Statistical Criminal Profiling

Predicting Homicide Offender Characteristics using a Bayesian Network

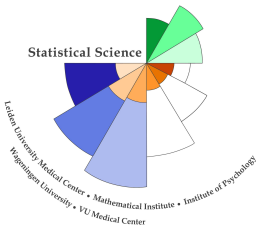
Lotte Pas (s1042378)

Thesis advisors: Prof. Dr. P.J.F. Lucas and Dr. B.J.A. Mertens

MASTER THESIS

Defended on July 25, 2018

Specialization: Statistical Science



STATISTICAL SCIENCE
FOR THE LIFE AND BEHAVIOURAL SCIENCES

Index

Abstract

Criminal profiling is a rapidly growing field of research, in which statistics get more and more incorporated alongside of the traditional behavioural profiling approach that uses psychological theories to predict the behaviour of an offender. A model was built to predict the offender characteristics from crime and victim characteristics for single-victim-single-offender homicides in the Netherlands. Using the Dutch Homicide Monitor, eight different Bayesian network structure learning algorithms were combined into one model; arcs that were present in at least three separate structure learning algorithms were represented in the combined model and its direction was determined by the highest cumulative arc strength. The graphical representation of the model gives insight into the dependence relationships between crime, victim, and offender characteristics, and therefore could be used to confirm existing and develop new hypotheses on criminal psychology. Moreover, with an appropriate threshold resulting in a prediction error of less than 10 percent, the combined Bayesian network might be suitable for actual implementation by the police. This practical implication and the restrictions of the model are discussed, and recommendations for future research are given.

1. Introduction

Criminal profiling traditionally depends on expert knowledge and experience. The FBI Behavioural Analysis Units (BAU) are well known for their expertise in this field. The job of a behavioural analyst, often called criminal profiler, is to understand the behaviour of "individuals who threaten national security or public safety" ("FBI Behavioral Analysis Jobs", 2017). Criminal profilers are often portrayed in the media – for example in popular television programs such as *Criminal Minds* – as special agents who solve the most complex cases within a few days. White, Lester, Gentile, and Rosenbleeth (2011) described this incorrect public perception of criminal profiling and the problem of having a jury with this wrongful perception. In real life it is not possible to solve cases in the same way and within the same time they are solved in novels, television programs, and films. Most special agents do not walk around with a gun, but sit behind their desks to coordinate investigative and operational teams, conduct criminological research, and assist the federal, state, local, and foreign law enforcement agencies investigating the most serious violent crimes ("FBI Behavioral Analysis Jobs", 2017). Their work is not only based on personal experience and knowledge, but more and more includes data-driven analysis techniques due to the development of computer and information technologies (Baumgartner, Ferrari, & Palermo, 2008).

The goal of this research was to develop an analysis tool that might be helpful during the criminal investigation process, by building a Bayesian network that predicts certain characteristics of homicide offenders. The importance of this research lies in the potential practical implications of the model, as well as in the rather unexplored application of Bayesian network learning in the field of criminal profiling. The research question read: "To what extent are homicide offender characteristics predictable using a Bayesian network learned from data about solved single-victim-single-offender homicides in the Netherlands from 1992-2016?" The aim of the research was twofold. On the one hand, focus was on the statistical methods to build the most appropriate Bayesian network for this criminal profiling purpose, and on the other hand, attention was given to the interpretation of the graphical structure of the resulting network and its practical implications. A Bayesian network approach was chosen because it is a user friendly, multivariate statistical model that gives a straightforward insight into the relationships between the variables, both qualitative and quantitative.

This thesis has the following structure. First, more information will be provided about criminal profiling (Section 2) and Bayesian network learning (Section 3) respectively. Then, the data and analysis techniques used in this research will be explained (Section 4) before showing the resulting network and (Section 5) its potential practical implications during criminal investigations (Sections 6 and 7). Lastly, conclusions will be drawn (Section 8) and discussed (Section 9).

2. Criminal Profiling

Criminal profiling was first used in the Mad Bomber case in New York almost a century ago (Aydin & Dirilen-Gumus, 2011). James Brussel, a psychiatrist, constructed a detailed offender profile in which

he even predicted the clothing of the offender correct. The foundations of criminal profiling were laid by psychiatrists and psychologists, who often interviewed convicted criminals in an attempt to understand their behaviour. The FBI started a working group about criminal profiling, that developed the first theories and classifications based on interviews with 36 convicted homicide offenders who were all together responsible for 118 victims (Burgess, Hartman, Ressler, Douglas, & McCormack, 1986; Ressler, Burgess, & Douglas, 1988; Ressler, Burgess, Douglas, Hartman, & D'Agostino, 1986; Ressler, Burgess, Hartman, Douglas, & McCormack, 1986). The widely known FBI homicide offender classification of organised and disorganised offenders by Ressler, Burgess, and Douglas (1988) was the first theory made public. However, as Beauregard and Proulx (2002) mentioned, neither clinical nor statistical analysis methods that were used by those authors were explained, and Chifflet (2014) described criminal profiling theories in general as "uncertain at best" (p. 238). Thus, although the FBI's offender typologies are often used in homicide studies, the usefulness of criminal profiling is frequently criticised.

2.1 Related research on criminal profiling

Strano (2004) discussed some critiques against the psycho-investigative technique of criminal profiling. Above all, the invisible line between instinct and intuition on the one side, and scientific procedures on the other, seems to be the problem. The scientific reliability and validation of the different criminal profiling techniques remains tenuous (Strano, 2004). Today, however, a criminal profiler will not be accepted as an expert witness in court if he or she cannot prove that the methodology used to profile the offender was reliable (Bosco, Zappalà, & Santtila, 2010). Criminal profiling is nowadays only accepted when it is scientifically accountable; criminal profiling is science, not an art, according to Bosco et al. (2010).

Another critique against criminal profiling is that the actual accuracy of criminal profiling is still scarcely investigated (Chifflet, 2015). Wilson and Soothill (1996) already pointed out that it is difficult to find a good measurement criterion to validate criminal profiling, something that Chifflet (2015) confirmed more recently. Is measuring the accuracy enough? Or should utility and investigative relevance of the profile also be examined? And should the skills of the profiler be considered? The ambiguity around the validation of criminal profiling leads to insecurity about criminal profiling being a sound discipline that could be used safely (Chifflet, 2015). Kocsis stated in this context: "possibly the greatest mystery surrounding criminal profiling has been its growth despite an absence of robust scientific evidence to validate it" (Kocsis, 2006, p. 458).

Not merely the concept of criminal profiling, but also the proficiency of criminal profilers is criticised. Kocsis (2003) claimed that investigative experience is not the key skill needed for criminal profiling; his sample of students outperformed the selected group of psychologists (yet the criminal profilers still performed better than the students), so Kocsis concluded that the capacity for logical reasoning is the most important competence for a criminal profiler. However, Kocsis's data and

analysis methods were questioned by Bennell, Jones, Taylor, and Snook (2006), who suggested that more research in this field is necessary before conclusions can be drawn on the expertise of professional criminal profilers. Although the proficiency of criminal profilers is still a subject of discussion in the current literature, the field has become more professionalised over the last decade (Turvey & Esparza, 2016).

Despite the critiques against criminal profiling, most authors do realise that the fact that criminal profiling has its limitations does not mean that criminal profiling could not be useful *besides* the standard police investigation. Wilson and Soothill (1996) even suggest a symbiotic relationship between criminal profiling and thorough and well-planned investigation; you cannot have the one without the other. Classical police investigation searches for an offender that has some sort of relationship with the victim, since most homicides are committed by someone within the social environment of the victim. Criminal profiling might be useful when no such relationship is found, or when there is no rational cause and no crime scene evidence (Ainsworth, 2001; Aydin & Dirilen-Gumus, 2011). Sturup, Karlberg, and Kristiansson (2015) found that cases in which there was no eyewitness, the victim was not intoxicated with alcohol, the victim had a criminal record in the past five years, or the victim was killed with a firearm, were the least likely to be solved and needed extra, intensive and lasting resources. Criminal profiling could be one of these extra resources when it is not seen as a technique to obtain evidence but as a decision-aid tool during the investigation process.

Unfortunately, because forensic science has focused mainly on generating evidence that is admissible in court, less attention has been spent on the possible contribution in earlier stages of the forensic process (Morelato et al., 2013). Morelato et al. (2013) state that although the concept of intelligence-led policing is widely known nowadays, not much is known about the extensiveness and effectiveness of its practical application. In other words, perhaps the discussion should not be about the admissibility of an offender profile as evidence in court, but about the usefulness of criminal profiling during the criminal investigation.

2.2 Statistical criminal profiling

Based on their systematic review of the literature about criminal profiling, Dowden, Bennell, and Bloomfield (2007) concluded that "the statistical sophistication of these studies is sorely lacking, with most including no statistics or formal analyses of data" (p. 44). Years later, also Briggs (2015) concluded from her literature review that more research into computerised criminal profiling is needed. As described before, an offender profile is traditionally constructed from psychological and classification theories based on in-depth interviews with convicted homicide offenders. The classification theories are often supported by a cluster analysis or multidimensional scaling (for examples, see Adeyiga & Bello, 2016; Beauregard & Proulx, 2002; Salfati & Canter, 1999; Salfati & Park, 2007). Although nowadays more advanced statistical techniques are available to construct an offender profile, criminal profiling traditionally does not incorporate much statistics. This is

unfortunate, especially since statistical modelling works better in terms of prediction accuracy than a frequency approach – such as cluster analysis – in which the percentage of offenders with certain characteristics is extracted from a set of similar offences (Aitken et al., 1996; Francis et al., 2004; Ter Beek, Van den Eshof, & Mali, 2010).

Daéid (1997) investigated the differences in approach to criminal profiling between the United States and the United Kingdom. The FBI's National Centre for the Analysis of Violent Crime (NCAVC) records behavioural traits of convicted offenders and uses this database combined with other forensic evidence to predict the type of offender that fits the committed crimes. The FBI probably possesses the majority of the actual knowledge about criminal profiling, but this knowledge often "remains in the cultural baggage passed down over the years from one profiler to another" (Strano, 2004, p. 496). According to Daéid (1997), statistical analyses of crime, victim, and offender characteristics were incorporated earlier in the developmental stages of criminal profiling in the United Kingdom. Their approach was retrospective; known offender characteristics and information from solved cases were the starting point. When not much information was available, non-parametric statistics such as the chi-square test for association between various variables, were used. Statistics were used when more in-depth information was available; logistical regression and Bayesian belief networks turned out to be useful in a criminal profiling context (Daéid, 1997). More recently, Aitken (2006) came to the same conclusion: logistic regression and Bayesian networks are two statistical techniques that could help police investigators to identify suspects. He emphasised however, that the predictions of homicide offender characteristics made by those techniques do not provide any evidence. The probabilistic calculations are meant to give a direction during the police investigation, and they should not be used as evidence in court according to Aitken. However, the opinion of most statisticians is that Bayesian networks – if applied correctly – could be used in court to represent decision problems (Fenton & Neil, 2013; Taroni, Biedermann, Bozza, Garbolino, & Aitken, 2010). Comparable is the use of statistics in the medical field, where decisions of life or death are made based on statistical analyses.

Logistic regression has proven it usefulness in criminal profiling, and is not merely applicable in analyses of lethal crimes. Davies, Wittebrood, and Jackson (1997) found that with a logistic regression the criminal antecedents (burglary, violent offences, and rape recidivism) of a stranger rapist could be obtained from his behaviour during the offence. Likewise, Ter Beek et al. (2010) used a logistic regression model to predict the probabilities of stranger rape offender characteristics based on different victim and crime characteristics.

Bayesian networks however, are less often used for criminal profiling purposes compared to logistic regression. Although the field of Bayesian networks is relatively new, it is still surprising that not more research is done to the possibilities of Bayesian networks in criminal profiling. Mears and Bacon (2009) conclude that much of the decision making in the criminal justice system happens within a "black box" (p. 152). Analyses such as Bayesian networks could help give an expression to

why the decisions are made during the investigation process. Besides, not only are Bayesian networks fairly user friendly, they also give a clear and intuitive graphical representation of the data. Daéid (1997) expressed this perspicuously:

Bayesian belief networks allow a more graphical representation of the relationship between victim/scene characteristics with those of a known offender giving greater visual clarity than logistical regression. Associations identified between victim/scene characteristics and those of known offenders can be assigned probabilities either from a generated database (objectively) or by experience or expert opinion (subjectively). Using Bayes's Theorem it is then possible to calculate the probabilities of an offender having a particular characteristic given victim characteristics. (p. 30).

Yet there are a few authors that built and used Bayesian networks for criminal profiling purposes. Aitken et al. (1996) constructed a seven node network based on expert knowledge. The structure as well as the prior probabilities were determined in discussion with a senior detective. According to those authors, the main advantages of Bayesian networks for criminal profiling are: its clear graphical representation of the relationships between variables, its potential use in operational conditions, and its transparency in combining subjectivity and objectivity in statistical analyses – which is not often the case with other statistical techniques.

Baumgartner et al. (2008) used data instead of expert knowledge to construct a Bayesian network; they applied an adapted K2 learning algorithm to a police database on cleared single-victim-single-offender homicides, and predicted 21 offender variables from 36 evidence variables. Their Bayesian network outperformed a team of expert criminologists in predicting the offender variables; the Bayesian network predicted 86 percent of the offender characteristics correctly, versus 53 percent for the experts. Although the K2 algorithm (Cooper & Herskovits, 1992) is one of the oldest score-based algorithms, and many new algorithms were found to perform better since, the Bayesian network built by Baumgartner et al. already appeared to be useful in predicting homicide offender characteristics. Moreover, as the authors argue, their Bayesian network could be used to develop hypotheses on criminal psychology, since it reveals the most significant relationships among the variables.

The most recent study to Bayesian networks used in criminal profiling is conducted by Stahlschmidt, Tausendteufel, and Härdle (2013). Based on 252 cases of sex-related homicides in Germany, they constructed a Bayesian network with 53 nodes (variables) and 83 edges (relationships between variables) after combining different structure learning algorithms. Their final Bayesian network was undirected, because the eight structure learning algorithms did not agree on the directions of all arcs. Therefore, they claimed that they were not able to learn the parameters for this final network from their data, since a Bayesian network's parameters can only be estimated from data if the structure is completely directed (Scutari & Ness, 2018). However, their combined undirected graph represents different dependence relationships than the graphs following from the separate structure learning algorithms. For example, a directed structure A → B ← C implies independence between A

and C, whereas the same structure but then undirected, A – B – C, implies a dependency between A and C. For more information on the dependence relationships of a Bayesian network, see Section 4.2. For the moment, from Stahlschmidt et al. could be concluded that more research is needed to discover the real dependence relationships. Nevertheless, the authors showed that not only a model averaging approach, but also all of their Bayesian network structure learning algorithms separately outperformed logistic regression in terms of prediction performance. They concluded that the error rate of less than 10 percent is a promising start for a real-life implementation of their model by the police. Moreover, they suggested that an online learning procedure could be developed, such that when new cases become available, the model will be adjusted automatically.

Sometimes an even more advanced data-driven approach is used for criminal profiling. Since law enforcement agencies nowadays gather a tremendous amount of data, knowledge management (i.e. "simplifying and improving the process of sharing, distributing, creating, capturing and understanding knowledge", Gottschalk, 2006, p. 381) becomes more necessary than ever. It is even argued that the most important resource during a criminal investigation is knowledge, and that the success of an investigation depends on the available knowledge (Gottschalk, 2006). Advanced statistical techniques could help manage the knowledge that is available to police investigators. For example, Brahan, Lam, Chan, and Lcung (1998) set up a knowledge management system called AICAMS (artificial intelligence crime analysis and management system) that provided investigations into machine-learning and neural network techniques to assist during a criminal investigation. A few years later, Strano (2004) described the project he started in Italy, the Neural Network for Psychological Criminal Profiling (NNPCP), in which he uses artificial intelligence (more specifically, a neural network and data mining) for criminal profiling in homicide cases and other serious crimes. The neural network produces a psychological, psychopathological, and motivational profile of the offender, based on crime scene analyses, victimology, and forensic, autopsy, and police reports.

3. Bayesian networks

Aitken et al. (1996) gave a good definition of a Bayesian network in a criminal profiling context:

A Bayesian belief network is a graphical representation of the relationships amongst the various characteristics, offender, victim and crime. In this context, a graph is a set of nodes and directed arcs. Each node represents a particular characteristic; two nodes are linked by an arc whose direction represents a causal or influential relationship. The absence of an arc between two nodes implies that the two characteristics associated with these nodes are conditionally independent of each other, that is, they are independent conditional on knowledge of the values of the other characteristics. There is also a restriction that the directed arcs cannot form a closed loop, so that it cannot be possible to start from a particular node and follow arcs to return to that node. (p. 250).

The construction of a Bayesian network for criminal profiling purposes consists of three phases (Aitken, 2006). In the first phase the victim, crime, and offender characteristics (nodes) that

could be incorporated in the model should be specified. During the second phase the structure of the Bayesian network is learned either from data by a structure learning algorithm, or by the knowledge of an expert in the field. In the third and last phase the prior (conditional) probabilities of each node are determined, again either data-driven or estimated by an expert.

In this research, the victim, crime, and offender characteristics were obtained from the available dataset (Section 4). Next, both during the second and third phase (Section 5), learning algorithms were used to extract the structure and parameters of the Bayesian network from data. But first, Section 3.1 gives a more formal definition of a Bayesian network, and Section 3.2 discusses its structural properties. Sections 3.2 and 3.3 respectively describe the process of structure learning and parameter learning.

3.1 Definition

Before applying a Bayesian network learning algorithm, the definition and the structure of a Bayesian network should be understood. A Bayesian network – or belief network or Bayes net – consists of a graphical structure and its complementary parameters (Scutari, 2010). Figure 1 shows an example of a Bayesian network structure for the variables used in this research.

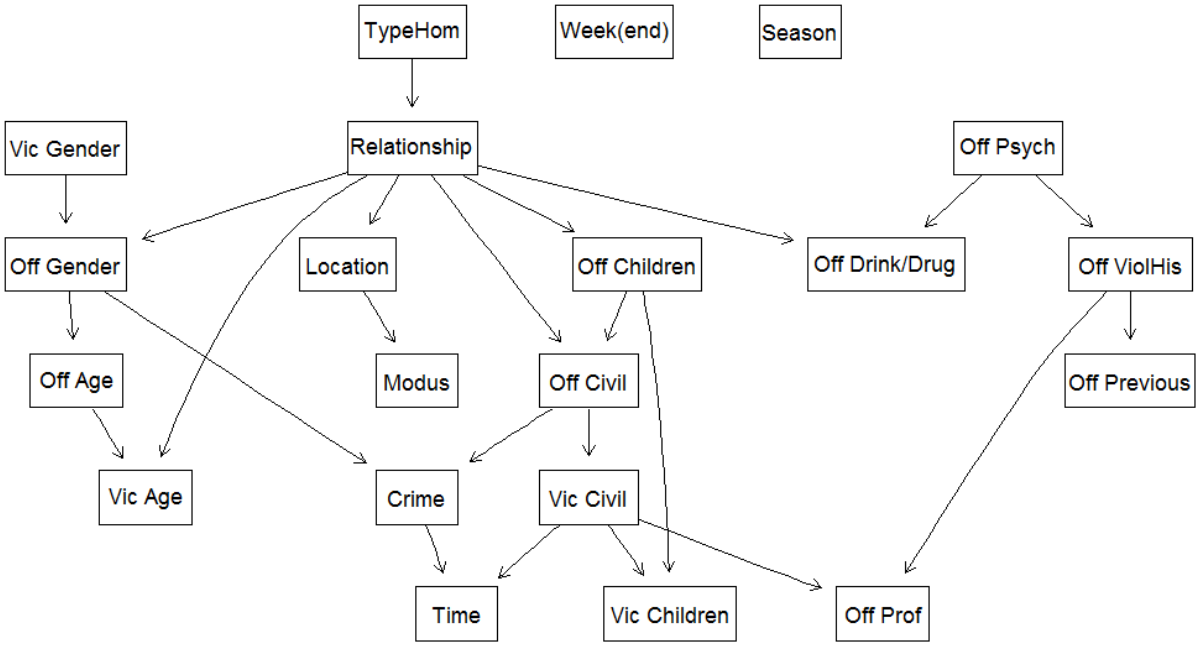
A Bayesian network $\mathcal{B}$ is defined as follows. Let $\mathcal{B} = (G, P)$ be a Bayesian network, where $G = (V, A)$ is a directed acyclic graph (DAG) with V the set of nodes corresponding to the variables $\mathrm{X} = \{\mathrm{X}_1, \dots, \mathrm{X}_n\}$ and A the set of arcs (or edges), with $A(G) \subseteq V(G) \times V(G)$, and where

$$\mathrm{P}(\mathrm{X}_1, \dots, \mathrm{X}_n) = \prod_{v \in V} P(X_v \mid X_{\pi(v)}), \quad [1]$$

is the joint probability distribution defined in terms of the set of conditional probability distributions $P(X_v \mid X_{\pi(v)})$ of the Bayesian network (Korb & Nicholson, 2011). In other words, the definition of a Bayesian network includes that the joint probability distribution, also known as the global probability distribution, is equivalent to the product of the local probability distributions – i.e. the conditional probability distributions – of all variables in G (Scutari, 2010). The local probability distribution of each variable is given by the *Markov property* of Bayesian networks – perspicuously explained by Korb and Nicholson (2011) – which states that every variable X_j directly depends only on its parents $X_{\pi(v_j)}$.

The directions of the arcs $A(G)$ represent the (in)dependence relationships between the variables $V(G)$. Sometimes the direction of an arc could be determined easily by logical reasoning, for example if there is a clear cause and effect relationship between two variables; setting the direction of the arc from cause to effect is most natural then. However, mathematically there is no difference between an arc directed from cause to effect and an arc directed from effect to cause. Fenton and Neil (2013, p. 173) clearly explain this by showing that a Bayesian network consisting of two nodes might differ in graphical structure by turning around the direction of the arc, but the marginal probabilities

Figure 1: Example of a Bayesian network structure



are identical in both cases and therefore the results will be the same if an observation is entered in either one of the networks.

Of course, Bayesian networks normally consist of more than two nodes, and unfortunately the arc directions representing the dependence relationships between three variables are not plainly reversible, since then the marginal probabilities of the network will not remain equivalent (see Section 3.2). Therefore, Pearl (1988) was the first to develop a structure learning algorithm; a technique to capture the dependence relationships (and thus arc directions) of the set of variables in an as compact as possible graphical representation (Korb & Nicholson, 2011). Using such an algorithm, the direction of an arc is not merely determined by the relationship between the two connected variables, because the other variables and dependencies between them may influence the possible directions of a specific arc since a Bayesian network by definition is a directed acyclic graph (Fenton & Neil, 2013). Before discussing structure learning algorithms in Section 3.3 and parameter learning in Section 3.4, first some characteristics of the structure of a Bayesian network will be explained in Section 3.2.

3.2 Bayesian network properties

While working with Bayesian networks the it is generally assumed that all direct dependences in the system that is modelled are represented by an arc (Korb & Nicholson, 2011). In other words, there are no direct dependences in the system that are not shown in the network. A graphical models that has this property is called an *Independence-map* (I-map), since all independences that are shown in the network (by the absence of an arc) are real. Bayesian networks are by definition directed I-maps (Flesch & Lucas, 2007). However, not all arcs in a Bayesian network necessarily represent real dependences in the system, since the conditional probability tables could be defined in such a way that

the dependence suggested by an arc is actually absent (Korb & Nicholson, 2011). For representation reasons as well as computational reasons, most of the time *minimal I-maps* – I-maps that are not I-maps anymore after deleting any arc – are preferred. In the most ideal situation a Bayesian network is a *perfect map* (P-map): the network is not only a (minimal) I-map, but also a *Dependence-map* (D-map). The latter is the case if every arc in the Bayesian network corresponds to a direct dependence in the system that is modelled. See Flesch & Lucas (2007) for a more extensive explanation of I-maps, D-maps, and P-maps.

A Bayesian network has a simple graphical representation, yet the graph provides a lot of information about the (in)dependence relationships between the variables. If two variables are directly dependent, there will be an arc connecting the nodes. Figure 1 shows for example a direct relationship between Victim Gender and Offender Gender, and also between Location and Modus. Two variables could also depend indirectly of each other, then there will be two or more arcs that connect the nodes via one or more other nodes. In Figure 1 Offender Psychological problems and Offender Previous convictions have an indirect dependence relationship, and the same goes for Type of Homicide and Victim Children.

For a set of three nodes, there are three possible structures that form the basics of the graphical and probabilistic representation of a Bayesian network (Scutari & Denis, 2015). They are often called *fundamental connections*, and are all three visible in Figure 1. A *serial connection* has a structure of the form A → B → C, such as the nodes Victim Gender, Offender Gender, and Offender Age. The structure A ← B → C is called a *divergent connection*, which is found for Offender Psychological problems, Offender Drink/Drug, and Offender Violent History. When a child receives directed arcs from two of its parents, and thus has the structure A → B ← C, this is called a *convergent connection*. Crime, Victim Civil status and Time appear to have such a connection.

Two variables are independent if there is no arc that connects their nodes. In Figure 1 for example, the variables Week(end) and Season are completely independent of all other variables. Yet, more often two variables are independent of each other given one or more other variables. If a node *blocks* all *paths* between two other nodes, it is said to *d-separate* them. A more formal definition of *d-separation* (directed separation): if **X**, **Y** and **Z** are three disjoint subsets of nodes in a DAG, then **X** and **Y** are d-separated by **Z** (denoted $\mathbf{X} \perp_G \mathbf{Y} \mid \mathbf{Z}$), if and only if in every path between each $X \in \mathbf{X}$ and each $Y \in \mathbf{Y}$ there is a node V that either is in **Z** and does not have converging arcs, or has converging arcs and neither V nor any of its children are in **Z**. If **Z** d-separates **X** and **Y**, X and Y are conditionally independent given **Z**. In other words: if a node has a serial or diverging connection, it blocks the path if that node is being conditioned on; and if a node has a diverging connection, it blocks the path if neither that node nor one of its descendants is conditioned on. Figure 1 shows for example that Offender Children and Offender Civil status d-separate Relationship and Victim Children, but not Relationship and Time (there is still a path from Relationship to Time via Offender Gender and Crime). From Figure 1 could also be concluded that Type of Homicide and Offender Age are d-

separated by Relationship, and thus are conditionally independent, but if Victim Age is being conditioned on, then Type of Homicide and Offender Age become dependent.

A more detailed explanation of Bayesian networks and their properties is beyond the scope of this research. The books of Korb and Nicholson (2011) and Scutari and Denis (2015) are highly recommended for anyone who wants to read a thorough and perspicuous explanation of the structure and properties of Bayesian networks. For reasons of convenience, only the learning algorithms that were used in this research are discussed in the next section.

3.3 Structure learning

Already since the 17th century mathematicians are developing methods to deal with uncertainty, though Bayesian network structure learning is a relative new and rapidly growing scientific field of research, with Judea Pearl being one of the pioneers (see Pearl, 1988). A thorough explanation and summary of the current literature about Bayesian network learning is given by Daly, Shen, and Aitken (2011), but also more recently new, more optimised and faster Bayesian network learning algorithms were proposed by several authors (amongst others: Contaldi, Vafaee, & Nelson, 2017; Kreimer & Herman, 2016; Li, Xing, Zhang, & Chen, 2017; Liu, Zhou, Lam, & Guan, 2017).

Structure learning algorithms could be divided into three categories: constraint-based algorithms, score-based algorithms, and hybrid algorithms. Constraint-based algorithms use conditional independence tests to obtain the Markov blankets (i.e. the parents, children, and other parents of its children) of the variables and subsequently construct the Bayesian network based on those. Score-based algorithms are heuristic optimisations that order possible structures on a goodness-of-fit score. Hybrid algorithms combine elements of constraint-based and score-based algorithms simultaneously to find the optimal network structure (Scutari & Ness, 2018). Eight different structure learning algorithms were applied to the data available in this research, and are discussed below: five constraint-based (PC-stable, Grow-Shrink, IAMB, Fast-IAMB, and Inter-IAMB), two score-based (Hill-Climbing, and Tabu Search), and one hybrid structure learning algorithm.

The PC-stable algorithm that was used here is a modern implementation of the PC algorithm, which was the first practical constraint-based structure learning algorithm (Scutari & Ness, 2018). The Grow-Shrink algorithm uses pairwise independence tests to recover the Markov blanket of each node and herewith constructs the Bayesian network in two phases, a “grow” phase and a “shrink” phase (Margaritis, 2003). The Incremental Association Markov Blanket (IAMB) algorithm is structurally similar to the Grow-Shrink algorithm, since it uses the same two phases, called “forward” and “backward” phase here. However, because the Grow-Shrink algorithm orders the variables according to their degree of association with the concerned node in the “grow” phase and the IAMB algorithm does not, the latter selects less false positives in the Markov blankets of each node (Tsamardinos, Aliferis, & Statnikov, 2003a). The Fast-IAMB and Inter-IAMB algorithms are variants of the IAMB algorithm; the first uses speculative stepwise forward selection to reduce the number of conditional

independence tests, the latter uses forward stepwise selection to avoid false positives in the Markov blanket detection phase (Scutari & Ness, 2018). For a more extensive explanation of constraint-based structure learning algorithms, see Scutari (2017). More specifically for the five constraint-based structure learning algorithms mentioned here, see Colombo and Maathuis (2014; PC-stable), Margaritis (2003; Grow-Shrink), Tsamardinos et al. (2003; IAMB), and Yaramakala and Margaritis (2005; Fast-IAMB and Inter-IAMB).

The score-based Hill-Climbing algorithm is a hill climbing greedy search on the space of the directed graphs (Scutari & Ness, 2018). Although hill climbing is the most widely known and perhaps simplest iterative technique used for optimisation, it is also well known that hill climbing could return a local minimum that is not necessarily the global minimum (Beretta, Castelli, Gonçalves, & Ramazzotti, 2017). Therefore, more advanced neighbourhood search methods were developed, one of which is Tabu search. The Tabu Search algorithm uses an adaptive memory to explore new areas during the search and thus avoids getting stuck in a local minimum (Beretta et al., 2017; Leegon, 2009).

A hybrid structure learning algorithm uses a constraint-based or local search algorithm in the "restrict" phase and a score-based algorithm in the "maximise" phase (Scutari & Ness, 2018). Which combination of algorithms was used in this research is explained in paragraph 6.3. In any case, it has been proven that hybrid algorithms can outperform numerous score-based and constraint-based algorithms. For example, Tsamardinos, Brown, and Aliferis (2006) have shown that the Max-Min Hill-Climbing algorithm outperforms the following other algorithms that are often used to construct Bayesian networks: the PC, Sparse Candidate, Three Phase Dependency Analysis, Optimal Reinsertion, Greedy Equivalence Search, and Greedy Search.

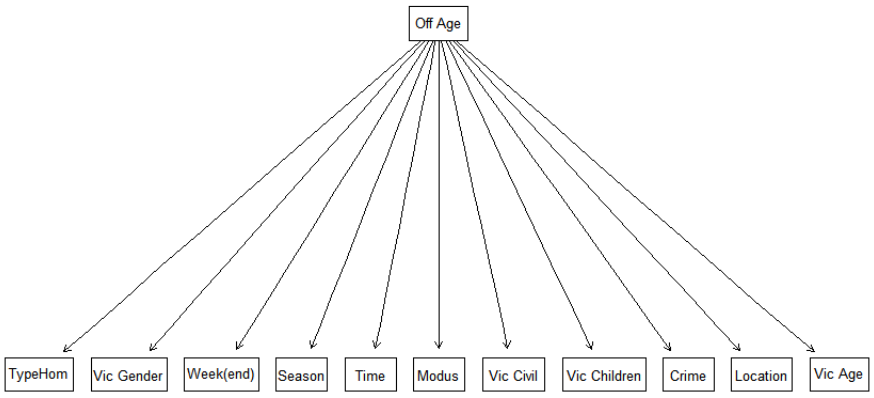
Though the eight mentioned structure learning algorithms are developed to gain insight into the dependence relationships between the variables in this particular dataset (i.e. their aim is to discover the true network structure), those networks do not necessarily have the best prediction accuracy. So called Bayesian network classifiers favour predictive power over a correct network structure and thence often result in more accurate predictions than structure learned Bayesian networks (Scutari & Ness, 2018). Therefore, the Naïve Bayes and Tree-Augmented Naïve Bayes algorithms were also applied to the available data in this research, to obtain a network that potentially has the best prediction accuracy. Naïve Bayes assumes all the variables to be conditionally independent of each other given the predicted variable (class variable). In other words, a naïve Bayesian network classifier always has the structure as shown in Figure 2a. This clearly is an unrealistic assumption, so the performance of naïve Bayes classifiers with its fairly good predictions is often called surprising (Friedman, Geiger, & Goldszmidt, 1997). Tree-augmented naïve Bayes is an improvement of naïve Bayes, which approximates the interactions between the feature variables (Friedman et al., 1997). For the Tree-augmented naïve Bayesian network in this research the Chow-Liu tree structure is imposed on the naïve Bayesian structure; whereas the Chow-Liu algorithm normally returns an undirected tree,

here the conditional mutual information of the feature variables on the class variable is used to determine the arc directions (Chow & Liu, 1968; Scutari & Ness, 2018). Figure 2b shows an example of the structure of a Tree-augmented naïve Bayesian classifier that was learned with the data used in this research. As with naïve Bayes, the class variable always depends on all feature variables in Tree-augmented Bayes, but here the feature variables are allowed to have dependence relationships with each other as well.

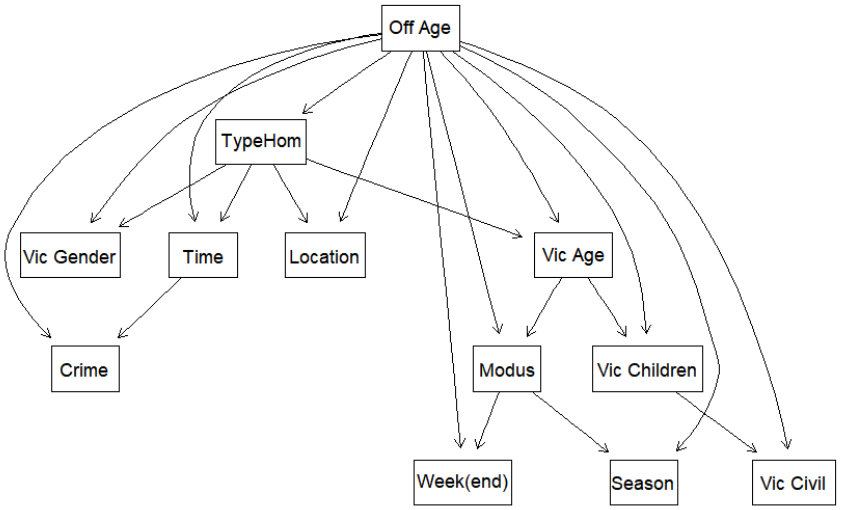
3.4 Parameter learning

If the structure of a Bayesian network is determined, the complementary parameters can be established. The joint probability distribution of a Bayesian network is uniquely defined by the (in)dependence relationships between the variables. Consequently, a Bayesian network leads to a reduction in the number of model parameters compared to deriving the joint distribution of the variables by the chain rule. This reduction is one of the reasons why Bayesian networks are known for relatively fast computations and straightforward inference and learning (Ben-Gal, 2008).

Figure 2: Structure examples of (a) naïve Bayesian classifier, and (b) Tree-Augmented naïve Bayesian classifier



(b) Tree-Augmented Naïve Bayesian Classifier



The parameters of a Bayesian network can be learned from data if all arcs in the structure are directed arcs (Scutari & Ness, 2018). The parameters are the values in the conditional probability distributions of the nodes in the network (Neapolitan, 2004). When using maximum likelihood estimation (MLE), those values could simply be the actual probabilities as found in the data, while with a Bayesian parameter estimation, the values could specify a conditional density that models the probabilities in the conditional probability tables (Daly et al., 2011). But if a network has many nodes and there is not much data available, it is likely that the probability tables of some nodes can be undefined if that specific case does not occur in the dataset. This problem can be solved by placing a prior distribution on the variables, for which often the Dirichlet distribution is used (Daly et al., 2011).

Although traditionally Bayesian networks were designed to give insight into the dependence relationships between discrete variables, nowadays parameter learning in Bayesian networks is possible for discrete as well as continuous variables. Neapolitan (2004) explained this procedure extensively in Chapter 6 of his work. He also mentioned that limited and/or missing data are problematic while learning the parameters of a Bayesian network. Three different assumptions can be applied to missing data: missing-completely-at-random (MCAR), missing-at-random (MAR), and missing-not-at-random (MNAR), which have different implications for how could be dealt with the missing values. Daly et al. (2011) explained those differences clearly:

Under a missing-completely-at-random (MCAR) assumption, the missing value mechanism depends neither on the observed data nor on the missing data. This means that the data with missing values could simply be discarded. This is an extremely easy situation to implement, but is a bad policy in general, in that some if not most of the data would be unavailable for learning. Under a missing-at-random (MAR) assumption, the missing value mechanism depends on the observed data. This means the missing data can be estimated from the observed data. This is more complicated than the MCAR situation, but all the data get used. Under a missing-not-at-random (MNAR) assumption, the missing value mechanism depends on both the observed and missing data. On account of this, a model of the missing data must be supplied. This is the most complicated situation, as a model may not be readily available, or could even be unknown. (p. 114).

Assuming MCAR is mostly unrealistic and assuming MNAR makes the estimation of parameters rather complicated (Neapolitan, 2004). Therefore, MAR is often assumed, and the Expectation Maximisation (EM) algorithm is used to estimate the parameters with missing data (Daly et al., 2011). Besides the EM algorithm, also Monte Carlo methods – in particular Gibbs sampling – could be used to handle missing values (Daly et al., 2011; Neapolitan, 2004). For an explicit explanation of the EM algorithm and how it is used in Bayesian network parameter learning, see paragraph 6.5 in Neapolitan (2004).

Just like with structure learning, much research is done into parameter learning recently. Computer science and machine learning researchers are looking for quicker and more optimal methods

to estimate the parameters with the largest possible precision (e.g. Lamine, Kalti, & Mahjoub, 2011; Su, Zhang, Ling, & Matwin, 2008; Van den Broeck, Mohan, Choi, & Pearl, 2015). In this research parameter learning was done using Bayesian parameter estimation as this was available in the *bnlearn* package (Scutari & Ness, 2018).

4. Criminal profiling using Bayesian network learning in this research

Criminal profiling is mostly used in (violent) serial homicide and/or sexual homicide cases (Strano, 2004). However, the number of those crimes is relatively small in comparison with the number of single homicide cases; in particular in the Netherlands, where serial homicide is a very rare phenomenon. Some profiling techniques even require multiple events, especially in geographical profiling (e.g. Canter & Larkin, 1993; Rossmo, 1993), but criminal profiling is also possible in single events. A Bayesian network is a suitable method that could be used in single cases; the recovered victim and crime characteristics could be initiated in the model, and posterior probabilities for the offender characteristics will be produced automatically. Section 4.1 explains more about the choice for a Bayesian network approach in this research. Thereafter, Sections 4.2 and 4.3 give more information about the dataset and variables used here.

4.1 Criminal profiling with a Bayesian network

Strano (2004) mentioned several advantages of criminal profiling based on a neural network, which also apply to criminal profiling based on a Bayesian network. There would not have to be a professionally trained profiler present at every crime scene, since the tool is easy to use and does not require knowledge about the psychological or psychopathological processes behind crimes. It could be employed by every crime scene investigator, and the profile could be generated rather quickly (possibly already at the crime scene on a laptop). Another advantage of this approach to criminal profiling, is the capacity of the model to become more precise as more data is added.

Artificial intelligence – and thus neural network – approaches to criminal profiling are very promising, yet a lot of data is needed to use those advanced techniques. The dataset used in this research is the most extensive one about homicides in the Netherlands, but the dataset is not complete. Since the data contains missing values and Bayesian networks are able to handle this, a Bayesian network approach to criminal profiling was used here. Furthermore, as mentioned before, in comparison to logistic regression, Bayesian networks have lower error rates (Stahlschmidt et al., 2013), and are more user friendly since they predict multiple offender characteristics at once and the probability distributions are easy to interpret.

The structure and parameters of a Bayesian network can be estimated by experts, or can be learned from data. Building a Bayesian network from expert knowledge is unfeasible when experts do not have enough insight into the relationships between the variables or cannot give a precise ordering of the variables (Stahlschmidt et al., 2013). Because there are a lot of victim and offence

characteristics as well as offender characteristics in the dataset used in this research, it might be hard or even impossible to construct the structure by hand (Daly et al., 2011). Besides, not enough research is done to their mutual (in)dependence relationships, so the Bayesian network structure and parameters in this research was learned from data instead of constructed based on expert knowledge.

4.2 Dataset

The dataset used in this research is the Dutch Homicide Monitor (Liem, Alink, Aarten, & Schönberger, 2018), which is part of a European initiative to create a joint database on lethal violence, the European Homicide Monitor (Ganpat, et al., 2011). The Dutch dataset contains detailed offender, victim, and crime information on all homicides (in the Dutch Criminal Code (WvSr) defined as murder or manslaugther, Art. 289 or 287-288 WvSr) that were committed in the Netherlands between 1992 and 2016. A case is included in this dataset when the Prosecutor charges the suspect with murder, manslaughter, or complicity in one of the two. Thus not every suspect will be convicted of one of those crimes in the end; not every charge leads to conviction. However, when the judge acquits the suspect of any crime, the case is left out of the dataset (Liem, et al., 2018).

The Dutch Homicide Monitor is composed of seven sources: (1) newspaper articles, extracted from Elsevier's 'homicide lists' that are published yearly by Gerlof Leistra and retrieved from searching through the LexisNexis database; (2) data from the National Police, which contains details surrounding the case and the arrest of the suspect; (3) data from the Public Prosecution Office (OM), which includes details about the charges against and conviction and sentencing of the offender; (4) case files, with background information about the case and the suspect (through Pro Justitia reports); (5) data from the Legal Services Department (DJI), which contains information on prisoners and their detention period; (6) files from the Research and Documentation Centre – criminal records (OBJD), with details about the criminal careers of the suspects; and (7) data from Statistics Netherlands (CBS), which gives insight into the family situation and history of offenders and victims of homicide (Ganpat, et al., 2011; Ganpat & Liem, 2012). These sources are layered; the lists of Leistra are verified by the data of the National Police, and so on. In this way all cases that meet the inclusion criteria end up in the Dutch Homicide Monitor.

The dataset is in the midst of an update; it is checked whether all homicides are correctly in the dataset, and more information about the cases is added. The data used in this research were extracted in January 2018, and consisted of 8004 individuals: 4321 perpetrators (54%) and 3683 victims (46%), in 3474 cases. To prevent the Bayesian network becoming too complex, only cases with one single victim and one single offender were selected. There was one case without homicide date, wherefore it was removed. Cases for which the victim and the perpetrator information was not consistent were removed as well. The resulting dataset consisted of 2725 homicide cases with one perpetrator and one victim in each case. Those cases were randomly divided into a training set (80%, N = 2180) and a test set (20%, N = 545). The models were learned on the training set, and their

prediction accuracies were evaluated on the test set. The training set of 2180 cases should be sufficient, as Baumgartner et al. (2008) showed that a dataset of at least 1000 cases should be sufficient for structural robustness of the Bayesian network as well as accurate predictions.

4.3 Variables

Various victim, crime, and offender characteristics were available in the dataset. Following Aitken et al. (1996) and Beek et al. (2010) the characteristics that were not present in at least 5 percent of the cases – for example information about eyewitnesses, the motives of the offender, and the education levels of the offender and victim– were excluded from the analysis. Table 1 shows a summary of the remaining 4 victim, 7 crime, and 10 offender variables that were used in this research.

The only continuous variable in this dataset was the age of the victims and perpetrators, which were discretised in the categories: < 18, 18-25, 26-45, 46-60, > 60. This was done because first of all the algorithms in the package *bnlearn* do not support Bayesian networks containing both discrete and continuous variables. A more important reason to discretise those variables, lies in the method used for parameter learning. When maximum likelihood estimation is used during the parameter learning, the local distributions of discrete nodes have no estimate for all parents configurations that were not observed in the data (Scutari & Ness, 2018). This could propagate to the prediction values, which is of course not desirable. Therefore Bayesian parameter estimation was used when learning the parameters, and this method was only implemented in *bnlearn* for discrete variables.

Table 1: Frequencies of victim, crime, and offender characteristics

N = 2725	Level	Frequency	Percentage
Victim			
Age	< 18	215	7.89
	18-25	435	15.96
	26-45	1296	47.56
	46-60	487	17.87
	> 60	250	9.17
	Unknown	42	1.54
Gender	Male	1627	59.71
	Female	1054	38.68
	Unknown	45	1.61
Civil status	Married/relation	221	8.11
	Single/widowed	142	5.21
	Unknown	2362	86.68
Children	Yes	252	9.25
	No	176	6.64
	Unknown	2297	84.29

Crime			
Season	Winter	678	24.88
	Spring	682	25.03
	Summer	742	27.23
	Fall	623	22.86
Week(end)	Fri-Sun	1291	47.38
	Mon-Thu	1434	52.62
Time	Morning (6.00-12.00)	134	4.92
	Afternoon (12.00-18.00)	205	7.52
	Evening (18.00-24.00)	304	11.16
	Night (00.00-06.00)	297	10.90
	Unknown	1785	65.50
Crime	Murder	917	33.65
	Manslaughter	604	22.17
	Unknown	1204	44.18
Modus operandi	Hanging/Strangling/Suffocation	342	12.55
	Firearm	758	27.82
	Knife/Sharp object	1023	37.54
	Other weapon	164	6.02
	Phys violence no weapon	175	6.42
	Other	111	4.07
	Unknown	152	5.58
Type of homicide	Partner killing	728	26.72
	Family/infanticide	367	13.47
	Robbery/Criminal milieu	249	9.14
	Other	353	12.95
	Unknown	1029	37.72
Location	Private home	1477	54.20
	Public place	902	33.10
	Other	164	6.02
	Unknown	182	6.68
Offender			
Age	< 18	62	2.28
	18-25	521	19.12
	26-45	1321	48.48
	46-60	348	12.77
	> 60	81	2.97
	Unknown	392	14.39
Gender	Male	2090	76.70
	Female	229	8.40

	Unknown	407	14.90
Relationship	Partner	536	19.67
	Ex-partner	196	7.19
	(Step-)child	130	4.77
	(Step/Grand-)parent	120	4.40
	Other relative	117	4.29
	Strangers	99	3.63
	Other	255	9.36
	Unknown	1272	46.68
Civil status	Married/relation	389	14.28
	Single/widowed	229	8.40
	Unknown	2107	77.32
Children	Yes	384	14.09
	No	188	6.90
	Unknown	2153	79.01
Professional status	Employed	145	5.32
	Unemployed	177	6.50
	Unknown	2403	88.18
Drink/drug	Yes	150	5.50
	No	314	11.52
	Unknown	2261	82.97
Psychological problems	Yes	284	10.42
	No	218	8.00
	Unknown	2223	81.58
Violent history	Yes	127	4.66
	No	154	9.32
	Unknown	2344	86.02
Previously convicted	Yes	166	6.09
	No	224	8.22
	Unknown	2335	85.69

5. Network construction

This section discusses the process of constructing the final, combined, network. Section 5.1 discusses the preliminaries. Section 5.2 demonstrates how the constraint-based networks were fit, and Sections 5.3 and 5.4 do the same for the score-based and hybrid networks respectively. Lastly, in Section 5.5, the combined network will be constructed.

5.1 Preliminaries

Similar to the work of Stahlschmidt, et al. (2013) multiple structure learning algorithms were applied to the training set and subsequently merged into one Bayesian network in order to get insight into the (in)dependence relationships between the variables. As said, eight structure learning algorithms were applied to the dataset: five constraint-based (PC-stable, Grow-Shrink, IAMB, Fast-IAMB, and Inter-IAMB), two score-based (Hill-Climbing, and Tabu Search), and one hybrid structure learning algorithm. All were available in version 4.2 of the package *bnlearn* (Scutari & Ness, 2018) in the statistical environment R (R Core Team, 2013).

The data did not contain any missing values, since the category 'Unknown' was added to every variable to prevent problems with missing values. The reason to handle missing values in this way, is because imputing the missing variables is possibly not suitable here. Every homicide case is different, and building a model based on imputed values, was considered undesirable. Moreover, in criminal investigation having no information about a variable could also indicate something; then, having no information, ís the information. If for example the gender of a victim cannot be determined, this possibly says something about the cruelty of the circumstances in which the homicide was committed; a victim must be rather physically assaulted if its gender could not be determined. So therefore it was chosen here to add the category 'Unknown' to every variable. This reasoning actually implicitly assumes the absence of missing values, since having no information is seen as information and not as a missing value.

In an attempt to find a model that describes the (in)dependence relationships between the variables as good as possible, the eight different structure learning algorithms were separately applied to the training set. All network constructions were performed with the significance level set to the usual 5 percent. For every constraint-based algorithm different conditional independence tests were applied and the model was chosen that showed the lowest mean 5-fold cross-validated prediction error over the offender variables. Something similar was done with the score-based algorithms; different scores were applied and the model was selected that on average predicted the offender variables best in a 5-fold cross-validation. In order to find a suitable hybrid network, different combinations of score-based and constraint-based or local search algorithms were applied. Again, the network that showed the lowest mean prediction error in a 5-fold cross-validation was selected. Finally, the five constraint-based networks, the two score-based networks and the one hybrid network that had the lowest prediction errors were merged into one combined network.

5.2 Evaluation measures

The performance of Bayesian networks can be evaluated in multiple ways. The model performance in this research was reported by means of the prediction error, the log-likelihood loss, and an ROC-analysis. Before discussing the performance of the eight different structure learning algorithms and the combined model in Sections 5.3-5.6 and comparing them with a naïve Bayesian classifier and tree-

augmented naïve Bayesian classifier in Section 6, the three different evaluation measures will be described here.

The prediction error of every network was derived by a 5-fold cross-validation. Let $D = (r_i, \ldots, r_m)$ be the test set containing observed values of X, i.e. $r_i = (x_{i1}, \ldots, x_{in})$ with x_{ij} the value of the j^{th} variable of the i^{th} case. Besides, let $\tau \in [0,1]$ be a *threshold* of a probabilistic prediction, then

$$c_i = \begin{cases} 1, & \text{if } P(X_c = x_{ic} \mid E_i) \geq \tau \\ 0, & \text{otherwise} \end{cases}, \quad [2]$$

with $X_c = x_c$ the value in record r_i for variable X_c (the offender characteristic), and E_i the evidence taken from record r_i consisting of the values x_{ij} corresponding to the crime and victim characteristics in the concerned case. The *total prediction error* of a Bayesian network $\mathcal{B}$ for offender characteristic X_c could then be defined as:

$$TPE(X_c) = 1 - \frac{\sum_{k=1}^{m} c_k}{m}. \quad [3]$$

However, a prediction error could also be defined with a probabilistic maximum. This is the approach that was used in this research, and reads as follows. Let $\tau \in [0,1]$ again be a *threshold* of a probabilistic prediction and

- $c_k \in \{0,1\}$: the *correct* classification of case *k*;
- $i_k \in \{0,1\}$: the *incorrect* classification of case *k*;
- $u_k \in \{0,1\}$: the *undefined* classification of case *k*.

It holds that for each case *k*: $c_k + i_k + u_k = 1$, $0 \leq c_k + i_k \leq 1$, and $c_k + i_k = 0$ if and only if $u_k = 1$. These measures are defined as follows:

$$c_k = \begin{cases} 1, & \text{if } \max_{x_c} P(X_c = x_c \mid E_k) \geq \tau \text{ and } \arg\max_{x_c} P(X_c = x_c \mid E_k) = x_{kc} \\ 0, & \text{otherwise} \end{cases}, \quad [4]$$

where the value x_{kc} is included in record r_k,

$$i_k = \begin{cases} 1, & \text{if } \max_{x_c} P(X_c = x_c \mid E_k) \geq \tau \text{ and } \arg\max_{x_c} P(X_c = x_c \mid E_k) \neq x_{kc} \\ 0, & otherwise \end{cases}, \quad [5]$$

and

$$u_k = \begin{cases} 1, & \text{if } \max_{x_c} \mathrm{P}(X_c = x_c \mid E_k) < \tau \\ 0, & \text{otherwise} \end{cases}. \quad [6]$$

Now the *partial prediction error* of a Bayesian network $\mathcal{B}$ for offender characteristic X_c is defined as:

$$PPE(X_c) = 1 - \frac{\sum_{k=1}^{m} c_k}{\sum_{k=1}^{m} c_k + i_k}. \quad [7]$$

Note that using the definitions of c_k, i, and u_k also other performance measures could be defined, such as the *total partial prediction error* of a Bayesian network $\mathcal{B}$ for offender characteristic X_c:

$$TPPE(X_c) = 1 - \frac{\sum_{k=1}^{m} c_k}{m}. \quad [8]$$

Because of the difference in definition in c_k, it is normally not the case that $TPE(X_c) = TPPE(X_c)$ even though the formulae are identical.

Also the log-likelihood loss was derived by a 5-fold cross-validation, but then repeated a hundred times using the function *bn.cv()* (Scutari & Ness, 2018) because in this way the standard deviation of the log-likelihood loss could be estimated. The *negative log-likelihood* is defined as

$$ll(X_c) = -\sum_{k=1}^{m} \ln P(X_c = x_{kc} \mid E_k). \quad [9]$$

Since $\ln P(X_c = x_{kc} \mid E_k) \leq 0$, it follows that $ll(X_c) \geq 0$. The log-likelihood loss is also called negative entropy or negentropy, and is the negated expected log-likelihood of the test set for the Bayesian network from the training set in each fold (Scutari & Ness, 2018). It holds that the lower the negative log-likelihood loss, the better the performance of the model.

Lastly, the performance of the combined Bayesian network will be evaluated by plotting Receiver Operating Characteristic (ROC) curves, and calculating the area under these curves (AUC), as is often done in Bayesian network analyses (Marcot, 2012; Bockhorst, Craven, Page, Shavlik, & Glasner, 2003). An ROC curve shows the relationship between the probability of finding a true positive (Sensitivity) and the probability of finding a false positive (1 – Specificity). The larger the area under an ROC curve, the more accurate the classifications of the model. Random models have an AUC of 0.5, and typically models with an AUC of 0.6, 0.7, 0.8, or 0.9 are considered bad, acceptable, good, or excellent respectively. Here, for every variable of the combined model each of the values was set out against all other values, and subsequently the ROC curve was plotted (see Section 6).

5.3 Constraint-based networks

The conditional independence tests that were compared could be divided into mutual information tests and Pearson's χ^2 tests. The following tests for mutual information were used: the asymptotic χ^2 test (MI), the semiparametric test (MI-SP), and the shrinkage estimator for mutual information (MI-SH). The applied Pearson's χ^2 tests were: the asymptotic χ^2 test (X2), and the semiparametric test (SP-X2). See Scutari (2010) and Scutari and Ness (2018) for more information about these conditional independence tests.

As said, for every offender variable the prediction error was derived with a 5-fold cross-validation, and for every learning algorithm the conditional independence test that had the lowest mean prediction error was selected for constructing the combined network. If a constraint-based structure learning algorithm resulted in an undirected graph that represented an equivalence class, the function *cextend()* (Scutari & Ness, 2018) was used to consistently extend it to a directed acyclic graph (DAG). For more explanation about the construction of a consistent extension of a partially oriented graph, see Dor and Tarsi (1992). Sometimes the (partially) undirected graph could not be consistently extended to a DAG, because with that specific skeleton and v-structures it was impossible to have an acyclic graph. If a consistent extension was not possible the concerned conditional independence test was not taken into consideration for that specific learning algorithm. With the

extension to a directed acyclic graph the Bayesian network parameters could be fit and the prediction error could be computed.

Figure 3 (in Appendix 1) shows the prediction errors per variable for the different conditional independence tests in the five constraint-based learning algorithms. It holds for every learning algorithm that averaged over all offender variables it did not make a large difference which conditional independence test was used. However, sometimes one of the conditional independence tests outperformed the others in predicting a specific variable. For example, as can be seen from Figure 3, the semiparametric test for mutual information (MI-SP) in the PC stable algorithm stood out in predicting the relationship between the victim and the offender; and within the inter-IAMB algorithm the Pearson's asymptotic χ^2 test (X2) predicted better than the other conditional independence tests whether or not the offender was previously convicted.

The conditional independence test that on average performed best within each learning algorithm was used for the construction of the combined model. The PC stable learning algorithm had the lowest mean prediction error with the semiparametric test for mutual information (MI-SP; 18.68%); the Grow-Shrink algorithm with the Pearson's asymptotic χ^2 test (X2; 21.46%); the IAMB with the shrinkage estimator for mutual information (MI-SH; 18.53%); the fast-IAMB with the Pearson's asymptotic χ^2 test (X2; 18.09%); and the inter-IAMB with the semiparametric test for mutual information (MI-SP; 17.74%).

The goodness of fit to the data could be measured by the log-likelihood loss, which is shown for the constraint-based networks in Figure 4 (in Appendix 1). As can be seen from the small range of the y-axes the different conditional independence tests of the networks fit the data approximately equally within every learning algorithm. Also the various algorithms do not differ much in their goodness of fit to the data, although the Grow-Shrink algorithm seems to fit the data slightly worse than the other algorithms. Disregarding the Grow-Shrink algorithm, for the constraint-based algorithms the lowest log-likelihood loss amongst the different conditional independence tests is approximately 16.5.

5.4 Score-based networks

Comparable to the constraint-based algorithms, the score-based algorithms were compared in their performance in terms of prediction error. For both learning algorithms (Hill-Climbing and Tabu) the following scores were used: the Aikaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), the logarithm of the Bayesian Dirichlet equivalent score (BDE), the logarithm of the Bayesian Dirichlet sparse score (BDS), the logarithm of the Bayesian Dirichlet score with Jeffrey's prior (BDJ), the logarithm of the locally averaged Bayesian Dirichlet score (BDLA), and the logarithm of the K2 score (K2).

Again, for every offender variable the 5-fold cross-validated prediction errors were computed and the score with the lowest mean prediction error was used in the construction of the combined

network. Figure 5 (in Appendix 1) shows the prediction errors per variable for the different scores in the two score-based learning algorithms. The Hill-Climbing algorithm had the lowest prediction error (16.92%) when the AIC score was used, whereas the Tabu algorithm performed best when the logarithm of the BDJ score was used (16.88%). Noteworthy is the variation in prediction error of the relationship between the victim and offender; especially in the Tabu algorithm the prediction performance fairly differs between the different score types (from circa 26 to 41 percent). Nonetheless, averaged over all variables the different score types perform quite the same, as can be seen from the black lines in Figure 5.

Figure 6 (in Appendix 1) shows the log-likelihood loss for the score-based networks. With a log-likelihood loss between 15 and 15.5 for all different scores, the score-based networks seem to fit the data somewhat better than the constraint-based networks, but between themselves the fit is almost identical.

5.5 Hybrid network

In the search for a hybrid network with a small prediction error, different combinations of score-based and constraint-based or local search algorithms were compared. The Hill-Climbing algorithm was used with the Aikake's Information Criterion (AIC), since this performed best in terms of prediction error as shown before; the same goes for the Tabu search and the logarithm of the Bayesian Dirichlet score with Jeffrey's prior (BDJ). Both those score-based algorithms were applied with two constraint-based algorithms and two local search algorithms.

The two constraint-based algorithms were Max-Min Parents and Children (MMPC), which is a forward selection technique for neighbourhood detection based on the maximisation of the minimum association measure observed with any subset of the nodes selected in the previous iterations, and Semi-Interleaved Hiton Parents and Children (HITON PC), which is a forward selection technique for neighbourhood detection that is designed to exclude nodes early based on the marginal association (Scutari & Ness, 2018). See Tsamardinos, Aliferis, and Statnikov (2003b), Tsamardinos, Brown, and Aliferis (2006), and Aliferis, Statnikov, Tsamardinos, Subramani, and Koutsoukos (2010) for more information about the MMPC and HITON PC algorithms. With both algorithms the semiparametric test for mutual information (MI-SP) was used, since this conditional independence test performed best in terms of prediction error averaged over all offender variables and all constraint-based learning algorithms discussed above.

The two local search algorithms, CHOW-LIU and ARACNE, use pairwise discrete mutual information coefficients to learn simple graph structures (Scutari & Ness, 2018). CHOW-LIU is an application of the minimum-weight spanning tree and the information inequality, and ARACNE is an improved version of CHOW-LIU that learns polytrees (Margolin et al., 2006; Scutari & Ness, 2018).

As Figure 7 (in Appendix 1) shows, the hybrid network that was constructed by using the ARACNE algorithm during the "restrict" phase and the Hill-Climbing algorithm during the

"maximise" phase had the lowest prediction error (17.22%) averaged over the offender variables. Though, the difference in mean prediction error with the other combinations of algorithms is not large.

The goodness of fit of the hybrid networks is with a log-likelihood around 15.5 comparable to that of the score-based networks (see Figure 8 in Appendix 1). The different combinations of score-based and constraint-based networks scarcely influence the log-likelihood loss. Since the constraint-based, score-based, and hybrid networks do not show large differences in goodness of fit to the data, the performance in terms of prediction errors rather than log-likelihood loss was used to select the structure learning algorithm and the resulting network that was used for the construction of the combined network.

5.6 Combined network

The different structure learning algorithms described above result in eight more and less extensive Bayesian network structures. The extensiveness of the graphical structures ranges from a network with 4 arcs (resulting from the Grow-Shrink algorithm) to a network with 50 arcs (resulting from the Tabu algorithm). Those eight networks were merged into one combined network.

As pointed out by Stahlschmidt et al. (2013), the number of potential edges in a Bayesian network grows exponentially in the number of variables. Though the number of cases ($n = 2180$ in the training set) is larger than the number of variables ($p = 21$), the number of cases is outgrown by the number of potential edges and their corresponding parameters θ. Similar to Stahlschmidt et al. the corresponding challenges for structural learning and prediction were addressed by running several different structure learning algorithms and selecting the arcs that persisted throughout at least three of the eight resulting graphs for the construction of the final model.

The eight applied structure learning algorithms did not uniformly agree on the presence of any of the arcs. However, seven of the eight structure learning algorithms resulted in an arc from "TypeHom" (the type of homicide) to "Relationship" (the relationship between the victim and the offender). Whereas the algorithms did not find much consensus on present arcs, they did agree on 340 missing arcs, out of the possible 420. For six of the arcs that were identified by at least three structure learning algorithms the direction remained unclear. Since the partially undirected graph could not be consistently extended to a DAG, the direction in the combined model for those arcs was established using the Bayes factor. The strength of those arcs was determined in the eight separate structure learning algorithms using Bayes factors, and the direction that had the highest cumulative strength was set in the combined model.

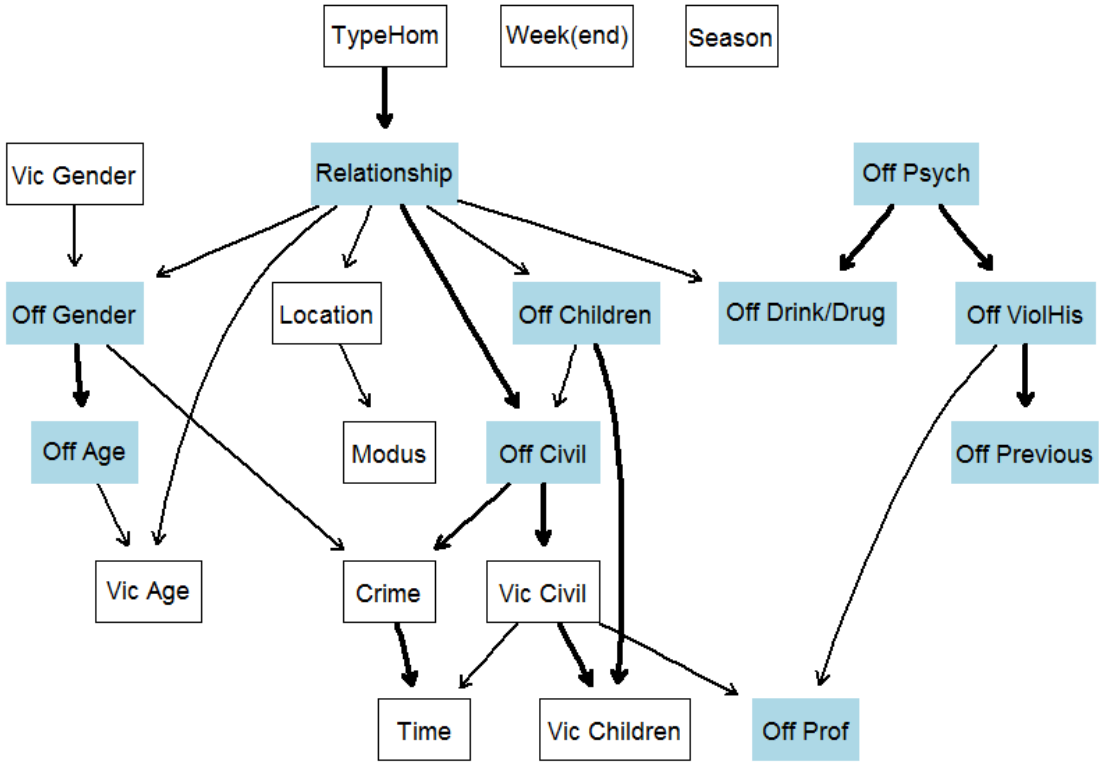
Figure 10 shows the resulting combined network; it consists of 19 nodes and 24 arcs. There were two variables that were not connected to any of the other nodes in the network: "Week(end)" (whether the homicide was committed during the week or in a weekend), and "Season" (the season in which the homicide was committed). Apparently the day on and season in which a homicide was committed do not have a probabilistic relationship with the further circumstances of a homicide. The

offender characteristics are highlighted in light blue, and the thickness of the arcs corresponds with the strength of the arc; the thicker the arc, the stronger confidence in the presence of that particular arc direction.

The relationship between the victim and offender appears to be the pivot in the network; almost all crime and victim characteristics are – whether or not directly – dependent of this variable. This is not surprising: earlier research has shown that the victim-offender relationship might play an important role in the circumstances under which a homicide is committed (Cao, Hou, & Huang, 2008; Chan, Heide, & Myers, 2012). Besides, the combined model shows some other argumentative dependence relationships between the variables. For example, the connections between the victim and offender genders and ages and their connections with the relationship between the victim and offender do not come unexpectedly (see Bell Holleran & Vandiver, 2016). The connection between the civil status and whether they had children or not for both victims and offenders is intuitive: for someone who is married the probability that this person has children will be different than for someone who is single. And the dependencies between the psychological problems of an offender and his violent history, drink/drug abuse, and previous convictions are reasonable as well.

However, not all dependencies are straightforward. Some could be considered to be rational, for example the dependence relationship between the offender gender and what type of crime was committed could be explained by reasoning that men in general tend to be somewhat more impulsive and thus are more likely to commit a manslaughter instead of a murder compared to women.

Figure 10: The combined Bayesian network structure



Nevertheless, there are some dependencies that are not as logical as others. For instance the relation between the modus and location of a crime. The modus appears to be only conditionally dependent of the location. In other words: when the location is known, knowing the modus does not add any information to the model, which might not be directly explicable. In any case, it may have become clear that this Bayesian network is not only able to confirm already existing theories, it could also give inspiration for new criminological theories.

The parameters of the combined network, as well as those of the other eight networks, were learned using Bayesian parameter estimation. As Figure 9 (in Appendix 1) shows, the combined network's goodness of fit to the data is comparable to that of the score-based and hybrid networks; the log-likelihood loss was 15.4. So however the goodness of fit of the combined model was better than the constraint-based algorithms, it did not outperform all separate structure learning algorithms.

In the next section the prediction errors and ROC curves of the combined model will be compared with that of the eight networks resulting from the different structure learning algorithms and two naïve Bayes classifiers.

6. Prediction

In this section the performance of the combined model is compared with that of the eight networks resulting from the different structure learning algorithms and two naïve Bayes classifiers (naïve Bayes and Tree-Augmented naïve Bayes) by computing their partial prediction errors – as defined in Equation 7 in Section 5.2 – on the test set and plotting their ROC curves. The main goal of the model is to predict the offender characteristics from the victim and crime characteristics of a new homicide case, such that the criminal investigators obtain a first direction in their search for the offender. The prediction of an offender characteristic in this context is the value of that specific characteristic that has the highest posterior probability according to the Bayesian network. Or in other words: with prediction here is meant the classification of the Bayesian network. While using the network, of course it is important to know how often the Bayesian network predicts those offender characteristics correctly. The prediction error here is the percentage of cases in which the offender characteristics were predicted incorrectly by the Bayesian network. The test set (N = 545) was used for this purpose, since this data is completely new to the model.

For all Bayesian networks – the eight separate networks and the combined network – the prediction error were calculated on the test set for each of the offender variables. Besides, the prediction errors of the naïve Bayesian and Tree-Augmented naïve Bayesian classifiers were computed, again for every offender variable. As explained in Section 3.2, Bayesian network classifiers favour predictive power over a correct network structure, and the variable of interest is by definition dependent of all other variables in the model. The naïve Bayesian classifier assumes conditional independence between the variables that serve as evidence, whereas the Tree-Augmented naïve Bayesian classifier uses Chow-Liu to approximate the dependence structure of the variables that serve

as evidence. Figures 2a and 2b in Paragraph 3.1 show with Offender Age as variable of interest the naïve Bayesian classifier and Tree-Augmented naïve Bayesian classifier respectively.

Following Stahlschmidt et al. (2013) a threshold was implemented, which needed to be surpassed by the maximum of the posterior distribution in order to accept the prediction. Using a

Figure 11: Prediction errors for the different structure learning algorithms average over the offender characteristics

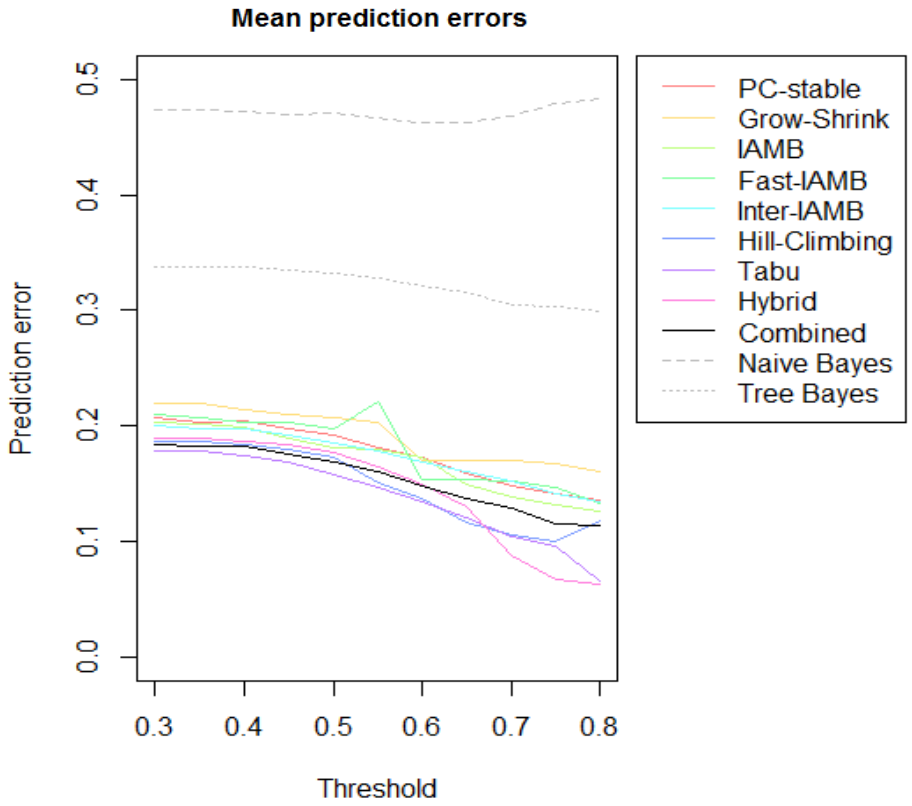
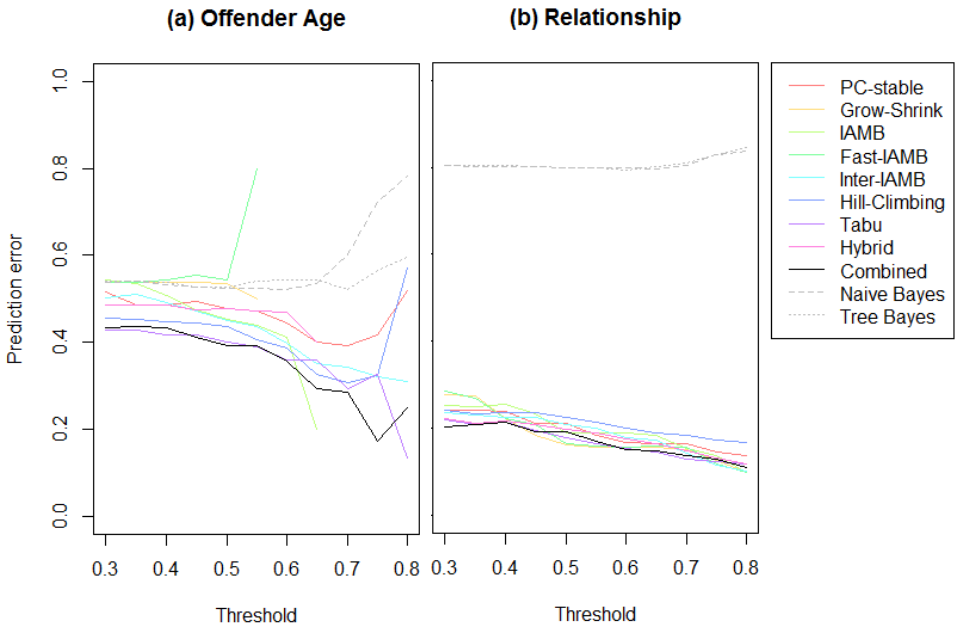
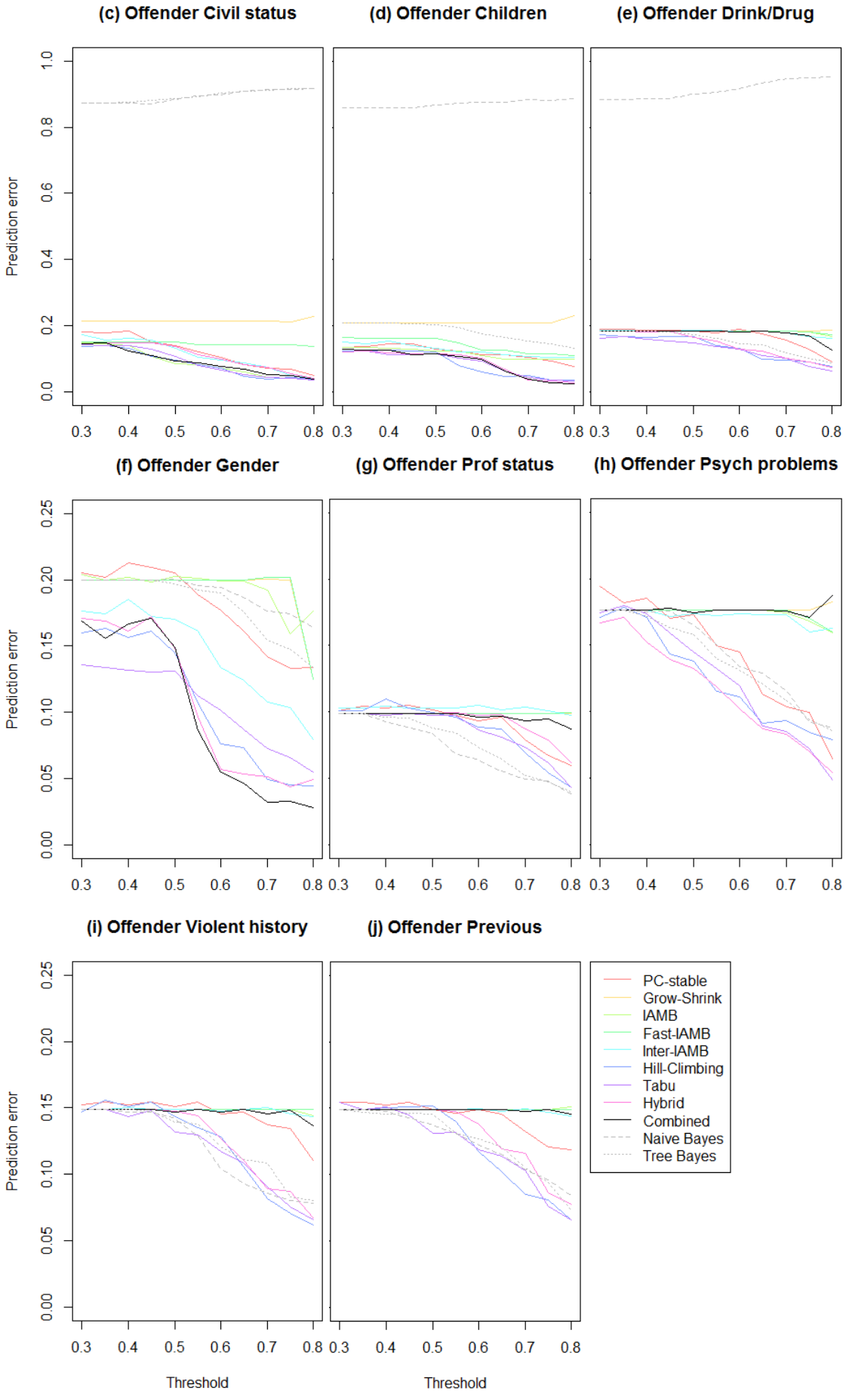


Figure 12: Prediction errors for the different structure learning algorithms per offender characteristic




(c) Offender Civil status
(d) Offender Children
(e) Offender Drink/Drug
(f) Offender Gender
(g) Offender Prof status
(h) Offender Psych problems
(i) Offender Violent history
(j) Offender Previous
Prediction error
Threshold
PC-stable
Grow-Shrink
IAMB
Fast-IAMB
Inter-IAMB
Hill-Climbing
Tabu
Hybrid
Combined
Naive Bayes
Tree Bayes

larger threshold leads to a smaller number of predictions but the ones that are made are more well-assured. So with shifting the threshold, one trades the number of predicted cases for the certainty of the predicted cases.

Figure 11 shows the mean prediction errors of all models averaged over the offender variables. As expected, for most models the prediction error decreased with a larger threshold. The decrease in prediction error was quite smooth for most models, but for example the prediction error of the hybrid network suddenly dropped below 10 percent at a threshold of 0.7, so this could be a plausible choice for the threshold.

Both naïve Bayes classifiers were clearly outperformed by all Bayesian networks. Further, the score-based algorithms seemed to perform somewhat better than the constraint-based algorithms, and the averaged prediction errors of the combined model were smaller than those of the constraint-based algorithms but larger than those of the score-based and hybrid algorithms. However, the prediction errors for the separate offender variables showed different patterns, dependent on the variable of interest.

The prediction errors of the models per variable are shown in Figure 12. Note the difference of the scale of the y-axis between Figure 12 a-e and f-j. Predicting Offender Age seems to be complex, but most models predict most variables with an error rate of approximately 20 percent. However, for some of the variables the naïve Bayesian classifiers have much higher prediction errors than the Bayesian networks. For other variables the performances of the classifiers and networks are comparable, and in one variable (Offender Professional status) the classifiers outperform the networks. For the Bayesian networks holds that the combined network does not necessarily outperform the other networks. Though, Offender Age and Offender Gender – the two offender variables that have the lowest amount of unknown values – are predicted most accurately by the combined model.

Another way to evaluate the model performance of a Bayesian network is by plotting ROC curves. An ROC-analysis was performed for the combined model as well as the naïve Bayesian and tree-augmented naïve Bayesian classifiers. For reasons of convenience beside the three most important offender variables – age, gender, and relationship with the victim – only two less important offender variables were evaluated. Figure 13 shows the ROC curves for the offender characteristics Age (Figure 13a), Gender (Figure 13b), Relationship with the victim (Figure 13c), Offender Civil Status (Figure 13d), and Previous Convictions (Figure 13e). The AUC for Offender Age is approximately 0.7 or higher for all age categories. Therefore could be concluded that the model gives fair predictions for the age of the offender; no excellent predictions, but not bad either. The combined model's predictions for Offender Gender will be more accurate; with an AUC of around 0.9 the combined model predicts the gender of the offender quite well. The relationship with the victim, the civil status of the offender, and whether or not the offender was previously convicted appear to be more complex to predict for the combined model. The AUC's lie around 0.6 or lower, which means bad predictions for this variable; the model does not perform much better than a random classifier – which would have an AUC of 0.5.

Figure 13: ROC curves for the combined model of offender variables (a) Age, (b) Gender, (c) Relationship with the victim, (d) Civil Status, and (e) Previous Convictions

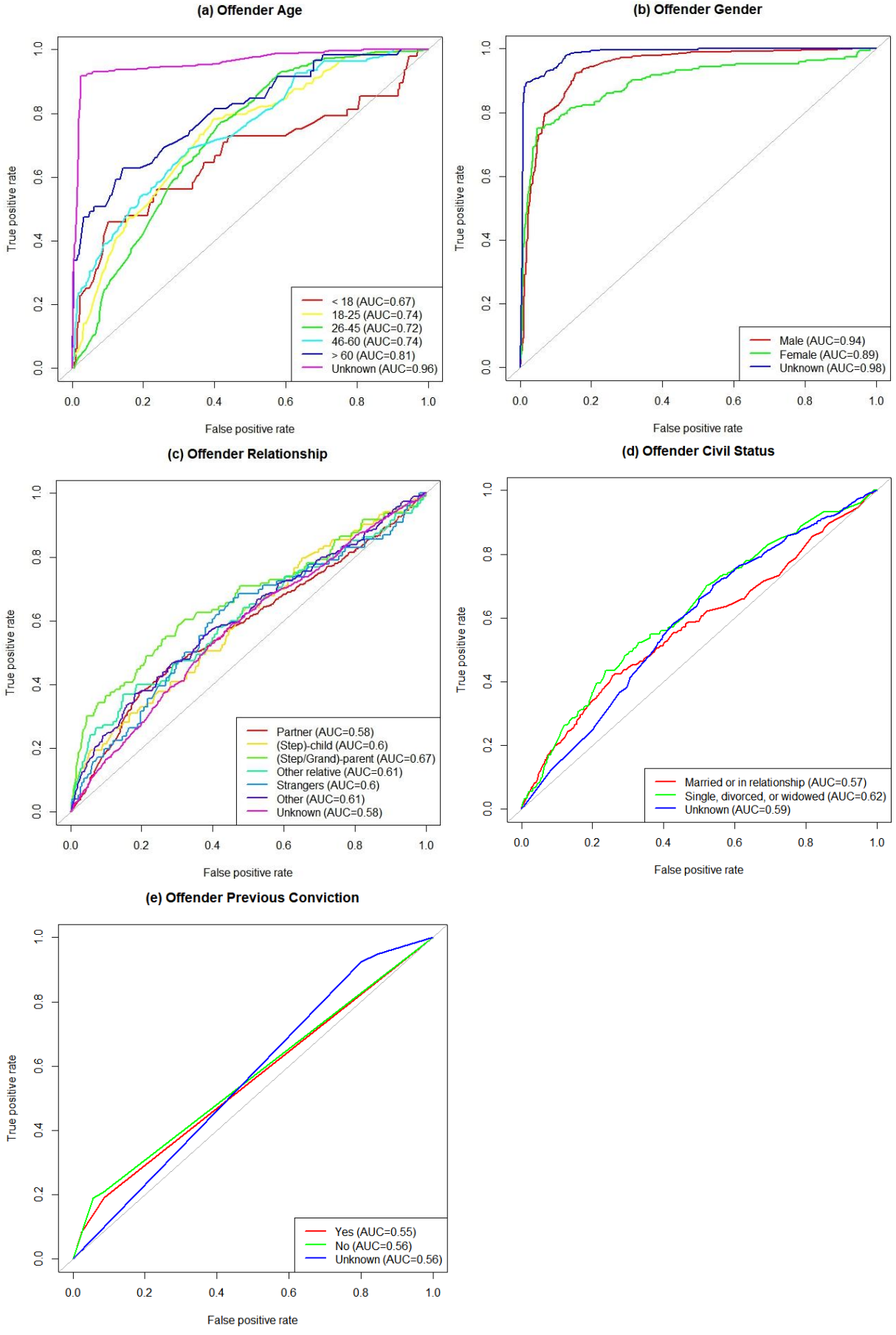


Figure 14: ROC curves for the naïve Bayesian classifier of offender variables (a) Age, (b) Gender, (c) Relationship with the victim, (d) Civil Status, and (e) Previous Convictions

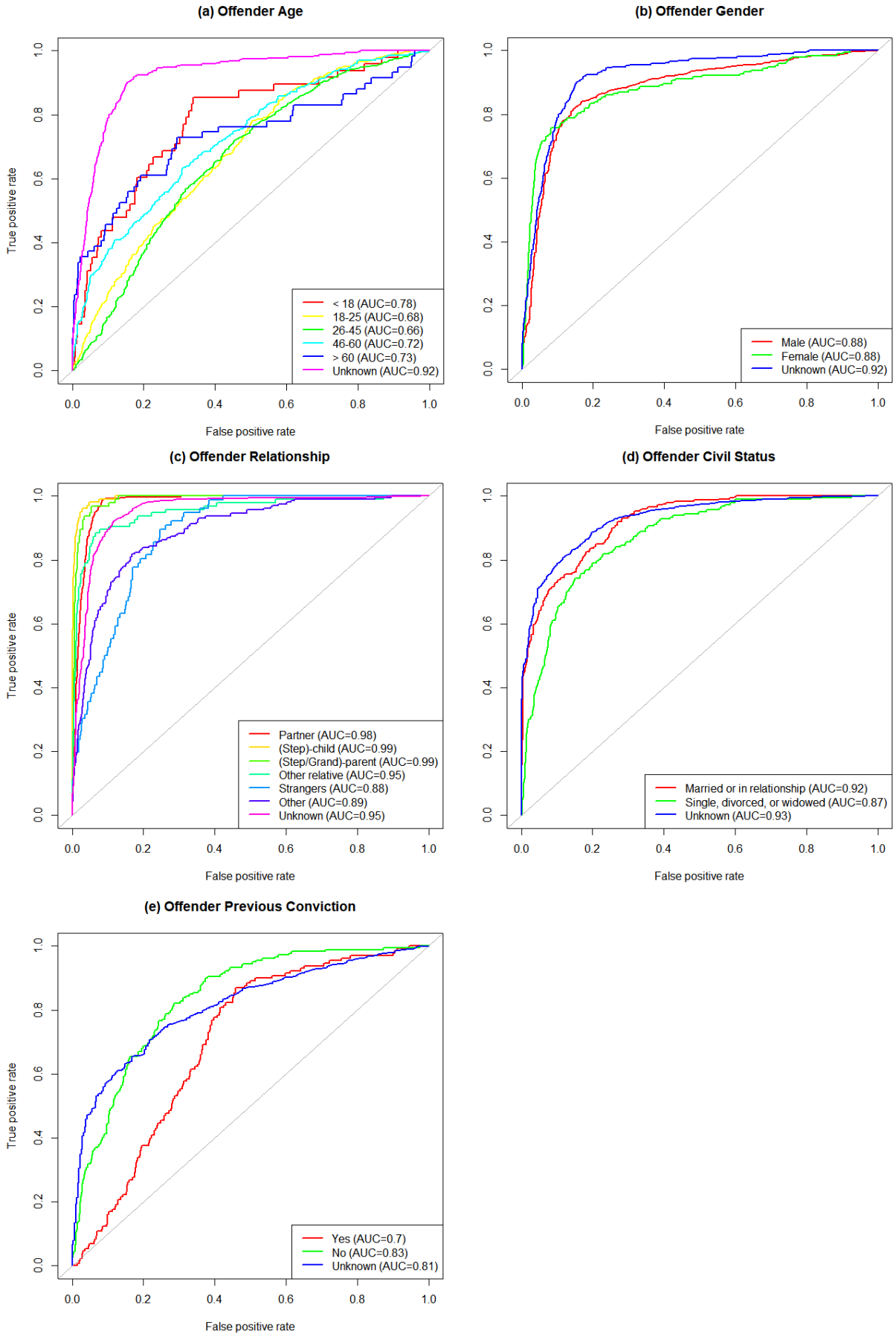
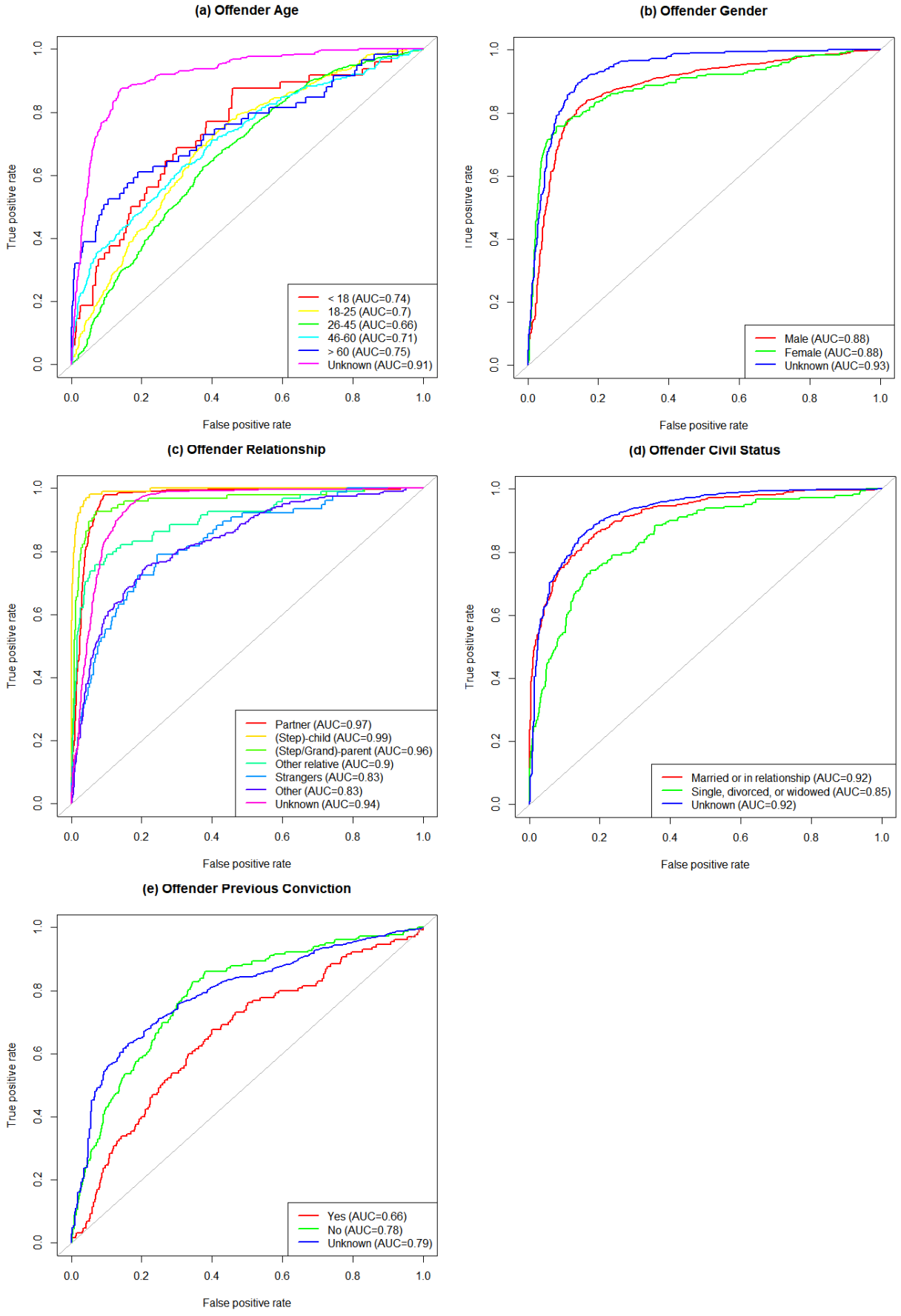


Figure 15: ROC curves for the tree-augmented naïve Bayesian classifier of offender variables (a) Age, (b) Gender, (c) Relationship with the victim, (d) Civil Status, and (e) Previous Convictions



The model performance of the combined model was compared with the performances of the naïve and tree-augmented naïve networks. The ROC curves for these networks are respectively shown in Figure 14 and Figure 15. As can be seen, the naïve Bayesian classifier performs slightly better than the tree-augmented naïve classifier in terms of AUC's. Moreover, the combined model seems to be better in predicting the age and gender of the offender, the two variables that have the least unknown values, but for all other variables, the naïve and tree-augmented naïve classifiers outperform the combined model.

7. Application

To demonstrate how the Bayesian networks could be used for criminal profiling during a criminal investigation, an example is given here. GeNIe (BayesFusion, LLC, 2017) is a graphical user interface that has a more user friendly graphical representation for Bayesian networks than R (R Core Team, 2013), so the conditional probability tables of the combined network were implemented in a network constructed in GeNIe. This network is shown in Figure 16 and is exactly the same network as the one in Figure 10, namely the combined model, but here the marginal probabilities of the nodes are shown. The offender characteristics are again highlighted in light blue.

Of course, all eight separate networks shown above could be used for prediction, but the combined network is chosen here to illustrate the example since this model had the lowest prediction errors for "Offender Age" and "Offender Gender", two variables that could help the police narrowing down their pool of suspects early in the investigation process. Consider the following case.

On Sunday July 6, the body of a 31-year old man was found in his home by his wife. She came home around 3.00 am from a night out with friends. The door was open when she arrived, so she called her husband, but no response came. She walked in and found her husband lying on the kitchen floor in a puddle of blood. As a nurse, she immediately checked his pulse, and unfortunately had to determine that he was already dead. The forensic pathologist estimated the time of death between 0.00 and 02.00 am. The victim was stabbed multiple times by a sharp object, probably one of the knifes out of his own kitchen since one was missing. After searching the house, the wife missed two laptops, some jewellery, and the cash money they kept in the top drawer of the closet in their living room. The couple did not have any children, the victim worked in the same hospital as his wife as a doctor, and because of the circumstances at the crime scene the criminal investigators classified the offence as a robbery that lead to manslaughter.

If the circumstances of this homicide are instantiated in the combined Bayesian network (see the instantiation of the white nodes in Figure 17), the model predicts probabilities for the offender characteristics (the blue nodes in Figure 17). The age of the offender is estimated to be between 26 and 45, and his gender is likely to be male. The relationship of the victim with the offender is predicted to be of the category "Unknown", but it would be also possible that the victim and offender are strangers. Further, the offender is probably in a relationship without having children. The rest of the offender

characteristics do not give a direction to the criminal investigation, since they are predicted to be unknown. Nevertheless, the predictions of the Bayesian network could give the criminal investigators some guidelines during their search for the offender.

Taking into account a threshold of 0.7 as discussed in Section 6, the model gives police investigators directions to search for a male suspect that is in a relationship and does not have children. Besides, the information of the suspect about his psychological history, alcohol and/or drug abuse, violent history, previous convictions, and professional status is likely to remain unknown. The model is not as certain about the age category of the offender, although it is approximately four times more likely that the offender falls in the age category 26-45 (60%) than that he falls in the second most likely age category (18-25; 16%). The model thus not only gives predictions, but also the probabilities of those predictions. This could be useful for criminal investigators, since they can take into account the amount of (un)certainty about one of the predicted offender characteristics.

Figure 16: Marginal probabilities of the combined network

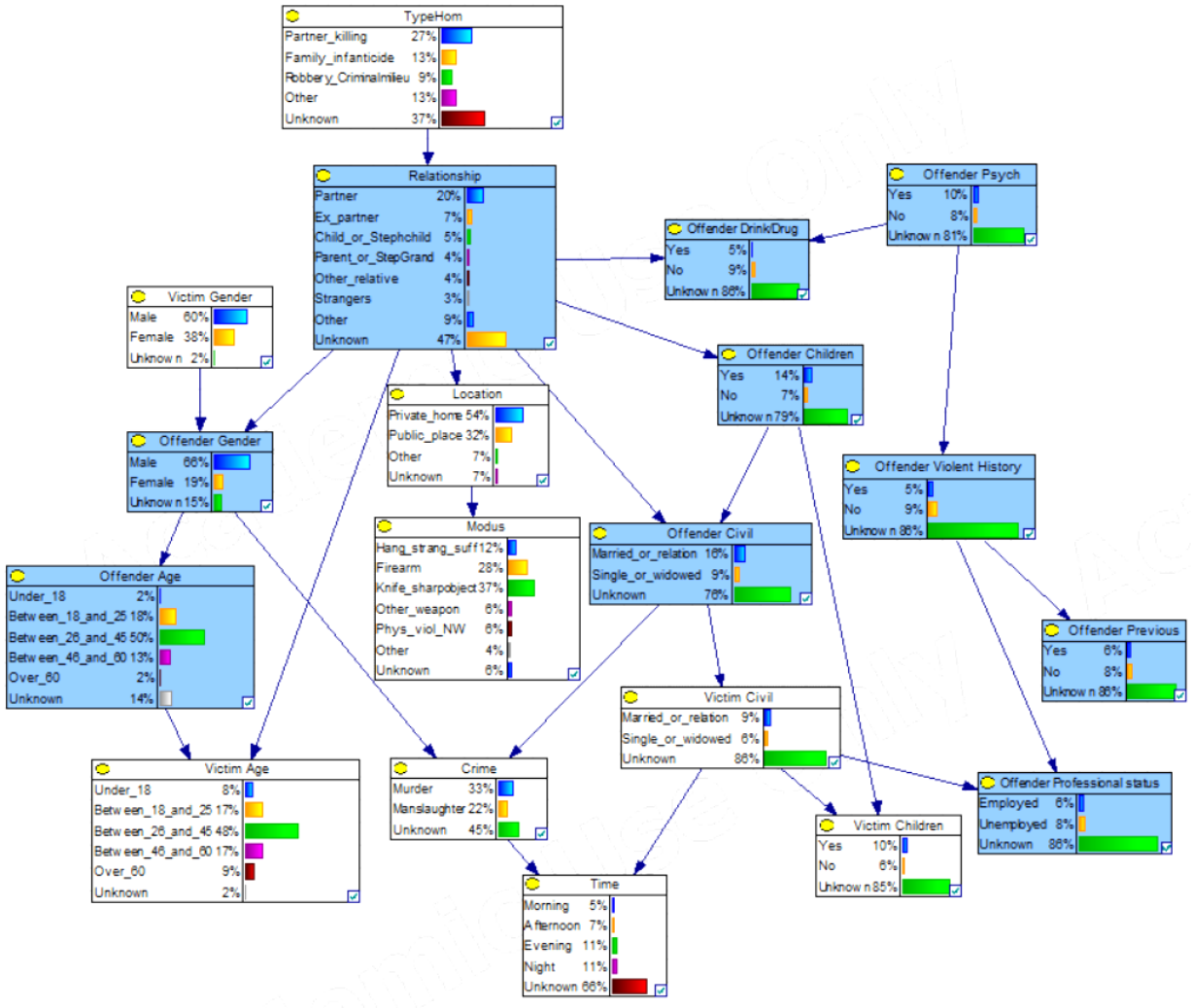
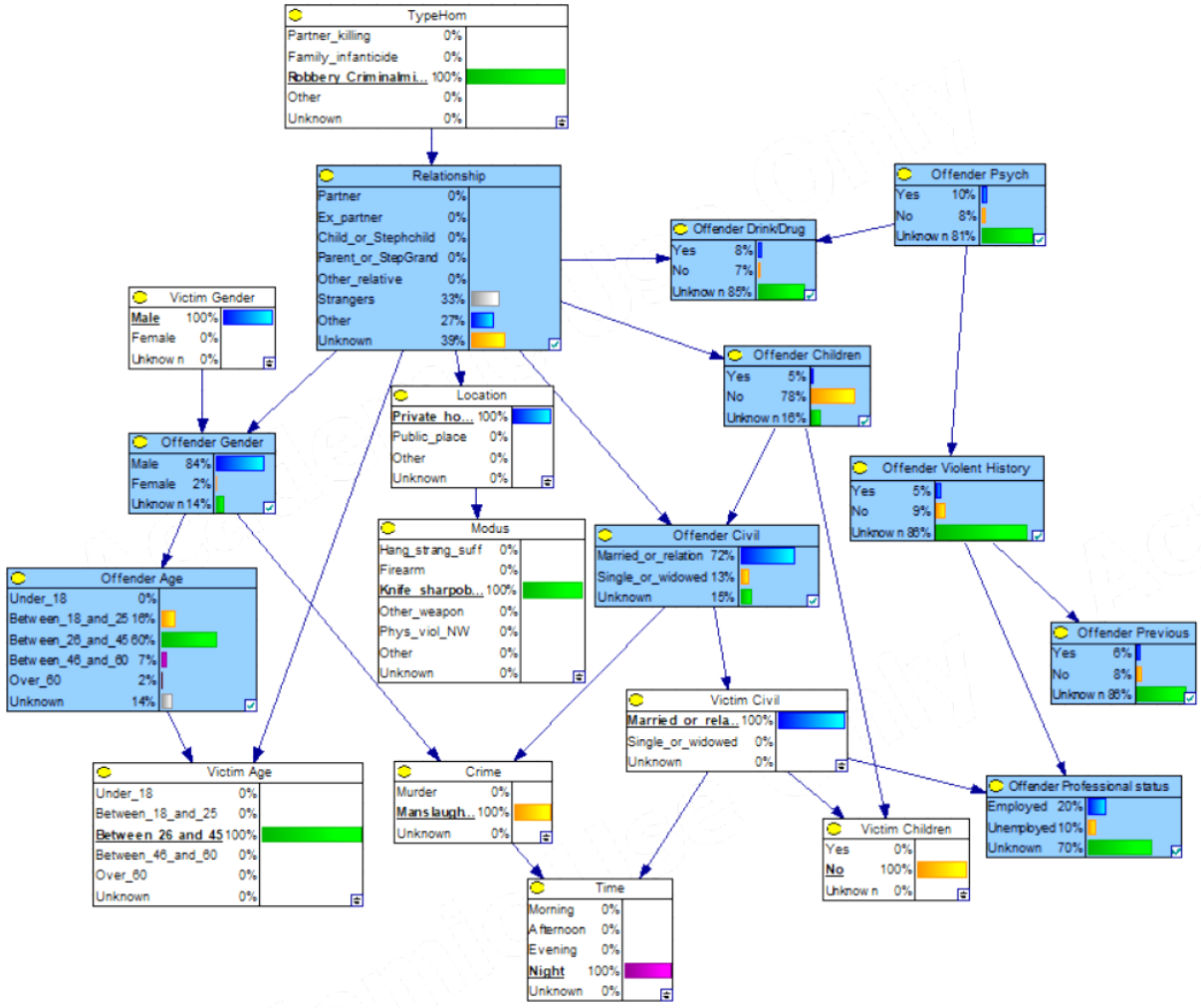


Figure 15: Posterior probabilities of the combined network after initiating case and victim characteristics



8. Conclusion

The research question – "To what extent are homicide offender characteristics predictable using a Bayesian network learned from data about solved single-victim-single-offender homicides in the Netherlands from 1992-2016?" – does not have a straightforward answer. The ROC curves of the combined model for the offender characteristics age, gender, relationship with the victim, civil status, and previous convictions imply a variable model performance, yet with a proper threshold for the minimum probability of the predicted value the performance of the combined model in terms of prediction errors looks promising. In comparison to the naïve and tree-augmented naïve Bayesian classifiers, the combined model performs better on average in terms of prediction errors. When the models are compared based on the ROC curves, the combined model performs best on two of the most important offender characteristics – age and gender (the two variables which have the least unknown values in the dataset) – but is outperformed by both classifiers on the other offender characteristics.

This research should be seen as a first attempt in the Netherlands to predict homicide offender characteristics from crime and victim characteristics using a Bayesian network; the model that was built forms a promising start, but could be improved and updated with more complete data for

example. The field of statistical criminal profiling is relatively young, especially in the Netherlands, so this research has given some useful insights in the usability of a Bayesian network in criminal profiling, and serves as a good starting point for future research.

An analysis tool was developed that might be helpful during the criminal investigation process, by building a Bayesian network that predicts characteristics of homicide offenders based on crime and victim characteristics. Bayesian network learning was applied in the context of criminal profiling, and similar to the work of Stahlschmidt et al. (2013) the resulting network – with the threshold for the minimum probability of the fitted value set to 0.7, which lowers the prediction error rate to 10 percent – is potentially implementable in the criminal investigation process. Besides, the network gives a helpful view on the dependence relationships between crime, victim, and offender characteristics; existing theories could be confirmed and the network could give inspiration for new theoretical considerations.

With regard to the different Bayesian network structure learning algorithms that were applied, it could be concluded that in terms of prediction error it does not make a large difference which learning algorithm (and within a specific algorithm which conditional independence test or score) is used. This goes for constraint-based as well as score-based and hybrid learning algorithms. However, some small distinctions could be made, for example this research has shown that the K2 algorithm, which was used in the comparable study of Baumgartner et al. (2008), is not the most suitable algorithm for this dataset if it is evaluated in terms of prediction errors. Further, the performance of the different learning algorithms depended quite much on which variable was predicted, so perhaps the choice for a specific learning algorithm should be made based on which offender characteristic one wants to predict. Nevertheless, the combination of the eight different structure learning algorithms lead to a combined model that performs above average for each of the offender characteristics. Therefore this combined model is considered to be the best option in predicting all homicide offender variables at a time. This advantage of predicting multiple offender characteristics all at once namely was the reason for choosing a Bayesian network approach in the first place.

9. Discussion

However this research could serve as a good starting point for criminal profiling using Bayesian networks in the Netherlands, there are some points for improvement, which are subject for future research at once. First, the number of structure learning algorithms is growing rapidly since this field of research is relatively new and develops quickly. For convenience reasons eight different structure learning algorithms which were available within the same R package, *bnlearn*, were used to construct the combined Bayesian network. However, as amongst others Gasse, Aussem, and Elghazel (2014) showed, there are more recently developed structure learning algorithms that outperform the algorithms that were used here, both in terms of goodness of fit to new data and in terms of closeness to the true dependence structure of the data. And for example Kelner and Lerner (2012) concluded that

the Bayesian network classifier learning algorithm they proposed, which uses risk minimisation cross validation with the 0/1 loss function, outperforms naïve Bayes and tree-augmented naïve Bayes in terms of prediction accuracy. Also De Campos, Corani, Scanagatta, Cuccu, and Zaffalon (2015) argued that their extension of tree-augmented naïve Bayes results in a higher prediction accuracy than naïve Bayes and tree-augmented naïve Bayes. Moreover, new parametrisation techniques are being proposed (see for example Zaidi et al., 2017; Zhou, Fenton, and Zhu, 2016). It would be interesting to compare the structures and performances in terms of prediction errors of more recently developed structure learning and (tree-augmented) naïve Bayes classifier algorithms and different parameter learning methods with those used in this research.

Second, the quality of the dataset could be improved. The Institute of Security and Global Affairs at Leiden University is still busy merging information from different sources, and completing every case as much as possible. Because of the large amount of missing values, some interesting variables were left out of this analysis. For example, it could be interesting to add information about whether or not there were witnesses at the crime scene, the motive of the offender, the educational level of the victim and/or offender, the number of stabs or even the location of the stabbing wounds in case a knife or sharp object was used, or the type of firearm if one was used. Moreover, much research is done to geographical criminal profiling because it could narrow down the search area for the police, adding information about the living area of the victim and offender could be interesting in future research.

Third, a different approach to the missing values could be considered. In this research, the category "Unknown" was added to each variable. However, this could lead to predictions that are not useful in the criminal investigation process; it does not help a police officer if the relationship between the victim and offender is predicted to be unknown. This choice was made because every homicide case is different and imputing missing values was considered undesirable. Notwithstanding, future research could compare the current model with a model in which the parameters are estimated using Expectation Maximisation to handle the missing values, as discussed in Section 3.3.

Fourth, some restrictions of the model should be taken into account. Baskin and Sommers (2011) found that most homicides remain unsolved and that suspects who knew their victims are more likely to be arrested. Since this model is merely based on solved homicides and most offenders in this dataset knew their victims, the model could be biased. However, Lammers (2014) concluded that 'arrest data are probably less selective than has been suspected in the past', so perhaps the model is not so much biased. But it is still possible that the model provides worse predictions for the more complicated cases, since they are relatively less represented among the data on which the model was based. The consequences hereof should be investigated, since as mentioned in Section 2, criminal profiling is often used when police investigators believe that there is no connection between victim and offender. Besides, the model is only based on single-victim-single-offender cases. However if a homicide was committed, the number of offenders is not always clear, so the model might be applied

in a case which later appears to be a multi-offender case. Bell Holleran and Vandiver (2016) found that single-offender homicides differ from multi-offender homicides, so it could be interesting to incorporate the latter in the model as well.

Lastly, as Stahlschmidt et al. (2013) already suggested, an online learning procedure could be developed, such that new cases are entered into the model automatically and the model thus adjusts to the newly added information.

References

Adeyiga, J. A., & Bello, A. O. (2016). A review of different clustering techniques in criminal profiling. *International Journal of Advanced Research in Computer Science and Software Engineering, 6*(4), 659-666.

Ainsworth, P. B. (2001). *Offender Profiling & Crime Analysis*. Oregon, United States: Willan Publishing.

Aitken, C. G. G. (2006). Statistics in forensic science. Part I: An aid to investigation. *Problems of Forensic Sciences, 65*, 53-67.

Aitken, C. G. G., Connolly, T., Gammerman, A., Zhang, G., Bailey, D., Gordon, R., & Oldfield, R. (1996). Statistical modeling in specific case analysis. *Science & Justice, 36*(4), 245-255.

Aliferis, F. C., Statnikov, A., Tsamardinos, I., Subramani, M., & Koutsoukos, X. D. (2010). Local causal and Markov blanket induction for causal discovery and feature selection for classification Part I: Algorithms and empirical evaluation. *Journal of Machine Learning Research, 11*(1), 171-234.

Aydin, F., & Dirilen-Gumus, O. (2011). Development of a criminal profiling instrument. *Procedia - Social and Behavioral Sciences, 30*(2011), 2612-2616.

Baumgartner, K., Ferrari, S., & Palermo, G. (2008). Constructing Bayesian networks for criminal profiling from limited data. *Knowledge-Based Systems, 21*(7), 563-572.

BayesFusion, LLC (2017). GeNIe version 2.2 Academic [Computer software]. Retrieved from https://download.bayesfusion.com/files.html?category=Business

Beauregard, E., & Proulx, J. (2002). Profiles in the offending process of nonserial sexual murderers. *International Journal of Offender Therapy and Comparative Criminology, 46*(4), 386-399.

Bell Holleran, L. L., & Vandiver, D. M. (2016). US homicides: Multi-offenders and the presence of female offenders. *Violence and gender, 3*(1), 27-35.

Ben-Gal, I. (2008). Bayesian networks. In F. Ruggeri, R. S. Kenett, & F. Faltin (Eds), *Encyclopedia of statistics in quality and reliability*. Oxford, United Kindom: Wiley & Sons.

Bennell, C., Jones, N. J., Taylor, P. J., & Snook, B. (2006). Validities and abilities in criminal profiling: A critique of the studies conducted by Richard Kocsis and his colleagues. *International Journal of Offender Therapy and Comparative Criminology, 50*(3), 344-360.

Beretta, S., Castelli, M., Gonçalves, I., & Ramazzotti, D. (2017). *A quantitative assessment of the effect of different algorithmic schemes to the task of learning the structure of Bayesian networks*. Manuscript submitted for publication. Retrieved from https://arxiv.org/pdf/1704.08676.pdf

Bockhorst, J., Craven, M., Page, D., Shavlik, J., & Glasner, J. (2003). A Bayesian network approach to operon prediction. *Bioinformatics, 109*(10), 1227-1235.

Bosco, D., Zappalà, A., & Santtila, P. (2010). The admissibility of offender profiling in courtroom: A review of legal issues and court opinions. *International Journal of Law and Psychiatry, 33*(3), 184-191.

Brahan, J. W., Lam, K. P. Chan, H., & Lcung, W. (1998). AICAMS: Artificial intelligence crime analysis and management system. *Knowledge-Based Systems, 11*, 355-361.

Briggs, S. G. (2015). Computerized criminal profiling: More research is needed. *UC Merced Undergraduate Research Journal, 8*(1). Retrieved from https://cloudfront.escholarship.org/dist/prd/content/qt50d3993q/qt50d3993q.pdf

Burgess, A. W., Hartman, C. R., Ressler, R. K., Douglas, J. E., & McCormack, A. (1986). Sexual homicide: A motivational model. *Journal of Interpersonal Violence, 1*(3), 251-272.

Canter, D., & Larkin, P. (1993). The environmental range of the serial rapists. *Journal of Environmental Psychology, 13*(1), 63-69.

Cao, L., Hou, C., & Huang, B. (2008). Correlates of the victim-offender relationship in homicide. *International Journal of Offender Therapy and Comparative Criminology, 52*(6), 658-672.

Chifflet, P. (2015). Questioning the validity of criminal profiling: An evidence-based approach. *Australian & New Zealand Journal of Criminology, 48*(2), 238-255.

Chow, C. K., & Liu, C. N. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory, 14*(3), 462-467.

Colombo, D., & Maathuis, M. H. (2014). Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research, 15*(1), 3921-3962.

Contaldi, C., Vafaee, F., & Nelson, P. C. (2017). The role of crossover operator in Bayesian network structure learning performance: A comprehensive comparative study and new insights. *GECCO'17 Proceedings of the Genetic and Evolutionary Computation Conference*, 769-776.

Cooper, G. F., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning, 9*(4), 309-347.

Daéid, N. N. (1997). Differences in offender profiling in the United States of America and the United Kingdom. *Forensic Science International, 90*, 25-31.

Davies, A., Wittebrood, K., & Jackson, J. L. (1997). Predicting the criminal antecedents of a stranger rapist from his offence behaviour. *Science & Justice, 37*(3), 161-179.

De Campos, C. P., Corani, G., Scanagatta, M., Cuccu, M., & Zaffalon, M. (2015). Learning extended tree augmented naïve structures. *International Journal of Approximate Reasoning, 68*, 153-163.

Dor, D., & Tarsi, M. (1992). A simple algorithm to construct a consistent extension of a partially oriented graph.

Dowden, C., Bennell, C., & Bloomfield, S. (2007). Advances in offender profiling: A systematic review of the profiling literature published over the past three decades. *Journal of Police and Criminal Psychology, 22*(1), 44-56.

FBI Behavioral Analysis Jobs. (2017). Retrieved from http://www.fbiagentedu.org/careers/intelligence/fbi-behavioral-analyst/

Fenton, N., & Neil, M. (2013). *Risk assessment and decision analysis with Bayesian networks*. Boca Raton, United States: CRC Press.

Flesch I., Lucas P. J. (2007). Markov Equivalence in Bayesian Networks. In P. Lucas, J. A. Gámez, & A. Salmerón (Eds), *Advances in Probabilistic Graphical Models* (pp. 3-38). Studies in Fuzziness and Soft Computing (214). Berlin, Germany: Springer.

Francis, B., Barry, J., Bowater, R., Miller, N., Soothill, K., & Ackerley, E. (2004). *Using homicide data to assist murder investigations* (Online Report 16/04). London, England: Home Office Research, Development and Statistics Directorate.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning, 29*(2), 131-163.

Ganpat, S., Granath, S., Hagstedt, J., Kivivuori, J., Lehti, M., Liem, M., & Nieuwbeerta, P. (2011). *Homicide in Finland, the Netherlands and Sweden: A first study on the European Homicide Monitor Data*. Stockholm, Sweden: Brottsförebyggande rådet/The Swedish National Council for Crime Prevention.

Ganpat S. M., & Liem M. C. A. (2012). Homicide in the Netherlands. In M. C. A. Liem, & W. A. Pridemore (Eds.), *Handbook of European homicide research: Patterns, explanations, and country studies* (pp. 329-342). New York, United States: Springer.

Gasse, M., Aussem, A., & Elghazel, H. (2014). A hybrid algorithm for Bayesian networks structure learning with application to multi-label learning. *Expert systems with Applications, 41*(15), 6755-6722.

Gottschalk, P. (2006). Stages of knowledge management systems in police investigations. *Knowledge-Based Systems, 19*(6), 381-387.

Kelner, R., & Lerner, B. (2012). Learning Bayesian network classifiers by risk minimization. *International Journal of Approximate Reasoning, 53*(2), 248-272.

Kocsis, R. N. (2003). Criminal psychological profiling: Validities and abilities. *International Journal of Offender Therapy and Comparative Criminology, 47*(2), 126-144.

Kocsis, R. N. (2006). Validities and abilities in criminal profiling: The dilemma for David Canter's Investigative Psychology. *International Journal of Offender Therapy and Comparative Criminology, 50*(4), 458-477.

Korb, K., & Nicholson, A. (2011). *Bayesian artificial intelligence*. Boca Raton, United States: CRC Press.

Kreimer, A., & Herman, M. (2016). A novel structure learning algorithm for optimal Bayesian network: Best parents. *Procedia Computer Science, 96*, 43-52.

Lamine, F. B., Kalti, K., & Mahjoub, M. A. (2011). The threshold EM algorithm for parameter learning in Bayesian network with incomplete data. International Journal of Advanced Computer Science and Applications, 2(7), 86-91.

Lammers, M. (2014). Are arrested and non-arrested serial offenders different? A test of spatial offending patterns using DNA found at crime scenes. *Journal of Research in Crime and Delinquency, 51*(2), 143-167.

Leegon, J. F. (2009). A comparison of Bayesian network structure learning algorithms on emergency department ambulance diversion data (Master's thesis, Vanderbilt University, Nashville, United States). Retrieved from http://etd.library.vanderbilt.edu/available/etd-07242009-135048/unrestricted/Leegon2009ThesisFinal.pdf

Li, G., Xing, L., Zhang, Z., & Chen, Y. (2017). A new Bayesian network structure learning algorithm mechanism based on the decomposability of scoring functions. *IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences, 100*(7), 1541-1551.

Liem, M. C. A., Alink, L. R. A., Aarten, P. G. M., & Schönberger, H. J. M (2018). *Dutch Homicide Monitor* (January 2018) [Data file, codebook, and background information]. Den Haag, the Netherlands: Leiden University.

Liu, H., Zhou, S., Lam, W., & Guan, J. (2017). A new hybrid method for learning Bayesian networks: Separation and reunion. *Knowledge-Based Systems, 121*, 185-197.

Marcot, B. C. (2012). Metrics for evaluating performance and uncertainty of Bayesian network models. *Ecological Modelling, 230*, 50–62

Margaritis, D. (2003). *Learning Bayesian network model structure from data* (Master's thesis, Carnegie Mellon University, Pittsburgh, United States). Retrieved from https://www.cs.cmu.edu/~dmarg/Papers/PhD-Thesis-Margaritis.pdf

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., & Califano, A. (2006). ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics, 7*(Suppl 1): S7.

Mears, D. P., & Bacon, S. (2009). Improving criminal justice through better decision making: Lessons from the medical system. *Journal of Criminal Justice, 37*(2), 142-154.

Morelato, M., Beavis, A., Tahtouh, M., Ribaux, O., Kirkbride, P., & Roux, C. (2013). The use of forensic case data in intelligence-led policing: The example of drug profiling. *Forensic Science International, 226*, 1-9.

Neapolitan, R. E. (2004). *Learning Bayesian Networks*. Upper Saddle River, United States: Pearson Prentice Hall.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco, United States: Morgan Kaufmann Publishers.

R Core Team (2013). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Ressler, R. K., Burgess, A.W., & Douglas, J. E. (1988). *Sexual homicide: Patterns and motives*. New York, United States: Free Press.

Ressler, R. K., Burgess, A. W., Douglas, J. E., Hartman, C. R., & D'Agostino, R. B. (1986). Sexual killers and their victims: Identifying patterns through crime scene analysis. *Journal of Interpersonal Violence, 1*(3), 288-308.

Ressler, R. K., Burgess, A.W., Hartman, C. R., Douglas, J. E., & McCormack, A. (1986). Murderers who rape and mutilate. *Journal of Interpersonal Violence, 1*(3), 273-287.

Rossmo, D. K. (1993). A methodological model. *American Journal of Criminal Justice, 17*(2), 1-21.

Salfati, C. G., & Canter, D. V. (1999). Differentiating stranger murders: Profiling offender characteristics from behavioral styles. *Behavioral Sciences and the Law, 17*(3), 391-406.

Salfati, C. G., & Park, J. (2007). An analysis of Korean homicide crime-scene actions. *Journal of Interpersonal Violence, 22*(11), 1448-1470.

Scutari, M. (2010). Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software, 35*(3), 1-22.

Scutari, M. (2017). Bayesian network constraint-based structure learning algorithms: Parallel and optimized implementations in the bnlearn R package. *Journal of Statistical Software, 77*(2), 1-20.

Scutari, M., & Denis, J-B. (2015). *Bayesian Networks: With Examples in R.* Boca Raton, United States: CRC Press.

Scutari, M., & Ness, R. (2018). bnlearn: Bayesian Network Structure Learning, Parameter Learning and Inference. R package version 4.3.

Stahlschmidt, S., Tausendteufel, H., & Härdle, W. K. (2013). Bayesian networks for sex-related homicides: Structure learning and prediction. *Journal of Applied Statistics, 40*(6), 1155-1171.

Strano, M. (2004). A neural network applied to criminal psychological profiling: An Italian initiative. *International Journal of Offender Therapy and Comparative Criminology, 48*(4), 495-503.

Sturup, J., Karlberg, D., & Kristiansson, M. (2015). Unsolved homicides in Sweden: A population-based study of 264 homicides. *Forensic Science International, 257*, 106-113.

Su, J., Zhang, H., Ling, C. X., & Matwin, S. (2008). Discriminative parameter learning for Bayesian networks. Proceedings of the 25th International Conference on Machine Learning, 1016-1023.

Taroni, F., Biedermann, A., Bozza, S., Garbolino, P, & Aitken, C. (2010). *Bayesian Networks for Probabilistic Inference and Decision Analysis in Forensic Science.* Oxford, United Kindom: Wiley & Sons.

Ter Beek, M., Van den Eshof, P., & Mali, B. (2010). Statistical modelling in the investigation of stranger rape. *Journal of Investigative Psychology and Offender Profiling, 7*(1), 31-47.

Tsamardinos, I., Aliferis, C. F., & Statnikov, A. (2003a). Algorithms for large scale Markov blanket discovery. *Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference* (FAIRS), 376-381.

Tsamardinos, I., Aliferis, C. F., & Statnikov, A. (2003b). Time and sample efficient discovery of Markov blankets and direct causal relations. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 673-678.

Tsamardinos, I., Brown, L. E., & Aliferis, C. F. (2006). The max-min Hill-Climbing Bayesian network structure learning algorithm. *Machine Learning, 65*(1), 31-78.

Turvey, B. E., & Esparza, M. (2016). *Behavioral Evidence Analysis: International Forensic Practice and Protocols.* Amsterdam, the Netherlands: Elsevier Science & Technology Books.

Van den Broeck, G., Mohan, K., Choi, A., & Pearl, J. (2015). Efficient algorithms for Bayesian network parameter learning from incomplete data. Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, 161-170.

White, J. H., Lester, D., Gentile, M., & Rosenbleeth, J. (2011). The utilization of forensic science and criminal profiling for capturing serial killers. *Forensic Science International, 209*, 160-165.

Wilson, P., & Soothill, K. (1996). Psychological profiling: Red, green or amber? *Police Journal, 69*(1), 12-20.

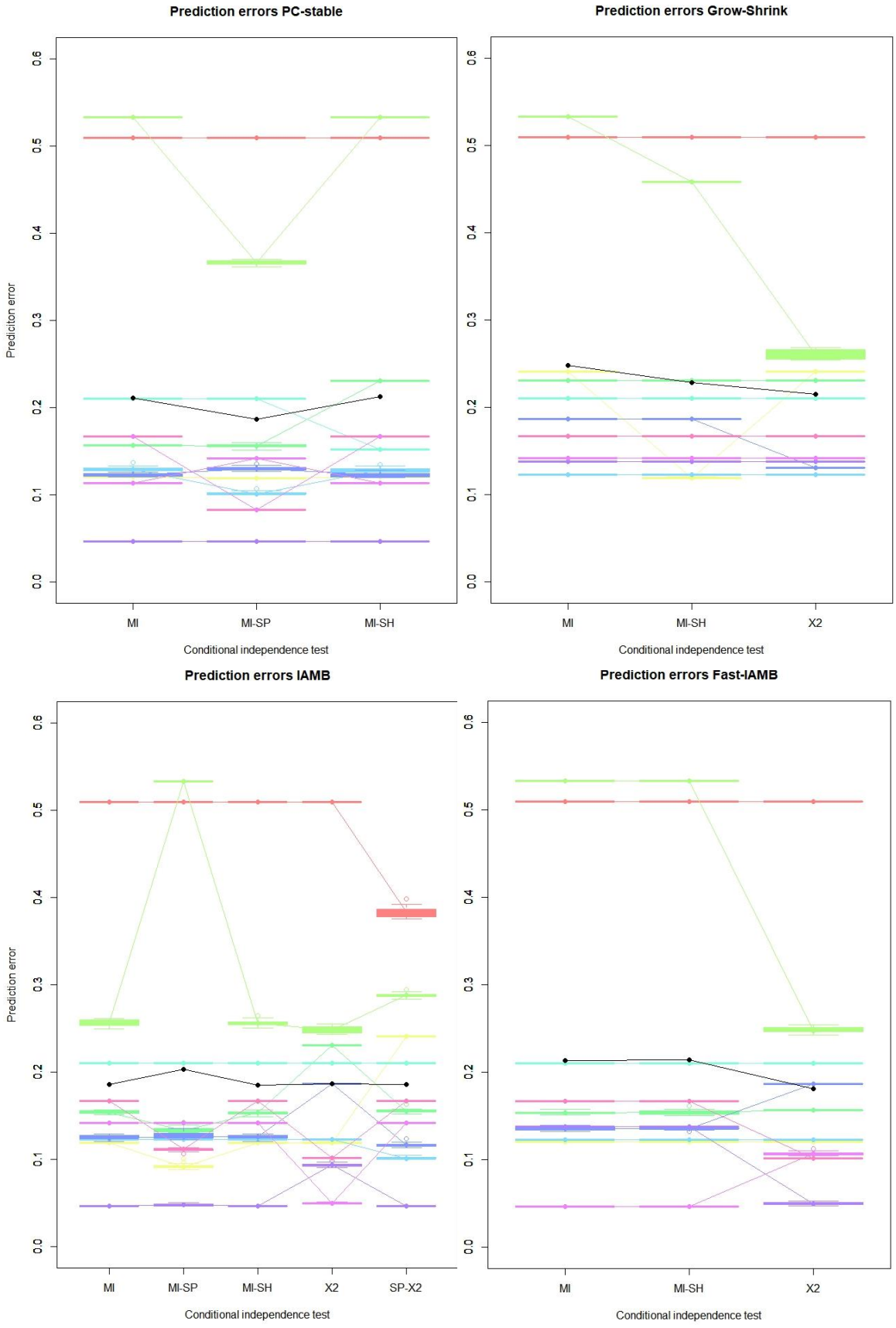
Yaramakala, S., & Margaritis, D. (2005). Speculative Markov blanket discovery for optimal feature selection. *Proceedings of the Fifth IEEE International Conference on Data Mining* (ICDM), 809-812.

Zaidi, N. A., Webb, G. I., Carman, M. J., Petitjean, F., Buntine, W., Hynes, M., & De Sterck, H. (2017). Efficient parameter learning of Bayesian network classifiers. *Machine Learning, 106*(9-10), 1289-1329.

Zhou, Y., Fenton, N., & Zhu, C. (2016). An empirical study of Bayesian network parameter learning with monotonic influence constraints. *Decision Support Systems, 87*, 69-79.

Appendix 1: Figures 3-9

Figure 3: Prediction errors of the constraint-based networks



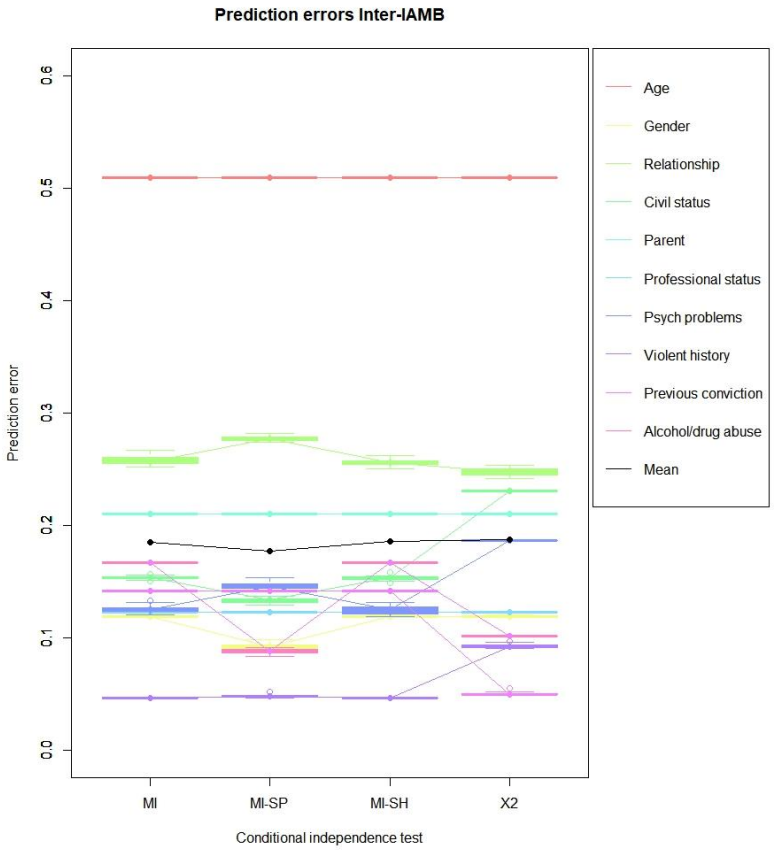

Prediction errors Inter-IAMB
Prediction error
0.0
0.1
0.2
0.3
0.4
0.5
0.6
MI
MI-SP
MI-SH
X2
Conditional independence test
Age
Gender
Relationship
Civil status
Parent
Professional status
Psych problems
Violent history
Previous conviction
Alcohol/drug abuse
Mean

Figure 4: Negative log-likelihood loss of the constraint-based networks

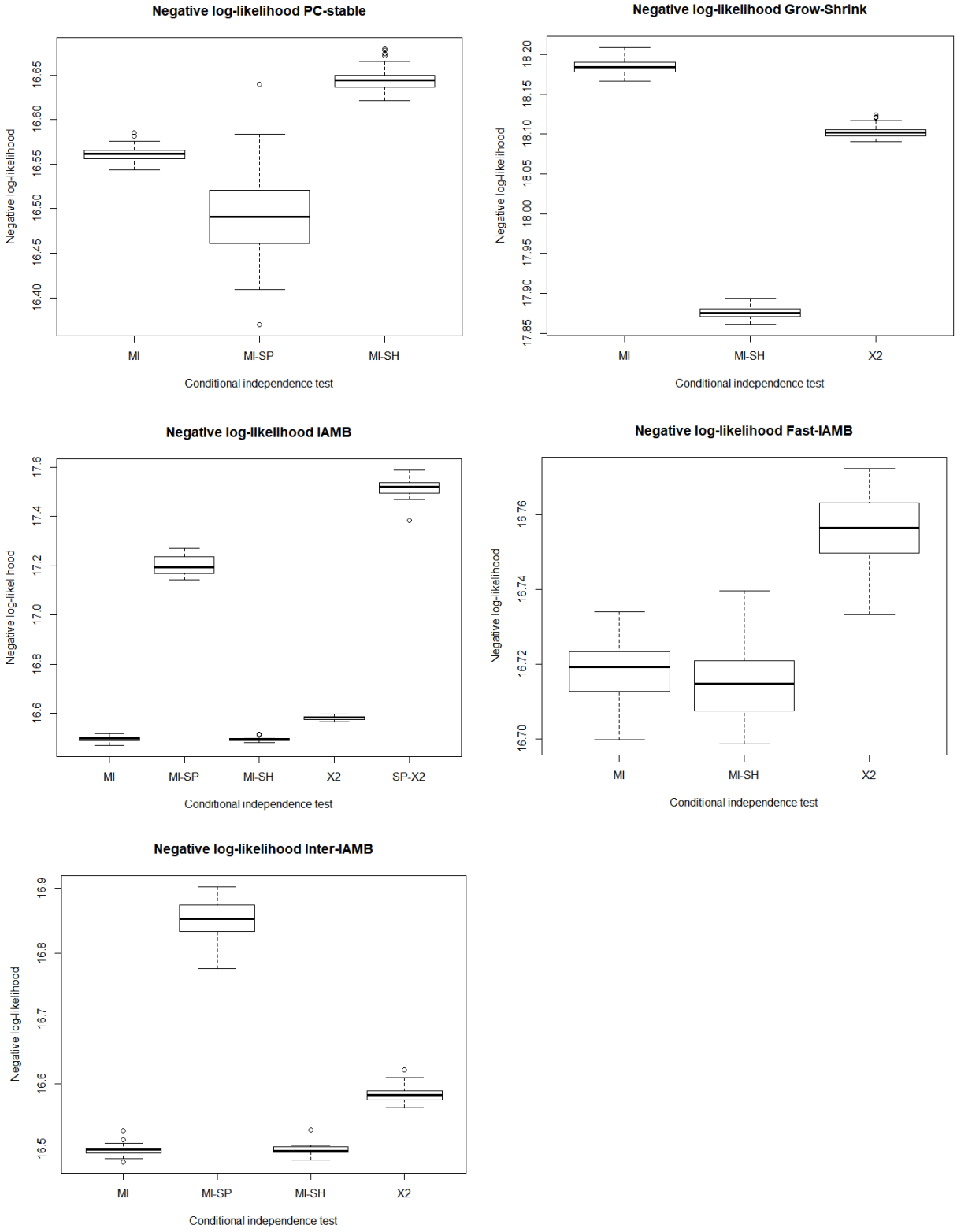


Figure 5: Prediction errors of the score-based networks

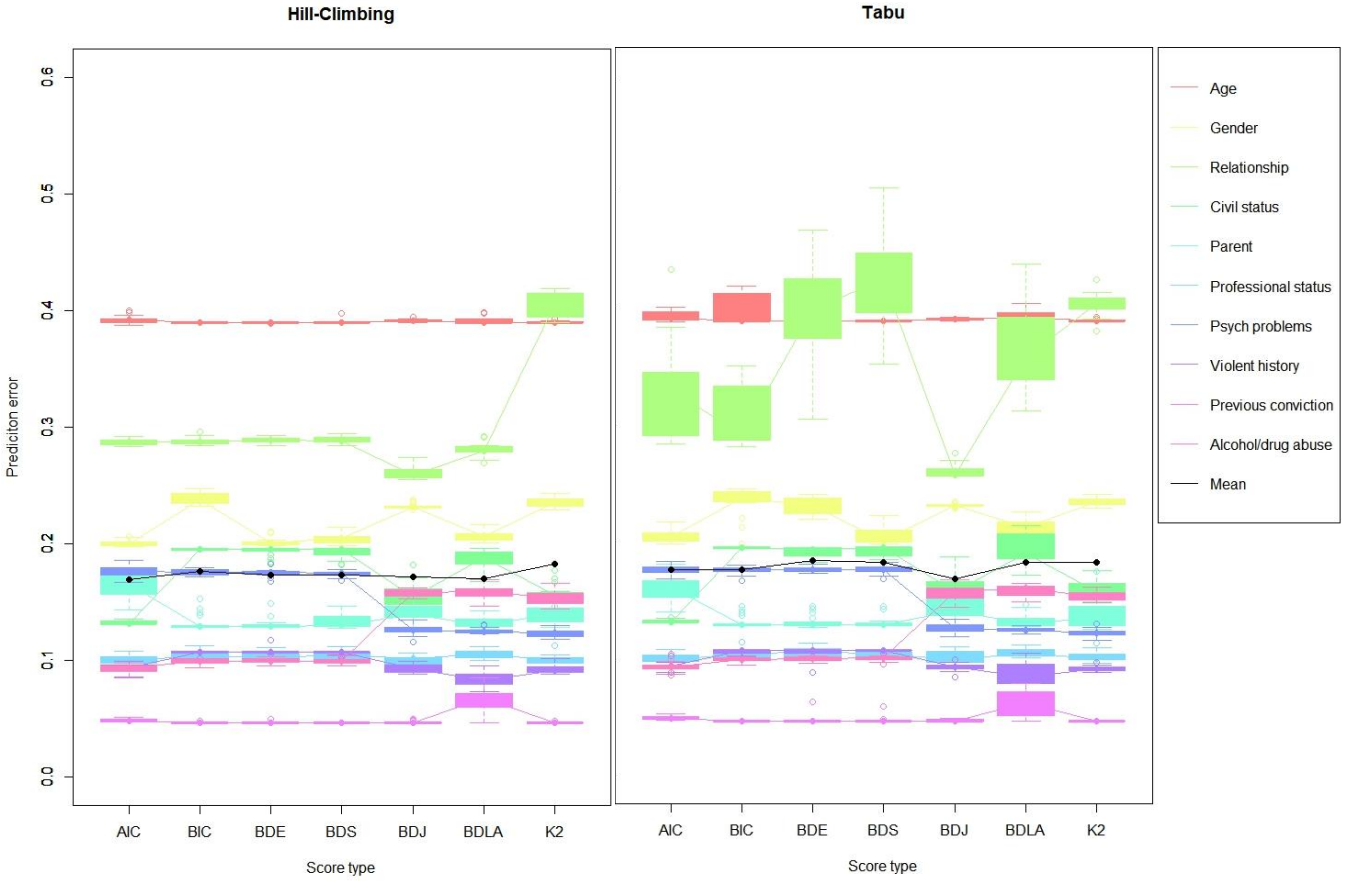


Figure 6: Negative log-likelihood loss of the score-based networks

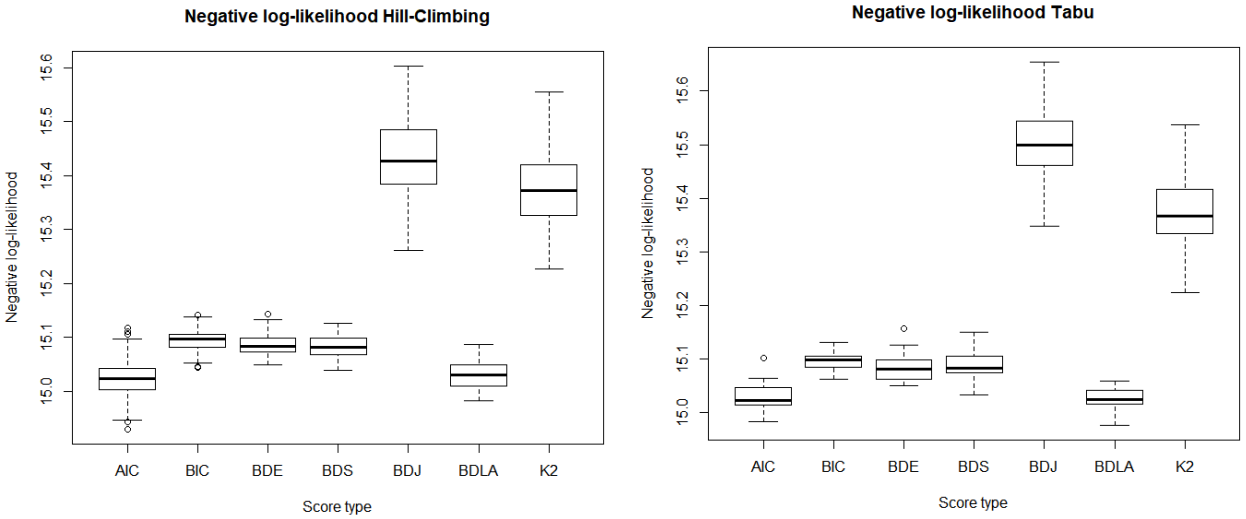


Figure 7: Prediction errors of the hybrid networks

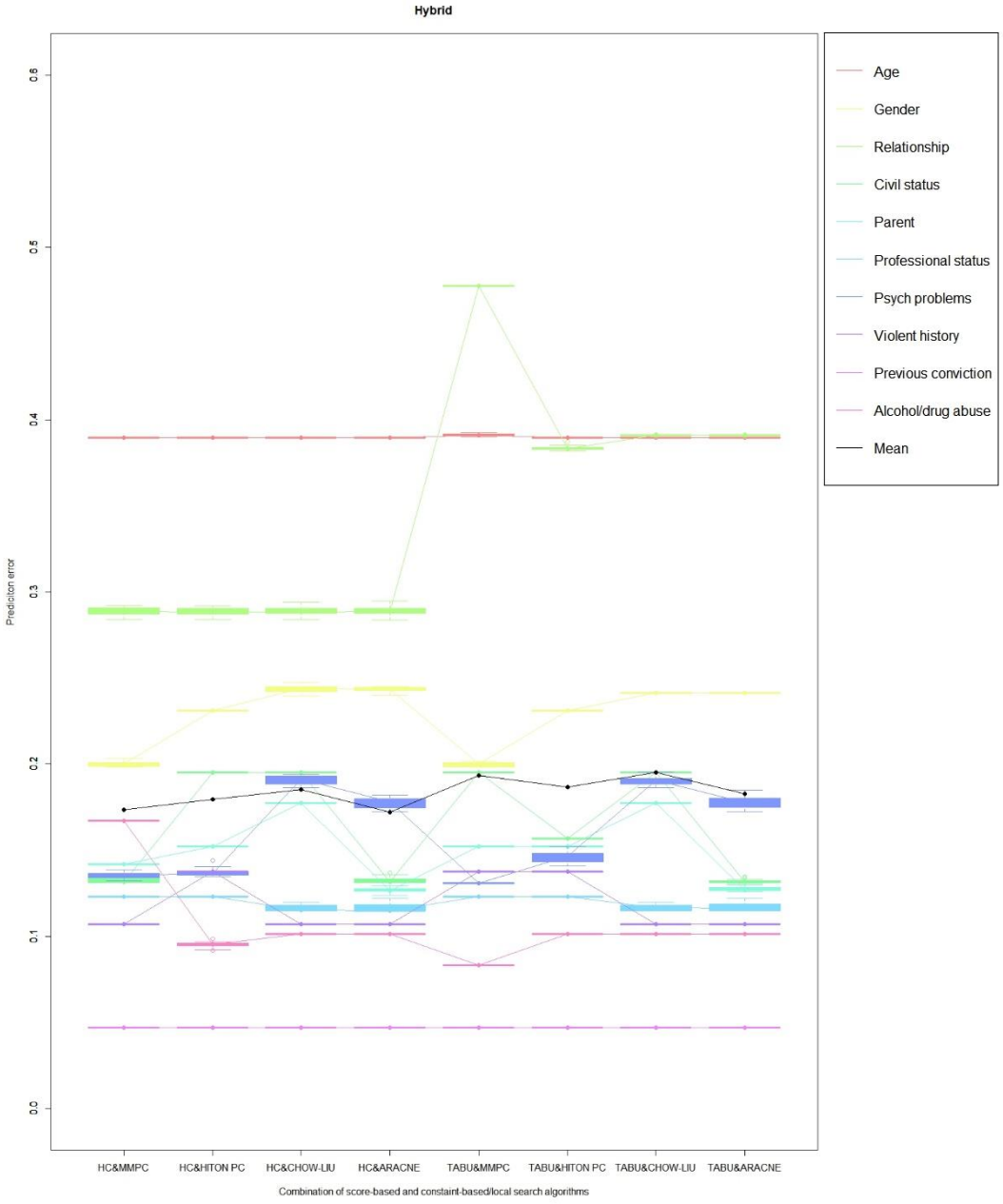


Figure 8: Negative log-likelihood of the hybrid networks

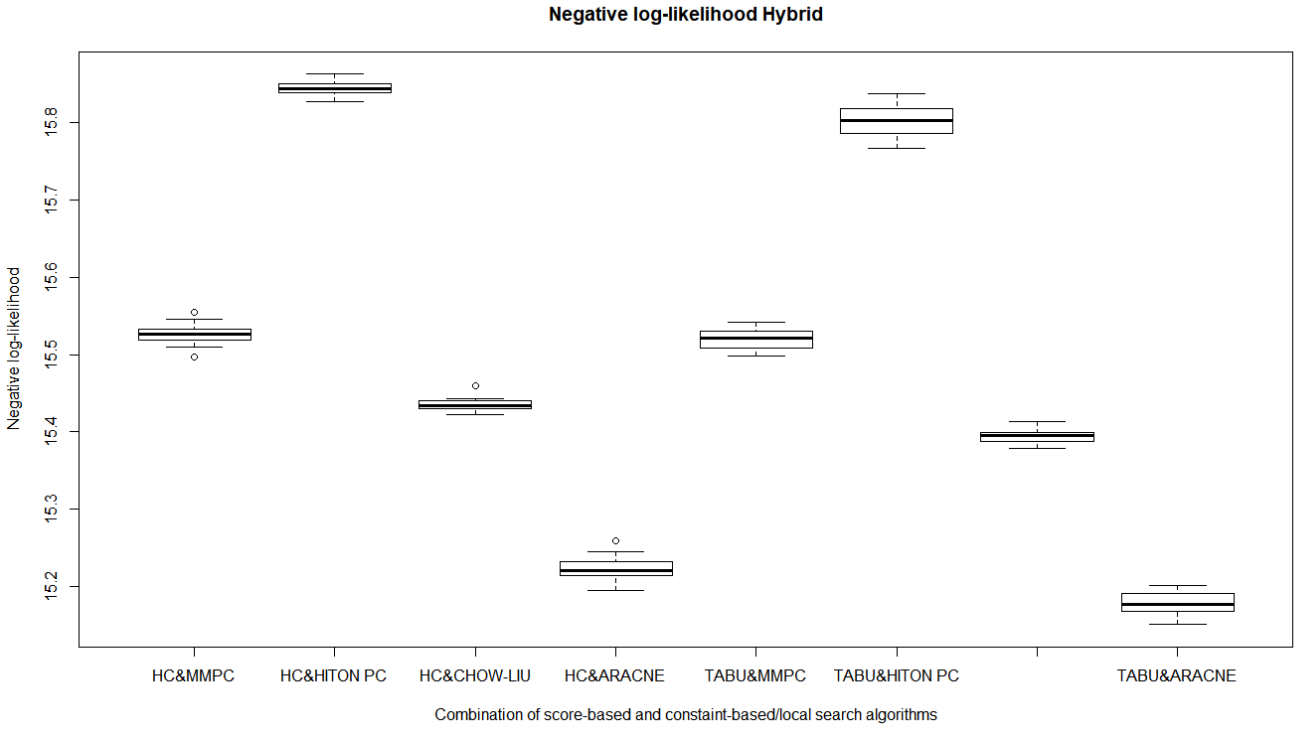


Figure 9: Negative log-likelihood of the combined network