

# Graphical Modelling and Statistical Learning for Sex-related Homicides

Masterarbeit

zur Erlangung des Grades Master of Science im gemeinsamen  
Masterstudiengang Statistik der Humboldt-Universität zu Berlin,  
Freien Universität Berlin, Technischen Universität Berlin  
und  
Charité-Universitätsmedizin Berlin

vorgelegt von

**Stephan Stahlschmidt**

Prüfer: Prof. Dr. Wolfgang K. Härdle

Berlin, 12. September 2011

---

## Abstract

We present a twofold analysis in the domain of sex-related homicides. Police profilers often help in criminal investigation of high profile cases, but empirical evidence in the domain is incomplete. We therefore at first apply a structural learning approach and secondly, try explicitly to predict the age of an unknown offender from information obtained from the crime scene.

We apply graphical modelling to obtain a factorisation of the probability function which governs the domain. This factorisation allows us to infer dependencies and independencies between the variables and therefore describes the domain. We apply several structure learning algorithms for Bayesian Networks and combine them to a final graphical model. In the second part, we compare several prediction techniques concerning their error rate in predicting the offender's age.

The graphical model broadly presents a distinction between an offender and a situation driven crime. A situation driven crime may be characterised by an offender lacking preparation and typically attacking a known victim in familiar surroundings. The offender tends to apply blunt force to gain control over the victim and does not show a high level of forensic awareness. In contrast offender driven crimes may be identified by the high level of forensic awareness demonstrated by the offender and the sophisticated measures applied to control the victim. Furthermore the graphical model indicates that these offenders are more likely to attack an unknown victim in unfamiliar surroundings and prepare their attack.

Applying several prediction techniques to the data results in a significant decrease in the root mean square error, if compared with a simple baseline model. However the actual performance of the best model, namely the lasso, is still not applicable in criminal investigation, as its average error of 8 years is too high.

---

# Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
<b>2</b>	<b>A note on Criminological Theory</b>	<b>9</b>
<b>3</b>	<b>The Data</b>	<b>11</b>
<b>4</b>	<b>Bayesian Networks</b>	<b>14</b>
4.1	Technique . . . . .	18
4.2	Structure Learning . . . . .	21
4.2.1	Algorithms . . . . .	23
4.3	Parameter Learning . . . . .	27
4.4	Implementation . . . . .	30
4.5	Results . . . . .	31
<b>5</b>	<b>Statistical Learning</b>	<b>39</b>
5.1	Model Selection . . . . .	40
5.2	Techniques . . . . .	43
5.2.1	Linear Regression . . . . .	43
5.2.2	Ridge Regression . . . . .	44
5.2.3	Lasso . . . . .	46
5.2.4	Regression Trees . . . . .	48
5.2.5	$k$ -nearest-neighbour . . . . .	51
5.2.6	Random Forest . . . . .	53
5.2.7	Support Vector Regression . . . . .	56
5.3	Implementation . . . . .	58
5.4	Results . . . . .	60
<b>6</b>	<b>Discussion</b>	<b>64</b>
<b>A</b>	<b>Appendix</b>	<b>73</b>

# 1 Introduction

Sex-related homicides tend to arouse wide media coverage and thus raise the urgency to find and prosecute the responsible offender. Especially the involvement of a child victim results in a broad discussion in the public sphere and the police is confronted with close attention. However, most homicides are cleared rather quickly and only some cases require a profound effort by the police. For these special cases, so-called profilers may assist in the ordinary criminal investigation.

Criminal profiling can be defined as the process of identifying a suspect's behavioural characteristics and principal personality from a crime scene. Police profilers firstly analyse the crime scene carefully and deduce the exact course of events. Based on this groundwork they try to discover why these events occurred and finally what type of person could have committed these acts. The method thereby relies on certain assumptions, most notably the belief that the criminal's personality can be retrieved from the crime scene.

Gaining a psychological and social profile of the suspect has several advantages for the police. Known characteristics of the offender can narrow the number of potential suspects by excluding those not showing the specific traits. This hopefully leads to a faster arrest of the criminal, but also reduces costs for the police and society. Furthermore the knowledge may lead to certain investigative strategies and, as people show different reaction to police interrogation approaches, prove useful during questioning of other suspects.

The wide and successful application of offender profiling has been enhanced by scientific background knowledge. Beauregard (2007) gives an overview of applied techniques. However most studies concentrate on a rather broad typology or predict only single variables, e.g. Davies (1997) and Salfati and Canter (1999) and consequently empirical knowledge covering the whole domain of sex-related homicides is

scarce. This also results from the low frequency of such cases, as there are on average about 40 such cases every year in Germany. Furthermore transferring these cases into data observation for a subsequent analysis is labour intensive, as the necessary information may only be gathered via a retrospective analysis reading important documents which result from the criminal investigation and the judicial proceedings. And at last, the term sex-related homicides includes a wide range of homicides, which could be as diverse as a paedophiliac offender assaulting a child or the killing of a victim, which happens to be naked in the surprise attack.

Because of these reasons the analysis of sex-related homicides is complicated and still missing a substantial effort to enlighten this field of research. The German Federal Criminal Police Office has recognized this need for information and conducted their own research on the offender's geographical behaviour (Dern et al., 2004). Knowing the distance between the crime scene and the offender's personal hub substantially narrows the geographical space to be covered in the search for the offender. The knowledge reduces the number of potential suspects and consequently accelerates the criminal investigation, as there are less potential suspects to be inspected.

In the thesis at hand we like to contribute to this topic by concentrating on the offender's age. Knowing the potential age group of an offender also reduces the number of potential suspects and therefore speeds up the criminal investigation. Firstly, we report on the structure in the domain of sex-related homicides and afterwards apply several modern prediction techniques to predict the offender's age. The structure learning part concentrates on extracting dependencies in the data, which are actually present in the data generating process. Obviously many more bivariate dependencies may be found via a statistical test, of which only a subset is present in the data generating process. We like to identify the structural form of the joint probabil-

ity function  $P(Y, \mathcal{X})$ , where  $Y$  denotes the offender's age and  $\mathcal{X} \in \mathbb{R}^P$  describes all other variables in the domain. This approach is known as unsupervised learning in the Machine Learning literature (Ripley, 1996). In the second part we concentrate on supervised learning, that is we like to infer from the knowledge of  $\mathbf{X} = \mathbf{x}$  the value of  $Y$ . We observe information on the crime scene, add this information into our prediction model and obtain a prediction of the offender's age. Obviously a model  $P(Y|\mathbf{X} = \mathbf{x})$  has to be learned to facilitate prediction. Furthermore in prediction we limit our variables  $\mathbf{X}$  to those which can be deduced from the crime scene. This limitation would be obstructive in the structure learning part, as we need to include all relevant factors to observe the actual structure. However, some factors may be hidden from the police in their criminal investigation and therefore can not be applied to a realistic approach to prediction.

In detail, we apply graphical modelling via Bayesian Networks (BN) to obtain the structure in the domain of sex-related homicides. A BN is a graphical representation of a factorised probability function in which a node is drawn for every variable and edges between the nodes describe dependencies between the variables. Such a graph marks an intuitive illustration of a probability function and facilitates an easy inspection of the dependencies. Furthermore the nodes may be endowed with local probability functions detailing how the variables influence each other. This allows for statistical inference by introducing evidence. For example some trace on the crime scene may be introduced into a BN by setting some variable value. By the structure and the local probability functions this information may be passed through the BN and alter the local probability functions along its way. The effect of the entered evidence may than be read of the altered probability functions. A BN may be generated via expert knowledge or learned from data. Due to the lack of domain knowledge we learn this structure from data and combine several learning algorithms into a final graphical model. In a subsequent step we conduct parameter learning and

obtain local probability functions from the data.

In supervised learning a function approximation  $\hat{f}(\mathbf{X})$  is to be found which generates predictions  $\hat{Y}$ . These predictions should not differ to a large extent from the true value  $Y$  to stay useful. How far the predictions differ from the true values is determined via a loss function  $L(Y, \hat{Y})$ , which returns an indication of how far the two values diverge. There exist several function approximations to predict  $Y$  and an obvious procedure would consist of choosing the one which implies the lowest loss, as detailed by the loss function. In order for the model to be applied successfully to new data, it is important to assess the loss on new observations. We apply therefore cross-validation to obtain an estimate of the expected loss and present the performance of the diverse functions approximations according to their loss resulting from cross-validation. Cross-validation is also applied to identify the value of any tuning parameter. We apply linear regression with a step procedure, ridge regression, least absolute shrinkage and selection operator (lasso), support vector regression,  $k$ -nearest-neighbour, regression tree and random forest.

The graphical model broadly presents a distinction between an offender and a situation driven crime. A situation driven crime may be characterised by an offender lacking preparation and typically attacking a known victim in familiar surroundings. The offender tends to apply blunt force to gain control over the victim and does not show a high level of forensic awareness. On the other hand offender driven crimes may be identified by the high level of forensic awareness demonstrated by the offender and the sophisticated measures applied to control the victim. Furthermore the graphic model indicates that these offenders are more likely to attack an unknown victim in unfamiliar surroundings and prepare their attack.

The prediction results do not indicate that predictions methods could

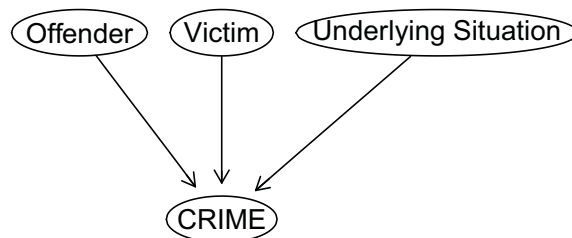
enhance the criminal investigation to a large extent. Whereas a simple base model entails a root mean square error (RMSE) of 0.303 predicting the log of the offender's age, the lasso archives 0.264. All other models generate a RMSE in between. These values translate to a decrease of the average failure in predicting the offender's age from 9 years to 8 years. Although this difference between the base model and the best performing lasso is significant in a  $t$ -test on a 5% level, it hardly matters for criminal investigation. The gain in performance of one year is not of much use, as the estimate is still too imprecise. These poor performance results may arise from two reasons. Firstly, the information on the crime scene may not be sufficient to determine the offender's age and secondly due to the heterogeneity of sex-related homicides more data may need to be collected to account for the complexity arising from the diverse homicides.

The thesis is organised as follows: Next in chapter 2 we present some information on criminological theory which allows for a better understanding of our motivation and our approach to unsupervised and supervised learning. Afterwards in section 3 we report on the data collection process. Section 4 explains the technique of BN and explains our implementation of structure and parameter learning. At the end of this section the final graphical model is discussed. Section 5 is dedicated to statistical learning. We first give details on model selection and explain the applied function approximations afterwards. The corresponding results and a discussion thereof follows. Finally, section 6 concludes.

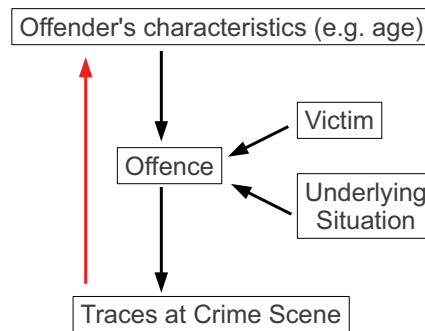
## 2 A note on Criminological Theory

Criminologist often refer to the so-called Criminal Event Perspective (Miethe and Regoeczi, 2004) to analyse a crime. Its schematic overview is given in Figure 1 and presents three mayor factors, which influence a crime: The offender, the victim and the underlying situation. Whereas the offender may be thought of as a driving force of the crime, the victim also influences the actions on the crime scene. For example, the victim's willingness to defend itself against the offender leads to more or less fighting on the crime scene. Furthermore the underlying situation described by the geographical and temporal structure exhibits influence on the crime. A criminal may want to leave the current location before starting the assault on the victim in order to avoid potential witnesses. But this geographical influence may be affected by a temporal one, as for example the possibility to encounter a potential witness in a park may be different at day time than at night time. The events at the crime scene therefore depend on all components and the crime may be understood as a result of the interaction of these three components. Consequently, all observable variables of the actual crime arise from this interaction.

As mentioned before the offender obviously exhibits the greatest influence on the crime. In detail, certain steps have to be accomplished by any offender. First the offender needs to gain control over the victim, impose sexual activities on the victim, murder the victim and



**Figure 1:** *Schematic overview on factors influencing a crime*



**Figure 2:** *Schematic overview on the structure exploited in prediction*

finally the offender may try to hide the crime and hinder the disclosure. However, there are many ways to accomplish these steps and the form applied by the offender may be interpreted as an expression of his personal characteristics. Criminologists therefore assume that the offender's characteristics affect the crime and, as the crime determines the crime scene and all traces found on the crime scene, the offender's personality may also be deduced from the crime scene. However, as denoted by the Criminal Event Perspective this influence is blurred by the victim's behaviour and the underlying situation, which also affect the crime and consequently the traces on the crime scene.

Knowing that the offender's characteristics influence the crime scene, this influence may be exploited for criminal investigation. Observing the crime scene may lead to knowledge of how the crime evolved. This information may then be applied to infer some characteristics of the unknown offender. This procedure is presented in Figure 2 and forms the basic idea of offender profiling.

Although the importance of the crime scene is widely acknowledged (Clages, 2003), the assumption of homology between the offender's characteristics and the crime scene lacks verification (Alison et al., 2002).

### 3 The Data

Parts of this section appear in Stahlschmidt et al. (2011).

The data used in this these is based upon support by the German police, which drew a sample of sex-related homicides from their internal documentation and provided access to the corresponding prosecutor's files. These files count between 1,500 and 10,000 pages, of which the crime scene report, the autopsy report, the psychiatric examination of the offender and the sentence contain almost all the essential information. Among them are, for example, the victim's injuries, the offender's age or information on the contact location. Although the documents cover all the important aspects of the crime and the offender's characteristics, this indirect access to information on the crime results in some distortion. Obviously police officers arrive at the crime scene only after the crime has been committed and therefore may only collect traces of the crime without observing it directly. Furthermore the police and judicial system act on their own principles, which constitutes a further influence on the available information. And most importantly, only traces found on the crime scene are to be considered as genuine. These are seldom sufficient and testimonies by witnesses and the offender have to be taken into account to reconstruct the details of the crime. But as this study is concerned with homicides, especially statements by the offender can only be checked against the traces at the crime scene and may leave room for speculation and misinformation.

Transferring information from prosecutor's files into nominal variables requires comparable information throughout all cases. Therefore the prosecutor's files have to be scanned to determine which content would be available for an empirical analysis. Comparative text analysis (Strauss and Corbin, 1990) is a popular technique to select the information satisfying this requirement and the variables presented in this study result from a comparative text analysis of 30 cases. However,

not all available information is of use and the amount of information transferred into variables is restricted to a consistent set of important factors in the domain of sex-related homicides. The resulting variable selection is guided by sociological and psychological theory extended by the police's hands-on experience. This theoretic background states that predominantly soft factors, such as the offender's disposition to commit a crime influence the actions at the crime scene (Mokros and Alison, 2002). These factors cannot be measured directly, but due to their complexity may only be expressed via proxy variables. Furthermore the occurrence of several offenders, victims and/or crime scenes in a single crime poses a challenge for the storage and analysis of the corresponding data. The same holds for serial crimes, in which every single crime enters the data separately, though marked by a dummy variable.

The information analysed and transferred into variables focuses on four main elements: The offender, the victim, the underlying situation and the actual offence. The offender is described by his social, psychological and economic characteristics. Furthermore information regarding his medium-term and short-term disposition to commit a crime including his criminal record and any preparation to commit the crime are collected. Information on the victim is not widely available, however indicators on her social and economic status, as well on her prior relationship status with the offender is present throughout all cases. The underlying situation with its geographical and temporal information provides the general setting of the crime. The actual offence can be split up into several categories. First, any pre-attack events regarding the offender or shared activities between the offender and the victim before the attack are taken into account. Afterwards the actual crime begins with the offender's attack on the victim, which differs, for example, in the time needed, the victim's resistance or the level of applied violence. Resulting injuries including the fatal ones are recorded and sexual activities imposed on the victim are observed.

Finally the offender's forensic awareness is measured and broadly divided into activities to hide his identity and activities to hide the crime. Further details on the variables are available in Tausendteufel et al. (2011)

The quality and quantity of the available information is highlighted by missing values and inter-rater reliability (Fleiss, 1971). Crimes resulting in limited traces entail a higher than average percentage of missing values. The same holds if the criminal refuses to testify, as several factors cannot be deduced from traces alone. Furthermore a high rate of missing values is accompanied by relatively low levels of inter-rater reliability. Raters seem to handle vague information differently. In general the data includes 6% missing values and Fleiss' measure of inter-rater reliability between four raters amounts to  $\kappa = 0.53$  with a percental match of 73%.

An overview of the variables is given in the appendix.

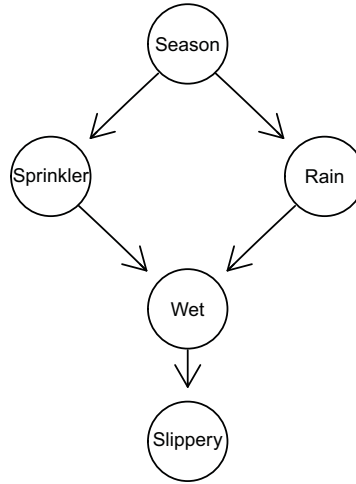
## 4 Bayesian Networks

Parts of this sections on Bayesian Networks appear in Stahlschmidt et al. (2011) or are based on Tausendteufel et al. (2011).

Bayesian Networks (BN) may serve three purposes: identification of an unknown probability function, illustrating it and drawing inference from this probability function. Learning a BN from data will identify the corresponding pdf in the domain. If the structure in the domain is known before the analysis a BN may be generated from expert knowledge. Apart from identifying the probability function a BN serves in illustrating this structure. The graph of a BN illustrates all factors via nodes and the dependencies via edges. The domain's structure can be easily read of the graphical model. Finally a BN may be exploited to draw inference from it. The graphical structure and the local probability functions in the nodes describe a system of cause and effect. Entering evidence in a node will alter the probability in other nodes of the BN and therefore statistical inference resulting from the introduced evidence may be recognised.

BN have successfully been applied to several distinct fields. For example Wright (1921) obtains graphical models for crop failure, whereas more recent examples include Heckerman (1990), who applies to BN to medical diagnosis, and Friedman et al. (2000) use BN to detect structure in biological networks. In Criminology, BNs mark a rather new tool, especially BNs driven from data and not derived from expert knowledge. Several statistical techniques have been applied to forensic data (Beauregard, 2007). However most studies concentrate on a rather broad typology or predict only single variables, e.g. Davies (1997) and Salfati and Canter (1999). Therefore Aitken et al. (1996) propose the application of BN derived from expert knowledge.

A BN consists of a directed acyclic graph (DAG) and probability functions in the nodes. It forms a graphical representation of a joint proba-



**Figure 3:** *BN describing the structure between five variables.*

bility function over several domain variables. The DAG represents the factorised probability function, where each node represents a variable of the domain and a directed edge represents a dependence between the corresponding variables. Conditional independence between variables results in a sparse graph in which only some edges persist. Each node may be endowed with a local probability function which depends on nodes pointing via directed edges towards it. This combination of a DAG and local probability functions describes, possibly causal, relations in the domain and therefore facilitates statistical inference via entering of evidence.

A BN therefore resembles the Bayes theorem

$$P(X|Y) = \frac{P(Y|X) P(X)}{P(Y)}$$

in that it also calculates a a-posterior distribution of the local probability functions after entering some information in the prior state of the BN. Figure 3 describes the classical example of Pearl (2000), which describes an artificial system between the season and the state of the ground, being slippery or not. The season directly affects the

use of the sprinkler during summer and raises the probability of rain during autumn or winter. The BN accounts for this dependencies via directed edges. The use of the sprinkler and the observation of rain directly influence if the ground is wet and consequently this state of the ground determines if the ground is slippery. The conditional independence between season and the wet ground accounts for the fact that although in may be summer the ground will only be wet, if the sprinkler has been applied.

This simple example illustrates the type of connections available in a BN and how information may be transferred across these connections. A serial connection includes three variables which are connected in a chain, e.g.  $Season \rightarrow Rain \rightarrow Wet$ . Information may be passed accordingly to the edges' direction, that is knowledge of the season influences the probability of rain, and this information leads to an update of the probability of a wet floor. This example illustrates causal reasoning. However, information may also follow diagnostic reasoning and contradict the edges' directions. Knowing that the floor is wet, alters the probability that it is raining and the notice of rain will influence the probability of being in winter. In a serial connection the edges' direction may show how causes affect effects, however the information flow is not limited to this direction.

The same holds for diverging connections. In a diverging connection some node points via directed edges at two other variables and the information may be transferred in either direction. As before the observation of rain influences the probability of the season. However, updating the probability on the season also influences the probability of using a sprinkler. Observing rain will rise the probability of being in winter. Being in winter lowers the probability of using a sprinkler. Observing the value of the variable in the middle in the serial and diverging connection blocks the information flow. The knowledge of the seasons will no longer affect the probability of observing

a wet floor, if it is known to be raining. Furthermore observing the season will block the information flow in the diverging connection  $Sprinkler \leftarrow Season \rightarrow Rain$ . Observing rain will no longer affect the probability of the use of the sprinkler, if it is known to be winter.

A converging connection, also called a collider, does not follow this reasoning. In this type of connection two non-adjacent nodes point via directed edges at the same node. In the sprinkler example, the nodes *Sprinkler* and *Rain* point both at the node *Wet*. Knowing that it is summer blocks any dependence between *Sprinkler* and *Rain*. However, if a wet floor is observed and therefore the node in the middle of a converging connection is known, the information from one end of this connection may influence the other and the path is opened. The knowledge of summer blocks the diverging connection  $Rain \leftarrow Season \rightarrow Sprinkler$ , but observing the state of the floor the state of rain will influence the state of the sprinkler. Observing a wet floor, the knowledge of rain will affect the probability of sprinkler use, as the information of a wet floor and rain will lower the probability of using a sprinkler. Two independent variables may become dependent via a third variable (Berkson, 1946).

BNs offer several advantages for the analysis of forensic data, as they describe the structure of a pre-specified domain. Hence the building of a BN mainly by data may be used for learning the structure of an unknown domain, e.g. certain types of homicides. Furthermore BNs may also be employed for prediction. A prediction of the offender's age could for example be obtained by entering evidence found on the crime scene into an appropriate BN. Furthermore, crime scenes often lack certain information or do not only render one course of events plausible, but several competing ones. By its very nature a BN can be exploited to order competing hypothesis according to their probability given the facts and allow for inclusion of soft evidence.

## 4.1 Technique

A graph  $\mathcal{G} = (\mathbf{V}, E)$  is defined by a set of nodes  $\mathbf{V} = \{V_1, \dots, V_p\}$  and a set of edges  $E \subseteq \mathbf{V} \times \mathbf{V}$ , which connect the nodes (Lauritzen, 1996). BNs form a particular subclass of graphical models and contain solely directed edges. The set of edges  $E$  in a BN includes the entry  $(V_i, V_j)$ , but not the entry  $(V_j, V_i)$  to denote a directed edge from node  $V_i$  to node  $V_j$ . Undirected edges are expressed as the entries  $(V_i, V_j)$  and  $(V_j, V_i)$  in  $E$ . In a directed edge  $(V_i, V_j)$  the node  $V_i$  is known as the parent of node  $V_j$ , and recursively the node  $V_j$  is said to be the child of  $V_i$ . The set of parents and children of a node  $V_i$  describe its adjacency and are also called neighbours of  $V_i$ . Extending the adjacency by all further parents of  $V_i$ 's children, the Markov blanket of  $V_i$  is specified. For example in Figure 3, the adjacency of node ‘‘Rain’’ consists of the parent node ‘‘Season’’ and the child ‘‘Wet’’, whereas the Markov blanket of the same node also includes the node ‘‘Sprinkler’’, as it constitutes a further parent node to the joint child ‘‘Wet’’. The descendants  $de(V_i)$  of any node  $V_i$  are defined by its children and any subsequent children. In order to distinguish clearly between descendants and non-descendants, we require the graph to omit circles. Consequentially the structure of a BN is known as a DAG (directed acyclic graph). A skeleton is a DAG without the arrow heads, such that all directed edges are converted into undirected edges. It includes several paths, describing a chain of nodes consecutively connected by edges. A chain of directed edges pointing all in the same direction is known as a directed path. If any two nodes point, via directed edges, at the same node without being adjacent, a collider arises. Figure 3 includes a single collider, namely the node ‘‘Wet’’.

A path  $\pi$  in a DAG  $\mathcal{G} = (\mathbf{V}, E)$  is said to be blocked by a set  $S \subseteq \mathbf{V}$  if node  $V_w \in S$  on the path  $\pi$  is not a collider or some other collider  $V_v \notin S$  on the path  $\pi$  exists and  $V_w \notin de(V_v)$ . Two disjoint subsets  $A$  and  $B$  of  $\mathbf{V}$  are  $d$ -separated by  $S$ , if all paths between  $A$  and  $B$  are blocked by  $S$  (Pearl, 2000). In Figure 3, the nodes *Season* and *Slip-*

*pery* are  $d$ -separate by the set  $\{Sprinkler, Rain\}$  or the single node *Wet*. However, the node *Rain* alone does not  $d$ -separate the nodes *Season* and *Slippery* as the directed path  $Season \rightarrow Sprinkler \rightarrow Wet \rightarrow Slippery$  is not blocked by it.

The probability function of a random vector  $\mathbf{X} = (X_1 \dots X_p)^\top \in \mathbb{R}^p$  with an arbitrary ordering of the variables may be factorised as

$$P(\mathbf{X}) = \prod_{i=1}^p P(X_i | X_1, \dots, X_{i-1}). \quad (1)$$

Assuming that the conditional probability of some variable  $X_i$  is affected by only its Markov parents  $PA_i \subseteq \{X_1, \dots, X_{i-1}\}$ , which describe a subset of its predecessors, (1) can be shortened to

$$P(\mathbf{X}) = \prod_{i=1}^p P(X_i | PA_i). \quad (2)$$

This assumption implies, conditional on the Markov parents  $PA_i$ , independence between  $X_i$  and its non-Markov parents predecessors  $\overline{PA_i} = \{X_1, \dots, X_{i-1}\} \setminus PA_i$ .

The probability distribution function (2) can be represented as a DAG establishing a tie between probability distribution functions and graphs. Variables  $X_i$  are displayed as nodes  $V_i$  and edges are drawn from the Markov parents  $PA_i$  towards their child  $X_i$ .

Returning to Figure 3, we factorise the joint pdf and, by certain independence statements, express it as

$$\begin{aligned} P(Season, Sprinkler, Rain, Wet, Slippery) &= P(Season) \\ &\quad \cdot P(Sprinkler | Season) \\ &\quad \cdot P(Rain | Season) \\ &\quad \cdot P(Wet | Sprinkler, Rain) \\ &\quad \cdot P(Slippery | Wet). \end{aligned}$$

Drawing all five nodes and the corresponding edges from the Markov parents to their children as denoted above, the DAG in Figure 3 is

obtained.

A DAG describes a probability distribution function graphically encoding dependencies in the distribution as edges. However, only if the probability function  $\mathcal{P}$  allows for a factorisation according to (2) relative to a DAG  $\mathcal{G}$ , we may call  $\mathcal{G}$  and  $\mathcal{P}$  Markov compatible and the DAG  $\mathcal{G}$  describes a so-called perfect map of  $\mathcal{P}$ . As a consequence conditional independences in the probability function can be inferred from  $d$ -separations in the compatible graph (Lauritzen et al., 1990). A necessary and sufficient condition for this Markov compatibility is the so-called local Markov condition requiring that every variable in  $\mathcal{P}$  may be independent of all its non-descendants conditional on its parents (Lauritzen, 1996).

If  $\mathcal{G}$  does not form a perfect map of  $\mathcal{P}$ , we may still obtain a valuable approximation of  $\mathcal{P}$ . In detail  $\mathcal{G}$  must meet two requirements to be described as a perfect map: correctness and completeness. For  $\mathcal{P}$  defined over three subsets of variables  $X_1, X_2$  and  $X_3$  with corresponding nodes  $V_1, V_2$  and  $V_3$  in  $\mathcal{G}$ , correctness of  $\mathcal{G}$  with respect to  $\mathcal{P}$  is defined as

$$X_1 \perp X_2 | X_3 \Rightarrow V_1 \perp\!\!\!\perp V_2 | V_3$$

and completeness of  $\mathcal{G}$  with respect to  $\mathcal{P}$  may be deduced from

$$X_1 \perp X_2 | X_3 \Leftarrow V_1 \perp\!\!\!\perp V_2 | V_3.$$

A correct graph contains  $d$ -separations for all independencies in the pdf and all  $d$ -separations of a complete graph are mirrored by independencies in the pdf. A correct and complete graph describes a perfect map of the corresponding pdf. A correct graph is also known as an independence map (I-map) and a complete graph may also be called a dependence map (D-map).

Several DAGs may exist, which are Markov compatible to some distribution  $\mathcal{P}$  and are correspondingly members of the same equivalence

class. An equivalence class is characterised by the same skeleton and the same set of colliders across its members, whereas the direction of any non-collider edge differs across the DAGs in the same equivalence class (Verma and Pearl, 1990). Learning a DAG via observational data is limited to finding the corresponding equivalence class and such a graph may be drawn as a completed partially directed acyclic graph.

## 4.2 Structure Learning

Structure learning refers to identifying the edges of a graphical model, where we assume that the i.i.d. data can be modeled as a sparse BN. The subsequent step of parameter learning endows the nodes with local probability functions or tables in order to transfer the DAG into a BN. As the space of DAGs grows exponentially in the number of variables, Chickering (1996) has shown that finding the correct structure of a BN is  $np$ -complete. Still several heuristic ideas exist to obtain the structure from observational data, which can be classified into constraint-based, score-based or hybrid approaches. Constraint-based approaches infer the existence of an edge by conditional independence tests and are vulnerable to errors in these tests. Furthermore the repeated application of independence tests inhibits any statement on the accuracy of the resulting graph, as the general significance level is unknown. Li and Wang (2009) have developed a constrained-based algorithm with a false discovery rate control which in comparison lacks power in disclosing existing edges. On the other hand score-based methods return a DAG, which possesses the highest score among all considered DAGs. Apart from choosing an appropriate score, these algorithms have to artificially narrow the search space in order to stay usable in large data sets. Finally, hybrid methods combine elements from constraint-based and score-based methods.

Although structure learning, defined as learning the existence of edges between nodes and consequently direct dependencies between the corresponding variables, is notoriously difficult, it constitutes a indis-

pensable step to reach the final structure of the graphical model. The evaluation of this step by comparing error rates across the diverse algorithms in predicting some variable does not turn out to be a feasible option. An optimised prediction model may not resemble the existing dependencies and independencies in the data generating process (Meinshausen and Bühlmann, 2006). Furthermore the available data is limited in that there are much more potential edges than observations. The 53 variables would lead to a complete graph of 1378 undirected edges, which existence we determine by analysing 252 observations.

The number of available data points is also short of a sufficient number of cases required by well-known structural learning algorithms (Zuk et al., 2006). The number of potential edges in a BN grows exponentially in the number of variables (Robinson, 1977) and although we have more observations than variables, we have considerably fewer observations than potential edges. This situation leads to the realm of “ $p \gg n$ ” and poses several challenges for structural learning which we address by combining several algorithms to find edges persisting throughout the resulting graphs.

We apply a combinatorial approach, which is loosely related to ensemble learning. In detail, we apply  $J = 8$  different structure learning algorithms to the data, which return an indicator  $ed_{ji} \in \{0, 1\}$  describing, if edge  $i$  has been included in the BN resulting from algorithm  $j$ . We combine these indicators  $ed_{ji}$  via the committee rule

$$ed_i^{Gen} = \mathbf{I} \left( \sum_{j=1}^J ed_{ji} > 0 \right),$$

where  $\mathbf{I}(\cdot)$  denotes the indicator function.  $ed_i^{Gen}$  determines the inclusion of an edge in the final graphical model shown in Figure 13 and consequently all edge included have been detected by at least one of the single algorithms. Obviously stricter committee rules lead to sparser combined graphs in which only edges found by several sin-

gle algorithms persist. Meinshausen and Bühlmann (2010) propose a related approach for structure learning, which generates variation by sub-sampling and, via application of a single penalized structure learning technique to the sub-samples, allows for false discovery control in the final result.

Apart from the inclusion of an edge Figure 13 also reports on how often an edge has been detected across the algorithms. This frequency

$$ed_i^{Fre} = \sum_{j=1}^J ed_{ji} ,$$

determines the thickness of an edge  $i$  in the combined graph. Instead of deciding on a result via a committee rule, the graph offers, by the displayed frequencies  $ed_i^{Fre}$ , a degree of confidence in the existence of any edge which guides the resulting discussion of the graph.

#### 4.2.1 Algorithms

We apply two score-based algorithms, five constraint-based algorithms and one hybrid algorithm. Their description in this thesis is restricted to how each of them obtains the undirected skeleton and we refrain from giving details on how the algorithms orientate the edges. There are two reasons to this. Firstly, examining observational data may only lead to observing the skeleton and colliders. All further edges' direction may not be deduced from observational data alone. Secondly, as the algorithms do not restrict their analysis to finding this so called partially directed acyclic graph, the edges' direction across algorithms contradict each other. We therefore restrict our analysis to the skeleton and explain how the algorithm generate them. Further details on how the algorithms set the orientation may be found in the quoted literature.

- The plain *Hill Climbing Greedy Search* algorithm (Heckerman, 1998) selects the BN which maximises some score criterion. At each step in the iterative process, the algorithm evaluates all

feasible steps and executes the step, which improves the score most. It stops, if the improvement in the score does not exceeds some threshold. Starting with some random structure, e.g. no edges, the algorithm may add an edge, erase an edge or change an edge's direction until it finds the action which implies the largest increase in the score and reiterates. As all Hill Climbing algorithms this algorithm may stop at some local maximum or fail to transcend some plateau in the score function.

- The *Sparse Candidate* algorithm (Friedman et al., 1999) also relies on some score measure to present a final BN, but limits the number of possible steps at each point in the iteration by some pre-processing step. It selects a set of potential Markov parents for every variable and thereby limits the number of potential BN structures which are subsequently evaluated by their score. Potential Markov parents are chosen via mutual Information. However, restricting set of Markov parents may result in suboptimal scores and the algorithm therefore reiterates the procedure. After choosing a set of potential Markov parents this information is utilised to generate a BN. This BN presents a set of Markov parents for every variable and the set of potential Markov parents is modified accordingly. Afterwards this modified set is employed to generate a new BN and the process reiterated until convergence.
- The *PC* algorithm (Sprites et al., 2000) does not minimise a score, but generates the BN via independence tests. It forms a constraint-based algorithm. Starting from a complete graph every edge is tested conditional on a set of neighbours. If the test negates an edge's existence, it is directly removed and therefore the set of neighbours reduced. The algorithm reiterates all persisting edges increasing at every iteration the set of neighbours. It starts with the empty set and increases the set of neighbours used in the independence test by one until no node does not contains a neighbourhood larger than the set to be conditioned

on.

- The *Three-Phase Dependency Analysis* algorithm (Cheng et al., 2002) passes through three phases. Firstly, the mutual information is calculated for every possible combination of two variables and a corresponding path included in the model whenever this mutual information exceeds some threshold value. Secondly, a direct edge is included between two variables whenever their mutual information exceeds a threshold. At this the mutual information is conditioned on the set of direct neighbours of the two variables which are on the path between the two variables. Lastly, a reduction phase is executed by rechecking the mutual information between two directly connected nodes conditioning on the set of direct neighbours on all paths between the two variables.
- The *HITON Parents-Children* algorithm (Aliferis et al., 2003a) is divided in two steps. For every variable  $X$  a set of potential neighbours is constructed. Other variables are admitted to this set, if they maximise some measure of association conditional on the actual state of the set of potential neighbours. In a reduction step the association between every potential neighbour and the variable  $X$  is re-examined conditional on the final set of potential neighbours and the potential neighbours excluded if the association does not exceed some threshold. Completing this last step results in a set of direct neighbours for every variable and this information may be exploited to construct the BN.
- The *Grow-Shrink Markov Blanket* algorithm (Margaritis and Thrun, 1999) may also be classified as a constraint-based approach. It concentrates entirely on the detection of the Markov blanket for every variable and based upon this information generates the corresponding BN. In the growing phase the algorithm adds, for every variable  $X$ , variables to a set  $S_X$  as long as these variables don't show to be independent from  $X$  given the present state of  $S_X$ . After testing all variables, the set  $S_X$  marks an in-

terim selection for the Markov blanket of  $X$ . In the subsequent shrinkage phase the algorithm retests the independence between a single variable of  $S_X$  and  $X$  given all other variables in  $S_X$ . This phase usually concludes with a shrunken set  $S_X$  which forms the Markov blanket of  $X$ . In the end the algorithm infers if some  $Y \in S_X$  constitutes a direct neighbour of  $X$  or a separate parent of some joint child by a further independence test between  $X$  and  $Y$  given all subsets of the Markov blanket of  $X$ .

- The *Incremental Association Markov Blanket* algorithm (Tsamardinos et al., 2003) also concentrates on the Markov blanket and combines the single Markov blankets into a BN. It basically follows the Grow–Shrink Markov Blanket algorithm and firstly allows variables in set, which describes a potential Markov blanket and afterwards repeats the independence test to exclude variables which were admitted erroneously. However in contrast to the previous algorithm it includes variables into the set of the potential Markov blanket according to their strength of mutual information. The first variable to include implies the largest mutual information and the second variable admitted contains the largest mutual information given the first variable included. The Grow–Shrink Markov Blanket algorithm includes any variable as long as it passes the independence test and does not account for the strength of the dependence.
- The *Max–Min Parents and Children* algorithm (Tsamardinos et al., 2006) aims to find the direct neighbours of some variable  $X$ , that is its parents and children. It builds a candidate set  $S_X$  by finding a subset  $R \subseteq S$  for every potential candidate  $Y$  which minimises the mutual information between  $X$  and  $Y$  and includes that variable  $Y$  that shows the highest mutual information. Obviously the name min–max originates from this procedure which at first tries to minimise the association and of all these minimised associations chooses the maximal one. After this growing phase a subsequent shrinking phase attempts to

exclude any erroneously added candidates. Therefore a subset  $T \subset S$  is searched for which renders the potential candidate of the Markov blanket independent of  $X$ . With all direct neighbours determined the BN may be generated and the algorithm determines the orientation of the edges via Hill Climbing. This last step makes the Max–Min Parents and Children algorithm a hybrid algorithm, as it employs mutual information and independence test in the generation of the skeleton and afterwards refers to some score metric to orientate the edges.

### 4.3 Parameter Learning

Parameter learning describes the endowment of nodes in a BN with probability functions conditional on their Markov parents. To decide which neighbours are actually parents and which neighbours are children of some variable  $X$ , the skeleton has to be transferred into a DAG. An popular approach to edge orientation is based on Verma and Pearl (1990) and Verma and Pearl (1992) and proposes the following rules to transfer a skeleton into a DAG:

1. For all pairs of nonadjacent variables  $X$  and  $Y$  with a common neighbour  $Z$ , test, if  $Z$   $d$ -separates  $X$  and  $Y$ . A collider  $X \rightarrow Z \leftarrow Y$  may be drawn, if this can be neglected.
2. Set the direction of an undirected edge  $Y - Z$  to  $Y \rightarrow Z$ , if there is an edge  $X \rightarrow Y$  and  $Z$  is not adjacent to  $X$ .
3. Set the direction of an undirected edge  $X - Z$  to  $X \rightarrow Z$ , if there is a chain  $X \rightarrow Y \rightarrow Z$ .
4. Set the direction of an undirected edge  $W - Z$  to  $W \rightarrow Z$ , if there are two chains  $W - Y \rightarrow Z$  and  $W - X \rightarrow Z$  and  $X$  is not adjacent to  $Y$ .
5. Set the direction of an undirected edge  $W - Z$  to  $W \rightarrow Z$ , if there is a chain  $W - X \rightarrow Y \rightarrow Z$ ,  $X$  is not adjacent to  $Z$  and  $W$  is adjacent to  $Y$ .

Whereas the first rule results in a partial directed acyclic graph (PDAG), Meek (1995) shows that the repeated application of rules 2, 3, 4 and 5 generates the corresponding maximally directed PDAG. This maximal directed PDAG may be transferred into a DAG by orienting any still undirected edges randomly accounting for the required acyclic character of the graph. This procedure may be encouraged by the fact that information may flow according to an edge's direction or contrary to an edge's direction. However, these randomly set directions may contradict domain knowledge and an alternative to setting the direction randomly consists in orienting the edge via expert knowledge.

As a result of these oriented edges, the Markov parents of every variable are known and therefore the parameters of the conditional probability function may be learned from the available data. There are two approaches, maximum likelihood estimation and Bayesian estimation. Furthermore these approaches change according to the data level, that is numeric or nominal data. We restrict our analysis to the maximum likelihood approach for nominal data, as this characterises our data. The maximum likelihood approach to discrete data makes use of two features. Firstly, the general likelihood function can be decomposed in a product of independent local likelihood functions and secondly, for tabular probability functions these local likelihood functions may be solved efficiently via sufficient statistics. Starting with the general likelihood function we first demonstrate how to decompose it and afterwards how to derive the actual maximum likelihood estimate via sufficient statistics.

The likelihood for the parameter vector  $\Theta = (\theta_{i,j})_{i=X_1,\dots,X_P;j=PA_1,\dots,PA_P}$  for the variables  $\mathbf{X} = (X_1, \dots, X_P)$  and their corresponding Markov

parents  $\mathbf{PA} = (PA_1, \dots, PA_P)$  may be decomposed as

$$\begin{aligned} L(\Theta) &= \prod_{n=1}^N P(\mathbf{X}_n; \Theta) \\ &= \prod_{n=1}^N \prod_{p=1}^P P(X_{n,p} | PA_{n,p}; \theta_{X_p | PA_p}) \end{aligned}$$

by factorization of the probability function and, by the structure of the BN, can be rewritten as

$$L(\Theta) = \prod_{p=1}^P \prod_{n=1}^N P(X_{n,p} | PA_{n,p}; \theta_{X_p | PA_p}).$$

This equation states via the definition of the likelihood for local parameters  $\theta_{X_p | PA_p}$

$$L(\theta_{X_p | PA_p}) = \prod_{n=1}^N P(X_{n,p} | PA_{n,p}; \theta_{X_p | PA_p})$$

a product of local likelihood functions

$$L(\Theta) = \prod_{p=1}^P L(\theta_{X_p | PA_p})$$

This global decomposition of the general likelihood function in a product of independent local likelihood functions facilitates the maximisation. All local likelihood functions may be maximised independently and their solution combined to reveal the general maximum likelihood estimator.

The local likelihood function may be rewritten by the sufficient statistics  $M(x_p, pa_p) = \sum_{i=1}^N \mathbf{I}(X_{p,i} = x_p \wedge PA_{p,i} = pa_p)$  which denotes how often a combination of the specific values  $x_p$  and  $pa_p$  of the variables  $X_p$  and  $PA_p$  exists across all observations. Obviously for nominal variables there are  $|X_p| \times |PA_p|$  combinations which defines the length of  $\theta_{X_p | PA_p}$ . We may factorise the local likelihood as

$$L(\theta_{X_p | PA_p}) = \prod_{pa_p \in PA_p} \prod_{x_p \in X_p} \theta_{x_p | pa_p}^{M(x_p, pa_p)}$$

and, under the constraint  $\sum \theta_{x_p|pa_p} = 1$ , where the sum refers to all values of  $PA_p$ , obtain the familiar estimate

$$\hat{\theta}_{x_p|pa_p} = \frac{M(X_p = x_p, PA_p = pa_p)}{M(PA_p = pa_p)}.$$

This parameter estimate for the specific values  $X_p = x_p$  and  $PA_p = pa_p$  may be obtained for all values of  $X_p$  and  $PA_p$  to observe  $\hat{\theta}_{X_p|PA_p}$  which results in a vector of length  $|X_p| \times |PA_p|$  denoting the parameter estimates for all combinations of the nominal variables  $X_p$  and its Markov parents  $PA_p$ .

Obtaining the parameters for a DAG turns the DAG into a BN, which may be analysed by entering evidence in some node and observe the subsequent changes in the probability distribution in other nodes. Evidence refers to setting some node to a specific value and observe its effect via inference algorithms. Although this mechanism may be exploited for prediction and even allow for entering soft evidence, defined as adjusting probabilities for some variable and observe the consequences, we restrict our analysis of the BN to the DAG structure. There are two reasons to this. Firstly due to the low number of observations we do not possess a test set of the data which we could analyse by entering certain information of the test set into the BN and compare the resulting BN probability distribution with the remaining information of the test set. Second entering fictive or simulated data in the BN does not help in assessing its value for real data. We therefore restrict the analysis to the DAG.

## 4.4 Implementation

We generate our final graphical model by applying several structure learning algorithms to the data and combine their resulting graphical models into a single skeleton. The Grow Shrink Markov Blanket, Incremental Association Markov Blanket, Max Min Parents and Children and the Hill Climbing algorithms are obtained from their implementation in the R package `bnlearn` (Scutari, 2010). The Sparse

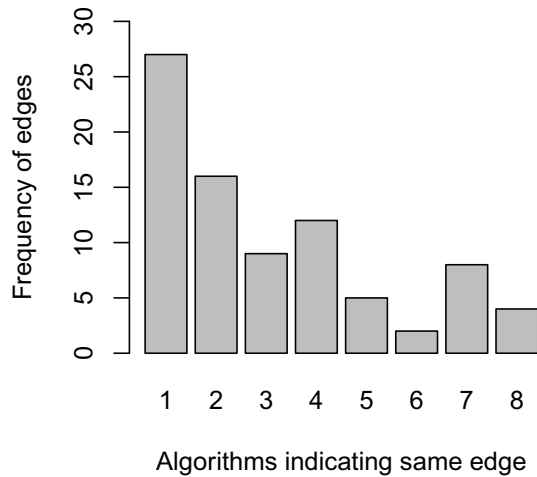
Candidate, PC, Three-Phase Dependency Analysis and HITON algorithms are used via their implementation in the MATLAB package `CausalExplorer` (Aliferis et al., 2003b). As we apply two different packages we can not make use of the same independence test across all constraint-based algorithms. `bnlearn` implements the  $\chi^2$ -test, whereas `CausalExplorer` relies on the  $g$ -test based on likelihoods. However, the  $\chi^2$  test describes a approximation of the  $g$ -test and we therefore do not expect any difference in the test results to be of great importance. Furthermore we apply the R package `pcalg` (Kalisch and Bühlmann, 2007) to direct the edges of the joint graph.

We set the Bayesian Information Criterion (Schwarz, 1978) as a score in the algorithms Hill Climbing and Sparse Candidate and apply a significance level of 5% in the independence tests used in the constraint-based algorithms.

All missing values are treated via Multiple Imputation. In detail we use chained equations based on Gibbs sampling, as implemented in the R package `mice` (van Buren and Groothuis-Oudshoorn, 2010). We use five imputations and join the diverse graphs firstly on an algorithm level to obtain a final graph from each algorithm. Only edges persisting in all five imputed data sets are accepted for the final graphical model of the respective algorithm. Only afterwards the graphical models of the eight algorithms are joint to the final graphical model presented in Figure 13.

## 4.5 Results

The application of the algorithms to our data yields several distinct graphs. We combine these graphs to a single one, in which the edge thickness is determined by how often an edge is found across the algorithms and indicates our confidence in an actual dependence between the corresponding variables in the data generating process. We omit the resulting edge direction and concentrate on the skeletons, as the

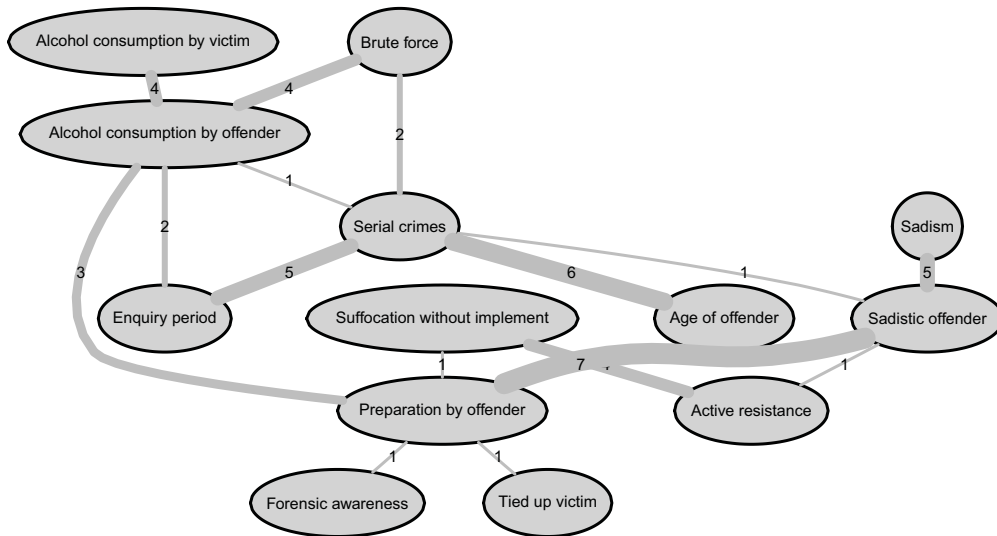


**Figure 4:** Bar chart stating how many algorithms indicate the same edge and the frequency of such edges

algorithms do not agree uniformly on all edge directions. However, nearly all directions may be deduced from sociological or psychological theory and may be examined via cross-tables. Figure 13 in the appendix presents the resulting graph, which consists of 53 nodes and 83 edges.

The single algorithms find between 20 and 68 edges and completely agree on 4 edges. A bar chart on the frequency of edges one or more algorithms, in changing combinations, agree upon is given in Figure 4. The maximal size of an adjacency in the final graph is 9, whereas the single algorithms provide adjacencies not larger than 8. The graph is considerably sparse taking into account the maximum of 1378 potential edges, which could arise from 53 variables.

The graph may be interpreted as showing the plain topology of an



**Figure 5:** Excerpt of Figure 13 showing variables which mark the difference between an offender and situation driven crime

extensively organised offender, an offender lacking organisation and a mixture type (Ressler et al., 1988). However, this categorisation has been criticised for focusing solely on the offender and consequently has been enlarged to the Criminal Event Perspective (Miethe and Regeczi, 2004). This theory stresses the influence of the victim and the underlying situation on the crime and thereby illustrates that for example, well prepared offenders may also show chaotic behaviour, if faced by unforeseen obstacles. The approach broadens the perspective to analyse a crime and we adapt it by including several variables describing the victim’s behaviour and the underlying situation as illustrated in Figure 1. Hence an interpretation of the graph will account for this extended perspective.

Starting with the node “Preparation of offender”, which is defined as the level of preparation to gain control over the victim, to hide the crime and to conduct the sexual assault, we observe 6 edges. The node

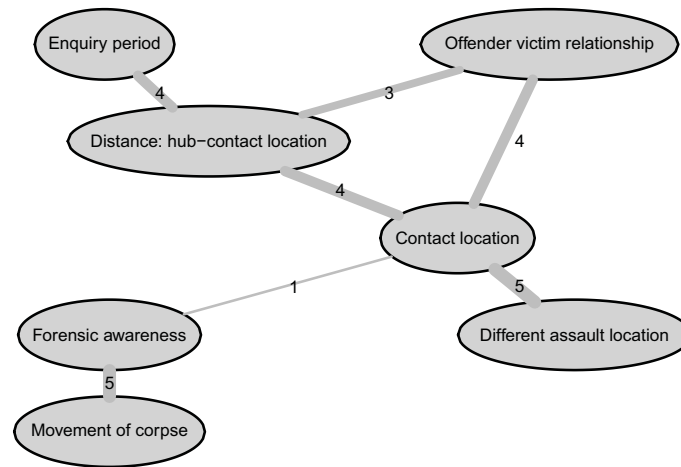
may be located in the fourth row from below to the right in Figure 13 or in the lower centre of Figure 5. Of the emerging edges from the node “Preparation of offender”, the edge towards the node “Sadistic Offender” sticks out by its thickness. The state of this node is defined via the psychiatric examination of the offender and is clearly connected to sadistic actions by the offender during the crime, included as the node “Sadism” in the graph. Examining the corresponding mosaic plots in Figure 9 shown in the appendix it may be concluded that a sadistic offender is much more likely to behave sadistically and shows a higher level of preparation. Furthermore a sadistic offender conducts serial crimes more often than a non-sadistic offender. The node “Serial crimes”, a dummy variable indicating if the specific crime is part of a wider series, exhibits profound edges to the offender’s age and the enquiry period. Serial criminals usually belong to an age group of 24 to 33 years and obviously such a crime carries a longer enquiry period.

Apart from the sadistic offender, the serial criminal marks the second ideal example of an offender driven crime. On the other hand, there are situation driven crimes. These crimes show low levels of organisation by the offender and for the most part do not involve neither sadistic nor serial criminals. Rather they display a strong influence of the consumption of alcohol, which can be read in the graph by the edge between “Preparation by offender” and the node “Alcohol consumption by offender”. This negative interaction is expanded by the node “Alcohol consumption by the victim” stating if the victim had consumed alcohol before the offender’s attack. These also include cases in which the offender and victim voluntary and before the offender’s attack engage in drinking. Most often either the victim and the offender have both consumed alcohol, which often leads to a situation driven crime, or neither the victim nor the offender have consumed alcohol, which characterises an offender driven crime. Details may be found in Figure 10 presented the appendix. Apart from alcohol, the situation driven crimes are also marked by the use of brute force by the offender

to gain and maintain control over the victim. The graphical model illustrates this interaction by the edge between “Alcohol consumption by offender” and the node “Brute force”, which reflects any injuries of the victim due to the application of blunt force.

Serial criminals with their high level of preparation generally do not rely on blunt force, but apply more sophisticated measures to control the victim. This negative interaction can be read off the mosaic plot corresponding to the edge between “Brute force” and “Serial crimes”. One such measure to control the victim applied by offenders in a criminal driven crime is described by the edge between “Preparation by offender” and the node “Tied up victim”. This node indicates if the victim is tied up by the offender and the corresponding cross-table reveals that offenders characterised by a high level of preparation are more likely to tie up their victim. Furthermore these offenders suffocate their victims less often with their hands, as highlighted by the cross-table corresponding to the edge between “Preparation by offender” and “Suffocation without implement”. In general criminals with a high level of preparation apply a more instrumental mode to gain and maintain control, whereas a low level of preparation leads to a more expressive crime, where the offender likely applies blunt force. However, the likelihood of suffocation by the offender rises in both cases, whenever the victim strongly resists the attack. This general influence of the victim on the crime is specified by the edge between the nodes “Suffocation without implement” and “Active resistance”, where active resistance is defined as resisting the assault physically, trying to escape or calling for help.

During the crime the level of planing carries over to the criminals’ behaviour, as a high level of planing is accompanied by a high level of forensic awareness. Forensic awareness describes measures to hide the crime by for example using gloves or cleaning the crime scene afterwards. The corresponding node “Forensic awareness” is connected to



**Figure 6:** *Excerpt of Figure 13 showing geographical variables and their adjacency which mark the difference between an offender and situation driven crime*

the node “Preparation by offender” highlighting this positive interaction. The corresponding mosaic plot is provided in Figure 11 in the appendix.

The node “Forensic awareness” links the degree of planning by the offender to certain geographical characteristics of the crime. The node may be found on the third row from below to the right in Figure 13 or to the left in Figure 6. Firstly, a criminal showing a high level of forensic awareness is more likely to hide the corpse at a separate location which serves solely for this purpose and complicates the prosecution. This interaction is reflected by the edge between “Forensic awareness” and “Movement of corpse”. Furthermore the node “Forensic awareness” is connected to the node “Contact location”. This node describes the location of the first contact between the offender and the victim before the assault and distinguishes between location indoors, such as the victim’s flat, the offender’s flat or a shared flat, and locations outdoors.

The corresponding mosaic plot reveals that offenders are less likely to show a high level of forensic awareness, if the contact takes place in their familiar surroundings, e.g. their own or a shared flat. On the contrary offenders meeting the victim in a rather unknown surrounding like the victim's flat or some location outdoors show a high level of forensic awareness and the corresponding crime is therefore most likely offender driven. The node "Contact location" exhibits a profound edge to the node "Offender victim relationship", which details the pre-attack relationship between the offender and the victim. Examining the corresponding cross-table reveals that the contact between the offender and an unknown victim is mostly established outdoors, whereas offenders meet any known victims rather indoors.

As an outdoor location is associated with a high level of forensic awareness, these outdoors contacts between the offender and the unknown victim may be attributed to the offender driven crime, whereas the indoor contact exhibits the characteristics of a situation driven crime and likely includes a victim known to the offender. An offender meeting the victim in his familiar surrounding obviously does not travel a great distance from his personal hub to the contact location, where a hub is defined as any location the offender is perfectly familiar with, e.g. his flat or work place. The graph therefore includes an edge between these two nodes. Furthermore the node "Distance: hub – contact location" is connected to the node "Enquiry period" and the corresponding mosaic plot details that a greater distance between the offender's personal hub and the contact location complicates the prosecution, as the enquiry period rises.

In general, it may be concluded, that the differentiation between an offender driven crime and a situation driven crimes carries over to the geographical variables. Well organised offenders meet the victim in general not in their familiar surrounding, but have rather travelled a

longer distance and hide the corpse at a separate location to impede the exposure of their crime. Less organised offenders meet the victim, which is most likely known to them, in rather familiar surroundings and do not travel a great distance. Furthermore they do not show a high level of forensic awareness or hide the corpse at a separate location. However, as before, the actual crime is not solely influenced by the criminal, as an examination of the edge between the nodes “Contact location” and “Different assault location” depicts. If the offender meets the victim in an outdoors location, in just over half of the crimes, the ensuing attack is conducted at a different location. The offender may not feel confident, that the contact location outdoors allows him to conduct the crime and is therefore forced to change the location. This change of location occurs only in less of a quarter of all crimes, in which the contact location is indoors. Figure 12 in the appendix provides the corresponding mosaic plot.

## 5 Statistical Learning

Statistical Learning may be divided into unsupervised and supervised learning. Unsupervised learning refers to inferring information of the joint probability function of several variables  $\mathbf{X} \in \mathbb{R}^P$  and BNs are one of the applied methods in unsupervised learning. Supervised learning refers to predicting some response variable  $Y$  via the predictors  $\mathbf{X}$ . A data set of predictors and the corresponding responses are gathered and this training data set  $\{Y, \mathbf{X}\}_{n=1}^N$  is analysed to obtain a functional approximation  $\hat{Y} = \hat{f}(\mathbf{X})$  of the function  $f(\mathbf{X})$ , that governs the relationship between the predictors and the response variable. If  $Y$  is real, we are looking at a regression task, if  $Y$  is nominal or ordinal, we deal with a classification task (Hastie et al., 2009).

For a regression task characterised by an adaptive error  $\varepsilon$  with expectation  $E(\varepsilon) = 0$  and variance  $\text{Var}(\varepsilon) = \sigma_\varepsilon^2$  and squared error loss function  $L(Y, f(X)) = (Y - f(X))^2$ , the optimal functional approximation is the conditional expectation  $f(x) = E(Y|X = x)$  also known as the regression function.

If we want to infer the function  $f(X)$  from the training data set, we may choose any function passing through the training points  $\{y_n, x_n\}$ . Obviously some of these functions will perform better on new and unseen data points and the search space must therefore be restricted to promising functions, which generalise well to new data. This restriction may be done by reducing the complexity of the model via a tuning parameter, that is the degree to which the model adapts to the specific characteristics of the training data. This is often accomplished by imposing some regular behaviour in small neighbourhoods of the input space. However, as the number of predictors grows, the curse of dimensionality will affect this approach and constitute a further obstacle.

## 5.1 Model Selection

As there is no natural choice of a model to predict the offender's age, we apply several models and choose the best one among them. We decide on the best one via a loss function, which describes how far off the prediction is from the true value. The offender's age marks the response variable  $Y \in \mathbb{R}$  which we like to predict via the predictors  $\mathbf{X} \in \mathbb{R}^P$ . We do so with a prediction model  $\hat{f}(\mathbf{X})$ , which arises from the analysis of some training set  $\mathcal{T} = \{Y_n, \mathbf{X}_n\}_{n=1}^N$  governed by a unknown, joint probability function  $P(Y, \mathbf{X})$ . A typical choice for the loss function consists in the squared error

$$L(Y, \hat{f}(\mathbf{X})) = (Y - \hat{f}(\mathbf{X}))^2,$$

which returns the loss in quadratic terms. Other popular choices include the absolute loss or the 0–1–loss for classification tasks.

With a loss function at hand, one may compare the fitted values with the actual observations:

$$\text{err} = \frac{1}{N} \sum_{n=1}^N L(y_n, \hat{f}(x_n)).$$

This statistic describes the training error, which may be decreased at will by increasing the model complexity. The model exploits the information in the training data set to a large extent and adapts to the specific structure of this data sample. Such a model may be characterised by a small bias, but large variance and will not generalise well to new data. By tying the specific form of the model to close to the available training data set, overfitting occurs and the trained model fails in predicting unobserved observations.

The so-called test error measures the prediction performance of the model on observations not included in the training data set  $\mathcal{T}$  and therefore new to the model:

$$\text{ERR}_{\mathcal{T}} = \text{E} \left[ L(Y, \hat{f}(\mathbf{X})) \mid \mathcal{T} \right], \quad (3)$$

where the pair  $\{Y, \mathbf{X}\}$  refer to random draws from the population and  $\mathcal{T}$  indicates that the model has been set up via exploring the training data. It therefore mimics the error rate one would expect by setting up a model on a specific training data set and observing its prediction performance on new observations.

The expected test error

$$\begin{aligned}\text{ERR} &= \text{E}[\text{ERR}_{\mathcal{T}}] \\ &= \text{E}\left[L\left\{Y, \hat{f}(\mathbf{X})\right\}\right]\end{aligned}\tag{4}$$

averages the test error over training data sets. It therefore eliminates the influence of the training data set on the model and does not analyse how well a model, build on some specific training data set, performs on new data. It serves however to illustrate the effect of increasing the model complexity to determine the model's generalisation performance. Assuming an additive error model  $Y = f(X) + \varepsilon$  with  $\text{E}[\varepsilon] = 0$  and  $\text{Var}(\varepsilon) = \sigma_{\varepsilon}^2$  and applying a squared error loss function the expected test error may be broken down into its elements:

$$\begin{aligned}\text{E}\left[\left(Y - \hat{f}(x_0)\right)^2\right] &= \sigma_{\varepsilon}^2 + \left[\text{E}\hat{f}(x_0) - \hat{f}(x_0)\right]^2 + \text{E}\left[\hat{f}(x_0) - \text{E}\hat{f}(x_0)\right]^2 \\ &= \sigma_{\varepsilon}^2 + \text{Bias}^2\left(\hat{f}(x_0)\right) + \text{Var}\left(\hat{f}(x_0)\right) \\ &= \text{Irreducible Error} + \text{Bias}^2 + \text{Variance}.\end{aligned}$$

Increasing the model complexity will lower the bias, but increase the variance. Optimising the training error with its strong adaption to the training data set will result in a high model complexity with high variance and a low bias. Such a model will therefore show a decreasing prediction error on the training data set, but perform poorly on new observations. The prediction performance of a model optimised by the test error will also start to improve as the model complexity is increased. It will however reach a minimum and thereafter the prediction performance will start to decrease. An optimal model will apply just the right amount of model complexity to reach the minimum test

error. This minimum will also hold for new observations.

Following the definition of the test error one could set apart a chunk of the training data and obtain the test error by applying the model, obtained without this chunk of data, to these separated observations. But data on sex-related homicides data is sparse and we may not want to exclude valuable observations from the model building process. We therefore apply cross-validation to obtain an indication of the model performance. We divide the training data set in 10 folders and build the model by excluding one them. The prediction performance of this model is afterwards evaluated for the data observations in this folder. Repeating this process for every of the ten folders returns the cross-validation estimate of the prediction error

$$\text{CV}(\hat{f}) = \frac{1}{N} \sum_{n=1}^N L\left(y_n, \hat{f}^{-\kappa(n)}(x_n)\right),$$

where  $\kappa$  describes an indexing function splitting the training data set in 10 folders and  $\hat{f}^{-\kappa(n)}$  denotes the model generated without  $\kappa$  part of the training data. Although cross-validation is a feasible approach to obtaining a prediction error without excluding data observations in the modelling process, it comes at a cost. The cross-validation estimate of the prediction error estimates the expected test error (4) and not the actual test error as defined in (3). By applying cross-validation we therefore trade the use of the whole training data set against an imprecise estimate of the test error.

Apart from obtaining an estimate for the test error, cross-validation may also serve to determine the value of any tuning parameter  $\alpha$ . As the value of the tuning parameter is chosen as to optimise the prediction performance, the parameter may be included in the calculation of the prediction error:

$$\text{CV}(\hat{f}, \alpha) = \frac{1}{N} \sum_{n=1}^N L\left(y_n, \hat{f}^{-\kappa(n)}(x_n, \alpha)\right),$$

where  $\widehat{f}^{-\kappa(n)}(x_n, \alpha)$  denotes the model obtained by setting some specific value  $\alpha$ .  $\text{CV}(\widehat{f}, \alpha)$  reports an estimate on the prediction error and  $\alpha$  may be chosen to minimise it.

## 5.2 Techniques

### 5.2.1 Linear Regression

A simple linear model assumes the regression function to be linear in the predictors  $\mathbf{X} \in \mathbb{R}^p$ . This assumption may be justified by describing a reasonable approximation of the true model. A mayor benefit of linear models is the clear and interpretable description of how the predictors influence the response variable  $Y$ . The model may be denoted as

$$\widehat{f}(\mathbf{X}) = \beta_0 + \sum_{p=1}^P X_p \beta_p,$$

where the parameter  $\beta_p$  describes the influence of the variable  $X_p$  on the response  $Y$ . A popular method to solve a linear model is described by the term least-squares. At this the residual sum of squares

$$\text{RSS}(\beta) = \sum_{n=1}^N \{y - \widehat{f}(x)\}^2$$

describing how well the model fits the data in squared terms, is minimised by choosing appropriate values for  $\beta$ . The unique solution for  $\beta$  is given by

$$\widehat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

Apart form the least-squares criterion the same solution may also be obtained via maximum likelihood estimation. This approach determines the appropriate parameter values  $\theta$  via the likelihood function

$$L(\theta) = \sum_{n=1}^N \text{P}(y_n | \theta).$$

In detail, the appropriate values  $\widehat{\theta}$  are those which maximise this function or its log transformation  $l(\theta) = \log(L(\theta))$ . In case of the linear

model  $f_\beta(\mathbf{X}) = \beta_0 + \sum_{p=1}^P X_p \beta_p$ , this results in the log-likelihood function

$$l(\beta) = -\frac{N}{2} \log(2\pi) - N \log \sigma - \frac{1}{2\sigma^2} \sum_{n=1}^N \{y_n - f_\beta(\mathbf{x}_n)\}^2$$

Maximising this log-likelihood function over the parameters  $\beta$  only affects the last term, as the first two terms do not include the parameters. This last term equals the RSS, a statistic minimised in least square, up to a scalar multiplier. This highlights the connection between the two methods for a linear model. More general, the application of least squares to a linear model with an additive error  $Y = f_\beta(\mathbf{X}) + \varepsilon$  with the error distributed as  $\varepsilon \sim N(0, \sigma^2)$  results in the same parameter values as the application of maximum likelihood to the Gaussian conditional likelihood

$$P(Y|\mathbf{X}, \beta) \sim N(f_\beta(\mathbf{X}), \sigma^2).$$

In spirit of the other Machine Learning techniques applied in this section we generate our linear model via an automatic model selection process. In detail we apply the AIC criterion (Akaike, 1974) in a stepwise selection procedure. A simple model with only a constant and a full model with all variables included define the lower and upper bound of the search space. Starting with some random model, at each step a variable is added or cleared from the list of predictors until the AIC criterion can not be improved by any such step. The variable added or cleared from the active predictor list is the one which results in the highest improvement of the AIC criterion at each point of the iteration. The result is given in Table 1.

### 5.2.2 Ridge Regression

Subset selection as conducted by the stepwise procedure just explained either includes variable in the set of predictor or cleans it from the list of active predictors. This marks a discrete process and is often accompanied by a high variance and therefore dissatisfying prediction

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	3.3175	0.0993	33.41	0.0000
t_gefaengnisvor._normNein	-0.1675	0.0353	-4.75	0.0000
tv_sozsit_kNein	0.1918	0.0530	3.62	0.0004
o_tatalter	0.0026	0.0008	3.27	0.0012
toi_krim_spur_kondNein	0.0719	0.0474	1.52	0.1302
toie_dauer_kue	0.0032	0.0012	2.67	0.0080
xu_zerlaubnisNein	-0.0950	0.0342	-2.78	0.0059
xu_zlistNein	0.1264	0.0515	2.46	0.0147
toit2_sex_manNein	-0.1210	0.0443	-2.73	0.0068
x_spurendnaNein	-0.0584	0.0343	-1.70	0.0905
x_l_loksons._Tatorte_drinnen	-0.0967	0.0593	-1.63	0.1041
xl_vgeschlecht_norm2Nein	-0.0756	0.0344	-2.20	0.0289
xl_spositionNein	-0.0790	0.0439	-1.80	0.0730
xl_vextremitaeten_norm2Nein	-0.0713	0.0337	-2.12	0.0354

**Table 1:** *Coefficients, standard errors, t values and their grading for the stepwise procedure to generate a linear model. Variable names are explained in Table 8*

performance. Shrinkage methods like ridge regression and the lasso mark a more continuous process and therefore exhibit less variance.

Ridge regression (Hoerl and Kennard, 1970) shrinks the regression coefficients by imposing a  $L_2$ -penalty on their size. In detail the parameter values  $\beta^{ridge}$  minimise the RSS which is enlarged by a penalty term:

$$\hat{\beta}^{ridge} = \arg \min_{\beta} \left\{ \sum_{n=1}^N \left( y_n - \beta_0 - \sum_{p=1}^P x_{np} \beta_p \right)^2 + \lambda \sum_{p=1}^P \beta_p^2 \right\}.$$

The parameter  $\lambda$  controls the amount of shrinkage and marks a tuning parameter to be determined via cross-validation. A large value

of  $\lambda$  will decrease the parameter values of  $\beta^{ridge}$ , as the coefficients are shrunken towards zero and towards each other. Ridge regression may be classified as a proportional shrinkage methods in which the smaller principal components of  $\mathbf{X}$  are shrunken to a larger extent than the larger principal components of  $\mathbf{X}$ . The reasoning behind this behaviour of ridge regression lies in the assumption that the response variable will vary most in the direction of high variance of the predictors and will vary less in the direction of small variance of predictors. By applying a stronger shrinkage on the smaller principal components of  $\mathbf{X}$  ridge regression decreases the noise resulting the low amount of data on these small principal components.

Rewriting the residual sum of squares criterion in matrix form

$$\text{RSS}(\lambda) = (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^\top \beta$$

and solving for  $\beta$ , the actual parameter values  $\hat{\beta}^{ridge}$  may be deduced from

$$\hat{\beta}^{ridge} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}.$$

In this notation it may be observed that a positive constant is added to the diagonal of  $\mathbf{X}^\top \mathbf{X}$  and this facilitates inversion, even if  $\mathbf{X}^\top \mathbf{X}$  may be singular. This was the main motivation, when ridge regression was presented.

Table 2 report on the coefficients  $\hat{\beta}^{ridge}$  for the prediction model of the log of the offender's age. There are no standard errors reported as the bias in ridge regression is an integral and welcomed part of the model.

### 5.2.3 Lasso

The lasso (Tibshirani, 1996) also applies a penalty to the RSS. This penalty, however, is a  $L_1$  penalty and results in a nonlinear solution for the corresponding parameter values  $\hat{\beta}^{lasso}$ . In detail, these parameter

	Estimate
t_gefaengnisvorerfahrung_normNein	-0.12
tv_sozsit_kNein	0.11
o_tatalter	0.00
toi_krim_spur_kondNein	0.04
toie_dauer_kue	0.00
xu_zerlaubnisNein	-0.06
xu_zlistNein	0.09
xu_zoffentlichNein	0.02
toit1_sex_vaginalNein	0.04
toit2_sexNein	0.03
toit2_sex_manNein	-0.10
x_spurendnaNein	-0.05
x_spurenpersdingeNein	-0.01
x_l_loksonstige_Tatorte_drinnen	-0.06
x_l_lokWohnung_Taeter	-0.00
x_k_loksonstige_Tatorte draussen	0.02
x_k_lokWohnung_Taeter	0.03
x_u_loksonstige_Tatorte draussen	0.00
x_u_loksonstige_Tatorte_drinnen	-0.01
x_kungleichuNein	0.03
x_ort_fallgemischt	-0.01
xl_vgeschlecht_norm2Nein	-0.06
xl_spositionNein	-0.06
toi_enteigenNein	0.01
xl_vextremaeten_norm2Nein	-0.05

**Table 2:** *Regression coefficients resulting from ridge regression. Variable names are explained in Table 8*

values are the solutions to the Lagrangian form

$$\hat{\beta}^{lasso} = \arg \min_{\beta} \left\{ \sum_{n=1}^N \left( y_n - \beta_0 - \sum_{p=1}^P x_{np} \beta_p \right)^2 + \lambda \sum_{p=1}^P |\beta_p| \right\}.$$

There is no closed form solution for this quadratic programming problem and a solution has to be determined numerically.

The penalty may also be denoted as

$$\sum_{p=1}^P |\beta_p| \leq t.$$

The lasso translates all coefficients by some constant factor truncating at zero. Therefore a sufficient small value of  $t$  will set some of the coefficients to zero and perform thereby a subset selection. On the other hand choosing some  $t_0 = \sum_{p=1}^P |\beta_p|$  will result in no shrinkage at all and the coefficients  $\hat{\beta}^{lasso}$  will not differ from the ordinary OLS coefficients. Setting  $t = t_0/4$  will shrink the least squares coefficients by 25% on average. Obviously the size of the penalty, denoted as  $t$  or rewritten as  $\lambda$ , is a tuning parameter and may be determined via cross-validation.

Table 3 reports on the coefficients  $\hat{\beta}^{lasso}$  for the prediction model of the log of the offender's age.

#### 5.2.4 Regression Trees

Tree-based models split the feature space in several distinct rectangles and fit a simple model, for example a constant, in each of them. To simplify the generation of a tree, usually only recursive binary partitions are applied to the feature space. As a consequence the partition of the feature space may be drawn as a binary tree. This facilitates the interpretation of the resulting tree and every observation classified by the tree in one of the rectangles may be inspected by following its path down the tree. Furthermore drawing the divided feature space

	x
t_gefaengnisvorerfahrung_normNein	-0.15
tv_sozsit_kNein	0.16
toi_krim_spur_kondNein	0.04
xu_zerlaubnisNein	-0.07
xu_zlistNein	0.10
xu_zoffentlichNein	0.01
toit1_sex_vaginalNein	0.03
toit2_sexNein	0.03
toit2_sex_manNein	-0.13
x_spurendnaNein	-0.05
x_l_loksonstige_Tatorte_drinnen	-0.07
x_kungleichuNein	0.03
xl_vgeschlecht_norm2Nein	-0.06
xl_spositionNein	-0.07
xl_vextremitaeten_norm2Nein	-0.06

**Table 3:** *Regression coefficients resulting from the lasso. Variables shrunk to zero are omitted. Variable names are explained in Table 8*

is only feasible for a low dimensional feature space, whereas trees do not imply such a limit.

We generate the regression tree by following the CART approach by Breiman et al. (1984). At first, the feature space  $\mathbf{X}$  is split in two regions and the mean of the response variable  $Y$  in every region is reported as the model response. The variable to be split and the split-point are chosen to archive the best model fit. Afterwards the two separate regions are split and this process continues until some

stopping criterion is fulfilled. In general the model may be denoted by

$$\hat{f}(\mathbf{X}) = \sum_{m=1}^M c_m \mathbf{I}\{(X_1, \dots, X_p) \in R_m\},$$

where  $M$  is the number of separate regions in the feature space,  $c_m$  denotes the model response in region  $m$  and indicator function reports if an observation falls into the region  $m$ . Any algorithm generating such a partition of the feature space must decide which variables to split and at what point to split them. If we set the sum of squares  $\sum_{i=1}^N (y_i - f(x_i))^2$  as a minimisation criterion,  $c_m$  will equal the average of every  $y_i$  in that region, that is

$$\hat{c}_m = \text{ave}(y_i | x_i \in R_m).$$

After deciding on this minimisation criterion the tree may be obtained via an greedy algorithm. Describing a pair of half-planes for some splitting variable  $u$  and split point  $v$  by

$$R_1(u, v) = \{X | X_u \leq v\} \quad R_2(u, v) = \{X | X_u > v\}$$

we chose the splitting variable  $u$  and split point  $v$ , which minimise

$$\min_{u,v} \left[ \min_{c_1} \sum_{x_i \in R_1(u,v)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(u,v)} (y_i - c_2)^2 \right].$$

At this the value of  $c_1$  is solved via  $\hat{c}_1 = \text{ave}(y_i | x_i \in R_1(u, v))$  and the value for  $\hat{c}_2$  respectively. As we apply a greedy algorithm, we scan through all combinations of  $u$  and  $v$  to find the best pair at each step. Having found this pair we repeat the splitting process in the separated regions. Obviously repeating the process too often will generate a overfitted tree, while a small tree may ignore important structure and the corresponding tree size serves as a tuning parameter determining the model's complexity.

We determine the optimal tree size via cost-complexity pruning and minimise the cost complex criterion

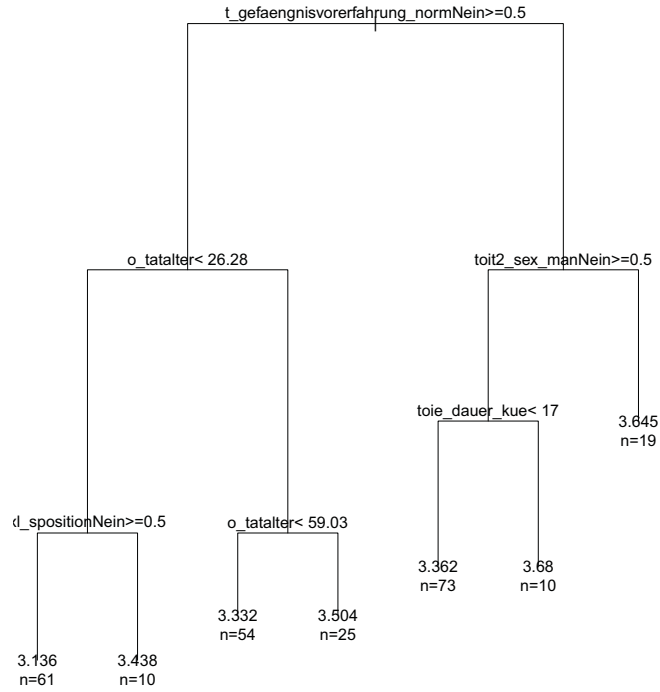
$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha |T|,$$

where  $|T|$  describes the number of terminal nodes,  $N_m = \#\{x_i \in R_m\}$  denotes the number of observations falling into the partition  $m$  of the feature space and  $Q_m(T) = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \hat{c}_m)^2$  describe the average sum of squares in the partition  $m$ . A subtree  $T_\alpha \subseteq T$  minimises  $C_\alpha(T)$  for every  $\alpha$ , where the tuning parameter  $\alpha$  determines the trade-off between goodness of fit and tree size. For every  $\alpha$  the unique subtree  $T_\alpha$  may be obtained via weakest link pruning. Starting with an excessively grown tree, weakest link pruning determines the internal node, which results in the smallest per-node increase in the sum of squares statistics  $\sum_m N_m Q_m(t)$ , collapses the node and repeats this procedure until a root tree emerges. This sequence of subtrees contains the sought after tree  $T_\alpha$  (Ripley, 1996). The optimal  $\alpha$  may be chosen via cross-validation and this method generates for the our purpose the tree in Figure 7.

### 5.2.5 $k$ -nearest-neighbour

$k$ -nearest-neighbour techniques (Cover and Hart, 1967) describe a powerful, yet simple technique for classification and regression. Although it may not result in the lowest error rate for a given classification problem, it is usually routinely reported, as the Bayes error rate asymptotically amounts to half of the error rate of the 1-nearest-neighbour classifier. It therefore roughly indicates an optimal lower bound, which could be archived at the most among all prediction techniques.

The techniques also differs in that it is memory-based and the only major modelling decision involved is the definition of the distance between the observations in the feature space. However, due to this characteristics, it is necessary to include the whole data set into an analyses of a new observation and this may become challenging in high-dimensional settings. Given the new observation,  $k$ -nearest-neighbour compiles the distance between this new observation and



**Figure 7:** Regression tree for the prediction of the offender's age. Variable names are explained in Table 8

the training data set and selects the  $k$  nearest neighbours. An average is deduced among their values for the responses, which gives the model response:

$$\hat{f}(x) = \frac{1}{k} \sum_{i=1}^k y_k,$$

where  $i$  denotes the  $k$  nearest neighbours according to the employed distance metric. The parameter  $k$  describes a tuning parameter governing the complexity described by its bias and variance. Small values of  $k$  result in a low bias, but high variance. The actual value of  $k$  for a certain data set may be deduced via cross-validation.

For numeric predictor  $\mathbf{X} \in \mathbb{R}^P$ , the Euclidean distance between a new observation  $x_0$  and the training observation  $x_n$

$$d_n = \|x_n - x_0\| = \sqrt{\sum_{p=1}^P (x_{n,p} - x_{0,p})^2}$$

would be a natural choice for the distance metric.

### 5.2.6 Random Forest

Random Forrest (Breiman, 2001) grow a large number of de-correlated trees on bootstrap samples of the data and their model responses constitutes a simple average across those trees:

$$\hat{f}^B(\mathbf{X}) = \frac{1}{B} \sum_{b=1}^B T(\mathbf{X}, \Theta_b)$$

where  $B$  denotes the number of trees and  $T$  a single tree with parametrisation  $\Theta_b$ . The parametrisation includes the split variables, the splitting points and the values at the terminal nodes.

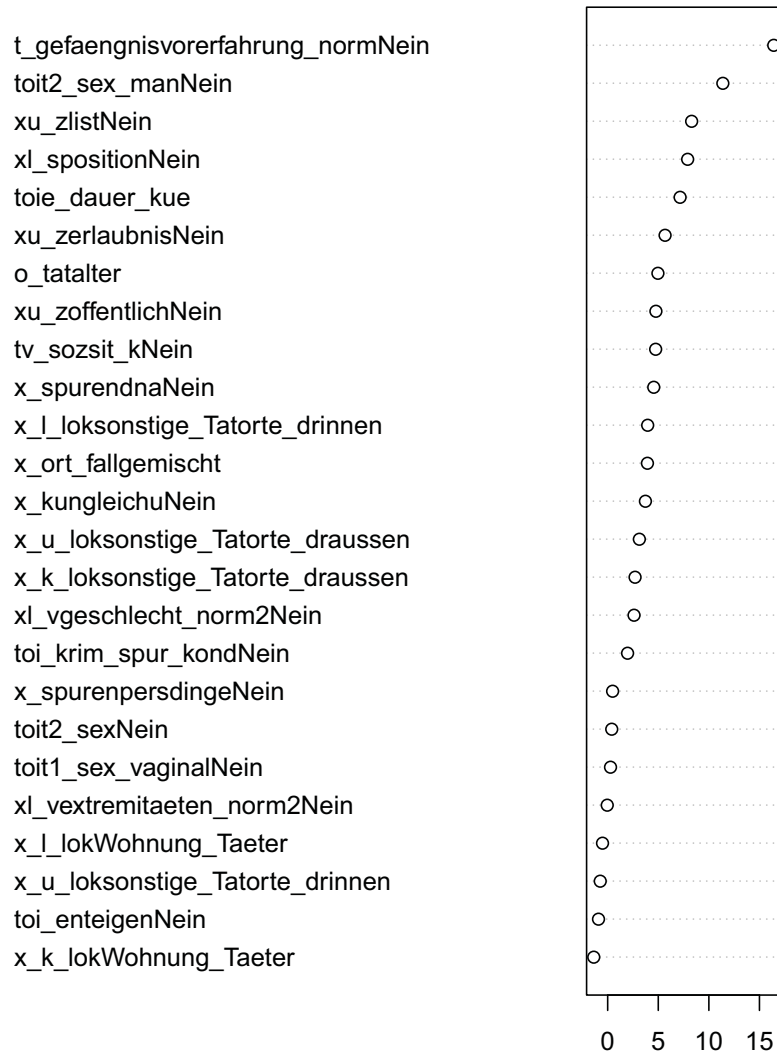
As the single trees are generated from bootstrap samples of the data, the techniques constitutes a modification of bagging (Breiman, 1996) on trees in which the correlation between the trees is reduced to minimise the variance of the predictor. A sufficiently deep grown tree incorporates low bias, but a high variance. However, as the trees in random forests are optimised on bootstrap samples of the data, there are identically distributed and the expectation of an average of such trees is the same as the expectation of a single tree. Consequently the bias is not affected by growing a large number of trees. On the other hand, the variance of the average can be reduced and therefore random forests results in better prediction. In detail, observing  $B$  identically distributed variables  $T_i, i \in \{1, \dots, B\}$  with variance  $\sigma^2$  and pairwise correlation  $\rho$ , the average of these  $B$  variables will exhibit the variance

$$\text{Var}(\text{ave}\{T_1, \dots, T_B\}) = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2.$$

Increasing  $B$  will lead to a decrease of the second term, whereas the first term may only be tackled by reducing the pairwise correlation  $\rho$ . Random forests lowers the correlation between its trees by choosing a set of predictors at random. It admits only a subset of  $m \leq p$  predictors as candidates for splitting variables. These  $m$  candidates are chosen at random at each node. Reducing  $m$  will lower the correlation between the trees and  $m$  therefore serves as a tuning parameter which may be resolved via cross-validation.

The algorithm firstly generates a large number of bootstrap samples, for example  $B = 500$  and grows a tree on each of them without relying on pruning, but by stopping when the minimum node size is reached. As a further difference to section 5.2.4 on regression trees, at each node  $m$  variables are selected at random as candidate splitting variables and the, in terms of reduced sum of squares, best variable with its according splitting point is chosen among them. This procedure results in an ensemble of trees  $\{T_b\}_{b=1}^B$  with parameters  $\Theta_b$ , which vary across the trees because of the bootstrap samples of the data and the random selection of  $m$  variables for splitting at each node. Every new observation is passed down all single trees and the average of all single tree results is returned as the model prediction of this new observation.

As the trees are grown on a bootstrap sample of the data, not every observation is used in the generation of a single tree  $T(\Theta_b)$  and the prediction power of the tree may be tested on the observations not used for the growing process. This error is also known as the out-of-bag (OOB) error and, as cross-validation, predicts the test error of the random forest. This OOB error may be exploited to gain knowledge on the variable importance, that is information on which variables have an effect on the prediction accuracy. In detail, for a grown tree  $T(\Theta_b)$  of a random forest the OOB sample is passed down the tree and the prediction error is recorded. Afterwards the values for some variable  $p$  are permuted in the OOB sample and therefore



**Figure 8:** Variable importance in random forest concerning the MSE increase after permutation. Values are divided by their standard errors. Variable names are explained in Table 8

the prediction power of this variable  $p$  minimised. This modified OOB sample is again passed down the tree and the difference in the prediction accuracy indicates the importance of the variable  $p$ . The values of an importance sampling for the data set analysed in this thesis is given in Figure 8.

### 5.2.7 Support Vector Regression

Support vector regression (Vapnik, 1995) adapts properties of support vector machine classification to the prediction of a real response variable. It also incorporates a margin and the predictors may also be mapped to a feature space. However, in support vector regression the observations outside the margin add to the cost, whereas in support vector machine classification observations outside the margin do not matter for the prediction, but slack variables on the wrong side of the linear decision boundary add up to the cost.

The role of the margin in support vector regression originates in the applied  $\epsilon$ -insensitive loss function  $V_\epsilon$ :

$$V_\epsilon(r) = \begin{cases} 0 & , \text{ if } |r| < \epsilon \\ |r| - \epsilon & , \text{ otherwise,} \end{cases}$$

where  $\epsilon$  describes a threshold and  $r$  denotes the difference between the true value and the fitted value. In  $\epsilon$ -support vector regression we search for a prediction function  $\hat{f}(X)$  that has at most  $\epsilon$  deviation from the true observation  $y$  for all training data and at the same time minimises complexity. In a simple linear model

$$f(X) = \langle \beta, X \rangle + \beta_0,$$

where  $\langle \cdot, \cdot \rangle$  denotes the dot product, this translates into minimising the Euclidean norm  $\|\beta\|_2$  and results in the optimization problem

$$\min \frac{1}{2} \|\beta\|_2^2 + C \sum_{n=1}^N (\xi_n + \xi_n^*) \quad \text{s. t.} \quad \begin{cases} y_n - \langle \beta, x_i \rangle - \beta_0 < \epsilon + \xi_n \\ \langle \beta, x_n \rangle + \beta_0 - y_n \leq \epsilon + \xi_n^* \\ \xi_n, \xi_n^* \geq 0. \end{cases}$$

We allow for some error by including slack variables  $\xi_i$  and  $\xi_i^*$ . This facilitates the optimisation.  $C$  denotes a tuning parameter, which equilibrates the complexity and the allowed amount of deviance.

This optimisation may be solved more easily in its dual formulation.

For this purpose we specify the Lagrange function

$$L = \frac{1}{2} \|\beta\|_2 + C \sum_{n=1}^N (\xi_n + \xi_n^*) - \sum_{n=1}^N \alpha_n (\epsilon + \xi_n - y_n + \langle \beta, x_n \rangle + \beta_0) \\ - \sum_{n=1}^N \alpha_n^* (\epsilon + \xi_n^* + y_n - \langle \beta, x_n \rangle - \beta_0) - \sum_{n=1}^N (\eta_n \xi_n + \eta_n^* \xi_n^*)$$

with the Lagrange multiplier  $\alpha_n, \alpha_n^*, \eta_n$  and  $\eta_n^*$ . Setting the partial derivatives of the primal variables  $\beta, \beta_0, \xi_n$  and  $\xi_n^*$  to zero and rearrange terms via substituting yields the dual optimisation problem

$$\max \left( -\frac{1}{2} \sum_{n,m=1}^{n,m=N} (\alpha_n - \alpha_n^*) (\alpha_m - \alpha_m^*) \langle x_n, x_m \rangle - \epsilon \sum_{n=1}^N (\alpha_n + \alpha_n^*) \right. \\ \left. + \sum_{n=1}^N y_n (\alpha_n - \alpha_n^*) \right) \\ \text{s. t. } \begin{cases} \sum_{n=1}^N (\alpha_n - \alpha_n^*) = 0 \\ \alpha_n, \alpha_n^* \in [0, C]. \end{cases}$$

Solving this dual optimisation problem for  $\alpha_n$  and  $\alpha_n^*$  facilitates the discovery of the support vectors, which may be applied to describe the support vector expansion of our predictor function  $\hat{f}(\mathbf{X})$ .

In detail, setting the partial derivation of the primal optimisation of  $\beta$  equal to zero, results in the expression

$$\beta = \sum_{n=1}^N (\alpha_n - \alpha_n^*) x_n$$

and we apply this expression to the linear model to obtain

$$\hat{f}(\mathbf{X}) = \sum_{n=1}^N (\alpha_n - \alpha_n^*) \langle \mathbf{x}_n, \mathbf{x} \rangle + \beta_0. \quad (5)$$

The constant  $\beta_0$  may be deduced from exploiting the Karush–Kuh–Tacker conditions. Its value may be derived from

$$\beta_0 = y_n - \langle \beta, x_n \rangle - \epsilon \quad \text{for } \alpha_n \in (0, C) \\ \beta_0 = y_n - \langle \beta, x_n \rangle + \epsilon \quad \text{for } \alpha_n^* \in (0, C).$$

The equation (5) denotes the so-called support vector expansion and describes a linear combinations of the training data. However, due to the constrains only some values  $(\alpha_n - \alpha_n^*)$  are nonzero and the corresponding observations denote the support vectors. The complexity of a function's representation by support vectors is independent of the dimensions of  $\mathbf{X}$ , but depends only on a limited number of support vectors. Furthermore the support vector expansion relies only on a dot product of data points and the kernel trick may therefore be applied to transfer data in a high dimensional feature space.

The prediction accuracy may be enhanced by mapping the observations  $\mathbf{X} \in \mathcal{X}$  in some feature space  $\mathcal{F}$  via some function  $\phi : \mathcal{X} \rightarrow \mathcal{F}$  and conduct the support vector regression in this feature space. However, the calculation of the dot product  $\langle \phi(x_n), \phi(x) \rangle$  of some very high dimensional vectors  $\phi(x)$  may be unfeasible to obtain in an acceptable time frame. The kernel trick

$$k(x_n, x) = \langle \phi(x_n), \phi(x) \rangle$$

resolves this issue as it leads to a calculation in the feature space without the need to actually compute the mapping in the feature space explicitly. This solution stems from the fact that certain kernel functions can be expressed as an inner product of vectors in some high dimensional space. In the thesis at hand the radial basis kernel function  $k(x_n, x_m) = \exp(-\gamma \|x_n - x_m\|^2)$  is applied to solve the support vector regression

$$f(x) = \sum_{n=1}^N (\alpha_n - \alpha_n^*) k(x_n, x) + \beta_0.$$

### 5.3 Implementation

We apply several prediction methods on the data in order to predict the offender's age from information obtained from the crime scene. We furthermore limit the set of predictors to information that could

be obtained in a criminal investigation. Obviously many variables applied in the generation of a BN are hidden to the police during their criminal investigation.

We rely thereby on several packages for R. The regression tree is calculated via the package `rpart` (Therneau and Atkinson, 1997) and the random forest via `randomForest` (Liaw and Wiener, 2002). We rely on the  $k$ -nearest-neighbor implementation from the `caret` package (Kuhn, 2008) and the support vector regression is calculated on a Gaussian radial basis kernel as implemented in `kernlab` (Karatzoglu et al., 2004). The simple linear model is generated via the package `MASS` (Venables and Ripley, 2002) and the penalized versions via `penalized` (Goeman, 2010).

As explained before we apply 10-fold cross-validation to set the tuning parameter and observe the prediction error. As our response variable is real, we search for the lowest root mean square error (RMSE) to decide on the tuning parameter in every model. We furthermore repeat the cross-validation 10 times to observe the distribution of the RMSE as we vary the allotment of observations into the 10 folders of cross-validation. This approach allows us to report a box plot as a final result instead of a single RMSE. In this process we make sure that every model obtains the same folders to ensure that any differences in the RMSE arise from the difference in modelling and do not arise from the allotment of observations into folders. This procedure is facilitated by the package `caret` (Kuhn, 2008).

Any missing values were imputed via a Gibbs sampling as implemented in the package `mice` (van Buren and Groothuis-Oudshoorn, 2010).

## 5.4 Results

We compare and evaluate the root mean square error

$$RMSE = \sqrt{\frac{1}{N} \sum_{n=1}^N \left( y_n - \hat{f}^{-\kappa(n)}(x_n) \right)^2}$$

derived from cross-validation for every model. Furthermore we generate a base line model, which adopts the simplest approach to prediction: Its model response constitutes in the sample mean  $\bar{Y} = \frac{1}{N} \sum_{n=1}^N Y_n$  without relying on information provided by the predictors  $\mathbf{X}$ . The mean age of an offender in our data is 29.93 years. This simple prediction method results in an RMSE of 0.303 for predicting the log of the offender's age, which our prediction methods, incorporating information from the predictors  $\mathbf{X}$ , outperform.

The prediction results are presented in Figure 14 in the appendix. We draw box plots resulting from repeated cross-validation and order the models according to their prediction performance, that is a small RMSE for predicting the logged offender's age. Furthermore in Table 4 we present the actual differences between the models and highlight all cells which report a significant difference as reported by a  $t$ -test on a 5% significance level.

As expected the base model performs worst in terms of the mean. However the regression tree exhibits the largest variance across the cross-validations performing much worse than the base model or on a par with the best models. This high variance is a typical behaviour of trees and was one of the reasons to create random forests. The next model is  $k$ -nearest-neighbours, which performs only marginally better than the base model or the regression tree. There is no significant difference between these three models and the poor performance of  $k$ -nearest-neighbour is surprising. However, this behaviour probably results from the chosen Euclidean distance to measure the distance between the observations. As most predictors are of nominal type, this

	tree	knn	rf	lmStepAIC	svr	ridge	lasso
base	0.002	0.006	0.029	0.031	0.033	0.039	0.039
tree		0.004	0.027	0.029	0.031	0.037	0.037
knn			0.023	0.026	0.027	0.033	0.034
rf				0.002	0.004	0.009	0.010
lmStepAIC					0.001	0.007	0.008
svr						0.006	0.007
ridge							0.001

**Table 4:** *Difference in means of the RMSE for every prediction approach. Significant differences on a 5% level are highlighted by grey background colour.*

distance measure may not be appropriate. However due to the general poor performance of the models, we don't assume an better suited distance measure to improve the performance of  $k$ -nearest-neighbours to a large extent.

All other models perform equally well with the exemption of random forest, which exhibits a significant worse performance than the best performer ridge regression and lasso. As can be seen in the Figure 14 the variance of random forests is decreased to a large extent if compared with a single regression tree. But also on average the performance of random forest outperforms a single regression tree significantly, although random forest consists of single trees, which are restricted in their choice of the optimal variable for every node. This aggregation of weak learners exhibits better results than a single optimized one of them. Our implementation of support vector regression does not significantly deviate in its performance from the best models. This performance however depends on the chosen kernel function, which we have optimised. Other kernel functions as for example a polynomial kernel do not perform as well as the implemented Gaussian radial basis function.

The linear model however perform best in predicting the offender's age. Even the simple linear model optimised via the *AIC* criterion does not deviate significantly from the more complicated penalized models. The lasso outperforms all other models although the performance of ridge regression can hardly be distinguished from the lasso. All linear models are quite restrictive in their assumptions as they only elaborate on a linear relationship between the predictors and the response variable. In most cases such a model marks an oversimplification and models being able to incorporate nonlinear relations like random forests or support vector regression outperform them on most data sets. However, linear models perform reasonable well, if the data set for training is small, sparsity arises or the data exhibits a low signal-to-noise ratio. With 252 observations for 25 predictors, our analysis may not be characterised by an excessive amount of data, but does include a reasonable amount of observations. So the good performance of linear models should arise due to sparsity or a low signal-to-noise ratio.

Observing the difference in RMSE resulting from comparing the base line model with the lasso, we note that the increase is rather minor. The base model exhibits a RMSE of 0.303, whereas the lasso lowers this value to 0.264. Performing the analysis on the untransformed response variable offender's age, this results in an increase of one year. The base line model includes a RMSE of about 9 years, whereas the lasso increases this performance to 8 years. Although this is a significant increase in terms of a *t*-test, it hardly matters for practise. No prediction model will help the police in catching an offender, if on average the prediction model is out in it's prediction by 9 or 8 years. There may be two reasons for this lack of performance. Firstly, the information drawn from a crime scene may not include a lot of information on the offender's age and therefore any model will fail. Secondly, the poor performance may result from a lack of data.

Sex-related homicides is a rather broad term including many different assaults and more data would be necessary to account for the heterogeneity. However, as explained in the beginning, data on sex-related homicides is hard to obtain and it may be infeasible to obtain a data set large enough to predict the offender's age reasonable well.

## 6 Discussion

Sex-related homicides arise wide media coverage and therefore extent the pressure on the police to catch the responsible offender. Although most cases are resolved rather quick, the police employs several specialists in offender profiling. These experts analyse the crime scene carefully and try to recover what has happened at the crime scene in great detail. If possible, they draw conclusion of the offender's characteristics from the knowledge gained from the crime scene. However, sex-related homicides occur infrequently and are of heterogeneous character and therefore knowledge drawn from empirical analysis is lacking. The thesis at hand therefore tries to contribute to this background knowledge by concentrating on the offender's age. Knowing the approximate age of an unknown offender constitutes a valuable information to the criminal investigation as the number of potential suspects is strongly reduced by this information.

We apply two different approaches. Firstly, a general structural learning approach without special emphasis on the offender's age and secondly we deliberately try to obtain a precise estimation of the offender's age from evidence found on the crime scene. The structural learning approach is based on graphical modelling. We make use of BN and learn a final graphical model by applying several structure learning algorithms to the data. Each algorithm presents a slightly different BN and we combine these BN to a single graphical model in which the edges' thickness describes how often an edge is found across the algorithms. This number indicates a level of confidence in the actual existence of a dependence between the corresponding variables.

In the second part of this thesis, we apply supervised learning in order to predict the offender's age. We apply several models, namely linear regression with a step procedure, ridge regression, lasso, regression tree, random forest,  $k$ -nearest neighbour and support vector regression. We optimise every model by obtaining the optimal value for

the corresponding tuning parameter as indicated by a 10-fold cross-validation. Cross validation also indicates the performance of every model on new data and the use the resulting expected test error to rank the models.

The BN indicates two type of crimes: An offender driven crime and a situation driven crime. In a situation driven crime the offender does not prepare the assault and consequently needs to apply brute force to gain control over the victim. The offender acts in familiar surroundings and is probably known to the victim. In a offender driven crime, the offender does plan the assault, but attacks in unfamiliar surroundings a victim, which he does not know. However there are may cases, which do not fit in these two classes.

The applied prediction techniques differed in their performance, but may be classified in two groups. Regression trees and  $k$ -nearest neighbour do not significantly deviate from the simple base line model. Whereas the application of all other models did improve the prediction criterion significantly. Whereas the poor performance of regression trees may be explained by its high variance, the performance of  $k$ -nearest neighbour is somewhat surprising. However, this probably results from the applied distance metric, which does not fit the data very well. All other models performed similarly, although random forest did worst and the application of the lasso resulted in the best prediction performance. However the actual increase in the performance in comparison with the simple base line model is small. It translates to an average error of 8 years instead of 9 years for the base model and the resulting decrease of one year does not suffice to adopt a data driven prediction approach in the criminal investigation. It may be worth noting, that an attempt to predict the geographical distance between the crime scene and the offender's personal hub from the same variables did also not succeed in gaining a acceptable performance. Although we don't present any details, we like to report

that a simple base model resulted in RMSE of about 79km, whereas the application of  $k$ -nearest neighbour resulted in a RMSE of about 60km. There may be two answers to this poor performance of prediction methods. Firstly, more data might be necessary, as sex-related homicides includes a wide range of unequal cases and secondly it might be concluded, that the evidence on the crime scene does not suffice for an assured and data-driven prediction of the offender's age and geographical distance.

---

## References

- Aitken, C. G. G., Gammerman, A., Zhang, G., Connolly, T., Bailey, D., Gordon, R. and Oldfield, R. (1996). Bayesian belief networks with an application in specific case analysis. In *Computational Learning and Probabilistic Reasoning* (ed A. Gammerman). Chichester: Wiley.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transaction on Automatic Control*, **19**, 716–723.
- Aliferis, C. F., Tsamardinos, I. and Statnikov, A. (2003a). HITON, a novel Markov blanket algorithm for optimal variable selection. In *Proceedings of the 2003 American medical informatics Association Annual Symposium*, 21–25.
- Aliferis, C. F., Tsamardinos, I., Statnikov, A. and Brown, L. E. (2003b). CausalExplorer: A Causal Probabilistic Network Learning Toolkit for Biomedical Discovery. In *Proceedings of the International Conference on Mathematics and Engineering Techniques in Medicine and Biological Science*, 371–376.
- Alison, L. , Bennell, C., Mokros, A. and Ormerod, D. (2002). The personality paradox in offender profiling *Psychology, Public Policy, and Law*, **8**, 115–135.
- Beauregard, É. (2007). The Role of Profiling in the Investigation of Sexual Homicide. In *Sexual Murderers: A Comparative Analysis and New Perspectives* (eds J. Proulx, É. Beauregard, M. Cusson and A. Nicole). Chichester: Wiley.
- Berkson, J. (1946). Limitation of the Application of Fourfold Table Analysis to Hospital Data. *Biometrics Bulletin*, **2**, 47–53.
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, **26**, 123–140.
- Breiman, L. (2001). Random forests. *Machine Learning*, **45**, 5–32.

- Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*. New York: Wadsworth
- van Buuren S. and Groothuis-Oudshoorn K. (2010). MICE: Multivariate Imputation by Chained Equations In *Journal of Statistical Software*, forthcoming.
- Cheng, J., Greiner, R., Kelly, J., Bell, D. A. and Liu, W. (2002). Learning Bayesian Networks from data. *The Artificial Intelligence Journal*, **137**, 43–90.
- Chickering, D. M. (1996). Learning Bayesian Networks is NP-complete. In *Learning from Data: Artificial Intelligence and Statistics V* (eds Fisher, D. and Lenz, H.–J.). New York: Springer.
- Clages, H. (2003). Erster Angriff. In *Handbuch der Kriminalistik* (eds Ackermann, H., Clages, H. and Roll, H.). Stuttgart: Richard Boorberg Verlag.
- Cover, T. and Hart, P. (1967). Nearest neighbour pattern classification. *IEEE Transaction on Information Theory*, **IT-11**, 21–27.
- Davies, A. (1997). Specific profile analysis: a data-based approach to offender profiling. In *Offender profiling: Theory, Research and Practise* (eds Jackson, J. L. and Bekerian, D. A.). Chichester: Wiley.
- Dern, H., Frönd, R., Straub, U., Vick, J. and Witt, R. *Geografisches Verhalten fremder Täter bei sexuellen Gewaltdelikten*. Wiesbaden: BKA.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, **76**, 378–382.
- Friedman, N., Linial, M., Nachman, I. and Peer, D. (2000). Using Bayesian Networks to Analyze Expression Data. *Journal of Computational Biology*, **7**, 601–620.

- Friedman, N., Nachman, I. and Peer, D. (1999). Learning Bayesian Network Structure from massive Datasets. *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 206–215.
- Goeman, J. J. (2011). L1 penalized estimation in the Cox proportional hazards model. *Biometrika Journal*, **52**, 72–84.
- Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*. New York: Springer.
- Heckerman, D. (1998). A tutorial on learning with Bayesian networks. In *Learning in Graphical Models* (ed M. I. Jordan). Dordrecht: Kluwer.
- Heckerman, D. (1990). Probabilistic similarity networks. *Networks*, **20**, 607–636.
- Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **42**, 80–86.
- Jensen, F. V. (1996). *Introduction to Bayesian Networks*. New York: Springer.
- Kalisch, M. and Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-Algorithm. *Journal of Machine Learning Research*, **8**, 613–636.
- Karatzoglou, A., Smola, A., Hornik, K. and Zeileis, A. (2004). kernlab - An S4 Package for Kernel Methods. *Journal of Statistical Software*, **11**, 1-20.
- Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, **28**, 1–26.
- Lauritzen, S. L., Dawid, A. P., Larsen, B. N. and Leimer, H. G. (1990). Independence properties of directed markov fields. *Networks*, **20**, 491–505.
- Lauritzen, S. L. (1996). *Graphical Models*. Oxford: Clarendon Press.

- Li, J. and Wang, Z. J. (2009). Controlling the false discovery rate of the association/causality structure learned with the PC algorithm. *Journal of Machine Learning Research*, **10**, 475–514.
- Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *R News* **2**, 18–22.
- Margaritis D. and Thrun, S. (1999). Bayesian Network induction via local neighborhoods. In *Advances in Neural Information Processing Systems 12* (eds Solla, S. A., Leen, T. K. and Müller, K.-R.). Cambridge: MIT Press.
- Meek, C. (1995). Causal inference and causal explanation with background knowledge In *Uncertainty in Artificial Intelligence 11* (eds Besnard P. and Hanks, S.). San Francisco: Morgan Kaufmann.
- Meinshausen, N. and Bühlmann, P. (2006). High-Dimensional graphs and variable selection with the LASSO. *The Annals of Statistics*, **34**, 1436–1462.
- Meinshausen, N. and Bühlmann, P. (2010). Stability Selection. *Journal of the Royal Statistical Society: Series B*, **72**, 417–473.
- Miethe, T. D. and Regoeczi, W. C. (2004). *Rethinking Homicide*. New York: Cambridge University Press.
- Mokros, A. and L. J. Alison (2002). Is offender profiling possible? *Legal and Criminological Psychology*, **7**, 25–43.
- Pearl, J. (2000). *Causality*. New York: Cambridge University Press.
- Ressler, R. K., Burgess, A. W. and Douglas, J. E. (1988). *Sexual homicide*. New York: Lexington Books.
- Ripley, B. (1996). *Pattern Recognition and Neural Networks* Cambridge: Cambridge University Press

- Robinson, R. W. (1977). Counting unlabelled acyclic digraphs. In *Lecture Notes in Mathematics: Combinatorial Mathematics V*. Heidelberg: Springer.
- Salfati, G. and Canter, D. V. (1999). Differentiating stranger murders: profiling offender characteristics from behavioral styles. *Behavioral Sciences and the Law*, **17**, 391–406.
- Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.
- Learning Bayesian Networks with the bnlearn R package. *Journal of Statistical Software*, **35**, 1–22.
- Sprites, P., Glymour, C. and Scheines, R. (2000). *Causation, prediction and Search*. Cambridge: MIT Press.
- Stahlschmidt, S., Tausendteufel, H. and Härdle, W. K. (2011). Bayesian Networks and Sex-related Homicides. *SFB 649 Discussion Paper*, 2011-045, Berlin: Humboldt–Universität zu Berlin
- Strauss, A. and Corbin, J. M. (1990). *Basics of qualitative research*. Thousand Oaks: Sage Publications.
- Tausendteufel, H., Stahlschmidt, S. and Kühnel, W. (2011). *Bestimmung des Täteralters bei sexuell assoziierten Tötungsdelikten auf der Basis von Tatgeschehensmerkmalen*. Wiesbaden: Bundeskriminalamt.
- Therneau, T. M. and Atkinson, E. J. (1997). An introduction to recursive partitioning using the rpart routine. *Technical Report 61*, Mayo Clinic, Section of Statistics.
- Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267–288.
- Tsamardinos, I., Aliferis, C. F. and Statnikov, A. (2003). Algorithms for large scale markov blanket discovery. In *The 16th International FLAIRS Conference*, 376–381.

- Tsamardinos, I., Brwon, L. E. and Aliferis, C. F. (2006). The max–min hill–climbing Bayesian Network structure learning algorithm. *Machine Learning*, **65**, 31–78.
- Vapnik, V. (1995). *The nature of Statistical Learning Theory*. New York: Springer.
- Venables, W. N. and Ripley B. D. (2002). *Modern Applied Statistics with S*. New York: Springer.
- Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. In *Proceedings of the Sixth Conference on Uncertainty in Artificial intelligence*, 220–227.
- Verma, T. and Pearl, J. (1992). An algorithm for deciding if a set of observed independencies has a causal explanation. In *Proceedings of the Eighth Conference on Uncertainty in Artificial intelligence*, 323–330.
- Wright, S (1921). Correlation and Causation. *Journal of Agricultural Research*, **20**, 558–585.
- Zuk, O., Margel, S. and Domany, E. (2006). On the number of samples needed to learn the correct structure of a Bayesian Network. In *UAI 2006*.

# A Appendix

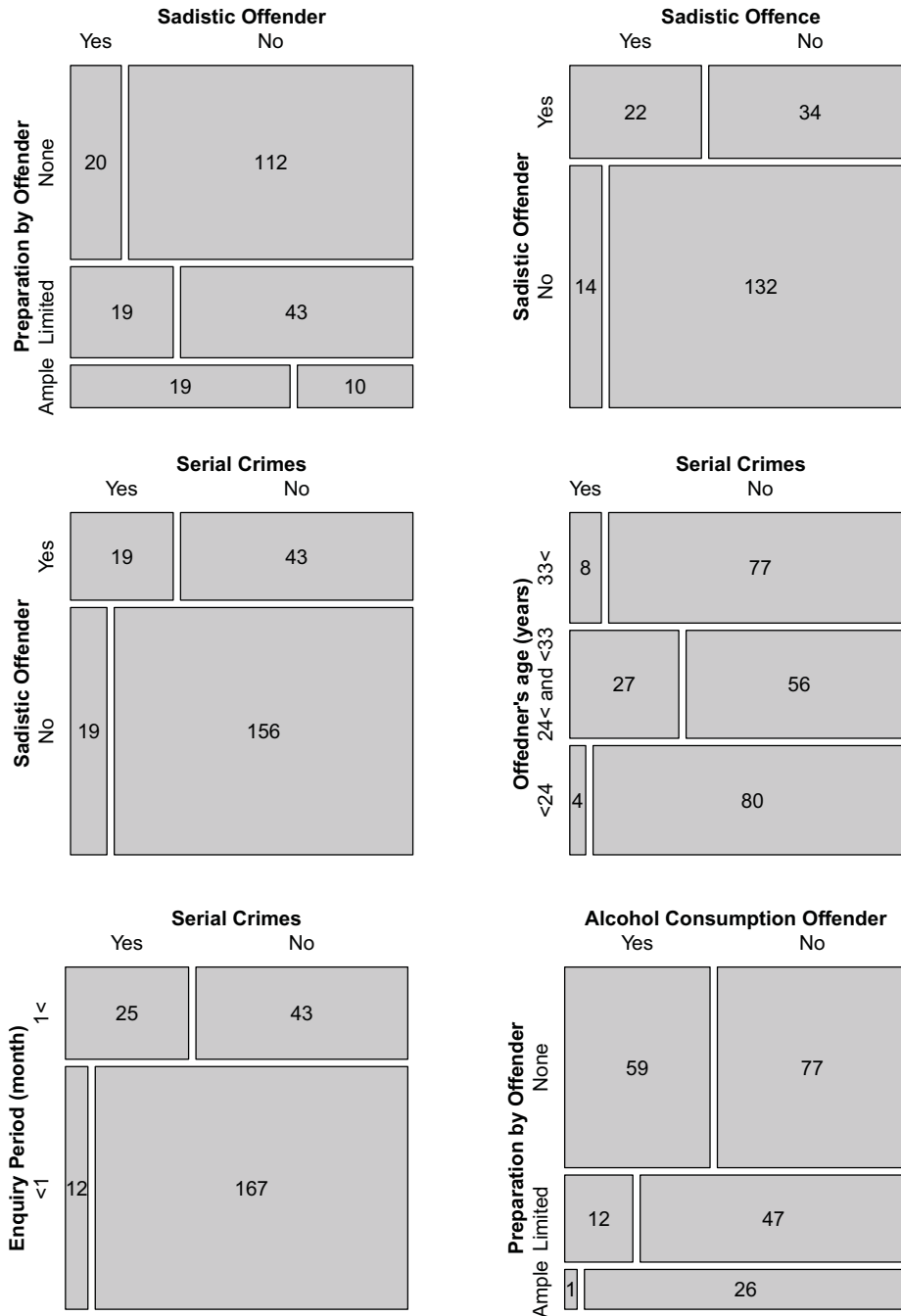


Figure 9: Mosaic plots corresponding to discussed edges

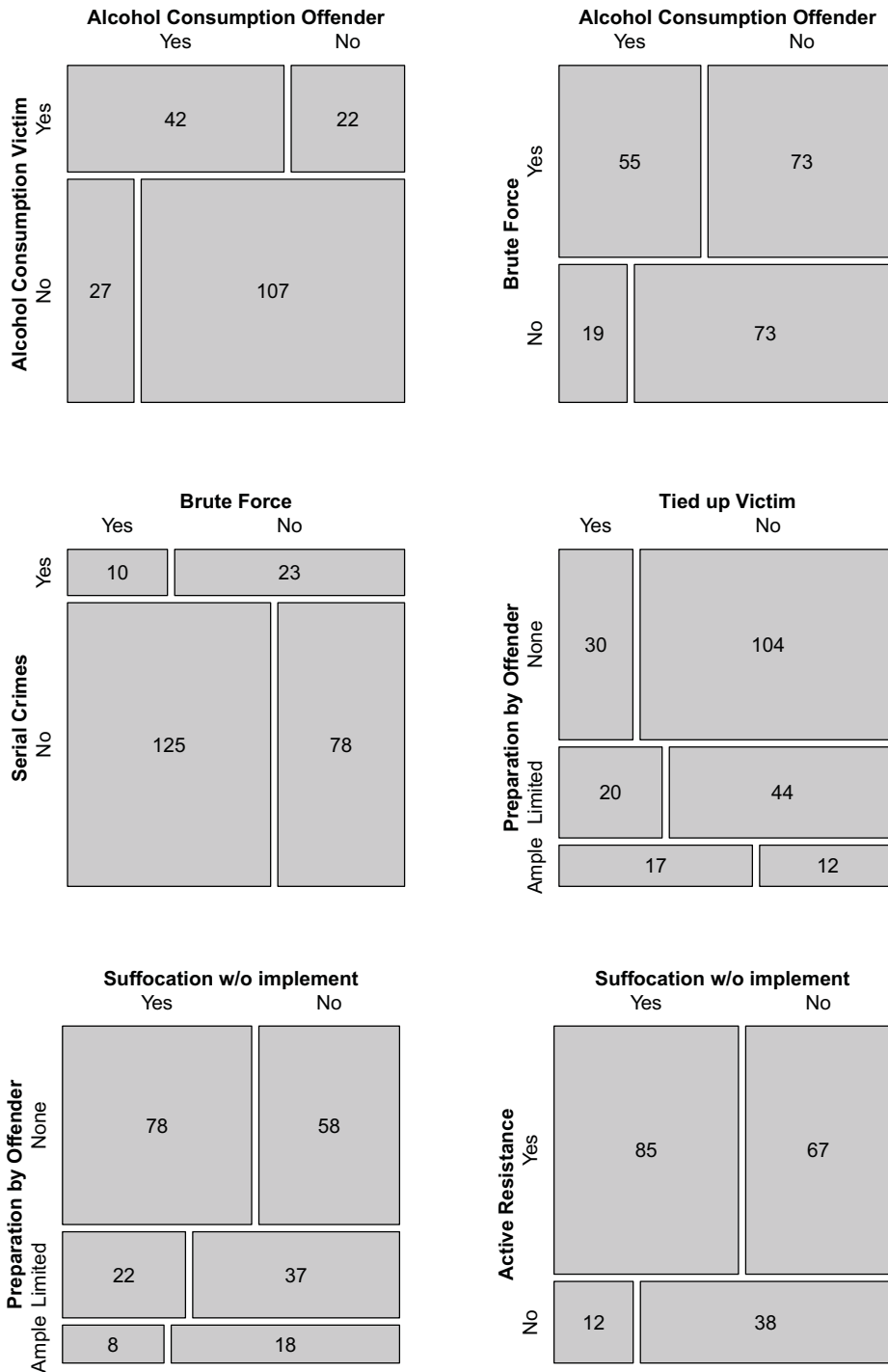


Figure 10: Mosaic plots corresponding to discussed edges

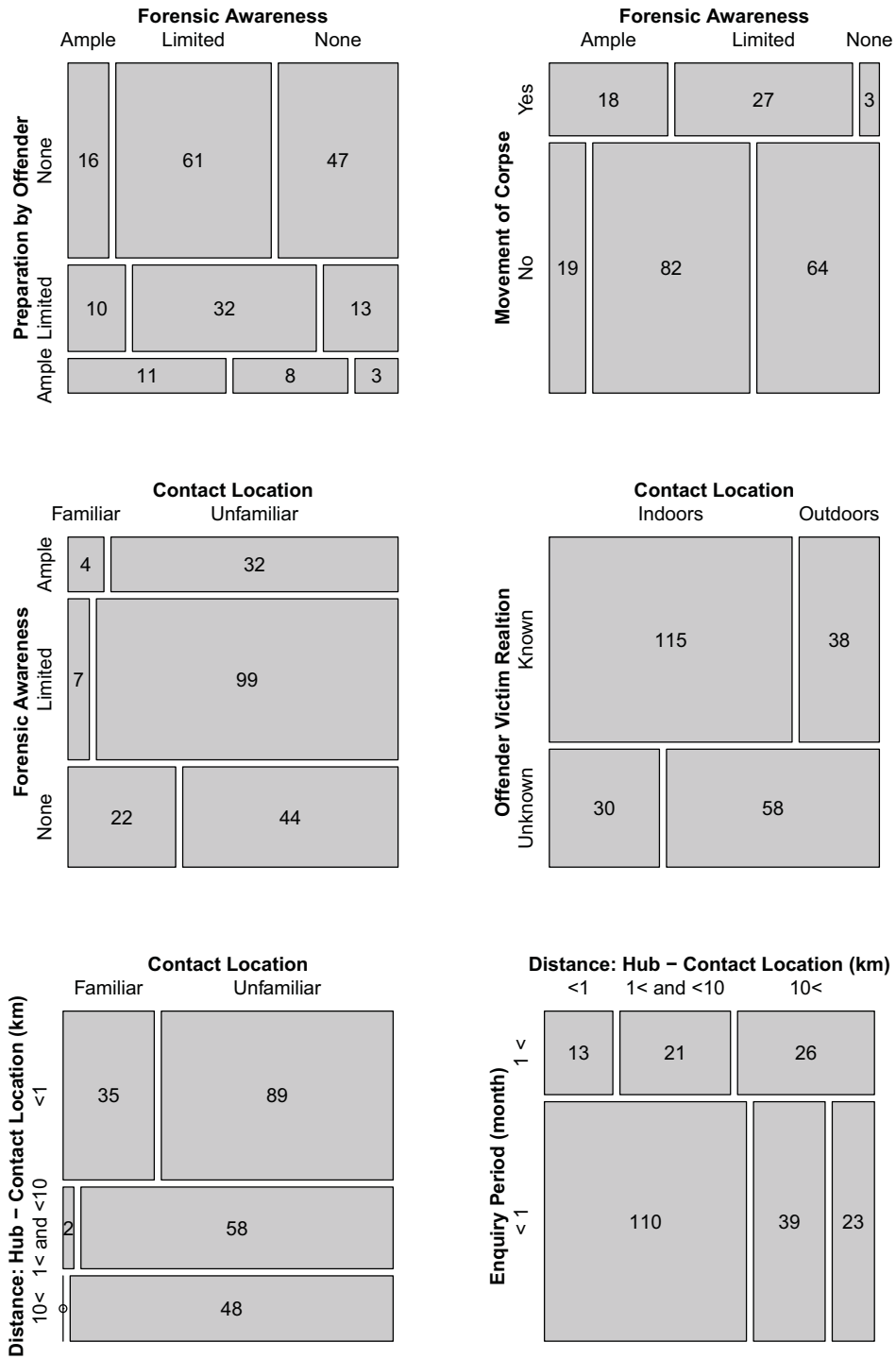


Figure 11: Mosaic plots corresponding to discussed edges

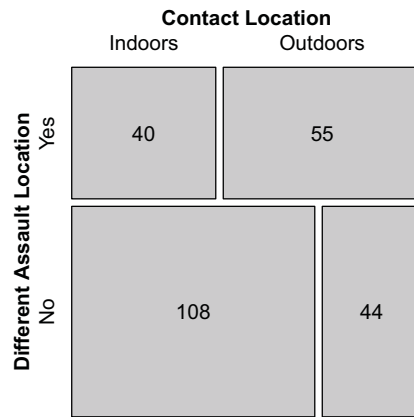
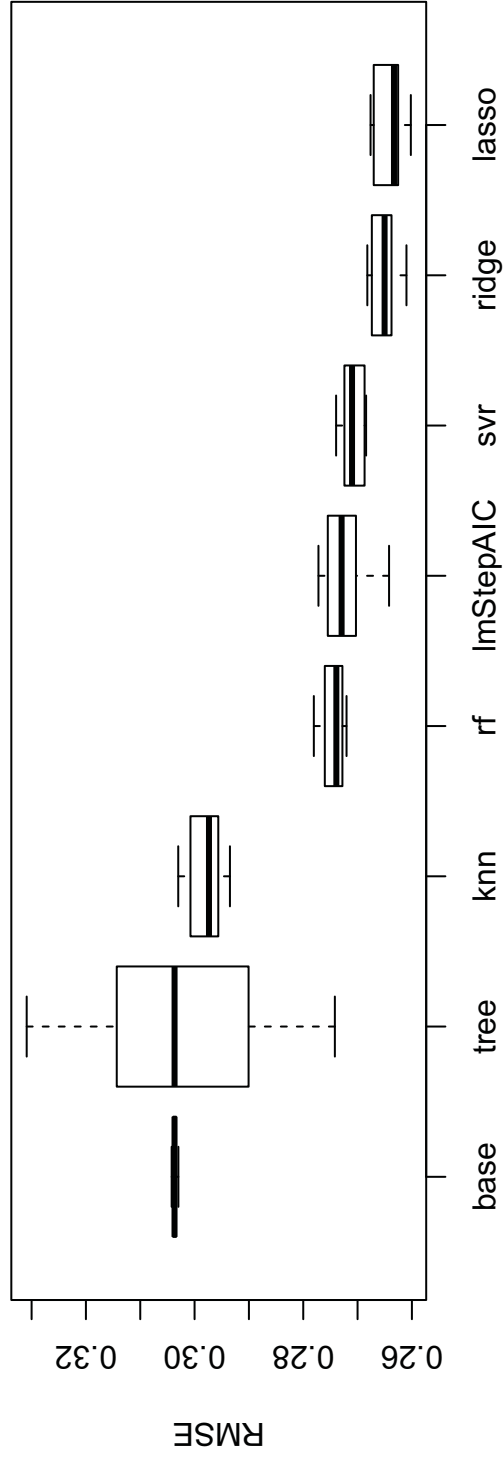


Figure 12: *Mosaic plots corresponding to discussed edges*





**Figure 14:** Comparison of RMSE for the prediction of  $\log(\text{offender's age})$  by the base model, regression tree,  $k$ -nearest-neighbour, random forest, linear model, support vector regression, ridge regression and lasso. The distributions indicated by box plots arise from repeated cross-validation.

Variable name	level	description
Gender of victim	binary	Which gender does the victim have?
Suffocation with implement	binary	Was the victim suffocated with an implement?
Demand sex. acts	binary	Did the offender demand sexual acts from the victim?
Fellatio	binary	Did the victim perform fellatio?
Age of victim	nominal	Which age group does the victim belong to?
Provocation: financial dispute	binary	Did any dispute arise from money issues before the attack?
Nationality of victim	binary	Is the victim German?
Alcohol consumption by victim	binary	Did the victim consume alcohol before the attack?
Intention of robbery	binary	Did the offender report a primary intention in robbery?
Criminal record: violence	binary	Has the offender been convicted for violence?
Inserted finger	binary	Did the offender insert a finger in the victim's sexual organs?
Criminal record: against property	binary	Has the offender been convicted for theft?
Number of offenders	binary	Did the offender act alone?
Brute force	binary	Did the offender apply blunt force to control the victim?
Enquiry period	binary	Did the criminal investigation take more than one month?
Provocation: insult	binary	Did the offender feel insulted by the victim prior to the assault?
Kissing	binary	Did the offender try to kiss the victim?
Robbery	binary	Did the offender steal belongings of the victim?
Provocation: Sex. demands	binary	Did the offender feel provoked by sexual demands of the victim?
Sex. frustration	binary	Did the offender feel sexually frustrated?
Consensual sex	binary	Did the victim and the offender engage in voluntary sex before the attack?
Alcohol consumption by offender	binary	Did the offender consume alcohol before the attack?

**Table 5:** *Description of variables for graphical modelling.*

Variable name	level	description
Disguised identity	binary	Did the offender try to hide his identity from the victim?
Crucial personal circumstances	binary	Was the offender affected by extraordinary events in his private life?
Offender: criminal environment	binary	Did the offender live in a criminal environment?
Suffocation without implement	binary	Did the offender suffocate the victim with his bare hands?
Sex. manipulation of corpse	binary	Did the offender conduct any se related manipulation on the corpse?
Serial crime	binary	Does the crime include a serial offender?
Offender victim relationship	nominal	Relationship status between offender and victim
Prostitution	binary	Does the victim work as a prostitute?
Passive resistance	binary	Did the victim obstruct the offender's actions in a passive form?
Appraisal of diminished responsibility	binary	Did the judge in the sentence acknowledge diminished responsibility
Deception	binary	Did the offender apply deception to gain control?
Preparation by offender	binary	Did the offender prepare the assault?
Active resistance	binary	Did the victim obstruct the offender's actions in a active form?
Sadism	binary	Did the offender act sadistically on the crime scene?
Age of offender	nominal	To which age group does the offender belong?
Distance: hub – contact location	nominal	Which distance traveled the offender before his encounter with the victim?
Threat	binary	Did the offender threat the victim to gain control?
Criminal record: sex offences	binary	Has the offender been convicted for sex offences?
Intercourse with corpse	binary	Did the offender engage in sexual intercourse with the corpse?
Provocation: break-up	binary	Did the offender fell provoked by a separation from the victim?
Sadistic offender	binary	Is the offender a sadist (via psychiatric examination)
Forensic awareness	nominal	To which extent does the offender try to hide the crime and traces?

**Table 6:** *Description of variables for graphical modelling.*

Variable name	level	description
Offender: permanent relationship	binary	Does the offender have a stable relationship with a spouse?
Tied up victim	binary	Did the offender tie up the victim to maintain control?
Anal intercourse	binary	Did the offender engage in anal intercourse with the victim?
Vaginal intercourse	binary	Did the offender engage in vaginal intercourse with the victim?
Offender: socially marginalised	binary	Did the offender belong to a socially marginalised group?
Offender: social isolation	binary	Was the offender socially isolated?
Movement of corpse	binary	Did the offender hide the corpse at a separate crime scene?
Contact location	nominal	Description of the contact location between offender and victim
Different assault location	binary	Victim and offender change the location before the offender's attack

**Table 7:** *Description of variables for graphical modelling.*

Variable name	level	description
t_gefaengnisvorerfahrung_norm	binary	Has the offender been sent to prison?
tv_sozsit_k	binary	Did the offender live in a criminal environment?
o_tatalter	real	The offender's age
toi_krim_spur_kond	binary	Did the police find a condom on the crime scene?
toie_dauer_kue	real	How much time passed between the initial meeting and the assault?
xu_zerlaubnis	binary	Was the offender invited to the assault location by the victim?
xu_zlist	binary	Did the offender apply deception to gain control?
xu_zoffentlich	binary	Was the assault location a publicly available?
toit1_sex_vaginal	binary	Did the offender engage in vaginal intercourse with the victim?
toit2_sex	binary	Did the offender engage on sexual activities with the corpse?
x_spurendna	binary	Did the police find DNA traces on the crime scene?
x_spurenpersdinge	binary	Did the police find any belongings of the offender?
x_l_loksonstige_Tatorte_drinnen	binary	Did the offender leave the corpse at an unspecified indoors location?
x_l_lokWohnung-Taeter	binary	Did the offender leave the corpse at his flat?
x_k_loksonstige_Tatorte_draussen	binary	Did the initial contact happen at an unspecified outdoors location?
x_k_lokWohnung-Taeter	binary	Did the initial contact happen at the offender's flat?
x_u_loksonstige_Tatorte_draussen	binary	Did the assault take place at an unspecified outdoors location?
x_u_loksonstige_Tatorte_drinnen	binary	Did the assault take place at an unspecified indoors location?
x_kungleichuNein	binary	Victim and offender change the location before the offender's attack
x_ort_fallgemischt	binary	Did the crime happen at indoors and outdoors locations?
xl_vgeschlecht_norm2	binary	Are there any injuries on sexual organs to be found?
xl_sposition	binary	Did the police find the victim in some extraordinary position?
toi_enteigen	binary	Did the offender steal belongings of the victim?
xl_vextremitaeten_norm2	binary	Are there any injuries on extremities?

**Table 8:** *Description of variables for prediction.*