal and spatial similarity variables from the sexual assault analyses in the current study.

Controlling the impact of prolific offenders. Similar to the break and enter data, the largest proportion of offenders had 2 crimes in their series (33%, $n = 23$). Overall, most offenders had between 2 and 4 crimes (75.4%, $n = 52$). The same outlier approach was used to identify prolific offenders and reduce the number of crimes in their series in an attempt to control their influence on the linking models produced. A total of eight series/offenders were identified as outliers (i.e., prolific offenders) in the serial sexual assault dataset (crimes per series = 10, 11, 12, 15, 18, 19, 29, and 37). The highest non-outlying value was 7. As such, a random sample of 8 crimes was selected from the eight prolific offenders' original crime series. This resulted in a reduced final dataset of 260

sexual assaults that was then used in all subsequent analyses (series ranging in length from 2 to 8 crimes, $M = 3.74$, $SD = 2.00$).[20]

Data Analysis Procedures

Phase 1: Calculation and descriptive analysis of similarity scores. The data coding and preparation processes outlined above resulted in two separate master datasets – one for serial break and enters and one for serial sexual assaults – where the rows represented each individual crime ($N = 170$ break and enters; $N = 260$ sexual assaults) and the columns consisted of information pertaining to each behaviour (coded as 1's and 0's), an *x*-coordinate, a *y*-coordinate, and a date (with the latter 3 columns present in the Saint John break and enter master file only). These master datasets were used to create all possible pairs of crimes prior to calculating behavioural similarity, spatial similarity, and temporal proximity scores.

For each dataset, these crime pairs were then split randomly in half to construct a development and test sample for model validation purposes (see Table 1 below for a breakdown of the sample sizes for each dataset). Over-fitting occurs when a model produced using one sample performs worse when it is applied to new cases (Thomas et al., 2005; Wang, Qin, & Zhang, 2010).[21] As such, models were constructed using the data from the development samples and applied to the crime pairs in the test samples to estimate the extent of over-fitting with the current data. Examining how well the model applies to the test sample provides a less biased index of the model's overall level of

[20] Similar proportions and problems were identified as discussed in the previous section when exploring the pattern in missing data for the date and address variables using this reduced final dataset.

[21] A variety of other methods could be used to validate the models in the current study, including: *k*-fold cross-validation, leave-one-out cross-validation, and bootstrapping (Efron & Tibshirani, 1993; Refaeilzadeh, Tang, & Lui, 2008). Split-half validation was deemed an appropriate method for the current study because the sample sizes are large (e.g., thousands of crime pairs for each crime type) and because this procedure is commonly used in linking research (e.g., Bennell & Canter, 2002; Bennell & Jones, 2005; Tonkin Woodhams, et al., 2012).

predictive accuracy and also offers an indication of how the linking models developed in the current study might perform if applied to new crimes of the same type (Bennell, 2002). For all subsequent analyses, performance indices (e.g., AUCs) of each linking model across the development and test sample were compared to determine the extent to which generalizability (or "shrinkage" in its predictive accuracy from development to test; Thomas et al., 2005) may be an issue. The same development and test samples were used for all statistical models developed.

Table 1. Final sample size breakdowns for the crime pairs constructed using both datasets.

	Serial Break and Enters			Serial Sexual Assaults		
Sample	L	UL	Total	L	UL	Total
Development	80	7,102	7,182	248	16,588	16,834
Test	81	7,102	7,183	247	16,587	16,836
Total	161	14,204	14,365	495	33,175	33,670

Note. L = linked crime pairs; UL = unlinked crime pairs.

Across-crime behavioural similarity scores as well as spatial and temporal proximity scores (break and enter sample only) were then calculated for each crime pair. These scores provided the basis for all subsequent analyses. In line with previous linking studies (e.g., Bennell & Canter, 2002; Bennell et al., 2009; Melnyk et al., 2011; Slater et al., 2015; Tonkin et al., 2008; Tonkin, Woodhams, et al., 2012; Woodhams & Labuschagne, 2012), Jaccard's coefficient (*J*) was used as the behavioural similarity coefficient in the current study.

As previously mentioned, Jaccard's coefficient can range from 0 (i.e., no similar behaviours are displayed across the two crimes) to 1 (i.e., the exact same behaviours are displayed across the two crimes; Bennell & Canter, 2002). For every pair of crimes in each dataset, a *J*-score was calculated for each of the behavioural domains (e.g., target selection, entry behaviours, internal behaviours, property stolen, for the serial break and enter dataset) using the dichotomously coded behaviours. A specially designed computer program called *B-LINK* was used to calculate the *J*-scores for all possible crime pairs in each dataset (Bennell, 2002). *B-LINK* accepts dichotomously coded data as input and provides output files with columns containing the identification number for each crime included in a pair (e.g., 1-1 for offender 1, crime 1), a code indicating whether the crime pair is linked (1) or unlinked (0), and the *J*-score for each crime pair for a particular behavioural domain. These data were then exported into SPSS in order to conduct the main analyses.

Similar to previous linking research, spatial proximity was operationalized as the distance in kilometres between any two crimes (i.e., the ICD; Bennell & Canter, 2002). ICD was calculated for the serial break and enters using the columns containing the geo-coded *x* and *y* coordinates from the master data file. A program called *S-LINK* (Bennell, 2002) was used to calculate ICDs for all possible pairs of break and enters, which works in a similar fashion to *B-LINK*. Inter-crime distances were then exported into SPSS and merged with the file containing the *J*-scores for the break and enters.

Similar to past linking research, temporal proximity was operationalized as the time lag in days between any two crimes (e.g., Markson et al., 2010; Tonkin et al., 2011). Using SPSS syntax, all possible unique pairs of rows from the master serial break and

enter data file were created. The temporal proximity variable was then computed using the CTIME.DAYS function in SPSS Compute, which provided the number of days between the two dates listed for each break and enter pair. This information was then merged with the file containing the *J*-scores for each behavioural domain and the ICD.

The merging process described above resulted in two final datasets used in all subsequent analyses – one for serial break and enters and one for serial sexual assaults – where the rows represented each crime *pair* and the columns consisted of information pertaining to an ID number for the crimes in each pair, linkage status (1 = linked; 0 = unlinked), *J*-scores for each behavioural domain, ICD, and temporal proximity scores (again, with the latter two variables present in the break and enter data only).

Before generating any linking models, initial descriptive analyses were conducted to examine the *J*-scores, ICDs, and temporal proximity scores. Scores for linked and unlinked crime pairs were graphed and assessed for normality, and appropriate statistical tests were used to determine whether linked crime pairs were characterized by significantly larger *J*-scores, smaller ICDs, and smaller temporal proximity scores than unlinked crime pairs.

Phase 2: Developing and evaluating main effects linking models. In order to determine how traditional linking methods perform using the current data, Phase 2 involved the construction of the main effects linking models (using the development samples) and the evaluation of these models (using the test samples) for each crime type separately. As previously mentioned, logistic regression was chosen as the main effects analysis of choice for the sake of consistency with Monahan et al.'s (2001) study and previous linking research (e.g., Bennell & Canter, 2002; Bennell & Jones, 2005; Markson

et al., 2010; Slater et al., 2015; Tonkin, Woodhams, et al., 2012; Tonkin, et al., 2011; Woodhams & Labuschagne, 2012; Woodhams & Toye, 2007). In essence, the models constructed and evaluated in this phase provided the baseline index of linking accuracy in the current study, as they represented the gold standard against which all CT-based models were compared. A number of steps were required to develop these models, each of which are described in more detail below.

Step 1: Logistic regression assumptions. The outcome variable used in the logistic regression analyses consisted of the dichotomous variable "linkage status", reflecting whether a crime pair is linked or unlinked, whereas the predictors consisted of the various linking features (as measured by the continuous *J*-scores, ICDs, and temporal proximity scores). Prior to the main analyses, the data were screened to ensure there were no data entry errors and that the data met the assumptions of logistic regression analysis (Menard, 2002). All data were assessed for: a sufficient ratio of cases-to-predictors, multicollinearity, linearity in the logit, and independence of errors (Field, 2013; Menard, 2002; Tabachnick & Fidell, 2007).

For each predictor variable, a minimum of 10 cases in the smallest group being predicted (i.e., the linked group) is recommended for logistic regression analysis. Models were first assessed to ensure that this case-to-predictor ratio was met (Peduzzi, Conacto, Kemper, Holford, & Feinstein, 1996).

Second, to assess multicollinearity, bivariate correlations were calculated between all predictors in each dataset (Tabachnick & Fidell, 2007). Likewise, a multiple linear regression was run for each dataset that included all predictors in order to obtain collinearity statistics (Field, 2013).

Third, logistic regression assumes that a linear relationship exists between the predictors and the logit transformation of the dependent variable, linkage status (Tabachnick & Fidell, 2007). The Box-Tidwell approach to assessing linearity in the logit was used to assess this assumption, where a multiple logistic regression (enter method) is run that includes each original predictor as well as each predictor's interaction with its natural logarithm (Hosmer & Lemeshow, 2000; Tabachnick & Fidell, 2007). If any of the interaction terms are significant, then the assumption is violated and transformations should be attempted on the predictors in question (Field, 2013; Tabachnick & Fidell, 2007).

Finally, logistic regression also assumes that the cases included in the analysis are independent of one another (Tabachnick & Fidell, 2007). Non-independence can inflate the Type 1 error rate associated with the significance tests of the predictors (Field, 2013; Tabachnick & Fidell, 2007). Although each case entering the logistic regression analyses in the current study represents a unique crime *pair*, the crimes that comprise each pair do overlap with one another. Recognizing that non-independence may be an issue with linking data, Tonkin, Santtila, et al. (2012) systematically examined the impact of constructing logistic regression linking models using a sample of residential burglaries where the crime pairs were either dependent (as is the situation in the current study) or independent of one another. Results of their analyses revealed very minor differences between the linking models produced (and the resulting significance tests) under either methodological condition, suggesting that non-independence is likely not a large issue with linking data of this nature. With that said, more weight was given to the ROC analyses versus the logistic regression significance tests when evaluating the predictive

performance of each linking model in the current research (particularly since significance tests are very highly influenced by large sample sizes; Field, 2013).

Step 2: Simple and forward stepwise logistic regression analyses. After assessing assumptions, simple logistic regression models were constructed using the development sample to provide a direct estimate of the extent to which each linking feature can be used to distinguish between linked and unlinked crime pairs (i.e., to examine the main effect of each linking feature separately). More specifically, using the serial break and enter data as an example, separate simple logistic regression analyses were conducted with the following linking features serving as the predictor variable: (1) target selection characteristics, (2) entry behaviours, (3) internal behaviours, (4) property stolen, (5) ICD, and (6) temporal proximity. These simple logistic regressions were then followed by a forward stepwise logistic regression analysis to determine the best combination of predictors for distinguishing linked from unlinked crimes (i.e., to examine the main effects of multiple linking features collectively). Summary statistics assessing the goodness-of-fit for each model and tests examining the contribution of each predictor to the overall model were examined to provide an initial indication of how well the different linking models performed (Field, 2013; Menard, 2002).[22]

As part of these analyses, predicted probabilities were saved for both the development and test samples, reflecting the likelihood that a crime pair will be classified as linked when using that pair's similarity scores on the predictors included in a particular model (Bennell & Canter, 2002; Menard, 2002). These probabilities were then used to evaluate the level of predictive accuracy achieved by each model in Step 3 below.

[22] However, as mentioned previously, more weight was placed on the subsequent ROC analyses when evaluating each model given the large sample sizes of the current datasets.

Step 3: Evaluating the predictive accuracy of the main effects models. As mentioned earlier, ROC analysis is the preferred method for evaluating the predictive accuracy of traditional main effects linking models (e.g., Bennell & Canter, 2002; Bennell & Jones, 2005; Bennell et al., 2009; Markson et al., 2010; Slater et al., 2015; Tonkin et al., 2008; Tonkin, Santtila, et al., 2012; Tonkin, Woodhams, et al., 2012; Woodhams & Labuschange, 2012). All ROC analyses were performed using SPSS. The data entered into this ROC analysis consisted of the predicted probabilities calculated in the previous step for every pair of crimes, along with the data representing whether these crime pairs are actually linked or unlinked.

As previously mentioned, the ROC curve plots the proportion of hits (i.e., crime pairs correctly identified as linked) against the proportion of false alarms (i.e., crime pairs incorrectly identified as linked) for all possible decision thresholds (Bennell, 2005). In addition to providing a graph that visually displays the ROC curve for each model, the ROC subroutine in SPSS provides the corresponding AUCs and their associated standard errors, *p*-values, and 95% confidence intervals (CIs) surrounding the AUCs. The AUC for a given model reflects the extent to which the similarity scores providing the basis for that particular model overlap across the linked and unlinked distributions (Bennell et al., 2009). As such, the AUC provides a numerical index of the discrimination accuracy achieved with that particular model, often ranging from a substantial amount of overlap between the two distributions (chance levels of discrimination accuracy; e.g., 0.50) to no overlap between the two distributions (perfect discrimination accuracy; e.g., 1.00; Bennell, 2005).

For each logistic regression model developed in the previous analyses, a separate ROC curve was constructed for the development samples and the test samples. For example, 14 ROC curves were constructed for the serial sexual assault data, two curves (development vs. test) for each of the following models: (1) control behaviours, (2) environmental behaviours, (3) escape behaviours, (4) sexual behaviours, (5) style behaviours, (6) victim selection characteristics, and (7) the forward stepwise model. As previously mentioned, constructing separate ROC curves for the development and test samples allowed for an examination of the level of shrinkage that may occur when applying the model to new data.

Similar to Tonkin, Woodhams, et al. (2012), the difference in magnitude between the development and test AUCs for any given model was examined to provide an index of the extent to which these models may (or may not) be over-fitting the development sample data and consequently, whether or not they are likely to generalize well to a new sample of crime pairs. The magnitude of each AUC was also compared to published guidelines to infer the level of predictive accuracy achieved (Swets, 1988). Although most previous linking research (e.g., Melnyk et al., 2011) has compared the overlap in 95% CIs surrounding each AUC to determine whether one linking model results in significantly higher (or lower) discrimination accuracy than another, the current dissertation also conducted significance tests (*z*-tests) to compare AUCs across all models.[23] This is because research has shown that overlapping 95% CIs do not always reflect a non-significant difference between two parameter estimates (see Knezevic, 2008).

[23] The procedure outlined in Hanley and McNeil (1982) was used to compare the difference between AUCs (using the *z*-test calculator available at: http://vassarstats.net/roc_comp.html).

Phase 3: Developing and evaluating the standard CT and ICT models. Based on the consensus of previous linking research discussed earlier (e.g., Bennell & Canter, 2002; Bennell & Jones, 2005; Markson et al., 2010; Tonkin et al., 2008; Tonkin, Woodhams, et al., 2012; Woodhams & Toye, 2007), it was expected that the baseline, or "gold standard", linking models developed in Phase 2 would further demonstrate that it is possible to discriminate between linked and unlinked crimes by relying on the various linking features discussed above. As a result, this next phase attempted to determine whether similar levels of predictive accuracy could be achieved using a CT approach (Monahan et al., 2001; Steadman et al., 2000). As mentioned, given the fact that CTs take an interactive approach to linking (i.e., where different predictors are used to predict linkage status across different crime pairs) this approach may perform better than the one-size-fits-all regression approach (i.e., where the same predictors are used to predict linkage status across all crime pairs) (Monahan et al., 2001; Steadman et al., 2000). The steps taken to develop and evaluate the standard CT and ICT linking models are described in more detail below.

Similar to other studies examining the value of using CTs to make diagnostic decisions (e.g., Hsu & Kang, 2007; Monahan et al., 2001; Silver & Chow-Martin, 2002; Steadman et al., 2000; Sullivan & van Zyl, 2008)**,** the CT linking models were developed using the CHAID algorithm available in SPSS (Kass, 1980). Recall that, using the chi-square test of independence, the CHAID algorithm assesses the relationship of each predictor (e.g., target selection, entry behaviours, internal behaviours, property stolen, ICD, temporal proximity) with the outcome (i.e., linkage status) and splits the crime pairs according to the levels of the predictor that has the most significant chi-square value (i.e.,

the 'best' relationship with linkage status; Magidson, 1993; Monahan et al., 2001). For each resulting sub-group, this procedure is repeated (i.e., each sub-group is further split using the next most significant predictor), until the resulting sub-groups, or end nodes, are as similar as possible on the outcome of interest (i.e., whether they represent linked or unlinked crime pairs).

In terms of the *iterative* aspect of the ICT, this entire process was repeated using only the sub-groups that remain unclassified after the first CT is created, until no further cases could be classified into linked and unlinked sub-groups on the basis of the predictors (Monahan et al., 2001; Steadman et al., 2000). A variety of steps were taken to construct the standard CT and ICT models, which are described below.

Step 1: Assessing CHAID assumptions. Since the CHAID algorithm is non-parametric, it requires fewer assumptions than other analytic approaches, such as linear regression analysis (Hill & Lewicki, 2006). With that said, because CHAID often produces multiple splits at each node, the analysis generally requires a larger sample size than many traditional analytical methods for the results to be meaningful (Kass, 1980; van Diepen & Franses, 2006). As a result, a variety of sample size guidelines have been proposed in the literature, including: at least 1,000 cases in total (Sonquist, Baker, & Morgan, 1971), a minimum of 200 observations per predictor variable (Perrault & Barksdale, 1980; van Middelkoop, Borgers, & Timmermans, 2003), and at least 33 observations per predictor variable (MacLachlan & Johansson, 1981). As previously shown in Table 1, both datasets met all of these sample size recommendations.

Step 2: Setting appropriate parameters for model development. Next, a variety of user-specified decisions were made prior to running the CHAID analyses. These

included: selection of the particular chi-square test used for splitting the data according to the predictors, the level of significance set for these tests, the maximum number of intervals that the continuous predictors can be separated into, the minimum number of cases that must be present in each successive node, and the maximum tree depth allowed (Monahan et al., 2001; SPSS, 2012). In general, most of these decisions were dependent on sample size.

First, in terms of the chi-square test used for determining node splitting, the default chi-square test in SPSS for determining node splitting is the Pearson chi-square; however, the likelihood ratio chi-square test was selected in the current study because it is more robust (SPSS, 2012). Second, although the default significance level for partitioning nodes in SPSS is $p \leq .05$, it was decided to adjust this to $p \leq .01$ in both datasets because the samples are quite large and a more conservative significance level made it less likely that the resulting models would capitalize on chance.[24] Third, very little guidance exists for modifying the maximum number of categories permitted to separate the continuous predictors. As such, the default level of 10 was used for both datasets.[25]

Fourth, the SPSS default for the minimum number of cases included in each node is 100 for parent nodes and 50 for child nodes. Decreasing the minimum number of cases in parent and child nodes is recommended when sample sizes are small (SPSS, 2012); however, the minimum number of cases in each node should not exceed the base rate of linked crime pairs in the sample (Bennell, Woodhams, & Beauregard, 2015). Given that

[24] The default alpha level for the logistic regression analyses in Phase 2 was also changed to $p \leq .01$ to ensure results remain comparable across linking approaches.

[25] Generally, the literature discussing CHAID analyses recommends that parameters are selected that result in a CT that is both manageable in terms of size and makes logical sense (Weller, Harris, Ware, & Jarvis, 2006). As such, it was decided to maintain the default for this particular parameter since no specific guidelines exist in relation to this particular parameter.

the sample sizes are quite large in both datasets, and that the base rate of linked crime pairs in the development samples exceeds the default minimum number of cases specified for child nodes (i.e., 50), it was decided to maintain the defaults levels of 100 and 50 for parent and child nodes, respectively.

Finally, although the SPSS default for tree depth is three, tree depth was set at the number of predictors involved in the analysis. This was done to ensure that each predictor had at least one chance to be included in the CT (Bennell et al., 2015; Tonkin, Woodhams, et al., 2012).

Step 3: Developing the standard CTs. Once the above parameters were set, the standard CTs were constructed by applying the CHAID algorithm to the development samples. It is important to note that, before beginning the CHAID analysis, it is possible to request that the exact same model be simultaneously applied to the test samples while it is being constructed using the development samples. That is, the SPSS CHAID sub-routine provides output that is separated into the CT results for the development samples (reflecting the original models) and the CT for the test samples (reflecting the application of that particular CT model to the test sample cases). This option was selected for all CHAID analyses run in this dissertation.

Step 4: Developing the ICTs. To develop the ICTs, the CHAID algorithm was first applied to the development samples and the end nodes were labelled as linked, unlinked, or unclassified. Using the criteria first outlined by Steadman et al. (2000), the number of linked crime pairs in a given terminal node needed to be: (1) greater than twice the base rate of linked pairs for the node to be labelled a linked sub-group, (2) less than

half the base rate of linked pairs to be labelled an unlinked sub-group, and (3) equal to or in-between those two values to be labelled an unclassified sub-group.

Next, for each dataset, the crime pairs comprising the unclassified sub-groups in the first iteration were pooled together and analyzed a second time using the CHAID algorithm (i.e., a second iteration of the CT was produced; Monahan et al., 2001). The same criteria were then used to classify sub-groups resulting from the second iteration of the CT. If no further sub-groups could be classified as linked or unlinked, the analyses were terminated after the second iteration (and only the one standard CT was produced). In contrast, if additional sub-groups could be classified as linked, unlinked, and unclassified after the second iteration, then a third iteration of the CT was produced using the pooled unclassified sub-groups that remained from iteration two. This continued (e.g., iteration four, iteration five, etc.) until no further cases could be classified.

Since the models applied to the test samples were identical to the models constructed using the development samples, the exact same CT structure was produced at each re-analysis for the test samples (i.e., the same predictors were used to split the sub-groups into the same categories for each predictor across all iterations). The only difference between the development and test sample CTs was the number of linked and unlinked cases contained in each corresponding node. Given this, the labels in the test samples for each sub-group contained in the terminal nodes (i.e., the number nodes classified as linked, unlinked, and unclassified) can differ from the labels given to the nodes in the development samples across each iteration (Bennell et al., 2015). The extent to which the labels/classification of terminal nodes differ from development to test CTs provides an index of the extent to which the two-threshold approach can result in

consistent linking decisions when the CT (and two-threshold approach) is applied to a new sample of crime pairs.

Similar to Phase 2, predicted probabilities reflecting the likelihood that a crime pair will be classified as linked when using the sequence of predictors included in each CT were saved during each successive run of the CHAID algorithm. For the standard CT models, only one set of predicted probabilities was calculated for all crime pairs in the development and test samples and used in subsequent ROC analyses to evaluate the standard CT models. For the ICT models, however, a given crime pair could have more than one predicted probability associated with it (because it was unclassified in earlier iterations). Thus, the predicted probability calculated from the last iteration that each crime pair entered into was used in the ROC analyses to evaluate the ICT models.

Step 5: Evaluating the predictive accuracy of the standard CTs and ICTs. To allow for direct comparisons with the logistic regression models, ROC analysis was again used to evaluate the predictive accuracy of the standard CT and ICT models. The data entered into these analyses consisted of the predicted probabilities calculated above for every pair of crimes, along with the data representing whether the crime pairs are actually linked or unlinked. Similar to Phase 2, separate ROC curves were constructed for the development and test samples to allow for an estimation of over-fitting. The magnitude of the resulting AUCs was also compared to previously mentioned published guidelines to determine the level of predictive accuracy achieved by the standard CT and ICT models (Swets, 1988).

Step 6: Comparing the standard CTs, ICTs and logistic regression models. A final component of this phase was to compare the predictive accuracy of the standard

CTs, ICTs, and logistic regression models. This was done to determine the extent to which the standard CT and/or the ICT models are comparable to the logistic regression models at correctly classifying linked and unlinked crimes. The AUCs achieved using the standard CT and ICT method were compared to the AUC achieved using the logistic regression method. Significance tests (*z*-tests) were also conducted to determine whether any statistically significant differences existed between these linking approaches. Finally, significance tests were also conducted to compare the AUCs for each approach across the development and test samples in order in order to determine whether certain linking approaches are potentially less generalizable than others.

Likewise, to provide a descriptive comparison of the ability of the different linking approaches to classify crime pairs into definitive groups, the number of cases classified as linked, unlinked, and unclassifiable was calculated for each model (Monahan et al., 2001). That is, using the two thresholds for classifying linked and unlinked crimes (i.e., twice and half the base rate of linked crimes in the test samples), the frequency and total percentage of crime pairs classified into each of the sub-groups by the logistic regression models versus the standard CT and ICT models was compared. Similarly, the percentage of cases *correctly* classified (as linked, unlinked, and overall) was calculated and compared across each model.

Phase 4: Developing and evaluating the multiple CT/ICT linking models. Phase 4 also adopted the methodology of Monahan et al. (2001) to examine the extent to which constructing multiple CT/ICT models to provide a combined estimate of linkage for each crime pair could potentially enhance the predictive accuracy of the CT-based linking approach. The procedure generally comprised an extension of Phase 3 in that it

involved the construction of multiple CT/ICT models using the development samples and the evaluation of these models using the test samples. With that said, Phase 4 also included some additional intermediate components: the steps taken to combine the information obtained from the multiple CT/ICT models (Monahan et al., 2000; Silver & Chow-Martin, 2002).[26]

Step 1: Constructing multiple CT/ICT models. Analogous to Phase 3, the multiple CT/ICT models were developed using the CHAID algorithm (Kass, 1980). Since CHAID assumptions were assessed in Phase 3, they were not re-assessed. Similarly, the parameters for model construction were set in the same manner as outlined in Phase 3. The multiple models were constructed by forcing a different predictor variable as the initial splitting variable in the first iteration of each CHAID analyses (Monahan et al., 2001). As such, the number of CT/ICT models that were constructed was determined by the number of predictor variables included in the dataset. For example, in the case of the serial sexual assault dataset, six CTs were constructed, with the following variables being forced as the first splitting variable in Iteration 1 of each respective tree: (1) control behaviours, (2) environmental behaviours, (3) escape behaviours, (4) sexual behaviours, (5) style behaviours, and (6) victim selection characteristcs. The same thresholds used in Phase 3 to identify linked, unlinked, and unclassifiable nodes were used, and unclassifiable nodes were pooled and entered into successive iterations of each CHAID analysis. Finally, similar to Phase 3, each of the

[26] As previously mentioned, this approach was labeled the multiple CT/ICT models approach because the presence/absence of second iterations of the multiple CTs will determine whether multiple standard CTs, multiple ICTs, or a mixture of both standard CTs and ICTs are combined to create the multiple model approach to linking.

CT/ICT models were simultaneously applied to the test samples while being constructed using the development samples.

Step 2: Combining the multiple CT/ICT models. For each crime type, the results were combined by creating a composite linkage score for each crime pair based on how they were classified in each CT/ICT model. As outlined by Monahan et al. (2001), three steps were taken to calculate these composite linkage scores.

First, for each of the single CT/ICT models (i.e., six in the case of the serial sexual assault data), the terminal nodes (and as a result, the crime pairs in those nodes) in each iteration were labelled as linked (greater than twice the base rate of linked crimes), unlinked (less than half the base rate of linked crimes), or unclassifiable (equal to or in-between these two thresholds) (Monahan et al., 2001).

Second, a scoring procedure proposed by Ohlin (1951) and Burgess (1928) was used to provide every crime pair with a score for each model that reflected their sub-group classification for that particular model (i.e., as linked, unlinked, or unclassifiable). That is, for each CT/ICT model developed, every crime pair was given a score of -1 (if it was classified as unlinked on that particular model), 0 (if it was deemed unclassifiable on that particular model), or +1 (if it was classified as linked on that particular model).

Finally, once the crime pairs were scored on each model, a composite linkage score for every crime pair was calculated by summing across their scores for each CT/ICT model.[27] More specifically, using the serial sexual assault data as an example, where six CT models were constructed, each crime pair received six scores. As such, the composite linkage scores had a possible range of -6 (indicating that a crime pair was

[27] Similar to Monahan et al. (2001), to ensure that a sufficient level of internal reliability existed, Cronbach's alpha was calculated between the scores produced by each of the single CT/ICT models prior to summing the scores to create the composite score.

classified as unlinked by all models) to +6 (indicating that a crime pair was classified as linked by all models). These scores were then used to evaluate the performance of the multiple CT/ICT model for linking the serial sexual assaults.

Step 3: Constructing the empirically optimal multiple CT/ICT models. Based on the fact that they are the product of multiple estimates of linkage, it is likely that the multiple CT/ICT models will result in improved predictive accuracy above that achieved by the single standard CT or ICT models developed in Phase 3 (Monahan et al., 2001; Silver & Chow-Martin, 2002). From a practical perspective, it is worthwhile to examine whether including information from all CT/ICT models in the calculation of the composite linkage scores is in fact necessary to achieve this improved level of linking accuracy (Monahan et al., 2001). In other words, it may be possible to achieve the same level of accuracy while relying on fewer models (i.e., by asking a smaller number of questions about a particular crime pair to arrive at an accurate linking decision). Following Monahan et al. (2001), a forward stepwise logistic regression was therefore conducted with the linkage scores (for each crime pair) from a single CT/ICT model serving as the predictor variables and whether the crime pairs are actually linked or unlinked serving as the outcome variable (Monahan et al., 2001; Silver & Chow-Martin, 2002). Composite linkage scores for all crime pairs were then recalculated using only the scores from the models included as significant predictors of linkage status in the forward stepwise regression analysis. This revised model was labelled the empirically optimal multiple CT/ICT model.

Step 4: Evaluating the original multiple CT/ICT models and empirically optimal multiple CT/ICT models. ROC analysis was then used to evaluate the predictive accuracy

of the two multiple CT/ICT models (the original models including all model scores in the composite calculation and the empirically optimal models containing a subset of the model scores in the composite calculation). In contrast to previous phases, however, the data entered into the ROC analyses consisted of the composite linkage scores (original vs. empirically optimal) for every pair of crimes (rather than their predicted probabilities), along with the data representing whether these crime pairs are actually linked or unlinked.

Similar to previous phases, separate ROC curves were constructed for the development and test samples. The magnitude of the resulting AUCs was compared to the previously mentioned published guidelines to determine the level of predictive accuracy achieved by both composite score models (Swets, 1988). The AUCs associated with the empirically optimal multiple CT/ICT models were compared to the AUCs associated with the original multiple CT/ICT models to ensure that their levels of predictive accuracy are comparable (i.e., that $p > .05$ for the significance tests). Likewise, the magnitude of shrinkage from development to test sample was calculated and statistically compared (z-tests) for both composite score models.

Step 5: Comparing the performance of all linking models. As a final component of the current study, the relative predictive accuracy of the logistic regression models, standard CT models, single ICT models, and multiple CT/ICT models was compared for both samples of crime. To do this, all possible pairwise comparisons of the model AUCs were performed in the same manner as described in all previous phases (z-tests).

CHAPTER 6

Linking Serial Break and Enters: Results

Phase 1: Calculation and Descriptive Analysis of Similarity Scores

As previously explained, behavioural similarity between all crime pairs was calculated using Jaccard's (*J*) coefficient for each behavioural domain separately. It is common practice in linking research to remove variables with lower frequencies of occurrence prior to calculating the various *J*-scores, ranging anywhere from variables that occur in less than 1% of crimes (e.g., Santtila et al., 2005) to variables that occur in less than 10% of crimes (e.g., Tonkin & Woodhams, 2015; Winter et al., 2013).

The rationale often provided for doing this is that low frequency behaviours will not be particularly useful for linking most crimes (e.g., Tonkin & Woodhams, 2015). Given that research has not systematically examined this issue, the empirical impact of removing variables occurring at various frequency intervals was assessed prior to conducting any main analyses. This was done by calculating *J*-scores when variables that occurred in less than 1 percent, 5 percent, 10 percent, 15 percent, and 20 percent of the crimes in the sample were progressively removed from the calculation (see Appendix A for a list of the variables included in each domain and their frequency of occurrence for the final data where prolific offenders were controlled).

As shown in Table 2, the impact of removing lower frequency variables on the calculation of *J*-scores was consistent across all behavioural domains. Median *J*-scores tended to increase as more behaviours are removed from the calculation (with the

exception of property stolen, where median *J*'s decreased). Likewise, standard deviations tended to increase as more behaviours are removed.[28]

Table 2. Median *J*-scores and standard deviations for all break and enter pairs when variables at various frequency intervals were removed from the calculations.

	Behavioural Domain							
	Target Selection		Entry Behaviours		Internal Behaviours		Property Stolen	
Removal Interval	Mdn	*SD*	Mdn	*SD*	Mdn	*SD*	Mdn	*SD*
None	.36	.21	.33	.23	.27	.21	.11	.20
≤ 1 %	.36	.21	.33	.23	.27	.22	-	-
≤ 5 %	.38	.21	.33	.23	.29	.22	.11	.22
≤ 10 %	.40	.21	.33	.22	.30	.23	.11	.23
≤ 15 %	.40	.22	.40	.24	.33	.24	.00	.23
≤ 20 %	.44	.23	.40	.25	-	-	.00	.24

Note. Missing cells correspond to instances where the domain did not contain any variables to remove at that particular frequency interval (e.g., the internal behaviours domain did not contain any variables with frequencies between 16 and 20 percent).

To determine the impact of variable removal on the predictive accuracy of the behavioural domains, the *J*-scores for each domain at the various intervals displayed in Table 2 were entered into separate ROC analyses. A forward stepwise logistic regression model was also constructed at each frequency interval and the predicted probabilities

[28] Medians (instead of means) were reported here because Kolmogorov-Smirnov tests of normality confirmed that all distributions were significantly different from normal (*D*'s ranged from 0.08 to 0.29, all *p*'s <.001). Medians and standard deviations for each behavioural domain were also calculated separately for linked and unlinked crime pairs. The same patterns as presented here emerged across both sets of crime pairs.

from each stepwise model were entered into separate ROC analyses to determine the impact of removing variables on the predictive accuracy of the model containing multiple behavioural features.[29] As displayed in Table 3 below, AUCs tended to remain the same or decrease as more behaviours were removed from the calculation of behavioural similarity (with the exception of the entry behaviours domain, which tended to stay the same or increase slightly). Importantly, any differences in predictive accuracy were not statistically significant (i.e., the confidence intervals overlap across all models for each domain and significance tests revealed that all *p*'s >.05). Given these results, all variables were retained in each domain when calculating similarity scores in the current study.

[29] Some domains did not have variables to remove at certain frequency intervals (e.g., the internal behaviours domain did not contain any variables with frequencies between 16 and 20 percent). As such, for these domains, the stepwise models included the most recently calculated *J*-scores (e.g., the internal behaviours domain at the 15% removal interval). This was done because the goal was to determine how *all* domains would perform together when using a particular removal criterion (e.g., removing variables occurring in less than 20% of cases), regardless of whether a particular domain had variables to remove at that level.

Table 3. Predictive accuracies of the *J*-scores for each behavioural domain and the stepwise logistic regression model across all removal intervals for the serial break and enter data.

	Model									
	Target Selection		Entry Behaviours		Internal Behaviours		Property Stolen		Forward Stepwise LR	
Removal Interval	AUC (*SE*)	95% CI	AUC (*SE*)	95% CI	AUC (*SE*)	95% CI	AUC (*SE*)	95% CI	AUC (*SE*)	95% CI
None	.57*** (.02)	.53 – .62	.53 (.02)	.49 – .58	.60*** (.02)	.55 – .64	.54 (.02)	.49 – .59	.62*** (.02)	.57 – .66
≤ 1 %	.57*** (.02)	.53 – .62	.53 (.02)	.49 – .58	.60*** (.02)	.55 – .65	-	-	.62*** (.02)	.57 – .66
≤ 5 %	.58*** (.02)	.53 – .62	.53 (.02)	.49 – .58	.60*** (.02)	.56 – .65	.54 (.02)	.49 – .59	.62*** (.02)	.57 – .66
≤ 10 %	.57*** (.02)	.53 – .62	.54 (.02)	.49 – .58	.59*** (.02)	.55 – .64	.54 (.02)	.50 – .59	.61*** (.02)	.57 – .66
≤ 15 %	.57** (.02)	.52 – .61	.54 (.02)	.49 – .59	.59*** (.02)	.54 – .64	.53 (.02)	.49 – .58	.61*** (.02)	.56 – .65
≤ 20 %	.56** (.02)	.52 – .61	.54 (.02)	.49 – .58	-	-	.54 (.02)	.49 – .58	.61*** (.02)	.56 – .65

Note. LR = logistic regression.
**$p < .01$. ***$p < .001$.

Final descriptive statistics for linked and unlinked break and enters are displayed in Table 4 as a function of each individual domain included in subsequent analyses.[30] These distributions were also plotted graphically in Figure 4.[31] On average, linked crime pairs had larger *J*-scores across all behavioural domains than unlinked crime pairs (with the exception of the entry behaviours domain). Linked crime pairs also had smaller ICDs and temporal proximity scores than unlinked crime pairs. However, as also demonstrated, substantial overlap exists in the linked and unlinked distributions across all predictors, indicating that perfect discrimination accuracy is unlikely using the current sample.

Table 4. Descriptive statistics for all linked and unlinked serial break and enters.

	Mdn (*SD*)		Range	
Domains (Measurement)	L	UL	L	UL
Target Selection *(J)*	.40 (.23)	.36 (.21)	.08 – 1.00	.00 – 1.00
Entry Behaviours (*J*)	.33 (.25)	.33 (.22)	.00 – 1.00	.00 – 1.00
Internal Behaviours (*J*)	.38 (.24)	.27 (.21)	.00 – 1.00	.00 – 1.00
Property Stolen (*J*)	.11 (.25)	.10 (.20)	.00 – 1.00	.00 – 1.00
ICD (kms)	1.88 (3.26)	3.59 (3.57)	0.00 – 15.94	0.00 – 22.66
Temporal Proximity (days)	16.00 (689.83)	1,527.00 (1,237.97)	0.00 – 2,689	0.00 – 4,996

Note. *J* = Jaccard's coefficient; L = linked crime pairs; UL = unlinked crime pairs.

[30] Again, medians (instead of means) are reported here because all distributions (linked and unlinked) were significantly different from normal (*D*'s ranged from and all *p*'s <.001), with the exception of the internal behaviours distribution for linked crime pairs ($D = 0.07$, $p = .081$). However, for the sake of consistency, the median was also reported for this distribution.

[31] The unlinked distributions in Figure 4 are based on a random sample of unlinked crime pairs that is equal in size to the total number of linked crime pairs in the sample ($n = 161$) because paired samples significance tests were ultimately conducted to examine differences between distributions.

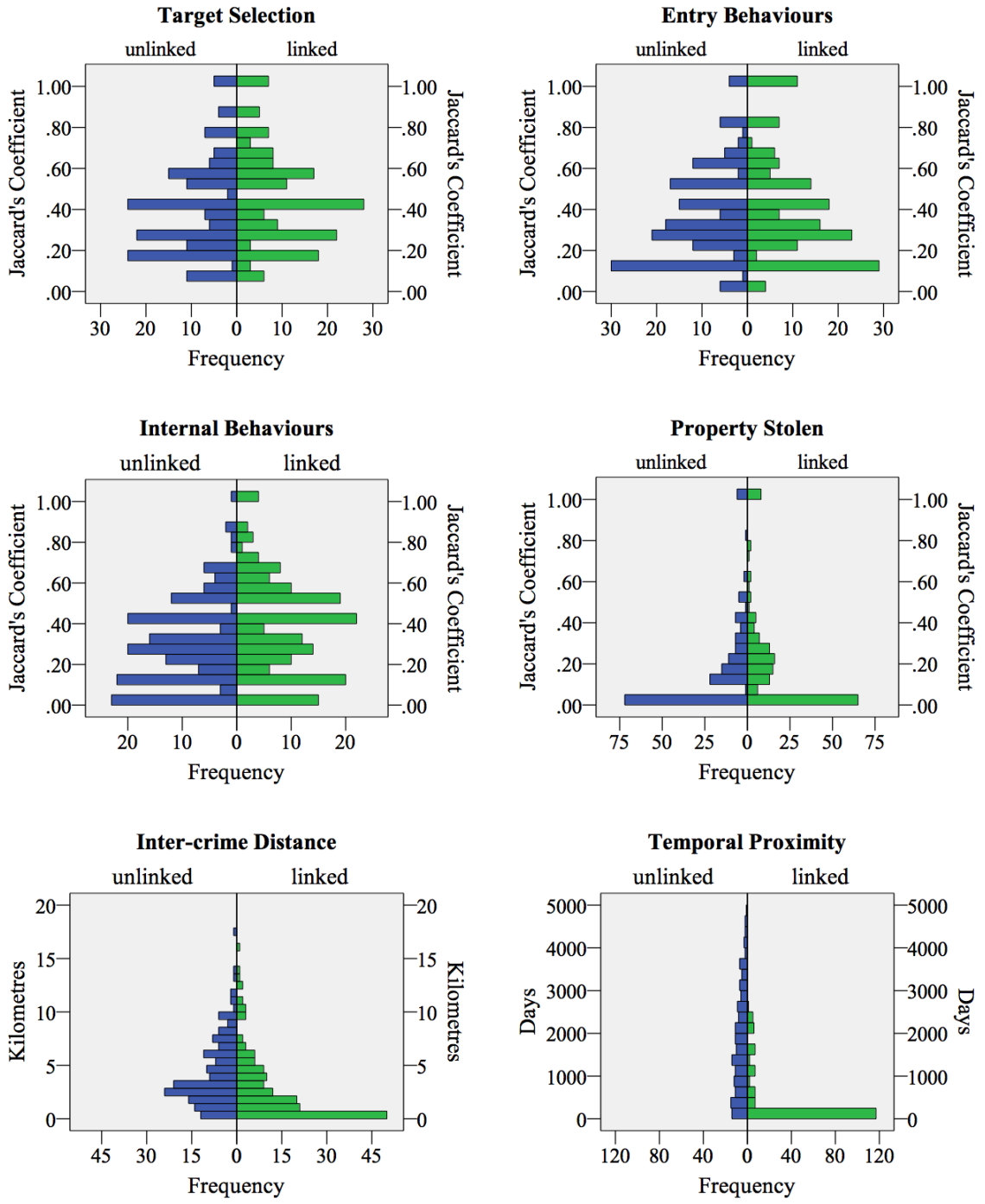


Figure 4. Frequency distributions displaying the range of scores for all linked (n = 161) and an equal random sample (n =161) of unlinked break and enters for each variable.

Significance tests were then carried out to examine these differences between linked and unlinked crime pairs. Similar to previous linking research, paired samples tests were conducted given the dependent nature of the linked and unlinked crime pairs (e.g., Markson et al., 2010; Woodhams & Labuschagne, 2012). Consequently, a random sample of 161 unlinked crime pairs was extracted to facilitate paired analyses. As evident in Figure 4, Kolmogorov-Smirnov tests of normality confirmed that all distributions were significantly different from normal at the $p < .05$ level (D's ranged from 0.07 to 0.32; p's ranging from <.001 to .033), with the exception of the linked distribution for the internal behaviours domain $D(161) = 0.07$, $p = .081$. As a result, a non-parametric equivalent to the paired samples t-test, the Wilcoxon signed-rank test, was conducted on all domains.

Results of the Wilcoxon tests are presented below in Table 5. J-scores for the internal behaviours domain were significantly larger for linked crime pairs than unlinked crime pairs. Likewise, ICDs and temporal proximity scores were significantly smaller for linked crime pairs than unlinked crime pairs. The differences between linked and unlinked crime pairs on all other behavioural domains were non-significant at the Bonferroni corrected significance level ($p<.008$), although the target selection domain was significant at the $p<.05$ level. Effect sizes (r index) were also calculated for each test.[32] As per Cohen's (1988, 1992) guidelines, most effect sizes were in the small range ($r <.30$), with the exception of ICD (medium effect; $r = .33$) and temporal proximity (large effect; $r = .74$).

[32] $r = z/\sqrt{N}$ (Field, 2013; Rosenthal, 1991); $r <.30$ = small effect size; r between .30 and .50 = moderate effect size; $r > .50$ = large effect size (Cohen, 1988, 1992; Field 2013).

Table 5. Results of the non-parametric comparisons of similarity scores for linked and unlinked break and enters across each linking feature.

	Median				
Domains (Measurement)	L	UL	Wilcoxon (*z*)	*p*-value	Effect Size (*r*)
Target Selection *(J)*	.40	.36	-2.02	.044	.16
Entry Behaviours (*J*)	.33	.33	-0.67	.501	.05
Internal Behaviours (*J*)	.38	.27	-3.31*	.001	.26
Property Stolen (*J*)	.11	.13	-0.88	.381	.07
ICD (kms)	1.88	3.33	-4.25*	<.001	.33
Temporal Proximity (days)	16.00	1,596	-9.41*	<.001	.74

Note. *J* = Jaccard's coefficient; L = linked crime pairs; UL = unlinked crime pairs; $r = z/\sqrt{N}$ (Field, 2013; Rosenthal, 1991).
*p <.008 (Bonferroni significance correction = .05/6 =.008).

Phase 2: Developing and Evaluating the Main Effects Linking Models

Next, the main effects logistic regression models (including and excluding temporal proximity as a predictor) were developed and evaluated using the steps outlined in the methodology chapter.

Logistic regression assumptions. As mentioned, a minimum of 10 linked crime pairs is required for each predictor variable included in the logistic regression analyses. This requirement was met for all logistic regression models developed. A maximum of six predictors were entered into the stepwise logistic regression model, requiring a minimum number of 60 linked cases (linked n = 80 in the current data). Bivariate correlations were also calculated between all predictors to assess multicollinearity. As shown in Table 6, all correlations were in the small range, with most near zero (r's ranging from -.03 to .15), suggesting that multicollinearity and singularity were not a

concern with the serial break and enter data (Tabachnick & Fidell, 2007).[33] A linear regression was also run where all predictors were entered into the model in order to obtain collinearity statistics (Field, 2013). These indices confirmed that multicollinearity was not present in the data - all VIFs were less than 10 (Bowermann & O'Connell, 1990). Specifically, VIFs ranged from 1.00 (temporal proximity) to 1.05 (internal behaviours domain). Likewise, all tolerance values were well above Menard's (1995) guideline of .02 for problematic tolerance levels (ranging from .95 for internal behaviours to 1.00 for temporal proximity). As such, multicollinearity was not an issue with the break and enter data (Tabachnick & Fidell, 2007).

Table 6. Correlations between all behavioural domains included in the serial break and enter analyses.

Variable	1	2	3	4	5
1. Target Selection	-				
2. Entry Behaviours	.05*	-			
3. Internal Behaviours	.15*	.04*	-		
4. Property Stolen	.06*	-.01	.15*	-	
5. ICD	-.02	-.03*	.09*	.03*	-
6. Temporal Proximity	-.01	.00	.01	-.01	.04*

*p <.003 (Bonferroni significance correction = .05/15 =.003).

As previously explained, the Box-Tidwell approach was used to assess linearity in the logit, where a multiple logistic regression (enter method) is conducted that includes each original predictor as well as each predictor's interaction with its natural logarithm

[33] A similar pattern was found when the correlations were examined for each of the behavioural domains within the linked and unlinked crime pairs separately.

(Hosmer & Lemeshow, 2000; Tabachnick & Fidell, 2007). Prior to creating the interaction terms, a constant of 2 was added to the raw variables to calculate its natural logarithm given that zero values were present for most predictors (Tabachnick & Fidell, 2007). Table 7 presents the relevant output testing the linearity in the logit assumption for each predictor using all cases in the dataset. As shown, property stolen, ICD, and temporal proximity violated the linearity in the logit assumption (i.e., the interaction terms were significant at $p<.05$).[34]

Table 7. Interaction terms testing the linearity in the logit assumption for the serial break and enter data.

Interaction Term	B (*SE*)	Wald (*df*)	*p*-value
Target Selection x Ln	0.89 (3.65)	0.06 (1)	.808
Entry Behaviours x Ln	1.95 (3.02)	0.42 (1)	.518
Internal Behaviours x Ln	0.09 (3.13)	0.01 (1)	.978
Property Stolen x Ln	5.35 (2.69)	3.95 (1)	.047
ICD x Ln	-0.06 (0.01)	21.64 (1)	<.001
Temporal Proximity x Ln	-2.67e-4 (2.50e-5)	111.55 (1)	<.001

In an attempt to address these violations, square root, logarithmic, and inverse transformations were applied to the three offending variables and a number of simple and forward stepwise logistic regression models were developed using the untransformed and

[34] Similar results were obtained when the linearity in the logit assumption was tested on the development and test samples separately; however, the property by Ln interaction term was not significant in the development sample ($p = .534$).

transformed data.[35] Using the predicted probabilities calculated for these various models, ROC analysis was then conducted in order to compare the predictions made by the logistic regression models using the transformed and untransformed variables.

As expected, the parameter estimates varied across all models; however, the significance of each predictor's coefficient (Wald's test) did not change from significant to non-significant across any of the models (and the R^2 values varied minimally). As demonstrated in Table 8, the AUCs for the simple logistic regression models for each of the offending variables were exactly the same across the transformed and untransformed data. For the stepwise logistic regression models, target selection, internal behaviours, ICD, and temporal proximity were included in the untransformed model. In contrast, target selection was excluded from all transformed models (although internal behaviours, ICD, and temporal proximity remained in all transformed stepwise models).[36] Although the AUCs varied slightly across the stepwise models using transformed versus untransformed data (see Table 8), these differences were not significant (i.e., the confidence intervals overlapped substantially and significance tests revealed that all p's >.05).[37] The decision was made to retain the untransformed variables largely because the AUCs did not vary significantly across the models using transformed versus untransformed data. However, this decision also ensured that the interpretability of the CT models was maintained, given that the same data must be used in the logistic

[35] A constant of 2 was added to each variable prior to applying the transformations given the occurrence of zero scores.
[36] Tables displaying the logistic regression results associated with the AUCs in Table 8 were not included for brevity. Likewise, the AUCs were focused on here given the ultimate concern is whether the predictive accuracy of the different models varies significantly as a function of using the transformed versus untransformed data.
[37] The results presented reflect models developed using all cases; however, similar results were found when the models were constructed using the development sample and applied to test sample (i.e., no significant differences were found between models when using transformed versus untransformed data).

regression and CHAID analyses to allow for a proper comparison of statistical approaches.

Table 8. Results of ROC analyses comparing the predictions made by logistic regression models constructed using the transformed versus untransformed variables for the serial break and enter data.

Model	AUC (*SE*)	95% CI
Property Stolen Only		
Untransformed	.54 (.02)	.49 – .59
Square root	.54 (.02)	.49 – .59
Logarithmic	.54 (.02)	.49 – .59
Inverse	.54 (.02)	.49 – .59
ICD Only		
Untransformed	.69 (.02)	.64 – .74
Square root	.69 (.02)	.64 – .74
Logarithmic	.69 (.02)	.64 – .74
Inverse	.69 (.02)	.64 – .74
Temporal Proximity Only		
Untransformed	.88 (.02)	.84 – .91
Square root	.88 (.02)	.84 – .91
Logarithmic	.88 (.02)	.84 – .91
Inverse	.88 (.02)	.84 – .91
Forward Stepwise		
Untransformed	.87 (.02)	.84 – .90
Square root	.88 (.02)	.85 – .91
Logarithmic	.88 (.02)	.85 – .92
Inverse	.85 (.02)	.81 – .89

Note. all *p*'s <.001, except property stolen AUCS (all *p*'s >.05).

Simple and forward stepwise logistic regression analyses. Next, simple logistic regression models were constructed to examine the predictive accuracy of each linking feature separately (using the development cases only). The results of these single linking feature models are presented in Table 9. As shown, all model chi-square tests were statistically significant, suggesting that each predictor was able to distinguish linked from unlinked crime pairs at a level greater than the constant only model. The signs of the regression coefficients for each predictor were also in the expected direction, suggesting higher degrees of behavioural similarity, shorter distances, and fewer days between crimes for linked crime pairs relative to unlinked crime pairs. Wald statistics indicated that all regression coefficients were significantly different from zero, further suggesting that all linking features were significant predictors of linkage status. Both R^2 indices (Nagelkerke's and Hosmer and Lemeshow's) indicated that temporal proximity was the best predictor relative to all other predictors.[38]

[38] As explained by Field (2013), the Hosmer and Lemeshow R_L^2 indicates how much the fit of the model improves as a result of the predictors in comparison to a baseline model that solely includes the constant. The R_L^2 can range from 0 (indicating that the predictors do not do a good job at predicting the outcome) to 1 (indicating that the predictors are able to perfectly predict the outcome). The R_L^2 tends to be preferred over other R^2 indices because it does not depend on sample size (unlike R_N^2) and it has been argued that it is more conceptually similar to the R^2 index used in linear regression than other options (Field, 2013; Menard, 2002). Likewise, Menard (2000) found that many other R^2 indices for logistic regression models increase as the base rate of the target outcome increases, however, the R_L^2does not.

Table 9. Results of separate simple logistic regression analyses for each predictor included in the serial break and enter sample.

Model	Constant (*SE*)	*B* (*SE*)	Wald	χ^2	R_N^2	R_L^2
Target Selection	-5.19 (0.25)	1.67 (0.49)	11.47**	10.82**	.01	.01
Entry Behaviours	-4.85 (0.22)	0.93 (0.46)	4.16*	3.96*	.01	4.51e-3
Internal Behaviours	-5.12 (0.22)	1.93 (0.47)	16.95***	15.89***	.02	.02
Property Stolen	-4.65 (0.15)	0.96 (0.46)	4.37*	3.86*	.01	4.39e-3
ICD	-3.79 (0.18)	-0.19 (0.05)	15.73***	20.83***	.03	.02
Temporal Proximity	-2.44 (0.16)	-2.85e-3 (3.46e-4)	67.94***	185.81***	.22	.21

Note. χ^2 = model chi-square; R_N^2 = Nagelkerke index; R_L^2 = Hosmer and Lemeshow's index; *df* = 1 for all significance tests.
*p<.05. **p<.01. ***p<.001.

A forward stepwise logistic regression model was then constructed using the development sample to determine the optimal combination of variables for predicting linkage status using the main-effects linking approach. As shown in Table 10, the stepwise procedure proceeded through four steps, including temporal proximity, internal behaviours, ICD, and target selection characteristics in the final model. The model chi-square test was statistically significant, suggesting that the predictors included in the model were able to distinguish linked from unlinked crime pairs at a level greater than the constant only model. Wald statistics for all regression coefficients were also significant, suggesting that all linking features included in the stepwise model were significant predictors of linkage status. As shown in Table 10, an R_L^2 of .25 was found for the stepwise logistic regression model, indicating that the linking features included in the stepwise model improve the prediction of linkage status to a moderate degree.

Table 10. Results of the forward stepwise logistic regression analysis performed on the break and enter development sample when temporal proximity was included in the model.

Model	B (*SE*)	Wald (*df*)	χ^2 (*df*)	R_N^2	R_L^2
Stepwise			223.76 (4)	.27	.25
Temporal Proximity	-2.79e-3 (3.42e-4)	66.38 (1)			
Internal Behaviours	1.93 (0.50)	14.72 (1)			
ICD	-0.15 (0.05)	11.85 (1)			
Target Selection	1.32 (0.51)	6.59 (1)			
Constant	-3.12 (0.35)	78.90 (1)			

Note. χ^2 = model chi-square; R_N^2 = Nagelkerke index; R_L^2 = Hosmer and Lemeshow's index; all *p*'s <.001 with the exception of the target selection coefficient (*p* =.01).

Next, a second forward stepwise logistic regression model was constructed excluding temporal proximity as a predictor variable. Recall that a large date range was used in the current study in order to obtain an adequate sample size (crimes committed between 2000 and 2013). As shown earlier in Table 4, the median number of days between unlinked crime pairs was 1,527 days. Most linking studies examining temporal proximity as a predictor have sampled crimes over a shorter timeframe, ranging from 1 to 5 years (e.g., Burrell et al., 2012, 2015; Davies, Tonkin, Bull, & Bond, 2012; Markson et al., 2010; Tonkin, Woodhams, Bull, Bond, & Palmer, 2011; Tonkin & Woodhams, 2015).[39] The median number of days between unlinked crime pairs in these studies is

[39] The exception to this is Tonkin, Santtila, et al. (2012), where the sample was comprised of Finnish residential burglaries committed between 1990 and 2001; however, the median number of days between crimes for linked pairs and unlinked pairs was not reported in this study.

therefore much smaller (e.g., ranging from 83.50 to 522 days).[40] Consequently, it is possible that the discriminative value of temporal proximity in the current dataset is inflated. As such, all multivariable models were also developed without temporal proximity.

As shown in Table 11, the analysis proceeded through three steps, with the following predictors included in the final model: internal behaviours, ICD, and target selection (i.e., entry behaviours and property stolen were excluded from the stepwise model). The model chi-square test was statistically significant, suggesting that the predictors included in the model were able to distinguish linked from unlinked crime pairs at a level greater than the constant only model. Wald statistics for all predictors were also significant, suggesting that all linking features included in the model were significant predictors of linkage status. However, both Nagelkerke's R_N^2 and Hosmer and Lemeshow's R_L^2 values suggested that a relatively low level of predictive accuracy could be achieved based on this stepwise model.

[40] Medians were not reported in Davies et al. (2012).

Table 11. Results of the forward stepwise logistic regression analyses performed on the break and enter development sample when temporal proximity was excluded from the model.

Model	B (*SE*)	Wald (*df*)	χ^2 (*df*)	R_N^2	R_L^2
Stepwise			46.28 (3)	.06	.05
Internal Behaviours	1.84 (0.47)	15.18 (1)			
ICD	-0.20 (0.05)	17.12 (1)			
Target Selection	1.41 (0.50)	7.92 (1)			
Constant	-4.99 (0.33)	235.66 (1)			

Note. χ^2 = model chi-square; R_N^2 = Nagelkerke index; R_L^2 = Hosmer and Lemeshow's index; all *p*'s <.001 with the exception of the target selection coefficient (*p* =.005).

ROC analyses. The predicted probabilities from the simple and stepwise logistic regression models presented in Tables 9, 10, and 11 were then entered into separate ROC analyses to compare the predictive accuracy of the various models produced. Likewise, as explained previously, separate ROC curves were constructed for the development and test samples in order to examine the shrinkage in the AUC and provide an estimate of the robustness of the logistic regression models. The ROC curves are visually displayed in Figure 5 and the AUCs and associated confidence intervals are presented in Table 12 for both the development and test samples.

Using Swets' (1988) guidelines, moderate-to-high levels of predictive accuracy were achieved for the model including temporal proximity across both the development and test samples. All other single linking feature models were associated with low levels of linking accuracy in both the development and test samples (with the exception of the ICD AUC in the test sample, which was moderate: AUC = .71). Although both stepwise

models (including and excluding temporal proximity as a predictor) achieved moderate levels of predictive accuracy across the development and test samples, a significantly higher level of accuracy was observed for the stepwise model including temporal proximity as a predictor than the stepwise model excluding temporal proximity as a predictor in both the development sample ($z = 4.66$, $p < .001$) and test sample ($z = 3.75$, $p < .001$).

For the development sample, significance tests indicated that there were no significant differences between the AUCs for the top two models (temporal proximity and the stepwise model including temporal proximity; p's $> .05$); however, the AUCs for these two models were significantly higher than the AUCs for the remaining models (z's ranging from 4.66 to 8.64, all p's $< .001$). Likewise, the stepwise model excluding temporal proximity had a significantly higher AUC than the models comprised of entry behaviours ($z = 2.98$, $p = .003$), property stolen ($z = 3.20$, $p = .001$), and target selection characteristics ($z = 2.33$, $p = .012$). Finally, the ICD only model had a significantly higher AUC than the models comprised of entry behaviours ($z = 2.33$, $p = .012$) and property stolen ($z = 2.54$, $p = .011$). No other AUCs were significantly different from one another (p's $> .05$). A similar pattern of results was found in the test sample; however, the stepwise model excluding temporal proximity also had a significantly higher AUC than the internal behaviours model ($z = 2.78$, $p = .005$) and the ICD model had a significantly higher AUC than the models containing target selection behaviours ($z = 3.22$, $p = .001$) and internal behaviours ($z = 3.22$, $p = .001$).

When examining the pattern of AUCs across the development and test samples, slight shrinkage was evident across all models, with the exception of the ICD only model,

where the AUC increased by .04. The level of shrinkage was not statistically significant across any of the remaining models (shrinkage ranged from .01 to .08 and all confidence intervals overlapped from development to test and all *p*'s >.05 for significance tests of the AUCs). As such, all logistic regression models seemed to generalize well to the test sample crime pairs.

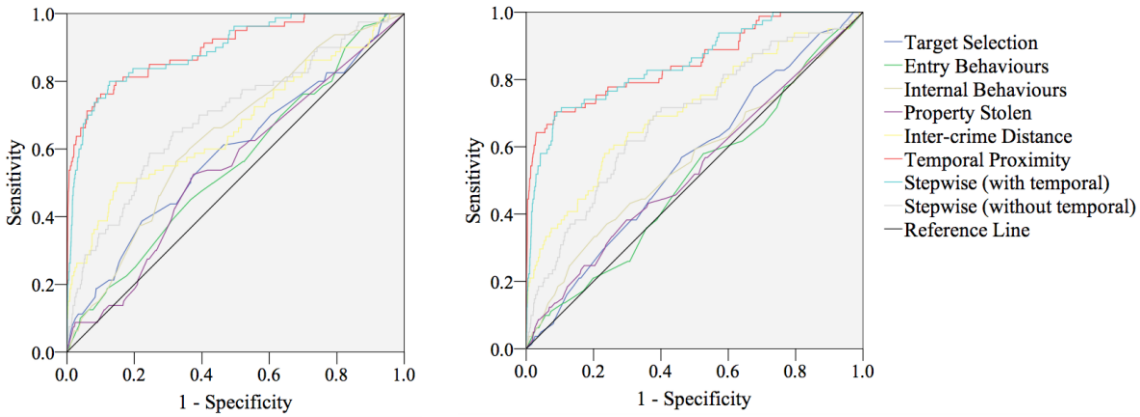


Figure 5. ROC curves for the development (left) and test (right) samples displaying the discrimination accuracy for the simple and forward stepwise logistic regression models constructed using the serial break and enter data.

Table 12. Development and test sample AUCs and their associated 95% confidence intervals for the simple and forward stepwise logistic regression models constructed using the serial break and enter data.

	Development Sample		Test Sample	
Model	AUC (*SE*)	95% CI	AUC (*SE*)	95% CI
Target Selection	.59 (.03)**	.52 – .65	.56 (.03)	.50 – .62
Entry Behaviours	.56 (.03)	.49 – .62	.51 (.03)	.44 – .57
Internal Behaviours	.64 (.03)***	.58 – .70	.56 (.04)	.49 – .63
Property Stolen	.55 (.03)	.48 – .61	.53 (.03)	.47 – .60
ICD	.67 (.04)***	.60 – .74	.71 (.03)***	.64 – .77
Temporal Proximity	.90 (.02)***	.86 – .94	.85 (.03)***	.80 – .90
Stepwise LR[a]	.89 (.02)***	.86 – .93	.85 (.02)***	.80 – .90
Stepwise LR[b]	.70 (.03)***	.63 – .76	.69 (.03)***	.62 – .75

[a]Stepwise logistic regression including Temporal Proximity, Internal Behaviours, ICD, and Target Selection. [b]Stepwise logistic regression including Internal Behaviours, ICD, and Target Selection. *p<.05. **p<.01. ***p<.001.

Phase 3: Developing and Evaluating Standard CT and ICT Models

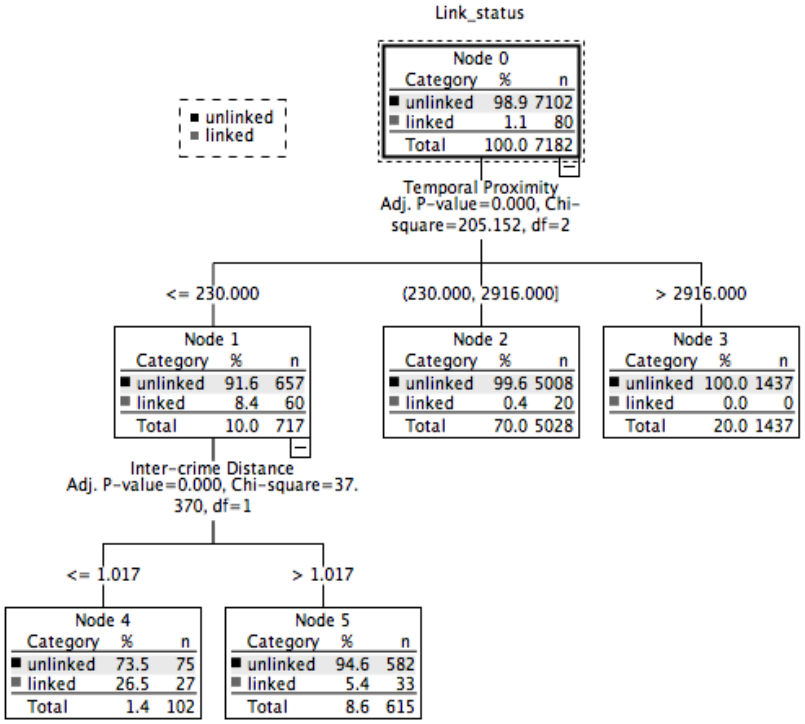
The next phase involved the development and evaluation of CT-based models. The CHAID parameters selected and sample size requirements were previously discussed in the methodology chapter.

CHAID analyses. To develop the standard CT, the CHAID algorithm was applied to the development sample and the resulting model was simultaneously forced on the test sample. Specifically, the following predictors/linking features were entered into the first CHAID analysis: target selection characteristics, entry behaviours, internal

behaviours, property stolen, temporal proximity, and ICD. The standard CT that resulted from this first iteration of CHAID is presented in Figure 6 for the development (left) and test (right) samples. As shown, temporal proximity was the first predictor selected, splitting the crime pairs into three nodes. This was followed by ICD, splitting the crime pairs from Node 1 into two additional nodes. In total, the CT split the crime pairs into six nodes/sub-groups, four of which were terminal nodes.

As also shown in Figure 6, the base rate of linked crime pairs in the development and test samples was 1.10%. Following Monahan et al. (2001), terminal nodes containing greater than 2.20% linked crime pairs were classified as linked, crime pairs containing less than 0.55% linked crime pairs were classified as unlinked, and nodes containing a proportion of linked crime pairs that were equal to or between these two thresholds (0.55 – 2.20%) were deemed unclassified. For the development sample, this resulted in two linked nodes (Nodes 4 and 5) and two unlinked nodes (Nodes 2 and 3). No nodes were left unclassified. Applying these cut-offs to the test sample resulted in the exact same classification pattern for each respective node. Given that no nodes were left unclassified, an ICT could not be constructed.

Iteration 1: Development Sample



Iteration 1: Test Sample

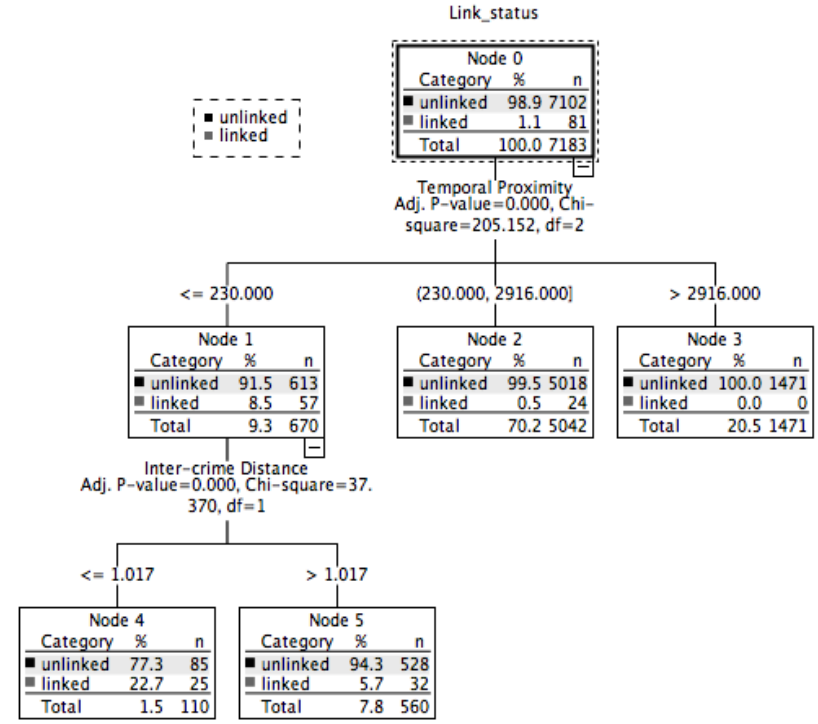
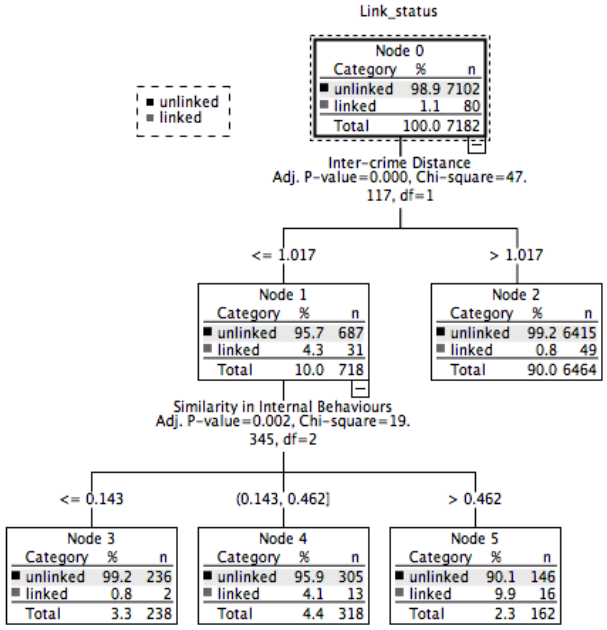


Figure 6. Iteration 1 (the standard CT) of the CHAID analyses for the serial break and enter data when temporal proximity was included in the analyses.

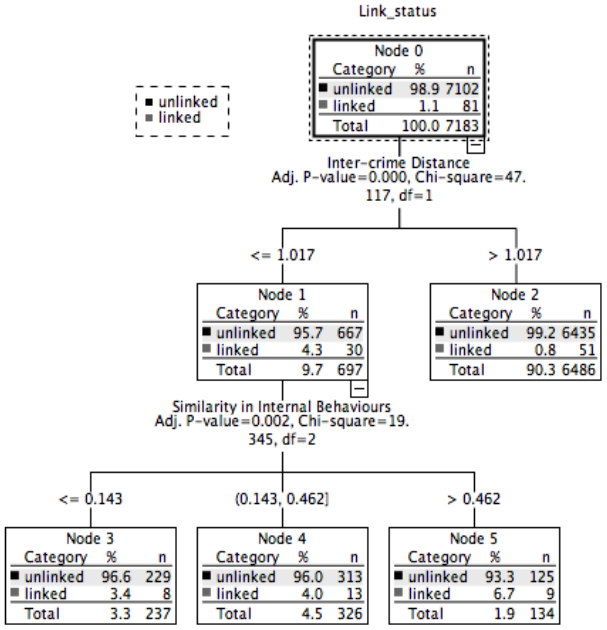
As previously explained, the CHAID analyses were then re-run without temporal proximity as a predictor. The standard CT (Iteration 1) that resulted from the CHAID analyses, which excluded temporal proximity as a predictor, is presented in Figure 7 (for the development and test samples). As shown, ICD was the first predictor selected, splitting the crime pairs into two nodes. This was followed by internal behaviours, splitting the crime pairs from Node 1 into three additional nodes. In total, the CT split the crime pairs into five nodes/sub-groups, four of which were terminal nodes. Using the same classification thresholds as the original CT, two nodes in the development sample were classified as linked (Nodes 4 and 5), and two nodes were unclassified (Nodes 2 and 3). In the test sample, three nodes were classified as linked (Nodes 3, 4, and 5) and one node (Node 2) was deemed unclassified. Compared to the standard CT including temporal proximity as a predictor, a substantially lower number of crime pairs could be classified as linked or unlinked using the standard CT excluding temporal proximity. In total, 6,702 crime pairs (93.32%) in the development sample and 6,486 crime pairs (90.30%) in the test sample were deemed unclassified by the standard CT, which excluded temporal proximity.

In an attempt to construct an ICT model excluding temporal proximity, crime pairs in the unclassified nodes in the development sample were pooled (Nodes 2 and 3) and the CHAID algorithm was applied to these cases a second time (and the resulting CT was simultaneously applied to the unclassified cases from Node 2 in the test sample). As shown in Figure 7, a second iteration (and therefore ICT) was not produced.

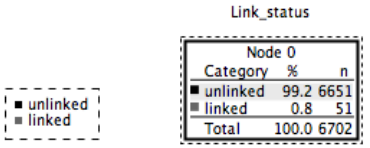
Iteration 1: Development Sample



Iteration 1: Test Sample



Iteration 2: Development Sample



Iteration 2: Test Sample

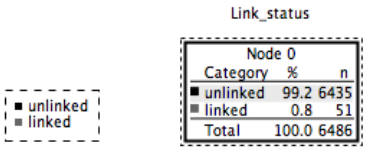


Figure 7. Iterations 1 (the standard CT) and 2 of the CHAID analyses for the break and enter data when temporal proximity was excluded from the analyses.

ROC analyses. The predicted probabilities from each of the CT-based models (standard CT including temporal proximity and standard CT excluding temporal proximity) were then entered into separate ROC analyses in order to compare their predictive accuracies to that achieved by the stepwise logistic regression model. The results of the ROC analyses comparing the standard CTs to the stepwise logistic regression models are visually presented in Figures 8 (models including temporal distance) and 9 (models excluding temporal distance). As expected, higher ROC curves were achieved for the models including temporal proximity as compared to the models excluding temporal proximity.

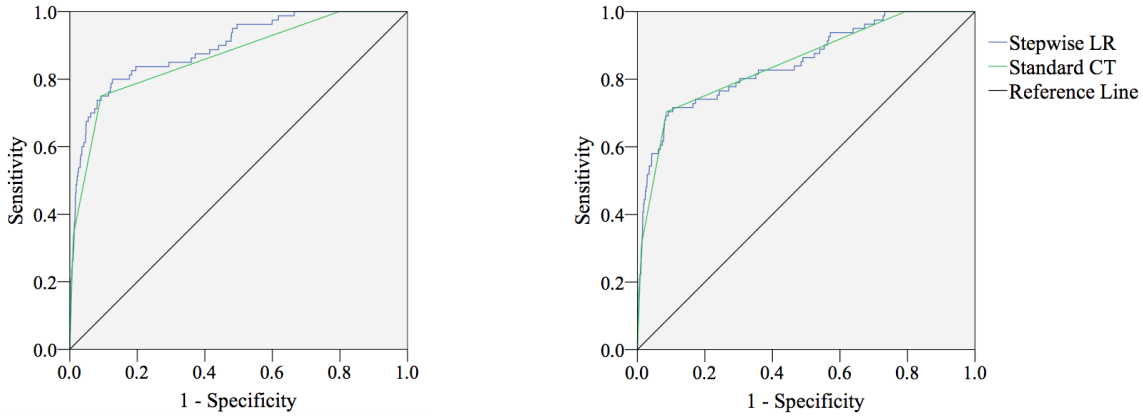


Figure 8. ROC curves for the development (left) and test samples (right) displaying the discrimination accuracy for the logistic regression and standard CT model constructed using the break and enter data when temporal proximity was included in the analyses.

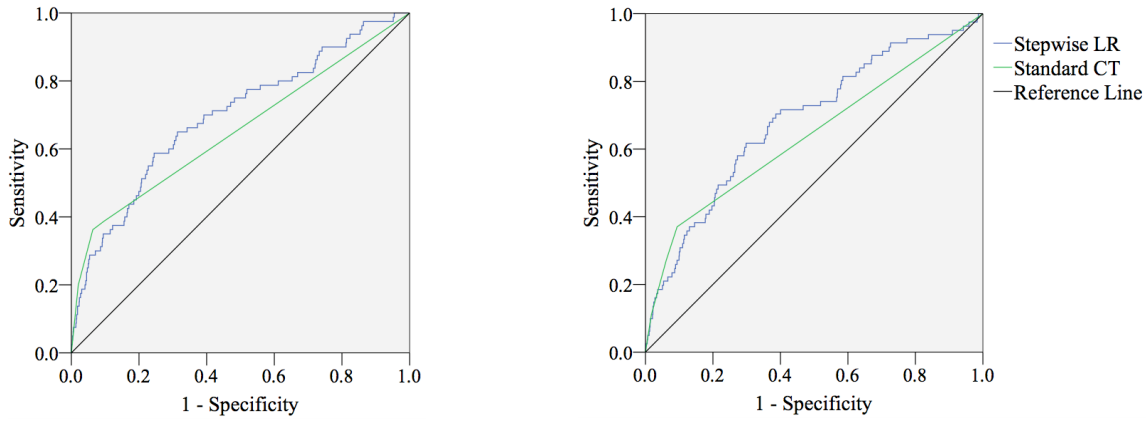


Figure 9. ROC curves for the development (left) and test samples (right) displaying the discrimination accuracy for the logistic regression and standard CT model constructed using the break and enter data when temporal proximity was excluded from the analyses.

Table 13 displays the AUCs and their associated confidence intervals for all models developed thus far. In terms of the development sample, moderate levels of predictive accuracy were achieved by the logistic regression and standard CT models including temporal proximity. Although the logistic regression model including temporal proximity achieved a slightly higher AUC than the corresponding standard CT model, the difference between the two models was not significant (their confidence intervals overlapped substantially and *p*'s >.05 for the significance tests). Similar results were found in the development sample when comparing the logistic regression to the standard CT model excluding temporal proximity (*p*'s >.05). Similar to the stepwise logistic regression results explored earlier, the CT model that included temporal proximity resulted in a significantly higher level of predictive accuracy relative to the CT model that excluded temporal proximity ($z = 5.20$, $p<.001$). The exact same pattern of results was found in the test sample.

Table 13. Development and test sample AUCs and their associated 95% confidence intervals for the stepwise logistic regression models and standard CT models constructed using the serial break and enter data.

	Development Sample		Test Sample	
Model	AUC (*SE*)	95% CI	AUC (*SE*)	95% CI
Including Temporal Proximity				
Stepwise LR[a]	.89 (.02)	.86 – .93	.85 (.02)	.80 – .90
Standard CT	.87 (.02)	.82 – .91	.85 (.02)	.80 – .90
Excluding Temporal Proximity				
Stepwise LR[b]	.70 (.03)	.63 – .76	.69 (.03)	.62 – .75
Standard CT	.65 (.04)	.58 – .73	.64 (.04)	.57 – .71

Note. All *p*'s <.001.
[a]Stepwise logistic regression including Temporal Proximity, Internal Behaviours, ICD, and Target Selection. [b]Stepwise logistic regression including Internal Behaviours, ICD, and Target Selection.

Comparisons of the AUCs *across* the development and test samples suggest that generalizability was not issue with any of the models. The AUCs decreased only slightly from the development sample to the test sample for all models, with the largest degree of shrinkage evident for the logistic regression model including temporal proximity (shrinkage = .04). The shrinkage observed was not statistically significant for any model (i.e., all AUCs overlapped from development to test for each model and all *p*'s >.05 for significance tests). These results suggest that all CT and logistic regression models generalize relatively well to the test sample crime pairs.

Comparing classification abilities. The next step was to compare the classification abilities of the logistic regression and CT-based models using the two-threshold approach. Specifically, the number of crime pairs labelled as linked (i.e., those with predicted probabilities >2.20%), unlinked (i.e., those with predicted probabilities

<0.55%), and unclassified (i.e., those with predicted probabilities equal to or between 0.55 and 2.20%) by each model was calculated and compared. The classification abilities of the models including and excluding temporal proximity as a predictor are presented in Table 14 as a function of the observed state (i.e., whether the crime pairs were truly linked or unlinked).[41]

As shown in Table 14, when temporal proximity was included in the models, 83.89% of cases could be classified as linked or unlinked using the logistic regression model, whereas 100% of cases could be classified as linked or unlinked using the standard CT. A total of 71.11% and 91.37% of the unlinked crime pairs were *correctly* classified by the logistic regression and CT model, respectively. In contrast, the logistic regression and CT models correctly classified a total of 71.60% and 70.37% of the linked crime pairs, respectively. This resulted in an overall correct classification rate of 70.12% for the logistic regression model and 91.13% for the CT model.

In contrast, the classification abilities of the models were quite different when examining the models that excluded temporal proximity as a predictor. Generally speaking, both models were able to classify a substantially smaller proportion of crime pairs as linked or unlinked using the two-threshold classification approach; however, the logistic regression model tended to outperform the CT model. For instance, the logistic regression model correctly classified 27.47% of unlinked crime pairs, whereas the CT model was unable to correctly classify any unlinked crime pairs. In contrast, the logistic regression model performed slightly worse than the CT model when it came to classifying linked crime pairs – with the models correctly classifying 23.46% and 37.04%

[41] The results are presented for the test sample only given that the previous analyses confirmed that the models generalize well to the test sample crime pairs.

of linked crime pairs, respectively. The majority of cases were left unclassified in both the logistic regression and CT model (63.89% and 90.30%, respectively). This resulted in an overall correct classification rate of 27.43% for the logistic regression model and 0.45% for the CT model.

Table 14. Classification table for the test sample using the two-threshold approach on the break and enter logistic regression and classification tree models when temporal proximity was included (top) and excluded (bottom) from the models.

	Including Temporal Proximity							
	Logistic Regression Model				Classification Tree Model			
Observed	Unlinked	Unclassified	Linked	Percent CC[a]	Unlinked	Unclassified	Linked	Percent CC[a]
Unlinked	4,979	1,151	972	71.11	6,489	0	613	91.37
Linked	17	6	58	71.60	24	0	57	70.37
Total	4,996	1,157	1,030	70.12	6,513	0	670	91.13
	Excluding Temporal Proximity							
	Logistic Regression Model				Classification Tree Model			
Observed	Unlinked	Unclassified	Linked	Percent CC[a]	Unlinked	Unclassified	Linked	Percent CC[a]
Unlinked	1,951	4,535	616	27.47	0	6,435	667	0.00
Linked	8	54	19	23.46	0	51	30	37.04
Total	1,959	4,589	635	27.43	0	6,486	697	0.45

Note. Total $N = 7{,}183$; linked $n = 81$; unlinked $n = 7{,}102$.

[a]Percent CC = percent correctly classified for unlinked pairs, linked pairs, and overall (total).

Phase 4: Developing and Evaluating Multiple CT/ICT Linking Models

The final phase involved constructing and evaluating the multiple CT/ICT models. As mentioned, the process began by developing separate CT/ICT models: each one forcing a different predictor as the first splitting variable in the first CHAID analysis run. The steps involved in developing, combining, and evaluating the multiple CT/ICT models when temporal proximity was included as a predictor are presented in detail below. A condensed summary of this process is then provided for the multiple CT/ICT models that were developed when temporal proximity was excluded as a predictor.[42]

Constructing multiple CT/ICT models. The CT with temporal proximity forced as the first variable was already constructed in Phase 3 and is displayed in Figure 6. The five additional CTs (each beginning with a different predictor) for the multiple models including temporal proximity as a predictor are presented in Figures C1 through C6 in Appendix C. None of these analyses resulted in a second iteration; consequently, only standard CT models were created for all the individual models comprising the multiple models approach.

The results of each of these analyses are summarized in Table 15. As shown, tree depth (i.e., the number of levels produced from the root node) ranged from 2 to 3, the number of nodes ranged from 6 to 15, and the number of terminal nodes ranged from 4 to 10. With the exception of the initial splitting variable, all CTs were highly similar in terms of which predictors were selected at each level. All CTs contained temporal proximity at the second level, and all but one CT (the property stolen CT) included ICD

[42] For the sake of brevity, a condensed summary was provided for the set of analyses excluding temporal proximity as a predictor since the same pattern of results emerged in terms of the relative predictive accuracy across the various statistical approaches (i.e., stepwise logistic regression, CT and multiple CT/ICT).

at the third level. The CT where property stolen was forced as the initial splitting variable included temporal proximity at the second level, followed by internal behaviours at the third level.

Table 15. Characteristics of the multiple standard CTs produced using the serial break and enter data when each predictor was forced as the initial splitting variable in the CHAID analysis and temporal proximity was included in the models.

First Variable	Depth	Nodes (#)	Terminal Nodes (#)	Variables Included
Target Selection	3	12	8	Target Selection, Temporal Proximity, ICD
Entry Behaviours	3	10	7	Entry Behaviours, Temporal Proximity, ICD
Internal Behaviours	3	15	10	Internal Behaviours, Temporal Proximity, ICD
Property Stolen	3	12	8	Property Stolen, Temporal Proximity, Internal Behaviours
ICD	2	9	6	ICD, Temporal Proximity
Temporal Proximity	2	6	4	Temporal Proximity, ICD

The AUCs and percent of cases classified as linked or unlinked by each CT model are presented in Table 16. As shown, most CTs classified a high number of the crime pairs as linked or unlinked. In fact, all CTs in the development sample classified 100 percent of cases as linked or unlinked, with the exception of the CT beginning with property stolen, which classified 79.74% of cases. In contrast, more variability in the percent classified was seen in the test sample (ranging from 78.88% for the internal behaviours CT to 100% for the temporal proximity CT).

Predictive accuracies for the CTs were also moderate-to-high for the development sample (AUCs ranging from .86 for the entry behaviours CT to .91 for the internal behaviours CT) and the test sample (AUCs ranging from .75 for the target selection CT to .85 for the temporal proximity CT). None of the AUCs were significantly different from one another in the development sample (according to both the confidence intervals and significance tests: all *p*'s >.05). However, in the test sample, although the confidence intervals overlapped, significance tests revealed that the AUC for the CT beginning with temporal proximity was significantly higher than the AUC for the CT beginning with target selection ($z = 2.41$, $p = .016$). All other AUCs were not significantly different from one another (*p*'s >.05).

Shrinkage was much more apparent across these multiple CT models than any of the previous models examined thus far. Shrinkage ranged from .02 (for the temporal proximity CT) to .14 (for the target selection CT). Significance tests revealed significant shrinkage in the target selection CT (shrinkage =.14; $z = 3.53$, $p < .001$), internal behaviours CT (shrinkage = .07; $z = 1.98$, $p = .048$), and the property stolen CT (shrinkage = .09; $z = 2.36$, $p = .018$). All other differences were non-significant ($p > .05$). These differences between the development and test AUCs are visually depicted in Figure 10.

Table 16. Predictive accuracies and percent classified for the CT models where each predictor was forced as the first splitting variable in the CHAID analyses conducted using the serial break and enter data when temporal proximity was included in the models.

	Development Sample		
First/Forced Variable	Classified (%)	AUC (*SE*)	95% CI
Development Sample			
Target Selection	100.00	.89 (.02)	.85 – .93
Entry Behaviours	100.00	.86 (.02)	.82 – .91
Internal Behaviours	100.00	.91 (.02)	.88 – .94
Property Stolen	79.74	.89 (.02)	.85 – .92
ICD	100.00	.89 (.02)	.85 – .93
Temporal Proximity	100.00	.87 (.02)	.82 – .91
Test Sample			
Target Selection	95.99	.75 (.04)	.67 – .82
Entry Behaviours	88.77	.82 (.03)	.77 – .88
Internal Behaviours	78.88	.84 (.03)	.79 – .89
Property Stolen	92.70	.80 (.03)	.74 – .86
ICD	91.83	.84 (.03)	.78 – .89
Temporal Proximity	100.00	.85 (.02)	.80 – .90

Note. All p's <.001.

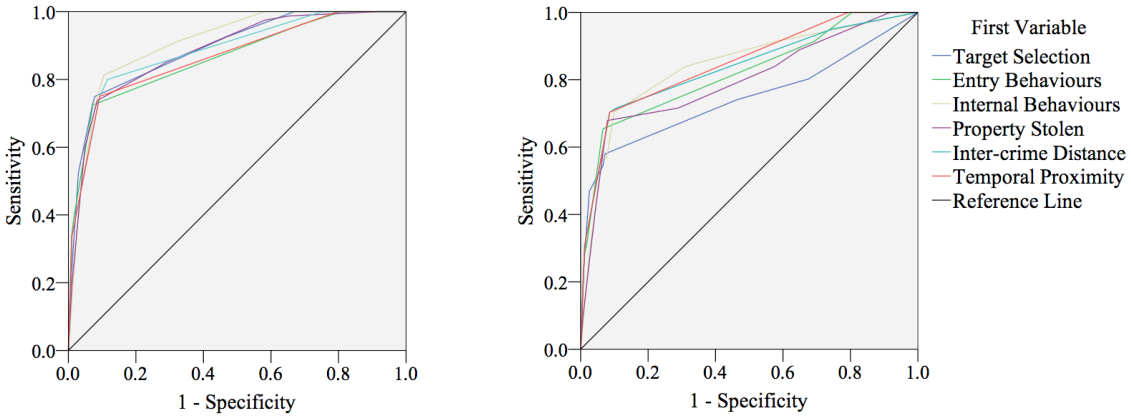


Figure 10. ROC curves for the development (left) and test samples (right) displaying the discrimination accuracy for the individual standard CT models comprising the multiple models approach for the break and enter data when temporal proximity was included in the models.

Combining the multiple CT models. Next, each crime pair was given a score based on how they were classified in each model using the two-threshold classification approach described previously. Specifically, for each model, crime pairs labelled as linked were assigned a score of 1, crime pairs labelled as unlinked were assigned a score of -1, and crime pairs labelled as unclassified were assigned a score of 0. A composite score was then created for each crime pair based on how they were classified on the models when they were combined (summing across all of their scores). Prior to creating the multiple model composite score, correlations were examined between the scores provided by each model to determine whether or not summing across scores to create the multiple model composite score is justifiable (Monahan et al., 2001). As shown in Table 17, correlations between the scores on each of the different models were moderate-to-

high (ranging from r = .67 to .93). Likewise, Cronbach's alpha (α = .96) for these six scores indicated a high level of internal reliability.

Table 17. Correlations between scores on each individual CT model comprising the multiple models approach for the break and enter data when temporal proximity was included in the models.

CT Score/Model	1	2	3	4	5
1. Target Selection	-				
2. Entry Behaviours	.78	-			
3. Internal Behaviours	.79	.72	-		
4. Property Stolen	.76	.67	.69	-	
5. ICD	.83	.74	.78	.74	-
6. Temporal Proximity	.93	.83	.85	.82	.89

Note. All p's <.001.

Composite scores ranged from -6 (indicating that the crime pair was classified as unlinked on all six CT models) to +6 (indicating that the crime pair was classified as linked on all six CT models), with a median score of -6.00 (SD = 3.26). The same range in composite scores was found when looking at linked and unlinked crime pairs separately; however, the median composite score was -6.00 (SD = 3.13) for unlinked cases and +6.00 for linked cases (SD = 4.87). Overall, these scores indicate that, on average, unlinked crime pairs were more often classified as unlinked (versus linked or unclassified) and linked crime pairs were more often classified as linked (versus unlinked or unclassified) across the different CT models.

The total number of crime pairs possessing each composite score value, and the corresponding percentage of those cases that are linked crime pairs, is presented in Table

18. Generally speaking, if more unlinked pairs are found at the bottom end of the distribution (e.g., with a score closer to -6) and more linked pairs are found at the higher end of the distribution (e.g., with a score closer to +6), then a higher level of discrimination accuracy would be expected for the composite score (multiple CT/ICT model approach). As shown, the largest proportion of linked cases was found in the composite score category of 6 for both the development and test sample. The smallest proportions of linked cases were found in the middle categories, with no linked cases possessing scores between -1 and 3 in the development sample and no linked cases possessing scores between -2 and 2 in the test sample. Overall, 74% of all linked crime pairs had a score of 1 or higher (indicating they were in the linked category more often across the six models than the other categories), whereas 91% of all unlinked crime pairs possessed a score of -1 or lower (indicating they were in the unlinked category more often across the six models than the other categories).

Table 18. Distribution of composite linkage scores and the percentage of linked cases included in each composite score category for the break and enter development and test samples when temporal proximity was included in the models.

	Development Sample		Test Sample	
Score	Total Cases (#)	Linked (%)	Total Cases (#)	Linked (%)
- 6	3,734	0.27	4,889	0.18
- 5	2,303	0.39	1,346	0.30
- 4	427	0.70	101	0.99
- 3	48	4.17	94	4.26
- 2	1	0.00	20	5.00
- 1	0	0.00	15	6.66
0	0	0.00	2	0.00
1	0	0.00	0	0.00
2	1	0.00	32	3.13
3	62	0.00	0	0.00
4	52	0.00	224	0.45
5	264	4.55	0	0.00
6	291	15.46	459	14.46

Constructing the empirically optimal combined CT model. Next, the empirically optimal combined CT model was created. First, a forward stepwise logistic regression analysis was run with the scores for each model entered as the predictors (six predictors) and linkage status as the outcome variable.[43] As shown in Table 19, all six

[43] Consistent with Monahan et al. (2001), logistic regression assumptions were not assessed for this analysis given that the goal was simply to determine a subset of models to include in the final multiple CT model (and not to identify stable logistic regression parameters for prediction purposes). However, multicollinearity did not seem to be an issue (i.e., there were no issues with model parameters changing drastically as new predictors entered the model, and predictors entering the stepwise model early on were not excluded in later steps as new predictors entered the model).

score variables were selected into the stepwise logistic regression model. Given that these results suggest that the empirically optimal multiple CT model is the one that includes the combination of all six scores, the multiple CT model was not modified further for the break and enter data.

Table 19. Results of forward stepwise logistic regression analysis on the scores for each of the six CT models comprising the original composite scores for the break and enter data when temporal proximity was included in the models.

Model	B (*SE*)	Wald (*df*)	χ^2 (*df*)	R_N^2	R_L^2
Stepwise			484.89 (6)	.29	.27
Target Selection CT score	1.83 (0.40)	20.73 (1)			
Internal Behaviours CT score	1.15 (0.25)	21.44 (1)			
Entry Behaviours CT score	0.88 (0.22)	16.43 (1)			
Temporal Proximity CT score	-2.99 (0.54)	31.25 (1)			
ICD CT score	0.68 (0.27)	6.49 (1)			
Property Stolen CT score	0.67 (0.26)	6.88 (1)			
Constant	-4.08 (0.11)	1402.17 (1)			

Note. χ^2 = model chi-square; R_N^2 = Nagelkerke index; R_L^2 = Hosmer and Lemeshow's index; all *p*'s <.001, except the regression coefficients for the property stolen CT score (*p* = .009) and the internal behaviours CT score (*p* = .011).

Evaluating the multiple CT model. ROC analysis was then conducted to examine the predictive accuracy of the multiple CT model. The ROC curves produced from this analysis are presented in Figure 11 for the development and test samples. As shown, the multiple CT model was associated with a moderate degree of predictive accuracy for both the development sample (AUC = .88, 95% CI [.84, .93]) and test sample (AUC = .85, 95% CI [.80, .90]). Although the AUC decreased from development to test (shrinkage = .03), the confidence intervals overlapped substantially (and *p* >.05 for

the significance test), suggesting that the multiple CT model generalized relatively well to the test sample.

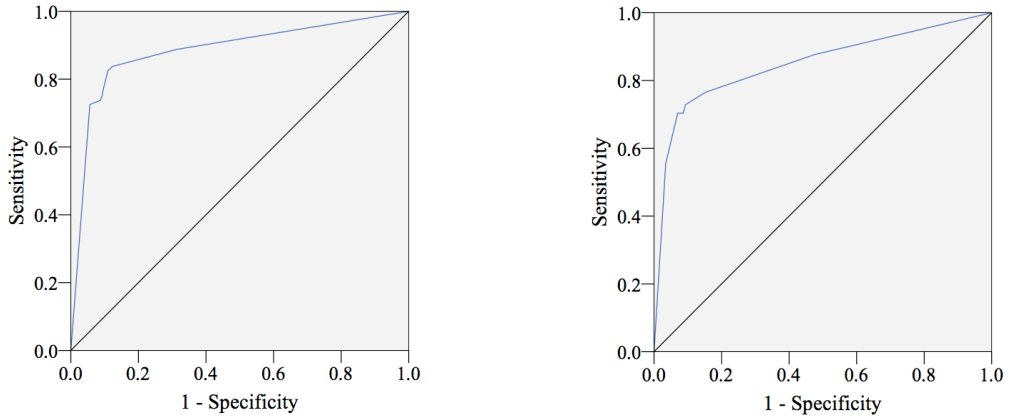


Figure 11. ROC curves for the development (left) and test samples (right) displaying the discrimination accuracy for the multiple CT model for the break and enter data when temporal proximity was included in the models.

Constructing the multiple CT/ICT models without temporal proximity. As mentioned, the multiple CT/ICT approach was also tested when temporal proximity was excluded as a predictor. This meant that a total of five CT/ICTs were constructed and combined.[44] The CT with ICD forced as the first variable was already constructed in Phase 3 and is displayed in Figure 7. The characteristics of all five CT/ICTs constructed when temporal proximity was excluded from the models are summarized in Table 20. As shown, tree depth ranged from 2 to 4, the number of nodes ranged from 5 to 13, and the number of terminal nodes ranged from 4 to 8. With the exception of the initial splitting variable, all CT/ICTs were highly similar in terms of which predictors were selected at

[44] As mentioned, an abbreviated version of the multiple model results will be presented here for the sake of brevity.

each level (e.g., ICD and internal behaviours were included in most CT/ICTs). The only ICT produced was for the internal behaviours forced model, where the unclassified crime pairs after Iteration 1 were split on the basis of their ICD at Iteration 2.

Table 20. Characteristics of the multiple CT/ICTs produced using the serial break and enter data when each predictor was forced as the initial splitting variable in the CHAID analysis and temporal proximity was excluded from the models.

First Variable	Depth	Nodes (#)	Terminal Nodes (#)	Variables Included
Target Selection	2	7	4	Target Selection, ICD
Entry Behaviours	3	8	5	Entry Behaviours, ICD, Internal Behaviours
Internal Behaviours[a]	2	6	4	Internal Behaviours, ICD
Property Stolen	4	13	8	Property Stolen, Internal Behaviours, ICD, Entry Behaviours
ICD	2	5	4	ICD, Internal Behaviours

[a] A second iteration was produced for the unclassified cases when internal behaviours was used as the initial splitting variable in the first CT. The second iteration included two nodes split by ICD, both of which were terminal nodes.

The AUCs and percent of cases classified as linked or unlinked by each CT model are presented in Table 21. As shown, most CT/ICTs did not perform as well as the multiple models that included temporal proximity. For instance, the CT/ICTs excluding temporal proximity classified a small number of the crime pairs as linked or unlinked, ranging from 6.68% to 51.39% in the development sample and 6.67% to 13.91% in the test sample. Predictive accuracies for the CT/ICTs were also low-to-moderate for the

development sample (AUCs ranging from .65 for the ICD CT to .75 for the property stolen CT) and low for the test sample (AUCs ranging from .61 for the property stolen CT to .67 for the target selection CT). For the development sample, significance tests revealed that the property stolen CT had a significantly higher AUC than the ICD CT (z = 2.16, p = .031). All other development sample AUCs were not significantly different from one another. For the test sample, no significant differences were found between AUCs (all p's >.05). Likewise, a significant amount of shrinkage was observed for the property stolen CT (shrinkage = .14, z = 3.03, p =.002). The amount of shrinkage observed in all other AUCs was not significant (all p's >.05).

Table 21. Predictive accuracies and percent classified for the CT/ICT models where each predictor was forced as the first splitting variable in the CHAID analyses conducted using the serial break and enter data when temporal proximity was excluded from the models.

	Development Sample		
First/Forced Variable	Classified (%)	AUC (*SE*)	95% CI
Development Sample			
Target Selection	10.00	.70 (.03)	.63 – .77
Entry Behaviours	17.79	.69 (.03)	.61 – .75
Internal Behaviours	31.69	.73 (.03)	.67 – .79
Property Stolen	51.39	.75 (.03)	.69 – .81
ICD	6.68	.65 (.04)	.58 – .73
Test Sample			
Target Selection	9.70	.67 (.03)	.60 – .74
Entry Behaviours	8.70	.62 (.04)	.55 – .69
Internal Behaviours	13.91	.65 (.04)	.58 – .72
Property Stolen	6.67	.61 (.04)	.54 – .68
ICD	9.70	.64 (.04)	.57 – .71

Note. All p's <.001.

Each crime pair was then scored based on how they were ultimately classified in each of the five models excluding temporal proximity. The correlations between the scores produced by these models were moderate-to-high (*r*'s ranging from .32 to .90) and Cronbach's alpha was also satisfactory ($\alpha = .81$). The composite scores that were created from summing across the scores for the five models ranged from -3 to +5, with a median score of 0.00 (*SD* = 1.49). The same range and median in composite scores was found when looking at linked and unlinked crime pairs separately; however, the standard deviations varied (*SD* = 1.46 for unlinked cases and *SD* = 2.34 for linked cases).

Although the highest proportion of linked cases in both the development and test sample was found in the composite score category of 6, only 46% of all linked crime pairs had a score of 1 or higher (indicating they were in the linked category more often across the five models than the other categories). Likewise, only 42% of all unlinked crime pairs possessed a score of -1 or lower (indicating they were in the unlinked category more often across the five models than the other categories). In fact, most (62%) crime pairs had a score of 0 (i.e., they were unclassified on all models). Overall, this indicates that the discrimination accuracy of the multiple models approach is likely to be low (which is not surprising given the performance of the individual models prior to summing their scores).

Next, all CT/ICT score variables were entered into a stepwise logistic regression analysis to construct the empirically optimal combined CT model. Only two scores (the ICD CT score and the property stolen CT score) were included in the revised model; however, the Cronbach's alpha for these two scores alone was quite low ($\alpha = .48$). As

such, the original composite scores were not revised and the original multiple CT/ICT was retained.

Finally, ROC analysis was conducted to examine the predictive accuracy of the multiple CT/ICT model excluding temporal proximity. The ROC curves produced from this analysis are presented in Figure 12 for the development and test samples. As shown, the multiple CT/ICT model excluding temporal proximity was associated with a moderate level of predictive accuracy for the development sample (AUC = .77, 95% CI [.71, .82]) and a low level of predictive accuracy for the test sample (AUC = .65, 95% CI [.58, .71]). A statistically significant amount of shrinkage in the AUC was observed when the multiple CT/ICT model excluding temporal proximity was applied to the test sample (shrinkage = .12; $z = 2.62$, $p = .009$), although the 95% CIs did overlap slightly.

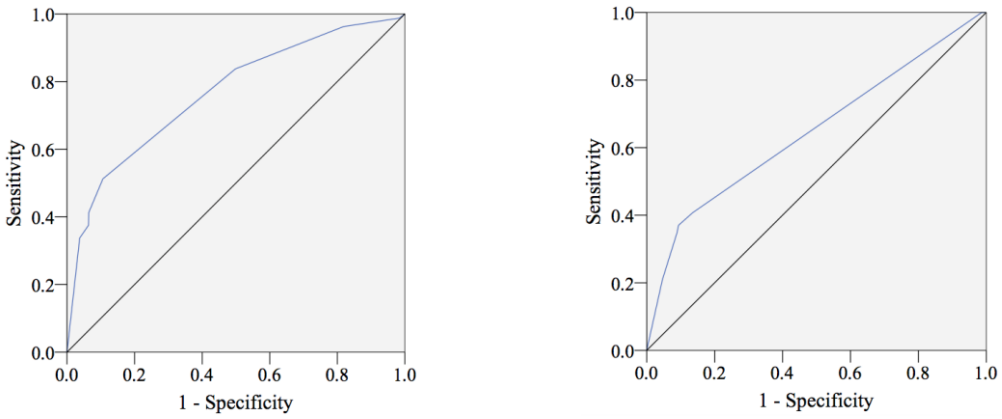


Figure 12. ROC curves for the development (left) and test samples (right) displaying the discrimination accuracy for the multiple CT/ICT model for the break and enter data when temporal proximity was excluded from the models.

Comparing the performance of all linking models. The final step was to compare the predictive accuracies of the stepwise logistic regression models, standard CT models, and multiple CT/ICT models developed thus far.[45] As shown in Figure 13 and Table 22, all models that included temporal proximity were associated with a moderate degree of predictive accuracy across both the development and test samples. For the development sample, the stepwise logistic regression model resulted in the highest level of predictive accuracy (AUC = .89), followed by the multiple CT model (AUC = .88), and the standard CT model (AUC = .87). However, all of the confidence intervals overlapped and the significance tests revealed that these AUCs were not significantly different from one another (all *p*'s >.05). For the test sample, all of the AUCs were the same across all models (AUC = .85), with development-to-test shrinkage ranging from .02 to .04. As a whole, generalizability was not a concern with any of the models including temporal proximity (i.e., the 95% CIs surrounding the AUCs for all models overlap and all *p*'s >.05 for the significance tests).

[45] The classification ability of the multiple model approach (comprised of composite scores) was not compared to the classification ability of the other approaches (comprised of predicted probabilities) since a comparable threshold to use on the composite scores was not identified by Monahan et al. (2001).

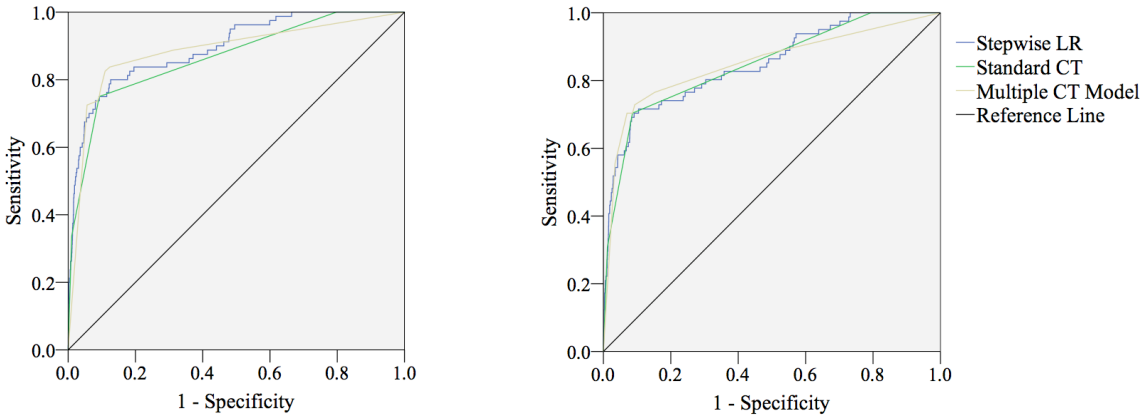


Figure 13. ROC curves for the development (left) and test samples (right) displaying the discrimination accuracy for the logistic regression model, standard CT model, and the multiple CT model created with the break and enter data when temporal proximity was included in the models.

Table 22. Development and test sample AUCs and their associated 95% confidence intervals for the stepwise logistic regression models, standard CT models, and multiple CT/ICT models created with the break and enter data.

	Development Sample		Test Sample	
Model	AUC (*SE*)	95% CI	AUC (*SE*)	95% CI
Including Temporal Proximity				
Stepwise LR[a]	.89 (.02)	.86 – .93	.85 (.02)	.80 – .90
Standard CT	.87 (.02)	.82 – .91	.85 (.02)	.80 – .90
Multiple CT	.88 (.02)	.83 – .93	.85 (.03)	.80 – .90
Excluding Temporal Proximity				
Stepwise LR[b]	.70 (.03)	.63 – .76	.69 (.03)	.62 – .75
Standard CT	.65 (.04)	.58 – .73	.64 (.04)	.57 – .71
Multiple CT/ICT	.77 (.03)	.71 – .82	.65 (.04)	.58 – .72

Note. all *p*'s<.001.
[a]Stepwise logistic regression including Temporal Proximity, Internal Behaviours, ICD, and Target Selection. [b]Stepwise logistic regression including Internal Behaviours, ICD, and Target Selection.

As shown in Table 22 and Figure 14, the models excluding temporal proximity were associated with low-to-moderate levels of predictive accuracy across both the development and test samples. For the development sample, the multiple CT/ICT model resulted in the highest level of predictive accuracy (AUC = .77), followed by the stepwise logistic regression model (AUC = .70), and the standard CT model (AUC = .65). Although all confidence intervals overlapped, the multiple CT/ICT model was associated with a significantly higher level of predictive accuracy than the standard CT model (z =2.62, p =.009), however, the other comparisons were not significant (both p's >.05). For the test sample, the AUCs ranged from .64 to .69, with the stepwise logistic regression

model achieving the highest AUC. None of the AUCs were significantly different from one another when the models were applied to the test sample (all *p*'s >.05). Although shrinkage was not a concern for the logistic regression or standard CT models (both *p*'s >.05), as mentioned previously, a significant amount of shrinkage was found for the multiple CT/ICT model excluding temporal proximity (shrinkage = .12).

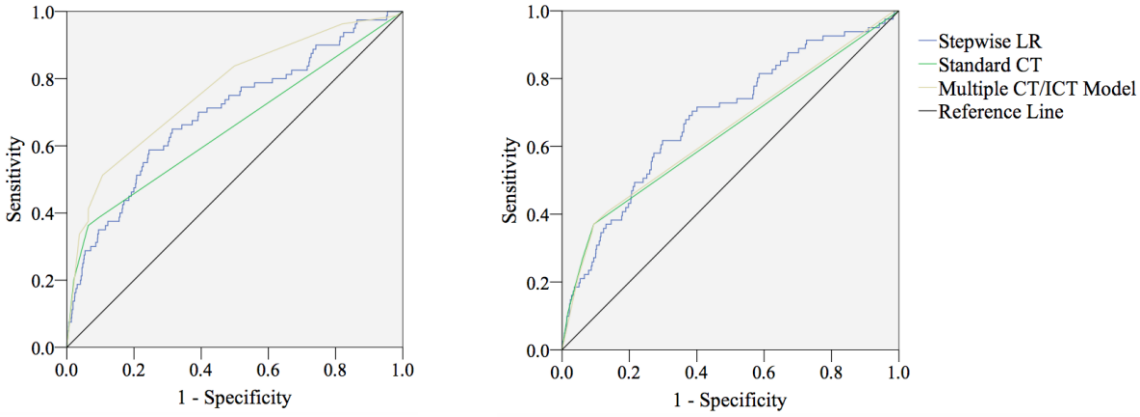


Figure 14. ROC curves for the development (left) and test samples (right) displaying the discrimination accuracy for the logistic regression model, standard CT model, and the multiple CT/ICT model created with the break and enter data when temporal proximity was excluded from the models.

Discussion

The analyses presented above provide a number of insights into BLA using property crime. Generally speaking, and as expected, the results indicated that it is possible to link serial break and enters from Saint John, NB, Canada in a relatively accurate fashion. The first hypothesis was partially supported: the initial descriptive analyses revealed that break and enters committed by the same offender had higher levels

of similarity in target selection characteristics and internal behaviours, shorter ICDs, and fewer days between them, than break and enters committed by different offenders. It is important to note, however, that the magnitude of these differences varied as a function of the features considered: larger effect sizes were associated with temporal proximity, ICD, and internal behaviours, whereas a smaller effect was associated with target selection characteristics. The lowest effect sizes were associated with entry behaviours and property stolen, where a similar level of behavioural similarity was observed between linked and unlinked crime pairs.

Based on previous research, it was also hypothesized that there would be noticeable differences in the discrimination abilities of the different statistical approaches employed. That is, it was expected that the stepwise logistic regression and standard CT approaches would be associated with similar levels of predictive accuracy, whereas the multiple CT/ICT approach would result in higher levels of predictive accuracy. This hypothesis was not supported using the current data. Regardless of whether a stepwise logistic regression, standard CT, or multiple CT/ICT approach was used, all models resulted in the same moderate level of predictive accuracy when applied to the test sample (e.g., test AUCs = .85 for all three models including temporal proximity as a predictor). Although the multiple CT/ICT model excluding temporal proximity as a predictor was associated with a significantly higher level of accuracy than the standard CT model (but not the stepwise logistic regression model) for the development sample, a significant amount of shrinkage in the predictive accuracy achieved was found when the multiple CT/ICT model was applied to the test sample. As such, similar levels of

predictive accuracy were found across the logistic regression, standard CT, and multiple CT/ICT models when the models were applied to the test sample.

Likewise, contrary to what was hypothesized based on Tonkin, Woodhams, et al. (2012), there were no issues with shrinkage for the standard CTs developed in the current study. However, although the amount of shrinkage observed for the multiple CT/ICT model including temporal proximity was not significant, a significant amount of shrinkage was observed for the multiple CT/ICT model excluding temporal proximity as a predictor.

Finally, although it was expected that hidden patterns of behavioural consistency and distinctiveness would become apparent using the CT approach (relative to the main effects logistic regression approach), this hypothesis was not supported with the current sample of break and enters.

Potential reasons for these findings, as well as the proposed advantages of CT-based models in light of these findings, are explored in more detail below.

The Behavioural Consistency and Distinctiveness of Canadian Serial Burglars

As previously mentioned, this is the first study to examine our ability to link property crimes from a Canadian police service. As expected based on previous linking research, crimes committed by the same offenders were higher in behavioural similarity (for some domains, such as internal behaviours and target selection characteristics) and had shorter distances between them (both in space and time) than crimes committed by different offenders. This suggests, in a general way, that linking serial break and enters using behavioural information is possible within the Canadian context.

The results of this study do, however, appear to suggest that our ability to link crimes is dependent on the type of behavioural information that is relied upon. As demonstrated from the results of the descriptive analyses and simple logistic regressions, the best predictors of linkage status were temporal proximity (test AUC = .85) and ICD (test AUC = .71). All remaining behavioural domains (i.e., internal behaviours, target selection, entry behaviours, and property stolen) were unable to discriminate between linked and unlinked crime pairs at levels greater than chance (test AUCs ranging from .51 to .56, all *p*'s >.05). Similar results were found concerning the relative importance of predictors in the CT analyses: with temporal proximity being selected as the first predictor to split the cases, followed by ICD. The remaining predictors were not included in the standard CT produced when temporal proximity was included. When temporal proximity was excluded, however, cases were first split using ICD followed by internal behaviours.

Some recent research, which has provided separate tests of the offender consistency and distinctiveness assumptions across subsets of variables, offers some insight into why domains, such as entry and target selection behaviours, may have little predictive accuracy. For instance, Bouhana, Johnson, and Porter (2014) found that, although individual burglars seemed to be relatively consistent in their choices of how to enter properties, the relatively low level of distinctiveness observed *between* offenders with respect to these factors prevented these features from being useful for linking purposes (similar results were found for target selection characteristics and the time of day that the crime was committed). Likewise, both experimental and non-experimental research examining burglars' expertise has found that different offenders prefer similar

target selection strategies and entry behaviours, and tend to steal similar types of property (see Nee, 2015 for an overview of this research). These preferences would decrease the extent to which different offenders display different behaviours, thus reducing the discriminatory power of these behaviours for linking purposes.

Generally speaking, the findings from the current study are in line with previous linking research: ICD and temporal proximity tend to be the best predictors of linkage status relative to all other behavioural domains when linking property crimes (e.g., see Bennell et al., 2014 for a review). For example, studies of UK residential burglaries have found AUCs in the low range according to Swets' (1988) guidelines for target selection, entry behaviours, property stolen, and internal behaviours (i.e., AUCs between .53 and .66), whereas AUCs for ICD and temporal proximity have ranged from .82 to .94 (Bennell et al., 2014).[46] Researchers commonly suggest that space and time variables may outperform modus operandi (MO) variables because they are arguably less impacted by situational factors, they are more likely to be accurately recorded by the police, and/or more reliably reported by victims (because they are less subjective; e.g., Bennell & Canter, 2002; Tonkin & Woodhams, 2015).

It is important to note, however, that studies of residential burglary committed in locations outside of the UK have found higher levels of predictive accuracy for some MO domains. For instance, Tonkin, Woodhams, et al. (2012), found test AUCs in the moderate range for entry (AUC = .70), target (AUC = .77), and internal behaviours (AUC = .78) when they examined residential burglaries committed in Finland. Reasons for these variations are currently unclear, but may be attributable to a number of factors.

[46] It is important to note that the vast majority of linking research examining residential burglaries has used UK data (Bennell et al., 2014).

For example, as explained by Tonkin, Santtila, et al. (2012), there is more variability in the type of housing in Finland relative to the UK, which may allow for more inter-offender distinctiveness. In addition, differences in the way MO domains were defined may explain the variations. For instance, the number of offenders and how the offender(s) exited the crime scene were included in the internal behaviours domain in the Finnish dataset, whereas these factors were not included in previous research using UK burglaries.[47]

Related to this last point, studies examining linking accuracy with residential burglaries have frequently defined behavioural domains in a slightly different manner (e.g., including different behaviours in the domains for target selection and internal behaviours, as mentioned above). As such, it is difficult to gauge how comparable the behavioural domains (and the results emerging from those domains) truly are across studies examining residential burglaries. As concrete evidence of this difficulty, in their linking study of serial car thefts from the UK, Davies et al. (2012) compared an old approach to operationalizing target selection behaviours in car thefts (from Tonkin et al., 2008) to a new approach, which included updated behavioural features (e.g., whether the car was equipped with a built-in immobilizer). They found that their new target selection domain could lead to a considerable increase in predictive accuracy (i.e., the AUC increased from .62 to .76).

Although differences in how individual behaviours (and domains) are defined are oftentimes a consequence of practical constraints on the data available to researchers in

[47] It is important to note, however, that these additional behaviours were included in the internal behaviours domain used in the current study and this higher level of predictive accuracy was not observed. As such, future research is needed to determine reasons why higher levels of accuracy may be obtained in certain jurisdictions relative to others.

different jurisdictions, future research should attempt to systematically examine the consequences of defining behavioural domains in different ways within a sample of residential burglaries from the same jurisdiction, and across samples of residential burglaries from different jurisdictions. Only then can we begin to understand the extent to which differences in domain performance are due to data collection artefacts versus true differences in offender consistency and distinctiveness.

The predictive accuracy of ICD. Although the results reviewed above seem to parallel the results of studies examining residential burglaries in other jurisdictions, especially the UK, one interesting finding in the current study noticeably sets it apart from most previous linking research: the level of discrimination accuracy associated with ICD. The finding that crimes committed closer together are more likely to have been committed by the same offender is relatively robust across previous studies that have examined property crime (e.g., Bennell & Jones, 2005; Markson et al., 2010; Tonkin, Santtila, et al., 2012; Tonkin & Woodhams, 2015). Although ICD, when considered alone, was a moderate predictor of linkage status in the current study (test AUC = .71), the level of predictive accuracy achieved was considerably lower than that achieved in previous research (e.g., AUCs for ICD have ranged from .83 to .94 for residential burglary studies; Bennell et al., 2014). Likewise, although included in the standard CT that was produced, ICD was not selected as the initial splitting variable when temporal proximity was included in the CT model.

For ICD to have high discriminatory power, the distances between crimes committed by the same offender have to be shorter than the distances between crimes committed by different offenders. A closer examination of ICDs in the current study,

however, suggests that this condition was not always met in the current dataset (and obviously met to a lesser degree than is the case in other studies). Figure 15 displays the location of all break and enters included in the current dataset, with the crime series of each offender identifiable by its own shape and colour combination. As illustrated by the lines connecting some offenders' series of crimes, a large number of offenders travelled relatively large distances to commit their crimes. These relatively large intra-offender ICDs would have contributed to the poor(er) performance of ICD as a linking predictor, relative to other studies.

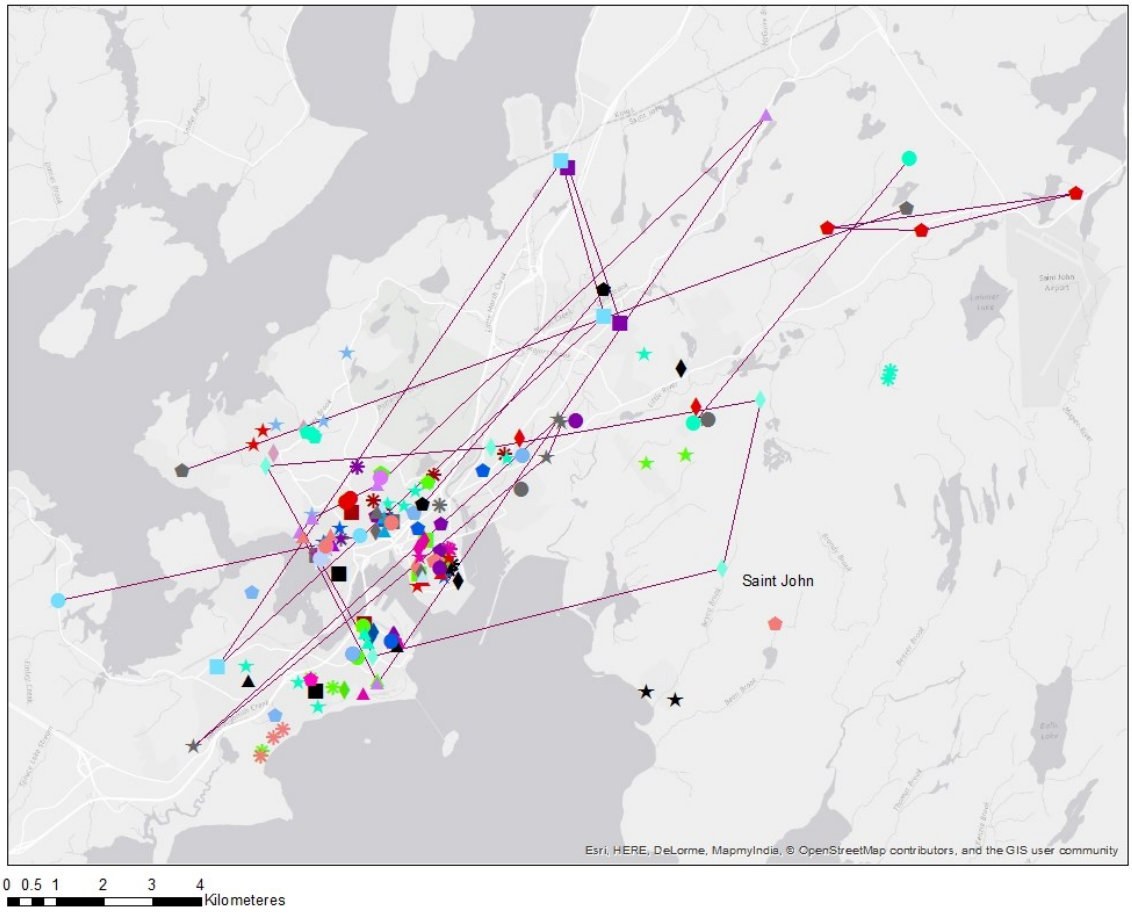


Figure 15. Map displaying the location of all Saint John break and enters.

The relatively large distances travelled by some offenders in the current study are likely the result of a number of factors. First, as mentioned, this study included crimes committed over a longer timeframe (i.e., 13 years) than most linking research that has evaluated ICD (although some previous research has employed a similar timeframe; e.g., 11 years for the sample of Finnish residential burglaries used in Tonkin, Santtila, et al., 2012 and Tonkin, Woodhams et al., 2012). This extended timeframe was necessary to obtain a large enough sample size. However, the use of a long timeframe may mean that offenders included in the study operated from different anchor points over time (e.g., they changed residences), which would likely impact the distance between their crimes.

Second, recent research conducted by Drawve, Walker, and Felson (2015) demonstrated that, as offenders age, the distance that they travel to commit their crimes increases. Drawve and his colleagues explain that this finding is likely the result of older offenders having more freedom in their choice of routine activities, which increases their awareness space and inevitably, the geographical area available to them for criminal activities. Most (63%) offenders in the current dataset were older offenders (i.e., above 18 years of age). Coupled with the longer timeframe adopted, it is possible that this is also contributing to the larger travel distances exhibited by some offenders in this study.

Third, the literature surrounding offender spatial behaviour in more recent years supports the idea that, spatially speaking, prolific property offenders commit their crimes in a similar pattern to animal foragers (e.g., Johnson & Bowers, 2004; Johnson, Summers, & Pease, 2009; Summers, Johnson, & Rengert, 2010). This research generally suggests that offenders responsible for more burglaries tend to commit a subset of them in a

certain area over a short timeframe and then, likely in an effort to decrease their detection, they move to a different area to continue their criminal activity. This behavioural pattern would lead to linked crimes that are spatially dispersed across the geographical landscape. Given that many of the studies examining residential burglary to date focus on two crimes per offender (e.g., Bennell & Jones, 2005; Markson et al., 2010; Tonkin, Santtila, et al., 2012), the fact that the current dataset includes longer series may mean that we are capturing a wider range of "foraging" behaviour, which may explain why ICD is less predictive of linkage status in the current study.[48]

Finally, there may be environmental differences between Saint John, NB and previously studied jurisdictions (e.g., Manchester, UK; Bennell & Jones, 2005) that account for the lower level of predictive accuracy associated with ICD in the current study. For example, differences in terms of the amount of green space in each location, the spread of targets, and public transportation accessibility may all contribute to the differences in findings. Moreover, the fact that Saint John, NB spans a larger geographic area (315.82 km^2) compared to jurisdictions examined in other studies may account for the lower level of predictive accuracy associated with ICD in the current study. Indeed, Burrell and colleagues (2012) found that the predictive accuracy of ICD in their study decreased substantially when analyzing data at the borough (i.e., division) level (AUC = .75) compared to the force-wide level (AUC = .92).

[48] In fact, when all crimes were included in the analyses (i.e., prolific offenders were not controlled for using the outlier approach), the predictive accuracy of the single feature logistic regression model for ICD *decreased* (AUC = .63) and it was not included as a predictor in the CT model that was developed. This finding contrasts with other research, which has found increases in predictive accuracy when all crimes are included compared to when only a subset of crimes is included (e.g., Woodhams & Labuschagne, 2012). Likewise, an increase (not decrease) in predictive accuracy was observed for all other predictors in this study when all break and enters were included in the analyses.

In order to determine the extent to which these issues might be impacting the ICD findings in the current study, future research needs to systematically examine the role that these variables play in influencing the predictive accuracy of ICD. Regardless of the explanations provided for this finding, however, the current study does suggest that there may in fact be jurisdictional differences in the utility of ICD within the linking context.

The Proposed Advantages of a CT-based Decision Support Tool for Linking Serial Break and Enters

As discussed earlier, there are many potential advantages of a CT-based versus logistic regression-based approach to linking crimes (Monahan et al., 2001; Tonkin, Woodhams, et al., 2012). More specifically, compared to the logistic regression approach, it has been argued that a CT-based approach: (1) will result in similar (standard CT) or superior (ICT and multiple CT/ICT) levels of predictive accuracy, (2) will be easier to understand in terms of the statistical processes underlying the approach, and (3) will provide a more transparent picture of the information that is considered to arrive at a final linking decision (that can also be more easily used and articulated to others). Each of these potential advantages is evaluated below in light of the findings.

Predictive accuracy of CT-based versus LR-based models. One of the proposed advantages of a CT-based approach relative to a logistic regression approach is that it will result in similar (standard CT) or superior (ICT and multiple CT/ICT) levels of predictive accuracy (Monahan et al., 2001). This was only partially supported by the current study. As expected, the same level of predictive accuracy was achieved for the stepwise logistic regression and the standard CT models, but a superior level of predictive accuracy was not achieved for the multiple CT/ICT approach (instead, all three

approaches resulted in the same level of predictive accuracy when applied to the test sample, regardless of whether temporal proximity was included in, or excluded, from the model).

A second iteration was not produced for either of the standard CT models developed (i.e., the standard CTs including and excluding temporal proximity). As such, it was not possible to examine whether ICTs result in higher levels of predictive accuracy than the standard CT or logistic regression models. This finding is consistent with Tonkin, Woodhams, et al. (2012), given that they were also unable to develop an ICT using their sample of break and enters. This begs the question: why has BLA research failed to produce ICTs like those produced by Monahan et al. (2001)? There are at least two explanations for this.

First, the patterns of behavioural consistency and distinctiveness that exist within the BLA context may not be complex enough to warrant the added intricacies of an ICT (or multiple CT/ICT) approach to linking. Second, in comparison to Monahan et al. (2001), some methodological differences exist in the current study and Tonkin, Woodhams, et al.'s (2012) study, which may account for these differences. For instance, Monahan et al. (2001) had access to a great deal of rich information on the offenders in their sample (e.g., they entered 106 risk factors into their analyses; Monahan et al., 2001), whereas (due to the severely limited data included in police files) only 6 predictors were used in the current study (and only 5 were included in Tonkin, Woodhams, et al., 2012). The fact that far fewer variables are available for analysis in BLA studies, may limit the need for statistical approaches (i.e., ICTs) that are designed to handle extremely complex relationships between variables.

The differences in the data used in BLA research versus the MacArthur study may also account for the lack of any improvements to linkage decisions using a multiple model approach to BLA. Recall that Monahan et al. (2001) were able to construct 10 multiple ICT models by identifying nine competitor variables in the CHAID analysis. To do this, they rank-ordered their 106 risk factors in terms of which factor would be selected first to split the cases if the factor that was originally selected was excluded from the analysis. They then selected nine risk factors to force as the initial splitting variable in the remaining CTs, ensuring that they did not overlap in terms of the construct being measured. Although the general methodological framework was the same in the current study, the decision was made to follow previous BLA research using a smaller number of predictors that included behaviours divided on the basis of their presumed function. It is likely that the multiple models developed and combined in the MacArthur Study are much more meaningful in nature than the current study, given that they could tap into different constructs related to violence as a result of the vast array of predictors they had at their disposal.

Importantly, despite being associated with similar AUCs, applying the two-threshold approach to the stepwise logistic regression and standard CT models demonstrated that the standard CT was able to correctly classify more cases than the logistic regression model when temporal proximity was included in the model. This was found despite the fact that similar levels of predictive accuracy were evident across these two models (i.e., the AUCs were exactly the same at test). That is, using the CT model, more linking decisions could be made without compromising accuracy. This is the opposite of what was found in Monahan et al. (2001). They found that the logistic

regression model resulted in better classification decisions using the two-threshold approach and improvements in classification abilities were only found using the ICT model. When temporal proximity was removed from the models, a similar pattern to Monahan et al. (2001) was found: although both the logistic regression and standard CT models classified considerably fewer crime pairs overall, the logistic regression model was able to correctly classify more crime pairs than the standard CT. Future research is needed to understand why these differences in results emerged. Overall, if the two-threshold approach is deemed appropriate for use in practice, it seems that a standard CT approach may lead to better linkage decisions than a logistic regression approach (when temporal proximity is included in the model).

Finally, it was also hypothesized that the shrinkage in the AUC observed from development to test would be more of an issue with the CT-based approaches than the logistic regression approach (as per the results of Tonkin, Woodhams, et al., 2012). The opposite, however, was found in the current study when temporal proximity was included in the models: the standard CT and multiple CT approaches were more robust from development to test than the logistic regression approach (although shrinkage was not significant for any approaches). In contrast, when temporal proximity was excluded from the model, a significant amount of shrinkage was observed for the multiple CT/ICT model.

It is unclear exactly why the standard CT shrinkage results differ from that of Tonkin, Woodhams, et al. (2012); however, the parameters used for CT development in the current study were slightly different, with those employed by Tonkin, Woodhams, et al. likely resulting in more complex CTs (e.g., they specified that nodes could have a

much lower number of cases [20 in parent and 6 in child nodes versus 100 and 50 in the current study], they used a less stringent *p*-value for splitting variables into nodes [$p < .05$ versus $p < .01$ in the current study], and they specified that continuous predictors could be split into a larger number of intervals [64 interval maximum versus the default of 10 intervals used in the current study]).[49] This fact may account for the differences observed in shrinkage. It is important to note, however, that when Tonkin, Woodhams, et al. (2012) varied their parameters (i.e., used less stringent parameters similar to the current study), the amount of shrinkage observed was still significant. Regardless of why the differences between studies emerged, generally speaking, shrinkage does not seem to be an issue with most of the current models, indicating that all the models developed may apply comparably well to other break and enters from Saint John (with the exception of the multiple CT/ICT model excluding temporal proximity).

Understanding of the statistical processes. A second proposed advantage of a CT-based approach over a logistic regression approach is that it will be easier for practitioners to understand the underlying statistical procedures that are applied to the data to create the models. It is difficult to determine whether this advantage holds true on the basis of the current study. Instead, this is an empirical question that can only be answered by future research. For instance, future research should present a sample of practitioners with each of the different linking models, an explanation of the statistical

[49] In fact, when the exact same parameters were applied to the current data using the same domains as Tonkin, Woodhams, et al. (2012), a much more complex CT emerged that split the cases into 24 nodes based on ICDs, entry behaviours, and property stolen. Although this more complex CT was moderately accurate at development (AUC = .86; 95% CIs = .83 to .90) significant shrinkage occurred when it was applied to the test sample (AUC = .67; 95% CIs = .61 to .74). This suggests that, at least in the current dataset, the parameters selected by Tonkin, Woodhams, et al. are likely capturing more random noise or idiosyncrasies than meaningful (or at least generalizable) patterns of behavioural consistency and distinctiveness.

processes followed to develop the models (and the rationale behind them), and then pose a series of questions to the practitioners testing their understanding of these principles.

If this were done, we believe it is likely that practitioners will find it easier to understand some components of the CT-based approach relative to the logistic regression approach. For instance, the fact that the chi-square test of independence compares the proportion of linked versus unlinked crimes across various levels of the predictor variables is likely to be somewhat easier for practitioners to grasp compared to, say, the process of maximum likelihood estimation used in logistic regression or why there is a "constant" in the logistic equation. Conversely, it is equally likely that, as the complexity of the CT-based approach increases, practitioners' understanding of the approach will decrease. As such, it is important to examine the extent to which practitioners' understanding of the various CT-based approaches varies, particularly if future research finds any predictive value to a multiple CT/ICT approach to BLA.

Transparency of the decision processes and ease of use. A third argument in support of a CT-based decision support tool is that it provides the user with a more transparent representation of the decisions made to link crimes, which is also easier to use and explain to others. Although it is true that the logistic regression equation can be visually displayed to the user in a way that is comparable to a CT (i.e., it can be just as "transparent" in an objective sense), it is likely that police practitioners (or even someone experienced in statistical analysis) would find it difficult to articulate to another individual how the logistic regression equation is used to determine the ultimate linkage decision.

For illustrative purposes, consider the stepwise logistic regression equation that was constructed using the current sample of break and enters:

$$\text{Log}\left(\frac{p}{1-p}\right) = \text{-2.82 – 2.79e-3(temporal proximity) + 1.93(internal } J\text{) – 0.15(ICD) + 1.32(target selection } J\text{)}$$

Although it is clear from this equation that temporal proximity, similarity in internal behaviours, ICD, and similarity in target selection characteristics are all used to arrive at a linkage decision (i.e., the predicted probability that is ultimately produced and compared to some predetermined threshold), it is not readily apparent from the above equation *how* the numerical components of the equation (e.g., the logit coefficients) relate to the decision that is made, nor would it be easy to articulate the meaning of the logit coefficients to others, particularly if these other individuals are lacking a background in statistical analysis techniques.

In contrast, using the standard CT produced for the break and enters, it would be relatively straightforward for a practitioner to articulate to someone else both *what* behavioural information was relied on to make a decision and also *how* the numerical information presented in the CT was used to inform their linkage decisions. To explicitly illustrate this, the CT for the test sample is presented again in Figure 16 in the manner in which it might be presented to practitioners (using the two thresholds proposed by Monahan et al. 2001).

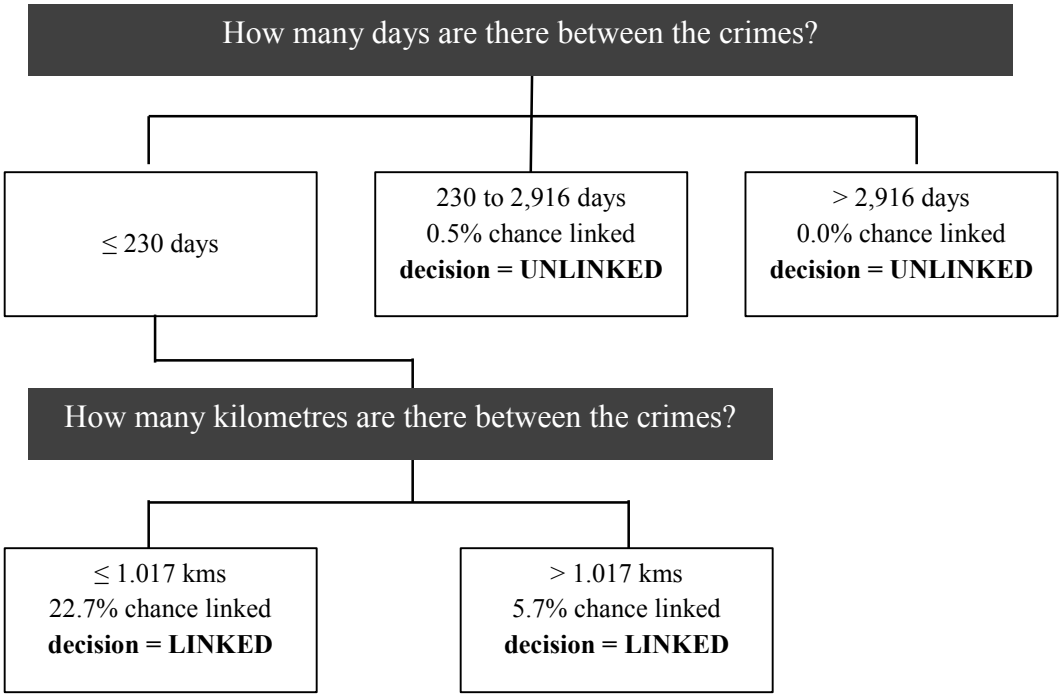


Figure 16. Standard classification tree produced using the serial break and enter data presented in a user-friendly format.

When explaining their decision process, a crime linkage practitioner would indicate that their first task was to determine the number of days between the crimes under examination. If the number of days was greater than 2,916, they would indicate that their decision was that the crimes were unlinked because there is a 0% chance of the crime pair being linked (based on the proportion of crime pairs in that node that were linked when this particular linking model was developed and applied to the validation sample). If, however, the number of days between the two crimes was less than or equal to 230, they would then indicate that they proceeded by calculating the number of kilometres between the crime pairs before making their decision. Using the thresholds proposed by Monahan et al. (2001), the practitioner would make a linked decision in both

cases (regardless of the size of the ICD). However, it is important to note that the decision may change if a different threshold was adopted in practice (e.g., crime pairs with ICDs greater than 1.017 kms may be deemed unclassifiable or unlinked, depending on the thresholds ultimately adopted). However, whether it is in fact easier for practitioners to use and articulate the decision-making process underlying a CT-based approach is an empirical question that needs to be further explored in future research by examining the usability of CT-based models with practitioners.

The Proposed Advantages of CTs for Capturing Aspects of Offending Behaviour

As also highlighted previously, another potential advantage of a CT-based approach to BLA is that it is better able to capture complexity in offending behaviour. That is, assuming that subsets of offenders do in fact differ in the way they are behaviourally consistent and distinctive (e.g., that some offenders display greater consistency in some behaviours whereas other offenders display greater consistency in other behaviours), the interactive nature inherent in a CT-based approach is better able to capture these differences relative to the main effects logistic regression approach. Moreover, it was hypothesized that such a CT-based representation of behavioural consistency and distinctiveness would reveal patterns in offending behaviour that are not captured using a main effects approach to linking, providing us with a better understanding of serial offender behaviour.

The results of this study only provide partial support for this proposed advantage of CT models in the BLA context. More specifically, using the thresholds proposed by Monahan et al. (2001) there were, in fact, two pathways to making a linked decision using the standard CT (see Figure 16), whereas there was only a single pathway to

making a linked decision when using the stepwise logistic regression model (as shown in the logistic regression equation presented above). With that said, both pathways that arose out of the standard CT relied on the same general behavioural indicators to arrive at a linked decision: temporal proximity and ICD. Moreover, the standard CT model relied on fewer predictors than the logistic regression model (two versus four predictors, respectively). Although the use of fewer predictors in the CT model can arguably be seen as an advantage, because a similar level of accuracy is achieved using a more parsimonious model, these findings do not necessarily support the assertion that the CT approach is more useful because it captures meaningful, or complex, subtypes of behavioural consistency and distinctiveness that lead to improved linkage decisions.

For instance, as discussed earlier, one of the possible advantages of CTs within the context of serial burglary is that they might be better equipped than a main effects approach to capture the behavioural consistency patterns displayed by commuters (i.e., offenders who travel large distances to commit their crimes). If this were true, however, ICD would have been split into additional nodes at higher distances and the nodes including crime pairs with large ICDs would have been split into additional sub-nodes based on other behavioural information (e.g., internal behaviours, target selection, etc.). However, the ICD variable was not split on the basis of other information when temporal proximity was included in the standard CT. Likewise, when temporal proximity was excluded from the model, only the node containing crime pairs with *shorter* ICDs was further split on the basis of internal behaviours.

Although it may be that serial break and enter behaviours are not complex enough for these types of interactions to matter, Tonkin, Woodhams, et al. (2012) found that

crime pairs with higher ICDs could be further classified on the basis of internal behaviours, with higher levels of internal behaviour similarity leading to a linked decision. There are at least two potential explanations for the different levels of complexity observed in the current study and the study conducted by Tonkin, Woodhams, and colleagues. First, as previously mentioned, the parameters used by Tonkin, Woodhams, et al. (2012) do create more complex CTs than the parameters used in the current data. Second, compared to predictors in the current study, all predictors in Tonkin, Woodhams, et al.'s study were considerably better at distinguishing linked from unlinked crime pairs. This improved predictive accuracy at the main effects level may in fact lead to an improvement in predictive accuracy when interactions are explored between the predictors as well.[50]

Overall, the simplistic nature of the CTs developed in the current study does not point to any clear patterns in offending behaviour that have previously gone unnoticed in the linking research. As mentioned, both "pathways" to making a linked decision using the CT relied on the same predictors in the current study: temporal proximity and ICD. These are the two predictors that have often been simultaneously included in stepwise logistic regression models developed in past linking research on property crimes (e.g., Markson et al., 2010; Tonkin, Santtila, et al., 2012), although additional predictor variables are occasionally also included in those models (e.g., target selection and target acquisition in Davies et al., 2012).

[50] The possible reasons for increased accuracy of some behavioural domains in Tonkin, Woodhams et al. (2012) using Finnish burglaries were discussed earlier.

Summary

Taken together, the results of this study suggest that, relative to a main effects logistic regression approach, a simple standard CT-based approach can lead to linking decisions that are similar in predictive accuracy and superior in terms of classification accuracy when using a two-threshold approach (when temporal proximity is included in the model). Moreover, it does not appear that there is any added value to creating more complex (e.g., ICT or multiple CT/ICT) CT-based models when linking serial residential break and enters, at least using the current data. Generally speaking, most of the value of a CT-based approach to linking serial break and enters seems to lie in the ease with which one can observe and possibly articulate the linking decision process to another individual. However, as mentioned, research will need to be conducted to determine if this is in fact the case with practitioners attempting to use a CT-based model.

It has been argued in the current dissertation and elsewhere (e.g., Tonkin, Woodhams, et al., 2012) that all of the potential practical advantages of the CT-based approach may result in an increased acceptance of a CT-based actuarial tool into crime linkage practice. This, however, is an empirical question that needs to be determined in future research. In addition to the research that should be conducted to examine the extent to which a CT is easier to understand and use, future research should also ask practitioners whether they would be more likely to adopt a CT-based approach into practice, and what some of the barriers may be for using either of the approaches (CT-based or logistic regression-based) in practice. The next chapter examines the extent to which the results presented here generalize to a sample of serial sexual assaults. As mentioned earlier, it may be that a CT-based approach is better suited for interpersonal

crime types given the increased complexity in sexual offending behaviour (as a result of the presence of a victim, for instance).

CHAPTER 7

Linking Serial Sexual Assaults: Results

Phase 1: Calculation and Descriptive Analysis of Similarity Scores

Similar to Chapter 6, the first step with the serial sexual assault data involved calculating *J*-scores for each behavioural domain when variables that occurred in less than 1 percent, 5 percent, 10 percent, 15 percent, and 20 percent of the sexual assaults in the sample were progressively removed from the calculation. Appendix B presents a list of the variables included in each domain and their frequency of occurrence in the reduced dataset (i.e., after the impact of prolific offenders was reduced). Similar to the break and enter data, the impact of removing lower frequency variables on the calculation of *J*-scores is consistent across all behavioural domains for the serial sexual assaults (see Table 23). Median *J*-scores tended to increase as more behaviours were removed from the calculation (with the exception of some intervals for the sexual and style domains, where median *J*'s increased at higher levels). Likewise, standard deviations tended to increase as more behaviours were removed.[51]

[51] Again, medians are presented instead of means because all distributions were significantly different from normal (*D*'s ranging from 0.08 to 0.53, all *p*'s <.001).

Table 23. Median *J*-scores and standard deviations for sexual assault crime pairs when variables at various frequency intervals are removed from the calculations.

	Behavioural Domain											
	Control Behaviours		Environmental Behaviours		Escape Behaviours		Sexual Behaviours		Style Behaviours		Victim Selection	
Removal Interval	Mdn	*SD*	Mdn	*SD*	Mdn	*SD*	Mdn	*SD*	Mdn	*SD*	Mdn	*SD*
None	.13	.21	.26	.20	.00	.09	.11	.22	.14	.20	.25	.22
≤ 1 %	.13	.21	.26	.20	-	-	-	-	.14	.20	-	-
≤ 5 %	.14	.21	.27	.20	.00	.11	.11	.22	.17	.21	.25	.22
≤ 10 %	.17	.25	.28	.20	-	-	.13	.23	.17	.25	.25	.23
≤ 15 %	.20	.30	.30	.21	-	-	.13	.25	.00	.29	.25	.24
≤ 20 %	.25	.31	.33	.22	-	-	.00	.28	.00	.34	.29	.25

Note. Empty cells correspond to instances where the domain did not contain any additional variables to remove within that particular frequency interval (e.g., the sexual behaviours domain did not contain any variables with frequencies between 0 and 1 percent).

To determine the impact of variable removal on the predictive accuracy of the behavioural domains, ROC analyses were conducted on the *J*-scores separately for each domain at the various removal intervals presented in Table 23. A forward stepwise logistic regression model was also constructed at each frequency interval, entering all the domains into the analysis.[52] ROC analyses were then conducted using the predicted probabilities from the stepwise logistic regression models.

As shown in Table 24, AUCs tended to stay the same or decrease as more behaviours were removed from the calculation of behavioural similarity (*J*-scores). Both the confidence intervals and significance tests for the AUCs for some behavioural domains indicate that the predictive accuracy decreases significantly as variables at higher frequency levels are removed (e.g., for the control domain, the AUCs were significantly lower at the 15 and 20 percent intervals versus the 5, 1, and 0 percent intervals; *z*'s ranging from 2.54 to 3.82 and *p*'s ranging from <.001 to .016). Most importantly, however, the stepwise logistic regression results did not meaningfully differ across removal intervals, both in terms of the variables included and the predictive accuracy achieved by the stepwise model (*p*'s >.05 for all AUC comparisons for the stepwise model). Given that it is ultimately the stepwise model that will be used for comparison purposes in this study, and that the AUCs were not statistically different from one another, all variables were retained in each domain when calculating similarity scores in the current study.

---

[52] Similar to the break and enter data, some domains did not have variables to remove at certain frequency intervals (e.g., the escape domain only had behaviours to remove at the 5% frequency interval level). As such, for these domains, the stepwise models included the most recently calculated *J*-scores. For example, the 1% interval stepwise model included the escape, sexual, and victim selection domains where no variables were removed. However, the escape domain was excluded altogether from the stepwise models at the 15% and 20% intervals because all escape behaviours would have been removed from the domain at these levels (i.e., all escape behaviours had frequencies that were less than 15%).

Table 24. Predictive accuracy of logistic regression models developed across all variable removal intervals for serial sexual assaults.

	Model													
	Control Behaviours		Environmental Behaviours		Escape Behaviours		Sexual Behaviours		Style Behaviours		Victim Selection		Forward Stepwise	
Removal Interval	AUC (*SE*)	95% CI	AUC (*SE*)	95% CI	AUC (*SE*)	95% CI	AUC (*SE*)	95% CI	AUC (*SE*)	95% CI	AUC (*SE*)	95% CI	AUC (*SE*)	95% CI
None	.86 (.01)	.85-.88	.81 (.01)	.78-.83	.59 (.02)	.56-.62	.76 (.02)	.73-.79	.83 (.01)	.81-.86	.87 (.01)	.86-.89	.90 (.01)	.88-.92
≤ 1 %	.86 (.01)	.85-.88	.81 (.01)	.78-.83	-	-	-	-	.83 (.01)	.81-.86	-	-	.90 (.01)	.88-.92
≤ 5 %	.87 (.01)	.85-.89	.80 (.01)	.78-.83	.59 (.02)	.56-.62	.76 (.02)	.73-.78	.82 (.01)	.80-.85	.87 (.01)	.86-.89	.90 (.01)	.88-.92
≤ 10 %	.84 (.01)	.82-.86	.81 (.01)	.78-.83	-	-	.75 (.02)	.72-.78	.82 (.01)	.79-.84	.87 (.01)	.86-.89	.90 (.01)	.88-.91
≤ 15 %	.82 (.01)	.80-.84	.80 (.01)	.77-.82	-	-	.76 (.02)	.73-.78	.78 (.01)	.76-.81	.86 (.01)	.85-.88	.89 (.01)	.87-.91
≤ 20 %	.81 (.01)	.78-.83	.80 (.01)	.77-.82	-	-	.70 (.02)	.67-.73	.77 (.01)	.75-.80	.86 (.01)	.85-.88	.88 (.01)	.86-.90

Note. All *p's* <.001

Final descriptive statistics for linked and unlinked sexual assaults are displayed in Table 25 as a function of each individual domain included in subsequent analyses.[53] These distributions were also plotted graphically (Figure 17). As shown, the distributions for linked and unlinked sexual assault pairs overlapped to a lesser degree than the distributions for the serial break and enter data. A substantial proportion of linked pairs had a *J*-score of 1 for each behavioural domain and a substantial proportion of unlinked pairs had a *J*-score of 0 for each behavioural domain. This suggests that a high level of discrimination accuracy will be achieved using the serial sexual assault data. However, the overlap in distributions (Figure 17) does suggest that perfect discrimination accuracy is unlikely using the current sample of serial sexual assaults.

Table 25. Descriptive statistics for all linked and unlinked serial sexual assaults.

	Mdn (*SD*)		Range	
Domains (Measurement)	L	UL	L	UL
Control Behaviours *(J)*	.67 (.36)	.13 (.19)	.00 – 1.00	.00 – 1.00
Environmental Behaviours (*J*)	.85 (.34)	.26 (.19)	.00 – 1.00	.00 – 1.00
Escape Behaviours (*J*)	.00 (.39)	.00 (.07)	.00 – 1.00	.00 – 1.00
Sexual Behaviours (*J*)	.80 (.43)	.10 (.21)	.00 – 1.00	.00 – 1.00
Style Behaviours (*J*)	1.00 (.38)	.14 (.19)	.00 – 1.00	.00 – 1.00
Victim Selection (*J*)	.67 (.29)	.25 (.21)	.10 – 1.00	.00 – 1.00

Note. *J* = Jaccard's coefficient; L = linked crime pairs; UL = unlinked crime pairs.

[53] Again, medians instead of means were displayed given that all linked and unlinked distributions were significantly different from normal (*D*'s ranged from 0.10 to 0.53; all *p*'s <.001).

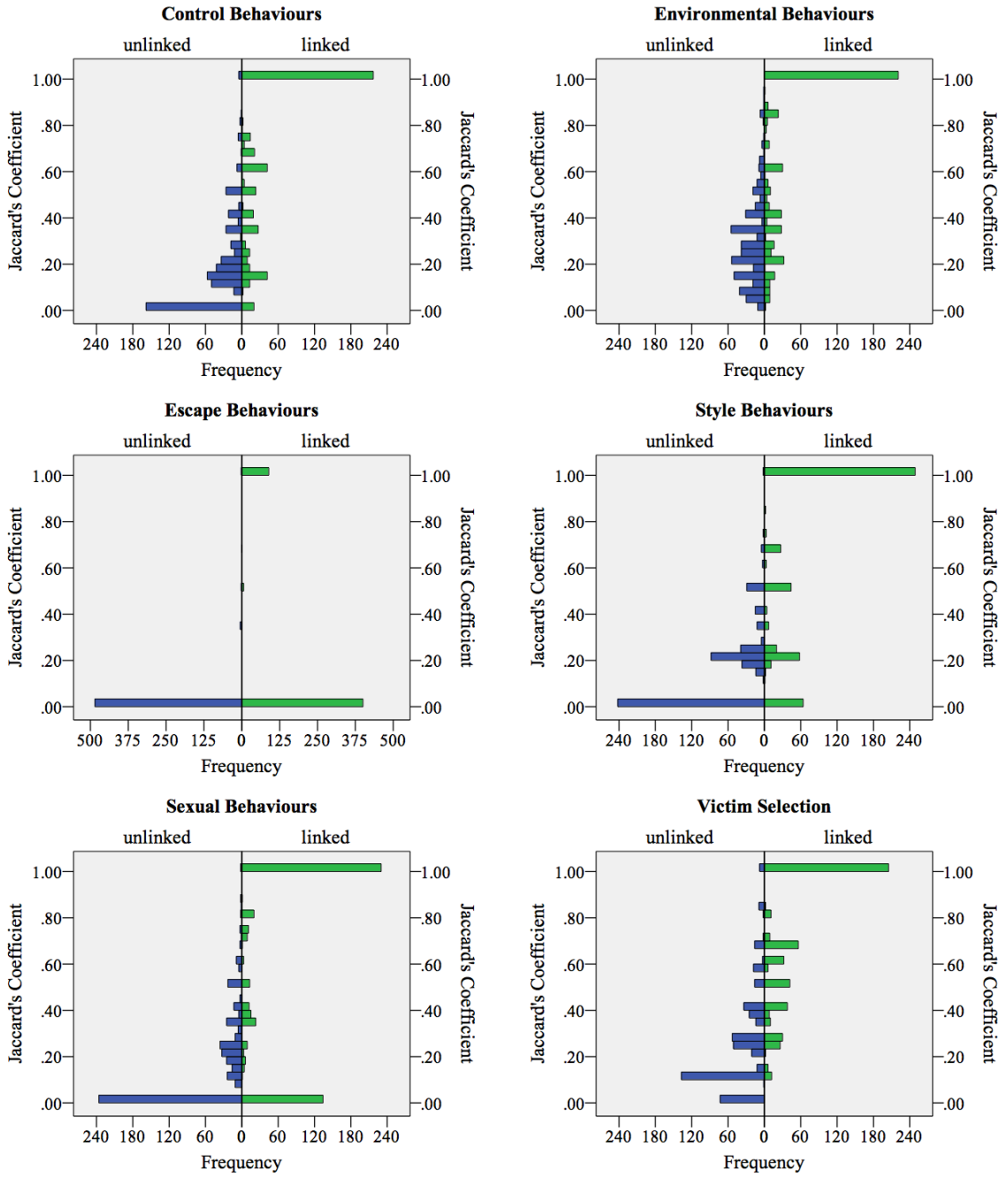


Figure 17. Frequency distributions displaying the range of scores for all linked sexual assaults and an equal random sample of unlinked sexual assaults (n = 495) for each domain included in subsequent analyses.

Significance tests were then carried out to systematically examine the abovementioned differences between linked and unlinked crime pairs. Similar to the break and enter data, a subset of unlinked crime pairs equal to the total number of linked crime pairs (n = 495) was randomly extracted from the complete dataset to facilitate paired samples analyses. As mentioned (and evident in Figure 17), Kolmogorov-Smirnov tests of normality confirmed that both the linked and unlinked distributions for each behavioural domain were significantly different from normal (D's ranged from 0.10 to 0.53; all p's <.001). As a result, Wilcoxon signed-rank tests were conducted to examine the median differences between the linked and unlinked distributions for each domain.

Results of these analyses are presented below in Table 26. As demonstrated, J-scores were significantly larger for linked crime pairs than unlinked crime pairs for all behavioural domains. As per Cohen's (1988, 1992) guidelines, all effect sizes were in the large range (r >.50), with the exception of the escape domain (r = .41). The control and victim selection domains had the largest effect sizes of all domains (r's = .76).

Table 26. Results of the non-parametric comparisons of similarity scores for linked and unlinked serial sexual assault across each linking feature/domain.[54]

	Mdn (*J*)			
Domains	L	UL	Wilcoxon (*z*)	Effect Size (*r*)
Control Behaviours	.67	.14	-16.83	.76
Environmental Behaviours	.85	.26	-16.08	.72
Escape Behaviours	.00	.00	-9.10	.41
Sexual Behaviours	1.00	.10	-15.35	.69
Style Behaviours	.80	.14	-16.54	.74
Victim Selection	.67	.25	-16.87	.76

Note. *J* = Jaccard's coefficient; L = linked crime pairs; UL = unlinked crime pairs; $r = z/\sqrt{N}$ (Field, 2013; Rosenthal, 1991); all *p*'s <.001.

Phase 2: Developing and Evaluating the Main Effects Linking Models

Next, the main effects logistic regression models were developed and evaluated.

Logistic regression assumptions. As mentioned, a minimum of 10 cases in the smallest group being predicted (i.e., the linked group) is needed for each predictor variable included in the analysis (Peduzzi et al., 1996). This requirement was met for all logistic regression models developed: a maximum number of 6 predictors were entered into the stepwise logistic regression model, requiring at least 60 linked cases (linked *n* = 248 for the development sample).

[54] Although the results in Table 26 seem strange for escape behaviours (i.e., the median for both linked and unlinked pairs is equal to zero, but the effect size is moderate), this can be explained by the statistical test used to compare the distributions presented in Figure 17. The Wilcoxon signed-rank test is a test of the *median difference score*, not the medians themselves. That is, it tests the null hypothesis that the *median of ranked difference scores* is equal to zero, after difference scores equal to zero are excluded from the calculation of the test statistic (Field, 2013). As shown in Figure 17, many difference scores (that is, J_{linked}-$J_{unlinked}$ for the linked vs. unlinked distributions) would be equal to zero. The remaining difference scores, however, are different from zero (most remaining *J*-scores for linked crimes are equal to 1, whereas most remaining *J*-scores for unlinked crimes are equal to zero), which explains why the effect in Table 26 was found.

Bivariate correlations were also calculated between all predictors to assess multicollinearity. As shown in Table 27, correlations ranged from .05 to .31 (all *p*'s <.001). The largest correlation was between the victim selection and control domains (*r* = .31). The fact that no correlations were in the large range (i.e., did not exceed .80), suggested that multicollinearity and singularity were not a concern with the serial sexual assault data (Tabachnick & Fidell, 2007).[55] A linear regression was also run where all predictors were entered into the model in order to obtain collinearity statistics (Field, 2013). These indices confirmed that multicollinearity was not present in the data - all VIFs were less than 10 (Bowermann & O'Connell, 1990). Specifically, VIFs ranged from 1.06 (environmental domain) to 1.16 (control domain). Likewise, all tolerance values were well above Menard's (1995) guideline of .02 for problematic tolerance levels (ranging from .82 for victim selection to .95 for the environmental domain).

Table 27. Correlations between all domains included in the serial sexual assault analyses.

Domains	1	2	3	4	5
1. Control Behaviours	-				
2. Environmental Behaviours	.16	-			
3. Escape Behaviours	.14	.11	-		
4. Sexual Behaviours	.16	.05	.14	-	
5. Style Behaviours	.20	.13	.15	.23	-
6. Victim Selection	.31	.19	.24	.20	.19

The Box-Tidwell approach was again used to assess linearity in the logit. Similar to the serial break and enter data, a constant of 2 was added to the raw variables prior to

[55] A similar pattern was found when the correlations were examined for each of the behavioural domains within the linked and unlinked crime pairs separately.

calculating each variable's natural log given that zero values were present for all variables (Tabachnick & Fidell, 2007). Table 28 presents the relevant output testing the linearity in the logit assumption (the interaction between each raw variable and its natural logarithm) for each linking feature using all cases in the dataset. As shown, all predictors, with the exception of the control domain, violated the linearity in the logit assumption (i.e., the interaction terms were significant).[56]

Table 28. Interaction terms testing the linearity in the logit assumption for the serial sexual assault data.

Interaction Term	*b* (*SE*)	Wald (*df*)	*p*-value
Control Behaviours x Ln	-1.354 (2.037)	0.44 (1)	.506
Environmental Behaviours x Ln	7.188 (2.510)	8.20 (1)	.004
Escape Behaviours x Ln	27.563 (5.217)	27.91 (1)	<.001
Sexual Behaviours x Ln	20.987 (1.932)	117.99 (1)	<.001
Style Behaviours x Ln	3.536 (1.829)	3.73 (1)	.053
Victim Selection x Ln	-6.890 (2.495)	7.63 (1)	.006

In an attempt to address these violations, square root, logarithmic, and inverse transformations were applied to all offending variables and a number of simple and forward stepwise logistic regression models were developed using the untransformed and transformed data.[57] Using the predicted probabilities calculated for these various models,

[56] The Style x Style Ln interaction term approached significance when all data was used ($p = .053$). However, when examining the linearity in the logit assumption using the development sample data only, the *p*-value for this interaction term decreased ($p = .004$). Given this, the impact of transforming the style domain was also examined. For all other domains, no differences were found when the assumption was tested on the development sample data only.

[57] Again, a constant of 2 was added to each variable prior to applying the transformations given the occurrence of zero scores.

ROC analyses were then conducted to compare the predictions made by the logistic regression models using the transformed versus untransformed variables. As expected, the parameter estimates varied across all models; however, the significance levels for each predictor's coefficient (Wald's test) and the variables included or excluded from the forward stepwise logistic regression model did not vary as a function of whether transformed or untransformed data was used. The AUCs, standard errors, and their associated 95% confidence intervals were exactly the same for the simple and forward stepwise logistic regression analyses using the untransformed, square root, logarithmic, and inverse transformed variables. The same results were found when the regression models were constructed using all cases versus only those cases in the development sample.[58] As a result, the untransformed variables were used in all subsequent analyses.

Simple and forward stepwise logistic regression analyses. The results of the single linking feature logistic regression models constructed using the development sample data are presented in Table 29. As shown, all model chi-square tests were statistically significant, suggesting that each predictor was able to distinguish linked from unlinked crime pairs at a level greater than the constant only model. The signs of the regression coefficients for each predictor were also in the expected direction, suggesting that linked crime pairs were characterized by higher degrees of behavioural similarity than unlinked crime pairs across all behavioural domains. Wald statistics indicated that all regression coefficients were significantly different from zero, further suggesting that all linking features were significant predictors of linkage status. Finally, Nagelkerke's R^2

[58] Given that only the relative differences between these AUCs, standard errors, and confidence intervals are important to these assumption tests (rather than their absolute values), a table presenting this information was not included here in an effort to avoid redundancy (since the values for the untransformed and all transformed analyses were identical).

values indicated that the style domain performed best relative to the other single linking feature models (R_N^2 = .60); this was followed by the control, victim selection, environmental, sexual, and escape domains (R_N^2 range = .11 to .28). Hosmer and Lemeshow's R_L^2 values suggest a similar pattern to the predictors, only slightly deviating from the Nagelkerke indices. Overall, these results suggest that most domains are moderate predictors of linkage status, with the exception of the escape domain.

Table 29. Results of separate simple logistic regression analyses for each predictor included in the serial sexual assault sample.

Model	Constant (*SE*)	*B* (*SE*)	Wald (*df*)	χ^2 (*df*)	R_N^2	R_L^2
Control	-6.10 (0.14)	5.39 (0.21)	677.16 (1)	672.89 (1)	.28	.26
Environmental	-6.93 (0.18)	6.05 (0.26)	538.64 (1)	566.13 (1)	.23	.22
Escape	-4.46 (0.07)	4.33 (0.22)	391.48 (1)	261.45 (1)	.11	.10
Sexual	-5.74 (0.13)	4.70 (0.21)	508.64 (1)	484.32 (1)	.20	.19
Style	-6.00 (0.13)	5.36 (0.20)	728.49 (1)	709.36 (1)	.29	.27
Victim Selection	-6.88 (0.17)	5.94 (0.25)	582.28 (1)	638.83 (1)	.26	.25

Note. χ^2 = model chi-square; R_N^2 = Nagelkerke index; R_L^2 = Hosmer and Lemeshow's index; all *p*'s < .001.

Next, a forward stepwise logistic regression model (with each individual behavioural domain entered) was constructed using the development sample to determine the optimal combination of variables for predicting linkage status. As shown in Table 30, the analysis proceeded through six steps, including style behaviours, control behaviours, victim selection, sexual behaviours, environmental behaviours, and escape behaviours in the final model. The model chi-square test was statistically significant, suggesting that the predictors were able to distinguish linked from unlinked sexual assault pairs at a level

greater than the constant only model. Wald statistics for all predictors were also significant, suggesting that all linking features included in each stepwise model were significant predictors of linkage status (all *p*'s <.001). Similar R_N^2 and R_L^2 values were found for the model (.50 and .48, respectively). These R^2 indices suggest improved discrimination accuracy relative to that achieved by any of the single linking feature models alone.

Table 30. Results of the forward stepwise logistic regression analyses performed on the sexual assault development sample.

Model	B (*SE*)	Wald (*df*)	χ^2 (*df*)	R_N^2	R_L^2
Stepwise			1243.04 (6)	.50	.48
Style	2.65 (0.28)	88.79 (1)			
Control	3.01 (0.29)	104.35 (1)			
Victim Selection	1.82 (0.35)	26.90 (1)			
Sexual	1.77 (0.29)	37.18 (1)			
Environmental	1.83 (0.34)	28.63 (1)			
Escape	1.76 (0.36)	24.70 (1)			
Constant	-8.02 (0.22)	1364.08 (1)			

Note. χ^2 = model chi-square; R_N^2 = Nagelkerke index; R_L^2 = Hosmer and Lemeshow's index; all *p*'s <.001.

ROC analyses. The predicted probabilities from each of the models presented in Tables 29 and 30 were then entered into separate ROC analyses (seven ROC analyses were completed in total). As explained previously, separate ROC curves were constructed for the development and test samples in order to examine the shrinkage in the AUCs and provide an estimate the generalizability of each model. The ROC curves are visually

displayed in Figure 18 and the AUCs and their 95% confidence intervals are presented in Table 31.

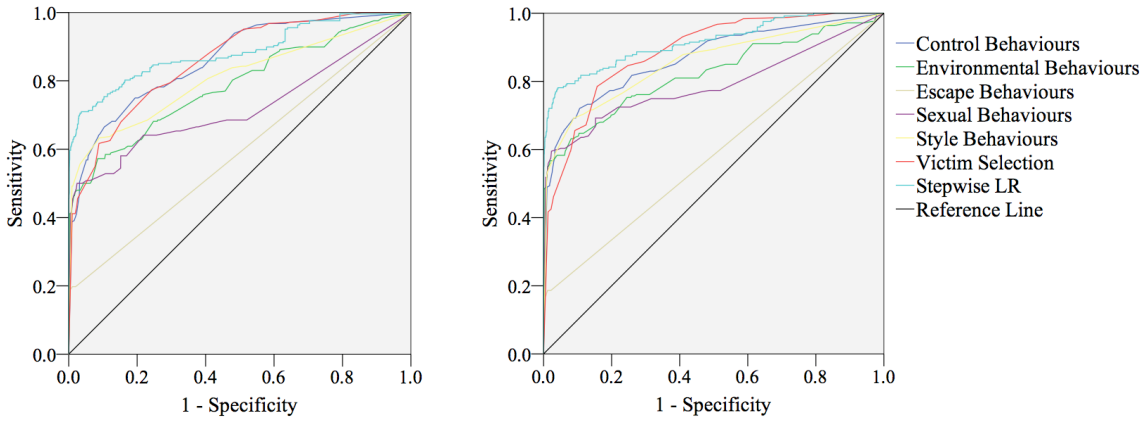


Figure 18. ROC curves for the development (left) and test (right) samples displaying the discrimination accuracy for all logistic regression models constructed using the serial sexual assault data.

Table 31. Development and test sample AUCs and their associated 95% confidence intervals for all logistic regression models constructed using the sexual assault data.

	Development Sample		Test Sample	
Model	AUC (*SE*)	95% CI	AUC (*SE*)	95% CI
Control	.86 (.01)	.84 – .89	.87 (.01)	.84 – .90
Environmental	.79 (.02)	.75 – .82	.82 (.02)	.79 – .86
Escape	.59 (.02)	.55 – .63	.58 (.02)	.54 – .63
Sexual	.72 (.02)	.68 – .76	.79 (.02)	.75 – .83
Style	.81 (.02)	.77 – .84	.86 (.02)	.83 – .89
Victim Selection	.86 (.01)	.83 – .88	.89 (.01)	.87 – .91
Stepwise	.88 (.01)	.86 – .91	.92 (.01)	.89 – .94

Note. All p's <.001.

As demonstrated in Figure 18 and Table 31, the ROC results mirror the results of the logistic regression analyses – with the stepwise model producing the largest AUC value (development AUC = .88, moderate predictive accuracy; Swets, 1988). All other domains were associated with moderate levels of predictive accuracy (development AUCs ranging from .72 to .86), with the exception of the escape domain, which had a low level of predictive accuracy (development AUC = .59).

For the development sample AUCs, there were no significant differences between the AUCs for the three top models (stepwise, control, and victim selection, all p's >.05). The stepwise model was significantly higher than all other AUCs (environmental, escape, sexual, and style; z's ranging from 3.20 to 12.19, all p's <.001). The control and victim selection domains also resulted in significantly higher levels of predictive accuracy than

the environmental, escape, sexual, and style domains (*z*'s ranging from 2.23 to 11.10, *p*'s ranging from <.001 to .026). Both the style and environmental domain achieved a significantly higher level of predictive accuracy than the sexual and escape domains (*z*'s ranging from 2.77 to 8.66, *p*'s ranging from <.001 to .006). Finally, the sexual domain achieved a significantly higher level of predictive accuracy than the escape domain ($z = 4.88$, $p<.001$). All other comparisons were not statistically significant (*p*'s >.05).

For the test sample, with the exception of the victim selection AUC, the AUC for the stepwise model was significantly higher than the AUCs for all other models (*z*'s ranging from 2.64 to 15.06, *p*'s ranging from <.001 to .008). The victim selection and control models significantly outperformed the environmental, sexual, and escape models (*z*'s ranging from 2.27 to 13.17, *p*'s ranging from <.001 to .023). The style model outperformed the sexual ($z = 3.06$, $p =.002$) and escape models ($z =11.50$, $p <.001$). Although the environmental and sexual models were statistically similar in their levels of predictive accuracy, both significantly outperformed the escape model (*z*'s = 9.51 and 8.15, respectively, both *p*'s <.001). All other comparisons were not statistically significant.

When examining the pattern of AUCs across the development and test samples, slight (and non-significant) shrinkage was only evident for the escape domain (shrinkage = .01, $p >.05$). In fact, all other AUCs increased from development to test (increases ranging from .01 to .07). For example, the stepwise model moved from a moderate level of predictive accuracy (AUC = .88) to a high level of predictive accuracy (AUC = .92) and this increase was statistically significant ($z = 2.16$, $p = .031$). Statistically significant increases in predictive accuracy were also found the sexual ($z = 2.76$, $p =.006$) and style

($z = 2.22$, $p = .026$) domains. All other increases were not statistically significant (all p's >.05). Generally speaking, all logistic regression models seemed to generalize well to the test sample crime pairs.

Phase 3: Developing and Evaluating Standard CT and ICT Models

The next phase involved the development and evaluation of CT-based models.[59]

CHAID analyses. Similar to the analyses using the break and enter data, the CHAID algorithm was applied to the development sample and the resulting standard CT was simultaneously forced on the test sample. Specifically, the following predictors/linking features were entered into the CHAID analysis: control behaviours, environmental behaviours, escape behaviours, sexual behaviours, style behaviours, and victim selection characteristics.

The standard CT that resulted from this first iteration of CHAID is presented in Figure 19 for the development sample.[60] As shown, the standard CT produced using the serial sexual assault data was considerably more complex than that produced using the serial break and enter data. The CT had 4 levels with 37 nodes, 23 of which were terminal nodes. All variables appeared at least once in the CT, with the exception of the escape domain. The control domain was the first predictor selected, splitting the crime pairs into six nodes. At the second level, Nodes 1 through 3 were further split using the victim selection domain, Node 4 was further split using the sexual domain, and Nodes 5 and 6 were split using the style domain. At the third level, Nodes 8, 12, and 19 were further

[59] CHAID parameters and sample size requirements (assumptions) were previously discussed in the methodology section (Chapter 5).

[60] Given that the CT produced is large and illegible in its original format, the CT for the development sample only was presented for visualization purposes. The extent to which the CT applies to the test sample can be inferred from subsequent AUC analyses and the table examining the classification ability of the CTs when applying the two thresholds to the test sample.

split using the sexual domain, whereas Nodes 18 and 20 were further split using the victim selection domain. The environmental domain was then used to split the last node included in the third level (Node 22), and was the only variable that was used at the fourth level, further splitting the crime pairs included in Node 32.

Overall, victim selection was the most commonly selected domain in the CT: splitting cases in three areas at the second level of the tree and two areas at the third level of the CT. The sexual domain was also selected quite frequently: splitting cases in one area at the second level and three areas at the third level of the tree. The style domain split cases in two areas at the second level of the CT, and the environmental domain split cases at one area of the third level and was the only variable splitting cases at the fourth level of the CT. The control domain only appeared once in the CT, as the initial splitting variable.

The base rate of linked crime pairs in the development and test samples was 1.50%. Following Monahan et al. (2001), terminal nodes containing greater than 3.00% linked crime pairs were classified as linked, crime pairs containing less than 0.75% linked crime pairs were classified as unlinked, and nodes containing a proportion of linked crime pairs that was equal to or fell between these two thresholds (0.75–3.00%) were deemed unclassified. For the development sample, this resulted in eight linked nodes (Nodes 10, 14, 25, 28, 30, 33, 34, and 36), nine unlinked nodes (Nodes 7, 9, 11, 15, 23, 24, 26, 27, and 29), and six unclassified node (Nodes 13, 16, 17, 21, 31, and 35).

A similar pattern was found when applying these cut-offs to the test sample, although some nodes were labelled differently. One node moved from being labelled linked to unclassified (Node 30), two nodes moved from being unclassified to unlinked

(Nodes 17 and 35), and one node moved from being labelled unlinked to unclassified (Node 27). Importantly, no nodes moved from being linked to unlinked (or vice versa) when the model was applied to the test sample.

In total, 2,556 crime pairs (15.18%) in the development sample and 1,220 crime pairs (7.25%) in the test sample were deemed unclassified in this standard CT. To construct the ICT, the CHAID algorithm was applied to these unclassified cases a second time. Given that a much simpler CT was produced for this second iteration, Iteration 2 is displayed in Figure 20 for both the development and test sample.

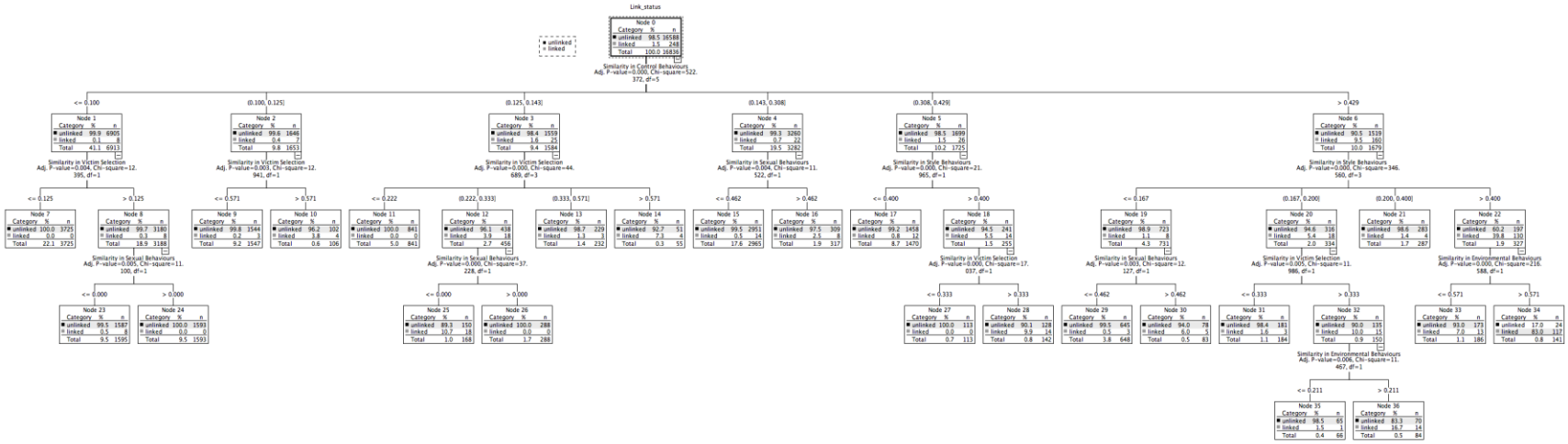
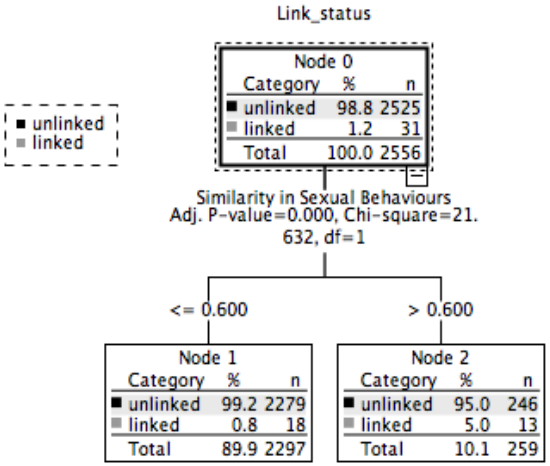
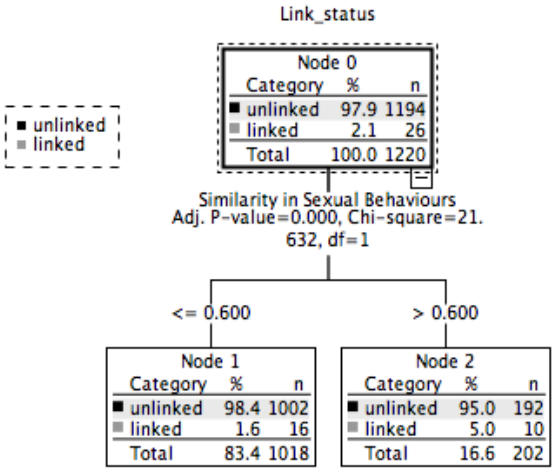


Figure 19. Iteration 1 (the standard CT) of the CHAID analyses for the sexual assault data development sample.

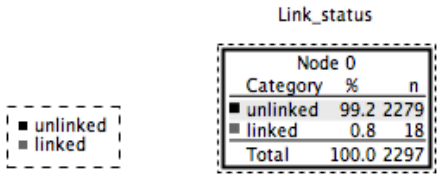
Iteration 2: Development Sample



Iteration 2: Test Sample



Iteration 3: Development Sample



Iteration 3: Test Sample

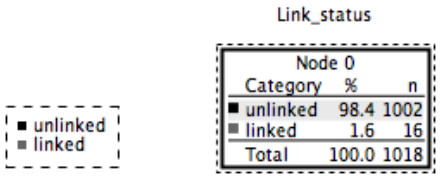


Figure 20. Iterations 2 and 3 of the CHAID analyses for the serial sexual assault data.

As shown in Figure 20, the second iteration produced a much simpler CT with two nodes, both of which were terminal nodes. The sexual domain was selected to split the crime pairs included in the second iteration. Using the same classification cut-offs as Iteration 1, this resulted in one unclassified node (Node 1) and one linked node (Node 2) for both the development and test sample CTs. As such, the ICT was able to classify an additional 259 sexual assault pairs (1.50%) in the development sample and 202 (1.20%) of the pairs in the test sample (as compared to the standard CT).[61]

The unclassified cases were run through another CHAID analysis; however, as shown in Figure 20, no further cases could be classified in Iteration 3. As such, the final ICT consisted of two CHAID iterations. Overall, the two CTs that comprised the ICT make logical sense in that most nodes classified as linked were associated with higher *J*-scores for the predictors included in the CT models than nodes classified as unlinked. However, for Iteration 1, there was one split at the second level where lower levels of similarity in sexual behaviours resulted in a linked decision (for both the development and test samples).

ROC analyses. The predicted probabilities from each of the CT-based models (standard CT and ICT) were then entered into a series of ROC analyses, conducted separately for the development and test samples. The ROC curves are visually displayed in Figure 21 and the AUCs, standard errors, and associated 95% confidence intervals are presented in Table 32.

[61] The classification abilities of all statistical approaches when adopting the two thresholds are compared in more detail on pp. 164-166.

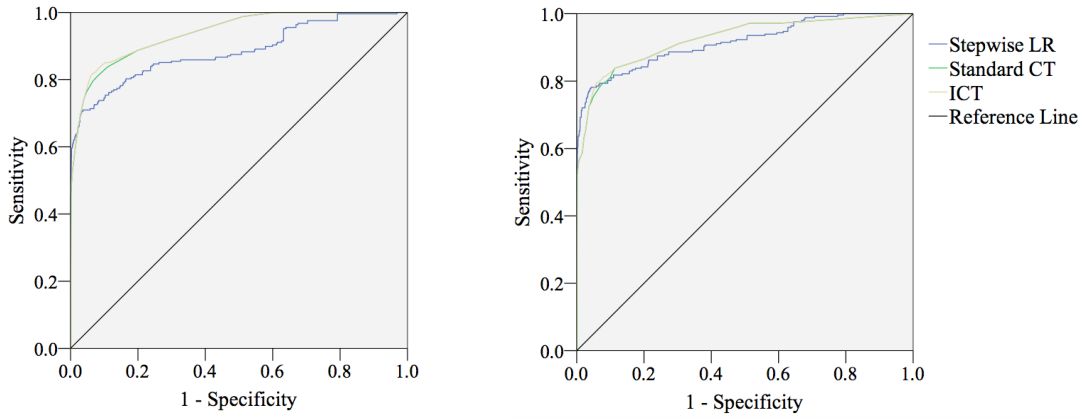


Figure 21. ROC curves for the development (left) and test samples (right) displaying the discrimination accuracy for the logistic regression, standard CT, and ICT models constructed using the serial sexual assault data.

Table 32. Development and test sample AUCs, standard errors, and their associated 95% confidence intervals for the stepwise logistic regression model and CT-based models for the serial sexual assault data.

	Development Sample		Test Sample	
Model	AUC (*SE*)	95% CI	AUC (*SE*)	95% CI
Stepwise LR	.88 (.01)	.86 – .91	.92 (.01)	.89 – .94
Standard CT	.94 (.01)	.92 – .95	.92 (.01)	.90 – .95
ICT	.94 (.01)	.93 – .96	.93 (.01)	.90 – .95

Note. All *p*'s <.001.

As shown in Figure 21 and Table 32, all CT-based models were associated with a high level of predictive accuracy (Swets, 1988). Both the standard CT and ICT were associated with high levels of predictive accuracy in both the development and test samples. Comparisons of the AUCs *across* the development and test samples suggest that generalizability was not an issue with any of the CT-based models. Although AUCs

decreased from development to test for the standard CT model (shrinkage = .02), and the ICT model (shrinkage = .01), these differences were not statistically significant (all *p*'s >.05). These results suggest that all CT-based models generalize relatively well to the test sample crime pairs.

AUCs were also compared to assess the relative performance of CT-based models to logistic regression models. For the development sample, both CT-based models resulted in significantly higher levels of predictive accuracy (both AUCs = .94) than the stepwise logistic regression model (AUC = .88; $z = 3.40$, $p < .001$). For the test sample, however, all three models had equally high levels of predictive accuracy (all AUCs were above .90 and all *p*'s >.05). Although shrinkage was not a large concern with the CT-based models, the fact that the AUC significantly increased from development to test for the logistic regression model ($z = 2.16$, $p = .031$), whereas it decreased slightly for the CT-based models, does suggest that shrinkage may be more of a concern with CT-based versus regression-based linking models.

Comparing classification abilities. The two-threshold classification approach was also applied to the logistic regression model in order to compare classification abilities across regression and CT-based models. Specifically, the number of crime pairs labelled as linked (i.e., those with predicted probabilities > 3.00%), unlinked (i.e., those with predicted probabilities < 0.75%), and unclassified (i.e., those with predicted probabilities equal to or between 0.75 and 3.00%) by each model was calculated and compared. Classification abilities for all three models (stepwise logistic regression,

standard CT, and ICT) are presented in Table 33 as a function of the observed state (i.e., whether the crime pairs were truly linked or unlinked).[62]

As shown in Table 33, the logistic regression, CT, and ICT models were able to classify a total of 85.02%, 92.75%, and 93.95% crime pairs as linked or unlinked, respectively. The logistic regression, CT, and ICT models *correctly* classified a total of 79.91%, 88.50%, and 88.50% of the unlinked crime pairs, respectively. In contrast, a total of 78.14%, 74.09%, and 79.91% linked crime pairs were *correctly* classified by the logistic regression, CT, and ICT models, respectively. The proportion of cases overall that were *correctly* classified was highest in the ICT model (88.35%), followed by the CT model (88.29%), and the logistic regression model (79.88%).

[62] Only the test sample data is displayed and discussed given that the previous ROC analyses demonstrate that all three models cross-validate.

Table 33. Classification abilities for the test sample when applying the two-threshold approach to the stepwise logistic regression, standard CT, and ICT models developed using the serial sexual assault data.

	Logistic Regression Model			
Observed	Unlinked	Unclassified	Linked	Percent CC[a]
Unlinked	13,254	2,506	827	79.91
Linked	39	15	193	78.14
Total	13,293	2,521	1,020	79.88
	Standard Classification Tree Model			
Observed	Unlinked	Unclassified	Linked	Percent CC[a]
Unlinked	14,680	1,194	713	88.50
Linked	38	26	183	74.09
Total	14,718	1,220	896	88.29
	Iterative Classification Tree Model			
Observed	Unlinked	Unclassified	Linked	Percent CC[a]
Unlinked	14,680	1,002	905	88.50
Linked	38	16	193	79.91
Total	14,718	1,018	1,098	88.35

Note. Total N = 16,834; linked n = 247; unlinked n = 16,587.

[a] Percent CC = percent correctly classified for unlinked pairs, linked pairs, and overall (total).

Phase 4: Developing and Evaluating Multiple CT/ICT Linking Models

The final phase involved constructing the CT-based multiple models. Six separate CT/ICT models were first developed: each one forcing a different predictor as the initial splitting variable in the first iteration of each CT. The ICT with control behaviours forced as the first variable was constructed in Phase 3 and is displayed in Figures 19 and 20. The five additional CTs (each beginning with a different predictor) and their second iterations are presented in Figures D1 through D9 in Appendix D. All CTs proceeded through two

iterations, with the exception of the CT where the style domain was forced as the first variable (only a standard CT was produced). The characteristics of the CTs produced in Iteration 1 and Iteration 2 are summarized in Table 34.

Table 34. Characteristics of the multiple classification trees produced in Iteration 1 and Iteration 2 when each predictor was forced as the initial splitting variable for the first iteration of each tree developed using the serial sexual assault data.

First Variable	Depth	Nodes (#)	Terminal Nodes (#)	Variables Included
Iteration 1 CT				
Control	4	37	23	All Except Escape
Environmental	5	38	24	All Except Escape
Escape	5	34	21	All Variables
Sexual	4	43	28	All Except Escape
Style	5	35	21	All Except Escape
Victim Selection	4	34	22	All Except Escape
Iteration 2 CT				
Control	1	3	2	Sexual
Environmental	1	4	3	Sexual
Escape	1	3	2	Sexual
Sexual	1	3	2	Environmental
Victim Selection	1	3	2	Style

Note. A second iteration was not produced when the style domain was forced as the initial splitting variable at the first iteration.

The CTs produced using the serial sexual assault data were much more complex than those produced using the break and enter data. In Iteration 1, the depth of each CT was 4 or 5 levels, the number of nodes in each tree ranged from 34 to 43, the number of terminal nodes in each CT ranged from 21 to 28, and all variables were included in every

CT, with the exception of the escape domain. The only CT that the escape domain was included in was the CT where escape was forced as the first splitting variable.

All second iterations were much less complex, with the depth of each CT equal to 1, the number of nodes in each CT ranging from 3 to 4, and the number of terminal nodes ranging from 2 to 3. Most CTs produced at Iteration 2 were split on the basis of the sexual domain, with the exception of the sexual domain CT (where the Iteration 2 CT was split based on the environmental domain) and the victim selection CT (where the Iteration 2 CT was split based on the style domain).

The predicted probabilities from each single CT/ICT developed at this stage were then submitted to separate ROC analyses. Figure 22 displays the ROC curves achieved for the six CT/ICTs. Each model's AUC and its overall ability to classify cases into linked versus unlinked subgroups is displayed in Table 35. As demonstrated, most CT/ICTs classified a high number of the crime pairs as linked or unlinked (ranging from 86.4 to 92.3 percent in the development sample and 76.1 to 93.9 percent in the test sample). Predictive accuracies for the CT/ICTs were also high for the development sample (AUCs ranging from .93 to .95) and the test sample (AUCs ranging from .90 to .93). The AUCs were not significantly different from one another in either the test or the development sample. Likewise, although the AUCs decreased from development to test for all models (shrinkage in the AUC ranged from .01 to .04), the only model with a statistically significant amount of shrinkage was the victim selection CT model ($z = 2.38$, p=.017).

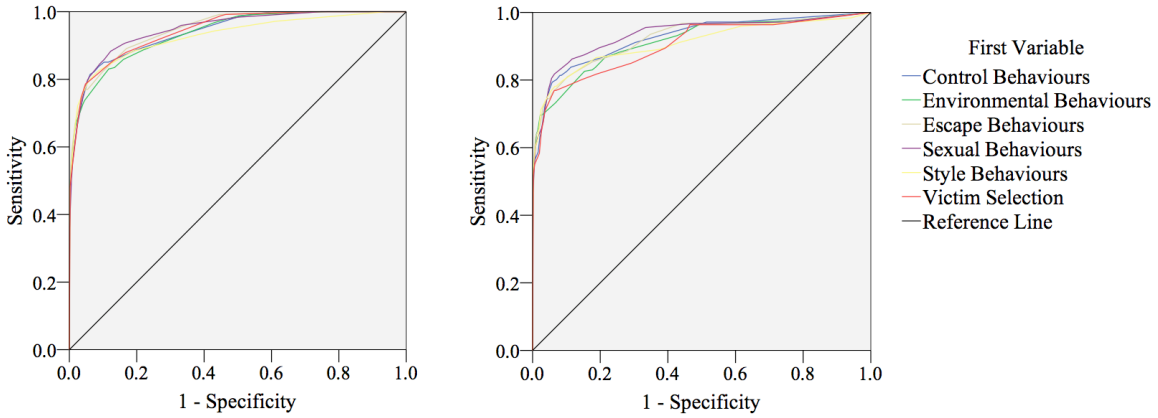


Figure 22. ROC curves for the development (left) and test samples (right) displaying the discrimination accuracy for the multiple CT/ICT models that were developed by forcing each predictor as the first splitting variable in the first iteration of the CHAID analyses conducted using the serial sexual assault data.

Table 35. Predictive accuracies and classification abilities of the CT/ICT models where each predictor was forced as the first splitting variable in the CHAID analyses conducted using the serial sexual assault data.

First Variable	Classified (%)	AUC (*SE*)	95% CI
Development Sample			
Control	86.40	.94 (.01)	.93 – .96
Environmental	88.10	.94 (.01)	.92 – .95
Escape	87.80	.95 (.01)	.93 – .96
Sexual	92.30	.95 (.01)	.93 – .96
Style	86.80	.93 (.01)	.91 – .95
Victim Selection	88.20	.94 (.01)	.93 – .96
Test Sample			
Control	93.90	.93 (.01)	.90 – .95
Environmental	76.10	.91 (.01)	.89 – .94
Escape	79.50	.92 (.01)	.90 – .94
Sexual	92.50	.93 (.01)	.91 – .95
Style	82.50	.91 (.01)	.88 – .93
Victim Selection	88.70	.90 (.01)	.88 – .93

Note. All p's <.001

Next, each crime pair was provided a score based on how they were classified in each model using the two-threshold classification approach described previously. Specifically, for each model, crime pairs labelled as linked were assigned a score of 1, crime pairs labelled as unlinked were assigned a score of -1, and crime pairs labelled as unclassified were assigned a score of 0. A composite score was then created for each crime pair based on how they were classified on the models combined (summing across all of their scores).

Prior to creating the multiple model composite score, correlations were examined between the scores provided by each model to determine whether summing across these

scores to create the multiple model composite score was empirically justified. As shown in Table 36, correlations between the scores on each of the different models were moderate-to-high (ranging from r = .40 to .78). Likewise, Cronbach's alpha (α = .86) for these six scores indicated a satisfactory level of internal reliability. As such, the scores were summed to create the composite score.

Table 36. Correlations between scores on each individual CT/ICT model comprising the multiple model approach for the serial sexual assault data.

Forced Variable/Model	1	2	3	4	5
1. Control	-				
2. Environmental	.40	-			
3. Escape	.53	.49	-		
4. Sexual	.57	.40	.58	-	
5. Style	.60	.39	.45	.47	-
6. Victim Selection	.59	.42	.78	.53	.44

Note. All p's <.001.

Composite scores ranged from -6 (indicating that the crime pair was classified as unlinked on all six CT models) to +6 (indicating that the crime pair was classified as linked on all six CT models), with a median score of -6.00 (SD = 2.55). The same range in composite scores was found when looking at linked and unlinked crime pairs separately; however, the median composite score was -6.00 (SD = 2.30) for unlinked cases and +6.00 for linked cases (SD = 3.61). Overall, these scores indicate that, on average, unlinked crime pairs were more often classified as unlinked (versus linked or unclassified) and linked crime pairs were more often classified as linked (versus unlinked or unclassified) across the different ICT/CT models.

The total number of crime pairs possessing each composite score value and the corresponding percentage of those cases that are linked crime pairs is presented in Table 37. As previously mentioned, if more unlinked pairs are found at the bottom end of the distribution (e.g., with a score closer to -6) and more linked pairs are found at the higher end of the distribution (e.g., with a score closer to +6), then a higher level of discrimination accuracy would be expected for the composite score (i.e., multiple CT/ICT model approach). As shown, the largest proportion of linked cases was found in the composite score category of 6 and the smallest proportion of linked cases was found in the composite score category of -6 for both the development and test sample. A total of 64.92% of all linked crime pairs in the development sample and 65.99% of all linked crime pairs in the test sample had a score of 6 (indicating they were classified as linked in all six CT/ICT models). The majority of linked cases in both the development and test samples (82.66% and 79.35%, respectively) had a score of 1 or higher (indicating they were in the linked category more often across the six models than the other categories). Likewise, the overwhelming majority of all unlinked crime pairs in the development and test samples (92.99% and 93.40%, respectively) possessed a score of -1 or lower (indicating they were in the unlinked category more often across the six models than the other categories).

Table 37. Distribution of composite linkage scores and the percentage of linked cases included in each composite score category for the serial sexual assault development and test samples.

	Development Sample		Test Sample	
Score	Total Cases (#)	Linked (%)	Total Cases (#)	Linked (%)
- 6	9,969	0.07	8,855	0.05
- 5	2,200	0.32	2,521	0.24
- 4	1,206	0.75	1,770	0.62
- 3	898	0.22	1,429	0.70
- 2	786	1.15	485	0.62
- 1	404	0.74	478	2.30
0	365	1.64	329	1.82
1	143	3.50	213	0.94
2	156	2.56	145	7.59
3	146	4.11	124	1.61
4	201	5.97	164	3.67
5	120	14.17	98	12.24
6	242	66.53	223	73.09

Constructing the empirically optimal combined CT/ICT model. Next, the empirically optimal multiple CT/ICT model was created. First, a forward stepwise logistic regression analysis was run with the scores for each model entered as the predictors (six predictors) and linkage status as the outcome variable. As shown in Table 38, scores for five of the six models entered into the stepwise model (the scores for the model beginning with the control domain was not selected into the stepwise logistic regression model). Cronbach's alpha ($\alpha = .83$) for these five score variables indicated a

satisfactory level of internal reliability. As such, a modified composite score was calculated for each crime pair using only the scores form these five models.[63]

Table 38. Results of the forward stepwise logistic regression analysis conducted on the scores for each of the six classification tree models comprising the original composite score for the sexual assault data.

Model	B (*SE*)	Wald (*df*)	χ^2 (*df*)	R_N^2	R_L^2
Stepwise			2653.34 (5)	.53	.51
Sexual ICT score	0.99 (0.10)	98.71 (1)			
Environmental ICT score	1.21 (0.08)	206.34 (1)			
Victim Selection ICT score	0.57 (0.10)	33.81 (1)			
Style CT score	0.69 (0.09)	57.08 (1)			
Escape ICT score	0.30 (0.11)	6.94 (1)			
Constant	-3.71 (0.08)	2445.79 (1)			

Note. χ^2 = model chi-square; R_N^2 = Nagelkerke index; R_L^2 = Hosmer and Lemeshow's index; all *p*'s <.001, except the escape ICT score regression coefficient (*p* = .008).

Modified composite scores ranged from -5 (indicating that the crime pair was classified as unlinked on all five CT/ICT models) to +5 (indicating that the crime pair was classified as linked on all five CT/ICT models), with a median score of -5.00 (*SD* = 2.12). The same range in composite scores was found when looking at linked and unlinked crime pairs separately; however, the median composite score was -5.00 (*SD* = 1.92) for unlinked cases and +5.00 for linked cases (*SD* = 2.99). Again, these scores

[63] Given that the control CT score variable was not included in the revised multiple model, multicollinearity indices were examined by running the predictors and outcome through a linear regression. These indices did not point to issues of mulitocollinearity. A partial correlation was then run to examine the relationship between the control CT score and the outcome while controlling for all other score variables. Although the control score was moderately correlated with the outcome (r_{pb} = .31), the relationship reduced substantially while controlling for the other score variables (partial r_{pb} = .03). As such, it seems that all other score variables are able to explain some unique amounts of variability in the outcome, and the addition the control ICT scores did not lead to meaningful improvements in the model when the other scores were already included in the model.

indicate that, on average, unlinked crime pairs were more often classified as unlinked (versus linked or unclassified) and linked crime pairs were more often classified as linked (versus unlinked or unclassified) across the different CT/ICT models.

The total number of crime pairs possessing each modified composite score value, and the corresponding percentage of those cases that are linked crime pairs, is presented in Table 39. A similar distribution was found with the modified composite scores as the original composite scores, with the largest proportion of linked cases possessing a score of 5 for both the development and test sample (65.32% and 65.99%, respectively). Similar to the original composite score, the majority of linked crime pairs in the development and test samples (82.66% and 81.16%, respectively) had a score of 1 or higher (indicating they were in the linked category more often across the four models than the other categories). Likewise, the majority of unlinked crime pairs in the development and test samples (93.51% and 92.73%, respectively) had a score of -1 or lower on the modified composite score variable (indicating they were in the unlinked category more often across the five models than the other categories).

Table 39. Distribution of modified composite linkage scores and the percentage of linked cases included in each modified composite score category for the serial sexual assault development and test samples.

	Development Sample		Test Sample	
Score	Total Cases (#)	Linked (%)	Total Cases (#)	Linked (%)
- 5	10,478	0.08	8,947	0.04
- 4	2,210	0.41	2,694	0.22
- 3	1,266	0.63	1,775	0.60
- 2	825	0.48	1,484	1.01
- 1	771	1.23	525	1.52
0	279	1.43	430	1.16
1	248	3.63	291	3.09
2	166	3.01	169	4.14
3	216	5.09	176	3.98
4	132	13.64	114	10.53
5	245	66.12	229	71.18

Evaluating the multiple CT/ICT models. ROC analysis was then conducted to examine the predictive accuracy of the original and modified composite scores (i.e., the original multiple CT/ICT model and the empirically optimal CT/ICT model). The ROC curves produced from these analyses are presented in Figure 23 for the development and test samples, and the AUCs, standard errors, and 95% confidence intervals are presented in Table 40. Both the original and empirically optimal multiple CT/ICT models resulted in the same level of predictive accuracy for both the development and test samples (all AUCs = .95).

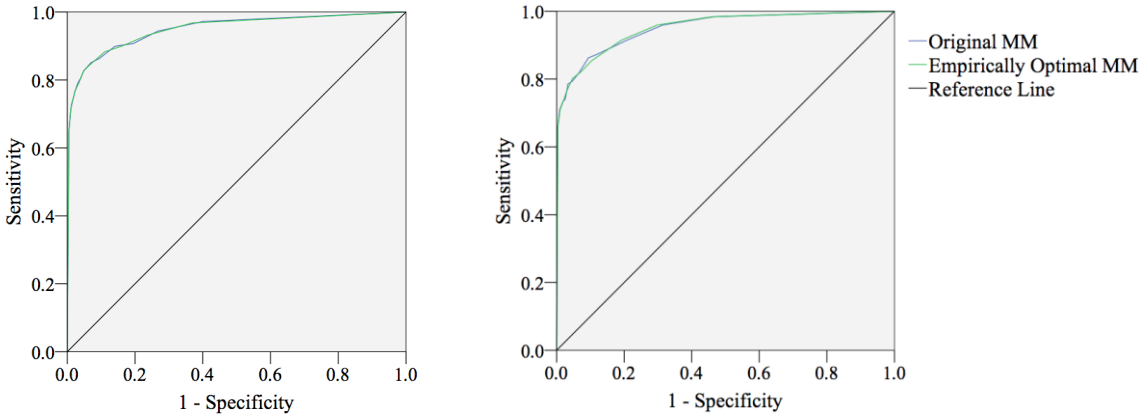


Figure 23. ROC curves for the development (left) and test samples (right) displaying the discrimination accuracy for the original and empirically optimal multiple CT/ICT models for the serial sexual assault data.

Table 40. Development and test sample AUCs and their associated 95% confidence intervals for the original and empirically optimal multiple CT/ICT models for the sexual assault data.

	Development Sample		Test Sample	
Multiple CT/ICT Model	AUC (*SE*)	95% CI	AUC (*SE*)	95% CI
Original MM	.95 (.01)	.93 – .97	.95 (.01)	.94 – .97
Empirically Optimal MM	.95 (.01)	.93 – .97	.95 (.01)	.94 – .97

Comparing the performance of all linking models. The final step was to compare the predictive accuracies of the logistic regression model, standard CT model, ICT model, and the original and empirically optimal multiple CT/ICT models for the sexual assault data (see Figure 24 and Table 41). For the development sample, both

multiple CT/ICT models (original and empirically optimal) resulted in the highest level of predictive accuracy (AUCs = .95), followed by ICT and standard CT (both AUCs = .94), and then the stepwise logistic regression model (AUC = .88). The AUCs of all CT-based models at the development stage were significantly higher than the AUC for the logistic regression model (*z*'s ranging from 3.40 to 4.08, all *p*'s <.001); however, none of the CT-based models performed better than the others (all *p*'s >.05 for CT-based comparisons). This pattern changed slightly when examining the test sample. There were no significant differences between the AUCs achieved by any of the models at test (all *p*'s >.05); however, the AUCs for the multiple models only slightly overlapped with the AUCs of all other models.

Comparisons of the AUCs *across* the development and test samples suggest that both multiple models are relatively stable, achieving the same level of predictive accuracy from development to test. Shrinkage seems to be the largest concern with the standard CT model (shrinkage = .02) and the ICT model (shrinkage = .01), whereas the predictive accuracy of the logistic regression model significantly improved from development to test (increase = .04; $z = 2.16$, $p = .031$). Overall, these results suggest that generalizability is not a concern across any of the models (i.e., all remaining confidence intervals overlapped and all *p*'s >.05).

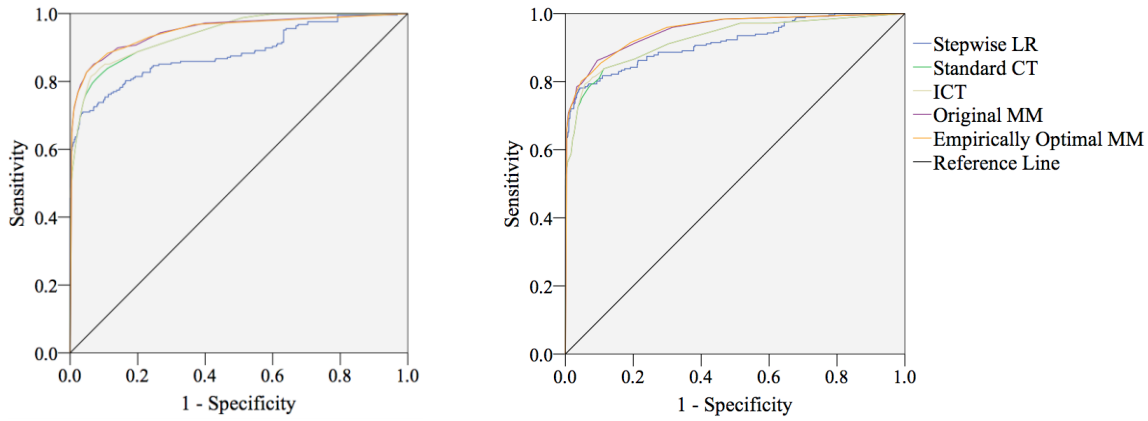


Figure 24. ROC curves for the development (left) and test samples (right) displaying the discrimination accuracy for the logistic regression model, standard CT model, ICT model, and both multiple CT/ICT models constructed using the serial sexual assault data.

Table 41. Development and test sample AUCs and their associated 95% confidence intervals for the logistic regression model, standard CT model, ICT, and the original and empirically optimal multiple models (MM) created with the sexual assault data.

	Development Sample		Test Sample	
Model	AUC (*SE*)	95% CI	AUC (*SE*)	95% CI
Stepwise LR	.88 (.01)	.86 – .91	.92 (.01)	.89 – .94
Standard CT	.94 (.01)	.92 – .95	.92 (.01)	.90 – .95
ICT	.94 (.01)	.93 – .96	.93 (.01)	.90 – .95
Original MM	.95 (.01)	.93 – .97	.95 (.01)	.94 – .97
Empirically Optimal MM	.95 (.01)	.93 – .97	.95 (.01)	.94 – .97

Note. All *p*'s<.001.

Discussion

Generally speaking, the analyses presented above offer a number of insights into our ability to link serial sexual assaults. Moreover, the CT/ICTs produced in the current study explicitly demonstrate the complexities inherent in serial sexual assaulters' behavioural consistency and distinctiveness. Overall, the results indicated that it is possible to link serial sexual assaults in an accurate fashion, with all multivariable models achieving AUCs in the high range when applied to the test sample (i.e., above .90; Swets, 1988).

The first hypothesis was generally supported: the initial descriptive analyses revealed that sexual assaults committed by the same offender had higher levels of behavioural similarity across all domains than sexual assaults committed by different offenders. The magnitude of the effect was large for all domains except escape (medium effect size); however, control and victim selection were associated with the largest effect sizes.

Based on previous research, it was also hypothesized that there would be noticeable differences in the discrimination abilities of the different statistical approaches employed. Specifically, it was expected that the stepwise logistic regression and standard CT approaches would exhibit similar levels of predictive accuracy, the ICT would result in a higher level of predictive accuracy, and the multiple CT/ICT approach would result in the highest level of predictive accuracy. This hypothesis was only partially supported with the serial sexual assault data. As expected, when applied to the test sample, the stepwise logistic regression and standard CT had the same high level of predictive accuracy (AUC = .92). However, while the ICT and multiple CT/ICT models were

associated with slightly higher levels of predictive accuracy (AUCs = .93 and .95, respectively) than the logistic regression and standard CT model, there were no significant differences in the magnitude of these AUCs.

Contrary to what was hypothesized based on Tonkin, Woodhams, et al. (2012), there were no serious issues with shrinkage across any of the statistical approaches adopted. Indeed, the AUCs were relatively stable across the development and test samples (increasing or decreasing slightly). No significant differences were observed in predictive accuracy when any of the models were applied to the test sample.

Finally, as hypothesized, more complex patterns of behavioural consistency and distinctiveness, which were not apparent when using the main effects logistic regression approach, did emerge when using a CT-based approach to BLA with sexual assaults. That is, multiple pathways to making a linked decision were identified, often relying on different subsets of behavioural information.

All of these findings are explored in more detail below.

The Behavioural Consistency and Distinctiveness of Canadian Serial Sexual Offenders

As mentioned, few studies have examined our ability to link serial sexual assaults committed in Canada (for exceptions, see Grubin et al., 2001; Oziel, Goodwill, & Beauregard, 2015). Moreover, BLA studies using sexual assaults that have adopted a similar methodological approach to the current study have not used behavioural domains similar to those used in this study (e.g., control domain, sexual domain, escape domain, etc.). Instead, existing studies have typically examined our ability to link sexual crimes on the basis of similarity across all behavioural items combined into a single domain

(e.g., see Bennell et al., 2014 for a summary of most of these studies and Slater et al., 2015 for the most recently published study employing this approach). Generally speaking, previous research that has examined overall behavioural similarity (i.e., measured as a single domain) has found that it is possible to distinguish linked from unlinked sexual offences to a moderate degree (AUCs ranging from .75 [Bennell et al., 2009] to .89 [Winter et al., 2013]). The results of the current study add further support to this literature; however, a higher degree of predictive accuracy was achieved by all multivariable models developed in this study (i.e., all AUCs were above .90).

The only study that has examined behavioural consistency in sexual offenders using domains that are somewhat comparable to the domains employed in the current study is Grubin et al.'s (2001) study. Measuring consistency across pairs of linked sexual assaults, Grubin et al. found greater behavioural consistency in control and escape than style and sexual domains. Grubin et al. explained that a greater degree of consistency in control and escape behaviours may be observed because offenders conceivably have more control over the behaviours comprising these domains than those comprising the sexual and style domains. Indeed, crime linkage researchers have often argued that behaviours that are more dependent on the offender's interaction with the victim will be less useful for linking purposes than behaviours that are less dependent on how the interaction unfolds (e.g., Woodhams et al., 2007).

Some recent research examining behavioural consistency and distinctiveness across different temporal phases in a serial sexual assault generally confirms that behaviours that rely less on the victim are more valuable for linking purposes than behaviours where the victim is involved (Ozeil et al., 2014). For example, Ozeil and

colleagues found higher levels of linking accuracy using domains comprised of pre-offence behaviours (i.e., behaviours engaged in by the offender before the offence, such as bringing a rape kit and/or weapon to the crime) compared to domains comprised of behaviours that were further along, temporally speaking, in the crime sequence (e.g., the strategies used to approach and attack the victim). They indicated this was likely because behaviours that are found further along the assault chain of events are more susceptible to situational influences. The univariate analyses presented in this chapter partially support this assertion. For instance, the domains with the highest AUCs in the test sample were victim selection (i.e., conceivably "pre-offence"), control, style, and environment (test AUCs > .80), whereas the sexual (test AUC = .79) and escape domains (test AUC = .58) were associated with the lowest AUCs.

It is interesting to note the poor performance of the escape domain given the findings of Grubin et al. (2001). It is unclear why differences emerged between the current study and Grubin et al.'s study, but it may be due to the fact that the behaviours comprising the current escape domain were operationalized differently from those used in Grubin et al.'s escape domain. For instance, Grubin et al.'s domain contained the following six variables: concerns about escape, gloves, mask, obvious precautions, tells victim not to look at him, and destroys semen. In contrast, the variables used in the current study included: used disguise during crime, wore gloves, tried not to leave semen, prevented his face from being seen, and used condom. Arguably, the escape domain in the current study contains more behaviours that would depend on how the interaction with the victim unfolds (e.g., tried not to leave semen, used condom, etc.). In fact, all domains used in the current study contained a mixture of behaviours, some likely being

more offender-driven, or pre-offence, and others likely being more situation-dependent (e.g., the control domain includes a variable, "bindings were brought with the offender," and a variable, "used restraints"). Nonetheless, with the exception of the escape domain, all individual domains achieved moderate levels of linking accuracy, suggesting that linking is possible even when the situation may not be entirely similar from crime-to-crime.

There are at least two possible explanations for the increased predictive accuracy observed in the current study compared to previous studies. First, at least some of the offenders in the sexual assault dataset had a history of psychiatric problems (n = 16; 23%). Recent research conducted by Woodhams and Komarzynska (2014), which examined behavioural consistency among mentally ill sexual offenders, suggests that these individuals exhibit highly consistent and distinctive behaviours across their offences, which can result in extremely high levels of linking accuracy (AUC = .996). To the degree that these types of individuals exist within the current dataset of serial sex offenders, this might explain the high AUCs found in this study. Future research can test this possibility if research collects more detailed information about the mental health status of offenders.

Second, unlike previous studies that relied on police data, the serial sexual assault data in the current study were collected through offender interviews. Given that police data is not collected for research purposes, the quality of data analyzed in previous studies may be poor (Alison, Snook, & Stein, 2001). For instance, because there is often no standardized protocol followed by investigators when collecting information, the data collected and recorded in police databases often varies from crime-to-crime, decreasing

both the accuracy of the account and the richness of the behavioural information available (even when standardized police protocols are used, research has found that data reliability can still be an issue; e.g., Snook, Luther, House, & Bennell, 2012). Although the data collection protocol used in the current study (i.e., semi-structured offender interviews) can also be potentially problematic (e.g., offenders may distort their accounts, fail to have insight into their own behaviours, or fail to remember certain things), offender interviews have the potential to capture a great deal of rich and reliable information (Brookman, 2010; Copes & Hochstetler, 2010; Elffers, 2010). Likewise, collecting data via an interview protocol *for research purposes* allows the researcher to standardize the process. Thus, the improved predictive accuracy in the current sample may be partially accounted for by differences in data collection protocols, an issue that is worthy of future research.

The Proposed Advantages of a CT-based Decision Support Tool for Linking Sexual Assaults

As discussed earlier, there are a number of potential advantages of a CT-based approach to linking versus an approach based on main effects logistic regression (Monahan et al., 2001; Tonkin, Woodhams, et al., 2012). The extent to which the proposed advantages are supported by the serial sexual assault analyses is discussed next.[64]

Predictive accuracy of CT-based versus LR-based models. The first argument in support of a CT-based approach is that it will result in a level of predictive accuracy that is similar to the level achieved by a stepwise logistic regression approach, and that as

[64] Given that the discussion in Chapter 6 surrounding some of the proposed advantages of CT-based models also applies here (e.g., the discussion surrounding whether or not it is true that the statistical processes underlying CT-based models are more easily understood than those underlying the stepwise logistic regression process), these issues will not be discussed in this chapter to avoid redundancy.

the complexity of the CT model increases (e.g., ICT, multiple CT/ICT model), the CT approach will progressively outperform the main effects regression approach (e.g., see Monahan et al., 2001). As hypothesized, compared to the break and enter models developed in Chapter 6, the serial sexual assault models that were developed in this chapter were more reflective of this hypothesized pattern. That being said, the improvements in predictive accuracy that were observed as the complexity in the CT-based model increased, from standard CT to ICT to multiple CT/ICT, were minimal and non-significant.

The classification results using the two-threshold approach adopted by Monahan et al. (2001) also did not completely parallel the findings from that study. As previously mentioned, Monahan et al. found that their ICT was able to classify, and *correctly* classify, more crime pairs overall than the logistic regression and standard CT models. They also found that the standard CT approach classified, and *correctly* classified, a *lower* number of cases than the logistic regression model. In the current study, although a greater number of crime pairs could be classified, and *correctly* classified, using the ICT compared to the logistic regression and standard CT models, the standard CT had better classification rates than the logistic regression model. Moreover, the standard CT and ICT correctly classified the exact same number of unlinked crime pairs, although the ICT correctly classified a slightly higher proportion of linked crime pairs (5.82% more than the standard CT). As such, it seems that both CT-based approaches can offer improved classification accuracies relative to a logistic regression approach, and that an ICT approach can offer a slight improvement for the classification of linked pairs than the

standard CT approach (although the overall levels of predictive accuracy achieved by the models are statistically comparable).

Another point that is important to highlight is that, unlike Tonkin, Woodhams, et al. (2012) and the break and enter analyses presented earlier, an ICT was developed using the serial sexual assault data. This arguably supports the idea that more complex patterns of behavioural consistency and distinctiveness are present in interpersonal crimes, possibly warranting an ICT approach. However, the ICT developed in the current study was still far less complex than that developed by Monahan et al. (2001) and it did not lead to the increase in predictive accuracy that was hypothesized based on Monahan et al.'s study. For instance, only two iterations were produced using the unclassified cases in the current study, whereas Monahan et al. produced an ICT with four iterations.

Again, these differences may be the result of some of the methodological issues discussed previously. For instance, although it is possible that the high level of predictive accuracy observed in the current study (i.e., AUCs >.90) is due to the fact that rich behavioural information was obtained from the offender interviewers, the number of predictors ultimately used in the current research undoubtedly limited the variation that could be observed in the CTs that were produced, and consequently, the need for complex CT-based approaches (e.g., ICTs or multiple CT/ICTs). Indeed, the additional iterations in Monahan et al. (2001) included nine additional variables that were not included in their original standard CT.

That being said, it is also important to consider the differences in magnitude of the AUCs found in the current research compared to those reported in the MacArthur Study. Relative to the AUCs achieved by the standard CT and ICT in the McArthur Study

(AUCs = .79 and .82, respectively), the AUCs observed with the sexual assault data in the current study were already extremely high for the standard CT and ICT (test AUCs = .92 and .93 for the CT and ICT, respectively). Consequently, although adopting a multiple model approach in the MacArthur Study resulted in a .06 increase in the AUC (i.e., ICT AUC = .82 vs. multiple model AUC = .88), the level of predictive accuracy achieved by the MacArthur multiple ICT model was still substantially lower than that achieved by the multiple CT/ICT model using the current sexual assault data (test AUC = .95). Indeed, given that the AUCs for the standard CT and ICT were already quite high in the current study, this makes it more difficult to observe an increase in predictive accuracy that is comparable to that achieved in the MacArthur study.

Finally, it is also important to reiterate that, despite the fact that the CTs produced from the sexual assault data were highly complex, there were no issues in terms of generalizability of the CTs to the test samples. It is unclear why this is the case and why these finding contradict those of Tonkin, Woodhams, et al. (2012). As discussed previously, it may be because of the fact that different CHAID parameters were selected in the current study.

Overall, although slight increases in predictive accuracy were observed across the different CT approaches, it still appears as though a standard CT approach might be the best CT-based model using the current sample of serial sexual assaults. The results from this study suggest that, despite the fact that there are complex behavioural patterns to capture in serial sexual assaults, attempting to capture this complexity using the ICT and multiple CT/ICT method may not add much value in terms of our ability to link these crimes. Moreover, although it seems that the multiple CT/ICT model approach may be

slightly more robust when applied to a new sample of crimes (i.e., no shrinkage was evident), the slight increase in accuracy and decrease in shrinkage from the standard CT to the multiple CT/ICT model likely does not outweigh the complexities involved in developing (and using) a multiple CT/ICT model approach to linking sexual assaults. In fact, if a multiple model approach was adopted in practice, it is likely that the other advantages of a CT approach would be lost (e.g., ease of use, understanding of the process involved, etc.), for very little added predictive power.

Transparency of the decision processes and ease of use. Another argument supporting a CT-based decision support tool is that it provides the user with a more transparent representation of the decision-making process that is also easier to explain to others (in comparison to a logistic regression equation). Even though the CT and ICT produced using the serial sexual assault data are considerably more complex than the CTs produced using the break and enter data, the same conclusions can be made here as in Chapter 6. Although it may be more difficult to use the CT or ICT models developed in the current study in raw CT form (as compared to the CT developed using the break and enter data), a fairly basic computer program could be developed that leads the user through one of the pathways to an ultimate decision. The pathway followed could then be printed so the user can explicitly explain to others when they relied on a certain behavioural feature (e.g., "the first thing I looked at was how similar the crimes were in their control behaviours…") and how they proceed to the next question (e.g., since the similarity in control behaviours was greater than .46, I then proceeded to look at how similar the crimes were in their style behaviours…") until they arrived at an ultimate linkage decision (linked, unlinked, or unclassified). Once again, however, whether it is in

fact easier to use and articulate the decision-making process underlying a CT-based approach to linking is an empirical question that needs to be explored in future research that tests the usability of CT-based models with police practitioners.

The Proposed Advantages of CTs for Capturing the Complexities in Sexual Offending Behaviour

Finally, another argument presented in favour of a CT-based approach is that it better reflects the complexity in offending behaviour than the traditional main effects logistic regression approach. That is, assuming that subsets of offenders do in fact differ in the extent to which they are behaviourally consistent and distinctive, the interactive nature of a CT-based approach can capture these differences, leading to multiple, tailored pathways for making a linkage decision.

This argument was supported using the serial sexual assault data. Although both the logistic regression and CT-based approaches included most behavioural domains in the model (the standard CT and ICT model did not include the escape domain, although the logistic regression model did), it does appear that the CT-based models were better able to capture the complexity inherent in sexual offending. Moreover, this complexity was apparent even when we used the same CT-building parameters that were used for the break and enter data (where lower levels of complexity were found).

Indeed, as shown in Table 42, a total of seven pathways to identifying a crime pair as linked were identified when applying the standard CT to the test sample of serial sexual assaults.[65] The pathways in Table 42 are rank-ordered by the percentage of linked pairs found within each pathway. As shown, if one were to rely on high similarity in

[65] There were an additional 6 pathways to making a linked decision using the ICT. Given that the CT and ICT resulted in similar levels of predictive accuracy, only the standard CT pathways will be discussed here for the sake of brevity.

control behaviours, high similarity in style behaviours, and high similarity in environmental behaviours (pathway 1), they could be relatively confident that they were dealing with crimes committed by the same offender (84.9% chance). In contrast, if one were to rely on moderate similarity in control behaviours, moderate similarity in victim selection behaviours, and *lower* similarity in sexual behaviours (pathway 7), they should be much less confident that they were dealing with crimes committed by the same offender (although they would still make that decision if using Monahan et al.'s (2001) thresholds).

Table 42. The seven different pathways in the CT leading to a "linked" decision for pairs of serial sexual assaults when using the two-threshold approach proposed by Monahan et al. (2001) on the test sample.

Pathway	Percent Linked
1 control behaviours (high similarity: $J > .429$) → style behaviours (high similarity: $J > .400$) → environmental behaviours (high similarity): $J > .571$)	84.9
2 control behaviours (moderate-to-high similarity: $J = .308$-$.429$) → style behaviours (high similarity: $J > .400$) → victim selection (moderate-to-high similarity: $J > .333$)	12.0
3 control behaviours (high similarity: $J > .429$) → style behaviours (low-to-moderate similarity: $J = .167$-$.200$) → victim selection (moderate-to-high similarity: $J > .333$) → environmental (moderate-to-high similarity: $J > .211$)	11.2
4 control behaviours (high similarity: $J > .429$) → style behaviours (high similarity: $J > .400$) → environmental behaviours (low-to-moderate similarity): $J \leq .571$)	9.5
5 control behaviours (moderate similarity: $J = .125$-$.143$) → victim selection (high similarity: $J > .571$)	5.3
6 control behaviours (lower similarity: $J = .100$-$.125$) → victim selection (high similarity: $J > .571$)	3.7
7 control behaviours (moderate similarity: $J = .125$-$.143$) → victim selection (moderate similarity: $J = .222$- $.333$) → sexual behaviours (lower similarity: $J \leq .000$)	3.4

It was also argued earlier in this dissertation that identifying these pathways might lead to a better understanding of serial sexual offending behaviour. Although it is not clear what these pathways mean at this time, what can generally be concluded is that serial sexual offenders do seem to differ from one another in terms of the types of

behaviours for which they are consistent and distinctive. For example, offenders falling along pathway 1 are highly similar in their control, style, and environmental behaviours. This may arguably reflect the fact that these offenders engage in high levels of pre-offence planning and fantasy.[66] However, other sexual offenders are much less consistent in their control behaviours, yet they seem to be consistent with respect to the types of victims they select (e.g., pathways 5 and 6). It is possible that these offenders are not predisposed to engage in specific control behaviours across their crimes (e.g., they are opportunistic rather than planners), despite the fact that they have a highly specific preference for a certain type of victim (consequently allowing us to link their crimes on the basis of victim selection similarity).

Not only can different offenders be consistent and distinctive in different ways (i.e., due to predispositions to act a certain way), certain offenders may display low levels of similarity in certain offence behaviours due to situational constraints. For instance, it may be that a subset of offenders who fall along pathways 5 and 6 display low levels of control similarity because they encounter a rather resistant victim during one of their crimes, demanding a different behavioural response (e.g., restraining the victim with something from the crime scene); however, we may still be able to link their two crimes based on their high levels of victim selection similarity.

Indeed, researchers often speculate that domains achieve high levels of linking accuracy because, *overall,* the domain may be more offender-driven than situation-driven (e.g., the control domain; Grubin et al., 2001; Woodham & Toye, 2007). However, what these tentative descriptions demonstrate is that, depending on the offender, *the same*

[66] Indeed, it has been proposed that sexual fantasy is often fairly stable and may therefore partially explain why high levels of consistency and distinctiveness in sexual offending behaviour might exist for some offenders (Gee & Belofastov, 2014).

behavioural domain can arguably be either offender-driven or situation-driven. That is, the same variable can mean different things for different offenders (i.e., those falling along different pathways).

It is important to stress that the explanations attached to these pathways are only speculative at this time. Indeed, future research is needed before the true meaning of these pathways is known. For instance, it may be useful to interview offenders in order to determine which pathways they fall along and what the patterns in consistency and distinctiveness represented by those pathways actually mean to those offenders. In this way, it might be possible to identify subtypes of offenders that can be distinguished from one another on the basis of these pathways. Alternatively, it is possible that distinct subtypes of offenders cannot be identified, and that the pathway an individual offender follows varies across their crimes as a function of some identifiable factor (or factors). Indeed, one offender's crimes may fall along different pathways across the offender's crime series. Situational factors may cause this to happen, or an offender's behaviour may evolve over time due to learning and/or evolving sexual fantasies (Gee & Belofastov, 2014).

While it is not possible to understand the psychological meaning of the identified pathways without additional research, at the very least, identifying these pathways provides researchers with heuristics that can guide future research (e.g., interviews with offenders) so that we can better understand the complexities that exist in offending behaviour over time. It is also important to stress that it was only possible to identify these pathways because of the CT-based approach that was adopted in this study (i.e., the pathways were not evident using the main effects logistic regression approach).

On a final note, recall that the domains used in the current study were defined in a relatively atheoretical manner. That is, behaviours that police could identify at the time of the crime were split into domains based on their presumed function (as mentioned, this is a standard approach to forming domains in many linking studies; e.g., Bennell & Canter, 2002; Burrell et al., 2012, 2015; Davies et al., 2012; Woodham & Toye, 2007; Tonkin, Santtila et al., 2012; Tonkin, Woodhams, et al., 2012). This may make it difficult to infer any sort of underlying psychological meaning from the pathways. Future research may explore the benefits of defining behavioural domains in different ways. For instance, as briefly mentioned, Oziel and colleagues (2015) divided sexual offence behaviours into four domains based on when the behaviours occurred in the sexual assault sequence (i.e., "temporal" domains): pre-crime, victim selection, approach, and assault. It is possible that defining the domains in this sequential manner, and subjecting them to a CT analysis, may provide greater insight into sexual offending behaviour (and may, ultimately, improve our ability to link crimes). Regardless of how domains are created, providing practitioners with credible explanations for why these pathways exist could increase the extent to which they are willing to incorporate statistically-based linking approaches into their BLA practices.

Summary

Taken together, the results of this study suggest that, relative to a main effects logistic regression approach, a simple standard CT-based approach can: (1) lead to linking decisions that are similar in predictive accuracy, (2) result in superior level of correct classification accuracy when using a two-threshold approach, and (3) capture more complex patterns in behavioural similarity and distinctiveness (leading to more

idiographic linking decisions). Moreover, it does not appear that there is any added predictive value associated with more complex (e.g., ICT or multiple CT/ICT) CT-based models when linking the current sample of serial sexual assaults. There also does not seem to be any issues with the generalizability of the models developed here.

Generally speaking, most of the value of a CT-based approach to linking serial sexual assaults seems to lie in: (1) the ease with which one can observe and possibly articulate the linking process to another individual and (2) the fact that the CT approach provides a more idiographic alternative to making linkage decisions in comparison to the main effects logistic regression approach. However, as previously stressed, research is needed to determine whether practitioners attempting to use a CT-based model can easily articulate their decision-making process to others who may not be as familiar with the procedure. Likewise, conducting more research with offenders to understand *why* the various CT pathways exist may facilitate the acceptance of a CT-based approach into practice. Finally, whether these potential advantages of the CT-based approach do, in fact, result in an increased acceptance of this tool is an empirical question that ultimately needs to be addressed in future research.

CHAPTER 8

General Discussion

This chapter begins with a summary of the results in relation to the various goals of this dissertation. Potential reasons for the divergent findings across the two datasets are also discussed. This is then followed by an overview of the implications of the current research in its entirety for crime linkage researchers and police practitioners. The chapter ends with a discussion of the limitations of the current research. Future research directions are proposed throughout the chapter.

Is BLA Possible in the Canadian Context?

One of the central goals of this dissertation was to determine the extent to which BLA is possible using Canadian serial crime. Results revealed that it is possible to link serial break and enters and serial sexual assaults from Canadian jurisdictions using certain behavioural domains. However, there were also some noticeable differences between the current research and past research. For instance, the predictive accuracy of ICD was considerably lower in the current sample of break and enters than in previous research. Some potential reasons for this discrepancy were explored in Chapter 6, such as possible differences between the jurisdictions examined in the current study and previous research (e.g., geographical landscape) or how the various behavioural domains were operationalized. Likewise, the level of predictive accuracy achieved by multivariable models using the sexual assault data was higher than that found in previous studies. Some potential reasons for this were also explored in Chapter 7, such as the fact that the data collection methodology employed in the current research (i.e., offender interviews) differed from that typically employed in past BLA research (i.e., police databases).

Linking property versus interpersonal crimes. Generally speaking, compared to the levels of accuracy achieved by the multivariable models produced from the break and enter data, a higher level of discrimination accuracy was achieved by the multivariable models developed using the serial sexual assault data. This is not particularly surprising and is consistent with past research, which has suggested that we may be better able to link interpersonal versus property crimes using MO behaviours (e.g., see Bennell et al., 2014 for a summary of this research that reaches this general conclusion).

Higher levels of linking accuracy may be found with sexual assaults for a variety of reasons. The degree to which sexual offending behaviour is driven by deeply engrained fantasies (Gee & Belofastov, 2014) may be one reason why sex offenders are more consistent/distinct that property offenders. Another potential explanation that was discussed relates to research conducted by Woodhams and Komarzynska (2014), which found that mentally ill offenders demonstrate particularly high levels of consistency and distinctiveness. Given that a notable proportion of the current sample of sex offenders had a history of psychiatric problems, it may be this partially explains the higher levels of linking accuracy that we observed for this sample. At this point, both of these potential explanations are speculative and future research is needed to compare the differences in psychopathology between these two types of offenders.

As mentioned earlier, there is one additional difference between the sexual assault and break and enter data that might account for the discordance in findings. Differences in the data collection methodologies may account for variations in the quality of the information collected, which could have resulted in higher levels of linking accuracy

being found for the sexual assault data. Like most previous BLA research, the break and enter data were collected using information contained in police databases and no strict protocol or investigative checklist was used to guide the original data collection process in Saint John (by the investigating officers). As a result, the detail of data (and potentially its quality) varied from crime to crime. As discussed, a different approach was used to collect the sexual assault data (i.e., standardized interviews with offenders, plus additional cross-referencing with police reports). The use of a standardized data collection protocol to capture break and enter behaviour would likely result in more reliable data, which could potentially enhance our ability to link serial break and enters. Future research should explore this possibility, in addition to exploring the value of drastically different data collection procedures, such as conducting interviews with burglars.

Finally, there are also some potential theoretical reasons for the greater consistency and distinctiveness in the behaviour displayed by serial sexual assaulters in comparison to serial burglars. For instance, the CAPS model that was briefly discussed in the literature review (Mischel & Shoda, 1995) has been used several times in the crime linkage literature to explain the conditions under which offender consistency and distinctiveness might be expected (e.g., Woodhams et al., 2007). One of the main tenets of the CAPS model is that individuals' behaviour operates under two systems: the hot and cool system (Metcalfe & Mischel, 1999; Mischel, 2009). Logically, the hot system is *emotional*: characterized by faster, reflexive behavioural responses. In contrast, the cool system is *cognitive*: characterized by slower, more reflective, self-controlled, or calculated behavioural responses. The impulsive, reflexive behaviours that characterize

the "hot" system are more likely to be consistent across situations (Furr & Funder, 2004). Likewise, under conditions of stress, the hot emotional system is predicted to override the cool system (Metcalfe & Mischel, 1999). As explained by Bennell and Woodhams (2014), it is possible that a higher level of predictive accuracy can be achieved with the sexual assaults versus the break and enters because the "hot" emotional system is more likely to be activated for offenders committing interpersonal crimes. This may due to the stress that may be caused in interpersonal crimes by the presence of an unwilling victim.

Does a CT-based Approach Lead to Improvements in Predictive Accuracy?

A second goal of this dissertation was to determine whether the results of past research exploring variations of a CT-based approach to violence risk assessment (e.g., Monahan et al., 2001) could be extended to the BLA field. More specifically, the current research sought to determine if a CT-based approach could produce levels of predictive accuracy that were similar to that of the traditional logistic regression approach (e.g., standard CT and ICT) and whether adopting a multiple CT/ICT model approach could further enhance the predictive accuracy of a CT-based approach beyond that achieved by a single CT or ICT.

The results across both datasets revealed that a CT-based approach (single standard CT or ICT) could produce levels of predictive accuracy that were comparable to the logistic regression approach; however, the current research failed to find a meaningful increase in predictive accuracy when the multiple model approach was adopted. Some possible reasons for these findings were explored in Chapters 6 and 7, such as the fact that, in comparison to Monahan et al. (2001), substantially fewer predictors were available in the current dissertation to construct the CTs.

A third goal was to determine if the issues observed in Tonkin, Woodhams, et al. (2012) extended to the current samples of crime. Specifically, Tonkin, Woodhams, et al. found that CT-based linking methods suffered from issues of generalizability (i.e., shrinkage in the AUC when cross-validating the mode on test sample data). The current research did not support these findings: all CT-based models were generalizable to the test samples for both the serial break and enter and serial sexual assault datasets. Some possible reasons for differences between the current findings and those of Tonkin, Woodhams, et al. were explored in Chapter 6, including the fact that different (and potentially less idiosyncratic) CHAID parameters were selected in the current study when constructing the CT-based models.[67]

Does a CT-based Approach Reveal Previously Hidden Complexities in Behavioural Consistency and Distinctiveness?

The final goal of this study was to compare the behavioural patterns that emerged under a main effects logistic regression versus a CT-based approach to linking in order to determine the extent to which a CT-based approach could enhance our understanding of serial offender behaviour. Overall, results indicated that a CT-based approach to linking serial break and enters did not reveal any strikingly different behavioural patterns relative to the traditional main effects approach (only two pathways to making a linked decision, each using the same predictors, were produced using the standard CT). In contrast, the standard CT produced using the serial sexual assault data was considerably more complex, revealing 7 pathways to making a linked decision, many of which relied on

[67] Again, when the parameters used by Tonkin, Woodhams, et al. (2012) were applied to the break and enter data used in this study, significant shrinkage was observed. It is, however, important to note that when Tonkin, Woodhams, and colleagues used less stringent parameters that were more in line with the parameters used in the current research, a significant amount of shrinkage was still found.

different combinations of predictors. With that said, it was difficult to discern any meaning from these pathways. Some future research endeavours to remedy this were proposed in Chapter 7, including using the identified pathways as a heuristic to interview offenders in the future in an attempt to determine why these differences in offender consistency and distinctiveness emerged.

The possibility that behavioural patterns observed during sexual assaults are more complex than those observed during break and enters may explain why the complexity of the CTs differed across the two datasets. Indeed, as explained in Chapter 6, it is possible that only simple CTs were produced using the serial break and enter data because there is not enough inter-individual variation in behavioural consistency/distinctiveness patterns to warrant an analytical approach that attempts to capture such complexity. In contrast, research using sequential analysis to examine the nuances in the behaviours exhibited over the course of a sexual assault has generally demonstrated that sexual assaults involve a highly complex event where an offender's behaviour can change quite rapidly depending on how the interaction unfolds (e.g., Fossi, Clarke, & Lawrence, 2005; Lawrence, Fossi, & Clarke, 2010; Winter, 2014).

For instance, Fossi et al. (2005) examined the complexities in behaviour for a very specific type of sexual assault: stranger rapes occurring in the victim's bedroom. They found that the amount of victim resistance was dependent on the underlying style of rapist (i.e., whether the rapist committed multiple sexual acts vs. single sexual acts) and that the offender's subsequent behavioural reactions to this resistance were, in turn, dependent on when in the assault sequence the resistance occurred (in addition to the offender's own underlying motivations). At the very least, this research supports the idea

that behaviour in sexual crimes is highly complex and interactive. Given that victims are typically not present within the context of break and enters (or if they are, they interact with the offender minimally), the level of behavioural complexity observed in serial sexual assaults is likely to be absent within these sorts of criminal events.

Indeed, supporting the argument that the behavioural patterns observed across sexual assaults are more complex than those observed across break and enters, many of the behaviours in the break and enter data occurred at relatively high frequencies (e.g., above 50%). In contrast, with the exception of variables in the environmental domain (reflecting where the various phases of the offence occurred; e.g., weekday, daytime, inside, outside, residential area), very few behaviours included in the remaining sexual assault domains occurred at frequencies higher than 50%. This may have impacted the differences in behavioural consistency and distinctiveness overall between the two samples. Taken together, these differences in the frequencies of behaviours observed across the two datasets may partially account for why a CT-based approach is more suited to sexual assaults than break and enters.

It is important to reiterate, however, that even though there may in fact be more complexity inherent in the behaviours exhibited by sex offenders versus property offenders, the differences in the data collection methodologies across the two sets of analyses complicates any attempt to determine why these differences exist. No linking research has been published using a sample of property crimes where the information was gathered directly from interviews with offenders. Such a study would allow for more direct comparisons with the serial sexual assault findings reported in Chapter 7. The

potential implications of conducting this sort of research are explored in more detail below.

Implications of this Research

The research presented in this dissertation has a number of related implications for both linking researchers and the police. These are reviewed next.

Implications for BLA researchers. The results of the current research have identified a number of issues that are relevant considerations for researchers examining statistical approaches to BLA in the future. These include issues surrounding: (1) the data collection methods employed (i.e., police databases versus offender interviews), (2) the sampling approach taken (i.e., constant subset of crimes per offender, all crimes per offender, or the new outlier approach), and (3) the frequency of occurrence for individual behaviours.

Data collection methods. The research presented in this dissertation suggests that a higher level of linking accuracy might be achieved when using data that are obtained from offender interviewers as compared to police files. Although there are other confounding factors between the two studies presented here that could account for these results (e.g., crime type), it may be the case that higher quality information (in terms of the amount, accuracy, and/or reliability of the information) can be obtained from offender interviewers than police databases, for the reasons discussed earlier. As suggested by Woodhams and Komarzynska (2014), a single study on the same sample of crimes would be needed to determine the extent to which different data collection protocols do, in fact, account for these findings.

To date, no linking studies have been published on residential burglary using offender interview data, despite the fact that research has generally shown that a great deal of behavioural information can be obtained from conducting interviews with residential burglars (e.g., Nee, 2010). Given the issues identified with the police files used in this dissertation, it may be a worthwhile endeavour to conduct interviews with offenders in the future for linking research purposes. For instance, collecting data on the same crimes from offenders and from police files would allow us to systematically determine how the information obtained from offenders versus police files differ, providing more concrete explanations for *why* improvements are found when using information from offender interviews. For example, it may be that the police are not systematically collecting certain types of information even though it may be crucial for linking purposes.

Identifying what differences exist between offender interview data and police records data may inform strategies that can be undertaken to improve information available in police databases, bringing it up to par with the information that can be gleaned from offender interviews. For instance, police could potentially be informed of important behavioural features that they are not currently capturing (or are at least not doing so in a systematic manner). Likewise, this research could be used to inform better victim interviewing procedures to ensure that the information required for linking purposes is ultimately incorporated into police databases.

Indeed, in line with the arguments presented by others (e.g., Grann & Långström, 2007; Liu, Yang, Ramsay, Li, & Coid, 2011), it may be wise to place more emphasis in future research on identifying the most appropriate predictors for a given prediction task

rather than comparing different statistical approaches using the same data. Although the latter task is valuable in its own right, it is hard to deny that the linking models we produce will only ever be as good as the data that are available to us. As such, future research efforts that strive to identify how improvements to police data can be made on the basis of the information gleaned from offender interviews should be prioritized. Although this line of research will undoubtedly be difficult, time consuming, and costly, it would move this field of research forward considerably.

Sampling approach. Recall that two sampling approaches have been used in previous linking research to date, either: (1) a subset of all crimes are retained so that the number of crimes contributed by each offender is constant (e.g., 2 crimes per offender), or (2) all crimes are retained so that the number of crimes per offender varies (e.g., from 2 to 20 per offender). The current research proposed a new method for sampling crimes for the purpose of constructing linking models. This method was labelled the outlier approach because it identified the most prolific (outlying) offenders in each dataset and retained a (Winsorized) subset of their crimes for the final sample. This was done in an attempt to provide a compromise between the two arguments that typically accompany traditional sampling approaches: the need to control the impact of prolific offenders on the linking models produced and the need to ensure that the linking models reflect the serial offender population for generalizability purposes.

As briefly mentioned in the methodology section (Chapter 5), analyses were conducted using all of the sampling approaches outlined above. When the analyses were conducted using two crimes per offender and all crimes per offender, the linking models that emerged were quite different from the models presented in Chapters 6 and 7, both in

terms of the sequence of variables used to construct the models and in terms of the levels of predictive accuracy achieved by the models. More specifically, in comparison to results using the outlier approach, higher levels of predictive accuracy were observed in both the break and enter and serial sexual assault data when all crimes were used, whereas lower levels of predictive accuracy were observed when only two crimes per offender were used.[68] As such, it may be the case that the outlier approach to controlling prolific offenders provides the most realistic levels of predictive accuracy, or possibly the least biased levels of predictive accuracy in comparison to the other two approaches.

Regardless, the issues raised here warrant future research comparing the use of different sampling procedures. Likewise, future research should more carefully examine the differences in behavioural consistency and distinctiveness patterns observed between more versus less prolific offenders to determine why the differences between sampling approaches emerged in this study. For instance, although it may be partially the result of a sample size issue (e.g., more linked crime pairs may lead to increased reliability of parameter estimates), it may also be the case that prolific offenders are more behaviourally consistent and/or distinctive than less prolific offenders (at least for a subset of their crimes). Indeed, more prolific offenders may develop some form of expertise in offending overtime, eventually leading to more consistent behaviours in their later crimes. Only two studies have examined this proposed expertise effect on behavioural consistency, each providing conflicting results (Tonkin et al., 2008; Woodhams & Labuschage, 2012). In fact, it may be that prolific offenders are less consistent because, overtime, they change their offending behaviour to avoid detection.

[68] As mentioned earlier, the only exception being that the predictive accuracy of the ICD alone model decreased when all break and enters from Saint John were included in the dataset.

As such, future research is needed to further explore this issue and how (and why) it has an impact on the linking models that are ultimately produced.

Regardless, the sampling approach that is taken (e.g., all crimes per offender versus a constant subset of crimes) is not a trivial matter. It can fundamentally change the linking models that are produced, the levels of linking accuracy that can be achieved, and, ultimately, the conclusions that we reach on the basis of this research (e.g., how useful certain behaviours are for linking purposes, how accurate certain linking models may be, etc.). As such, these issues need to be more carefully considered and justified in future research endeavours. It may, in fact, be that a different sampling approach is most optimal depending on the goal of the research (e.g., whether the goal is to develop linking models that are accurate for practice or whether the goal is to examine and improve our understanding of serial offender behavioural consistency and distinctiveness).

Frequency of behaviours. Recall that a number of previous linking studies have removed behaviours with low frequencies from the analyses (e.g., ranging from 1% to 10% frequencies; Santtila et al., 2005; Tonkin & Woodhams, 2015; Winter et al., 2013). The current research, however, demonstrated that, as behaviours with lower frequencies are progressively removed from the calculation of Jaccard's coefficient, the level of predictive accuracy achieved using these domains decreases. Although the decreases in predictive accuracy were never significant, this does suggest that researchers should reconsider omitting low frequency variables unless there is a strong rationale for doing so.

The current research also found that that there was a wide range of frequencies across the behaviours included in each of the domains, with some of the behaviours

occurring quite commonly and others occurring infrequently. Although researchers do not always publish the list of variables that they use in their research, and their respective frequencies, having access to this information would make it easier to compare how similar the datasets are across these studies. Moreover, examining differences in behavioural frequencies across behaviours that are included in these studies may provide further insights into why cross-jurisdictional differences may be emerging.

Likewise, having access to a list of the variables included in these studies could also provide insights into how similar the different domains are in terms of the behaviours that define them. As mentioned, Davies et al. (2012) found that changing the way in which the target selection domain was operationalized in a sample of car thefts led to differences in linking accuracy. It is certainly plausible that the domains used in serial residential burglary studies to date operationalize many of the same domains (e.g., entry, internal, target selection, etc.) differently. Future research should attempt to examine the consequences of doing this, as it may explain divergent results between studies. For instance, are results truly the result of cross-jurisdictional differences in offenders or are the results a function of the information included in the behavioural domains (or both)? As mentioned by Tonkin (2014), unless this issue is more systematically examined within the same study, it is difficult to conclude that certain behavioural domains are (or are not) useful for linking purposes. At the very least, researchers examining similar crime types should begin sharing this information with one another to provide greater insight into these differences.

Implications for police practice. There are also a number of implications for police practice arising from this research. These implications are related to: (1) the ability

of practitioners to link crimes using statistical approaches, (2) the availability of a (possibly) more user-friendly statistical tool for linking purposes, and (3) the need for improved police data collection practices. Each of these implications is discussed in more detail below.

Linking crimes using statistical approaches. The most obvious implication of this research for police practice is that it demonstrates that BLA practitioners in Saint John and Quebec may be able to link serial crimes in a relatively accurate fashion if they employ one of the statistical approaches to linking developed in the current research. It is possible that adopting these statistical approaches could lead to an improvement in the linking decisions made by practitioners. However, it is important to note that whether or not the statistical models developed in the current study lead to worthwhile improvements in the linking decisions made in practice is an empirical question that can only be answered through future research.

Indeed, although published research has suggested that the unstructured clinical judgements made by humans are inferior to statistical BLA approaches (e.g., Bennell et al., 2010; Canter et al., 1991; Santtila, Korpela, et al., 2004), some unpublished research has found comparable accuracy rates between statistical and human linking decisions (Tonkin, 2012).[69] Regardless of the findings of past research, to determine the true value of the statistical tools developed in the current research, an implementation study within a jurisdiction (e.g., Saint John) would need to be conducted, where the real-world decisions of practitioners using the tool are compared to the decisions of practitioners who do not use the tool.

[69] See Mugford and Martineau (2014) for a full discussion of the methodological differences between these studies that may have led to these divergent findings.

In addition to comparing the number of correct linking decisions that are made in practice when using these tools, it is also important to compare the number of *errors* made by each approach. For instance, one common argument against using statistical BLA tools in practice is that the number of false positive decisions made by these tools is too large to warrant their use. However, the large number of false positives that are made using these tools is not a function of the tool itself; instead, it results from the low base rate of linked crime pairs that is an inevitable reality of most BLA tasks. Since practitioners still make linking decisions in practice despite the low base rate event they are trying to predict (i.e., linked crimes), a comparison of both the number of correct and incorrect decisions made using each approach (i.e., statistical versus traditional clinical judgement) should ultimately inform decisions about which approach to adopt.

The availability of a more user-friendly statistical tool. A related implication of the current research is that the CT-based models developed here may provide practitioners with a more user-friendly statistical tool; one that they may be more willing to adopt into practice. As discussed previously, it is quite simple to explain how a linking decision is reached using the standard CTs developed from both datasets. Moreover, the standard CT developed using the sexual assault data does lead to more idiographic linking decisions, which practitioners may appreciate. Attempting to consider all possible interactions amongst linking predictors in a completely unaided fashion (which is something that is done in BLA practice currently) is likely to be difficult and lead to cognitive overload issues.

The extent to which a CT-based approach to linking is valuable for practice can only be determined, however, by future research. Not only do we need to determine

whether statistical approaches lead to better decisions (better rates of correct and incorrect linking decisions, as discussed previously), but future research needs to determine if the apparent advantages of a CT-based linking model, *specifically,* carry over into practice. For instance, is it easy to understand the pathways identified by a CT approach? Do practitioners find it easy to explain the decision-making process of a CT approach?

One potential barrier to using statistical approaches to make BLA decisions in practice (including the CT approach) is that they do not allow the practitioner the discretion to override a particular linkage decision if they feel it is justified (e.g., based on case-specific factors that may not be adequately captured by statistical linking models). For instance, most linking research has not attempted to incorporate situational factors into linking models, which may understandably decrease an offender's across-crime consistency, and the few attempts to do so have unfortunately been unsuccessful (e.g., Woodhams et al., 2008). It may be that integrating a professional override component into a CT-based tool is something that would need to be done for practitioners to use this approach (or any statistical approach). Indeed, many of the offender risk assessment tools currently used in practice have a professional override component (e.g., the Level of Service/Case Management Inventory; Andrews, Bonta, & Wormith, 2004). Although research has found that professional overrides can result in less accurate decisions when compared to the original actuarial prediction (e.g., Hanson, Harris, Scott, & Helmus 2007), they may still be useful if they counteract resistance to adopting a structured approach to BLA into practice. Indeed, although overrides may not be optimal, it may

still be the case that more accurate decisions are made when actuarial tools with an override option are used than if these tools were not implemented at all.

The need for improved data collection practices. The final implication of this research for police practice concerns the data quality issues discussed earlier. As mentioned, it is possible that the poor quality of the data available for the break and enter sample contributed to the lower levels of linking accuracy achieved. This indicates that there may be a need for better data collection practices by the police.

For instance, merely adopting a standardized break and enter protocol could, at the very least, improve the consistency across reports for the crimes included in the police database. Not only may this improve our ability to link crimes, it may also improve the ability of police to secure convictions (e.g., by reminding officers that it is important to record certain information that they may otherwise forget to record when no standard protocols are followed). Indeed, in their survey of crime analysts in the UK ($N = 18$), Burrell and Bond (2011) found that a key theme to the responses received was the need for better quality of data, both in terms of the amount and accuracy of the information recorded in police reports. The fact that crime analysts support the need for improvements in practices by front-line police officers may increase the chance that future research partnerships may successfully address this issue.

Limitations of the Current Research

There are a number of limitations of the current research that warrant discussion, some of which have already been raised. These include: (1) the cross-validation procedure used, (2) how the behavioural domains were operationalized, (3) the decisions made to develop the CT models, (4) the approach used to establish ground truth (i.e., that

the linked crimes were truly committed by the same offenders), and (5) the lack of practitioner input. The results of this research should be considered in light of these limitations and future research should be conducted in an attempt to address them.

The cross-validation procedure used. As mentioned, split-half validation was used in the current study to determine the extent to which the linking models that were developed apply to new crime pairs of the same type (e.g., new break and enters, sexual assaults). This approach is not ideal since it ultimately means that half of the data are not available to estimate (or test) the model. Ideally, a more robust validation procedure would have been used, such as leave-one-out cross-validation, which has been used in recent linking research (e.g., Woodhams & Labuschagne, 2012; Woodhams & Tonkin, 2015). This procedure was not used in the current study because there was no identifiable way to use this validation approach for the ICT (or multiple CT/ICT) models. Given that one of the central goals of this research was to compare the relative predictive accuracy of all statistical approaches, it was essential that the same validation procedure be used across all statistical approaches.

Relatedly, it is also important for future research to validate linking models on a *true sample of new crimes*, as this is something that has been done rarely in linking research to date. That is, although linking research commonly cross-validates statistical models on a test sample of new crime *pairs,* the crime pairs included in the test sample are still ultimately comprised of crimes that were used to develop the model.[70] Only two studies to date (both unpublished) have applied linking models to a truly new sample of

[70] As mentioned, the methodology commonly followed in linking research that uses a logistic regression approach starts with constructing all possible pairs of crimes from a master dataset of single crimes. These crime pairs are then randomly split into a development and test sample. As such, information from the same single crimes is inevitably included in both the development and test samples.

crimes to determine the extent to which they may generalize across jurisdictions (Tonkin, 2012; Blaskovits, Bennell, & Emeno, 2013). Although Tonkin found that a number of logistic regression equations applied equally well to new data from different jurisdictions, Blaskovits and colleagues found a significant amount of shrinkage in the AUC when applying a model developed on a sample of UK burglaries (AUC = .92, 95% CIs [.85, .98]) to a sample of Finnish burglaries (AUC = .63, 95% CIs [.57, .69]). Future linking research should strive to cross-validate their BLA models in this way and, in cases when low levels of generalizability are found, determine why these issues might exist.

The operationalization of behavioural domains. In hindsight, another possible limitation of the current research was the way in which the behavioural domains were operationalized. Defining the behavioural domains in an atheoretical manner (based on their assumed function) may have led to lower levels of linking accuracy. This approach to defining behavioural domains may have also contributed to difficulties interpreting the different pathways that arise using the serial sexual assault data. If a secondary goal of linking research is to further our understanding of offender behaviour, then future research should examine different ways to use a CT approach (which does appear to capture the complexity in offending behaviour) with more theoretically-informed predictors (e.g., the temporal approach used by Ozeil et al., 2014 discussed earlier). Alternatively, it may be useful to explore the value of creating behavioural domains using different statistical procedures (e.g., multidimensional scaling, cluster analysis, principal component analysis, etc.). Some unpublished research suggests that forming domains through factor analysis may not add any predictive value beyond the approach to domain formation employed in the current dissertation (Tonkin, 2012); however, future research

should explore the extent to which this is true using other samples of crimes and different statistical procedures.

The decisions made regarding how to develop the CT models. Another limitation of the current research concerns the CHAID parameters that were used to construct the CT models. It is important to note that the combination of parameters selected in the current research, and the combination of predictors selected in Tonkin, Woodhams, et al.'s (2012) study, are only two of countless options for parameter selection. Indeed, as demonstrated earlier, changing the parameters will inevitably lead to changes in the models produced. The parameters in the current study were ultimately selected because it was believed that they simultaneously created a more parsimonious model (i.e., a model that was relatively simple and seemed to make intuitive sense while not capturing a great deal of "noise" in the data), while still also capitalizing on the ability of the CT approach to capture more complex "signals" in the data than the traditional main effect approach.

All of the CT results presented in this dissertation are entirely dependent on the parameters that were selected and it is certainly plausible that the CT models developed were not constructed in the most optimal way possible. As such, the results presented in these studies must be interpreted in light of this limitation. Future research should examine the extent to which systematically varying different combinations of CHAID parameters can impact the results. Future research should also attempt to determine which parameters are most suitable for developing CT models within the BLA context.

A second issue related to how the CTs were constructed is the use of the two thresholds (i.e., twice and half the base rate) proposed by Monahan et al. (2001). Of

course, if these thresholds were changed, different results would likely have emerged across both studies in terms of the number of linked, unlinked, and unclassified nodes (and consequently, this may have altered whether or not additional iterations of the CTs were produced). Given that no strong rationale for these thresholds was ever provided by Monahan et al. (2001), future research should examine if there are more suitable thresholds for constructing ICTs.

Finally, not only may there be more appropriate thresholds for *constructing* ICTs, but the thresholds adopted in these studies may also not be optimal for making linkage decisions in practice. Future research needs to more deliberately examine what the costs are of making the different types of incorrect linking decisions and what the benefits are of making the different types of correct linking decisions. For instance, practitioners could be surveyed to explore the thresholds they currently use when making linking decisions. They could also be asked to describe how they weight the relative costs and benefits of the various types of linking decisions they make. This would help inform us as to what decision thresholds are relevant within this particular context.

Establishing ground truth. Although ground truth was established on the basis of convictions in the serial sexual assault data (E. Beauregard, personal communication, December 2, 2015), the crimes comprising the Saint John data were only *cleared by charge*, as the court and police database do not communicate with one another about whether or not the offenders were ultimately convicted. The obvious consequence of using charges as the benchmark for ground truth is that we may be less certain that the crimes have been attributed to the correct offender, which could have contributed to the reduced linking accuracy observed in the Saint John results.

What is important to note, however, is that a more conservative sampling approach was used with the Saint John data than many previous linking studies in the sense that crimes that were *cleared otherwise* were excluded. The *cleared otherwise* crimes in the Saint John database are similar to crimes that have been *taken into consideration* (called TICs) in the UK, whereby the police may not have enough evidence to charge the individual with a certain offence, but they may clear the offence if the offender admits responsibility (affording the benefit of not being prosecuted for these crimes at a later date). These TICs have often been incorporated into the samples used in previous linking research (e.g., Bennell & Jones, 2005). As explained by Bennell (2002), some TICs may be based on the fact that the crimes for which the offender was convicted may be behaviourally similar to the crimes for which they were not convicted, but were attributed to them via this TIC process. As such, inclusion of TICs could lead to inflated estimates of behavioural consistency among offenders. As Snook, Luther, and MacDonald (2014) suggest, research should explore whether the inclusion of TICs compromises the validity of the resulting linking models. Moreover, research should examine the extent to which relying on charges versus convictions may impact the linking models produced.

The existence of multiple linking tasks. As previously mentioned, numerous types of linking tasks exist in practice, including: (1) searching for pairs of linked crimes in a large database of crimes, (2) being provided with an "index" crime (e.g., by a police investigator) and searching a large database for possible linkages to that index crime, and (3) being provided with a pre-set number of crimes (e.g., 5 or 6) and determining whether any of those crimes are likely to be linked or not. The current dissertation deals directly

with the first task. As such, the results and conclusions are applicable to the first task only. Future research is needed to determine the extent to which findings from studies employing the current methodology can extend to these different types of linkage tasks faced by police practitioners in the real world.

Lack of practitioner input. Relatedly, a final limitation of the current research is that a number of the fundamental questions about CT-based linking models can only be answered by testing them with practitioners who make linking decisions in practice. As mentioned, future research needs to be conducted that determines whether or not practitioners are willing to use CTs in practice, what barriers might exist for their use, and whether they do in fact lead to the advantages proposed (e.g., is it easier to understand the statistical approaches behind the CT model? Is it easier to explain the decisions made to another individual?).

Conclusion

Although research has consistently supported the use of empirically-derived BLA approaches, BLA practitioners continue to make linking decisions by relying on their investigative knowledge rather than systematic relationships between linking features and linkage status. The current study represents a valuable contribution to BLA research in that it attempted to address some common concerns that practitioners have with currently proposed BLA methods by introducing a novel approach to BLA using CTs. Given the appealing qualities of the CT approach found in the current research, linkage practitioners may be more open to adopting a statistical tool for BLA into practice. This is, however, a question that needs to be answered by future research examining the various issues outlined in this dissertation. Future research should also prioritize improvements to data

collection methods and explore the extent to which the statistical BLA models developed thus far can be successfully used in practice. To do this, more communication is needed between researchers and practitioners, something that has been happening in more recent years as a result of researcher-practitioner collaborations (e.g., the Crime Linkage International Network [C-LINK]; Bennell & Woodhams, 2014). It is partnerships such as that that will ultimately lead to the research that is needed to move the BLA field forward.

References

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., et al. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counselling Psychologist, 34*(3), 341-382. doi:10.1177/0011000005285875

Alison, L. J., Snook, B., & Stein, K. L. (2001). Unobtrusive measurement: Using police information for forensic research. *Qualitative Research, 1*, 241-254. doi:10.1177/146879410100100208

Andrews, D. A., Bonta, J., & Wormith, S. J. (2004). *The Level of Service/Case Management Inventory (LS/CMI)*. Toronto, Canada: Multi-Health Systems.

Arkes, H. R. (1981). Impediments to accurate clinical judgment and possible ways to minimize their impact. *Journal of Consulting and Clinical Psychology, 49*(3), 323-330. doi:10.1037/0022-006X.49.3.323

Banks, S., Robbins, P. C., Silver, E., Vesselinov, R., Steadman, H. J., Monahan J., … Roth, L. H. (2004). A multiple-models approach to violence risk assessment among people with mental disorder. *Criminal Justice and Behavior, 31*(3), 324-340. doi:10.1177/0093854804263635

Beauregard, E. (2005). *Hunting process of serial sex offenders: A rationale choice approach* (Unpublished doctoral dissertation). University of Montreal, Canada.

Beauregard, E., Leclerc, B., & Lussier, P. (2012). Decision making in the crime commission process: Comparing rapists, child molesters, and victim-crossover sex offenders. *Criminal Justice and Behavior, 39*, 1275-1295. doi:10.1177/0093854812453120

Bennell, C. (2002). *Behavioural consistency and discrimination in serial burglary* (Unpublished doctoral dissertation). University of Liverpool, Liverpool, UK.

Bennell, C. (2005). Improving police decision making: General principles and practical applications of receiver operating characteristic analysis. *Applied Cognitive Psychology, 19*(9), 1157-1175. doi:10.1002/acp.1152

Bennell, C., Bloomfield, S., Snook, B., Taylor, P., & Barnes, C. (2010). Linkage analysis in cases of serial burglary: Comparing the performance of university students, police professionals, and a logistic regression model. *Psychology, Crime, & Law, 18*(6), 507-524. doi:10.1080/10683160902971030

Bennell, C., & Canter, D. V. (2002). Linking commercial burglaries by modus operandi: Tests using regression and ROC analysis. *Science & Justice, 42*(3), 153-164. doi:10.1016/S1355-0306(02)71820-0

Bennell, C., & Jones, N. J. (2005). Between a ROC and a hard place: A method for linking serial burglaries by modus operandi. *Journal of Investigative Psychology and Offender Profiling, 2*(1), 23-41. doi:10.1002/jip.21

Bennell, C., Jones, N. J., & Melnyk, T. (2009). Addressing problems with traditional crime linkage methods using receiver operating characteristic analysis. *Legal and Criminological Psychology*, *14*(2), 293-310. doi:10.1348/135532508X349336

Bennell, C., Mugford, R., Ellingwood, H., & Woodhams, J. (2014). Linking crimes using behavioural clues: Current levels of linking accuracy and strategies for moving forward. *Journal of Investigative Psychology and Offender Profiling, 11*, 29-56. doi:10.1002/jip.1395

Bennell, C., Snook, B., MacDonald, S., House, J. C., Taylor, P. J. (2012). Computerized crime linkage systems: A critical review and research agenda. *Criminal Justice and Behavior, 39*, 620-632. doi:10.1177/0093854811435210

Bennell, C., & Woodhams, J. (2014). Crime linkage research: Where to from here? In J. Woodhams & C. Bennell (Eds.), *Crime linkage: Theory, research, and practice* (pp. 369-372). Boca Raton, FL: CRC Press.

Bennell, C., Woodhams, J., & Beauregard, E. (2015). *Investigating individual differences in the expression of behavioural consistency in crime series using ICT analyses*. Manuscript in preparation.

Berk, R., Sherman, L., Barnes, G., Kurtz, E., & Ahlman, L. (2009). Forecasting murder within a population of probationers and parolees: A high stakes application of statistical learning. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 172*(1), 191-211. doi:10.1111/j.1467-985X.2008.00556.x

Blaskovits, B. L., Bennell, C., & Emeno, K. (2013, September). *The effectiveness of actuarial tools for linking serial burglaries.* Poster presented at The Society for Police and Criminal Psychology conference, Ottawa, Ontario.

Bosson, J., & Labarere, J. (2006). Determining indications for care common to competing guidelines by using classification tree analysis: Application to the prevention of venous thromboembolism in medical inpatients. *Medical Decision Making, 26*(1), 63-75. doi:10.1177/0272989X05284105

Bouhana, N., Johnson, S. D., & Porter, M. (2014). Consistency and specificity in burglars who commit prolific residential burglary: Testing the core assumptions

underpinning behavioural crime linkage. *Legal and Criminological Psychology*. Advance online publications. doi:10.1111/lcrp.12050

Bowerman, B. L., & O'Connell, R. T. (1990). *Linear statistical models: An applied approach* (2nd ed.). Belmont, CA: Duxbury

Brookman, F. (2010). Beyond the interview: Complementing and validating accounts of incarcerated violent offenders. In W. Bernasco (Ed.), *Offenders on offending: Learning about crime from criminals* (pp. 84-105). Portland, OR: Willan.

Burgess, E. (1928). Factors determining success or failure on parole. In A. A. Bruce (Ed.), *The workings of the indeterminate sentence law and the parole system in Illinois* (pp.205-248). Springfield, IL: Illinois State Board of Parole.

Burrell, A., & Bull, R. (2011). A preliminary examination of crime analysts' views and experiences of comparative case analysis. *International Journal of Police Sciences & Management, 13*, 2-15. doi:10.1350/ijps.2011.13.1.212

Burrell, A., Bull, R., & Bond, J. (2012). Linking personal robbery offences using offender behaviour. *Journal of Investigative Psychology and Offender Profiling*, *9*, 201-222. doi: 10.1002/jip.1365

Burrell, A., Bull, R., Bond, J., & Harrington, G. (2015). Testing the impact of group offending on behavioural similarity in serial robbery. *Psychology, Crime & Law, 21*, 551-569. doi:10.1080/1068316X.2014.999063

Canter, D. V. (1995). Psychology of offender profiling. In R. Bull & D. Carson (Eds.), *Handbook of psychology in legal contexts* (pp. 343-355). Chichester, UK, Wiley.

Canter, D. V., Heritage, R., Wilson, M., Davies, A., Kirby, S.,… Holden, R.(1991). *A facet approach to offender profiling*. London, UK: Home Office.

Canter, D. V., & Larkin, P. (1993). The environmental range of serial rapists. *Journal of Environmental Psychology, 13*(1), 63-69. doi:10.1016/S0272-4944(05)80215-4

Charron, A., & Woodhams, J. (2010). A qualitative analysis of mock jurors' deliberations of linkage analysis evidence. *Journal of Investigative Psychology and Offender Profiling, 7*, 165-183. doi:10.1002/jip.119

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). New York, NY: Academic Press.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159. doi:10.1037/0033-2909.112.1.155

Copes, H., & Hochstetler, A. (2010). Interviewing the incarcerated: Pitfalls and promises. In W. Bernasco (Ed.), *Offenders on offending: Learning about crime from criminals* (pp. 49-67). Portland, OR: Willan.

Criminal Code R.S.C., c. C-46 (1985). Retrieved from http://laws-lois.justice.gc.ca/eng/acts/C-46/

Daubert v. Merrell Dow Pharmaceuticals, Inc. U.S., 125 L. Ed. 2d 469, 113 S. Ct. 2786 (1993).

Davies, K., Tonkin, M., Bull, R., & Bond, J. W. (2012). The course of case linkage never did run smooth: A new investigation to tackle the behavioural changes in serial car theft. *Journal of Investigative Psychology and Offender Profiling*, *9*, 274-295. doi:10.1002/jip.1369

Dawes, R. M. (1989). Experience and validity of clinical judgment: The illusory correlation. *Behavioral Sciences & the Law, 7*(4), 457-467. doi:10.1002/bsl.2370070404

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*(2), 95-106. doi:10.1037/h0037613

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243,* 1668-1674. doi:10.1126/science.2648573

Deslauriers-Varin, N., & Beauregard, E. (2014a). Consistency in crime site selection: An investigation of crime sites used by serial sex offenders across crime series. *Journal of Criminal Justice, 42*, 123-133. doi:10.1016/j.jcrimjus.2013.09.005

Deslauriers-Varin, N., & Beauregard, E. (2014b). Unravelling crime series patterns amongst serial sex offenders: Duration, frequency, and environmental consistency. *Journal of Investigative Psychology and Offender Profiling, 11*, 253-275. doi:10.1002/jip.1418

de Zoete, J., Sjerps, M., Lagnado, D., & Fenton, N. (2015). Modelling crime linkage with Bayesian networks. *Science & Justice, 55*, 209-219. doi:10.1016/j.scijus. 2014.11.005

Drawve, G., Walker, J. T., & Felson, M. (2015). Juvenile offenders: An examination of distance-to-crime and crime clusters. *Cartography and Geographic Information Science, 42*, 122-133. doi:10.1080/15230406.2014.963677

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* New York, NY: Chapman & Hall.

Elffers, H. (2010). Misinformation, misunderstanding and misleading as validity threats to offenders' accounts of offending. In W. Bernasco (Ed.), *Offenders on offending: Learning about crime from criminals* (pp. 13-22). Portland, OR: Willan.

Federal Bureau of Investigation (2008). *Serial murder: Multidisciplinary perspectives for investigators*. Washington, DC: US Department of Justice. Retrieved from http://www.fbi.gov/stats-services/publications/serial-murder/serial-murder-july-2008-pdf

Field, A. (2013). *Discovering statistics using SPSS (and sex, drugs, and rock 'n' roll)* (4th ed.). Thousand Oaks, CA: Sage Publications, Inc.

Fossi, J. J., Clarke, D. D., & Lawrence, C. (2005). Bedroom rape: Sequences of sexual behaviour in stranger assaults. *Journal of Interpersonal Violence*, *20*, 1444–1466. doi:10.1177/0886260505278716

Furr, R. M., & F under, D. C. (2004). Situational similarity and behavioural consistency: Subjective, objective, variable-centred, and person-centred approaches. *Journal of Research in Personality, 38*, 421-447. doi:10.1016/j.jrp.2003.10.001

Funder, D. C., & Colvin, C. R. (1991). Explorations in behavioral consistency: Properties of persons, situations, and behaviors. *Journal of Personality and Social Psychology, 60*(5), 773-794. doi:10.1037/0022-3514.60.5.773

Gardner, W., Lidz, C. W., Mulvey, E. P., & Shaw, E. C. (1996) A comparison of actuarial methods for identifying repetitively violent patients with mental illness. *Law and Human Behavior*, *20*(1), 35-48. doi:10.1007/BF01499131

Gee, D., & Belofastov, A. (2014). Sex crime linkage: Sexual fantasy and offense plasticity. In J. Woodhams & C. Bennell (Eds.), *Crime linkage: Theory, research, and practice* (pp. 33-53). Boca Raton, FL: CRC Press.

Grann, M., & Långström, N (2007). Actuarial assessment of violence risk: To weigh or not to weigh? *Criminal Justice and Behavior, 34*, 22-36. doi:10.1177/0093854806290250

Green, E. J., Booth, C. E., & Biderman, M. D. (1976). Cluster analysis of burglary M/Os. *Journal of Police Science and Administration, 4*(4), 382-388. Retrieved from http://www.apa.org/pubs/databases/psycinfo/index.aspx

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, & Law*, *2*, 293-323. doi:10.1037/1076-8971.2.2.293

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*, 19-30. doi:10.1037//1040-3590.12.1.19

Grubin, D., Kelly, P., & Brunsdon, C. (2001). *Linking serious sexual assaults through behaviour*. London, UK: Home Office.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*, 29-36. doi:10.1148/radiology.143.1.7063747

Hanson, R. K., Harris, A. J. R., Scott, T., & Helmus, L. (2007). *Assessing the risk of sexual offenders on community supervision: The Dynamic Supervision Project* (Corrections Research User Report No. 2007-05). Ottawa, Canada: Public Safety Canada.

Hazelwood, R. R., & Warren, J. I. (2003). Linkage analysis: Modus operandi, ritual, and signature in serial sexual crime. *Aggression and Violent Behavior*, *8*(6), 587-598. doi:10.106/S1359-1789(02)00106-4

Hewitt, A., Beauregard, E., & Davies, G. (2012). "Catch and release": Predicting encounter and victim release location choice in serial rape events. *Policing: An International Journal of Police Strategies & Management, 35*, 835-856. doi:10.1108/13639511211275814

Hill, T., & Lewicki, P. (2006). *Statistics: Methods and applications: A comprehensive reference for science, industry, and data mining.* Tulsa, OK: StatSoft, Inc.

Homel, R., MacIntyre, S., & Wortley, R. (2014). How house burglars decide on targets: A computer-based scenario approach. In R. Wortley & B. LeClerc (Eds.), *Cognition and crime: Offender decision making and script analyses* (pp. 26-47). New York, NY: Routledge.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression, second edition.* New York, NY: Wiley.

Hsu, C. H. C., & Kang, S. K. (2007). CHAID-based segmentation: International visitors' trip characteristics and perceptions. *Journal of Travel Research, 46*(2), 207-216. doi:10.1177/0047287507299571

Jaccard, P. (1908). Nouvelle recherches sur la distribution florale. *Bulletin de la Société Vaudoise des Sciences Naturelles, 44*, 223-270.

Johnson, S. D., & Bowers, K. J. (2004). The burglary as clue to the future: The beginnings of prospective hot-spotting. *European Journal of Criminology, 1*, 237-255. doi:10.1177/ 1477370804041252

Johnson, S. D., Summers, L., & Pease, K. (2009). Offender as forager? A direct test of the boost account of victimization. *Journal of Quantitative Criminology, 25*, 181-200. doi:10.1007/s10940-008-9060-8

Kangas, L. (2001). *Artificial neural network system for classification of offenders in murder and rape cases, executive summary.* Washington, DC: National Institute of Justice. Retrieved from https://www.ncjrs.gov/pdffiles1/nij/grants/190983.pdf

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 29*(2), 119-127. doi:10.2307/2986296

Keppel, R. D., & Weis, J. G. (2006). Improving the investigation of violent crime: The homicide investigation and tracking system. In R. D. Keppel (Ed.), *Offender Profling* (pp. 361-371). Toronto, Canada: Thomson Nelson.

Knezevic, A. (2008, October). StatNews #73: Overlapping confidence intervals and statistical significance. *StatNews*. Retrieved from: https://www.cscu.cornell.edu/news/statnews/stnews73.pdf

Kocsis, R. N., & Irwin, H. J. (1997). An analysis of spatial patterns in serial rape, arson, and burglary: The utility of the Circle Theory of environmental range for psychological profiling. *Psychiatry, Psychology, and Law, 4*(2), 195-206. doi:10.1080/13218719709524910

Labuschagne, G. N. (2012). The use of linkage analysis as an investigative tool and evidential material in serial offenses. In K. Borgeson & K. Kuehnle (Eds.), *Serial offenders: Theory and practice* (pp.187-215). Sudbury, MA: Jones & Bartlett Learning.

Labuschagne, G. N. (2014). The use of linkage analysis evidence in serial offense trials. In J. Woodhams & C. Bennell (Eds.), *Crime linkage: Theory, research, and practice* (pp. 197-224). Boca Raton, FL: CRC Press.

Lawrence, C., Fossi, J. J., & Clarke, D. D. (2010). A sequential examination of offenders' verbal strategies during stranger rapes: the influence of location. *Psychology, Crime & Law*, *16*, 381–400. doi:10.1080/10683160902754964

Leclerc, B., Smallbone, S. W., & Wortley, R. (2014). Interpersonal scripts and victim reaction in child sexual abuse: A quantitative analysis of the offender-victim interchange. In R. Wortley & B. LeClerc (Eds.), *Cognition and crime: Offender decision making and script analyses* (pp. 101-119). New York, NY: Routledge.

Lee, J., Harp, S. S., Horridge, P. E., & Russ, R. R. (2003). Targeting multicultural purchase and consumption segments in the leather handbag market: Product development and merchandising implications. *Family and Consumer Sciences Research Journal, 31*(3), 297-330. doi:10.1177/1077727X02250141

Liu, Y. Y., Yang, M., Ramsay, M., Li, X. S., & Coid, J. W. (2011). A comparison of logistic regression, classification and regression tree, and neural networks models in predicting violent re-offending. *Journal of Quantitative Criminology, 27*, 547-573. doi:10.1007/s10940-011-9137-7

MacLachlan, D. L., & Johansson, J. K. (1981). Market segmentation with multivariate aid. *Journal of Marketing, 45*(1), 74-84. doi:10.2307/1251722

Magidson, J. (1993). The use of the new ordinal algorithm in CHAID to target profitable segments. *The Journal of Database Marketing, 1*, 29-48.

Markson, L., Woodhams, J., & Bond, J. W. (2010). Linking serial residential burglary: Comparing the utility of modus operandi behaviours, geographical proximity and temporal proximity. *Journal of Investigative Psychology and Offender Profiling*, *7*(2), 91-107. doi:10.1002/jip.120

Martineau, M. M., & Corey, S. (2008). Investigating the reliability of the Violent Crime Linkage Analysis System (ViCLAS) crime report. *Journal of Police and Criminal Psychology, 23*(2), 51-60. doi:10.1007/s11896-008-9028-5

Melnyk, T., Bennell, C., Gauthier, D. J., & Gauthier, D. (2011). Another look at across-crime similarity coefficients for use in behavioural linkage analysis: An attempt to replicate Woodhams, Grant, and Price (2007). *Psychology, Crime & Law, 17*(4), 359-380. doi:10.1080/10683160903273188

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician, 54*, 17-24.

Menard, S. (1995). *Applied logistic regression analysis*. Series: Quantitative Applications in the Social Sciences. Thousand Oaks, CA: Sage

Menard, S. (2002). *Applied logistic regression analysis* (2nd ed.). Series: Quantitative Applications in the Social Sciences. Thousand Oaks, CA: Sage.

Metcalfe, J., & Mischel, W. (1999). A hot/cool-system analysis of delay of gratification: Dynamics of willpower. *Psychological Review, 106*, 3-19. doi:10.1037/0033-295X.106.1.3

Mischel, W. (2009). From *personality and assessment* (1968) to personality science (2009). *Journal of Research in Personality, 43*, 282-290. doi:10.1016/j.jrp.2008.12.037

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics and invariance in personality structure. *Psychological Review, 102,* 246–268. doi:10.1037/0033-295X.102.2.246

Monahan, J., Steadman, H. J., Silver, E., Appelbaum, P. S., Robbins, P. C., … Banks, S. (2001). *Rethinking risk assessment: The MacArthur study of mental disorder and violence*. New York, NY: Oxford University Press.

Mugford, R., & Martineau, M. (2014). The ability of human judges to link crimes using behavioural information: Current knowledge and unresolved issues. In J. Woodhams & C. Bennell (Eds.), *Crime linkage: Theory, research, and practice* (pp. 251-277). Boca Raton, FL: CRC Press.

Nee, C. (2010). Research on residential burglary: Ways of improving validity and participants' recall when gathering data. In W. Bernasco (Ed.), *Offenders on offending: Learning about crime from criminals* (pp. 231-245). Portland, OR: Willan.

Nee, C. (2015). Understanding expertise in burglars: From pre-conscious scanning to action and beyond. *Aggression and Violent Behavior, 20*, 53-61. doi:10.1016/j.avb.2014.12.006

Neuilly, M., Zgoba, K. M., Tita, G. E., & Lee, S. S. (2011). Predicting recidivism in homicide offenders using classification tree analysis. *Homicide Studies, 15*(2), 154-176. doi:10.1177/1088767911406867

Ohlin, L. E. (1951). *Selection for parole: A manual of parole prediction*. New York, NY: Russell Sage Foundation.

Oziel, S., Goodwill, A., & Beauregard, E. (2015). Variability in behavioural consistency across temporal phases in stranger sexual offences. *Journal of Police and Criminal Psychology, 30*, 176-190. doi:10.1007/s11896-014-9150-5

Peduzzi, P., Concata, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology, 49*(12), 1373-1379. doi:10.1016/S0895-4356(96)00236-3

Perrault, W. D., & Barksdale, H. C. (1980). A model-free approach for analysis of complex contingency data in survey research. *Journal of Marketing Research, 17*(4), 503-515. doi:10.2307/3150503

Podgorelec, V., Kokol, P., Stiglic, B., & Rozman, I. (2002). Decision trees: An overview and their use in medicine. *Journal of Medical Systems, 26*(5), 445-463. doi:10.1023/A:1016409317640

Rainbow, L. (2014). A practitioner's perspective: Theory, practice, and research. In J. Woodhams & C. Bennell (Eds.), *Crime linkage: Theory, research, and practice* (pp. 173-196). Boca Raton, FL: CRC Press.

Rainbow, L., & Gregory, A. (2011). What behavioural investigative advisors actually do. In L. Alison & L. Rainbow (Eds.), *Professionalizing offender profiling: Forensic and investigative psychology in practice* (pp.18-34). New York, NY: Routledge.

Refaeilzadeh, P., Tang, L., & Lui, H. (2009). Cross-validation. In L. Lui & M.T. Özsu (Eds.), *Encyclopedia of Database Systems* (pp. 532-538). Springer. doi:10.1007/978-0-387-39940-9

Reid, J. A., Beauregard, E., Fedina, K. M., & Frith, E. N. (2014). Employing mixed methods to explore motivational patterns of repeat sex offenders. *Journal of Criminal Justice, 42*, 203-212. Doi:j.jcrimjus.2013.06.008

Rokach, L., & Maimon, O. (2008). *Data mining with decision trees: Theory and applications*. Hackensack, NJ: World Scientific Publishing Co.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (2nd ed.). Newbury Park, CA: Sage

Rossmo, D. K. (2000). *Geographic profiling*. Boca Raton, FL: CRC Press.

Santtila, P., Fritzon, K., & Tamelander, A. L. (2004). Linking arson incidents on the basis of crime scene behavior. *Journal of Police and Criminal Psychology*, *19*(1), 1-16. doi:10.1007/BF02802570

Santtila, P., Junkkila, J., & Sandnabba, N. K. (2005). Behavioural linking of stranger rapes. *Journal of Investigative Psychology and Offender Profiling*, *2*(2), 87-103. doi:10.1002/jip.26

Santtila, P., Korpela, S., & Häkkänen, H. (2004). Expertise and decision-making in the linking of car crime series. *Psychology, Crime & Law, 10*(2), 97-112. doi:10.1080/1068316021000030559

Santtila, P., Pakkanen, T., Zappalà, A., Bosco, D., Valkama, M., & Mokros, A. (2008). Behavioural crime linking in serial homicide. *Psychology, Crime & Law, 14*(3), 245-265. doi:10.1080/10683160701739679

Silver, E., & Chow-Martin, L. (2002). A multiple models approach to assessing recidivism risk: Implications for judicial decision-making. *Criminal Justice and Behavior, 29*(5), 538-568. doi:10.1177/009385402236732

Slater, C., Woodhams, J., & Hamilton-Giachritsis, C. (2015). Testing the assumptions of crime linkage with stranger sex offenses: A more ecologically-valid study. *Journal of Police and Criminal Psychology, 30*, 261-273. doi:10.1007/s11896-014-9160-3

Snook, B., Luther, K., House, J. C., & Bennell, C. (2012). The violent crime linkage analysis system: A test of interrater reliability. *Criminal Justice and Behavior, 39*, 607-691. doi:10.1177/0093854811435208

Snook, B., Luther, K., & MacDonald, S. (2014). Linking crimes with spatial behaviour: A need to tackle some remaining methodological concerns. In J. Woodhams & C. Bennell (Eds.), *Crime linkage: Theory, research, and practice* (pp. 83-105). Boca Raton, FL: CRC Press.

Sonquist, J. A., Baker, E. L., and Morgan, J. N. (1971) *Searching for Structure*. Ann Arbor, MI: Institute for Social Research, The University of Michigan.

SPSS (2012). *IBM SPSS decision trees 21*. SPSS, Inc. Retrieved from http://www.sussex.ac.uk/its/pdfs/SPSS_Decision_Trees_21.pdf

State v. Fortin, 178 N.J. 540, 843 A.2d 974 (2004).

Statistics Canada (2012a). *Saint John, New Brunswick (Code 1301006) and Saint John, New Brunswick (Code 1301) (table). census profile. 2011 census. Statistics Canada catalogue no. 98-316-XWE.* Retrieved from http://www12.statcan.gc.ca/census-recensement/2011/dp-pd/prof/index.cfm?Lang=E

Statistics Canada (2012a). *Population and dwelling count highlights table, 2011 census. Population and dwelling counts, for Canada, provinces and territories, 2011 and*

2006 censuses. Retrieved from https://www12.statcan.gc.ca/census-recensement/2011/dp-pd/hlt-fst/pd-pl/Table-Tableau.cfm?LANG= Eng&T=101&SR=1&S=9&O=A

Steadman, H. J., Silver, E., Monahan, J., Appelbaum, P. S., Robbins, P. C., Mulvey, E. P., … Banks, S. (2000). A classification tree approach to the development of actuarial violence risk assessment tools. *Law and Human Behavior, 24*(1), 83-100. doi:10.1023/A:1005478820425

Sullivan, D. J., & van Zyl (2008). The well-being of children in foster care: Exploring physical and mental health needs. *Children and Youth Services Review, 30*(7), 774-786. doi:10.1016/j.childyouth.2007.12.005

Summers, L., Johnson, S. D., & Rengert, G. (2010). The use of maps in offender interviews. In W. Bernasco (Ed.), *Offenders on offending: Learning about crime from criminals* (pp. 246–272). Portland, OR: Willan.

Swets, J. A. (1986). Indices of discrimination of diagnostic accuracy: their ROCs and implied models. *Psychological Bulletin, 99*, 100-117. doi:10.1037//0033-2909.99.1.100

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*(4857), 1285-1293. doi:10.1126/science.3287615

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, *1*(1), 1-26. doi:10.1111/1529-1006.001

Tabachnick, B. G., & Fidell, L. D. (2007). *Using multivariate statistics* (5th edition). Boston, MA: Pearson.

Tonkin, M. (2011). *Behavioural case linkage: Generalisability, ecological validity, and methodology* (Unpublished doctoral dissertation). University of Leicester, United Kingdom.

Tonkin, M. (2014). Testing the theories underpinning crime linkage. In J. Woodhams & C. Bennell (Eds.), *Crime linkage: Theory, research, and practice* (pp. 107-139). Boca Raton, FL: CRC Press.

Tonkin, M., Grant, T., & Bond, J.W. (2008). To link or not to link: A test of the case linkage principles using serial car theft data. *Journal of Investigative Psychology and Offender Profiling, 5*(1-2), 59-77. doi:10.1002/jip.74

Tonkin, M., Santtila, P., & Bull, R. (2012). The linking of burglary crimes using offender behaviour: Testing research cross-nationally and exploring methodology. *Legal and Criminological Psychology, 17*, 276-293. doi:10.1111/j.2044-8333.2010.02007.x

Tonkin, M., & Woodhams, J. (2015). The feasibility of using crime scene behaviour to detect versatile serial offenders: An empirical test of behavioural consistency, distinctiveness, and discrimination. *Legal and Criminological Psychology.* Advance online publication. doi:10.1111/lcrp.12085

Tonkin, M., Woodhams, J., Bull, R., & Bond, J.W. (2012). Behavioural case linkage with solved and unsolved crimes. *Forensic Science International, 222*, 146-153. doi: 10.1016/j.forsciint.2012.05.017

Tonkin, M., Woodhams, J., Bull, R., Bond, J. W., & Palmer, E. J. (2011). Linking different types of crime using geographical and temporal proximity. *Criminal Justice and Behavior, 38*(11), 1069-1088. doi:10.1177/0093854811418599

Tonkin, M., Woodhams, J., Bull, R., Bond, J. W., & Santtila, P. (2012). A comparison of logistic regression and classification tree analysis for behavioural case linkage. *Journal of Investigative Psychology and Offender Profiling, 9*, 235-258. doi:10.1002/jip.1367

van Diepen, M., & Franses, P. H. (2006). Evaluating chi-squared automatic interaction detection. *Information Systems, 31*(8), 814-831. doi:10.1016/j.is.2005.03.002

van Middelkoop, M., Borgers, A., & Timmermans, H. (2003). Inducing heuristic principles of tourist choice of travel mode: A rule-based approach. *Journal of Travel Research, 42*(1), 75-83. doi:10.1177/0047287503254116

Wang, T., Qin, Z., Jin, Z., & Zhang, S. (2010). Handling over-fitting in test cost-sensitive decision tree learning by feature selection, smoothing and pruning. *The Journal of Systems and Software, 83*, 1137-1147. doi:10.1016/j.jss.2010.01.002

Weller, A. F., Harris, A. J., Ware, A., Jarvis, P. S. (2006). Determining the saliency of feature measurements obtained from images of sedimentary organic matter for use in its classification. *Computers & Geosciences, 32*(9), 1357-1367. doi:10.1016/j.cageo.2005.12.007

Wilcox, R. R. (2001). *Fundamentals of modern statistical methods: Substantially improving power and accuracy.* New York, NY: Springer-Verlag.

Winter, J. M. (2014). Exploring if(situation)…then(behavior) contingencies in interpersonal crimes. In J. Woodhams & C. Bennell (Eds.), *Crime linkage: Theory, research, and practice* (pp. 303-335). Boca Raton, FL: CRC Press.

Winter, J. M., Lemeire, J., Meganck, S., Geboers, J., Rossi, G., & Mokros, A. (2013). Comparing the predictive accuracy of case linkage methods in serious sexual

assaults. *Journal of Investigative Psychology and Offender Profiling, 10*, 28-56. doi: 10.1002/jip.1372

Woodhams, J., & Bennell, C. (2014). Consistency and distinctiveness of criminal behavior. In J. Woodhams & C. Bennell (Eds.), *Crime linkage: Theory, research, and practice* (pp. 11-31). Boca Raton, FL: CRC Press.

Woodhams, J., Hollin, C. R., & Bull, R. (2007). The psychology of linking crimes: A review of the evidence. *Legal and Criminological Psychology, 12*, 233-249. doi:10.1348/135532506X118631

Woodhams, J., Hollin, C. R., & Bull, R. (2008). Incorporating context in linking crimes: An exploratory study of the relationship between behavioural consistency and situational similarity. *Journal of Investigative Psychology and Offender Profiling, 5, 1-23.* doi:10.1002/jip.75

Woodhams, J., & Komarzynska, K. (2014). The effect of mental disorder on crime scene behavio, its consistency, and variability. In J. Woodhams & C. Bennell (Eds.), *Crime linkage: Theory, research, and practice* (pp. 55-82). Boca Raton, FL: CRC Press.

Woodhams, J., & Labuschagne, G. (2012). A test of case linkage principles with solved and unsolved serial rapes. *Journal of Police and Criminal Psychology, 27*, 85-98. doi:10.1007/s11896-011-9091-1

Woodhams, J., & Toye, K. (2007). An empirical test of the assumptions of case linkage and offender profiling with serial commercial robberies. *Psychology, Public Policy, and Law, 13*(1), 59-85. doi:10.1037/1076-8971.13.1.59

Wormith, S. J., Hogg, S., & Guzzo, L. (2012). The predictive validity of a general risk/needs assessment inventory on sexual offender recidivism and an exploration of the professional override. *Criminal Justice and Behavior, 39*, 1509-1535. doi:10.1177/0093854812455741

Appendix A:
Serial Break and Enter Behavioural Checklist

Numbers in parentheses after each behaviour correspond to that particular behaviour's frequency of occurrence in the final reduced dataset where the impact of prolific offenders was controlled using the outlier approach.

Target Selection (24 variables):

Detached dwelling (e.g., single home house) (47.6%)
Non-detached dwelling (e.g., townhouse) (7.1%)
Apartment (44.1%)
Residence is on basement level or first floor of multi-storey building (14.7%)
Second floor or above of multi-storey building (11.8%)
Downtown (16.5%)
Suburban (68.8%)
Rural (14.1%)
Secluded property (15.9%)
Multiple neighbours (95.3%)
Single neighbour (1.2%)
No neighbours (3.5%)
Multiple floors in home (55.9%)
Single floor home (44.1%)
Residents were present at the time of the break and enter (17.1%)
Residents were temporarily away (less than 24 hours) (55.9%)
Residents were away for more than one day (10.0%)
Unoccupied (i.e., vacant apartment/house, etc.) (4.7%)
Alarm present (2.9%)
Daytime (6am-6pm) (58.8%)
Evening (6pm-8pm) (5.3%)
Nighttime (8pm-6am) (21.2%)
Weekday (72.4%)
Weekend (20.0%)

Entry Behaviours (18 variables):

Forced entry (61.2%)
Bodily force (22.4%)
Tool used to gain entry (15.3%)
Entry unlocked (32.9%)
Used key (1.2%)
Confidence trick used to gain entry (0.0%)
Climbed above street level to gain entry (12.4%)
Broke glass (10.6%)
Access from ground (91.8%)
Access from upper level (7.1%)

Access from front (38.2%)
Access from rear (43.5%)
Access from side (11.8%)
Access from window (38.2%)
Access from door (61.2%)
Access from balcony (0.6%)
Brought tool to the scene (e.g., crowbar) (18.2%)
Used tool from scene (e.g., rock) (1.8%)

Internal Behaviours (22 variables):

Tidy search (64.1%)
Untidy search (27.1%)
No search (8.8%)
Intrusive search (53.5%)
Multiple rooms searched (55.3%)
Private rooms searched (58.8%)
Malicious damage (4.7%)
Forced interior door (5.9%)
Consumed food (1.2%)
Used facilities (e.g., toilet, sink, shower) (0.6%)
Secured premises (1.2%)
Fingerprints or other forensic evidence left (e.g., DNA, clothing, footprints etc.) (37.6%)
Forensic awareness (10.6%)
Tools used left at the scene (4.7%)
Stolen items abandoned (8.8%)
Different exit (25.9%)
Exit on foot (25.3%)
Exit by car (5.9%)
Exit on bike (1.2%)
1 offender (63.5%)
2 offenders (24.1%)
3 or more offenders (12.4%)

Property Stolen (43 variables):

Wallet (2.4%)
Fuel (0.0%)
Tobacco (4.1%)
Tools (4.1%)
Building supplies (2.4%)
Computer (5.3%)
Computer games, videogames, other computer software discs, DVDs, CDs, etc. (23.5%)
Videogame console or accessories (e.g., Wii fit board, controllers, etc.) (15.9%)
Firearm or other weapon (4.7%)
Vehicle (0.0%)

Sports equipment (3.5%)
Clothing (11.2%)
Bicycle (2.4%)
Large (hard-to-carry) electronics (e.g., flat screen television) (12.4%)
Small (easy-to-carry) electronics (e.g., laptop, iPod, cell phone, etc.) (45.9%)
Prescription or other drugs (5.3%)
Office equipment (0.0%)
Keys (1.8%)
Food (4.7%)
Alcohol (6.5%)
Watch/clock (10.6%)
Jewellery (30%)
Cash (26.5%)
Furniture (1.8%)
Books/magazines (0.0%)
Credit cards (4.7%)
Music equipment (2.9%)
Cosmetic or hygiene products (3.5%)
Bags (e.g., suitcases, gym bags, backpacks, handbags/purses, etc.) (12.9%)
Identity documents (e.g., passport, drivers license, etc.) (1.8%)
Other documents (e.g., electricity bills, etc.) (1.8%)
Antiques or art (0.0%)
Porcelain, china or silverware (2.4%)
Eyeglasses (2.4%)
Plates/bowls or other kitchen utensils (2.4%)
Other appliances (e.g., washing machine) (4.7%)
Nothing stolen (13.5%)
Stolen low value (stated or estimated value under 100 dollars) (14.1%)
Stolen somewhat valuable (between 100 and under 1000 dollars) (34.7%)
Stolen high value (over 1000 dollars) (35.9%)
One item stolen (14.1%)
Two to four items stolen (25.9%)
Five or more items stolen (44.7%)

Appendix B:
Serial Sexual Assault Behavioural Checklist

Numbers in parentheses after each behaviour correspond to that particular behaviour's frequency of occurrence in the final reduced dataset where the impact of prolific offenders was controlled using the outlier approach.

Control (32 Variables):

Approach strategy – seduction/persuasion (10.8%)
Approach strategy – money/gift (7.7%)
Approach strategy – games (8.8%)
Approach strategy – trick/false identity (38.1%)
Approach strategy – using drugs or alcohol (4.2%)
Approach strategy – threatens victim (5.4%)
Approach strategy – uses physical violence (19.2%)
Strategy to bring victim to crime site – seduction/persuasion (5.0%)
Strategy to bring victim to crime site – money/gift (7.3%)
Strategy to bring victim to crime site – games (10.4%)
Strategy to bring victim to crime site – trick/false identity (28.8%)
Strategy to bring victim to crime site – using drugs or alcohol (4.6%)
Strategy to bring victim to crime site – threatens victim (8.8%)
Strategy to bring victim to crime site – uses physical violence (28.8%)
Strategy to commit crime – money/gift (13.1%)
Strategy to commit crime – games (2.3%)
Strategy to commit crime - using drugs or alcohol (7.7%)
Strategy to commit crime – threatens victim (15.4%)
Strategy to commit crime – physical violence (41.9%)
Kidnap style attack (16.9%)
Offender attack method – raptor (33.8%)
Offender attack method – stalker (8.8%)
Offender attack method – ambusher (57.3%)
Knife used (15.0%)
Firearm used (2.3%)
Sharpened object used (6.2%)
Rope, chain or wire used (0.8%)
Vehicle used (17.3%)
Used restraints (9.6%)
Bindings used were retrieved from the scene of the crime (7.3%)
Bindings used were brought with the offender (2.3%)
Offender brought rape kit (8.1%)

Environmental (52 Variables):

Crime occurred on weekday (61.5%)
Crime occurred on weekend (23.8%)

Crime occurred over weekday/weekend (14.6%)
Crime occurred in the winter (15.0%)
Crime occurred in spring (15.4%)
Crime occurred in summer (41.5%)
Crime occurred in fall (6.9%)
Crime occurred during the day (50.0%)
Crime occurred at night (38.5%)
Crime occurred over day and night (11.5%)
Offender changed locations once (30.0%)
Offender changed location two or more times (27.7%)
All phases of crime took place in same location (42.3%)
Encounter area land use – residential (60.0%)
Encounter area land use – commercial (23.8%)
Encounter area land use – industrial (4.2%)
Encounter area land use – institutional (1.9%)
Encounter area land use – park (8.1%)
Encounter area land use – wilderness/uninhabited (1.5%)
Encounter type of location – inside (49.2%)
Encounter type of location – outside (49.6%)
Encounter type of location – in a vehicle (1.2%)
Encounter occurred at victim's residence (17.3%)
Attack area land use – residential (65.0%)
Attack area land use – commercial (17.3%)
Attack area land use – industrial (6.2%)
Attack area land use – institutional (1.5%)
Attack area land use – park (4.2%)
Attack area land use – wilderness/uninhabited (5.8%)
Attack type of location – inside (64.6%)
Attack type of location – outside (30.4%)
Attack type of location – in a vehicle (5.0%)
Attack occurred at victim residence (19.6%)
Crime area land use – residential (61.5%)
Crime area land use – commercial (15.8%)
Crime area land use – industrial (6.2%)
Crime area land use – institutional (1.5%)
Crime area land use – park (4.2%)
Crime area land use – wilderness/uninhabited (10.4%)
Crime type of location – inside (66.9%)
Crime type of location – outside (26.9%)
Crime type of location – in a vehicle (6.2%)
Crime location is victim's residence (19.2%)
Victim release area land use – residential (65.0%)
Victim release area land use – commercial (15.4%)
Victim release area land use – industrial (2.7%)
Victim release area land use – institutional (2.3%)
Victim release area land use – park (4.2%)

Victim release area land use – wilderness/uninhabited (9.6%)
Victim release type of location – inside (55.0%)
Victim release type of location – outside (44.2%)
Victim release location is victim's residence (21.2%)

Escape (5 Variables):

Used disguise during crime (11.2%)
Offender wore gloves (5.0%)
Offender tried not to leave semen (4.6%)
Offender prevented his face from being seen (13.1%)
Offender used condom (1.5%)

Sexual (17 Variables):

Vaginal intercourse with fingers (32.7%)
Sodomy with fingers (8.8%)
Vaginal intercourse with penis (34.2%)
Sodomy with penis (11.9%)
Vaginal intercourse with objects (2.7%)
Sodomy with objects (2.3%)
Caress/rubbing (13.1%)
Cunnilingus (17.3%)
Fellatio (18.5%)
Masturbation (28.1%)
Exhibitionism (20.0%)
Offender ejaculated at the scene (61.2%)
Sexual dysfunction of offender (8.8%)
Offender forced victim to engage in fellatio (39.2%)
Offender forced victim to masturbate (33.5%)
Offender forced victim to engage in sodomy (3.5%)
Offender forced victim to engage in self-touching (1.5%)

Style (23 Variables):

Speech - Did not talk to victim/no speech (17.3%)
Speech - Sensitive to victim (8.1%)
Speech - Inquisitive (4.6%)
Speech - Complimentary to victim (6.9%)
Speech - Hostility towards women (6.9%)
Speech - Hostility in general (2.3%)
Speech - Reassuring (6.2%)
Speech - Making demands of victim (11.9%)
Non-sexual body parts mutilated (1.2%)
Sexual body parts mutilated (2.7%)
More force than necessary used to commit the crime (Avery-Clark & Laws) (20.0%)

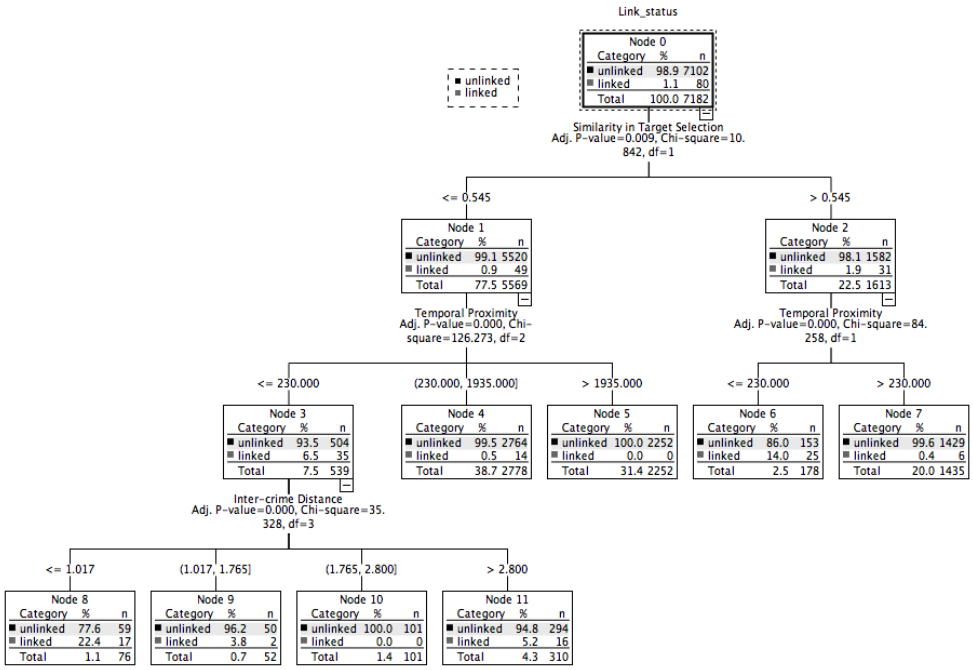
Offender did not undress (26.9%)
Offender put down pants and underwear (48.1%)
Offender took off pants and underwear (3.5%)
Offender undressed completely (21.5%)
Victim not undressed (10.4%)
Victim's clothes pushed up or pulled down (30.0%)
Victim's clothes torn or cut for access (4.2%)
Partial removal of victim's clothing (13.8%)
Complete or near complete removal of victim's clothing (27.7%)
Redressed victim (13.8%)
Offender stole something from victim (9.6%)
Offender broke into victim's home (10.0%)

Victim Selection (19 Variables):

Victim male (21.9%)
Victim female (78.1%)
Victim child (0-12) (30.8%)
Victim adolescent (13-17) (28.1%)
Victim adult (18 and over) (41.2%)
Victim under the influence of drugs/alcohol (18.5%)
Victim dressed in a provoking fashion (15.0%)
Victim at home before crime occurred (20.8%)
Victim at work before crime occurred (8.5%)
Victim commuting before crime occurred (30.4%)
Victim walking/jogging/hitchhiking/engaging in other recreational activity (23.1%)
Victim at a bar/nightclub (3.8%)
Victim visiting friend before crime occurred (10.8%)
Victim engaging in prostitution before crime occurred (1.5%)
Victim is alone when offender first enters contact (56.2%)
Offender hunting style – hunter (45.0%)
Offender hunting style – poacher (7.3%)
Offender hunting style – troller (23.1%)
Offender hunting style – trapper (24.6%)

Appendix C
Classification Trees Produced for the Multiple Models Approach (Break & Enters)

Iteration 1: Development Sample



Iteration 1: Test Sample

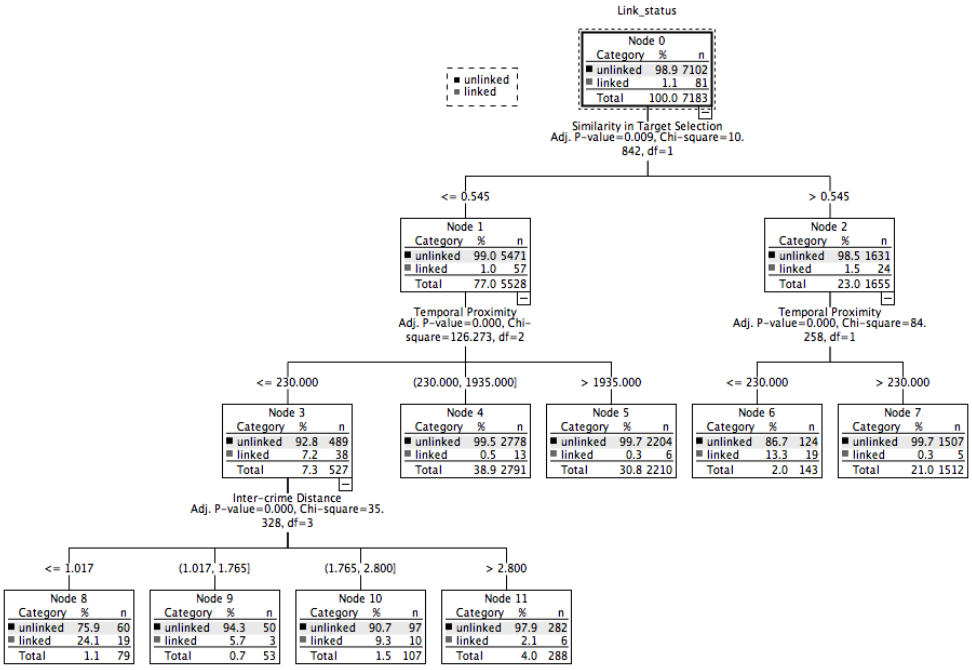
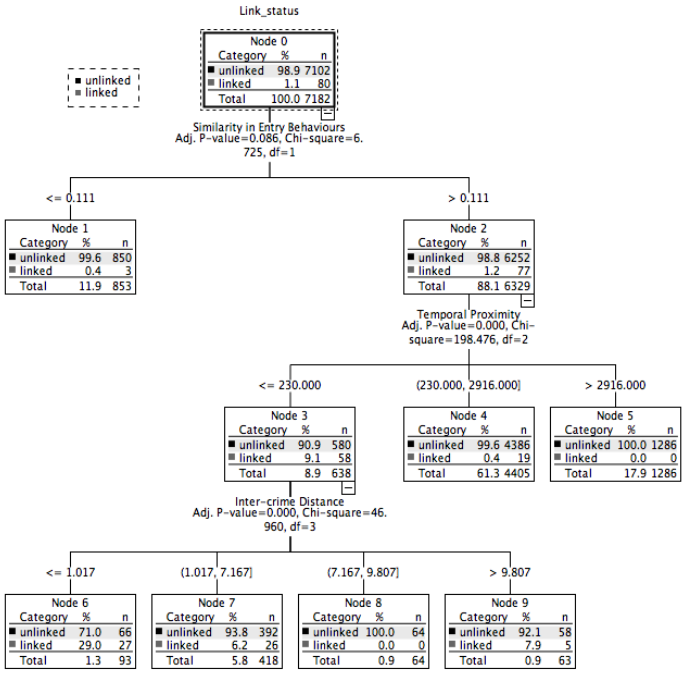


Figure C1. Iteration 1 of the target selection forced tree for the break and enter data.

Iteration 1: Development Sample



Iteration 1: Test Sample

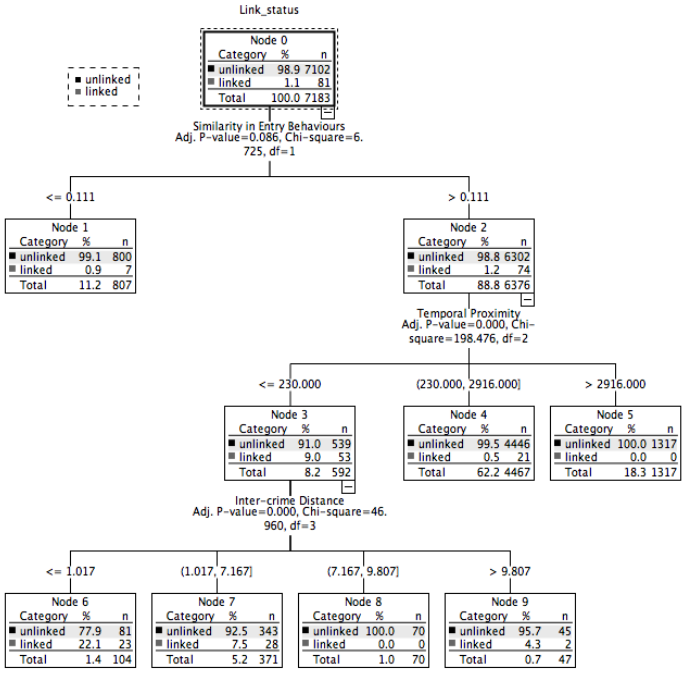
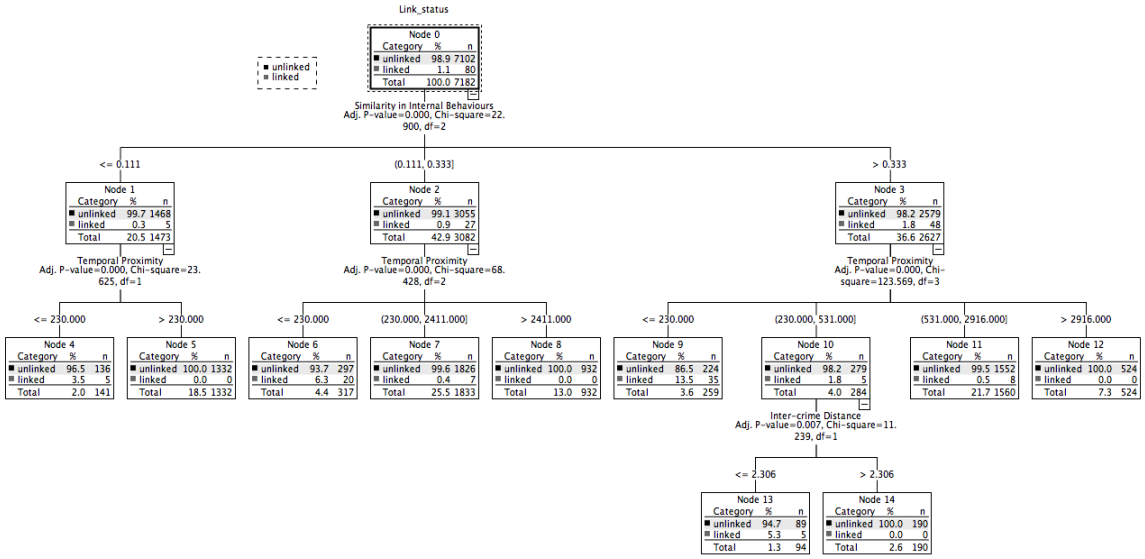


Figure C2. Iteration 1 of the entry behaviours forced tree for the break and enter data.

Iteration 1: Development Sample



Iteration 1: Test Sample

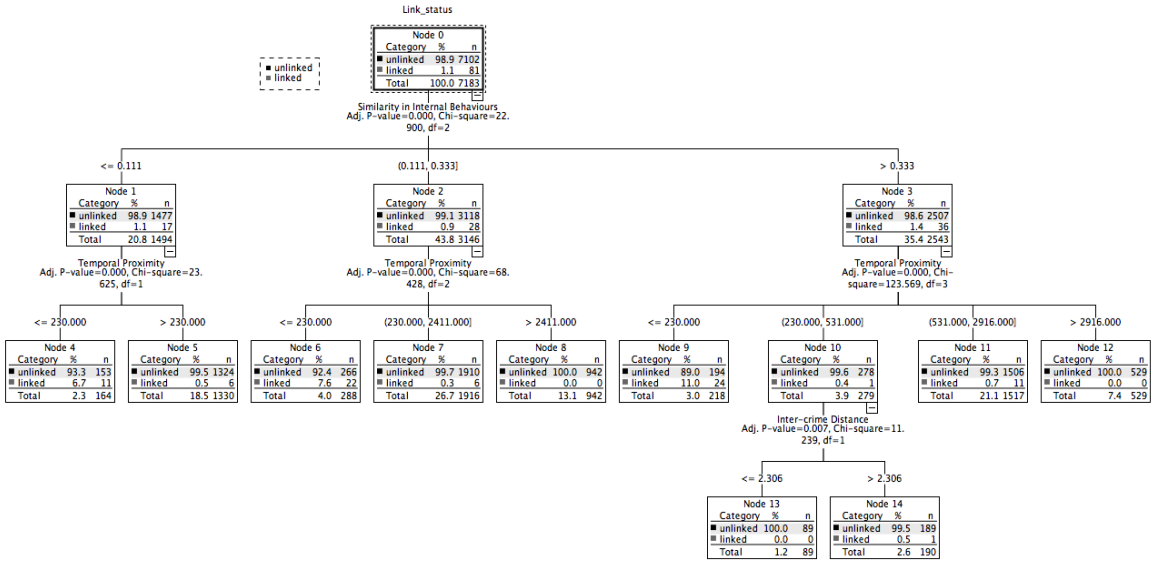
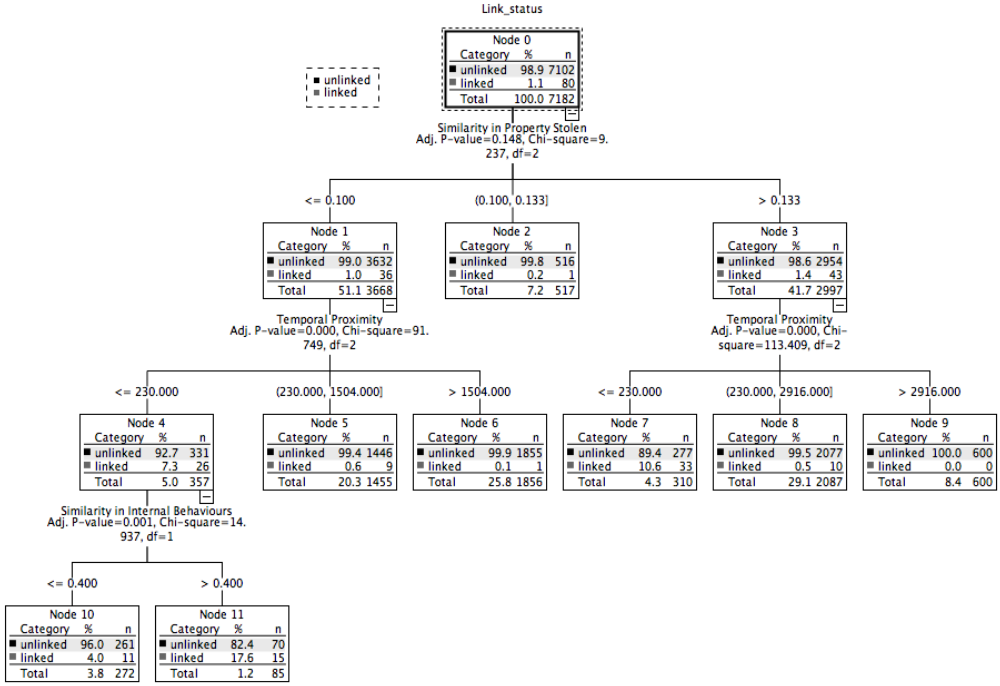


Figure C3. Iteration 1 of the internal behaviours forced tree for the break and enter data.

Iteration 1: Development Sample



Iteration 1: Test Sample

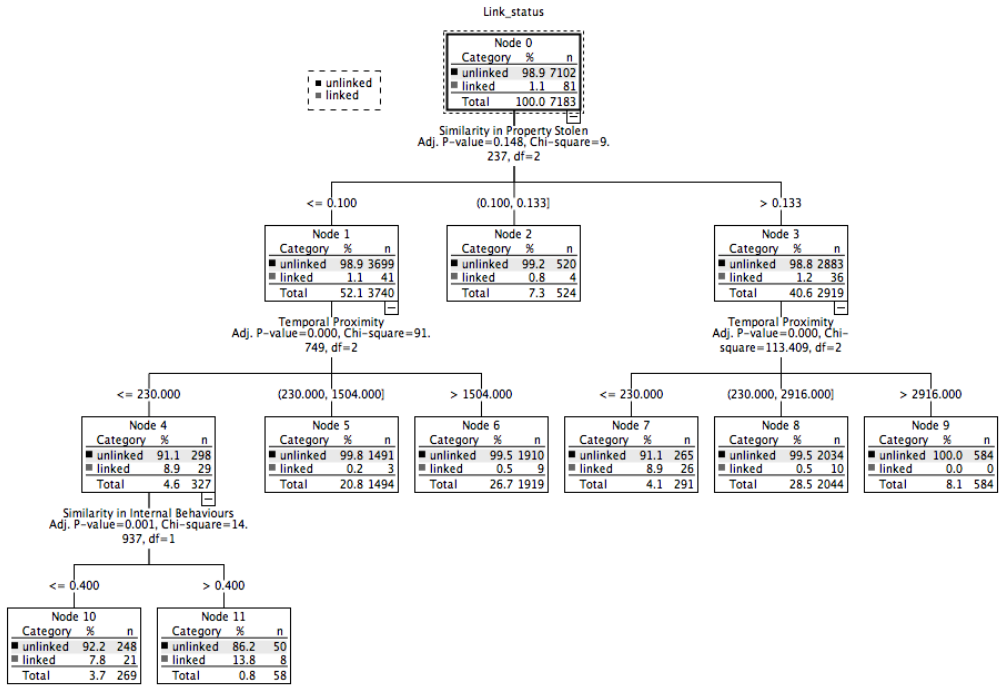
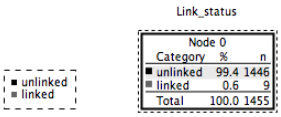


Figure C4. Iteration 1 of the property stolen forced tree for the break and enter data.

Iteration 2: Development Sample



Iteration 2: Test Sample

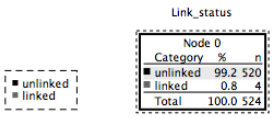
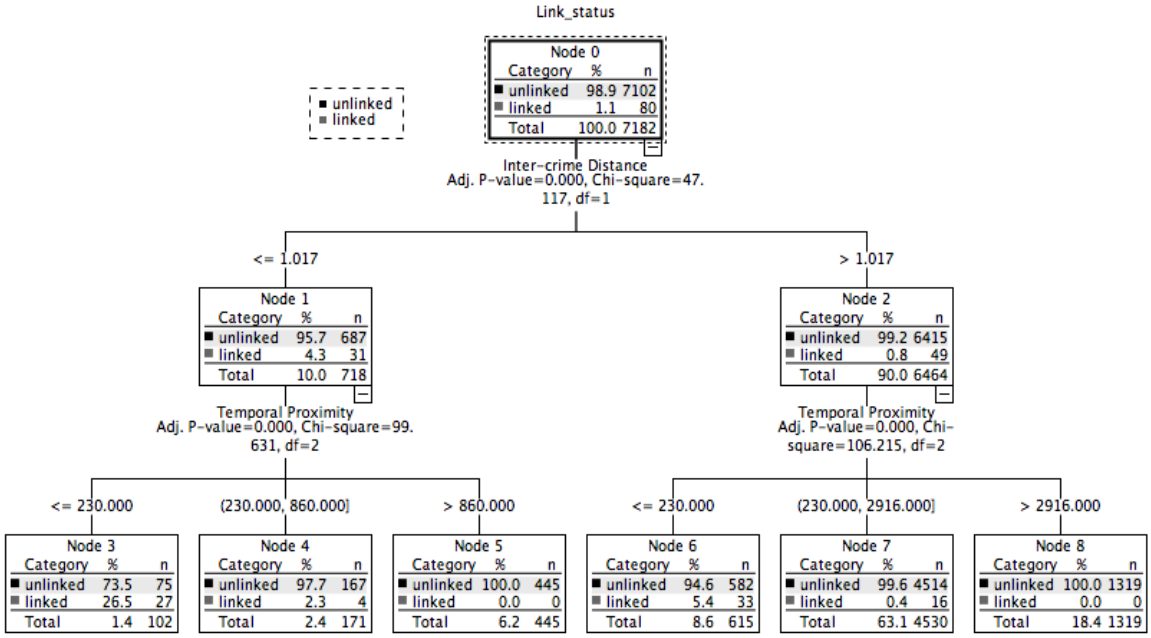


Figure C5. Iteration 2 of the property stolen forced tree for the break and enter data.

Iteration 1: Development Sample



Iteration 1: Test Sample

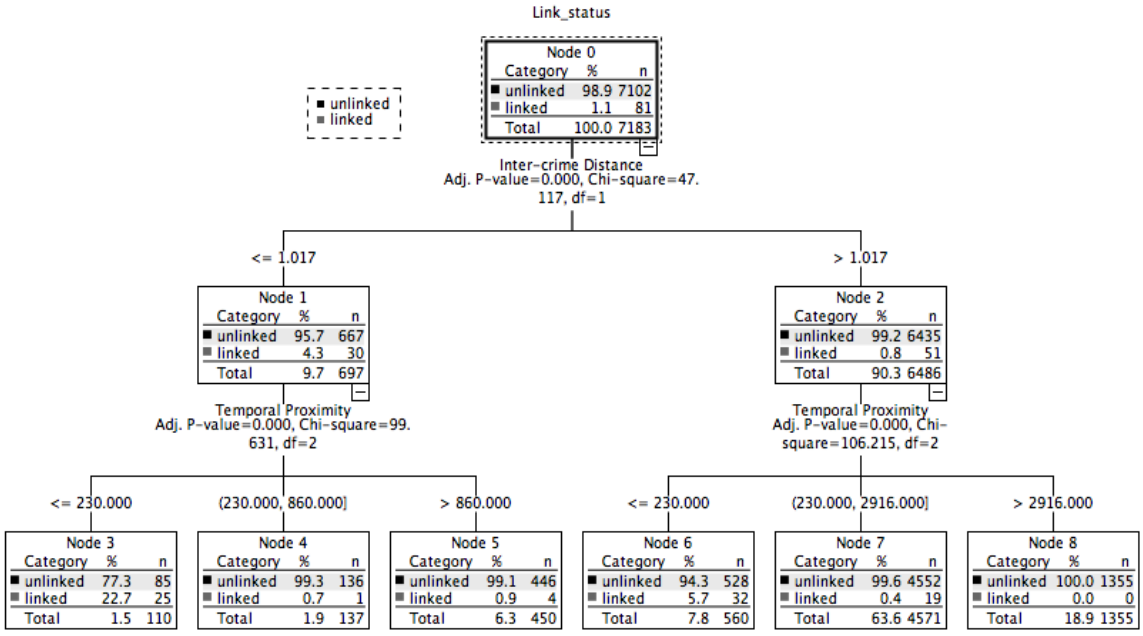


Figure C6. Iteration 1 of the inter-crime distance forced tree for the break and enter data.

Appendix D
Classification Trees Produced for the Multiple Models Approach (Sexual Assaults)

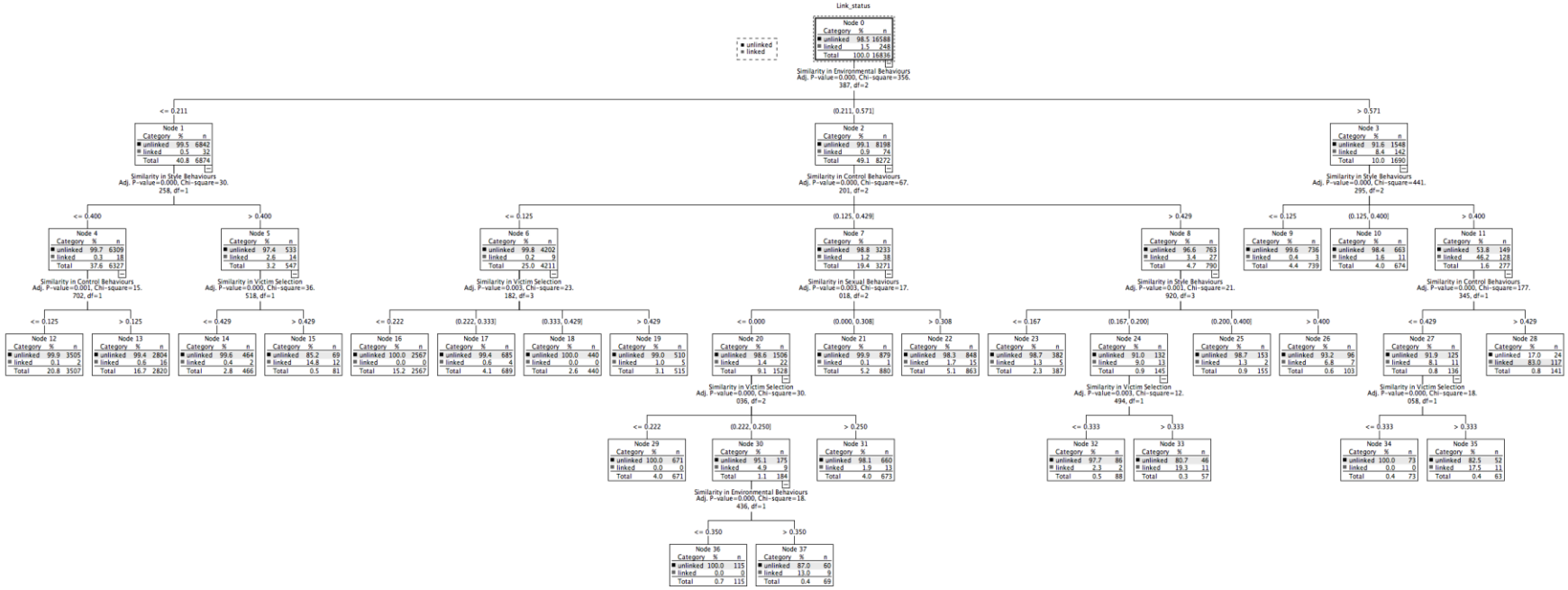
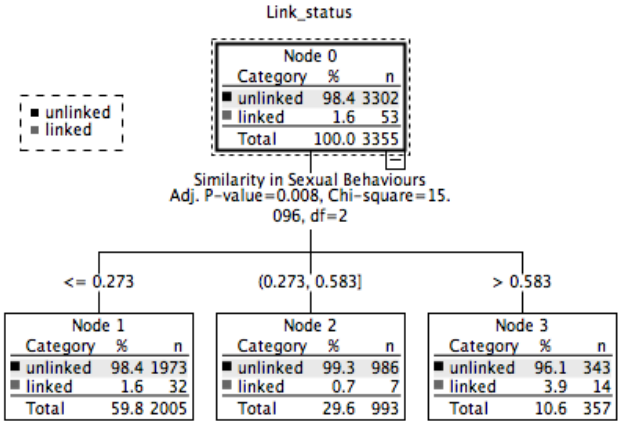
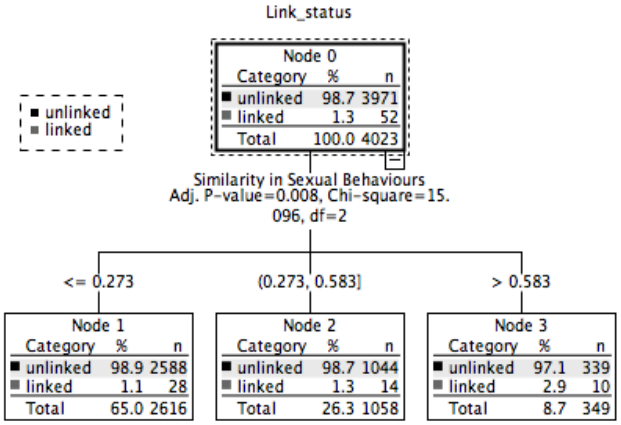


Figure D1. Iteration 1 of the environmental behaviours forced tree for the sexual assault data (development sample).

Iteration 2: Development Sample



Iteration 2: Test Sample



Iteration 3: Development Sample

Iteration 3: Test Sample

Figure D2. Iteration 2 and Iteration 3 of the environmental behaviours forced tree for the sexual assault data.

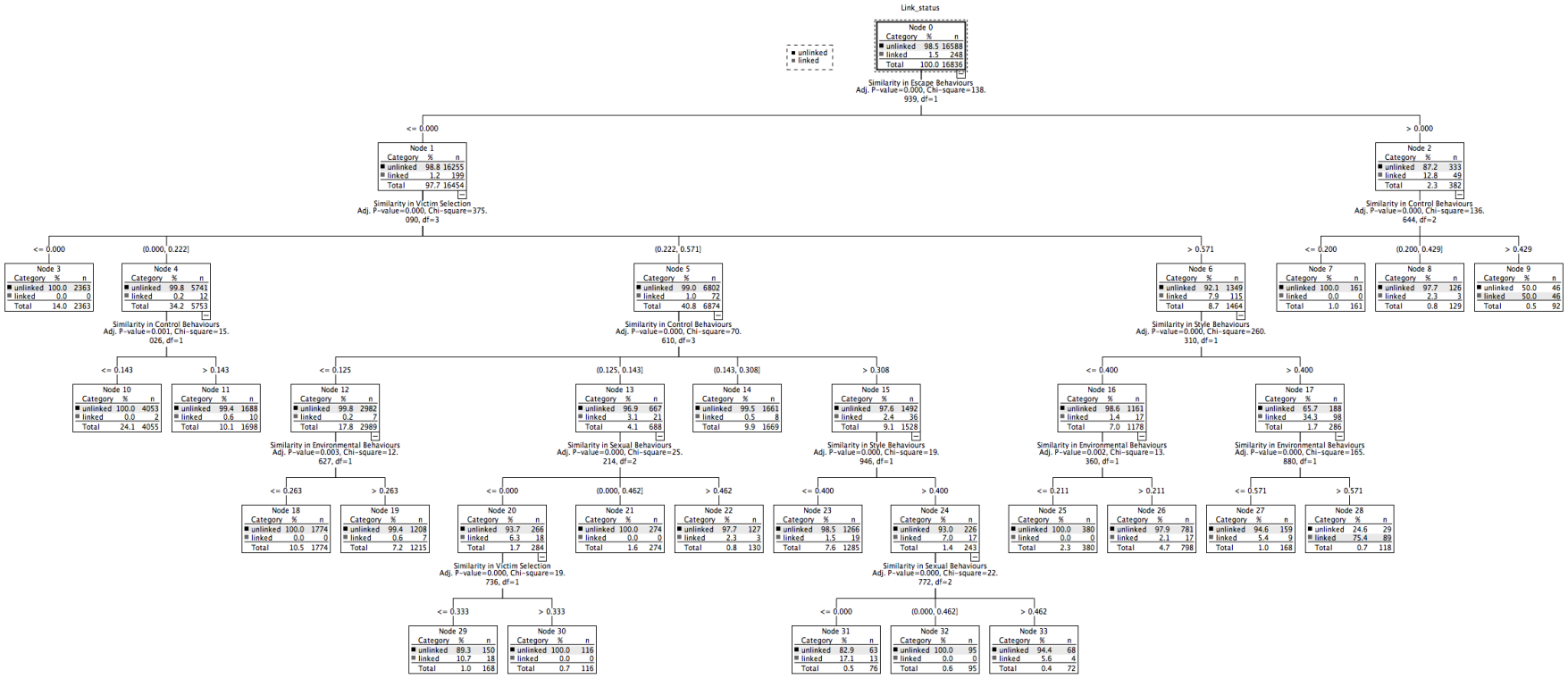
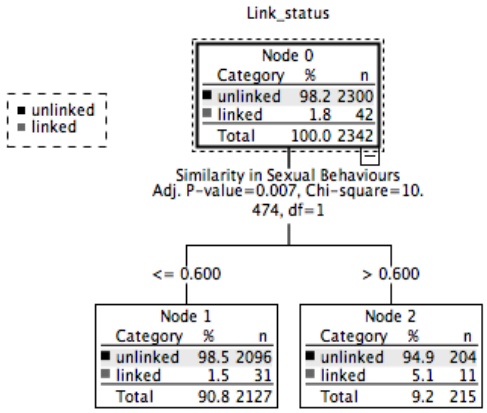
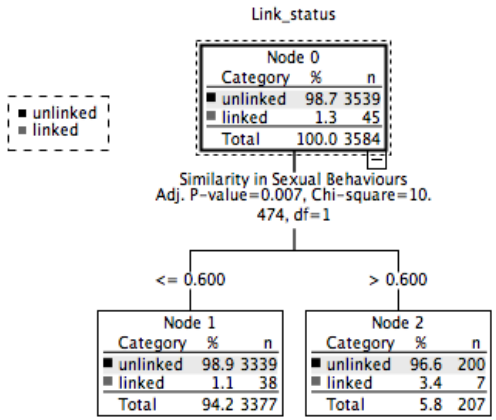


Figure D3. Iteration 1 of the escape behaviours forced tree for the sexual assault data (development sample).

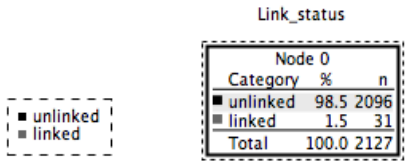
Iteration 2: Development Sample



Iteration 2: Test Sample



Iteration 3: Development Sample



Iteration 3: Test Sample

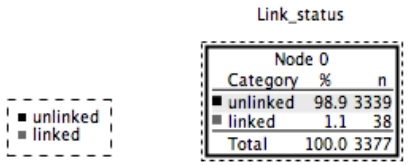


Figure D4. Iteration 2 and Iteration 3 of the escape behaviours forced tree for the sexual assault data.

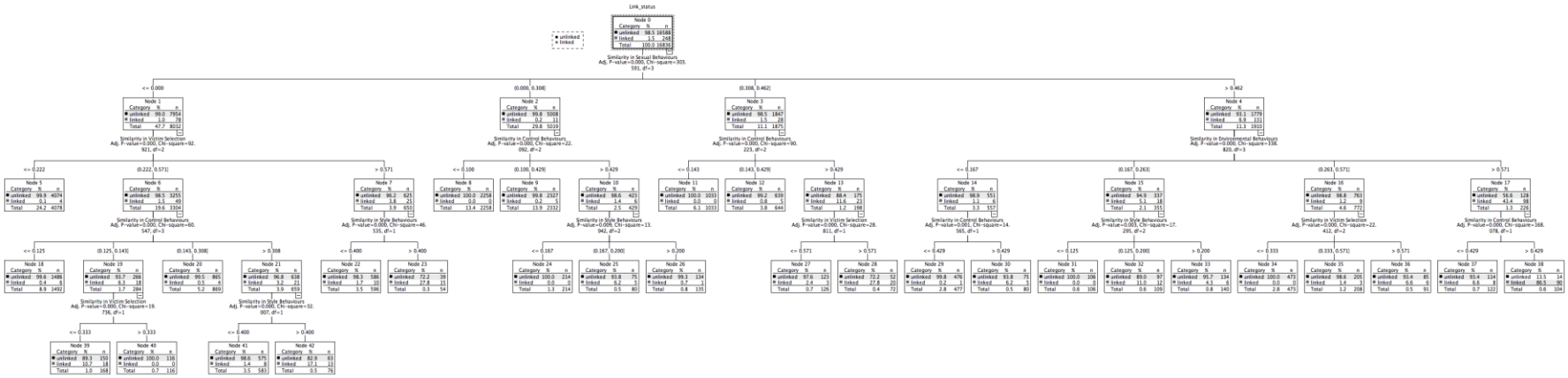
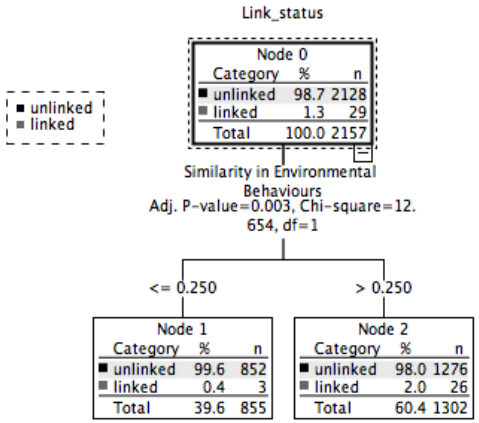
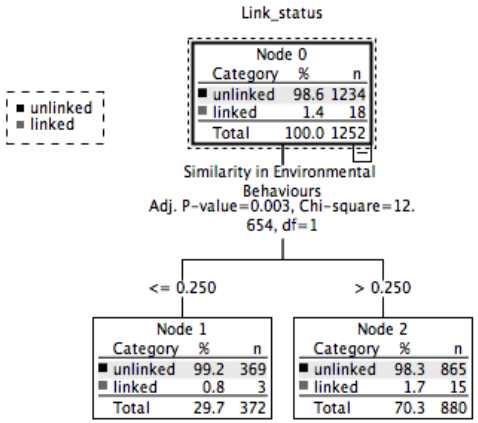


Figure D5. Iteration 1 of the sexual behaviours forced tree for the sexual assault data (development sample).

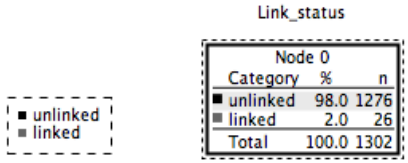
Iteration 2: Development Sample



Iteration 2: Test Sample



Iteration 3: Development Sample



Iteration 3: Test Sample

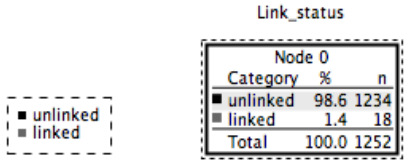
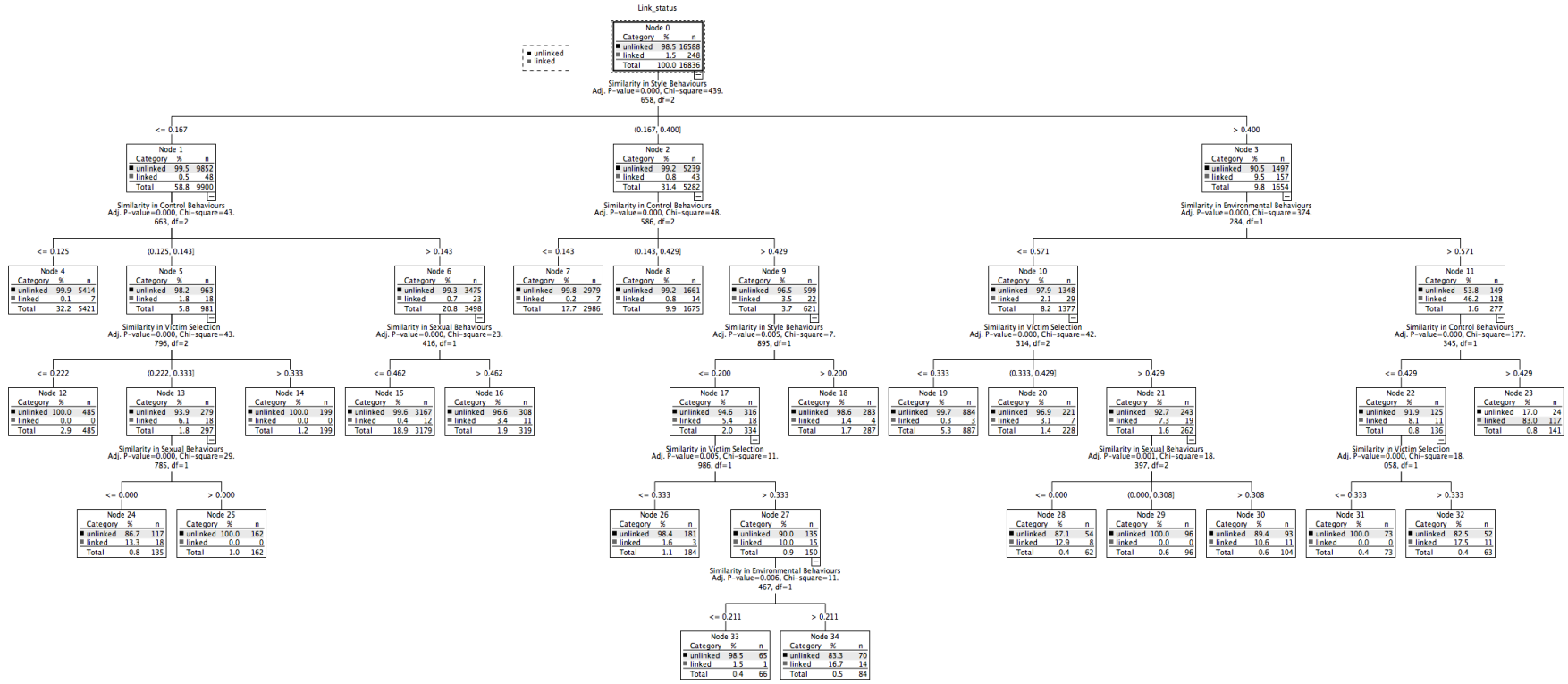


Figure D6. Iteration 2 and Iteration 3 of the sexual behaviours forced tree for the sexual assault data.



Iteration 2: Development Sample

Iteration 2: Test Sample

Figure D7. Iteration 1 (development sample only) and Iteration 3 of the style behaviours forced tree for the sexual assault data.

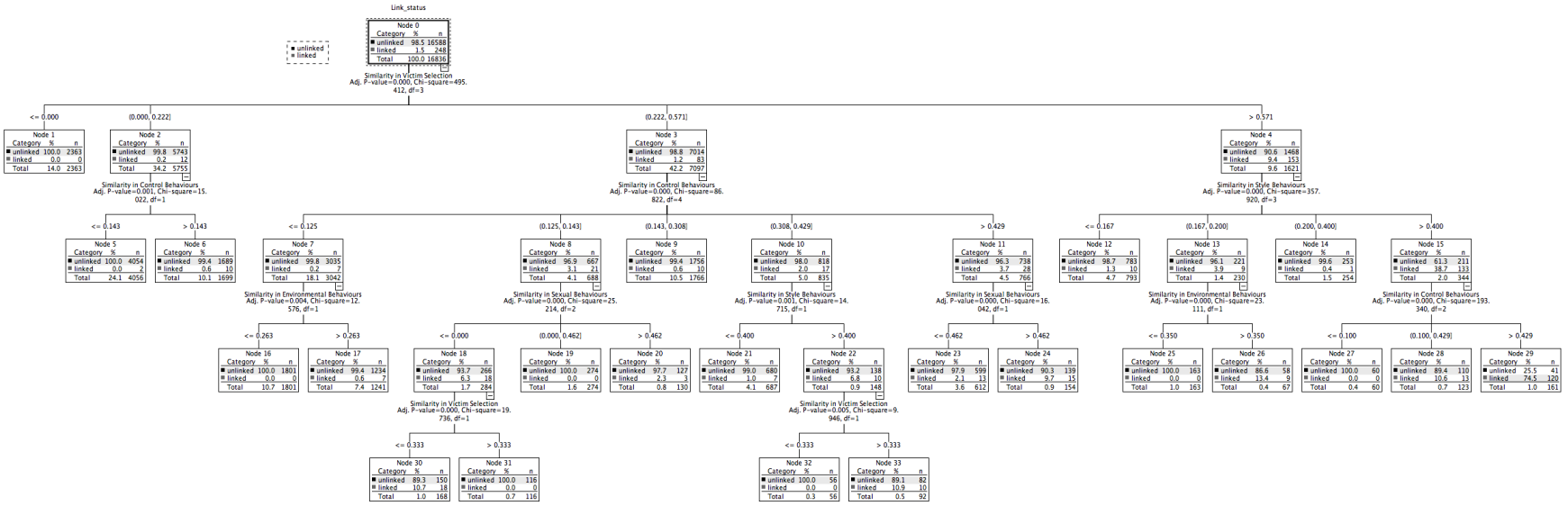
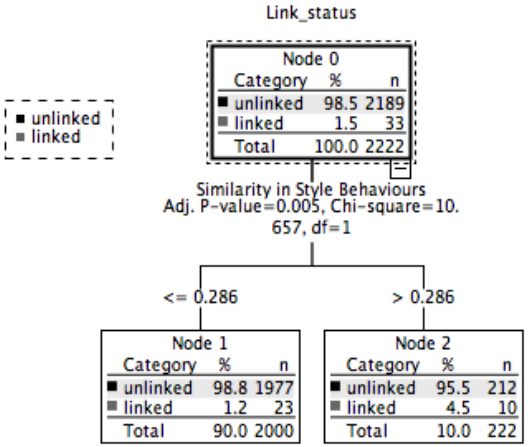
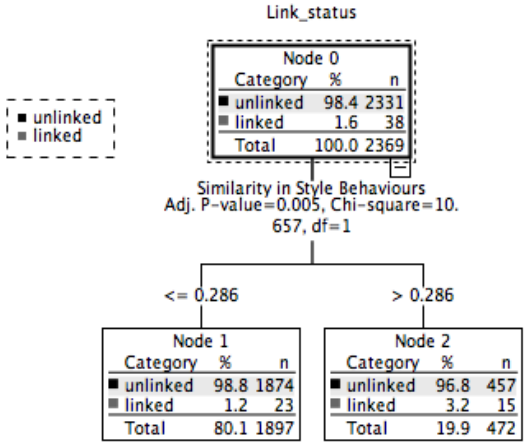


Figure D8. Iteration 1 of the victim selection forced tree for the sexual assault data (development sample).

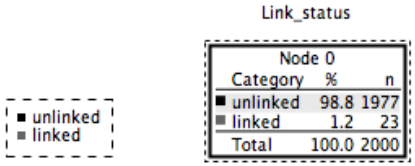
Iteration 2: Development Sample



Iteration 2: Test Sample



Iteration 3: Development Sample



Iteration 3: Test Sample

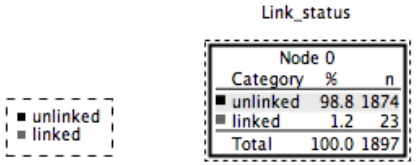


Figure D9. Iteration 2 and Iteration 3 of the victim selection forced tree for the sexual assault data.