TECHNISCHE UNIVERSITÄT WIEN
Vienna | Austria

DISSERTATION

From blackspots to blackpatterns:
Pattern recognition with road traffic accident data. Illustrated with single-vehicle accidents with a single occupation and personal injury that occurred outside the built-up area on the Austrian road network between 2012 and 2019

ausgeführt zum Zwecke der Erlangung des akademischen Grades eines
Doktors der technischen Wissenschaften

unter der Leitung von

Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Georg Hauger
Forschungsbereich Verkehrssystemplanung | IVS | TU Wien

und der Begutachtung durch

em. Univ.Prof. Dipl.-Ing. Dr.techn. Gerd Sammer
Institut für Verkehrswesen | IVe | BOKU Wien

Professore ordinario Silvio Nocera, PhD CE
Department of Architecture and Arts | Università Iuav di Venezia

eingereicht an der Technischen Universität Wien
Fakultät für Architektur und Raumplanung
von

Dipl.-Ing. Tabea Fian
01025833
Grabengasse 3/5, 2500 Baden

Baden, am 19.12.2021 Tabea Fian

Abstract

Besides the designation of a major accident cause and accident blackspots (i.e., accident accumulation points on the road network), we currently face a knowledge gap in the multivariate statistical investigation of co-occurring accident conditions. Official road traffic accident statistics in Austria indicate one explicit accident cause (or one explicit condition) for each road traffic accident (e.g., speeding). However, investigating co-occurring conditions (e.g., 'speeding', 'wet road', 'no safety belt applied' and 'probationary driving licence') is essential if we consider accidents as multicausal instead of monocausal events. It is, of course, impossible to depict all potential accident-related conditions. Still, the official Austrian road traffic accident database (UDM) provides a solid source to identify co-occurring accident-related variables. The UDM includes more than 100 accident-related variables, which can help understand accident conditions and causes in more detail. In-depth knowledge of accident conditions may be of interest in deriving (target-group specific) prevention measures to deal with the remaining number of fatal and severe road traffic accidents in Austria. Therefore, this thesis aims to detect recurring combinations of accident-related variables, which we designate as blackpatterns.

Consequently, this thesis applies a pattern recognition approach among single-vehicle accidents with single occupation and personal injury that occurred on the Austrian road network and outside the built-up area between 2012 and 2019 (n=20.293). It uses driver-, vehicle-, roadway- and situation-related variables to detect recurring variable combinations (blackpatterns). These variables (over 100 in total) are part of the official Austrian road traffic accident database (UDM). However, reprocessing the official database is essential to conduct pattern recognition methods with the data. It is to point out that this thesis explores blackpatterns underlying historical road traffic accident records. This thesis does not present an accident prediction model. It does not include data on traffic performance to derive statements on the overall probability of a road traffic accident.

The motivation of this thesis is to focus on the general applicability of the proposed methods. Firstly, we point out statistical characteristics of road traffic accident data (i.e., uncertainty, noise and bias, rare events, heterogeneity, and over-dispersion). Secondly, we discuss existing pattern recognition methods for road traffic accident data. Thirdly, we apply selected pattern recognition methods on the road traffic accident sample. These methods comprise binomial logistic regression, decision trees, Bayesian networks and a developed pattern recognition method based on the frequencies of variable combinations (PATTERMAX-method).

In a primary step, we conduct descriptive statistical analyses to estimate the relationship between each recorded accident-related variable and the target variable severe casualties (accidents with fatal or severe injury). We create contingency tables, calculate conditional and joint probabilities, apply Fisher's exact test and estimate the Phi coefficient. Also, we generate

a robust parameter estimation (95% confidence intervals showing the likelihood of a variable and severe or fatal accidents to occur) by applying a bootstrap resampling method on the newly established accident database. We calculate a so-called maximum combination value as an important measure towards blackpattern detection. This value tells us how often a specific variable co-occurs with (an)other accident-related variable(s). We then use binomial logistic regression to estimate each variable's impact on severe road traffic accidents with an odds ratio (i.e., the strength of the relationship between an accident-related variable and the target variable severe casualties compared to all observed variables). By knowing which variable appears to increase the risk of a severe road traffic accident, we can assess the overall impact of the detected blackpatterns.

As the next step towards blackpattern recognition, we grow decision trees using the CHAID-algorithm. Up to this point, binomial logistic regression and decision trees help us identify critical variables that aggravate an accident outcome and the degree of injury, respectively. However, since we are interested in gaining in-depth knowledge of recurring variable combinations (blackpatterns), we zoom further into the underlying data structures.

That being the case, we apply a probabilistic Bayesian network paradigm and a developed pattern detection method (PATTERMAX-method) to the data. Using these approaches, we finally detect blackpatterns and conclude the pattern recognition process with a statistical evaluation of whether the detected blackpatterns show a significant relationship with the target variable severe casualties. Like the beginning, so the end, and we calculate Fisher's exact test and the Phi coefficient.

We summarize the most aggravating accident-related variables and blackpatterns in the discussion chapter. Furthermore, we compare the applied pattern recognition methods. Finally, we highlight the advantages and limitations of the PATTERMAX-method in combination with binomial logistic regression to gain in-depth knowledge about accident circumstances. The combined application of both methods enables a precise detection and comparison of blackpatterns. For example, do blackpatterns among female drivers differ from blackpatterns among male drivers? Do accident patterns on regional roads within an 80 km/h speed limit differ from those on a 100 km/h speed limit? Additionally, the combined approach of the PATTERMAX-method and binomial logistics regression enables the assessment of the detected blackpatterns with the help of an odds ratio.

Within the research outlook, we propose expanding the investigation towards accidents with several parties involved. The newly established accident database might also serve as a reliable source for accident prediction. Especially, the estimated 95% confidence intervals may be of interest to establish a prediction model.

Kurzfassung

Neben der Benennung von Hauptunfallursachen und Unfallschwerpunkten im Straßennetz gibt es derzeit eine Wissenslücke bei der multivariaten statistischen Untersuchung von gemeinsam auftretenden Unfallbedingungen. Die amtliche Straßenverkehrsunfallstatistik in Österreich weist für jeden Straßenverkehrsunfall eine explizite Unfallursache (oder eine explizite Bedingung) aus (z.B. Geschwindigkeitsüberschreitung). Die Untersuchung von gleichzeitig auftretenden Bedingungen (z.B. "Geschwindigkeitsüberschreitung", "nasse Fahrbahn", "nicht angeschnallt" und "Probeführerschein") ist jedoch unerlässlich, wenn wir Unfälle als multikausale und nicht als monokausale Ereignisse betrachten. Es ist zwar nicht möglich alle möglichen Unfallbedingungen abzubilden, aber die offizielle österreichische Straßenverkehrsunfalldatenbank (UDM) bietet eine solide Quelle für die Identifizierung von gemeinsam auftretenden, unfallbezogenen Variablen. Die UDM enthält mehr als 100 unfallrelevante Variablen, die helfen können, Unfallbedingungen und -ursachen genauer zu verstehen. Ein vertieftes Wissen über die Unfallbedingungen kann von Interesse sein, um (zielgruppenspezifische) Präventionsmaßnahmen abzuleiten, um die verbleibende Zahl der tödlichen und schweren Straßenverkehrsunfälle in Österreich zu reduzieren. Ziel dieser Arbeit ist es, wiederkehrende Kombinationen von unfallbeschreibenden Variablen zu erkennen, die wir als Variablenmuster (blackpatterns) bezeichnen.

Diese Arbeit wendet daher einen Mustererkennungsansatz bei Unfällen mit einem Fahrzeug mit Einzelbesetzung und Personenschaden an, die sich zwischen 2012 und 2019 auf dem österreichischen Straßennetz außerorts ereignet haben (n=20.293). Es werden fahrer-, fahrzeug-, straßen- und situationsbezogene Variablen verwendet, um wiederkehrende Variablenkombinationen (blackpatterns) zu erkennen. Diese Variablen (insgesamt über 100) sind Teil der offiziellen österreichischen Straßenverkehrsunfalldatenbank (UDM). Um mit den amtlichen Daten Mustererkennungsmethoden durchführen zu können, ist jedoch eine Neuaufbereitung der amtlichen Datenbank notwendig. Die Neuaufbereitung der Datenbank stellt daher einen zentralen Bestandteil dieser Arbeit dar. Es ist wichtig hervorzuheben, dass in dieser Arbeit historische Straßenverkehrsunfälle untersucht werden und kein Unfallvorhersagemodell vorgestellt wird. Die Arbeit bezieht auch keine Daten zum Verkehrsgeschehen oder zur Verkehrsleistung ein. Es können daher keine Aussagen über die generelle Eintrittswahrscheinlichkeit eines Straßenverkehrsunfalls abgeleitet werden.

Die Motivation dieser Arbeit ist es, sich auf die allgemeine Anwendbarkeit der vorgeschlagenen Methoden zu konzentrieren. Zunächst wird auf die statistischen Eigenschaften von Straßenverkehrsunfalldaten hingewiesen (d.h. Unsicherheit, der sogenannte ‚evaluation bias', seltene Ereignisse, Heterogenität etc.). Zweitens werden bestehende Mustererkennungsmethoden für Straßenverkehrsunfalldaten diskutiert. Drittens werden ausgewählte Mustererkennungsmethoden auf die Stichprobe der Straßenverkehrsunfälle angewandt. Diese Methoden umfassen binomiale logistische Regression, Entscheidungsbäume,

Bayes'sche Netze und eine entwickelte Mustererkennungs-methode, die auf den Häufigkeiten von Variablenkombinationen basiert (PATTERMAX-Methode).

Zunächst werden deskriptive statistische Analysen durchgeführt, um die Beziehung zwischen jeder erfassten unfallbezogenen Variable und der Zielvariable „schwere Unfälle" (das sind Unfälle mit tödlichen oder schweren Verletzungen) zu schätzen. Es werden Kontingenztabellen erstellt, bedingte und gemeinsame Wahrscheinlichkeiten berechnet, der exakte Test nach Fisher angewandt und Phi-Koeffizienten geschätzt. Außerdem wird eine robuste Parameterschätzung durchgeführt (95 %-Konfidenzintervalle, welche die Wahrscheinlichkeit des Auftretens einer Variablen und schwerer Unfälle angeben), indem ein Bootstrap-Resampling-Verfahren auf die neu erstellte Unfalldatenbank angewandt wird. Weiters wird ein sogenannter höchster Kombinationswert als wichtiges Maß für die Erkennung von Variablenmustern berechnet. Dieser Wert gibt an, wie oft eine bestimmte Variable mit (einer) anderen unfallbezogenen Variable(n) gemeinsam vorkommt. Anschließend wird eine binomiale logistische Regression durchgeführt, um den Einfluss jeder Variable auf schwere und tödliche Straßenverkehrsunfälle mit einem Odds Ratio zu schätzen (d. h. die Stärke der Beziehung zwischen einer unfallbezogenen Variable und der Zielvariable „schwere Unfälle" im Vergleich zu allen beobachteten Variablen). Mit den Schätzungen, welche Variable das Risiko eines schweren oder tödlichen Straßenverkehrsunfalls zu erhöhen scheint, kann anschließend die Gesamtwirkung der noch zu entdeckenden Variablenmuster (blackpatterns) eingestuft werden. Als nächsten Schritt zur Erkennung von Variablenmustern werden Entscheidungsbäume mit dem CHAID-Algorithmus erstellt. Bis zu diesem Punkt helfen die binomiale logistische Regression und die Entscheidungsbäume dabei, kritische Variablen zu identifizieren, die den Unfallhergang bzw. den Grad der Verletzung erhöhen. Da der Fokus jedoch darauf liegt, vertiefte Kenntnisse über wiederkehrende Variablenkombinationen zu erlangen, werden die zugrunde liegenden Datenstrukturen noch tiefer analysiert. Zu diesem Zweck werden Bayes'sches Netzwerke und eine entwickelte Methode zur Mustererkennung (PATTERMAX-Methode) auf die Daten angewandt. Mit diesen Ansätzen werden schließlich wiederkehrende Variablenkombinationen detektiert. Die statistische Auswertung, ob die detektierten Muster einen signifikanten Zusammenhang mit der Zielvariablen „schwere Unfälle" aufweisen, schließt den Mustererkennungsprozess ab. Wie der Anfang, so das Ende, und es werden der exakte Test nach Fisher und der Phi-Koeffizient dazu verwendet.

Im Diskussionskapitel werden die schwerwiegendsten unfallbezogenen Variablen und Muster zusammengefasst. Außerdem werden die angewandten Mustererkennungsmethoden diskutiert. Abschließend werden Vorteile und Grenzen der PATTERMAX-Methode in Kombination mit der binomialen logistischen Regression aufgezeigt, um vertiefte Erkenntnisse über das Unfallgeschehen zu gewinnen. Im Rahmen des Forschungsausblicks wird die Ausweitung der Methoden auf Unfälle mit mehreren Beteiligten vorgeschlagen. Die neu erstellte Unfalldatenbank könnte auch als zuverlässige Quelle für die Unfallvorhersage dienen. Insbesondere die geschätzten 95%-Konfidenzintervalle könnten für die Erstellung eines Vorhersagemodells von Interesse sein.

Content

1. Introduction

1.1 Relevance and problem statement

Road traffic accidents result in substantial material and immaterial costs. Material costs, for example, include damage to vehicles, administrative costs, or medical costs. Immaterial costs refer to shorter lifetimes, sufferings, pain, or sorrow (European Commission, 2020, p. 38). The Austrian Accident Cost Accounting from 2017 (Herry Conuslt & KFV, 2017, p. 4) specifies the economic costs of one single fatal road traffic accident with 3.313.309 euros and accidents costs per severe injury with 429.517 euros (see table 1). Thus, traffic safety and the definition of appropriate prevention measures are significant issues in almost all traffic policy plans, such as the Austrian Road Safety Strategy 2030 (KFV & FGM, 2021). Therefore, information about accident accumulation points (i.e., blackspots), accident causes, and co-occurring conditions leading to fatal or severe accidents are essential for deducing appropriate prevention measures and policy decisions. However, when it comes to analysing conditions conducive to fatal and severe road traffic accidents (i.e., recurring accident patterns), we experience a knowledge gap.

			2016 at 2016 prices	2011 at 2011 prices	2004 at 2004 prices	1993 at 1993 prices
Total Accident Costs	including human suffering	Mio. EUR	9.701	10.088	10.518	
	excluding human suffering	Mio. EUR	5.203	5.278	5.184	3.818
Accident costs per fatality	including human suffering	EUR	3.316.309	3.016.194	2.461.345	
	excluding human suffering	EUR	1.390.800	1.401.085	1.287.004	805.233
Accident costs per severe injury	including human suffering	EUR	429.517	381.480	291.275	
	excluding human suffering	EUR	87.097	80.166	55.925	43.065
Accident costs per light injury	including human suffering	EUR	30.575	26.894	20.896	
	excluding human suffering	EUR	4.235	3.716	2.792	3.695
Material damage per accident		EUR	5.481	5.245	4.075	

Table 1: Total Accident Costs in Austria. Source: Herry Conuslt & KFV, 2017, p. 4

The definition of appropriate prevention measures to reduce the remaining number of fatal and severe accidents is challenging. Until today, besides accident causes, in-depth knowledge about recurring accidents is still missing. The thesis's hypothesis assumes that road traffic accidents do not represent monocausal events but a complex interplay between road users, vehicles, infrastructure, and the environment. Superordinate framework conditions (e.g., traffic policy, StVO, safety culture etc.) and how police officers record accidents strongly influence this interplay. For example, assessing the alleged main cause of the accident represents a subjective assessment by the police officer who fills out the accident data sheet on site. Depending on how differently police officers may be trained on accident surveys, there always exists a so-called evaluation bias going along with road traffic accident records. This thesis does not examine these superordinate parameters. The focus is placed exclusively on examining the officially available traffic accident data. However, we will point out challenges going along with road traffic accident data (see chapter 2.1).

Thus, this thesis ties in with a detailed investigation of recurring accident conditions. In addition to the cause of the accident, the thesis presents an exploratory research approach to analyse recorded accident-related variables. It investigates whether there exist recurring combinations of these variables (blackpatterns). Accident-related variables for single-vehicle accidents include driver-, vehicle-, roadway-, and situation-related variables. Other involved road users (pedestrians, cyclists, passengers) represent additional variables in other accident types. Thus, they do not occur in our selected road traffic accident sample (chapter 3.2 substantiates the sample selection).

The combined analysis of these variables is in accord with the Austrian "Safe System" approach. The "Safe System" approach is the philosophy of the Austrian Road Safety Strategy 2030. It finds its basis in the Swedish "Vision Zero" (Tingvall & Haworth, 1999) and the Dutch "Sustainable Safety" concept of the 1990s. Table 2 shows the principles of the "Safe Systems" approach, extended by a column on how this thesis contributes to its targets.

Question	*Traditional approach*	*Safety System approach*	*Embedding the Safety System approach into this dissertation*
What is the problem?	Accidents	Fatalities and severe injuries	Investigation of fatal and severe road traffic accidents
What causes the problem?	Misconduct	People make mistakes and are vulnerable	Investigation of single-vehicle accidents with single-occupancy. Thus, new insights on driving behaviour and mistakes are retrieved
Who is responsible?	Traffic participants	System designer and users	The thesis generated evidence-based knowledge on driver-, vehicle-, roadway-, and situation-related variables
Need for safety?	People do not want safety	People want safety	Among fatal and severe accidents, this thesis quantifies the share of involved drivers that did not apply safety belt
What is an appropriate target?	Optimal number of fatalities and severe injuries	Elimination of fatalities and severe injuries	Setting up appropriate intervention measures to reduce the remaining number of fatalities and severe injuries is challenging. Thus, in-depth knowledge about recurring accident-circumstances (i.e., blackpatterns) is required and retrieved through this dissertation

Table 2: Embedding the Safe System approach into this dissertation.
Source: Austrian Road Safety Strategy 2030 and author's amendments.

This thesis investigates recurring accident patterns (blackpatterns) among severe and fatal road traffic accidents (*severe casualties*), which strongly refers to the overall target of eliminating fatal and severe road traffic accidents. On the premise that people make mistakes and are vulnerable this thesis analyses driver-related accident variables and recurring combinations of these variables. For example, if the variables' probationary driving licence', 'impairment by alcohol', and 'speeding' occur combinedly, it counts the frequency of this combination. Furthermore, it analyses whether there exists a significant correlation between the variables' impairment by alcohol' and 'no safety belt applied'. To gain even more detailed information on driving behaviour, this thesis investigates single-vehicle accidents with single-occupancy. This particular sample of accidents helps to retrieve deeper insights into human failure in the context of road traffic accidents. Additionally, this thesis provides an evidence-based and analytical foundation for system designers and users by considering further accident-related variables such as vehicle characteristics, roadway conditions, and weather conditions. Before starting with the pattern recognition approaches, the following chapters present relevant information on Austria's road traffic accident development, an overview of major accident causes, and the definition of accident accumulation points (i.e., blackspots).

1.2 Development of road traffic accidents in Austria

In Austria, fatal road traffic accidents peaked in 1972 with 2.948 fatal road accidents, leading to the realisation of numerous intervention measures (e.g., speed limits or obligatory usage of seatbelts and helmets) to improve traffic safety. The result is a significant reduction in fatal road accidents. Figure 2 blends the number of fatal road traffic accidents with the established intervention measures.

Even if the number of fatal road accidents has decreased over the years, we can now see that it is still a significant challenge for transportation system planning to deal with the remaining number of fatal road accidents.

When comparing fatal accidents per million inhabitants in 2019 (see figure 1), Austria shows a rate of 47 traffic fatalities per 1 million inhabitants and ranks 11th with other EU countries. Sweden, Ireland and Malta had the lowest rate of fatal road traffic accidents in the European Union in 2019.

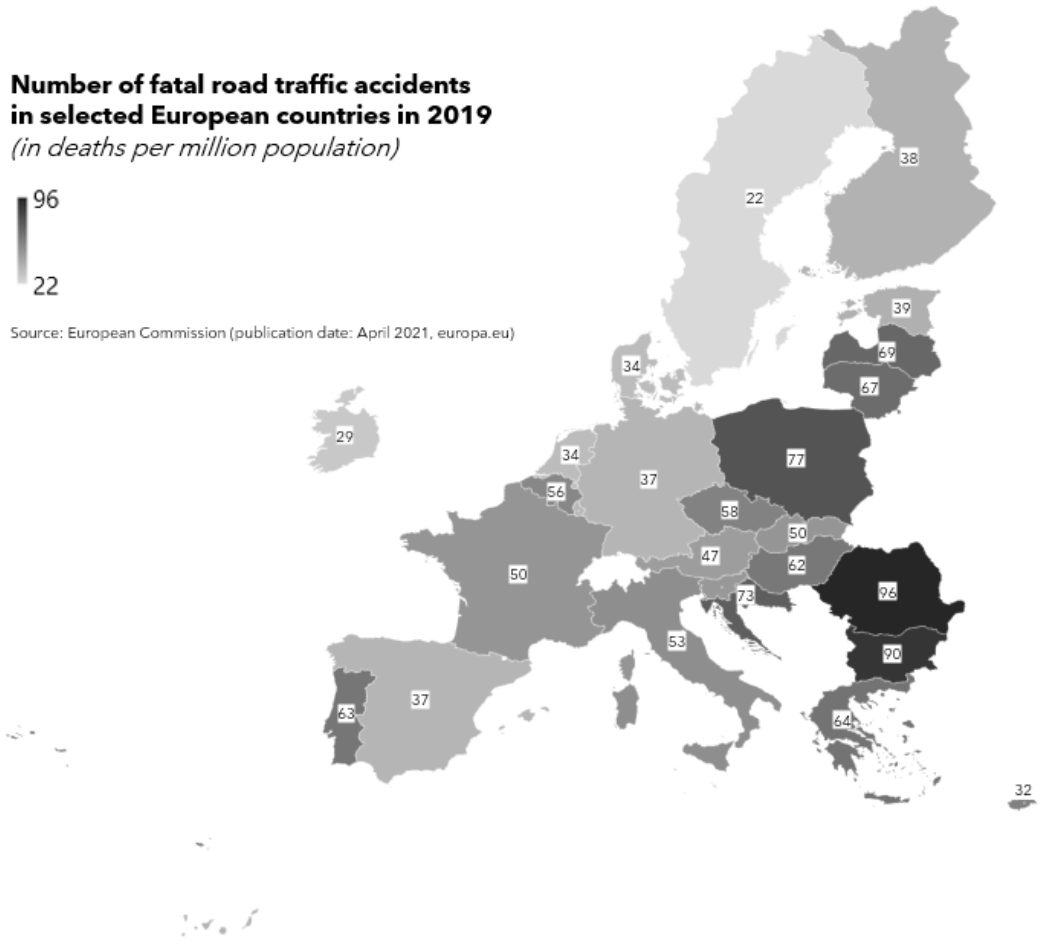

Figure 1: Number of fatal road traffic accidents in select European countries in 2019. Author's compilation. Source: Author's compilation based on European Commission (2019).

Knowledge about accident causes and blackspots provides substantial clues for deriving appropriate intervention measures. On average, 98 accidents can be recorded per day in 2019. A fatal road accident occurs every 21 hours. Therefore, an in-depth analysis of recurring variable combinations seems to be a fundamental approach to generate further understanding of accident circumstances.

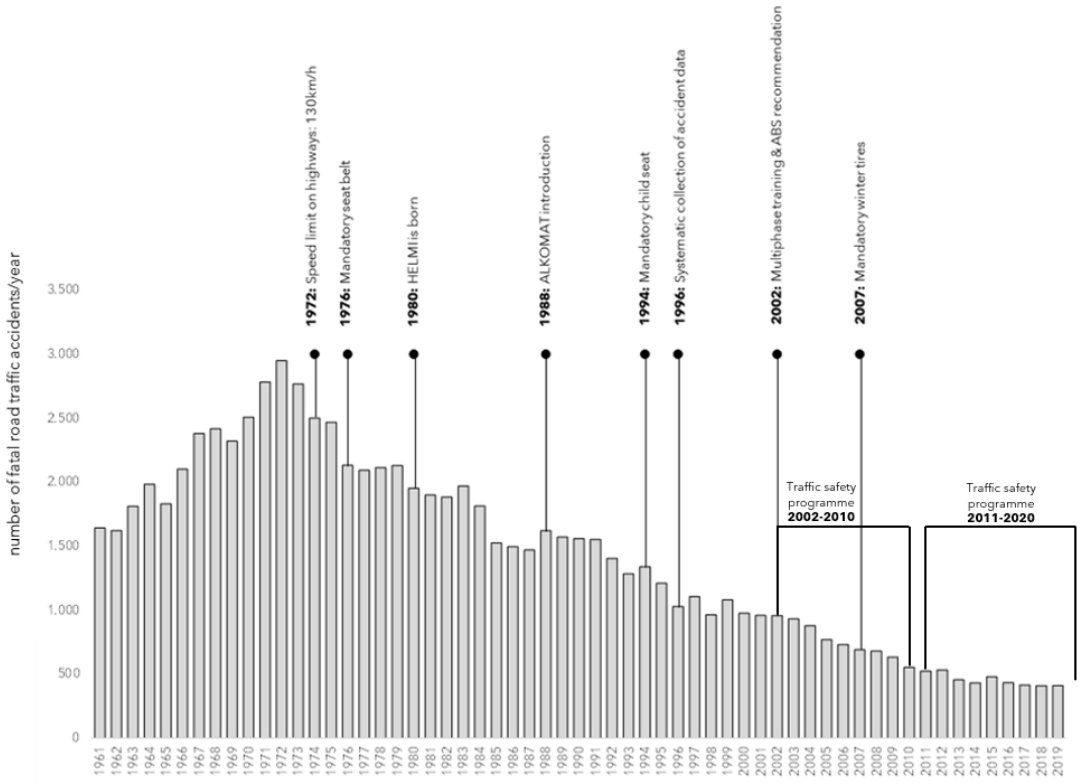

Figure 2: Development of fatal road traffic accidents in Austria with the realisation years of selected prevention measures and an indication of the two traffic safety programmes. Observation period: 1961 to 2019. Sources: Author's compilation based on Statistik Austria (2020) and KFV (Kuratorium für Verkehrssicherheit).

The following heatmap (see figure 3) illustrates the development of fatal road traffic accidents in the Austrian federal states per 100.000 inhabitants.

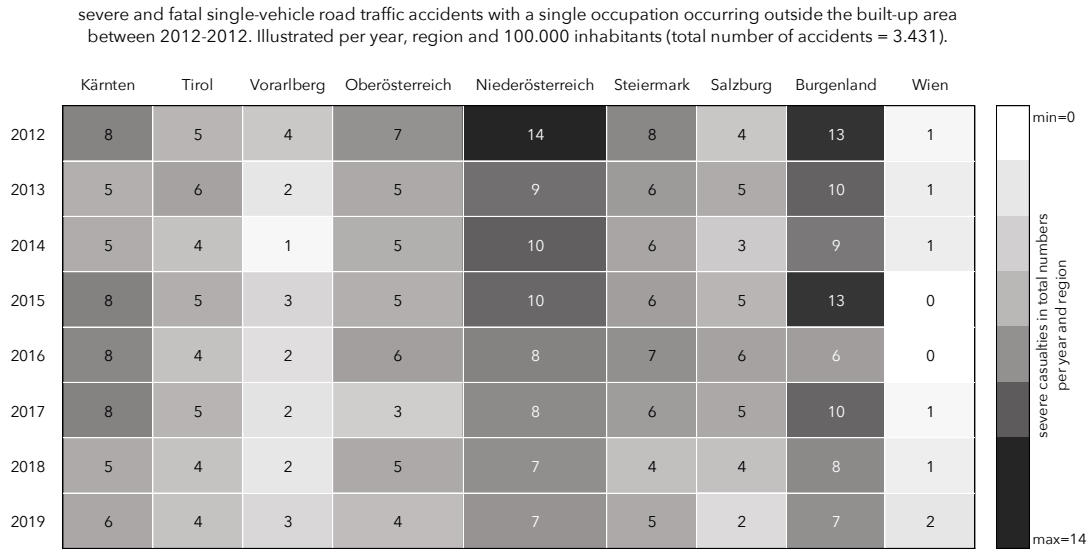

Figure 3: Development (2012-2019) of fatal road traffic accidents with single vehicles and a single occupation and personal injury on the Austrian road network outside the built-up area. Illustrated for the Austrian federal states per 100.000 inhabitants. Source: Author's compilation based on Statistics Austria (2020).

1.3 Major accident causes

The Austrian Ministry of the Interior (Bundesministerium für Inneres (BMI) II, 2020) defines the following alleged accident causes:

- unadjusted speed or speeding,
- distraction,
- priority violations,
- misconduct by pedestrians,
- health (e.g., cardiovascular failure),
- overtaking,
- disregarding bids and bans,
- inadequate safety distance,
- fatigue, and
- technical defects

As figure 4 illustrates, unadjusted speed, distraction and priority violations hold the highest shares among accident causes (above ten per cent). It is important to mention that these accident causes represent the accident causes across all types of road traffic accidents and not explicitly those for single-vehicle accidents with a single occupation. Therefore, other accident circumstances may be identified in the sample of this thesis. Thus, an advantage of this thesis

is to extract detailed knowledge for a specific accident type. If the method proves to be applicable, we will expand it towards further accident types.

One must emphasize that the designation of a major accident cause represents a primarily subjective assessment. For each road traffic accident, the police authorities determine the alleged cause of the accident on site when filling out the accident data sheet. Depending on how differently police officers may be trained on accident surveys, there will always exist a so-called evaluation bias going along with road traffic accident records. This thesis does not examine these superordinate parameters. The focus is placed exclusively on examining the officially available traffic accident data. However, we will point out limitations of road traffic accident data (see chapter 2.1).

The following is a selection of causes of accidents that require further specification.

Speeding: Speeding includes exceeding the maximum speed limit at the accident scene and speed not adapted to visibility, road, and weather conditions.

Distraction: Distraction includes lack of concentration, visual and mental distraction, and all non-driving activities such as eating, drinking, reading, smoking, picking up objects etc.

Impairment due to drugs, medication, overtiredness or health impairment: The determination of impairment may refer to medical assessment, the police officer's assessment, or questioning the driver.

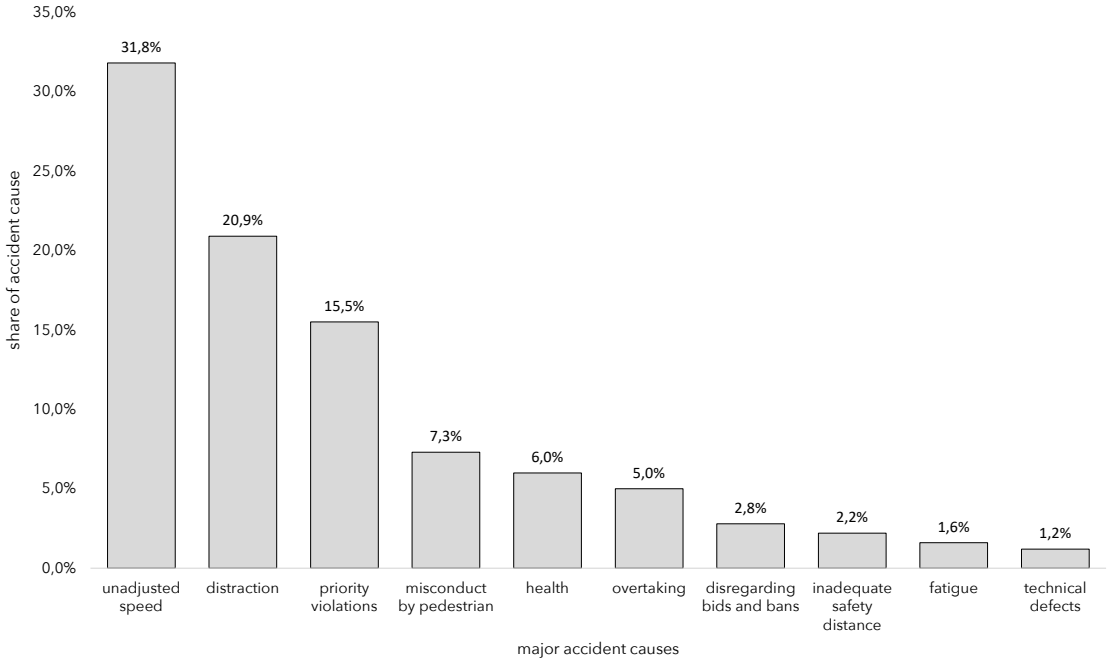

Figure 4: Major accident causes in Austria (2020). Source: BMI, 2020.

1.4 Blackspots

The definition of road accident accumulation points (blackspots) is part of the guideline on traffic safety inspection of the Austrian Research Association for Roads, Railways and Transport. The guideline designates a node or road section up to a length of 250 metres as an accident blackspot if:

- at least three similar accidents with personal injury have occurred in three years, and the relative coefficient[1] reaches or exceeds the value of 0,8 or
- at least five similar accidents (including accidents involving property damage) have occurred in one year.

An accident accumulation point exists if one of the two criteria applies.

In Austria, municipalities and districts calculate and visualize their respective accident accumulation points. The province of Upper Austria makes accident accumulation points accessible via a WebGIS application. Figure 5 illustrates the accident accumulation points in Upper Austria.

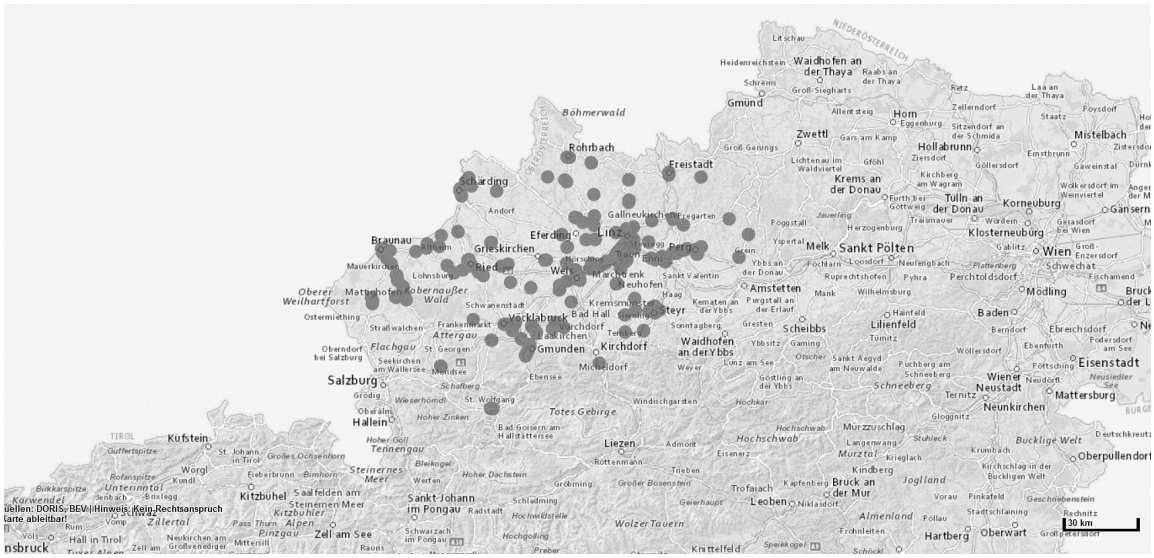

Figure 5: Accident accumulation points in Upper Austria in 2019. Source: doris.at.

[1]The relative coefficient is a value considering the number of accidents in relation to traffic volume.

An online map showing accident blackspots throughout Austria does not exist thus far. Statistics Austria (Statistik Austria, 2020) provides an interactive map showing the number of accidents by district (see figure 6).

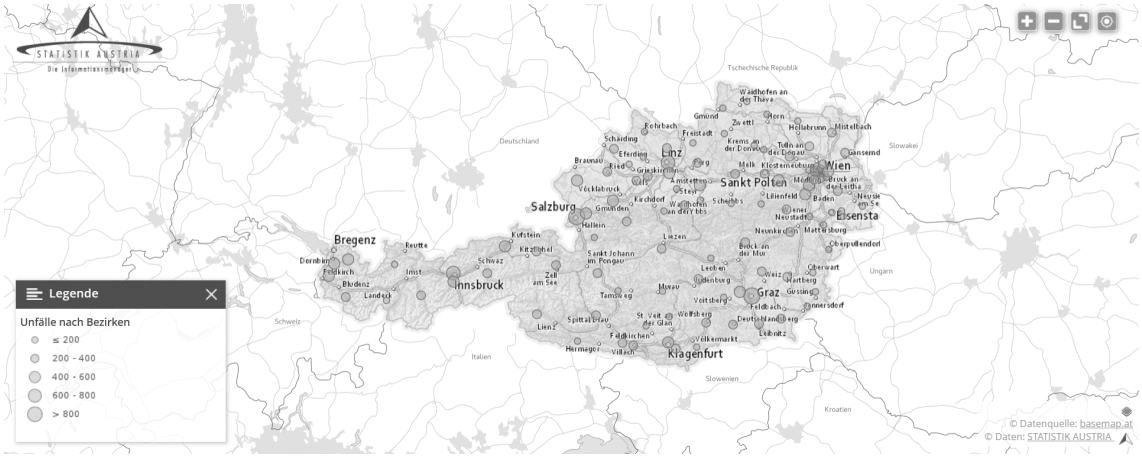

Figure 6: Number of accidents by districts in 2020. Source: statistik.at/atlas/verkehrsunfall.

1.5 Research gap: from blackspots to blackpatterns

Content-related research gap

In addition to accident causes and accident blackspots, this thesis presents an exploratory research approach to identify recurring combinations of accident-related variables: blackpatterns. Over the last few years, many analytical approaches have been developed in road traffic accident research and analysis. These approaches address the following categories:

- drivers (e.g., driving behaviour, impairments)
- vehicle (e.g., vehicle performance, technical defects)
- infrastructure (e.g., road surface condition)
- situation (e.g., light conditions or weather conditions)

This dissertation intersects driver-, vehicle-, roadway-, and situation-related variables and identifies multivariate road traffic accident patterns. A blackpattern is a multivariate and recurring combination of accident-related variables.

Data-related research gap

Due to the complexity of the road traffic accident database (over 180 potential characteristics for each road traffic accident), it is difficult to detect patterns and associations lying behind the data. Additionally, classic statistical approaches reach their limits to identify patterns within large datasets that do not correspond to a normal distribution. Therefore, it is necessary to implement and test new methods such as explorative pattern recognition methods on road traffic accident data. Currently, we experience a proliferation in the use of data science or machine learning techniques to identify patterns in large data sets that would otherwise be difficult to detect.

However, pattern recognition methods would not be directly applicable to the Austrian traffic accident database as it currently exists. As we illustrate in chapter 3.3, the original road traffic accident database makes it difficult to analyse one specific accident-related characteristic exclusively because of multiple characteristics showing up in one single entry or cell. For this reason, this thesis presents an approach to reprocess the original road traffic accident data. Furthermore, the thesis outlines how to apply suitable explorative methods on road traffic accident data on the premise that this dataset is subject to uncertainty and other limitations (see chapter 2.1).

1.6 Research question and scope of the thesis

The current accident-related information, the hypothesis that road traffic accidents cannot be considered monocausal events and the described research gap lead to the formulation of the following research question:

Based on road traffic accidents occurring in Austria between 2012 to 2019, how can multivariate and recurrent accident patterns that consider driver-, vehicle-, situation-, and roadway-related characteristics be identified? Do these patterns show a significant relation with severe and fatal accidents? To what extent do these patterns contribute to a better understanding of accident circumstances? Do these patterns enable a more precise delineation of accident prevention measures?

The research question includes the following hypothesis:

- A road traffic accident does not represent a monocausal event. Therefore, not all accident causes might be known yet. There might exist assumptions about accident causes and circumstances that are not evidence-based yet.

- The official Austrian road traffic accident database (UDM) includes many accident-related characteristics. We need more advanced statistical methods to investigate these characteristics' combined occurrences (blackpatterns), especially heuristic and probabilistic approaches.

The main scopes of the dissertation project are:

- Representation of driver-, vehicle-, roadway-, and situation-related variables and their correlation with accident severity (i.e., degree of injury)
- Identification of recurring variable combinations (blackpatterns) by applying heuristic and explorative methods (e.g., Bayesian Theorem)
- Creation of a consistent and standardized road traffic accident database that contributes to road traffic accident analysis and road traffic accident research

The added value to the field of traffic safety and transportation system planning comprises:

- Identifying recurring accident circumstances (blackpatterns)
- Evaluating the significance of the detected blackpatterns for severe and fatal road traffic accidents
- Creating a consistent and standardized road traffic accident database that may enable the development of traffic prediction models
- Providing evidence-based in depth-knowledge for the derivation of precise measures to improve road safety: not only regarding road conditions but especially regarding driving behaviour

This thesis explores patterns underlying historical road traffic accident records. The thesis does not present an accident prediction model. It does not include data on traffic performance to derive statements on the overall probability of a road traffic accident.

Also, the alleged accident cause represents a subjective assessment by the police officer who fills out the accident data sheet on site. Depending on how differently police officers may be trained on accident surveys, there exists a so-called evaluation bias going along with road traffic accident records. For this reason, we neglect the officially designated accident cause and exclusively focus on recorded accident circumstances regardless of which major accident cause was recorded by the police officer.

1.7 Thesis structure and applied methods

The following table provides an overview of this thesis's content and applied methods.

Chapter	*Content*	*Method*
1	accident costs, development of accidents, alleged accident causes, and accident blackspots	literature review, desk research, data acquisition and visualisation
2	characteristics and limitations of road traffic accident data	description of statistical properties such as data distribution, uncertainty, noise and bias in the context of road traffic accident data
2	pattern recognition methods for road traffic accident data	literature review and desk research
2	definition of an appropriate accident sample	sample selection based on the characteristics of different accident types (see chapter 3.1)
3	creation of a coherent road traffic accident database	data processing and application of the binary coding scheme
3	categorisation of accident-related variables	identification of driver-, vehicle-, roadway- and situation-related variables
3	definition of the target variable *severe casualties* (i.e., fatal accidents or accidents with severe injury)	reclassification of the recorded degree of injury
4	variable frequencies among degrees of injury	creation of contingency tables
4	probability measures	calculation of conditional probability and joint probability
4	correlation between an accident-related variable and *severe casualties*	calculation of Fisher's exact test and Phi coefficient
4	calculation of the highest number of a variable to occur in combination with other accident-related variables	calculation of the combination maximum
4	robust parameter estimation (95% confidence intervals) for each variable	bias-corrected and accelerated bootstrap (BCa) bootstrap resampling

Table 3: Structure and methods of the thesis.

Chapter	*Content*	*Method*
5	estimation of the strength of the relationship between an accident-related variable and the target variable *severe casualties* compared to all observed variables	binomial logistic regression
6	decision tree generation	generation of decision trees with the CHAID-algorithm
7	Bayesian network generation	generation of TAN-structured Bayesian networks
8	blackpattern detection (developed method)	application of the PATTERMAX-method
9	blackpattern evaluation	calculation of Fisher's exact test and Phi coefficient
10	Discussion	review and discussion on the retrieved insights as well as on the applied methods

Continuation of table 3: Structure and methods of the thesis.

Chapter one provides an overview of the thesis context, research gap, research questions and associated targets, and the scientific classification of the thesis.

Chapter two represents a theoretical chapter where we dive into road traffic accident data (i.e., uncertainty, noise and bias, rare events, heterogeneity, and over-dispersion). Also, we discuss pattern recognition methods within this chapter.

Chapter three looks at the existing accident types, and we present the reasons for choosing one specific accident type on which we test and run the pattern recognition approach. Moreover, we discuss the characteristics of the existing road traffic accident database and point out the reasons for the data reprocessing task. This reprocessing task leads to developing a binary database that includes more than 150 accident-related variables. Next, we categorise these accident-related variables into the following scheme: driver-related variables, vehicle-related variables, roadway-related variables, and situation-related variables. The third chapter concludes with the definition of the dependent variable.

After the three introductory chapters, we jump into analysis part I in chapter four. This chapter presents each accident-related characteristic in detail with the help of descriptive statistics. First, we show how often a variable occurs among all accidents (severe and fatal accidents and

accidents with slight injuries). Second, we only show how often a variable occurs among severe and fatal accidents. Based on the contingencies, we calculate the probability for a severe or fatal road accident given the respective accident-related variable.

Additionally, we apply Fisher's exact test to determine a possible relationship between an accident-related variable and the dependent variable (severe and fatal road accidents). Fisher's exact test shows whether there is a significant relationship between the two variables and outputs the Phi coefficient to determine the strength of the relationship. Also, we generate a robust parameter estimation (95% confidence intervals showing the likelihood of a variable and a severe or fatal accident to occur) by applying a bootstrap resampling method on the newly established accident database. Moreover, we calculate a so-called maximum combination value as the first value towards blackpattern detection. This value tells us how often a specific variable co-occurs with (an)other accident-related variable(s).

Chapter five uses binomial logistic regression to estimate each variable's impact on severe road traffic accidents with an odds ratio (i.e., the strength of the relationship between an accident-related variable and the target variable *severe casualties* (i.e., severe or fatal accidents) compared to all observed variables). By knowing which variable appears to increase the risk of a severe road traffic accident, we can assess the overall impact of the detected blackpatterns.

Furthermore, we grow decision trees using the CHAID-algorithm in chapter six. Decision trees generate a generalized tree-like structure of variable combinations that appear to increase the probability of a severe road traffic accident. At this point, binomial logistic regression and decision trees help us identify variables that aggravate an accident outcome and the respective degree of injury. However, since we are interested in gaining in-depth knowledge of recurring variable combinations (blackpatterns), we zoom deeper into the underlying data structures.

Consequently, we apply an explorative Bayesian network paradigm in chapter seven. Also, we apply a developed pattern detection method based on the frequency of variable combinations and joint probabilities (PATTERMAX-method) in chapter eight.

In chapter nine, the pattern recognition process concludes with a statistical evaluation of whether the detected blackpatterns show a significant relationship with the target variable *severe casualties*. Like the beginning, so the end, and we calculate Fisher's exact test and the Phi coefficient.

To conclude, we highlight the most aggravating accident-related variables and blackpatterns in chapter ten. Also, we compare the applied pattern recognition methods. The discussion highlights the advantages and the limitations of the PATTERMAX-method combined with binomial logistic regression to gain in-depth knowledge about accident circumstances. The combined application of both methods enables a precise detection and comparison of blackpatterns. For example, do accident patterns among female drivers differ from accident patterns among male drivers? Do accident patterns on regional roads within an 80 km/h speed

limit differ from those on a 100 km/h speed limit? Additionally, the combined approach enables the assessment of the detected blackpatterns with the help of an odds ratio.

Within the research outlook, we propose to expand the PATTERMAX-approach in combination with binomial logistic regression on other accident types. The newly established accident database might also serve as a reliable source for accident prediction. The estimated 95% confidence intervals may represent input variables for a prediction model.

1.8 Scientific classification of the thesis

This thesis represents a curiosity-driven, heuristic research approach to investigate patterns underlying historical road traffic accidents.

First, contingency tables, conditional and joint probability, Fisher's exact test, and binomial logistic regression determine an accident-related variable's impact on severe road traffic accidents (i.e., accidents with fatal or severe injury). Second, we proceed with an investigation of recurring combinations of accident-related variables (i.e., blackpatterns). We apply a probabilistic Bayesian approach, decision trees, and a developed pattern recognition method (the PATTERMAX-method).

We do not make any assumptions on a variable's impact on severe road traffic accidents in advance. Correlations and variable combinations are the results of objective analysis. Thus, the proposed methods represent a quantitative approach to disentangle variable relationships and to discover the data's inherent patterns.

Because of the assumption underlying this thesis that road traffic accidents do not represent monocausal events, the pattern recognition approach is related to the so-called INUS condition. INUS stands for 'insufficient, but necessary part of an unnecessary but sufficient' condition. The INUS condition explains the concept of cause in more detail by considering potential conditions leading to the impact under investigation. Thus, discarding mono-causality leads to the embedment of this thesis into John Mackie's (Mackie, 1965) construct of multi-causality, which he describes with the INUS condition.

As shown in chapter 1.3, official accident causes include only one clear accident condition (e.g., 'speeding'). Official road traffic accident statistics in Austria do not consider other potential conditions going along with it. It is, of course, impossible to depict all possible accidents conditions but the official road traffic accident database in Austria provides a source to identify co-occurring accident-related variables (blackpatterns). These variables are essential information when trying to understand accident conditions and causes, respectively, in more detail.

With the help of the INUS condition, we try to accommodate the notion of multiple causes and effects of an event. We now illustrate the INUS condition with a specific accident-related example. Let us assume that a severe road traffic accident is caused by 'speeding' but actually caused by 'speeding' and 'wet road surface'. According to the INUS condition, the statement

'Excessive speed was the cause of the accident.'

calls for the following interpretation:

- Excessive speed is not a sufficient part of the condition' excessive speed and wet road surface' because excessive speed only does not necessarily fulfil the condition
- Excessive speed, however, is a necessary part of the condition' excessive speed and wet road surface' because, without it, the condition cannot be fulfilled
- 'Excessive speed and wet road surface' is a non-necessary condition for a severe road traffic accident because other conditions can replace it (e.g., no safety belt applied and impairment by alcohol)
- 'Excessive speed and wet road surface' is a sufficient condition for a severe road traffic accident because it inevitably leads to a fatal or severe accident

It is possible to expand this example with more influencing factors (especially potential factors such as 'distraction', 'airbag not deployed', 'fatigue', 'probationary driving licence' etc.). We can conclude that, according to Mackie, a cause only represents a partial condition for one or more effects to occur.

This thesis, by no means, covers all possible effects. Still, it covers evidence-based and recorded effects from the road traffic accident database and thus significant effects for detecting accident-related patterns.

2. Characteristics of and pattern recognition methods for road traffic accident data

This thesis analyses geocoded road traffic accident records from Statistics Austria. As illustrated in figure 2, the systematic collection of road traffic accident data began in 1996. Since the accident data management (UDM) introduction in 2012, accident records have been available with geographic coordinates. Therefore, the investigation period of this thesis starts with the year 2012 and ends in 2019. The investigation period deliberately excludes the year 2020. The inclusion of 2020 might lead to distortions in the pattern discovery process as traffic performance and accident numbers may deviate from the years before due to the corona crisis.

Between 2012 and 2019, 303.700 (aggregated by the UDM-accident identification field "REFUND") road traffic accidents occurred on the Austrian road network. 110.666 road accidents occurred outside built-up areas, while 193.034 accidents occurred within built-up areas.

The road traffic accident records include variables describing each accident in detail, such as accident type, time and place of the accident, the alleged cause of the accident, ambient factors, driver characteristics, degree of injury, driver impairment, driver behaviour, road characteristics, and environmental conditions.

The major target of this thesis is to reveal recurring patterns underlying recorded road traffic accidents. A blackpattern is a recurring combination of co-occurring variables for a severe or fatal road traffic accident.

2.1 Characteristics of road traffic accident data

Road traffic accident data have specific characteristics, making it challenging to apply classical statistical methods to analyse them appropriately. These specific characteristics are:

- uncertainty,
- noise and bias,
- rare events,
- heterogeneity, and
- over-dispersion

Uncertainty

Road traffic accidents are uncertain and their analysis requires knowledge of the factors that influence them. For example, within the Austrian road traffic accident data, the specific location of an accident is subject to uncertainty. Recording errors can occur in the location data between the specification of the WGS 84 coordinates and the street kilometre data. The reason for the discrepancies arises from survey modalities. The WGS 84 coordinates of the accident sites result from marking the accident site on an electronic map. At the same time, the road mileage data are collected separately based on the physical, local kilometre marker. The more precisely the accident localisation is on the map, the better the match with the road kilometre.

Further uncertainty sources may involve input or reading errors while recording the accident. The variables are primarily categorical in the accident record sheet. Since humans are involved in recording the accidents on site and post-processing the recordings, typos, missing values, and noisy values may arise. Although several algorithms exist to detect and correct noisy data, most of these algorithms deal with continuous numeric data. Detecting and correcting erroneous values in categorical datasets remains a challenging task (Ayman & Ali, 2019, p. 27). Contingency tables are a remedy here, as we will see in chapter 4.

Noise and Bias

The simplest example of a biased sample arises directly from the quality of the accident data itself, namely the problem of underreporting in official accident statistics. That is, accidents with no or less severe injuries do not show up in the accident databases because these accidents are simply not recorded by the police. Most accident analysis models include accidents that result in human suffering. In contrast, accidents that result in property damage are neglected—this problem of underreporting leads to a biased sample of traffic accident records.

In this context, it is essential to address the evaluation bias. For example, the assessment of the alleged accident cause represents a subjective assessment by the police officer who fills out the accident data sheet on site. Depending on how differently police officers may be trained on accident surveys, there will always exist a so-called evaluation bias going along with road traffic accident records. For this reason, we neglect the officially designated accident causes and exclusively focus on all recorded circumstances of the accident (regardless of which major accident cause was recorded by the police officer).

Rare Events

Road traffic accidents records indicate the time and location of the accident. These parameters, for example, enable the calculation of the probability that a fatal road traffic accident will happen on a specific road segment within a specific period (e.g., daily, within seven days etc.). Since fatal road traffic accidents are rare events, we assume that the number of fatal road traffic accidents occurring on a specific road segment follows a Poisson distribution with a specific rate of fatal road traffic accidents per day. This specific rate (expectation value) is a constant within the Poisson distribution. However, we cannot expect this expectation value to be constant. This fact is a problematic restriction going along with Poisson distributed data. Thus, there exist approaches to vary the expectation value, as we will see in the section describing over-dispersion.

Heterogeneity

Another challenge in the analysis of road traffic accidents is the problem of heterogeneity in the data. Road traffic accident data do not represent homogenous data. The accidents vary in different variable combinations, making it difficult to detect patterns without computational support. The general problem of heterogeneity is that relationships among accident-related variables remain hidden (for example, co-occurring variables with a specific accident cause may not be significant in the entire road traffic accident dataset). The impact of accident-related variables depends on the co-occurrence of variables and thus on the accident conditions (for example, accident-related variables may be different for male or female drivers). Several studies apply clustering methods to reduce heterogeneity within road traffic accident data. However, a cluster always represents a generalized group with information loss regarding recorded variable combinations. In contrast to clustering, this thesis detailly detects co-occurring variables (blackpatterns) among road traffic accidents and counts how often these patterns occur in the period under review (2012-2019).

Over-dispersion

In general, we consider road traffic accident data as Poisson distributed. A data set exhibits over-dispersion when the variance is more than the mean. Poisson or binomial logistic regression represent commonly applied methods to quantify the relationship between accident-related variables and severe or fatal accidents. When modelling data with Poisson regression models, over-dispersion will almost always be the case. In Poisson regression models, the dispersion parameter ∅ is a constant. That is why Poisson regression is not the most reliable method for modelling road traffic accidents. As we will see later (chapter 2.2), an algorithm called negative binomial regression varies the dispersion parameter ∅ according to a Gamma distribution. The resulting dispersion parameter ∅ is a random variable, so negative binomial regression includes a dispersion parameter that addresses the unobserved heterogeneity in the accident data (Nwankwo & Godwin, 2015, p. 227).

2.2 Pattern recognition methods for road traffic accident data

Within the ongoing development of machine learning and advanced modelling approaches (algorithms and techniques to analyse, categorize, and predict), the analysis of road traffic accidents becomes an exciting field of research for road traffic safety. The target is to determine accident-related variables that contribute to fatal and severe road traffic accidents and to predict road traffic accidents (Gutierrez-Osorio & Pedraza, 2020). Hence, we experience an increase in methodological complexity in road traffic accident analysis. However, the requirement for science is to keep in mind the general applicability of the developed methods. This chapter presents applicable procedures for analysing road traffic accident data.

On the aspect to

- deeply analyse road traffic accident data,
- to characterize accident-related variables,
- to describe their impact on the degree of injury and
- to discover unrevealed patterns and driving behaviours

the following algorithms and computational modelling approaches evolve in the field of road traffic accident analysis:

- classification,
- regression,
- clustering, and
- association rules.

These approaches apply to the field of machine learning. Machine learning consists of two categories: supervised and unsupervised machine learning. As shown in figure 7, classification and regression belong to supervised machine learning, while clustering and association rules belong to unsupervised machine learning.

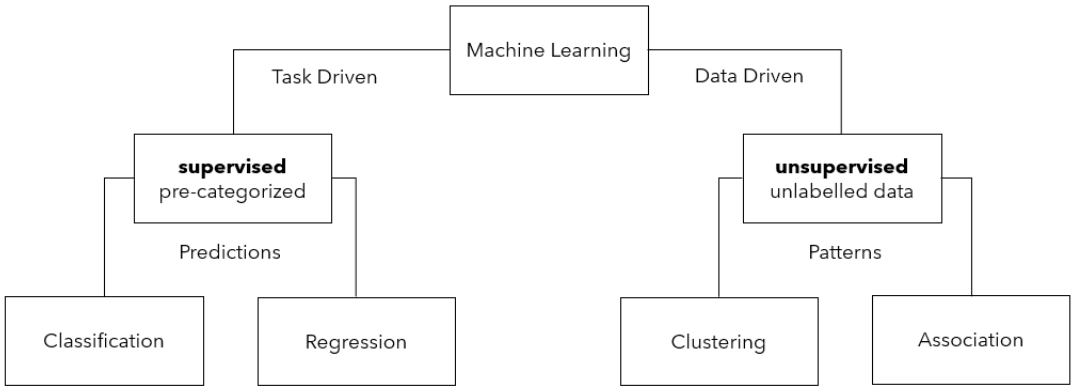

Figure 7: Pattern recognition methods.
Source: Author's compilation based on Cui, Ling, and Zhu (2018).

Table 4 illustrates the differences between supervised and unsupervised machine learning.

supervised machine learning	*unsupervised machine learning*
The algorithm learns from labelled data to predict the outcome from the input data.	It is the technique of using algorithms where there is no outcome variable to predict or classify.
The goal is to identify the relationship between the input and output variables and categorise new, unlabelled data.	Unsupervised data mining aims to find patterns in a dataset based on the relationship between data points themselves.
The methods for supervised machine learning include regression and classification.	The methods for unsupervised machine learning include clustering and association.
Supervised data mining tends to be highly scalable, and it is generally fast.	Unsupervised methods often raise several scalability issues, and they are relatively slow.

Table 4: Supervised vs unsupervised machine learning.
Source: Author's compilation based on Al Musawi (2018).

Various algorithms exist for pattern detection in road traffic accident data among the approaches mentioned above. As a result of the literature survey, figure 8 presents an overview of algorithms appropriate for road traffic accident analysis.

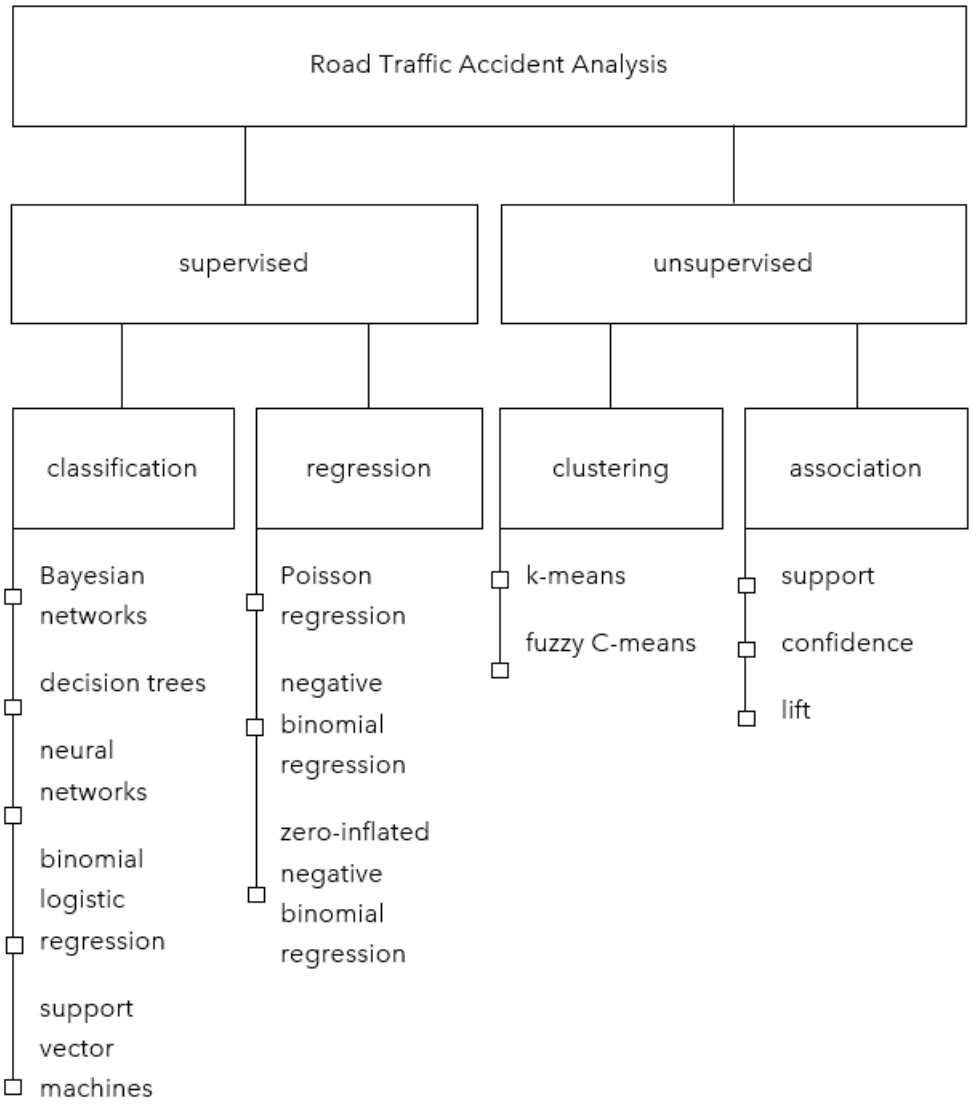

Figure 8: Pattern recognition methods in traffic accident analysis.

We will now have a detailed look at the four pattern recognition approaches for road traffic accidents: clustering, regression, classification, and association rules.

Clustering

Clustering leads to the partitioning of data into clusters. Objects with the most similarities remain in a group and have fewer or no similarities with the objects of another group. Clustering works with unlabelled data and represents an unsupervised data mining method. The following conditions define a cluster:

The m-Clustering of $X = x_1, x_2, \ldots, x_n$ is called the subdivision of X into m Cluster $C_1, \ldots, C_m$, so that

$$C_i \neq \emptyset, i = 1, \ldots, m$$

$$\cup_{i=1}^{m} C_i = X$$

$$C_i \cap C_j = \emptyset, i \neq j, i, j = 1, \ldots, m$$

These criteria mean that all clusters have at least one data point. No cluster is empty. The union of all clusters then corresponds back to the data. The intersection of C_i and C_j is always empty. That means differently formulated: Each data point from X is assigned to exactly one cluster, not two or more clusters (in the case of hard clustering). (Kovera, 2017)

The vectors x_i of cluster C_i are more similar to each other than the vectors in the other clusters. That kind of clustering is called "hard" or "crisp" (opposite = fuzzy or soft). Hard clustering states that every element is assigned to only one cluster (e.g., k-means clustering). In contrast, soft clustering states that one element is assigned to all available clusters with a different membership degree for each cluster (e.g., fuzzy c-means clustering). (Wiharto & Suryani, 2020, p. 43)

Figure 9 shows the principles of hard and soft clustering.

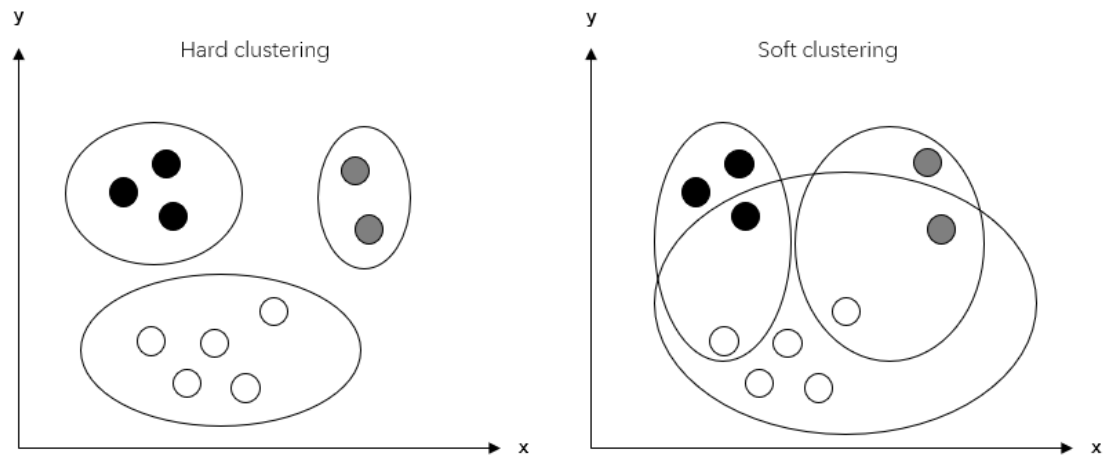

Figure 9: Comparison of hard and soft clustering.

Similarity and dissimilarity measures are core components of clustering. The distance measure tells us how similar or dissimilar two feature vectors or two subsets of X are.

Dissimilarity measure *d*:

A dissimilarity *d* on X is a function d: X × X → $\mathbb{R}$, where ∃ $d_0 \in \mathbb{R}$:

$$-\infty < d_0 \leq d(x, y) < +\infty, \forall x, y \in X$$

d_0 is the smallest dissimilarity measure

The dissimilarity measure *d* is a function representing feature vectors as real numbers. The dissimilarity of x and y (or the dissimilarity of two feature vectors) is bigger than d_0, and d_0 is the smallest dissimilarity measure. It may be that x and y account for precisely this smallest dissimilarity. In any case, the smallest dissimilarity is above minus infinity and below plus infinity. Ciaburro (2017)

Similarity measure *s*:

A similarity measure s on X is a function $s : X \times X \rightarrow \mathbb{R}$, where ∃ $s_0 \in \mathbb{R}$:

$$-\infty < s(\boldsymbol{x}, \boldsymbol{y}) \leq s_0 < +\infty, \forall \boldsymbol{x}, \boldsymbol{y} \in X$$

s_0 is the biggest measure of similarity

The similarity measure *s* is a function representing feature vectors as real numbers. The similarity of x and y is above minus infinity, and it is smaller than s_0, which applies to the biggest measure of similarity. The biggest similarity of x and y is below plus infinity.

Both conditions, the smallest measure of dissimilarity or the biggest measure of similarity, state that two feature vectors are similar. Thus, the similarity measure and dissimilarity measure enable the calculation of the distance between vector x and y. Ciaburro (2017)

Generally, clustering methods comprise hierarchical clustering and non-hierarchical clustering methods: hierarchical clustering and non-hierarchical clustering.

Hierarchical clustering creates homogenous groups of N objects based on distance. A stepwise procedure uses a series of predefined characteristics (i.e., an agglomerative algorithm using successive mergers or a divisive algorithm using successive division). It constructs a hierarchy or treelike structure to depict the cluster information. In contrast to hierarchical clustering, non-hierarchical clustering does not perform a treelike construction process but partitions a set of N objects into K distinct groups based on distance. Once the number of clusters is specified (known a priori or estimated as part of the procedure), non-hierarchical clustering assigns the objects into clusters.

The k-means cluster algorithm is commonly applied in road traffic accident analysis. It is a non-hierarchical clustering method and foresees the following steps:

- partition the N objects into K distinct clusters C_1, …, C_k
- for each i = 1, …, N:
    - assign object x_i to cluster C_k that has the closest centroid (mean)
    - update cluster centroids if x_i is reassigned to the new cluster (Helwig, 2017)

Kumar & Toshniwal (2017) use the k-means cluster algorithm to identify and categorize accident-prone locations by accident frequency (high-frequency, moderate-frequency, and low-frequency). Afterwards, they use association rules (see below) to characterize the identified accident locations. Saharan & Baragona (2017) use the k-means cluster algorithm to identify the factors associated with accidents of different levels of severity in Christchurch, New Zealand. The applied clustering approach provides new insights into the relationship between accident-related variables and accident severity. In their study, 'speed greater than 60 km/h' and 'did not see other people until too late' represent the two main variables contributing to fatal road accidents. Mauro, De Luca, and Dell'Acqua (2013) apply cluster analysis to aggregate accidents based on similarities. For each cluster, they identify one 'cluster representative' accident and a 'hazard index' to describe the danger level of each cluster. The generated clusters provide the basis for developing an accident prediction model. Assi, Rahman, Mansoor, and Ratrout (2020) use fuzzy clustering for road traffic accident analysis. They developed four different machine learning models to predict injury severity with 15 accident-related parameters: neural networks, support vector machines, fuzzy c-means clustering based on neural networks and fuzzy c-means clustering based on support vector machines. They conclude that fuzzy C-means clustering and support vector machines show a higher injury severity prediction accuracy, sensitivity, precision, and harmonic mean.

Regression

Regression models analyse the relationship between the number of (severe and fatal) road traffic accidents and influencing factors (x-variables) such as driver, vehicle-, roadway-, or situation-related attributes. The target variable (y-variable) often refers to fatal or severe road traffic accidents instead of the total number of accidents. Regression models are part of supervised machine learning.

Conventional linear regression models are not suitable for analysing road traffic accidents since they require a continuous and normally distributed target variable with a constant variance. The target variable (e.g., fatal and severe road traffic accidents) usually represents a discrete and Poisson-distributed variable (or count variable) in road traffic accident analysis. The Poisson distribution represents a probability distribution to show how many times an event is likely to occur within a specified period. Thus, the most common regression models in road traffic accident modelling are Poisson regression, negative binomial regression (NBR), and Poisson zero-inflated negative binomial regression (NINB). Poisson regression is a good starting point for analysing count data. It represents a generalized linear model which assumes that the outcome variable is Poisson distributed:

$$P(x) = \frac{\lambda^x \, e^{-\lambda}}{x!}$$

where:

x = 0, 1, 2, 3, …

λ = mean number of occurrences in the interval

e = Euler's constant 2.71828

A Poisson distribution is parameterized by λ, which happens to be the mean and variance simultaneously. Figure 10 illustrates the Poisson distribution for λ=1, λ=5, and λ=9.

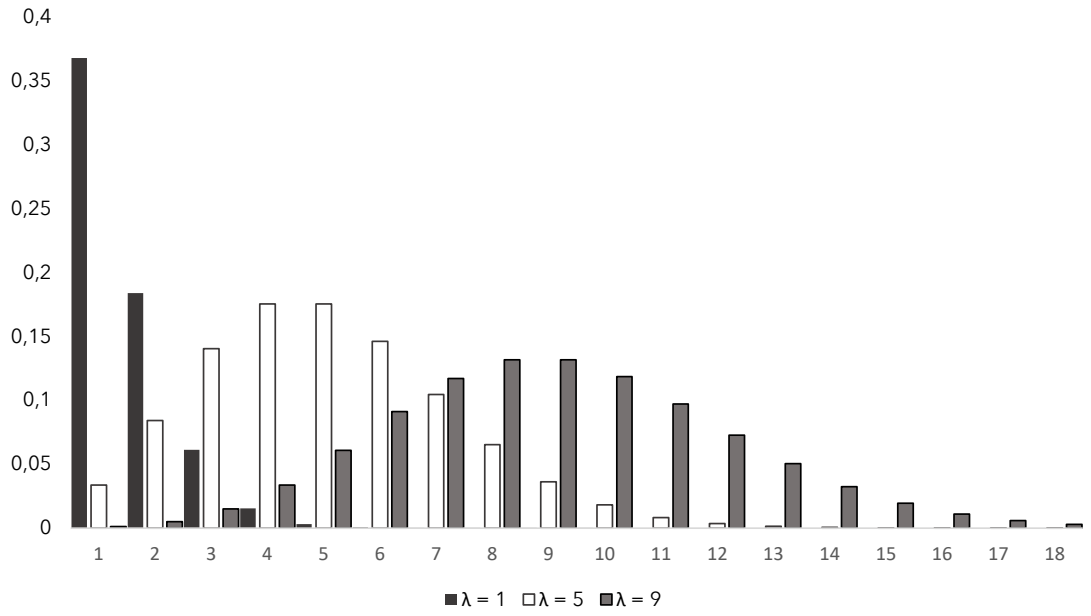

Figure 10: Poisson distribution for λ=1, λ=5, and λ=9.

One restriction with Poisson distribution is that it may not precisely define the variability of counts because of its constant expectation value λ. As discussed in the previous chapter on the characteristics of road traffic accident data (chapter 2.1), over-dispersion may apply when analysing road traffic accidents. Thus, the occurrence of accidents in an interval might vary. In negative binomial regression, mean and variance are not equal. Thus, negative binomial regression is a Poisson distribution with an adapted parameter λ, where λ is not a fixed but a random variable following a Gamma distribution (Yang & Berdine, 2015). Therefore, negative binomial regression appears to be a more reliable approach to analyse road traffic accidents because of its ability to catch over-dispersion rather well. However, negative binomial regression is not suitable for modelling overabundance zeros. Road traffic accident data include the presence of excess zero (as we will see in chapter 4). Suppose the count data exhibit over-dispersion and a substantial number of zeros. In that case, zero-inflated binomial Poisson regression is the model to choose because of its capability to model overabundance zeros.

The research paper of Basu & Saha (2017) reviews regression models for highway accidents. They state that many researchers aim to develop statistical models for accident prediction. Yet, several models do not consider the heterogeneity and potential over-dispersion of road traffic accident data (i.e., more variability (statistical dispersion) in an observed dataset than in a statistical model). They emphasize the necessity to reflect on over- and under-dispersion when applying regression models for road traffic accident data. They substantiate that Poisson regression is inappropriate for predicting road traffic accidents as it cannot handle over-dispersion in accident data. Ma & Yuan (2018) draw the same conclusion regarding the

applicability of Poisson regression on road traffic accident data. They compare Poisson regression, NBR, and NINB to describe the relationship between road traffic accidents and various predictor variables. They conclude that Poisson regression is not ideal as road traffic accident data show excessive dispersion.

Additionally, Prasejito & Musa (2016) illustrate that NINB represents the fittest regression model for analysing road traffic accidents. Also, Getahun & Dejen (2020) show that zero-inflated Poisson is preferable over the Poisson model. Aga, Woldemanuel, and Tadesse (2021) emphasize the necessity of applying negative binomial regression or zero-inflated negative binomial regression in the case of overabundance zeros.

Association

Association rule mining is an unsupervised machine learning method. It detects patterns in massive datasets by identifying co-occurring and correlated variables. Association rules identify if-then associations, with 'if' being an antecedent (i.e., an item within the data) and 'then' being a consequent (i.e., an item found in combination with the antecedent). Three criteria go along with detecting if-then association rules: support, confidence, and lift.

- Support identifies how often an item appears within the dataset
- Confidence identifies the number of items for which the if-then associations apply
- Lift compares the detected confidence with the expected confidence (how many times an if-then statement is expected to be true)

Feng, Zheng, Ren, and Xi (2020), Weng, Zhu, Yan, and Liu (2016), Montella (2011), Das (2014), Gao, Pan, Yu, and Wang (2018), and Priya & Agalya (2018) apply the association rule approach to explore the characteristics and contributory factors for different accident types under different conditions. For example, Das (2014) explores patterns in road traffic accidents that happen under rainy conditions.

Classification

Classification is a supervised data mining method that assigns unlabelled data to target classes or labels in a data collection. Standard classification methods in the context of road traffic accident analysis comprise

- decision trees (DT),
- neural networks (NN),
- logistic regression (LR),
- Bayesian networks (BN), and
- support vector machines (SVM).

Decision trees aim to generate the best possible tree structure from a known data set by creating rules to classify data. A decision tree represents a flowchart consisting of a root node, branches, internal nodes, and leaf nodes (see figure 11).

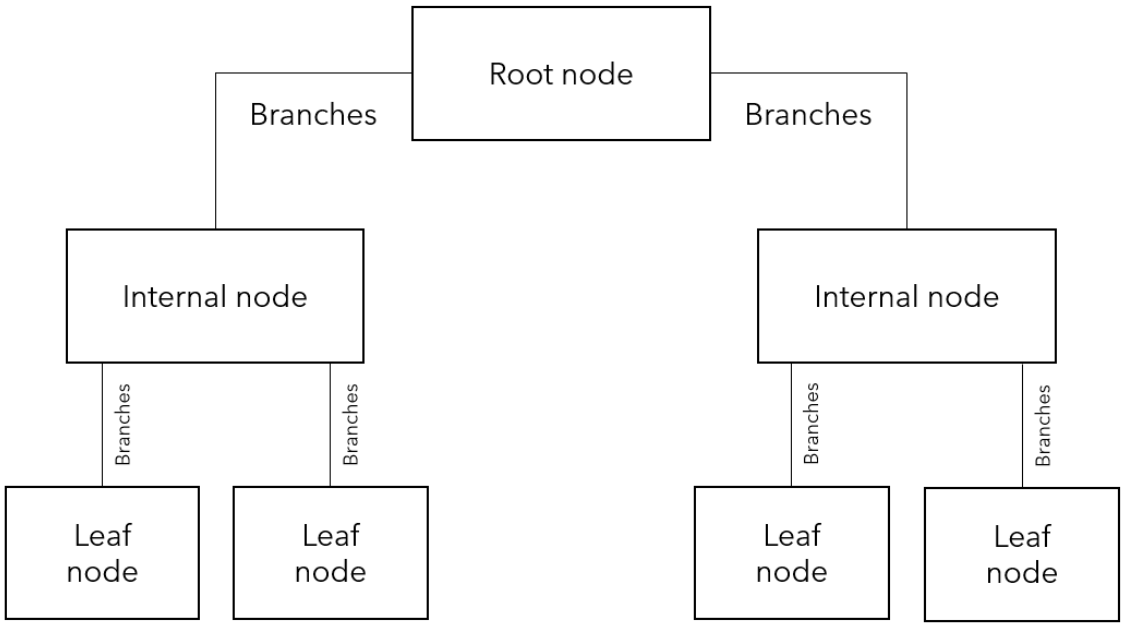

Figure 11: Decision tree structure.

Tree models with a discrete target variable represent classification trees, while decision trees with a continuous value represent regression trees. One well-known decision tree algorithm is CART (Classification and Regression Trees). CART represents a nonparametric model and an iterative method to grow a decision tree.

Starting from the root node (topmost node), each internal node carries out a test on a specific variable, and each branch represents the outcome of the test. In the end, each leaf node holds a class label. Thus, training the decision tree splits the data into two branches. Depending on the data in a node

- the node becomes a leaf, and we classify it, or
- a variable (or feature) divides the data into further branches and nodes.

At each node to be split, the algorithm successively selects the criterion that most appropriately splits the data. Each decision output at a node is therefore called a split (i.e., splitting the training data). The number of branches (edges) starting from a node is often called the branching factor or ratio. It is possible to represent a decision tree as a binary decision tree (branching factor $B = 2$), making it easier to train it. For example, CHAID- or CART-trees represent binary decision trees.

The CHAID-algorithm (Chi-square automatic interaction detection) is the pioneer among decision tree algorithms. Gordon V. Kass defined it in 1980. Based on the CHAID algorithm, further decision algorithms such as CART, ID3, C4.5, or random forest were developed. Figure 12 represents an exemplary CART consisting of this thesis's dataset on road traffic accidents.

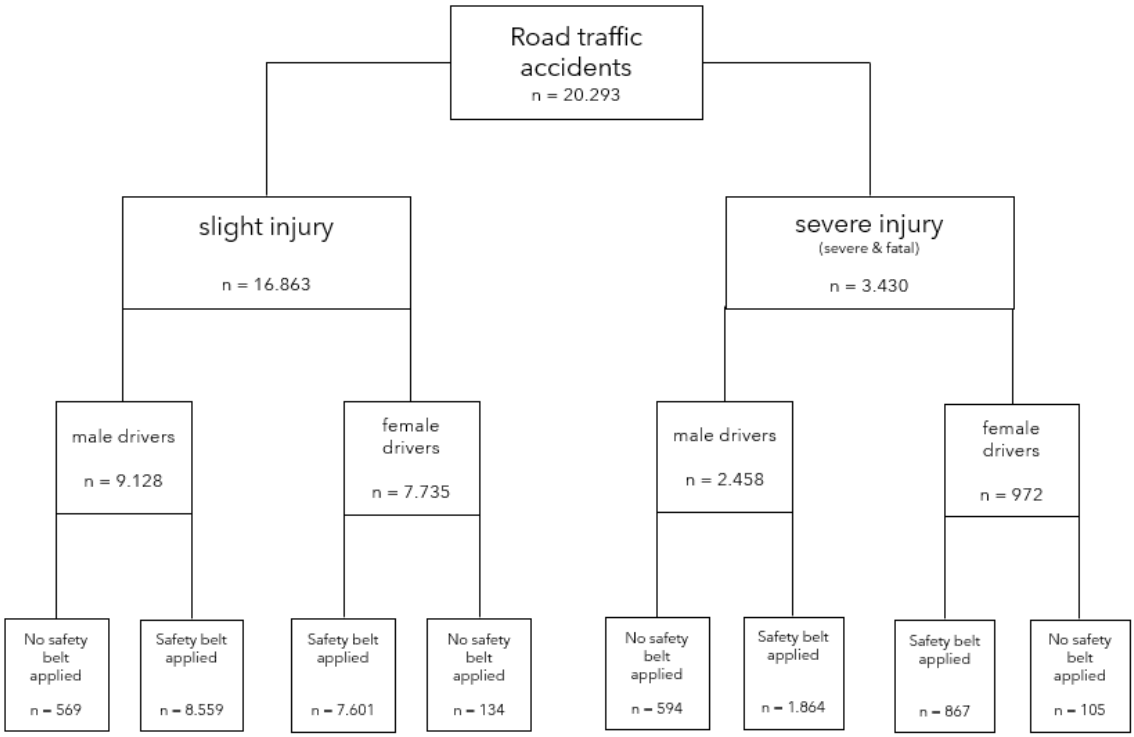

Figure 12: Illustration of a CART.

In this thesis, we illustrate the functionality of the CHAID algorithm and apply it to the road traffic accident dataset in chapter 6.

Several studies use the decision tree algorithm to study the impact of accident-related variables on the degree of injury. (Da Cruz Figueira, Pitombo, Meira, De Oliveira, and Camargo Larocca (2017) applied CART to identify probable accident causes and accident types for severe road traffic accidents on the highway BR-116 in Brazil. On the other hand, Zhou, Lu, Zheng, Tolliver, and Keramati (2020) point out the improved random forest accident forecasting performance because of bootstrap characteristics.

Neural networks are algorithms that replicate the human brain's information processing, information storage, and learning process. A neural network is an abstracted model of connected artificial neurons. It allows complex tasks from statistics, computer science and economics to be solved by computers. Neural networks are a very active field of research and are considered the basis of artificial intelligence and the central element of deep learning. They can interpret various data sources such as images, sounds, texts, tables, or time series by extracting information or patterns out of unknown data. In this way, they generate data-driven predictions. Neural networks vary in complexity but essentially exhibit the structures of directed graphs. We speak of deep learning if a neural network has deep network structures. Also, different types of neural networks exist, such as convolutional neural networks, recurrent neural networks, artificial neural networks, or modular neural networks. (Pan, 2016)

García de Soto, Bumbacher, Deublein, and Adey (2018) created an accident prediction model based on artificial neural networks for Swiss national roads from 2009 and 2021. Generally, they promote the applicability of artificial neural networks for road traffic accident prediction but point out the complexity in training the network because of overabundance zeros. As we know from regression, the issue of predicting the existence of zero is a core challenge in road traffic accident prediction. Pradhan & Sameen (2019, p. 102) also refer to this challenge when working with neural networks in road traffic accident analysis. They also provide a research review on neural networks for traffic accident prediction. They state that neural networks are suitable for handling nonlinear data. Still, they do not conclude with further specifications why the neural network represents a more suitable approach for road traffic accident prediction than other models.

Logistic regression is a form of regression analysis that predicts a nominal-scale categorical criterion. Logistic regression applies whenever the dependent variable has only a few equally significant values. If the dependent variable has only two values in the logistic regression, we apply binary logistic regression (also called binomial logistic regression). We apply multinomial logistic regression if the criterion has more than two categories. Unlike linear regression, logistic regression does not produce specific values for the dependent variable. Instead, it estimates how likely an item falls into one of the dependent variable's categories. Figure 13 illustrates the differences between linear regression and logistic regression.

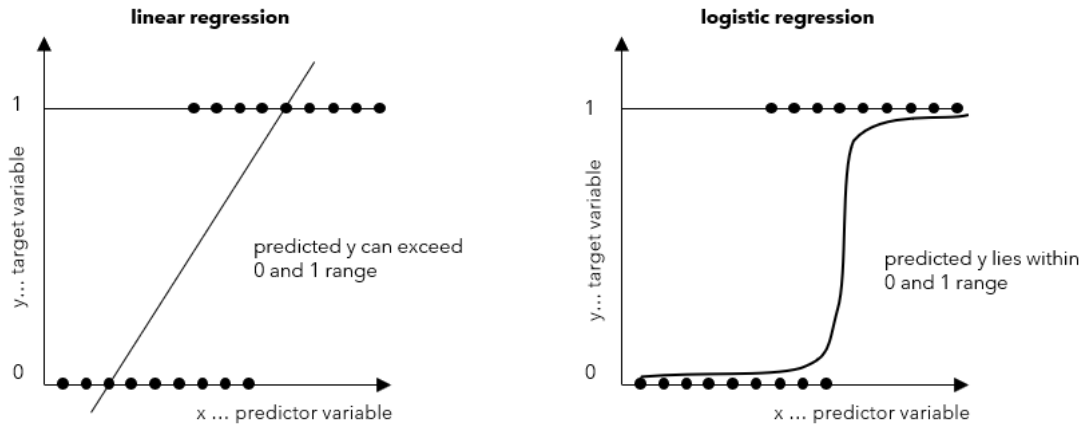

Figure 13: Linear and logistic regression.

The following formula defines the binary logistic regression model:

$$p\,(y\ =\ 1)\ =\ \frac{e^{\beta_0+\,\beta_1\times x_1}}{1+\ e^{\beta_0+\,\beta_1\times x_1}}$$

where:

p (y = 1) … probability for y = 1

e… Euler's number (basis of the natural algorithm)

x… predictor value

β_0 and β_1 … regression coefficients (the maximum likelihood method defines the coefficients)

Logistic regression represents an actively applied method in road traffic accident prediction as it investigates the relationship between accidents and contributing factors. Logistic regression provides insights into the parameter estimates, standard errors, significance, and the overall fitness of the model through an odds ratio (i.e., changes in the ratio of probabilities).

Alavi et al., (2017) apply logistic regression to explore the relationship between personality, driving behaviour, mental disorders, and road traffic accidents. Ahmed (2017) uses logistic regression to find essential variables in road traffic accidents. He detects three variables showing a significant association with fatal road traffic accidents: speed, car type, and location. Zong, Xu, and Zhang (2013) study argue that Bayesian networks (see chapter 7) show better performance in predicting accident severity than regression models.

While logistic regression appears to be a suitable method to estimate the influence of selected variables on severe road traffic accidents, it is not capable of detecting co-occurring variables (i.e., accident blackpatterns) among the data. Consequently, a mixed approach involving a pattern detection method and binary logistic regression might be a practical tool to investigate and evaluate blackpatterns.

Support vector machines represent a supervised learning method for classification and regression tasks. Figure 14 represents the principle of support vector machines. Support vector machines draw a decision boundary (hyperplane) to separate classes of data points. The objective is to define a hyperplane with a maximum margin, which means the maximum distance between data points of the classes. Thus, identifying the optimal position and orientation of the hyperplane is the crucial task of support vector machines. The decision boundary (hyperplane) is defined by the objects closest to it, also called support vectors. Vectors further away from the boundary are not essential for the calculation. Therefore, the algorithm does not load these vectors into the main memory, making support vector machines memory efficient and competitive to neural networks. (Schölkopf & Smola, 2002)

Yu, Wang, Zheng, and Wang (2013) establish a pattern recognition model for urban road traffic conditions. They classify transportation condition patterns in terms of blocking flow, crowded flow, steady flow, and unhindered flow. They conclude that support vector machines using kernel function separate different patterns from traffic flows with high classification accuracy.

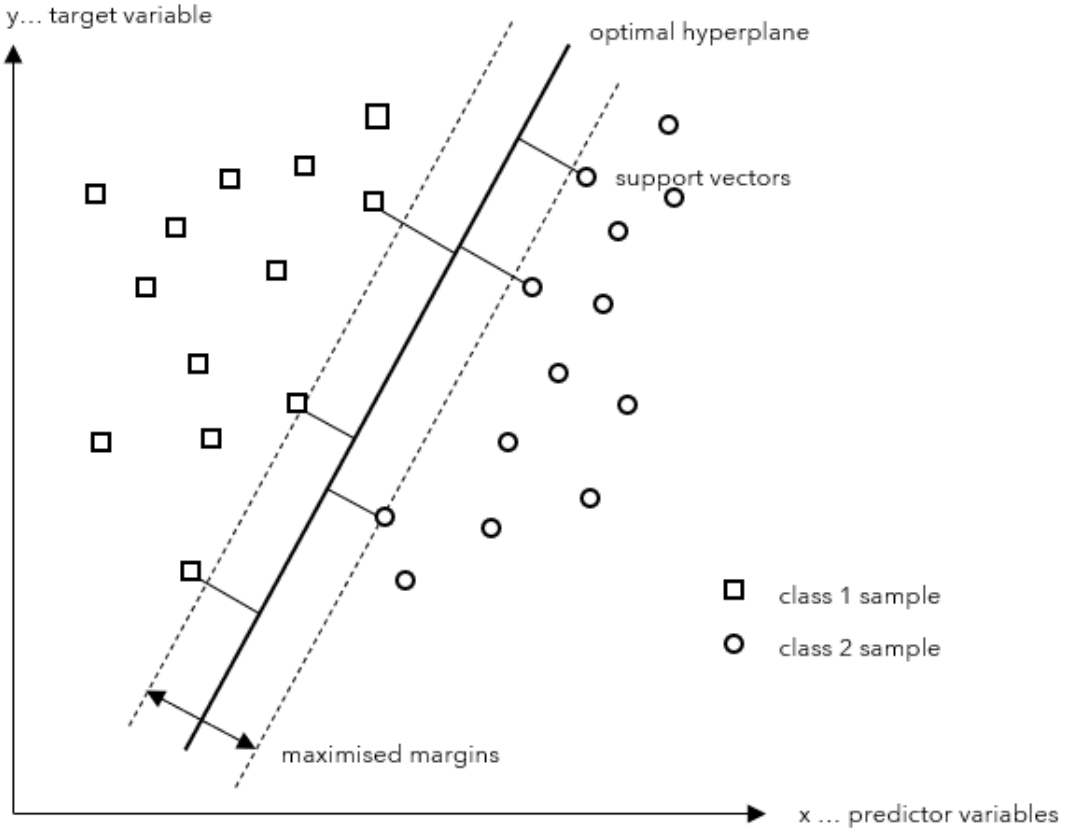

Figure 14: Principle of support vector machines.

The naïve Bayes classifier is a probabilistic classification algorithm, among others, to determine class probabilities based on observations. The classifier builds upon the Bayesian theorem: it calculates the probability of variable A happening, given that variable B has occurred. Therefore, it is a suitable method to calculate the probability of a specific combination of variables to occur (see figure 15). The model assumes that the presence of one variable does not affect another variable. The naïve Bayes classifier is a straightforward but powerful machine learning method. Chapter 7 provides a more detailed introduction to the naïve Bayes classifier.

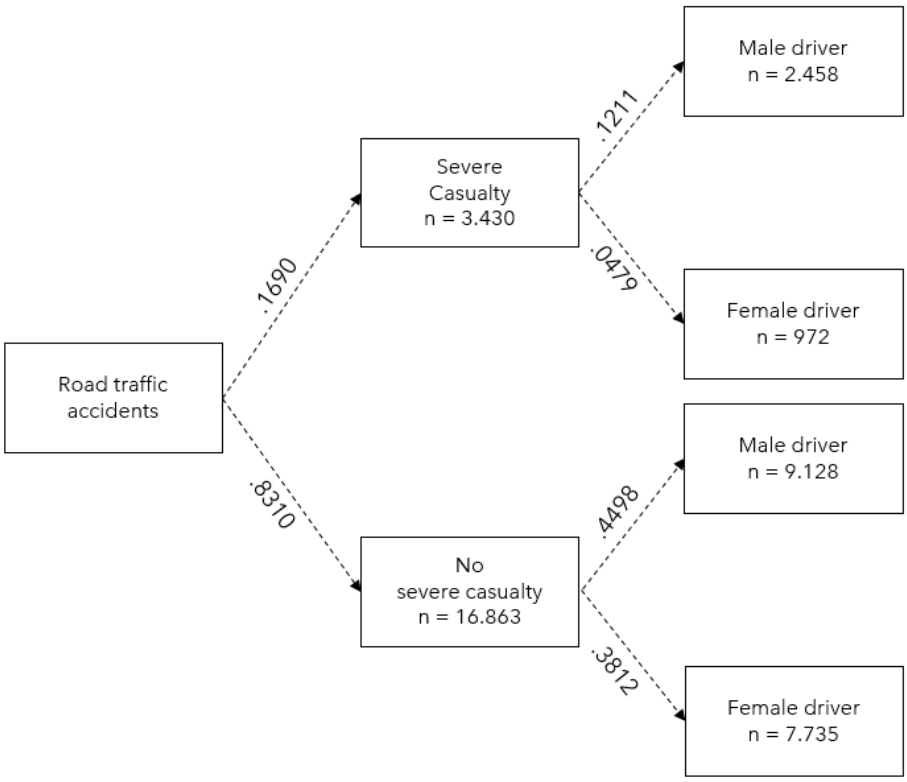

Figure 15: Bayesian network example.

Several studies use a combined methods approach for road traffic accident analysis and machine learning. A comparative analysis of the illustrated machine learning algorithms in traffic accident analysis is conducted by Chong, Abraham, and Paprzycki (2005). They conduct a comparison of four machine learning paradigms for accident pattern detection: neural networks, support vector machines, decision trees and a hybrid model (decision tree & neural network). For *severe casualties*, the hybrid model performed best.

Krishnaveni & Hemalath (2011) apply naïve Bayes classifier, random forest tree classifier, decision tree classifier, and a couple of other algorithms for classifying the level of injury. They conclude that the random forest tree classifier outperforms the other classification algorithms. Almamook, Keneth, Abdulbaset, and Alkasisbeh (2019) apply machine learning algorithms (AdaBoost (a statistical classification algorithm), logistic regression, naïve Bayes, and random forests) for predicting the accident severity (level of injury). The results show that the random forest classifier shows the best performance. Labib, Rifat, Hossain, Das, and Nawrine (2019) analyse road traffic accidents more deeply to determine their intensity by machine learning approaches in Bangladesh. Their analyses make use of the decision tree classifier and k-nearest neighbour. Lee, Yoon, Kwon, and Lee (2019) test the fandom forest classifier, artificial neural networks, and the decision tree classifier for analysing road geometry data, precipitation data, and traffic accident data over nine years. The proposed random forest

model shows the best results. Hayatu, Abdullahi, Ahmad, Ali, and Mohammed (2020) conduct a comparative analysis using four classification algorithms: random forest, decision tree, support vector machine, and k-nearest neighbour with random forest outperforming the other models. Shanthi & Ramani (2012) compare the performance of the classification algorithms naïve Bayes and random tree. The results reveal that the random tree classifier outperformed the individual approaches.

To sum up, the presented research focuses on the implementation and comparative investigation of multiple algorithms such as neural networks, support vector machines, decision trees, random forests, and hybrid models (e.g., decision tree & neural network). The targets are to describe the impact of accident-related variables on injury severity and create accident prediction models. Overall, the results show that decision trees such as random forests generate reliable results to predict accident severity. At this point, it is important to emphasize once more that this thesis does not address accident prediction. It intends to disentangle variable relationships and discover the accident data's inherent patterns and their probabilities and significance among severe and fatal road traffic accidents.

3. Data preparation for pattern recognition

Since road traffic accidents vary in type (e.g., single-vehicle accidents versus multiple-vehicle accidents) and consequently in structure, establishing a pattern recognition method for road accident records requires a suitable sample of accidents records. Also, the original data structure within the Austrian road traffic accident database (UDM) partly prevents an individual analysis of each accident-related variable. Therefore, processing the data is necessary to extract each variable into one column and prepare the pattern recognition process. The following subchapters explain both steps, the sample extraction and the required data processing procedure.

3.1 Accident types

In Austria, the typification of road traffic accidents comprises ten accident types, as shown in table 5. This thesis exclusively uses type 0 'Unfälle mit nur einem Beteiligten', which refers to single-vehicle accidents with a single occupation.

Main accident type	*Description (German)*
0	Unfälle mit nur einem Beteiligten (Abkommen, Auffahren auf Hindernisse u.a.)
1	Unfälle im Richtungsverkehr (Streifen, Auffahren u.a.)
2	Unfälle im Begegnungsverkehr (Frontalkollisionen)
3	Unfälle beim Abbiegen und Umkehren – richtungsgleich (Rechtsabbieger, Linksabbieger)
4	Unfälle beim Abbiegen und Umkehren – entgegengesetzte Richtung
5	Rechtwinkelige Kollisionen auf Kreuzungen beim Queren (geradeausfahrende Fahrzeuge)
6	Rechtwinkelige Kollisionen auf Kreuzungen beim Einbiegen
7	Unfälle mit haltenden und parkenden Fahrzeugen
8	Fußgängerunfälle (von rechts und links, auf Kreuzungen und in Straßenzügen)
9	Tierunfälle, Eisenbahnunfälle, Unfälle auf Parkplatz-, Tankstellen-Haus- oder Grundstücks-Ein oder Aus-fahrten, Kollision mit querenden Radfahrern

Table 5: Typification of road traffic accidents in Austria. Source: RVS 02.02.21.

3.2 Extraction of an appropriate road traffic accident sample

The selected accident type is of particular interest because it corresponds to one line within the road traffic accident database. The other accident types include multiple vehicles, passengers, and other road users (such as pedestrians or cyclists). Therefore, they span several lines within the road traffic accident database. In this case, extracting the drivers and involved vehicles and people becomes necessary, making the pattern recognition process more complex. Thus, we start with the most appropriate sample to test and evaluate the proposed pattern recognition methods, accident type 0. If the developed method proves to be suitable for pattern recognition with road traffic accident data, we will subsequently expand it towards other accident types.

The road traffic accident records of Statistics Austria consist of road traffic accidents with personal injury. Thus, this thesis investigates road traffic accidents resulting in personal injury. The analysis does not include accidents resulting in property damage. When analysing road traffic accident data, it is essential to consider one additional component: the spatial categorisation of where the accident occurs. Within the road traffic accident database, the spatial categorisation foresees the division of accidents into accidents occurring within the built-up area and accidents occurring outside the built-up area. The accident sample in this thesis will focus on single-vehicle accidents with a single occupation that occurred outside the built-up area between 2012-2019.

The extracted sample is defined as follows:

- Based on the road traffic accident records of Statistics Austria, this thesis investigates single-vehicle accidents with single occupation and personal injury that occurred outside the built-up area between 2012-2019 (n=20.293).

- Between 2012 and 2019, 303.700 (based on the field 'REFOUID' within the UDM dataset) road traffic accidents occurred on the Austrian road network. 110.666 road accidents occurred outside built-up areas, while 193.034 accidents occurred within built-up areas.

- The chosen sample amounts to 7 % of all road traffic accidents with a personal injury in Austria between 2012-2019 (n=303.700). Within the period under review, 110.666 accidents with personal injury occurred outside the built-up area, of which the extracted sample comprises 18 %.

Figure 16 shows the development of road traffic accidents in Austria from 2012 to 2019 and the number of accidents corresponding to the chosen accident sample.

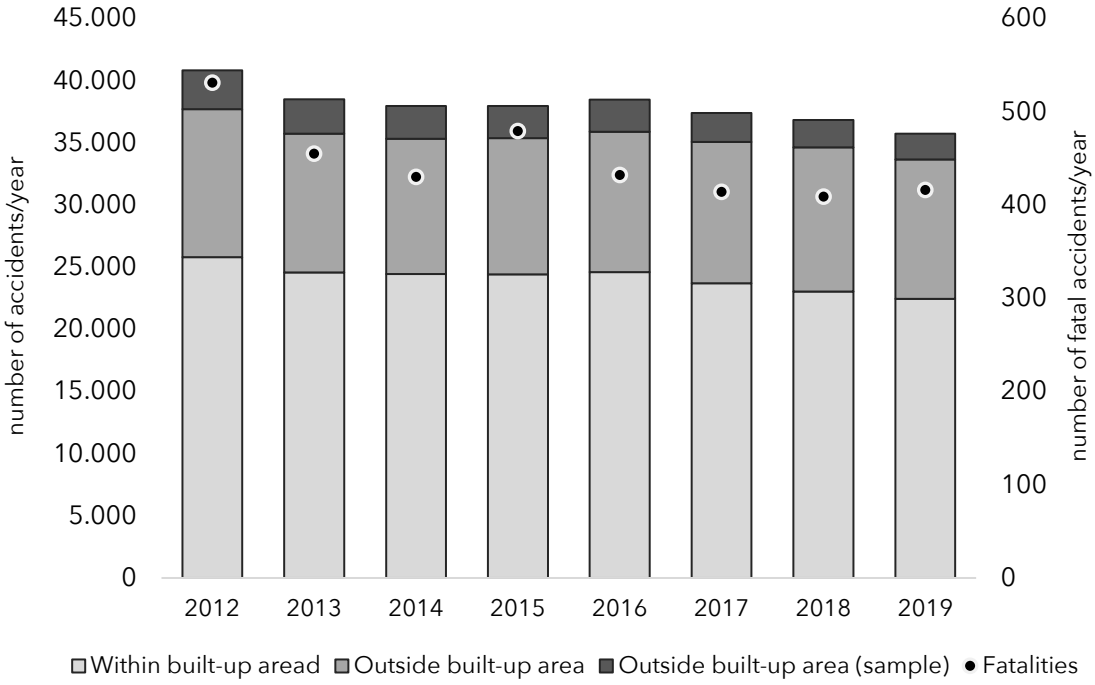

Figure 16: Development of road traffic accidents in Austria from 2012-2019.
Source: Author's compilation based on Statistics Austria, UDM.

Table 6 provides a more detailed picture of the development of fatal road traffic accidents on an annual basis. From 2012 to 2019, the investigated sample comprised 13 to 18 % of annual fatal road traffic accidents occurring outside the built-up area. As we will see later, this thesis focuses on investigating severe road traffic accidents (i.e., accidents resulting in severe injuries and fatal accidents). The target is to identify and quantify the impact of accident-related variables on severe and fatal road traffic accidents.

Year	Fatalities within built-up area n=940	Fatalities outside built-up area n=2.626	Fatalities outside built-up area (sample) n=404	Fatalities total n=3.566
2012	151	380	67 (18%)	531
2013	115	340	53 (16%)	455
2014	123	307	45 (15%)	430
2015	128	351	61 (17%)	479
2016	110	322	50 (16%)	432
2017	107	307	48 (16%)	414
2018	102	307	38 (12%)	409
2019	104	312	42 (13%)	416

Table 6: Development of fatal road traffic accidents in Austria.
Source: Author's compilation based on Statistics Austria, UDM.

The chosen sample is suitable for developing a pattern recognition method of road traffic accidents for multiple reasons. First and foremost, each accident record corresponds to one single line in the accident database. After testing and evaluating the pattern recognition method on this sample, the next step foresees the expansion of the proposed methods on additional accident types.

3.3 Creation of a binary road traffic accident database

One cell often contains multiple variable characteristics in the original road traffic accident records (UDM), making it difficult to analyse one single variable independently. Thus, recoding the accident records into a binary scheme is necessary for this thesis. Subsequently, each accident-related characteristic (more than 150) becomes an individual column.

Table 7 represents an excerpt of the original data structure for one variable and its characteristics (left side of the table). The table also shows the resulting binary data structure after processing the data (right side of the table). In the example illustrated, the overall variable "B_MANOEV" (driving manoeuvres before the accident) shows multiple characteristics embedded in one cell, representing the original structure of the traffic accident database from Statistics Austria. The data processing strategy transfers each variable characteristic (i.e., 6, 21,

22,30, 33, 34, 146 of the attribute B_MANOEV) in an individual column. Thus, each characteristic is an individual column in the resulting accident database consisting of binary digits (0 = false; 1 = true) in each cell. The headlines of the columns indicate the meaning of the variable. In the original database, it was necessary to consult an additional look-up table to identify the meaning of each accident-related characteristic. Consulting the look-up table could result in a very time-consuming and error-prone procedure (as you might associate the code with the wrong meaning since there exist more than 100 characteristics).

The conversion of the original dataset into a binary format was conducted twice to ensure the correctness of the processed database. In addition, randomly selected characteristics from the original database were compared with the corresponding frequencies within the newly established road traffic accident database. The resulting processed database enables an individual and more precise evaluation of each accident-related characteristic.

	6	21	22	30	33	34	146
B_MANOEV	Changing lanes	Sudden breaking	Skidding/ Drifting	Overtaking	Drifting left	Drifting right	Evasive manoeuvres
21;22;34	0	1	1	0	0	1	0
21;22;34	0	1	1	0	0	1	0
22;30;33	0	0	1	1	1	0	0
22;30;33	0	0	1	1	1	0	0
22;30;34	0	0	1	1	0	1	0
21;22;33	0	1	1	0	1	0	0
21;22;33	0	1	1	0	1	0	0
22;30;33	0	0	1	1	1	0	0
21;22;33	0	1	1	0	1	0	0
22;146	0	0	1	0	0	0	1
22;33;146	0	0	1	0	1	1	0
6;22;30;34	1	0	1	1	0	1	0
21;22;34	0	1	1	0	0	1	0
21;22;34	0	1	1	0	0	1	0
ORIGINAL				**PROCESSED**			

Table 7: Data processing scheme.

After applying the binary coding procedure to the entire road traffic accident records, the resulting dataset consists of 158 variables (i.e., 158 columns). With the help of this procedure, each road traffic accident receives a sequence of zeros and ones. Consequently, identifying identical variable combinations among the entire accident datasets becomes relatively easy. The achieved data structure represents an important step towards successful pattern recognition among the historical road traffic accidents.

Figure 17 shows the distribution of recorded characteristics per accident among the road traffic accident sample. The diagrams illustrate the overall distribution in Austria and the distribution within the Austrian federal states. The minimum number of recorded characteristics per accident is 14, and the maximum is 28. The median of recorded characteristics per accident in Austria is 21. The more characteristics the accidents include, the more complex the patterns recognition process becomes. The recorded characteristics within the binary traffic accident database represent sequences of zeros and ones for each accident. Thus, 0-1-sequences enable the identification of recurring patterns.

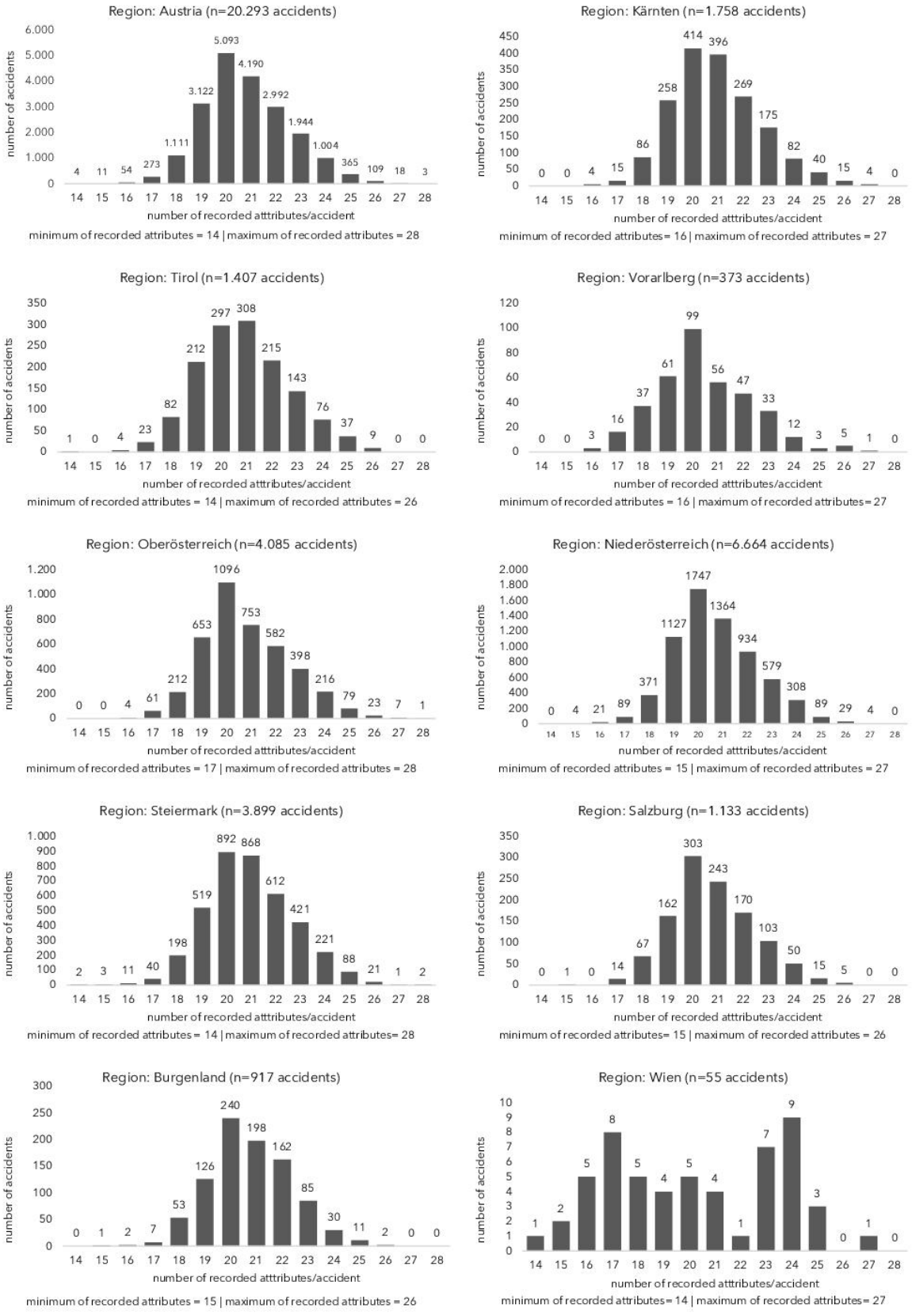

Figure 17: The number of recorded attributes per accident among the road traffic accident sample (n=20.293). The distribution is displayed for Austria and the Austrian federal states.

3.4 Creation of a categorisation scheme for accident-related variables

After recoding all accident-related characteristics and setting up a binary accident database, the next step in data preparation foresees the assignment of each variable to one of the following categories:

- driver-related variables (54 variables)
- vehicle-related variables (32 variables)
- roadway-related variables (50 variables)
- situation-related variables (22 variables)

Figure 18 illustrates the categorisation scheme for accident-related characteristics.

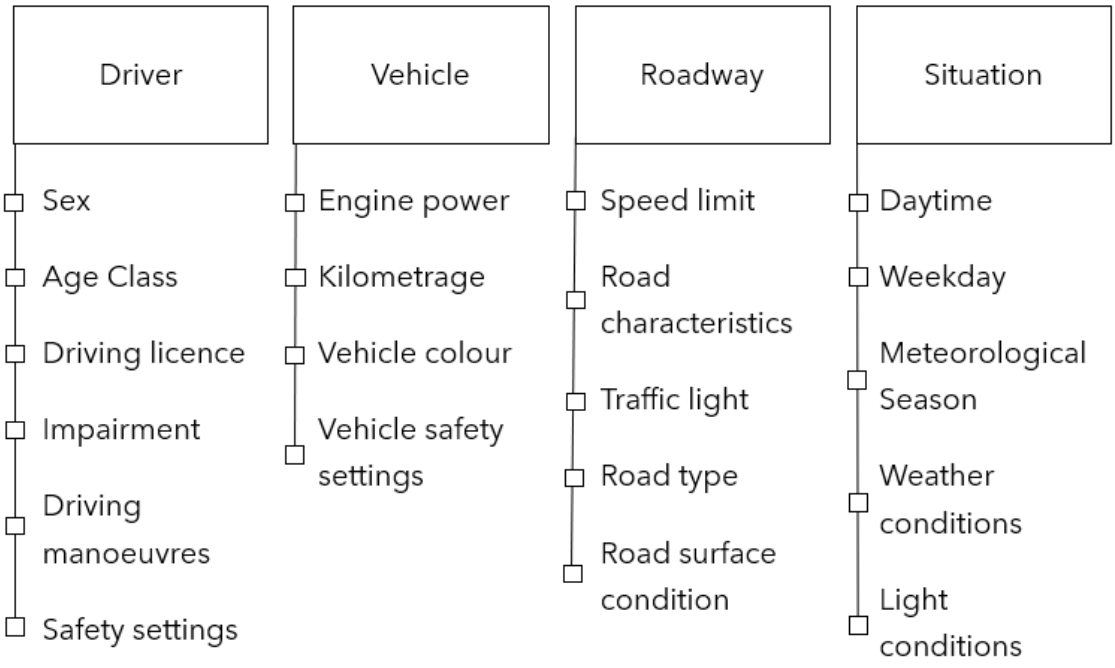

Figure 18: Categorisation scheme for accident-related variables.

Driver-related variables comprise personal data (i.e., sex, age class, nationality), driving experience (i.e., driving licence type and years of driving licence), impairments (i.e., alcohol, drugs, medicines, fatigue, health, excitement), driving manoeuvres before the accident, and safety settings (regarding seat belt).

Vehicle-related variables include engine power, kilometrage, vehicle colour and vehicle safety settings. Roadway-related variables include speed limit, road characteristics, traffic lights, road type, and road surface condition. Situation-related variables involve daytime, weekday, meteorological season, weather conditions, and light conditions.

3.5 Definition of the dependent variable

This thesis aims to quantify each accident-related variable's impact on the degree of injury. Within the original road traffic accident records of Statistics Austria, the degree of injury comprises the following categories:

- minor injury
- severe injury
- death at the accident site
- death within thirty days
- death after more than thirty days

In this thesis, the dependent variable shall combine severe injury and fatalities within the category *severe casualties*. Regarding the Austrian Road Safety Strategy 2021-2030 (KFV & FGM, 2021), it is equally important to reduce fatalities and the number of severe injuries. Also, both categories (severe and fatal accidents) entail high economic costs (as shown in table 1) and human suffering. These premises lead to the following reclassification of the degree of injury:

- casualties: minor injury, severe injury, death at the accident site, death within thirty days, death after more than thirty days
- *severe casualties*: severe injury, death at the accident site, death after thirty days, death after more than thirty years

Thus, the degree of injury comprises only two categories within this thesis. The resulting dependent variable is *severe casualties*. Figure 19 illustrates the reclassification scheme for the degree of injury. This classification corresponds to the definition of 'Schwerverunglückte' and 'Verunglückte' within the handbook of transportation system planning (Cerwenka, Hauger, Hörl, and Klamer, 2007, p.73).

Figure 19: Reclassification scheme for the degree of injury.

4. Road traffic accident data analysis I: Frequencies, Relationships, Probabilities, and Maximum Combinations Values

In preparation for our pattern recognition process, we apply descriptive statistics to analyse the accident-related characteristics and their relationship with the target variable *severe casualties*.

Based on the newly established binary road traffic accident database, it is possible to create contingency tables and to show how often an accident-related characteristic occurred in the observed period (2012-2019). Furthermore, we can illustrate the frequencies among the degree of injury (i.e., casualties and *severe casualties*).

Contingency tables also allow the determination of relative frequencies, conditional probabilities and joint probabilities. Conditional and joint probabilities are essential parameters for the subsequent creation of Bayesian networks (see chapter 7). In the contingency tables, joint probability refers to the probability of an accident-related characteristic to occur, given a severe or fatal road traffic accident.

Additionally, contingency tables help investigate whether an accident-related characteristic impacts the target variable *severe casualties*. Fisher's exact test and the Phi coefficient examine the relationship between an accident-related characteristic and *severe casualties*.

It is interesting how often an accident-related characteristic occurs in combination with another or multiple characteristics in pattern recognition. The maximum combination value determines how often one characteristic appears in the same combination with other characteristics. It provides the basis for the PATTERMAX-method.

The following sections provide an exemplary illustration of the calculations mentioned above:

- frequencies of accident-related characteristics: contingency tables
- conditional and joint probability: Laplace's equation
- relationship between an accident-related characteristic and *severe casualties*: Fisher's exact test and Phi coefficient
- co-occurring characteristics: maximum combination value

4.1 Contingency tables

Contingency tables (crosstabs) show each characteristic's frequency in the observation period (2012-2019) and its distribution among casualties and *severe casualties* (i.e., fatal or severe road traffic accidents). They provide the basis for

- calculating conditional and joint probability for accident-related characteristics and
- investigating the statistical relationship between accident-related characteristics and the target variable *severe casualties*.

Table 8 illustrates the contingency table for sex and degree of injury (all casualties and *severe casualties*). As we can see, our sample consists of 20.293 casualties which comprise 3.430 *severe casualties*.

Sex	C: Casualties	SC: Severe Casualties
M: Male	11.576	2.458
F: Female	8.706	972
U: Unknown	11	-
Total	*20.293*	*3.430*

Table 8: Contingency table of the road traffic accident dataset. n=20.293 (single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network between 2012-2019).

When interpreting table 8, there are a few aspects to consider. This thesis does not include information on traffic performance. It analyses a historical sample of road traffic accident data. Thus, we cannot conclude from table 8 that male drivers are more likely to have an accident because we do not know the total number of male drivers driving a vehicle outside the built-up area in Austria from 2012-2019. We can, however, conclude that single-vehicle accidents with personal injury and single occupation occurring outside the built-up area in Austria between 2012-2019 involve more male than female drivers. Among *severe casualties*, the share of male drivers is 21 % (2.458/11.576), and the share of female drivers is 11 %. However, when calculating the probability of a *severe casualty* to occur given male and female drivers, these percentages continue to fall apart. The following chapter will illustrate the concept of conditional and joint probability.

4.2 Conditional and joint probabilities

Table 9 and table 10 illustrate the conditional and joint probability calculation scheme. First, we transfer the data from the contingency table (see table 8) in probability expressions (see table 9).

Sex	C: Casualties	SC: Severe Casualties	Sex ∩ SC [sex and severe casualties]
M: Male	P (M)	P_M (SC)	P (M) x P_M (SC)
F: Female	P (W)	P_W (SC)	P (W) x P_W (SC)
U: Unknown	P (U)	P_U (SC)	P (U) x P_U (SC)
Total	*1*	*P (SC)*	

Table 9: Conditional (P_M (SC)) and joint probability (P (M) x P_M(SC)) – calculation scheme I.

In the next step (see table 10), we calculate the conditional and joint probability for a severe casualty given an accident-related characteristic (in this case, sex: male, female, and unknown sex).

Sex	C: Casualties	SC: Severe Casualties	Sex ∩ SC [sex and severe casualties]
M: Male	$P(M) = \frac{11.576}{20.293} = 0{,}570$	$P_M(SC) = \frac{2.458}{11.576} = 0{,}212$	P (M) x P_M(SC) = 0,570 x 0,212 = 0,121
F: Female	$P(W) = \frac{8.706}{20.293} = 0{,}429$	$P_W(SC) = \frac{972}{8.706} = 0{,}112$	P (W) x P_U(SC) = 0,429 x 0,112 = 0,048
U: Unknown	$P(U) = \frac{11}{20.293} = 0{,}000$	-	-
Total	1	$P(SC) = \frac{3.430}{20.293} = 0{,}169$	

Table 10: Conditional (P_M (SC)) and joint probability (P (M) x P_M (SC)) – calculation scheme II. n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network between 2012-2019 (3.431 are *severe casualties*).

As we already know, among *severe casualties,* 21 % of male drivers and 11 % of female drivers were observed. This picture changes when calculating conditional or joint probabilities. The overall probability of a fatal or severe accident is 17 % (3.431/20.293). The conditional probability for a *severe casualty* given a male driver is 12 %. Given a female driver, it is 5 %, which, when added up, represents the value of 17 %.

4.3 Fisher's exact test and Phi coefficient

Fisher’s exact test determines whether the observed values in a 2x2 field table are subject to randomness or not. The test compares two dichotomous variables by calculating the probability of obtaining the observed data in the 2x2 field table. (Mehta & Patel, 1983)

In this thesis, Fisher's exact test evaluates the statistical relationship between accident-related characteristics and *severe casualties*. The test is suitable for 2x2 field tables with frequencies less than five, which is likely to occur among accident-related characteristics. We use the Phi coefficient to describe the strength of the determined statistical relationship. Figure 20 represents the logical framework of Fisher’s exact test.

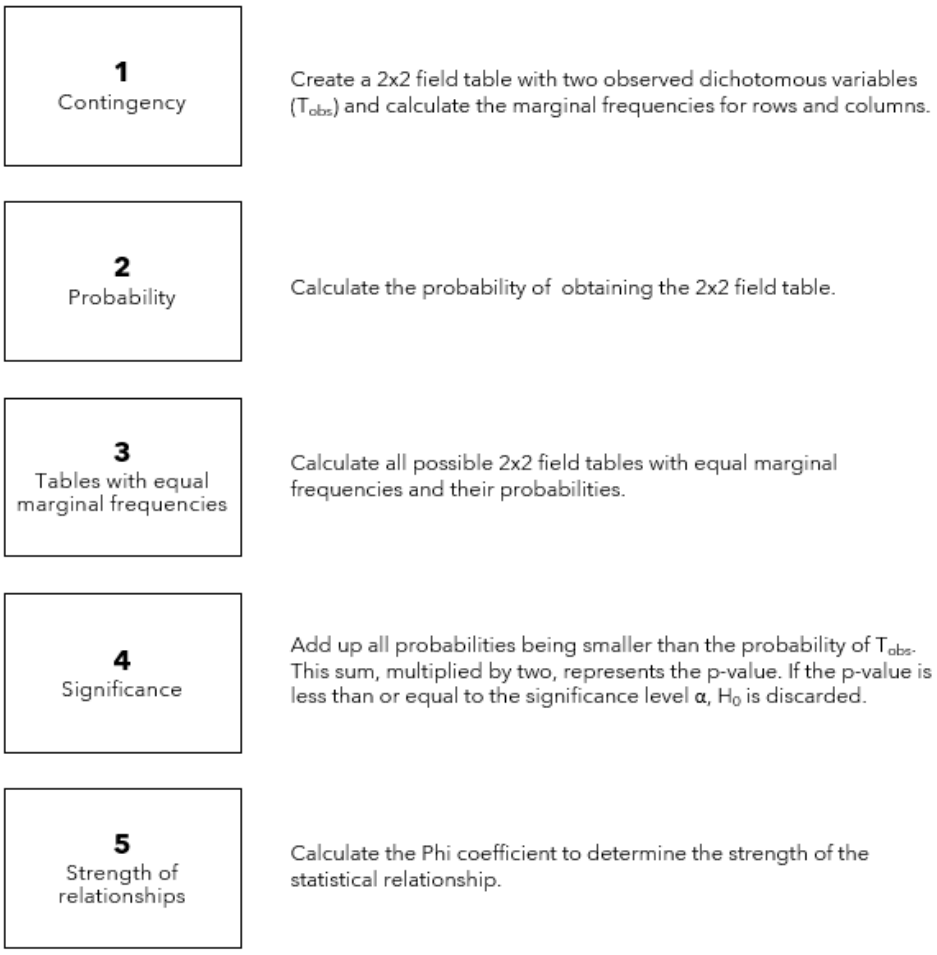

Figure 20: Logical framework for Fisher’s exact test.

We will now go through the calculation scheme for Fisher's exact test using a selected example (in this case, male drivers and the degree of injury). Table 11 illustrates the 2x2 field table containing the observed values from the accident sample.

	No Severe Casualty	Severe Casualty	Row total
Male drivers	9.118 (a)	2.458 (b)	11.576 (a + b)
No male drivers	7.745 (c)	972 (d)	10.604 (c + d)
Column total	16.863 (a + c)	3.430 (b + d)	20.293 (n = a + b + c + d)

Table 11: 2x2 field contingency table containing observed values (T_{obs}). n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network between 2012-2019 (3.431 are *severe casualties*).

Fisher's exact test calculates the probability of obtaining the observed 2x2 field table (p_{obs}) with the help of the hypergeometric probability function:

$$p = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!a!\,b!c!d!}$$

$$p_{obs} = \frac{11.576!\,10.604!\,16.863!\,3.430!}{20.293!9.118!\,2.458!7.745!972!} = 0{,}000$$

Step three in figure 20 shows that Fisher's exact test calculates the probabilities of all possible tables with equal marginal frequencies. The sum of all probabilities lower than the observed probability, multiplied by two, represents the p-value. At this point, it becomes evident that Fisher's exact test does not represent a manual procedure. A computer application is required to calculate all required probabilities automatically. However, to illustrate the process, three more tables below show how the resulting p-value is calculated (at least for a few steps). The following tables differ from the observed table (T_{obs}), but the marginal frequencies remain the same.

	No Severe Casualty	Severe Casualty	Row total
Male	9.068 (a)	2.448 (b)	11.576 (a + b)
Not Male	7.795 (c)	982 (d)	10.604 (c + d)
Column total	16.863 (a + c)	3.430 (b + d)	20.293 (n = a + b + c + d)

Table 12: Exemplary table (T_1) with equal marginal frequencies as $T_{obs.}$ n=20.293 (single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network between 2012-2019).

$$p_1 = \frac{11.576!\ 10.604!\ 16.863!\ 3.430!}{20.293!9.068!\ 2.448!7.795!982!} = 0{,}000$$

	No Severe Casualty	Severe Casualty	Row total
Male	9.143 (a)	2.445 (b)	11.576 (a + b)
Not Male	7.720 (c)	985 (d)	10.604 (c + d)
Column total	16.863 (a + c)	3.430 (b + d)	20.293 (n = a + b + c + d)

Table 13: Exemplary table (T_2) with equal marginal frequencies as $T_{obs.}$ n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network between 2012-2019 (3.431 are *severe casualties*).

$$p_2 = \frac{11.576!\ 10.604!\ 16.863!\ 3.430!}{20.293!9.143!\ 2.445!7.720!985!} = 0{,}000$$

	No Severe Casualty	Severe Casualty	Row total
Male	9.018 (a)	2.408 (b)	11.576 (a + b)
Not Male	7.845 (c)	1.022 (d)	10.604 (c + d)
Column total	16.863 (a + c)	3.430 (b + d)	20.293 (n = a + b + c + d)

Table 14: Exemplary table (T_3) with equal marginal frequencies as Tobs. n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network between 2012-2019 (3.431 are *severe casualties*). Source: Author's compilation.

$$p_3 = \frac{11.576!\ 10.604!\ 16.863!\ 3.430!}{20.293!9.018!\ 2.408!7.845!1.022!} = 0{,}000$$

The final formula for determining the p-value looks as follows (and would, of course, include the P-values of all possible 2x2 field tables).

$$p = 2x\ (p_1 + p_2 + p_3 + \ldots + p_n) = .000$$

For the observed table T_{obs}, the final p-value results in .000 and a phi coefficient of .133. A Phi coefficient of .133 represents a negligible relationship between male drivers and *severe casualties*. Table 15 shows the interpretation of the Phi coefficient.

ϕ	*Relationship*
+.70 and higher	very strong positive relationship
+.40 to +.69	strong positive relationship
+.30 to +.39	moderate positive relationship
+.20 to +.29	weak positive relationship
+.01 to +.19	no or negligible relationship
0	no relationship
-.01 to -.19	no or negligible relationship
-.20 to -.29	weak negative relationship
-.30 to -.39	moderate negative relationship
-.40 to -.69	strong negative relationship
-.70 and higher	very strong negative relationship

Table 15: Interpretation of the Phi coefficient.

Why can a phi coefficient be negative? This is easy to explain when looking at the Phi coefficient formula.

$$\phi = \frac{a \times d - b \times c}{\sqrt{(a+c)\ \times (b+d) \times (a+b) \times\ (c+d)}}$$

The Phi coefficient for table 11 looks as follows:

$$\phi = \frac{9.118 \times 972 - 2.458 \times 7.745}{\sqrt{(9.118+7.745) \times (2.458+972) \times (9.118+2.458) \times (7.745+972)}} = 0{,}116$$

A positive Phi coefficient indicates that the observed characteristic occurs comparatively often among *severe casualties*. A negative Phi coefficient does not mean that the characteristic does not occur among *severe casualties*. It comparatively occurs often among casualties with slight injuries.

4.4 Maximum combination value

In pattern recognition, it is interesting how often a variable occurs in combination with other variables. The maximum combination value (most frequent combination of the investigated variable with other variables) is an evaluation measure to determine whether a characteristic is part of a frequently occurring combination (blackpattern). The maximum combination value determines how often one characteristic appears in the same combination with other characteristics. It provides the basis for the PATTERMAX-method.

4.5 Bootstrapping and confidence intervals

The newly established binary accident database may serve as a source to develop an accident prediction model. We suggest a bootstrap resampling method for parameter estimation to establish robust predictive models. The bootstrap resampling method, introduced by Bradley Efron, draws samples out of an existing population with replacement. The estimation of confidence intervals represents a key application of the bootstrap methodology. Pei, Sze, Wong, and Yao (2016) apply the bootstrap resampling method to remove the effects of excess zeros on prediction performance. They conclude that the bootstrap resampling method generates more accurate and reliable parameter estimates, including reduced standard errors. By drawing samples out of an existing population, bootstrapping approximates a distribution of the observed data. Bootstrapping is a reliable method to use if the distribution of the observed data is unknown. It is a non-parametric method to estimate the parameters of a population. In this case, we use bootstrapping to calculate 95% confidence intervals for accident-related characteristics. The confidence intervals indicate a probability range for a characteristic among *severe casualties*. We use bias-corrected and accelerated bootstrap (BCa) to generate narrow confidence intervals. BCa is less prone to imbalances than other bootstrapping methods. In total, we draw 10.000 samples to calculate the 95% confidence intervals. Additional statistical figures for each accident-related variable comprise variance s^2, standard deviation σ, and standard error SEM.

4.6 Analysis of driver-related variables

Driver-related variables include sex, age class, driving licence, impairments, driving manoeuvres before the accident, and safety settings. Table 16 illustrates the driver-related variables and their detailed characteristics (54 in total).

Variable	*Characteristics*
Sex	male, female, unknown
Age Class	16 to 18, 19 to 24, 25 to 34, 35 to 44, 45 to 54, 55 to 64, 65+
Driving licence	no driving licence, probationary driving licence
Impairments	alcohol, distraction, fatigue, health, drugs, medicine, excitation
Driving manoeuvres	speeding, skidding/drifting, hitting an obstacle next to the road, hitting the guard rail, hitting a tree, misconduct by the pedestrian, hit and run, sudden braking, overtaking, cutting curves, hitting an obstacle on the road, changing lanes, inadequate safety distance, reverse driving, phoning, turning around, fall from the vehicle, getting in the lane, disregarding driving direction, priority violation, driving towards the left side of the road, forbidden overtaking, hitting a moving vehicle, disregarding driving ban, driving in parallel, opening the vehicle door, hitting a stationary vehicle, wrong-way driver, disregarding red light, dangerous stopping and parking, disregarding turning ban, missing indication of direction change, driving against one way, driving without mandatory light
Safety settings	no safety belt applied

Table 16: Driver-related variables and their characteristics.

The analysis of driver-related variables foresees the calculation of variable frequencies, conditional and joint probabilities, Fisher's exact test and Phi coefficient, the maximum combination maximum value, and confidence intervals.

As table 17 shows, male drivers hold a share of 57 % (n=11.576) among single-vehicle accidents with single occupancy and personal injury occurring outside the built-up area between 2012 and 2019 (n=20.293), while female drivers hold a share of 43 % (n=8.706). 0,05 % of the road traffic accident records do not indicate the sex of the driver (n=11). A severe or fatal accident with a male driver results in a joint probability of 12 %, whereas only 5 % with a female driver. Fisher's exact test is highly significant for both sex groups regarding their relationship with *severe casualties*. The Phi coefficient indicates that male drivers increase the risk to observe a severe casualty. However, the phi coefficient indicates this relationship to be negligible.

Variable X **Sex**	C: Casualties n	SC: Severe Casualties n	P (X ∩ SC) %	Fisher's exact test [Y=severe casualty] *p*	Phi coefficient [Y=severe casualty] ϕ	Comb Max [driver-related variables] n
Male	11.576	2.458	12,11%	,000	,133	817
Female	8.706	972	4,79%	,000	-,133	1.132
Unknown	11	1	-	-	-	-
Total	**20.293**	**3.431**				

Table 17: Single-vehicle accidents with single occupation and personal injury that occurred outside the built-up area between 2012 and 2019 broken down by sex. n=20.293 (3.431 are *severe casualties*).

The characteristic 'female driver' is part of the most frequent combination (blackpattern) of driver-related characteristics (a combination occurring 1.132 times). The most frequent combination of driver-related characteristics involving male drivers occurs 817 times. The most frequent blackpattern among female drivers represents 13 % of all accidents involving female drivers. In contrast, the most frequent blackpattern among male drivers represents a share of 7 % of all accidents among male drivers. Based on these numbers, the question arises if accidents involving female drivers result in fewer blackpatterns than accidents involving male drivers. Generally, the more often one blackpattern occurs and the fewer blackpatterns exist, the more targeted traffic safety work and strategies can be delineated. A detailed representation of blackpatterns among female and male drivers follow in chapter 5.1. Figure 21 illustrates the estimated 95% confidence intervals for both sex groups. The estimation is based on a bias-corrected and accelerated bootstrap resampling (BCa) with 10.000 samples. The confidence intervals indicate a probability range for a characteristic among *severe casualties*. For male drivers, the 95% confidence interval ranges from 11,9 % to 12,4 %, for female drivers from 4,5 % to 5,1 %. The standard error SEM for both variables is 0,13 %.

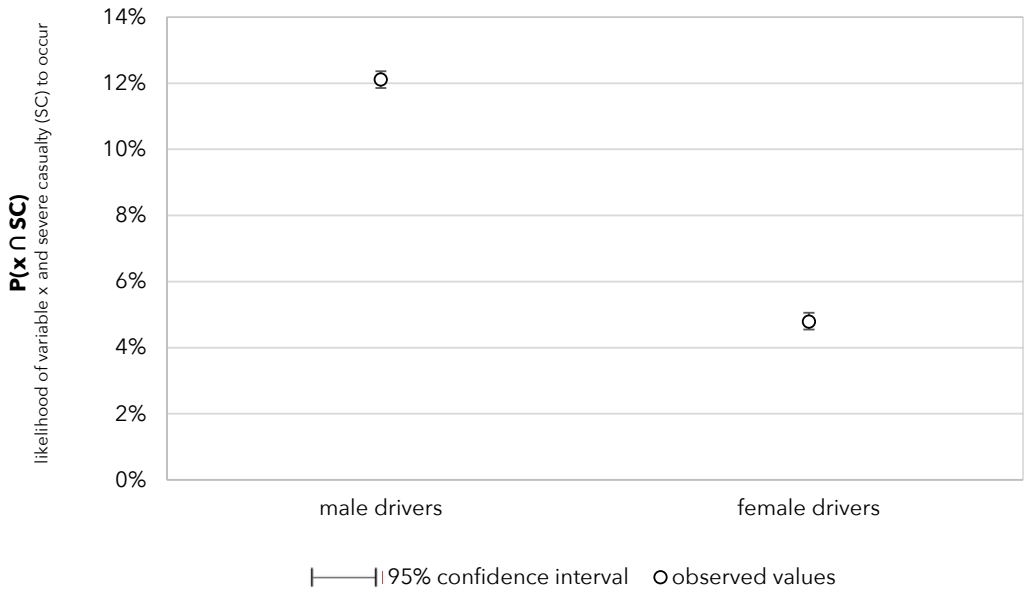

Figure 21: 95% confidence intervals for male drivers or female drivers. The confidence intervals estimate the likelihood of the variable and *severe casualties* to occur (range for joint probability). n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

We continue to with the analysis of different age classes. Table 18 shows that age classes 19 to 24 (n=806) and 25 to 34 (n=697) have the highest casualties and *severe casualties*. *Severe casualties* among both age classes tend to be almost equally likely. Interestingly, Fisher's exact test is not significant for *severe casualties* and age class 25 to 34, but highly significant for *severe casualties* and age class 19 to 24. Except for age class 35 to 44, all other age classes are significantly correlated with *severe casualties*. The phi coefficient shows a negligible relationship (values between ±,01 to ±,19) for all ages classes and *severe casualties*. Each age class shows a relatively high maximum combination value.

Variable X **Age class**	C: Casualties n	SC: Severe Casualties n	P (X ∩ SC) %	Fisher's exact test [Y=severe casualty] *p*	Phi coefficient [Y=severe casualty] ϕ	Comb Max [driver-related variables] n
16 to 18	1.465	162	0,80%	,000	-,044	171
19 to 24	6.547	806	3,97%	,000	-,085	1.132
25 to 34	4.323	697	3,43%	,120	-,011	830
35 to 44	2.488	468	2,31%	,008	,019	432
45 to 54	2.180	476	2,35%	,000	,046	382
55 to 64	1.404	323	1,59%	,000	,044	212
64 and higher	1.878	499	2,46%	,000	,082	303
unknown	8					
total	20.293	3.431				

Table 18: Single-vehicle accidents with single occupation and personal injury that occurred outside the built-up area between 2012 and 2019 broken down by age class. n=20.293 (3.431 are *severe casualties*).

Figure 22 shows the age distribution (in this case, on a metric scale) among accidents involving male or female drivers, broken down by accidents with minor injury accidents and severe and fatal accidents (*severe casualties*). The illustrated violin plot confirms that most accidents occur in the two age classes 19 to 24 and 25 to 34.

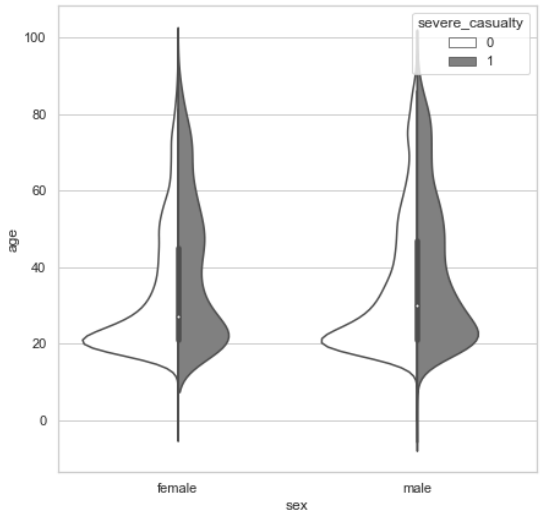

Figure 22: Age distribution in accidents involving male and female drivers, divided into accidents with a minor injury and severe or fatal accidents. n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network between 2012-2019 (3.431 are *severe casualties*). The violin plot represents a probability density function.

The 95 % confidence intervals for each age class are shown in figure 23 and present the following probability ranges:

- age class 16 to 18: from 0,68 % to 0,93 %, standard error of 0,06 %
- age class 19 to 24: from 3,37 % to 4,21 %, standard error of 0,12 %
- age class 25 to 34: from 3,21 % to 3,66 %, standard error of 0,12 %
- age class 35 to 44: from 2,11 % to 2,51 %, standard error of 0,10 %
- age class 45 to 54: from 2,14 % to 2,53 %, standard error of 0,99 %
- age class 55 to 64: from 1,43 % to 1,76 %, standard error of 0,09 %
- age class 65+: from 2,25 % to 2,66 %, standard error of 0,10 %

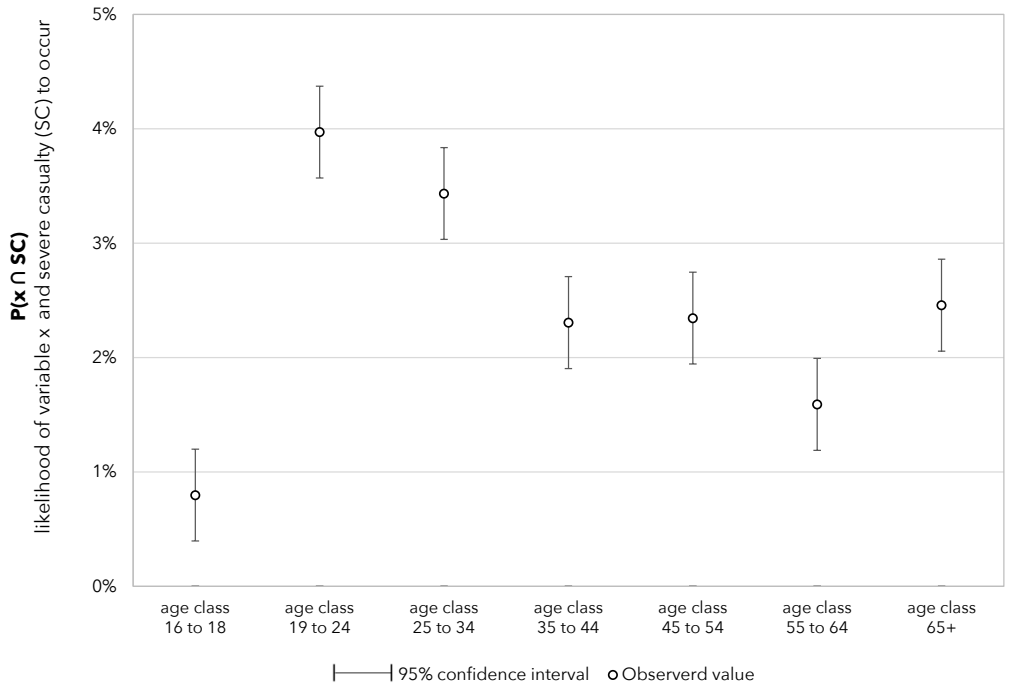

Figure 23: 95% confidence intervals for different age classes. The confidence intervals estimate the likelihood of the variables and *severe casualties* to occur (range for joint probability). n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

We now investigate the variable 'driving licence type' with the two characteristics 'probationary driving licence and 'no driving licence'. Table 19 shows the detailed results for both characteristics.

Variable X **Driving licence type**	C: Casualties n	SC: Severe Casualties n	P (X ∩ SC) %	Fisher's exact test [Y=severe casualty] *p*	Phi coefficient [Y=severe casualty] ϕ	Comb Max [driver-related variables] n
No driving licence	356	94	0,46%	,020	,034	15
Probationary driving licence	2.805	303	1,49%	,000	-,065	391

Table 19: Single-vehicle accidents with single occupation and personal injury that occurred outside the built-up area between 2012 and 2019 broken down by driving licence type. n=20.293 (3.431 are *severe casualties*).

The variable 'probationary driving licence' appears to correlate with *severe casualties* significantly. Yet, the strength of the relationship is negligible. Chapter eight will search for specific variable combinations (blackpatterns) that include 'probationary driving licence'. A better understanding of accident circumstances may help design appropriate intervention measures for this target group (young drivers, respectively). The number of casualties and *severe casualties* involving drivers without driving licences is surprisingly high (a total value of 94 in our observation period 2012-2019). Figure 24 illustrates the 95 % confidence intervals for both characteristics. The likelihood of observing a severe road traffic accident involving a driver with a 'probationary driving licence' ranges from 1,3 % to 1,7 %, with a standard error of 0,08 %. The 95% confidence interval for 'no driving licence' ranges from 0,4 % to 0,6 %, with a standard error of 0,05 %.

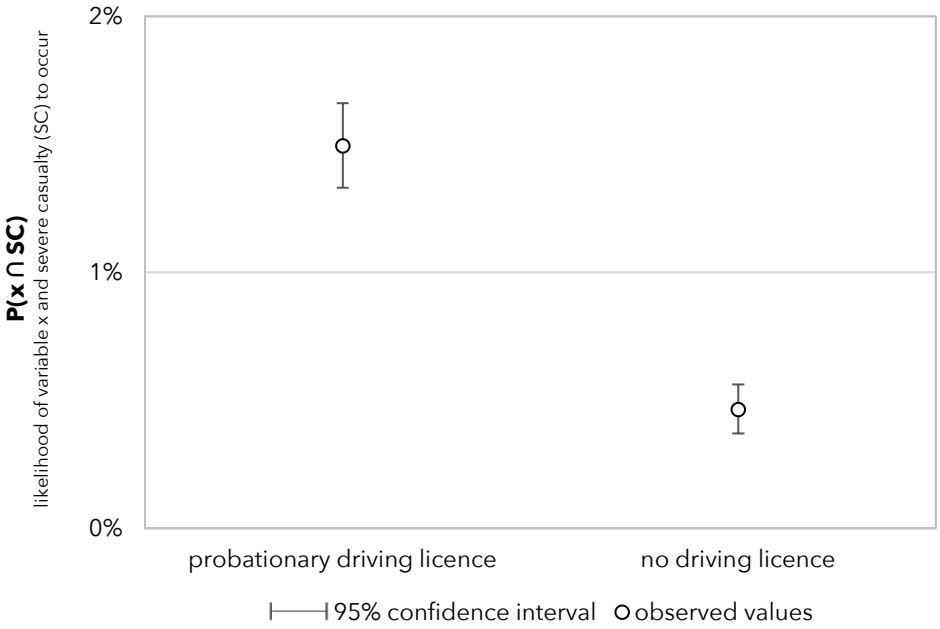

Figure 24: 95% confidence intervals for 'probationary driving licence' and 'no driving licence'. The confidence intervals estimate the likelihood of the variables and severe casualties to occur (range for joint probability). n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

Driver-related variables comprise different types of impairment, as table 20 shows. 'Alcohol' and 'distraction' are the most frequently observed types of impairment but do not appear to have a significant relationship with the target variable *severe casualties*. 'Fatigue' is the only type of impairment indicating a significant relationship with *severe casualties*.

Variable X **Impairment**	C: Casualties n	SC: Severe Casualties n	P (X ∩ SC) %	Fisher's exact test [Y=severe casualty] *p*	Phi coefficient [Y=severe casualty] ϕ	Comb Max [driver-related variables] n
Alcohol	2.858	481	2,37%	,934	-,001	246
Distraction	2.369	431	2,12%	,079	,012	93
Fatigue	1.518	317	1,56%	,000	,030	134
Health	432	91	0,45%	,021	,016	38
Drugs	66	15	0,07%	,247	,009	3
Medicines	50	10	0,05%	,570	,004	2
Excitation	7	2	0,01%	,337	,006	1

Table 20: Single-vehicle accidents with single occupation and personal injury that occurred outside the built-up area between 2012 and 2019 broken down by impairment. n=20.293 3.431 are *severe casualties*).

One aspect of discovering with blackpatterns (chapter eight) is if there exist differences in impairments between female and male drivers. To get an insight into the answer to that question, we illustrate violin plots including age (in this case, on a metric scale), impairment, and sex for the three most frequently observed types of impairment in figure 25. The violin plots suggest that 'alcohol' holds a higher share among male and mostly younger drivers. The age distribution among female drivers impaired by 'alcohol' is distributed across all age classes up to sixty years. The violin plot illustrating the distribution of 'fatigue' among age classes and both sex groups suggests that fatigue primarily occurs among younger age classes when looking at male drivers. The probability density function is almost equally distributed over the age classes for female drivers, with a slight drop around 30-40. 'Distraction' appears to occur mostly among younger drivers. All three types of impairment have their peaks within younger age classes.

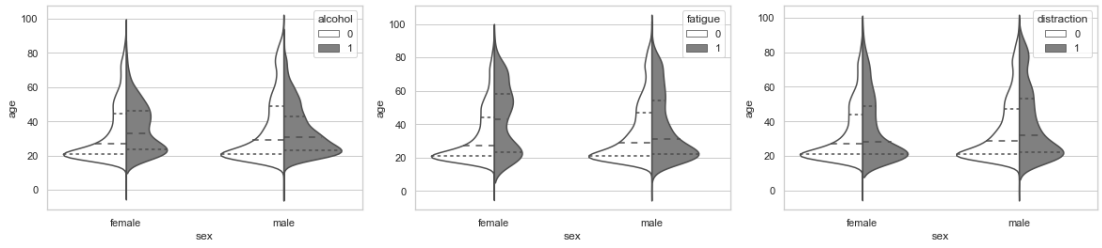

Figure 25: Distribution of age, impairment and sex among the observed road traffic accidents. n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network between 2012-2019 (3.431 are *severe casualties*). The violin plots represent a probability density function.

We conclude the detailed analysis of impairments with the illustration of the 95 % confidence intervals for the top three types of impairment. The confidence intervals represent the probability ranges for the type of impairment to occur with *severe casualties*. For 'alcohol', the range goes from 2,17 % to 2,58 %, with a standard error of 0,10 %. 'Distraction' ranges from 1,94 % to 2,32 % and has a standard error of 0,09 %. 'Fatigue' ranges from 1,39 % to 1,73 % and has a standard error of 0,08 %.

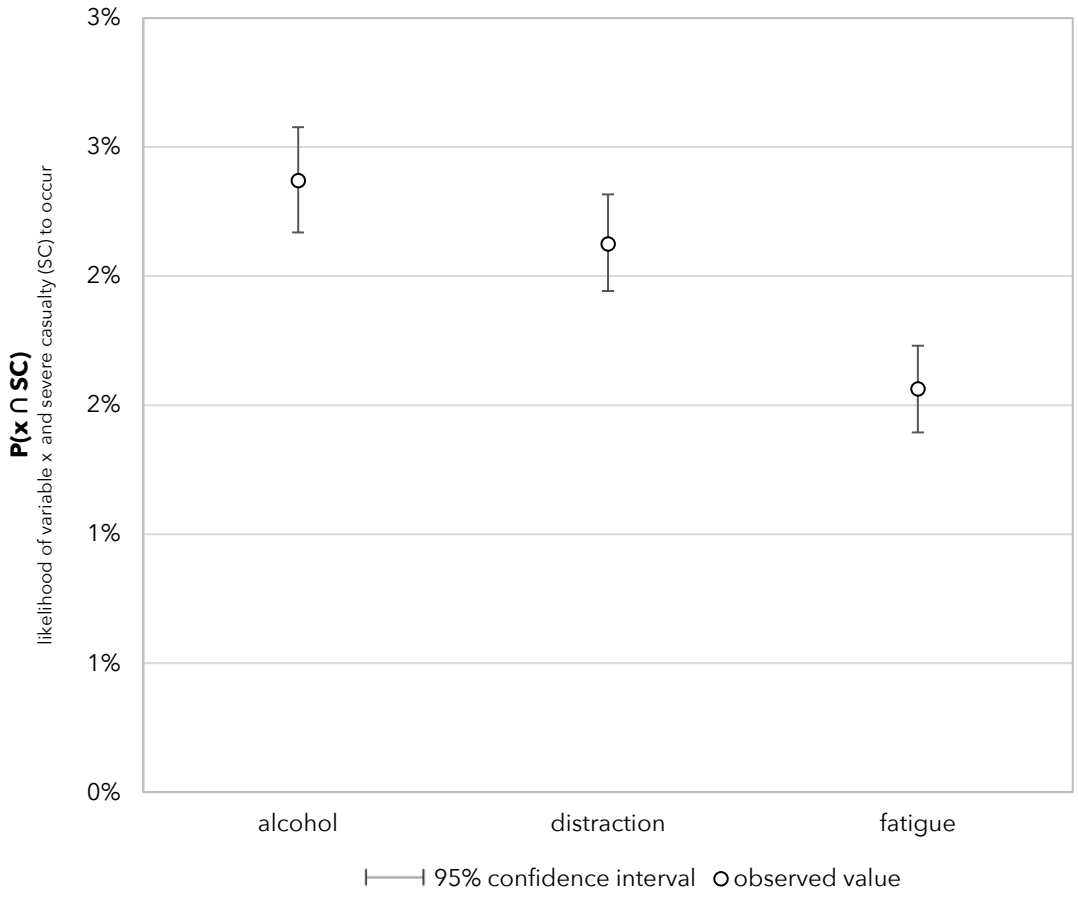

Figure 26: 95% confidence intervals for 'alcohol', 'distraction' and 'fatigue'. The confidence intervals estimate the likelihood of the variables and *severe casualties* to occur (range for joint probability). n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

After analysing sex, age classes, types of driving licence, and impairments, we evaluate recorded driving manoeuvres before the accident. Table 21 illustrates detailed results for different driving manoeuvres such as speeding, hitting obstacles, turning around, overtaking etc.

Variable X **Driving manoeuvres**	C: Casualties n	SC: Severe Casualties n	P (X ∩ SC) %	Fisher's exact test [Y=severe casualty] *p*	Phi coefficient [Y=severe casualty] ϕ	Comb Max [driver-related variables] n
Speeding	3.608	579	2,85%	,136	-,011	131
Skidding	1.823	239	1,18%	,000	-.032	80
Hitting an obstacle next to the road	1.512	280	1,38%	,086	,012	35
Hitting the guard rail	1.378	181	0,89%	,000	-,027	37
Hitting a tree	1.217	318	1,57%	,000	,062	23
Misconduct by pedestrians	503	79	0,39%	,505	-,005	12
Hit and run	371	53	0,26%	,186	-,010	22
Sudden braking	149	11	0,05%	,002	-.022	9
Overtaking	147	26	0,13%	,834	,002	8
Cutting curves	128	27	0,13%	,194	,009	4
Hitting an obstacle on the road	117	6	0,03%	,001	-,024	7
Changing lanes	58	9	0,04%	1,000	-,002	3
Inadequate safety distance	38	7	0,03%	,828	,002	1
Reverse driving	26	6	0,03%	,429	,006	2
Phoning	25	7	0,03%	,175	,010	1
Turning around	22	4	0,02%	,780	,001	3
Fall from the vehicle	22	11	0,05%	,000	,029	2
Getting in lane	18	4	0,02%	,529	,004	1
Disregarding driving direction	16	2	0,01%	1,000	-,003	1
Priority violation	15	4	0,02%	,302	,007	1
Driving towards the left-hand side of the road	9	3	0,01%	,184	,009	1
Forbidden overtaking	8	2	0,01%	,630	,004	1
Hitting a moving vehicle	8	0	0,00%	,367	-,009	2
Disregarding driving ban	5	2	0,01%	,201	,010	1

Table 21: Single-vehicle accidents with single occupation and personal injury that occurred outside the built-up area between 2012 and 2019 broken down by driving manoeuvre. n=20.293 (3.431 are *severe casualties*).

Variable X **Driving manoeuvres**	C: Casualties n	SC: Severe Casualties n	P (X ∩ SC) %	Fisher's exact test [Y=severe casualty] p	Phi coefficient [Y=severe casualty] φ	Comb Max [driver-related variables] n
Driving in parallel	5	1	0,00%	1,000	,604	1
Opening the vehicle door	5	2	0,01%	,201	,010	1
Hitting a stationary vehicle	3	0	0,00%	1,000	-,005	1
Wrong-way driver	1	0	0,00%	1,000	-,003	1
Disregarding red light	1	0	0,00%	1,000	-,003	1
Dangerous stopping and parking	0	0	-	-	-	-
Disregarding turning ban	0	0	-	-	-	-
Missing indication of direction change	0	0	-	-	-	-
Driving against one-way	0	0	-	-	-	-
Driving without mandatory light	0	0	-	-	-	-

Continuation of table 21: Single-vehicle accidents with single occupation and personal injury that occurred outside the built-up area between 2012 and 2019 broken down by driving manoeuvre. n=20.293 (3.431 are *severe casualties*).

'Speeding' and 'skidding' are the most frequently observed driving manoeuvres before the accident. Of these two variables, skidding shows a significant relationship with *severe casualties*. Also, the joint probability for speeding and a severe or fatal accident is more than twice as high as for skidding. 'Hitting the guard rail', 'hitting a tree' and 'hitting an obstacle on the road' appear to correlate with the target variable *severe casualties* significantly. Of these three, 'hitting the guard rail' and 'hitting a tree' show relatively high frequencies among *severe casualties,* whereas 'hitting an obstacle on the road' occurs only six times.

The two most frequently observed driving manoeuvres before the accident were 'speeding' and 'skidding'. Figure 27 illustrates the corresponding violin plots, including age (in this case, on a metric scale level) and sex. The violin plots do not show differences among female drivers and male drives. Both characteristics show their peak among younger age classes.

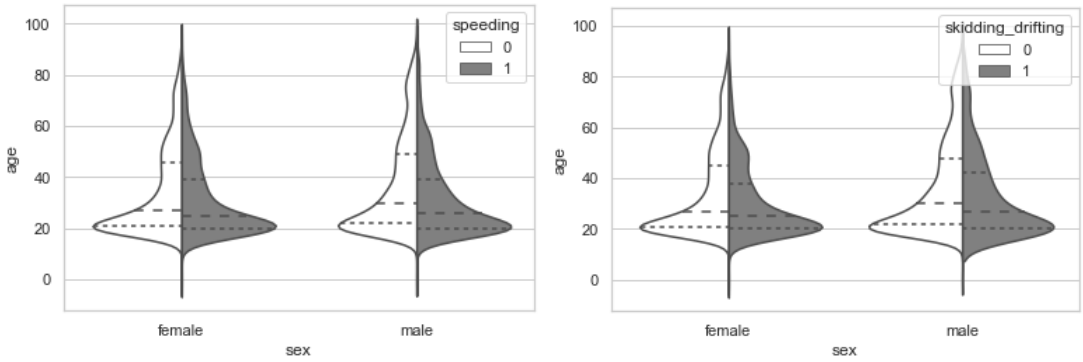

Figure 27: Distribution of age, driving manoeuvres and sex among the observed road traffic accidents. n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network between 2012-2019 (3.431 are *severe casualties*). The violin plot represents a probability density function.

The 95 % confidence interval for 'speeding' ranges from 2,65 % to 3,07 %, and the distribution has a standard error of 0,11 %. For 'skidding', the range spans 1,03 % to 1,33 % and has a standard error of 0,07 %. Figure 28 illustrates the confidence intervals for both characteristics.

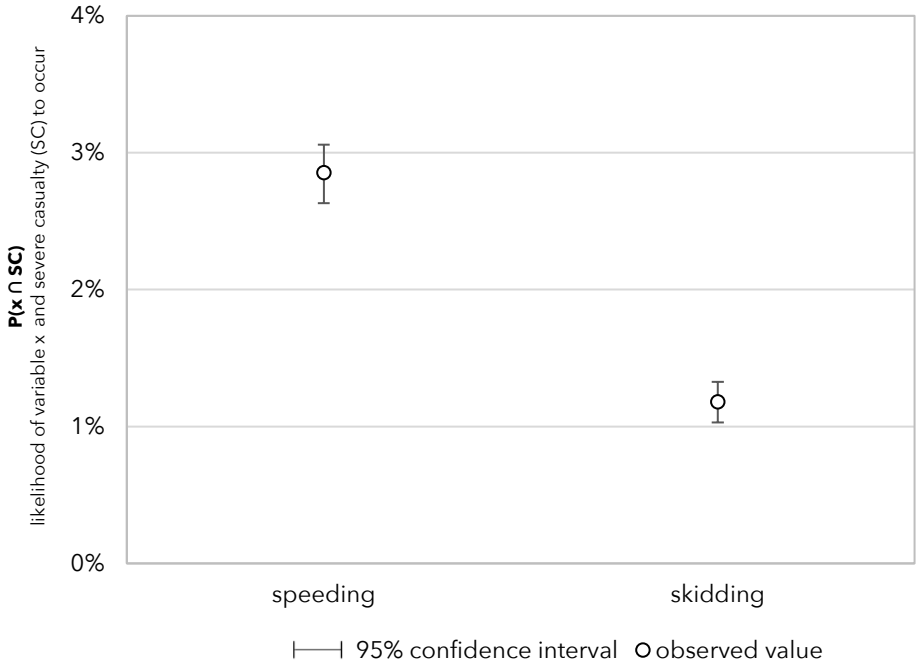

Figure 28: 95% confidence intervals for 'speeding and 'skidding'. The confidence intervals estimate the likelihood of the variables and *severe casualties* to occur (range for joint probability). n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

The investigation of safety settings concludes the analysis of driver-related variables. Table 22 illustrates the recorded values for the variable 'no safety belt applied'. Besides sex and age class, *severe casualties* show the highest joint probability given the 'no safety belt applied' variable. Table 22 reveals the impact of the characteristic 'no safety belt applied' on *severe casualties*. Compared to the other 54 driver-related characteristics, 'no safety belt applied' shows the highest phi coefficient (0,240) and goes along with a higher number of *severe casualties* than the variable 'no safety belt applied'.

Variable X **Safety settings**	C: Casualties n	SC: Severe Casualties n	P (X ∩ SC) %	Fisher's exact test [Y=severe casualty] *p*	Phi coefficient [Y=severe casualty] ϕ	Comb Max [driver-related variables] n
No safety belt applied	1.401	699	3,44%	,000	,240	60

Table 22: Single-vehicle accidents with single occupation and personal injury that occurred outside the built-up area between 2012 and 2019 broken down by safety settings. n=20.293 (3.431 are *severe casualties*).

Figure 29 suggests that the characteristic 'no safety belt applied' primarily occurs among young male and female drivers. For female drivers, the peak is lower among younger age classes.

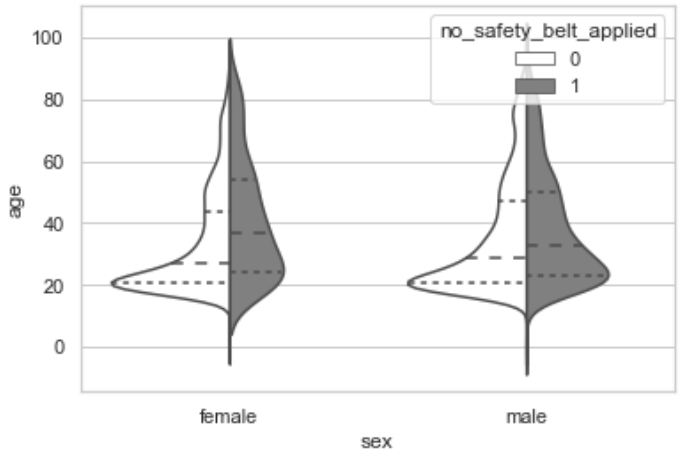

Figure 29: Distribution of age, 'no safety belt applied' and sex among the observed road traffic accidents. n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network between 2012-2019 (3.431 are *severe casualties*). The violin plot represents a probability density function.

The 95 % confidence interval for the characteristics 'no safety belt applied' ranges from 3,21 % to 3,68 %, and the distribution has a standard error of 0,02 % (see figure 30).

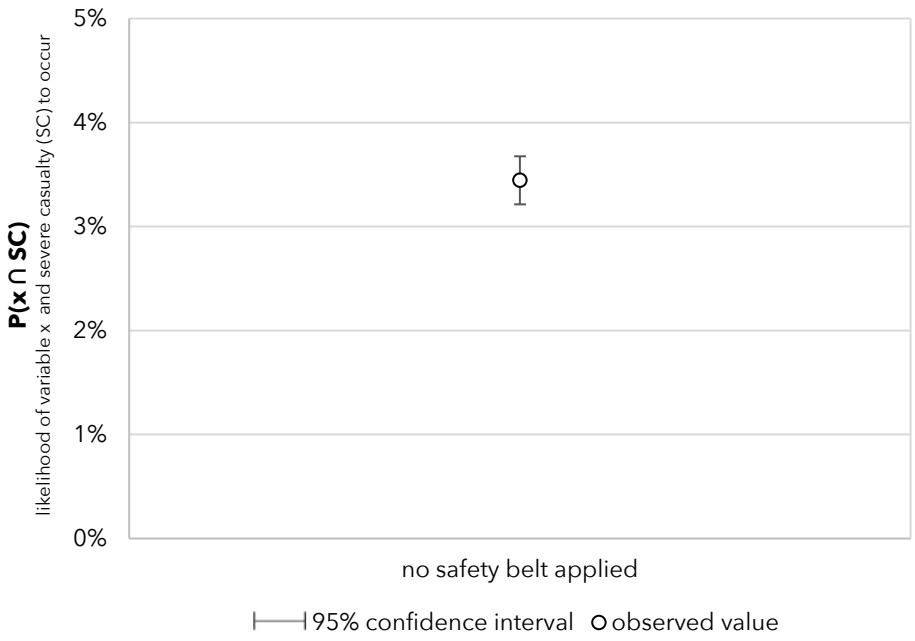

Figure 30: 95% confidence intervals for 'no safety belt applied'. The confidence intervals estimate the likelihood of the characteristics and *severe casualties* (range for joint probability). n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

The following figure 31 provides an overview of the presented driver-related accident characteristics. It intends to zoom into the shares of these characteristics among *severe casualties* (severe and fatal accidents) with male and female drivers. *Severe casualties* with male drivers have a total of 2.458, and *severe casualties* with female drivers have a total of 972. Among *severe casualties*, the share of female drivers owning a 'probationary driving licence' is higher. The share of male drivers not applying a safety belt is twice as high as for female drivers not applying a safety belt. The variables 'speeding' and 'fatigue' do not differ between *severe casualties* with female and male drivers. Regarding 'alcohol', the share of male drivers impaired by alcohol is three times higher than female drivers. 'Distraction' shows a higher share among *severe casualties* with female drivers.

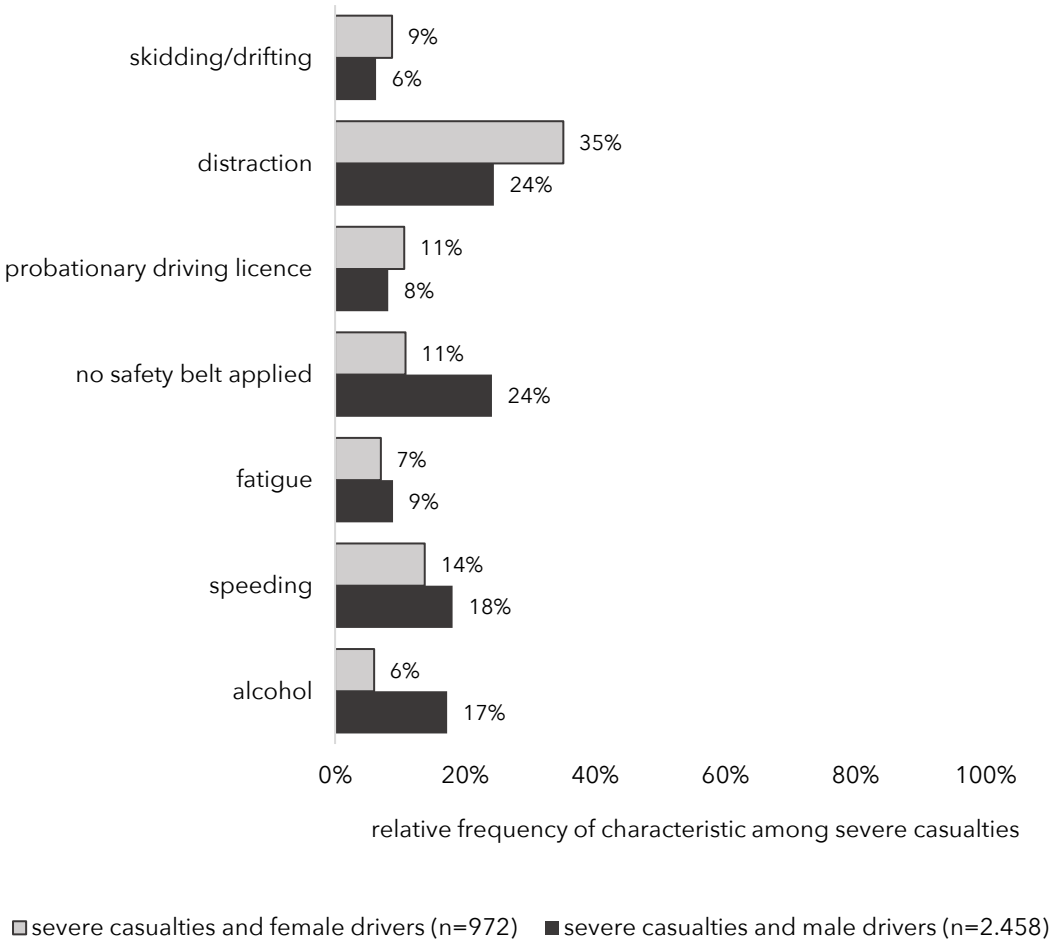

Figure 31: Relative frequencies (or conditional probabilities) of selected driver-related characteristics among *severe casualties* with male and female drivers. n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network between 2012-2019 (3.431 are *severe casualties*).

4.7 Analysis of vehicle-related variables

Vehicle-related variables comprise engine power, kilometrage, vehicle colour and vehicle safety settings. Table 23 illustrates vehicle-related variables and their characteristics. The analysis of vehicle-related variables includes variable frequencies and conditional and joint probabilities of the variable characteristics to occur among *severe casualties*. Fisher's exact test examines whether a statistical relationship exists between a characteristic and *severe casualties*. The Phi coefficient estimates the strength of this relationship. The maximum combination value illustrates the most frequent variable combination for each vehicle-related characteristic.

Variable	*Characteristic*
Engine power (kW)	0-24, 24-90, 90-110, 100+
Kilometrage (km)	0 to 15.000, 15.000 to 75.000, 75.000 to 100.000, 100.000 to 150.000, 150.000 to 200.000
Vehicle colour	beige, blue, brown, bronze, dark, yellow, gold, grey, green, bright, orange, red, black, silver, purple, white, others
Vehicle safety settings	insufficient vehicle security, insufficient load security, technical defects, vehicle fire, airbag deployed, airbag not deployed

Table 23: Vehicle-related variables.

We start with a detailed analysis of the variable 'engine power'. Table 24 shows that all engine power classes show a significant relationship with *severe casualties* except for engine power class '0 to 24 kW'. According to the Phi coefficient, the strength of this relationship is negligible. Engine power from '24-90 kW' shows the highest frequency among casualties and *severe casualties*, followed by engine power from '90-110kW' and '100+ kW' (both classes hold almost the same share among all casualties and *severe casualties*).

Variable X **Engine Power [kW]**	C: Casualties n	SC: Severe Casualties n	P (X ∩ SC) %	Fisher's exact test [Y=severe casualty] *p*	Phi coefficient [Y=severe casualty] ϕ	Comb Max [vehicle- related variables] n
0-24	11	3	0,01%	,411	,006	2
24-90	15.412	2.393	11,79%	,000	-,066	975
90-110	1.928	413	2,04%	,000	,039	201
110+	1.947	448	2,21%	,000	,053	256

Table 24: Single-vehicle accidents with single occupation and personal injury that occurred outside the built-up area between 2012 and 2019 divided by engine power. n=20.293 (3.431 are *severe casualties*).

Kilometrage (see table 25) appears to have no significant relationship with *severe casualties*. The probability of a severe or fatal accident appears to rise with vehicle kilometrage. The Phi coefficient shows a negligible strength of relationship for all kilometrage classes.

Variable X **Kilometrage [km]**	C: Casualties n	SC: Severe Casualties n	P (X ∩ SC) %	Fisher's exact test [Y=severe casualty] *p*	Phi coefficient [Y=severe casualty] ϕ	Comb Max [vehicle- related variables] n
0 to 15.000	156	24	0,12%	,662	-,004	13
15.000 to 75.000	605	89	0,44%	,154	-,010	51
75.000 to 100.000	387	70	0,34%	,541	,004	33
100.000 to 150.000	663	104	0,51%	,428	-,006	44
150.000 to 200.000	942	176	0,87%	,141	,010	56

Table 25: Single-vehicle accidents with single occupation and personal injury that occurred outside the built-up area between 2012 and 2019 divided by kilometrage. n=20.293 (3.431 are *severe casualties*).

The analysis of vehicle-related variables proceeds with vehicle colour and its relationship with *severe casualties* (see table 26). Except for blue and green, no vehicle colour has a significant relationship with *severe casualties*. *Severe casualties* with black, blue, grey, and red vehicles show the highest frequencies.

Variable X **Vehicle Colour**	C: Casualties n	SC: Severe Casualties n	P (X ∩ SC) %	Fisher's exact test [Y=severe casualty] *p*	Phi coefficient [Y=severe casualty] ϕ	Comb Max [vehicle-related variables] n
Beige	18	3	0,01%	1,000	,000	5
Blue	3.166	478	2,36%	,003	-,021	868
Brown	193	35	0,17%	,637	,003	52
Bronze	1	0	0,00%	1,000	-,003	1
Dark	30	6	0,03%	,626	,003	6
Yellow	129	18	0,09%	,408	-,006	37
Gold	18	3	0,01%	1,000	,000	5
Grey	2.702	462	2,28%	,784	,002	770
Green	1.219	262	1,29%	,000	,031	281
Bright	8	2	0,01%	,630	,004	2
Orange	130	24	0,12%	,647	,003	41
Red	2.272	381	1,88%	,857	-,001	602
Black	3.981	652	3,21%	,334	-,007	958
Silver	716	136	0,67%	,127	0,11	146
Purple	49	8	0,04%	1,000	-,001	11
White	1.907	323	1,59%	,977	,000	497
Others	1	1	0,00%	,169	,016	1

Table 26: Single-vehicle accidents with single occupation and personal injury that occurred outside the built-up area between 2012 and 2019 divided by vehicle colour. n=20.293 (3.431 are *severe casualties*).

The analysis of vehicle-related variables concludes with investigating vehicle safety settings (see table 27). Here we can see that the variable 'airbag not deployed' shows a comparatively high probability of occurrence (4 %) and a comparatively high combination maximum value (975). The other variables relating to vehicle safety settings (i.e., insufficient vehicle security, insufficient load securing, technical defects, and vehicle fire) show a low maximum combination value. Vehicle fire has a significant relationship with *severe casualties*: out of 18 accidents with a vehicle fire, 11 resulted in a *severe casualty*. Compared with all the Phi coefficient values we have seen thus far, the Phi coefficient between 'airbag not deployed' and *severe casualties* is relatively high. However, it still indicates a negligible relationship.

Variable X **Vehicle safety settings**	C: Casualties n	SC: Severe Casualties n	P (X ∩ SC) %	Fisher's exact test [Y=severe casualty] *p*	Phi coefficient [Y=severe casualty] ϕ	Comb Max [vehicle-related variables] n
Insufficient vehicle security	16	6	0,03%	,040	,015	2
Insufficient load securing	6	0	0,00%	,598	-,008	1
Technical defects	102	15	0,07%	,682	-,004	6
Vehicle fire	18	11	0,05%	,000	,035	1
Airbag not deployed	8.138	819	4,04%	,000	-,149	975

Table 27: Single-vehicle accidents with single occupation and personal injury that occurred outside the built-up area between 2012 and 2019 divided by vehicle safety settings. n=20.293 (3.431 are *severe casualties*).

Figure 32 shows the 95 % confidence interval for the characteristics 'airbag not deployed' to occur among *severe casualties*. The confidence interval represents a probability range from 3,80 % to 4,28 %. The distribution has a standard error of 0,12 %.

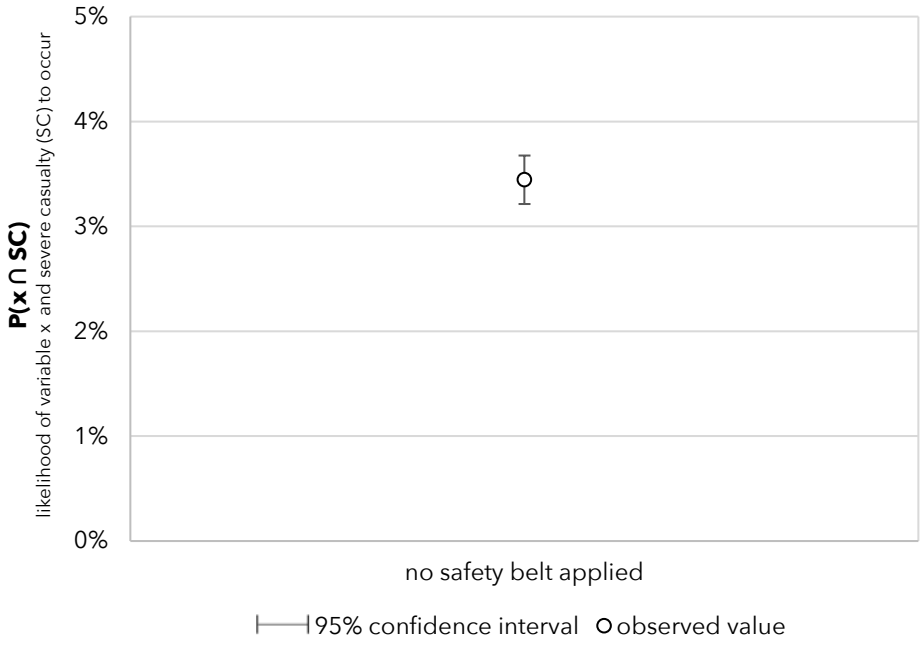

Figure 32: 95% confidence intervals for 'airbag not deployed'. The confidence intervals estimate the likelihood of the characteristics and *severe casualties* to occur (range for joint probability). n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

4.8 Analysis of roadway-related variables

Roadway-related variables comprise speed limit, road characteristics, traffic lights, road type, and road condition. Table 28 illustrates all roadway-related variables and their characteristics.

Variable	*Characteristics*
Speed limit (km/h)	Driving ban, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130
Road characteristics	Intersection, Roundabout, Deceleration Lane, Acceleration Lane, One-way, Construction site, Cycle path, Crosswalk, Pedestrian and cycle path, Parking Lane, Secondary Lane, Hard shoulder, Banquet, Straight Road, Tunnel, Gallery, Rest area, Traffic Island, Underpass, Middle separation, Bridge, Curve, Narrow Lane, Entry or exit, Tram or bus station
Traffic light	Traffic light in operation
Road type	Highway, Expressway, Regional Road, Other roads
Road condition	Dry road, Wet Road, Sand or grit on the road, Winter conditions, other conditions (oil, soil)

Table 28: Roadway-related variables.

The analysis of roadway-related variables and their characteristics foresees the calculation of variable frequencies, joint and conditional probabilities, Fisher's exact test and the Phi coefficient, and the maximum combination value. The following paragraphs summarize the key insights of analysing roadway-related variables.

Table 29 represents the distribution and statistics of different speed limits (km/h) regarding *severe casualties*. According to Fisher's exact test results, no speed limit shows a significant relationship with *severe casualties*. The speeds limits of 50 km/h, 70 km/h, 80 km/h, 100 km/h, and 130 km/h result in relatively high maximum combination values. 63 % (2.148) of *severe casualties* occur within a speed limit of 100km/h.

Variable X **Speed limit (km/h)**	C: Casualties n	SC: Severe Casualties n	P (X ∩ SC) %	Fisher's exact test [Y=severe casualty] *p*	Phi coefficient [Y=severe casualty] ϕ	Comb Max [roadway-related variables] n
Driving ban	2.270	380	1,87%	,833	-,002	350
5	1	1	0,00%	,169	,016	1
10	1	0	0,00%	1,000	-,003	1
20	2	0	0,00%	1,000	-,004	1
30	173	33	0,16%	,479	,005	13
40	40	8	0,04%	,533	,004	6
50	505	71	0,35%	,095	-,012	56
60	334	55	0,27%	,877	-,002	43
70	1.421	218	1,07%	,108	-,011	321
80	1.231	192	0,95%	,225	-,009	222
90	3	0	0,00%	1,000	-,005	1
100	12.292	2.148	10,58%	,008	,019	2.232
110	35	4	0,02%	,502	-,006	10
120	2	0	0,00%	1,000	-,004	1
130	1.983	321	1,58%	,377	-,006	488

Table 29: Single-vehicle accidents with single occupation and personal injury that occurred outside the built-up area between 2012 and 2019 broken down by speed limit. n=20.293 (3.431 are *severe casualties*).

Figure 33 shows the 95 % confidence intervals for the speed limits 50 km/h, 70 km/h, 80 km/h and 130 km/h. Since the speed limit 100 km/h shows a higher joint probability with *severe casualties*, we illustrate the respective confidence interval separately in Figure 35. The 95 % confidence represent the following probability ranges:

- speed limit 50km/h: 0,28 % to 0,43 %, standard error of 0,04 %
- speed limit 70km/h: 0,94 % to 1,22 %, standard error of 0,07 %
- speed limit 80km/h: 0,82 % to 1,08 %, standard error of 0,07 %
- speed limit 130km/h: 1,42 % to 1,74 %, standard error of 0,08 %

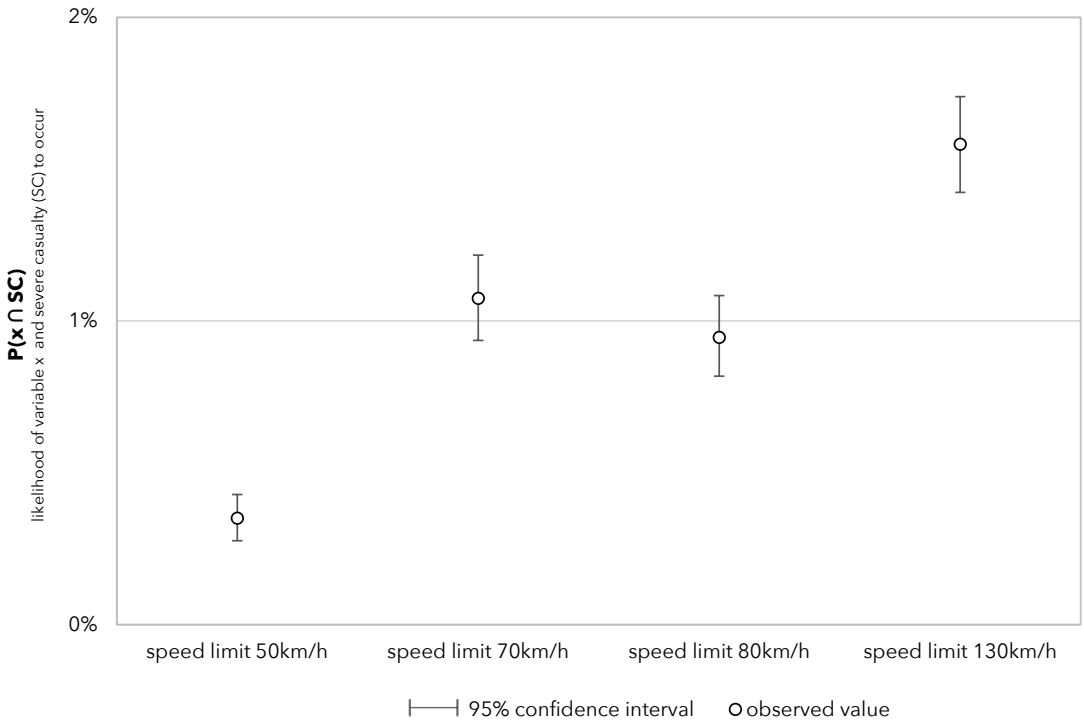

Figure 33: 95% confidence intervals for different speed limits (km/h). The confidence intervals estimate the likelihood of the characteristics and *severe casualties* to occur (range for joint probability). n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

Within the blackpattern recognition process, it is of interest to blend speed limits with road types. Table 30 shows the statistical key figures for casualties among different road types. 'Country roads' show the highest joint probabilities with *severe casualties* but no significant relationship with *severe casualties*. 'Expressways' and 'other roads' appear to correlate with *severe casualties* significantly.

Variable X **Road type**	C: Casualties n	SC: Severe Casualties n	P (X ∩ SC) %	Fisher's exact test [Y=severe casualty] *p*	Phi coefficient [Y=severe casualty] ϕ	Comb Max [roadway-related variables] n
Highway	2.593	417	2,05%	,239	-,008	488
Expressway	595	80	0,39%	,024	-,016	82
Country road	14.457	2.416	11,91%	,247	-,008	2.232
Other roads	2.220	463	2,28%	,000	,037	248

Table 30: Single-vehicle accidents with single occupation and personal injury that occurred outside the built-up area between 2012 and 2019 broken down by road type. n=20.293 (of which 3.431 are *severe casualties*).

Figure 34 illustrates the 95 % confidence intervals for the road types 'highway, 'expressway' and 'other road'. As the road type 'country road' is higher frequency, we illustrate it separately in figure 35. The probability ranges for the three road types are:

- highway: 1,88 % to 2,24 %, standard error of 0,09 %
- expressway: 0,31 % to 0,48 %, standard error of 0,04 %
- other road: 2,08 % to 2,47 %, standard error of 0,10 %

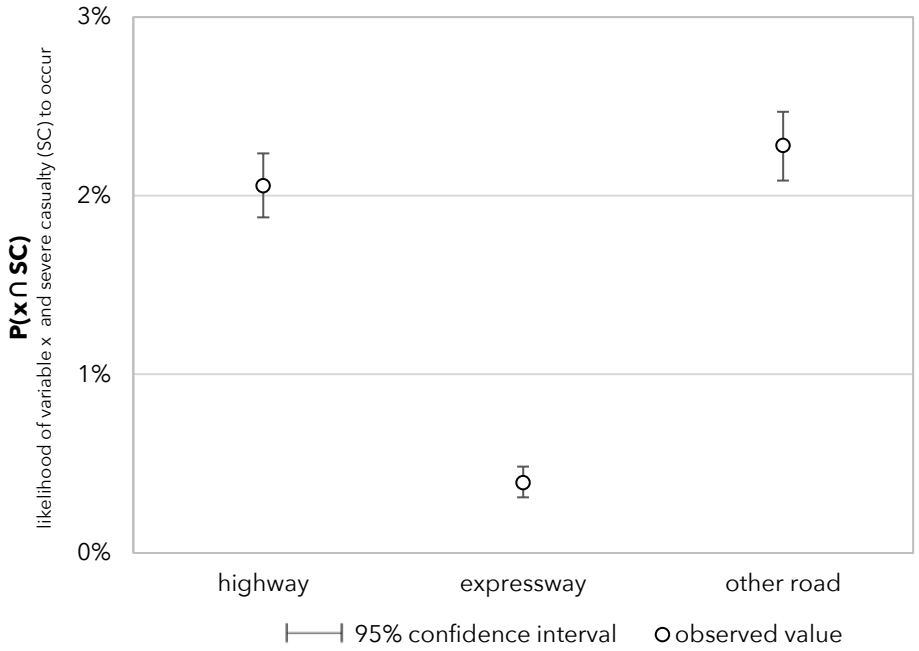

Figure 34: 95% confidence intervals for different road types. The confidence intervals estimate the likelihood of the characteristics and *severe casualties* to occur (range for joint probability). n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

In figure 35, we see the 95 % confidence intervals for the variables 'country road' and 'speed limit 100 km/h'. For 'country road', the 95 % confidence interval ranges from 11,65 % to 12,16 %, and the standard error of the distribution is 0,13 %. For 'speed limit 100 km/h', the 95 % confidence interval ranges from 10,31 % to 10,85 %, and the standard error of the distribution is 0,14 %.

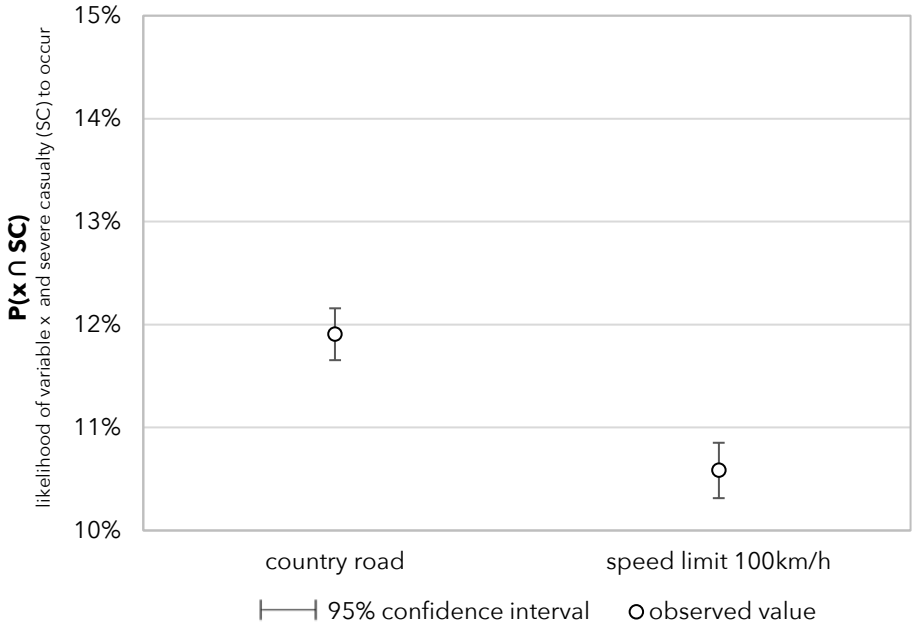

Figure 35: 95% confidence intervals for 'country road' and 'speed limit 100km/h'. The confidence intervals estimate the likelihood of the characteristics and *severe casualties* to occur (range for joint probability). n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

The analysis for roadway-related variables continues with the investigation of road characteristics (see table 31). The road characteristics 'straight road', 'curve', 'bridge', 'tunnel', 'gallery', 'middle separation' and 'entry or exit' show a significant relationship with *severe casualties*.

Variable X **Road characteristics**	C: Casualties n	SC: Severe Casualties n	P (X ∩ SC) %	Fisher's exact test [Y=severe casualty] *p*	Phi coefficient [Y=severe casualty] ϕ	Comb Max [roadway-related variables] n
Intersection	439	62	0,31%	,125	-,011	62
Roundabout	68	16	0,08%	,146	,010	11
Deceleration lane	10	2	0,01%	,681	,002	1
Acceleration lane	3	1	0,00%	,426	,005	1
One-way	144	33	0,16%	,054	,014	26
Construction site	157	21	0,10%	,286	-,008	10
Cycle path	4	0	0,00%	1,000	-,006	1
Crosswalk	3	0	0,00%	1,000	-,006	1
Pedestrian and cycle path	10	2	0,01%	,681	,002	3
Parking lane	7	0	0,00%	,610	-,008	1
Secondary lane	5	1	0,00%	1,000	,001	1
Hard shoulder	45	9	0,04%	,551	,004	7
Banquet	123	22	0,11%	,729	,002	22
Straight road	11.507	2.095	10,32%	,000	,040	2.232
Tunnel	89	26	0,13%	,004	,022	8
Gallery	15	8	0,04%	,001	,026	1
Rest area	26	6	0,03%	,429	,006	2
Traffic island	81	18	0,09%	,233	,009	4
Underpass	32	7	0,03%	,476	,005	3
Middle separation	777	104	0,51%	,008	-,019	137
Bridge	157	41	0,20%	,003	,022	7
Curve	8.399	1.264	6,23%	,000	-,042	1.437

Table 31: Single-vehicle accidents with single occupation and personal injury that occurred outside the built-up area between 2012 and 2019 broken down by road characteristics. n=20.293 (3.431 are *severe casualties*).

Variable X **Road characteristics**	C: Casualties n	SC: Severe Casualties n	P (X ∩ SC) %	Fisher's exact test [Y=severe casualty] *p*	Phi coefficient [Y=severe casualty] ϕ	Comb Max [roadway-related variables] n
Narrow lane	30	8	0,04%	,149	,010	3
Entry or exit	57	17	0,08%	,019	,018	5
Tram or bus station	8	2	0,01%	,630	,004	1

Continuation of table 31: Single-vehicle accidents with single occupation and personal injury that occurred outside the built-up area between 2012 and 2019 broken down by road characteristics. n=20.293 (3.431 are *severe casualties*).

Road conditions describe the condition of the road surface when the accident occurs. Except for 'other conditions', all road surface conditions result in a maximum combination value above 50 (see table 32). 'Dry road', 'wet road', and 'wintry conditions' appear to correlate with *severe casualties* significantly. The characteristic 'sand or grit on the road' does not significantly correlate with *severe casualties*.

Variable X **Road condition**	C: Casualties n	SC: Severe Casualties n	P (X ∩ SC) %	Fisher's exact test [Y=severe casualty] *p*	Phi coefficient [Y=severe casualty] ϕ	Comb Max [roadway-related variables] n
Dry road	10.441	2.126	10,48%	,000	,095	2.232
Wet road	5.705	872	4,30%	,000	-0,27	1.225
Sand or grit on the road	297	48	0,24%	,809	-,002	56
Wintry conditions	3.771	370	1,82%	,000	-,090	938
Other conditions (oil, soil)	95	17	0,08%	,796	,002	16

Table 32: Single-vehicle accidents with single occupation and personal injury that occurred outside the built-up area between 2012 and 2019 broken down by road condition. n=20.293 (3.431 are *severe casualties*).

Figure 37 shows the 95 % confidence intervals for the characteristics 'curve', 'wet road' and 'wintry conditions'. The confidence intervals for these characteristics result in the following probability ranges:

- curve: 5,96 % to 6,50 %, standard error of 0,14 %
- wet road: 4,06 % to 4,55 %, standard error of 0,12 %
- wintry conditions: 1,65 % to 2,00 %, standard error of 0,09 %

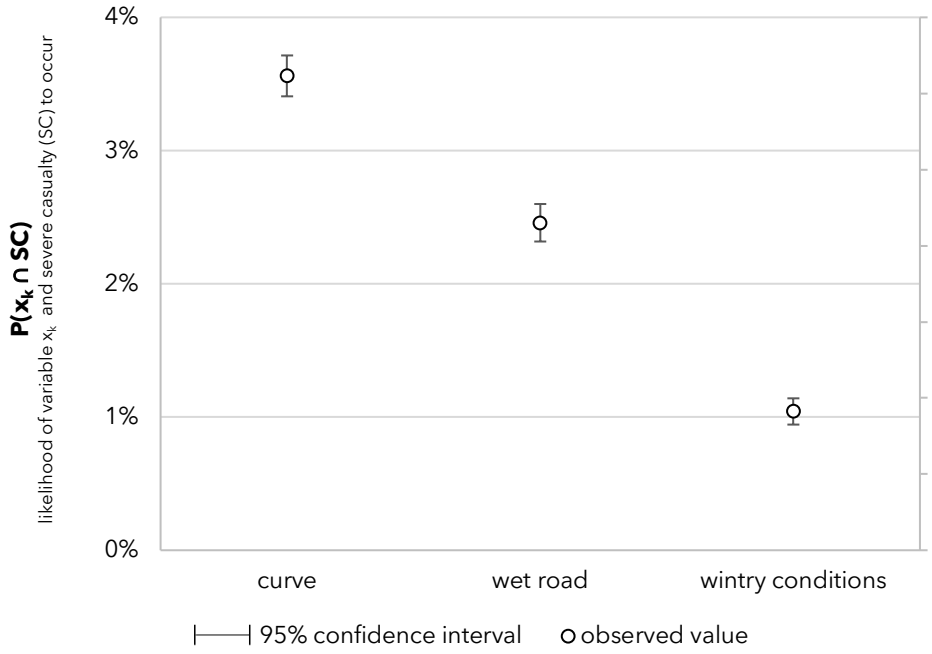

Figure 36: 95% confidence intervals for different road characteristics. The confidence intervals estimate the likelihood of the characteristics and *severe casualties* to occur (range for joint probability). n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

Figure 37 shows the 95 % confidence intervals for the characteristics 'intersection', 'middle separation' and 'sand or grit on the road'. The confidence intervals for these characteristics result in the following probability ranges:

- intersection: 0,24 % to 0,38 %, standard error of 0,04 %
- middle separation: 0,42 % to 0,61 %, standard error of 0,05 %
- sand or grit on the road: 0,18 % to 0,30 %, standard error of 0,03 %

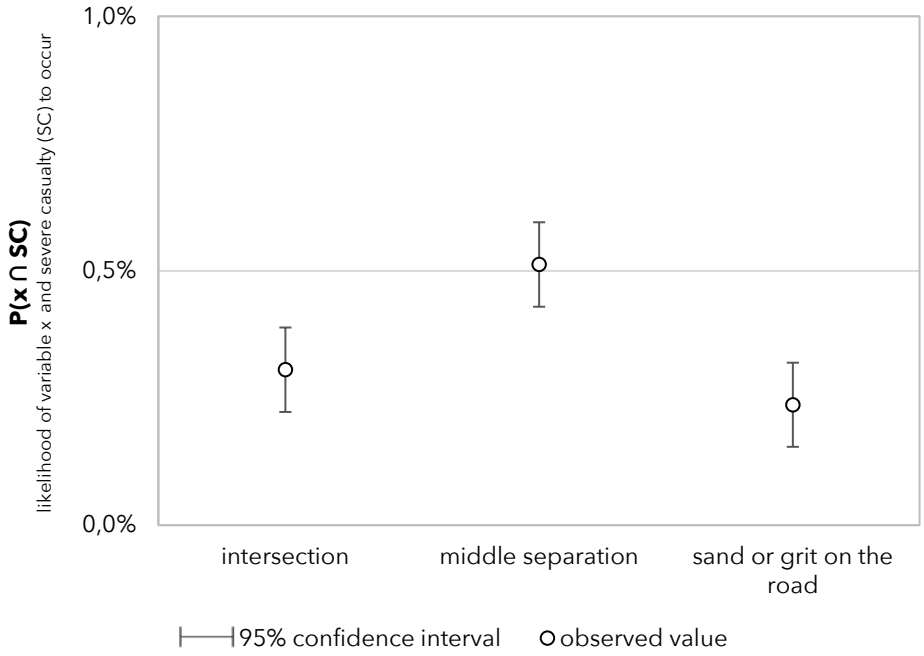

Figure 37: 95% confidence intervals for different road characteristics. The confidence intervals estimate the likelihood of the characteristics and *severe casualties* to occur (range for joint probability). n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

The analysis of roadway-related characteristics concludes with the analysis of traffic lights. The characteristic 'traffic light in full operation' does not significantly affect our target variable *severe casualties*.

Variable X **Traffic lights**	C: Casualties n	SC: Severe Casualties n	P (X ∩ SC) %	Fisher's exact test [Y=severe casualty] *p*	Phi coefficient [Y=severe casualty] ϕ	Comb Max [roadway-related variables] n
Traffic light in full operation	29	2	0,01%	,213	-,010	4

Table 33: Single-vehicle accidents with single occupation and personal injury that occurred outside the built-up area between 2012 and 2019 broken down by traffic lights. n=20.293 (3.431 are *severe casualties*).

Figure 38 illustrates the relative frequency of selected characteristics among *severe casualties* and casualties with a slight injury. The characteristics 'other road', 'speed limit 100 km/h', 'dry road' and 'straight road' hold a higher share among *severe casualties*.

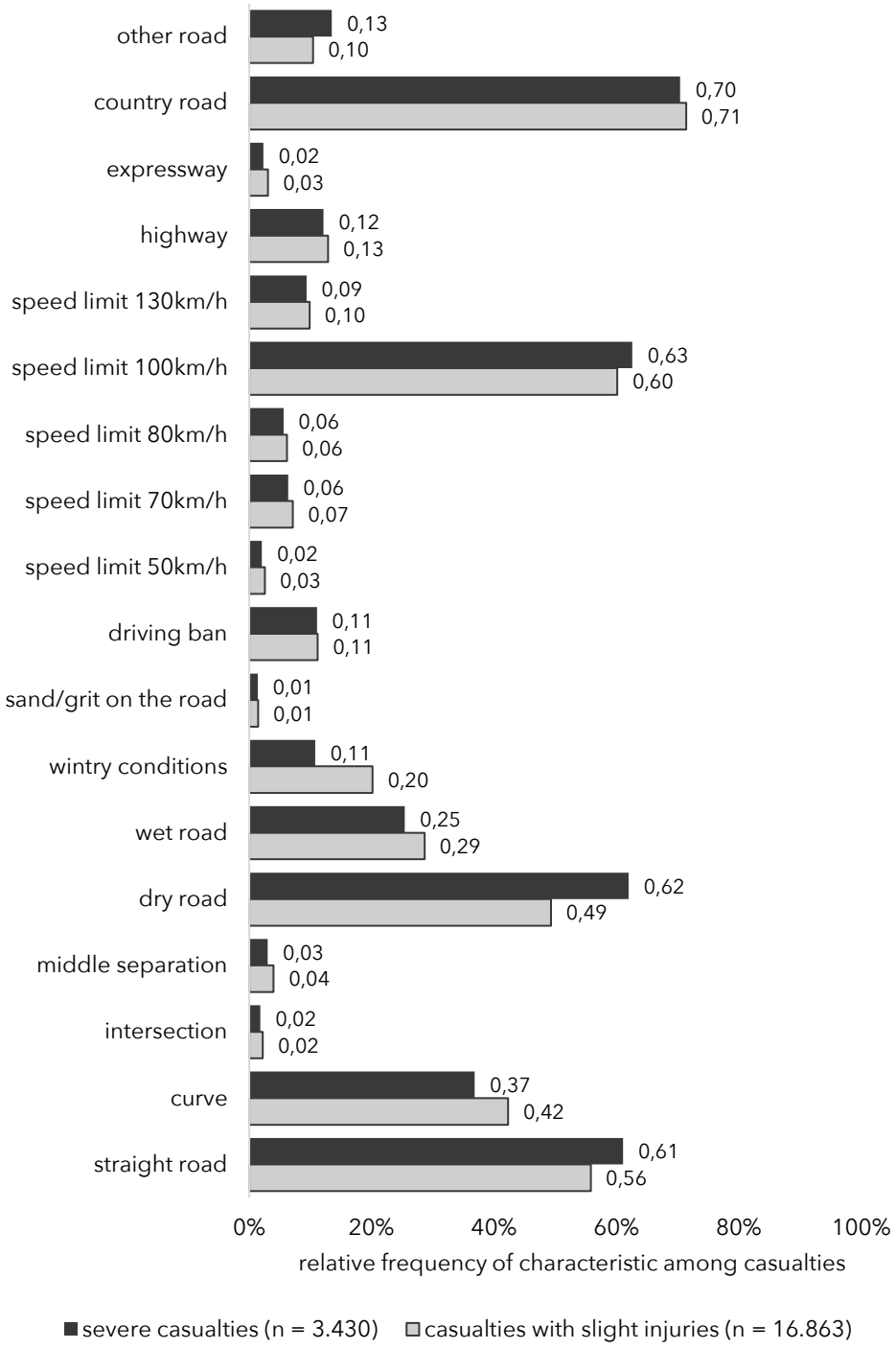

Figure 38: Relative frequency (or conditional probabilities) of roadway-related characteristics among casualties with slight injuries and severe casualties. n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network between 2012-2019 (3.431 are *severe casualties*).

4.9 Analysis of situation-related variables

Situation-related variables include the accident's time and weekday, the meteorological season, weather, and light conditions. Table 34 illustrates all situation-related variables and their characteristics.

Variable	*Characteristic*
Time	12 to 6, 6 to 12, 12 to 6, 6 to 12
Weekday	Monday to Thursday, Friday to Sunday
Meteorological Season	spring, summer, autumn, winter
Weather conditions	clear or overcast weather, rain, hail, freezing rain, snow, fog, high wind
Light conditions	daylight, dusk or dawn, darkness, artificial light, restricted view by vehicle, glare from the sun

Table 34: Situation-related variables.

The analysis of situation-related variables foresees the calculation of variable frequencies, conditional and joint probabilities, Fisher's exact test, the Phi coefficient, and the maximum combination value for each variable.

The analysis of situation-related variables starts with time investigation (see table 35). We can see that most accidents occur between 6 a.m. and 12 p.m. The minor accidents occur between 12 a.m. and 6 a.m., but the relative frequency of *severe casualties* within this timeframe is almost as high as within the other timeframes. The timeframe 12 p.m. to 6 p.m. shows the highest probability of a severe or fatal accident. At the same time, it is the only time category not showing a significant relationship with *severe casualties*.

Variable X **Time (h)**	C: Casualties n	SC: Severe Casualties n	P (X ∩ SC) %	Fisher's exact test [Y=severe casualty] *p*	Phi coefficient [Y=severe casualty] ϕ	Comb Max [situation-related variables] n
12 a.m. to 6 a.m.	3.367	713	3,51%	,000	,051	245
6 a.m. to 12 p.m.	6.283	889	4,38%	,000	-,049	586
12 p.m. to 6 p.m.	5.915	956	4,71%	,070	-,013	578
6 p.m. to 12 a.m.	4.728	873	4,30%	,001	,023	368

Table 35: Single-vehicle accidents with single occupation and personal injury that occurred outside the built-up area between 2012 and 2019 broken down by the time of the accident. n=20.293 (3.431 are *severe casualties*).

Figure 39 shows the 95 % confidence intervals for the time categories. The probability ranges for these categories comprise the following values:

- 0 a.m. to 6 a.m.: 3,28 % to 3,74 %, standard error of 0,00 %
- 6 a.m. to 12 p.m.: 4,13 % to 4,62 %, standard error of 0,13 %
- 12 p.m. to 6 p.m.: 4,47 % to 4,96 %, standard error 0,13 %
- 6 p.m. to 12 a.m.: 4,06 % to 4,55, standard error of 0,13 %

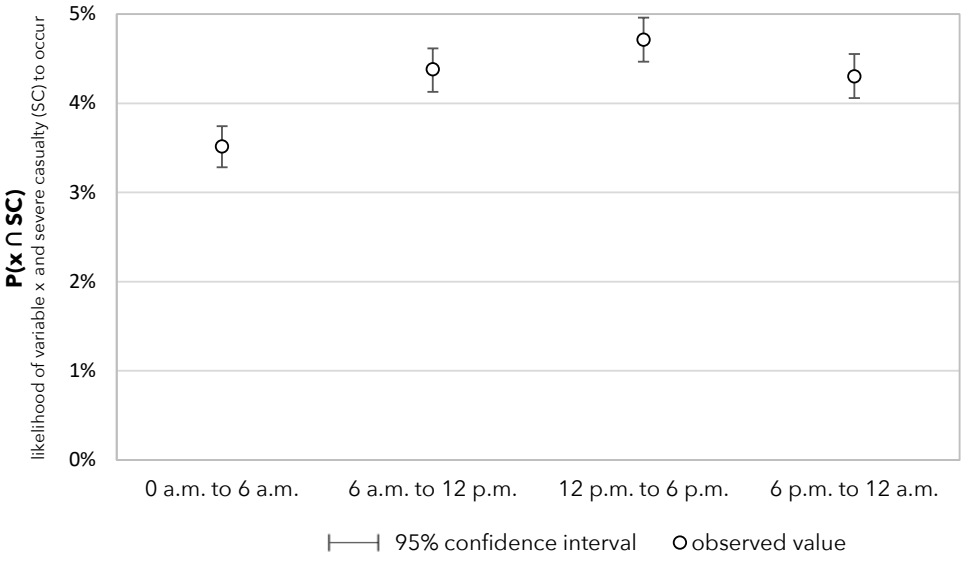

Figure 39: 95% confidence intervals for the time of the accident. The confidence intervals estimate the likelihood of the characteristics and *severe casualties* to occur (range for joint probability). n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

Table 36 summarises weekdays into two categories: Monday to Thursday and Friday to Sunday. A severe or fatal accident's probability is almost equal in both categories. Also, both categories show a significant relationship with *severe casualties*.

Variable X **Weekday**	C: Casualties n	SC: Severe Casualties n	P (X ∩ SC) %	Fisher's exact test [Y=severe casualty] *p*	Phi coefficient [Y=severe casualty] ϕ	Comb Max [situation-related variables] n
Mon to Thu	11.131	1.788	8,81%	,000	-,025	586
Fri to Sun	9.162	1.643	8,10%	,000	,025	430

Table 36: Single-vehicle accidents with single occupation and personal injury that occurred outside the built-up area between 2012 and 2019 broken down by weekday. n=20.293 (3.431 are *severe casualties*).

Figure 40 shows the 95 % confidence intervals for weekdays. The likelihood of a severe or fatal accident between Monday and Thursday ranges from 8,53 % to 9,09 %, with a standard error of 0,14 %. The likelihood of a severe or fatal accident between Friday and Sunday ranges from 7,81 % to 8,37 %, with a standard error of 0,14 %.

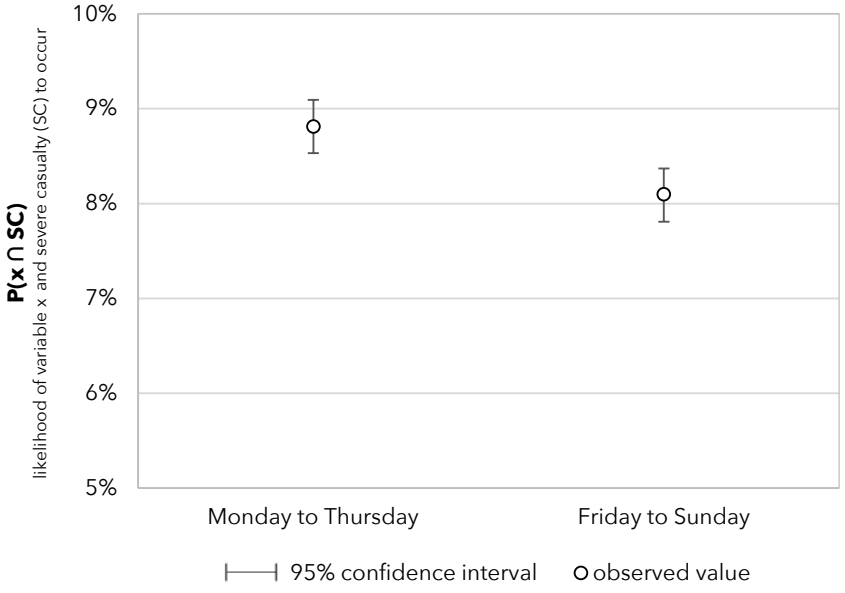

Figure 40: 95% confidence intervals for weekdays. The confidence intervals estimate the likelihood of the characteristics and *severe casualties* to occur (range for joint probability). n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

We proceed with the investigation of meteorological seasons. Even if most accidents occur in winter, a severe or fatal accident is more likely to occur in summer within the investigated sample (see table 37). Furthermore, summer and winter show a significant relationship with *severe casualties*.

Variable X **Meteorological Season**	C: Casualties n	SC: Severe Casualties n	P (X ∩ SC) %	Fisher's exact test [Y=severe casualty] *p*	Phi coefficient [Y=severe casualty] ϕ	Comb Max [situation-related variables] n
Spring	4.279	774	3,81%	,021	,016	435
Summer	4.821	896	4,42%	,000	,025	578
Autumn	4.802	885	4,36%	,001	,023	394
Winter	6.391	876	4,32%	,000	-0,58	586

Table 37: Single-vehicle accidents with single occupation and personal injury that occurred outside the built-up area between 2012 and 2019 broken down by meteorological season. n=20.293 (3.431 are *severe casualties*).

Figure 41 illustrates the likelihood of a severe or fatal accident to occur within a meteorological season. The 95 % confidence intervals for meteorological seasons show the following values:

- winter: 4,06 % to 4,56 %, standard error of 0,13 %
- spring: 3,58 % to 4,05 %, standard error of 0,12 %
- summer: 4,17 % to 4,68 %, standard error of 0,13 %
- autumn: 4,10 % to 4,62 %, standard error of 0,13 %

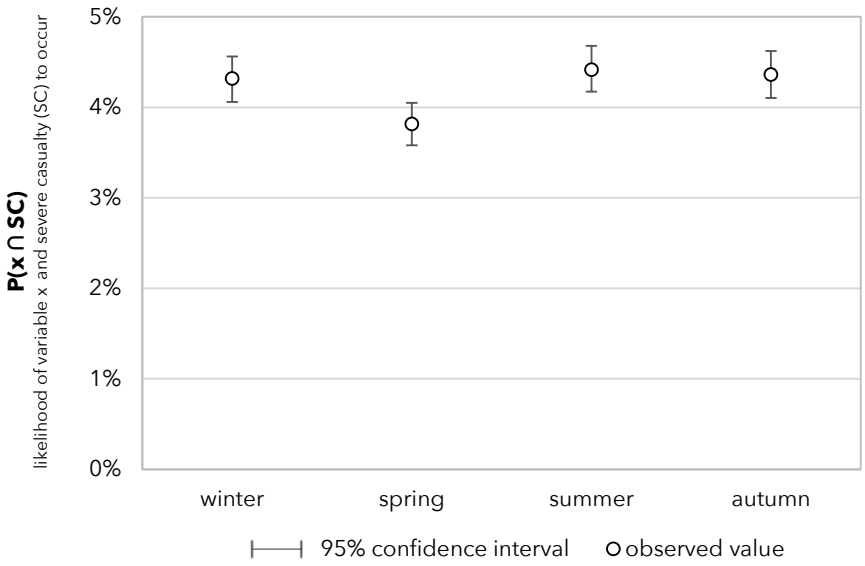

Figure 41: 95% confidence intervals for meteorological seasons. The confidence intervals estimate the likelihood of the characteristics and *severe casualties* to occur (range for joint probability). n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

In table 38, the weather analysis shows that 'clear or overcast weather' and 'snow' result in a significant relationship with *severe casualties*. Most severe or fatal accidents occur during 'clear or overcast weather', followed by 'rain' and 'snow'.

Variable X **Weather conditions**	C: Casualties n	SC: Severe Casualties n	P (X ∩ SC) %	Fisher's exact test [Y=severe casualty] *p*	Phi coefficient [Y=severe casualty] ϕ	Comb Max [situation-related variables] n
Clear or overcast weather	15.541	2.797	13,78%	,000	,053	586
Rain	3.013	458	2,26%	,007	-,019	110
Hail, freezing rain	124	17	0,08%	,398	-,007	12
Snow	1.913	175	0,86%	,000	-,067	147
Fog	636	102	0,50%	,588	-,004	37
High wind	377	52	0,26%	,113	-,011	17

Table 38: Single-vehicle accidents with single occupation and personal injury that occurred outside the built-up area between 2012 and 2019 broken down by weather conditions. n=20.293 (3.431 are *severe casualties*).

Figure 42 shows the 95 % confidence intervals for weather conditions. The likelihood of weather characteristics and severe or fatal accidents to occur comprise the following probability ranges:

- fog: 0,40 % to 0,60 %, standard error of 0,05 %
- high wind: 0,19 % to 0,33 %, standard error of 0,03 %
- hail: 0,04 % to 0,12 %, standard error of 0,02 %
- rain: 2,06 % to 2,45 %, standard error of 0,10 %
- snow: 0,74 % to 0,98 %, standard error of 0,06 %

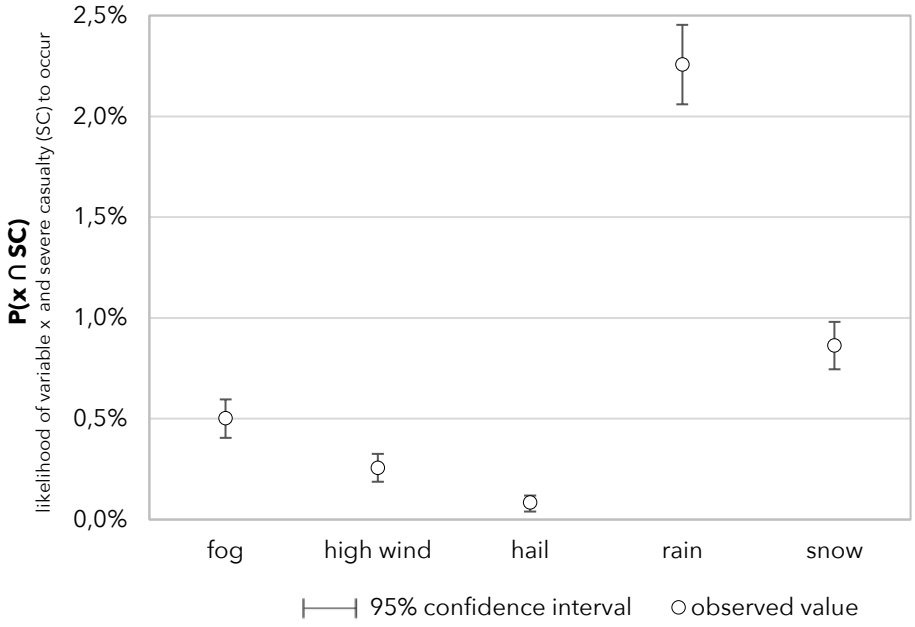

Figure 42: 95% confidence intervals for weather conditions. The confidence intervals estimate the likelihood of the characteristics and *severe casualties* to occur (range for joint probability). n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

As shown in table 39, the analysis of light conditions reveals a significant relationship between daylight and *severe casualties* and darkness and *severe casualties*.

Variable X **Light conditions**	C: Casualties n	SC: Severe Casualties n	P (X ∩ SC) %	Fisher's exact test [Y=severe casualty] *p*	Phi coefficient [Y=severe casualty] ϕ	Comb Max [situation-related variables] n
Daylight	11.546	1.790	8,82%	,000	-,043	586
Dusk or dawn	1.604	266	1,31%	,753	-,003	111
Darkness	6.828	1.311	6,46%	,000	,044	368
Artificial light	571	93	0,46%	,730	-,003	15
Restricted view by another vehicle	7	0	0,00%	,610	-,008	1
Glare from the sun	109	24	0,12%	,156	,010	8

Table 39: Single-vehicle accidents with single occupation and personal injury that occurred outside the built-up area between 2012 and 2019 broken down by light conditions. n=20.293 (3.431 are *severe casualties*).

Figure 43 illustrates the 95 % confidence intervals for the characteristics 'daylight', 'darkness' and 'dusk or dawn'. The probability ranges for the three characteristics comprise the following values:

- daylight: 8,53 % to 9,10 %, standard error of 0,14 %
- dusk or dawn: 1,16 % to 1,47 %, standard error of 0,08 %
- darkness: 6,20 % to 6,74 %, standard error of 0,14 %

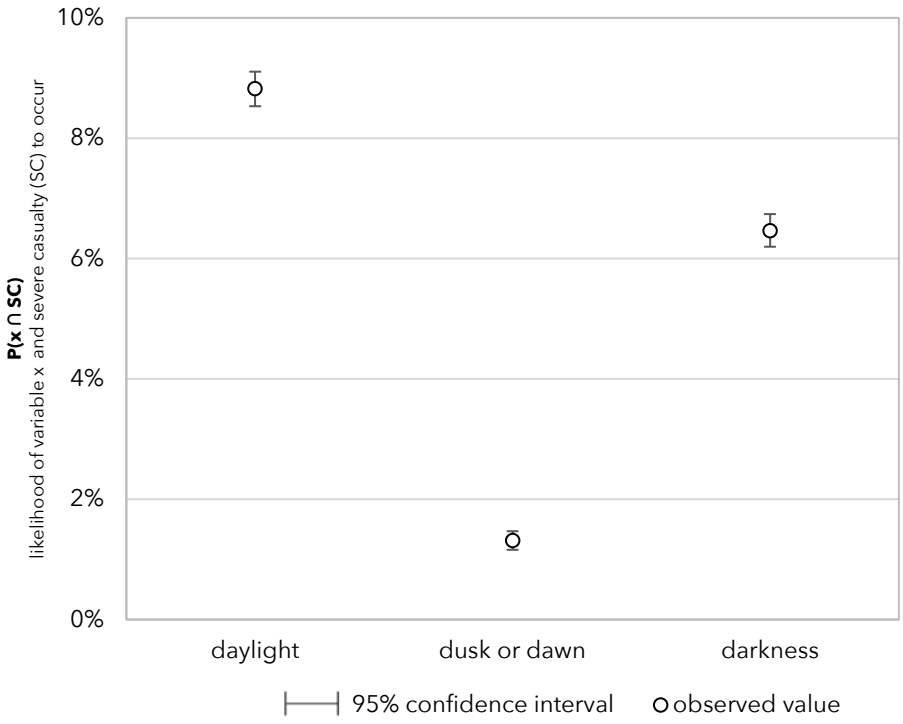

Figure 43: 95% confidence intervals for light conditions. The confidence intervals estimate the likelihood of the characteristics and *severe casualties* to occur (range for joint probability). n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

Figure 44 illustrates the share of selected situation-related variables among all casualties and *severe casualties*. *Severe casualties* hold a higher share for the variables 'clear and overcast weather', 'darkness', 'Friday to Sunday', '6 p.m. to 12 a.m.', and '0 a.m. to 6 a.m.'.

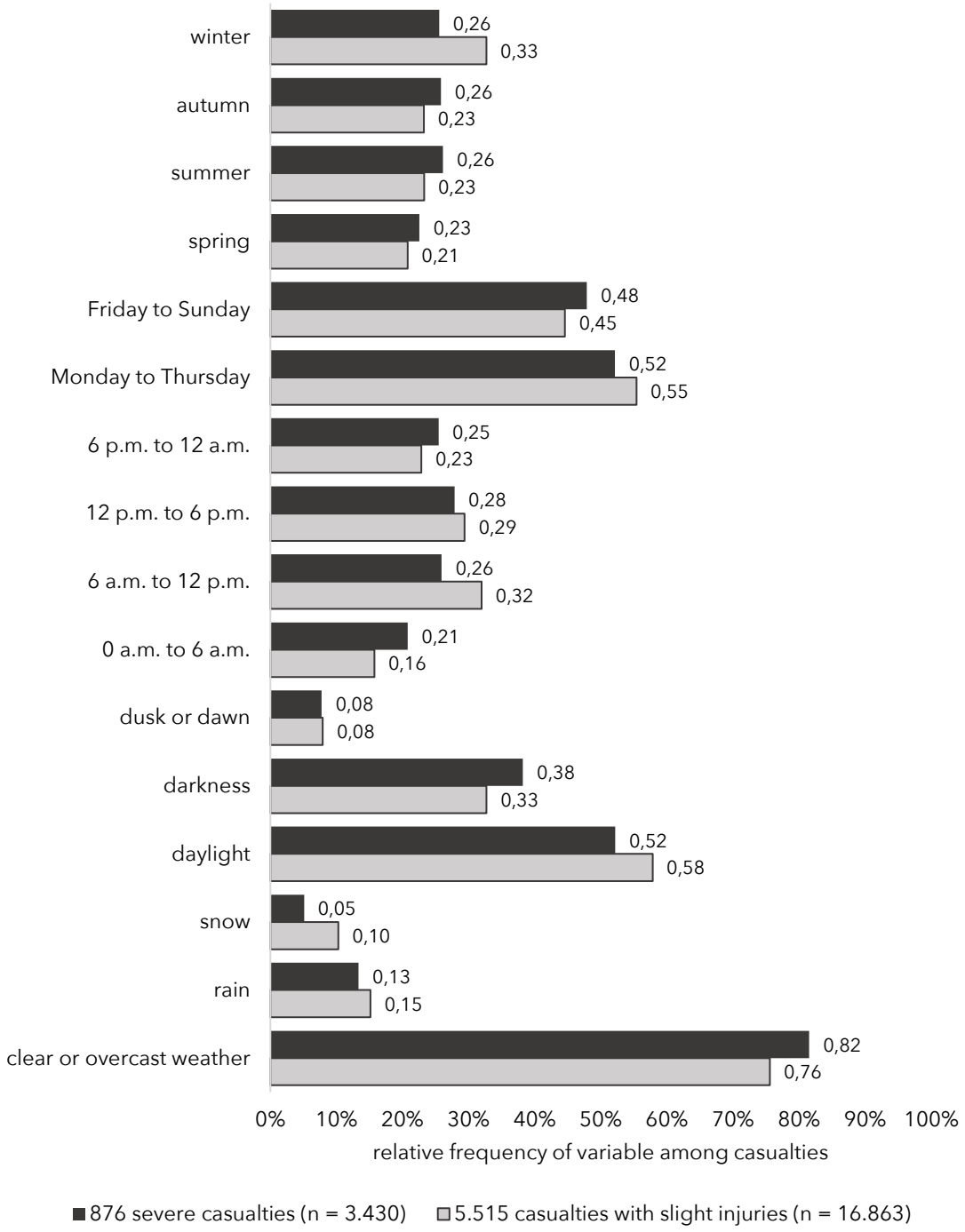

Figure 44: Relative frequency (or conditional probabilities) of situation-related variables among casualties with slight injuries and *severe casualties*. n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

5. Road traffic accident data analysis II: Logistic Regression

Because the target variable *severe casualties* is a dichotomous or binary variable, we apply binary logistic regression (also known as binomial logistic regression) to quantify the relationship and the impact of accident-related variables on *severe casualties*. For this study, the benefit of binary logistic regression is twofold:

- it helps us to exclude accident-related characteristics having no significant relationship with our target variable *severe casualties;*
- it helps us estimate the impact of an accident-related characteristic on *severe casualties* compared to all investigated accident characteristics.

For the subsequent pattern recognition procedures, this is essential information to identify blackpatterns that

- exclusively include accident-related characteristics having a significant relationship with our target variable *severe casualties;*
- can be assessed because of knowing the impact of each included accident-related characteristic on the target variable *severe casualties*.

In this case, we do not apply binary logistic regression to create a prediction model but to retrieve information that will help us to evaluate our detected blackpatterns (see chapter 8). This way, we can determine whether there exist blackpatterns, including accident-related variables with a relatively high impact on *severe casualties*.

First, we apply binary logistic regression separately in our four categories (driver, vehicle, roadway and situation). Second, we apply binary logistic regression with all accident-related characteristics (i.e., across all four categories). We will then compare and discuss the output of both approaches.

5.1 Generation of logistic regression models

Before generating our regression models, we zoom into some core terms in logistic regression and substantiate why we choose specific model settings. We will now explain the meaning of

- odds,
- logit,
- odds ratios
- and log odds ratios.

To illustrate the meaning of these measures, we will use the following example of our binary road traffic accident dataset. Table 40 shows the distribution of *severe casualties* (y, target variable) among drivers who did not apply a safety belt (x, independent variable).

	no safety belt applied (x=1)	safety belt applied (x=0)	*total*
severe casualty (y=1)	699 p_1=33 %	2.731 p_2=15 %	3.430 17 %
no severe casualty (y=0)	1.401 1-p_1=66 %	15.462 1-p_2=85 %	16.863 83 %
total	2.100 10 %	18.193 90 %	20.293 100 %

Table 40: 2x2 field table showing the dummy variables "severe casualty" (target variable) and "no safety belt applied" (independent variable). n=20.293 single-vehicle accidents with a single occupation that occurred on the Austrian road network outside the built-up area between 2012 and 2019.

The probability of a *severe casualty* involving a driver with 'no safety belt applied' is 33 %. To estimate the odds of a severe or fatal road traffic accident involving a driver with 'no safety belt applied', we calculate the ratio of the two probabilities (p_1/1-p_1). Thus, the odds (also referred to as chance) of a severe or fatal accident involving a driver with 'no safety belt applied' are 50 %. In comparison, the odds of a severe or fatal accident involving a driver with an applied safety belt are 18 %. A logit is the natural algorithm of a chance (p_1/1-p_1). The logit to be involved in a severe or fatal road traffic accident with 'no safety belt applied' is -0,69, which we retrieve by applying the following formula:

$$logit = ln\left(\frac{p}{1-p}\right) = ln\,(odds(p))$$

Thus, we can also describe odds as e^{logit} with e referring to Euler's number (2,71828). To better understand the behaviour between probability, odds and the logit function, we illustrate p, 1-p, odds and logit for the value range of p in table 41.

p	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90
1-p	0,90	0,80	0,70	0,60	0,50	0,40	0,30	0,20	0,10
Odds	0,11	0,25	0,43	0,67	1,00	1,50	2,33	4,00	9,00
Logit	-2,20	-1,39	-0,85	-0,41	0,00	0,41	0,85	1,39	2,20

Table 41: Value range p with corresponding p-1, odds and logit.

The table allows us to draw the following conclusions:

- Logit is symmetric around 0 (p=0,50).
- The more extreme the probability p deviates from 0,50, the more the logit changes.
- For large logits, p approaches 0 and 1, respectively, but without reaching these values
- Therefore, even for very large logits, the probabilities p are always in the bounds of 0 and 1.

Table 42 now shows probability, odds and logit for our example.

	p		*odds*		*logit*	
	no safety belt applied x=1	safety belt applied x=0	no safety belt applied x=1	safety belt applied x=0	no safety belt applied x=1	safety belt applied x=0
severe casualty y=1	0,33 p_1	0,15 p_2	0,50 $p_1/(1-p_1)$	0,18 $p_2/(1-p_2)$	-0,69 $\ln(p_1/(1-p_1))$	-1,71 $\ln(p_2/(1-p_2))$
no severe casualty y=0	0,66 $1-p_1$	0,85 $1-p_2$	2,00 $(1-p_1)/p_1$	5,67 $(1-p_2)/p_2$	0,69 $\ln((1-p_1)/p_1)$	-0,37 $\ln((1-p_1)/p_1)$

Table 42: Probability, odds and logit.

Odds ratio refers to the quotient of two odds ((p_1/1-p_1)/ (p_2/1-p_2)). Thus, the odds ratio of a severe or fatal road traffic accident involving a driver with 'no safety belt applied' is 2,78 when considering our reference group as drivers with 'no safety belt applied'. Log odds ratio or ln(odds ratio) refers to the natural algorithm of the odds ratio. Table 43 illustrates the odds ratio and log odds ratio four our example.

	odds ratio		*log odds ratio*	
	no safety belt applied (x=1)	safety belt applied (x=0)	no safety belt applied (x=1)	safety belt applied (x=0)
severe casualty (y=1)	2,78 $p_1/(1-p_1) / p_2/(1-p_2)$	0,36 $p_2/(1-p_2) / p_1/(1-p_1)$	1,02 $\ln(p_1/(1-p_1) / p_2/(1-p_2)$	-1,02 $\ln(p_2/(1-p_2) / p_1/(1-p_1))$
no severe casualty (y=0)	0,35 $(1-p_1)/p_1 / (1-p_2)/p_2$	2,84 $(1-p_2)/p_2 / (1-p_1)/p_1$	-1,05 $\ln((1-p_1)/p_1 / (1-p_2)/p_2)$	1,04 $\ln((1-p_2)/p_2 / (1-p_1)/p_1)$

Table 43: Odds ratio and log odds ratio.

Logistic regression

To apply binomial logistic model regression (also known as binary logistic regression), we use

$$\ln \frac{\pi(x)}{1-\pi(x)} = \beta_0 + \beta_1 \times x_1 + \ldots + \beta_k \times x_k$$

where:

$\ln \frac{\pi(x)}{1-\pi(x)}$ … target variable (logit)

$x_1 \ldots x_k$ … independent variables x_1 … x_k

$\beta_0 \ldots \beta_k$ … regression coefficients

In our example, we illustrate the formula for binomial logistic regression (see table 44).

	no safety belt applied x=1	safety belt applied x=0
severe casualty y=1	$\pi(1) = \frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}}$	$\pi(0) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$
no severe casualty y= 0	$1 - \pi(1) = \frac{1}{1 + e^{\beta_0 + \beta_1}}$	$1 - \pi(0) = \frac{1}{1 + e^{\beta_0}}$

Table 44: Binomial logistic regression with a binary independent variable x and a binary target variable y.

Thus, we can describe the logistic distribution as

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 \times x_1 + \beta_k \times x_k}}{1 + e^{\beta_0 + \beta_1 \times x_1 + \beta_k \times x_k}}$$

where:

π(x) = p(y = 1) … probability for y = 1:

e … Euler's number (basis of the natural algorithm)

$x_1 \ldots x_k$ … independent variables x_1 … x_k (predictor variables)

$\beta_0 \ldots \beta_k$ … regression coefficients

Dependent variable (y, target variable)

We use binary logistic regression to describe the relationship and impact of accident-related variables on our target variable *severe casualties*. The target variable is defined as follows:

- y=1: probability to observe a *severe casualty* among single-vehicle accidents with single occupation and personal injury on the Austrian road network outside the built-up area
- y=0: probability to observe *no severe casualty* among single-vehicle accidents with single occupation and personal injury on the Austrian road network outside the built-up area

Independent variables (x_1,…, x_k predictor variables)

Our model exclusively works with dummy variables representing detailed information on accident-related characteristics. In total, we work with 158 accident-describing characteristics. The dummy variables are defined as follows:

- x=1: characteristic is present
- x=0: characteristic is not present

Regression coefficient β

The regression coefficient β is the logarithm of the odds ratio, e^{β} represents the odds ratio. The maximum likelihood method defines the regression coefficient. We will not discuss maximum likelihood in detail in this thesis.

Exp(β)

Exp(β) represents the odds ratio for a one-unit increase in x_k. Thus, if we get an exp(β)-value of 6,012, for example, it indicates that a one-unit change in the variable corresponding to x_k will multiply the relative risk of *severe casualties* (compared to the base outcome) by 6.012.

Stepwise variable selection with Likelihood Ratio

We estimate our logistic regression model using the stepwise variable selection method *forward selection (Likelihood Ratio)*. Stepwise regression appears helpful for heuristic research approaches (i.e., we do not have a specific hypothesis which accident-related variable affects our target variable *severe casualties*).

The stepwise variable selection process focuses on finding the best model/equation when working with many variables (in our case, 158 variables in total). These variables will invariably have patterns of overlap of information about Y, our dependent variable *severe casualties*, which are difficult to see and understand. Stepwise regression results in a model/equation consisting of significant variables only. Also, the model ensures not to miss a significant variable.

Regression model evaluation

In logistic regression, an equivalent to R^2 that we know from linear regression does not exist. There exist various pseudo R^2 as measures of fit. Their interpretation requires caution as pseudo R^2 in logistic regression do not mean the same as R^2 in linear regression (where we estimate the model with a least-squares estimator). Hosmer and Lemeshow recommend not to use pseudo R^2 to evaluate logistic regression models. They argue that in the case of logistic regression, the measure of fit should strictly focus on comparing observed and predicted values from the fitted model. We take up upon their recommendation and use Hosmer-Lemeshow goodness-of-fit test. The Hosmer-Lemeshow test divides the sample into subgroups and checks the differences between observed and expected values—the smaller the difference, the better the model fit. Therefore, we are looking for a confirmation of the H_0. The H_0 states that the observed and expected proportions are the same across all samples. Therefore, we require a non-significant result for the Hosmer-Lemeshow test.

5.2 Logistic regression with driver-related variables

This chapter analyses the relationship and impact of driver-related characteristics on our dichotomous target variable *severe casualties* (0=no severe casualty, 1=severe casualty). Driver-related predictor variables comprise

- sex,
- age classes,
- impairments,
- type of driving licence,
- manoeuvres before the accident, and
- safety settings

with 56 characteristics (see chapter 4.6). We integrate each characteristic as dummy variables into our model (0: characteristic is not present, 1: characteristic is present). The regression model performs a 15-step variable selection process. Table 45 illustrates the model results (=step 15). The Hosmer-Lemeshow test results in p=0,360 and indicates no evidence of poor model fit.

The estimated model shows that the characteristic 'no safety belt applied' appears to have the highest impact on *severe casualties* (*exp(β)=5,14)*, followed by 'hitting an obstacle on the road', 'sudden braking', 'male drivers', and 'skidding/drifting'. The risk to observe a severe or fatal accident decreases with 'age class 65+' or 'when hitting the guard rail'.

Variable	*regression coefficient* β	*standard error* *SEM*	*Sig.* *p*	*exp(β)*
sudden braking	0,670	0,320	0,037	1,953
skidding/drifting	0,164	0,076	0,031	1,178
hitting a tree	0,510	0,073	0,000	1,666
hitting an obstacle on the road	1,180	0,423	0,005	3,254
hitting the guard rail	-0,306	0,086	0,000	0,736
age class 16 to 18	0,714	0,104	0,000	2,043
age class 19 to 24	0,636	0,060	0,000	1,889
age class 25 to 34	0,415	0,060	0,000	1,514
age class 35 to 44	0,247	0,068	0,000	1,280
age class 65+	-0,128	0,069	0,066	0,880
male driver	0,648	0,043	0,000	1,912
probationary driving licence	0,163	0,076	0,033	1,177
alcohol	0,410	0,059	0,000	1,507
no safety belt applied	1,638	0,060	0,000	5,143
hit and run	-0,431	0,158	0,006	0,650
constant	*-6,221*	*0,576*	*0,000*	*0,002*

Table 45: Driver-related logistic regression model. Input data: 20.293 single-vehicle accidents with single occupation occurring outside the built-up area on the Austrian road network between 2012-2019. The dataset includes 56 driver-related characteristics as dummy variables (0=characteristic is not present, 1=characteristic is present). The binary target variable is *severe casualties* (0=no severe casualty, 1=severe casualty).

Table 46 shows the characteristics excluded from the driver-related regression model. Unless the characteristic has a comb max value of over 50 (see analysis I), we will not integrate these variables into further analyses (i.e., pattern recognition with Bayesian networks and pattern recognition with the PATTERMAX method). For the generation of the decision trees, we will integrate all characteristics (as we will compare the decision tree outcomes with the logistic regression results).

Variable	*Excluded characteristics*
Sex	female
Age Class	45 to 54, 55 to 64, 65+
Driving licence	no driving licence
Impairments	distraction, fatigue, health, drugs, medicine, excitation
Driving manoeuvres	speeding, hitting an obstacle next to the road, misconduct by a pedestrian, overtaking, cutting curves, changing lanes, inadequate safety distance, reverse driving, phoning, turning around, falling from the vehicle, getting in the lane, disregarding driving direction, priority violation, driving towards the left side of the road, forbidden overtaking, hitting a moving vehicle, disregarding driving ban, driving in parallel, opening the vehicle door, hitting a stationary vehicle, wrong-way driver, disregarding red light, dangerous stopping and parking, disregarding turning ban, missing indication of direction change, driving against one way, driving without mandatory light
Safety settings	no safety belt applied

Table 46: Variables excluded from the driver-related regression model.

Excluded variables that show a comb max value of over 50 are speeding, distraction, fatigue, female driver, age class 45 to 54, age class 55 to 64 and age class 65+ (see chapter 4.6). We will additionally integrate these variables in our pattern recognition process. We include these variables because of their high maximum combination value. This value indicates that the respective variable or characteristic often occurs in a single pattern. Therefore, it might be an essential variable or characteristic for blackpattern detection.

5.3 Logistic regression with vehicle-related variables

We proceed with the estimation of a vehicle-related logistic regression model. Vehicle-related variables include

- kilometrage (km),
- engine power (kW),
- the vehicle colour, and
- vehicle safety settings

with 32 detailed characteristics (see chapter 4.7), which we include as dummy variables in our model (0=characteristic is not present, 1=characteristic is present). The model estimates the impact of these variables on our binary target variable *severe casualties* (0=no severe casualty, 1=severe casualty). Stepwise variable selection with *Likelihood Ratio* performs nine steps to estimate the model. Table 47 presents the estimated driver-related regression model. The Hosmer-Lemeshow goodness-of-fit test shows the good model fit with p=0,795. The regression model assigns 'insufficient load security' (*exp(β)=4,684)*, 'vehicle fire' (*exp(β)=5,664),* and 'airbag not deployed' (*exp(β)=2,403)* with the highest impact on *severe casualties*. Engine power of '24-90kW' appears to reduce the risk of *severe casualties*.

Variable	*regression coefficient* β	*standard error SEM*	*Sig. p*	*exp(β)*
engine power 24-90 kW	-0,118	0,087	0,175	0,889
engine power 90-110 kW	0,243	0,100	0,016	1,275
engine power 110+ kW	0,306	0,010	0,002	1,358
insufficient vehicle security	1,544	0,527	0,003	4,684
vehicle fire	1,734	0,486	0,000	5,664
airbag not deployed	0,877	0,043	0,000	2,403
vehicle colour: green	0,349	0,0734	0,000	1,418
constant	*-2,178*	*0,090*	*0,000*	*0,113*

Table 47: Vehicle-related logistic regression model. Input data: 20.293 single-vehicle accidents with single occupation occurring outside the built-up area on the Austrian road network between 2012-2019. The dataset includes 32 vehicle-related characteristics as dummy variables (0=characteristic is not present, 1=characteristic is present). The binary target variable is *severe casualties* (0=no severe casualty, 1=severe casualty).

The estimated model excludes the following vehicle-related characteristics. Unless a characteristic has a comb max value of over 50 (see analysis I), we will not integrate these characteristics into further analyses (i.e., pattern recognition with Bayesian networks and pattern recognition with the PATTERMAX method). For the generation of the decision trees, we will integrate all characteristics (as we will compare the decision tree outcomes with the outcomes of logistic regression).

Variable	*Excluded characteristics*
Engine power (kW)	0-24
Kilometrage (km)	0 to 15.000, 15.000 to 75.000, 75.000 to 100.000, 100.000 to 150.000, 150.000 to 200.000
Vehicle colour	beige, blue, brown, bronze, dark, yellow, gold, grey, bright, orange, red, black, silver, purple, white, others
Vehicle safety settings	insufficient load security, technical defects, airbag deployed

Table 48: Variables excluded from the vehicle-related regression model.

Based on the maximum combination value (see chapter 4.7), we additionally integrate the following characteristics in the pattern recognition process (Bayesian networks and PATTERMAX-method): kilometrage of 15.000-75.000km, kilometrage of 150.000-200.000km and the vehicle colours blue, brown, grey, red, black, silver and white. We include these variables because of their high maximum combination value. This value indicates that the respective variable or characteristic often occurs in a single pattern. Therefore, it might be an essential variable or characteristic for blackpattern detection.

5.4 Logistic regression with roadway-related variables

The generation of a roadway-related logistic regression model involves the following roadway-related variables:

- speed limit,
- road type,
- road characteristics,
- road condition, and
- traffic lights.

These variables consist of 50 detailed characteristics. We will integrate these characteristics as dummy variables (0=characteristic is not present, 1=characteristic is present) into our roadway-related regression model. The model performs a 12-step variable selection process based on the *Likelihood Ratio*. The Hosmer-Lemeshow-goodness-of-fit test results in a $p < 0,05$ in all steps (step 12: p=0,001). Thus, the roadway-related logistic regression model appears to be subject to randomness (i.e., the observed values deviate from the expected values). Therefore, the analysis of roadway-related variables does not allow us to conclude on the impact of these variables on *severe casualties*. However, we include roadway-related variables into our overall regression model (see chapter 5.6) and see if this model includes statistically sound roadway-related variables.

5.5 Logistic regression with situation-related variables

Situation-related variables include the following predictor variables

- time,
- weekday,
- meteorological season,
- weather conditions, and
- light conditions

and their characteristics. The regression model estimates the relationship and impact of 22 situation-related dummy characteristics (0=characteristic is not present, 1=characteristic is present) on the target variable *severe casualties* (0=no severe casualty, 1=severe casualty). The model performs stepwise variable selection based on *Likelihood Ratio* with eight steps in total. Table 49 represents the estimated situation-related regression model (i.e., the result of step eight). The Hosmer-Lemeshow goodness-of-fit test results in p=0,883, indicating that the estimated model shows a good fit. Weekdays 'Monday to Thursday' appear to decrease the risk of observing a severe or fatal road traffic accident. 'Snow' (exp(β)=1,921), the time between '0 a.m. and 6 a.m.' (exp(β)=1,415), and 'rain' (exp(β)=1,270) appear to have the positive impact on *severe casualties* when comparing situation-related characteristics.

Variable	*regression coefficient β*	*standard error SEM*	*Sig. p*	*exp(β)*
12 a.m. to 6 a.m.	0,347	0,072	0,000	1,415
12 p.m. to 6 p.m.	0,119	0,051	0,019	1,126
6 p.m. to 12 a.m.	0,199	0,064	0,002	1,220
Monday to Thursday	-0,077	0,038	0,042	0,926
winter	0,271	0,046	0,000	1,312
rain	0,239	0,055	0,000	1,270
snow	0,653	0,085	0,000	1,921
darkness	0,130	0,058	0,024	1,138
Constants	*-2,744*	*0,106*	*0,000*	*0,064*

Table 49: Situation-related logistic regression model. Input data: 20.293 single-vehicle accidents with single occupation occurring outside the built-up area on the Austrian road network between 2012-2019. The dataset includes 22 situation-related characteristics as dummy variables (0=characteristic is not present, 1=characteristic is present). The binary target variable is *severe casualties* (0=no severe casualty, 1=severe casualty).

Our estimated model excludes the following situation-related characteristics (see table 50).

Variable	*Excluded characteristics*
Time	6 to 12
Weekday	Friday to Sunday
Meteorological Season	spring, summer, autumn
Weather conditions	clear or overcast weather, hail, freezing rain, fog, high wind
Light conditions	daylight, dusk or dawn, artificial light, restricted view by vehicle, glare from the sun

Table 50: Characteristics excluded from the situation-related logistic regression model.

Unless these characteristics do not result in a comb value of over 50 (see chapter 4.9), we dismiss these characteristics for the pattern recognition process (i.e., Bayesian networks and PATTERMAX-method). This applies to the following characteristics: time between 6 a.m. and 12 p.m., Friday to Sunday, spring, summer, winter, clear or overcast weather, daylight and dusk or dawn. We include these variables because of their high maximum combination value. This value indicates that the respective variable or characteristic often occurs in a single pattern. Therefore, it might be an essential variable or characteristic for blackpattern detection. For the generation of the situation-related decision tree, we will integrate all 22 situation-related characteristics as we want to compare the outcomes of our logistic regression model with the outcome of the decision tree.

5.6 Logistic regression with all accident-related variables

The last part of the logistic regression chapter foresees the estimation of an overall logistic regression model. This model integrates 158 accident-describing characteristics from four categories (driver, vehicle, roadway and situation) as dummy variables.

The model performs a 34 -step variable selection process based on the *Likelihood Ratio*. The final model (step 34) results in p=0,952 for the Hosmer-Lemeshow goodness-of-fit test, which indicates that the model shows a good fit. Table 51 shows the estimated overall logistic regression model. We can see that the variables 'no safety belt applied' (exp(β)=5,015), 'gallery' (exp(β)=4,583), 'hitting an obstacle on the road' (exp(β)=3,394), 'airbag not deployed' (exp(β)=2,223), 'age class 16 to 18', exp(β)=2,317), and 'bridge' (exp(β)=2,166) have the highest impact on *severe casualties*. Even if some of these characteristics occur very rarely, they show a relatively high probability for a *severe casualty* in the case of their occurrence. 'Speed limit 50 km/h' and 'hitting a guard rail' appear to decrease the probability of *severe casualties*.

Variable	*regression coefficient β*	*standard error SEM*	*Sig. p*	*exp(β)*
0 a.m. to 6 a.m.	0,307	0,058	0	1,359
speed limit 50km/h	-0,329	0,144	0,022	0,719
speed limit 100km/h	0,114	0,046	0,013	1,12
intersection	0,45	0,148	0,002	1,569
curve	0,18	0,043	0	1,198
bridge	0,773	0,197	0	2,166
gallery	1,522	0,589	0,01	4,583
tunnel	0,515	0,258	0,046	1,674
one-way	0,507	0,219	0,02	1,66
dry road	0,232	0,047	0	1,261
wintry conditions	0,38	0,07	0	1,462

Table 51: Overall logistic regression model. Input data: 20.293 single-vehicle accidents with single occupation occurring outside the built-up area on the Austrian road network between 2012-2019. The dataset consists of 160 accident-describing characteristics, which we integrate as dummy variables (0=characteristic is not present, 1=characteristic is present) into the model. The binary target variable is *severe casualties* (0=no severe casualty, 1=severe casualty).

Variable	*regression coefficient β*	*standard error SEM*	*Sig. p*	*exp(β)*
darkness	0,165	0,049	0,001	1,18
county road	0,247	0,062	0	1,28
other road	0,397	0,082	0	1,487
engine power 24-90 kW	0,175	0,046	0	1,192
sudden braking	0,693	0,324	0,032	2
hitting a tree	0,365	0,075	0	1,441
hitting an obstacle on the road	1,222	0,426	0,004	3,394
hitting a guard rail	-0,313	0,091	0,001	0,731
vehicle fire	1,394	0,541	0,01	4,029
hit and run	0,552	0,161	0,001	1,737
age class 16 to 18	0,84	0,104	0	2,317
age class 19 to 24	0,743	0,057	0	2,101
age class 25 to 34	0,492	0,057	0	1,635
age class 35 to 44	0,308	0,065	0	1,361
drifting left	0,147	0,041	0	1,158
male driver	0,491	0,045	0	1,634
probationary driving licence	0,166	0,078	0,033	1,181
alcohol	0,65	0,062	0	1,916
no safety belt applied	1,612	0,062	0	5,015
airbag not deployed	0,803	0,046	0	2,233
vehicle colour: green	0,275	0,078	0	1,317
constant	*-9,285*	*0,611*	*0*	*0*

Continuation of table 51: Overall logistic regression model. Input data: 20.293 single-vehicle accidents with single occupation occurring outside the built-up area on the Austrian road network between 2012-2019. The dataset consists of 160 accident-describing characteristics, which we integrate as dummy variables (0=characteristic is not present, 1=characteristic is present) into the model. The binary target variable is *severe casualties* (0=no severe casualty, 1=severe casualty).

Table 52 shows the variables excluded from the overall logistic regression model.

Variable	*Excluded characteristics*
Sex	female driver
Age class	45 to 54, 55 to 64, 65+
Time	6 to 12, 12 to 18, 18 to 24
Weekday	Monday to Thursday, Friday to Sunday
Meteorological Season	spring, summer, autumn, winter
Weather conditions	clear or overcast weather, hail, freezing rain, fog, high wind, snow
Light conditions	daylight, dusk or dawn, artificial light, restricted view by vehicle, glare from the sun
Speed limit	driving ban, 5, 10, 20, 30, 40, 60, 70, 80, 90, 110, 120, 130
Road characteristics	roundabout, traffic light in operation, straight road, narrow road, deceleration lane, acceleration lane, cycle path, pedestrian and cycle path, entry or exit, banquet, secondary lane, hard shoulder, parking lane, rest area, underpass, traffic island, tram or bus, station, middle separation, construction site, crosswalk
Road conditions	wet road, sand or grit on the road, other road condition
Road type	Highway, expressway
Engine power (kW)	0-24, 90-110, 110+
Kilometrage (km)	0-15.000, 15.000-75.000, 100.000-150.000, 150.000-200.000
Driving manoeuvre	getting in the lane, reverse driving, overtaking, turning around, changing lanes, driving in parallel, cutting curves, wrong-way driver, disregarding driving direction, disregarding driving ban, forbidden overtaking, speeding, disregarding red light, priority violation, phoning, inadequate safety distance, skidding/drifting, opening the vehicle door, misconduct by a pedestrian, hitting an obstacle next to the road, hitting a moving vehicle, hitting a stationary vehicle, drifting right
Vehicle safety settings	insufficient vehicle security, insufficient load security, technical defects
Impairment	distraction, drugs, medicines, fatigue, excitation, health
Driving licence	no driving licence
Vehicle colour	beige, blue, brown, bronze, dark, yellow, golden, grey, bright, orange, red, black, silver, violet, white

Table 52: Variables excluded from the overall logistic regression model based on a stepwise variable selection with *Likelihood Ratio*.

6. Road traffic accident data analysis III: Decision Trees

In addition to binary logistic regression, we estimate decision trees to detect accident-related characteristics that appear to impact our dependent variable *severe casualties*. We are interested in comparing the outcome of binary logistic regression with the outcomes of decision trees. Even if decision trees do not analyse accident-related characteristics in such detail as logistic regression, we are curious to see whether the identified characteristics will be similar to those detected in our regression models. As in the previous chapter, we create a decision tree for each accident-describing category (driver, vehicle, roadway, situation). Afterwards, we will generate a decision tree including all the variables of the four categories. To grow the decision trees, we apply the CHAID-algorithm (Chi-square Automatic Interaction Detector). There exist several reasons to apply this algorithm, which we discuss in the following chapter.

6.1 Generation of decision trees

The CHAID-algorithm discovers the relationships between our accident-describing variables and their respective characteristics. Chi-square is the metric to detect the significance of a characteristic. This approach extends analysis part I, where we applied Fisher's Exact Test to estimate the relationship between accident-related variables and the target variable *severe casualties*. As in the previous analyses, the entire dataset consists of binary accident-related characteristics (0=characteristic is not present, 1=characteristic is present). We define *severe casualties* as our target variable (0=no severe casualty, 1=severe casualty). The CHAID-algorithm expects the target variable to be categorical.

The formula of chi-squared testing is:

$$\sqrt{((y - y')^2 / y')}$$

where:

y … the actual value

y' … the expected value

CHAID generates a 2x2 field table for each accident-related variable with the target variable *severe casualties*. We use the same example as in the logistic regression chapter (chapter 5) to illustrate the logic behind the CHAID-algorithm. Table 53 shows the observed values for the characteristics 'no safety belt applied' (independent variable) and severe casualties'(dependent variable). Table 54 shows the respective expected values.

	no safety belt applied (x=1)	safety belt applied (x=0)	*total*
severe casualty (y=1)	699 p_1=33 %	2.731 p_2=15 %	3.430 17 %
no severe casualty (y=0)	1.401 1-p_1=66 %	15.462 1-p_2=85 %	16.863 83 %
total	2.100 10 %	18.193 90 %	20.293 100 %

Table 53: 2x2 field table showing the *observed* values for *severe casualty* (target variable) and 'no safety belt applied' (independent variable). n=20.293 single-vehicle accidents with a single occupation that occurred on the Austrian road network outside the built-up area between 2012 and 2019.

	no safety belt applied (x=1)	safety belt applied (x=0)	*total*
severe casualty (y=1)	355 p_1=17 %	3.075 p_2=17 %	3.430 17 %
no severe casualty (y=0)	1.745 1-p_1=83 %	15.118 1-p_2=83 %	16.863 83 %
total	2.100 10 %	18.193 90 %	20.293 100 %

Table 54: 2x2 field table showing the *expected* values for *severe casualty* (target variable) and 'no safety belt applied' (independent variable). n=20.293 single-vehicle accidents with a single occupation that occurred on the Austrian road network outside the built-up area between 2012 and 2019.

The shown example results in a chi-square value of 1.165,52 and p=0,000.

As we have 158 accident-describing characteristics, CHAID creates 158 2x2 field tables and calculates each table's chi-square value. The table with the maximum chi-square value becomes the root node. In our case, only two decisions can emanate from this node: 0=characteristic is not present (subset 1), and 1=characteristic is present (subset 2). Both subsets contain different proportions of *severe casualties*. The target of the CHAID algorithm is to create sub informational datasets having a single decision such that *severe casualties* exclusively consist of ones or zeros. Thus, the CHAID algorithm continues to calculate chi-square values and decision nodes for each data subset until the sub informational datasets exclusively include the same decision (0 or 1) for the target variable.

As we will see in chapters 6.2 and 6.6, our illustrated example will result in the maximum chi-square value among driver-related characteristics and all accident-describing characteristics. It will become the root node in both cases. However, the CHAID-algorithm mainly selects those variables which also have a high impact on the target variable (exp(b) in the logistic regression model.

6.2 Decision tree with driver-related variables

The driver-related decision tree works with 54 driver-related characteristics representing dummy variables. The dichotomous target variable is *severe casualties* (0=no severe casualty, 1= severe casualty). Figure 45 shows the driver-related decision tree we generated with the CHAID-algorithm.

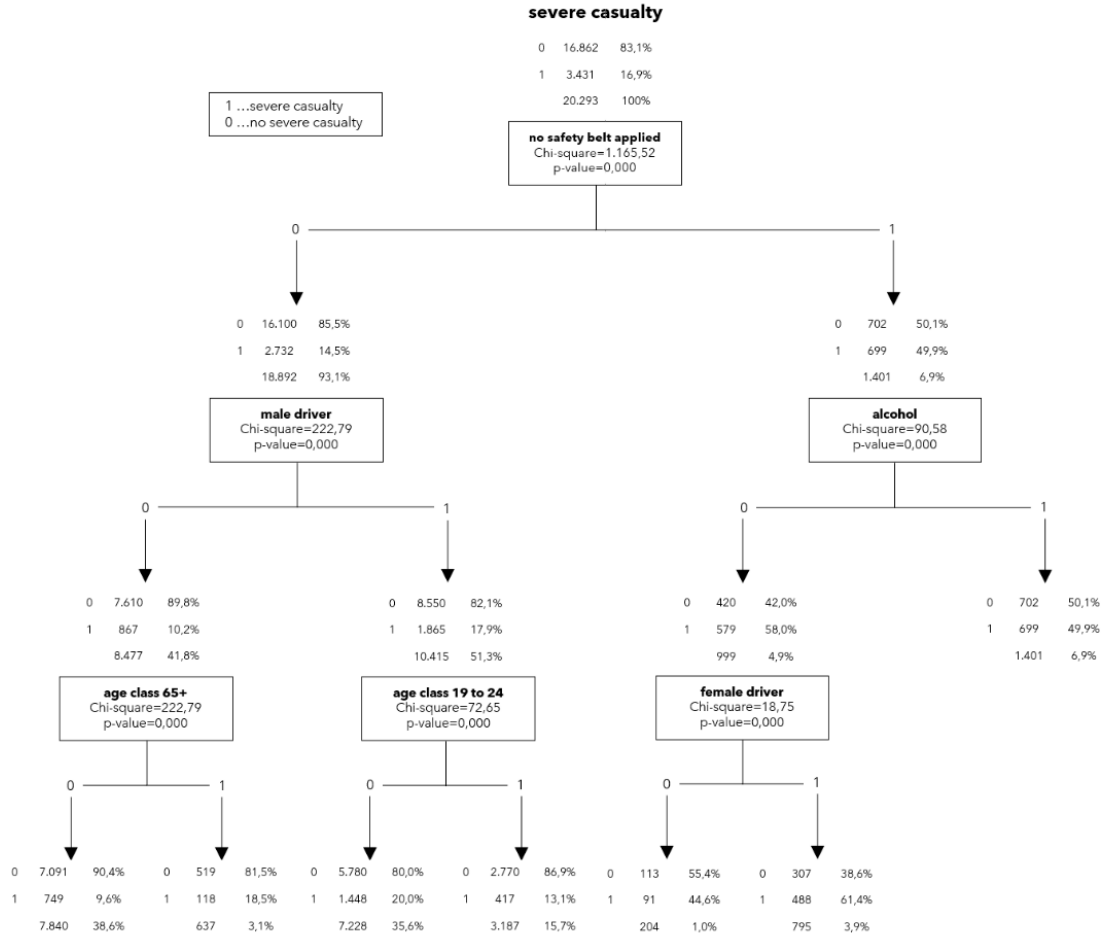

Figure 45: Driver-related decision tree. Input data: n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

The root note is 'no safety belt applied' results in a Chi-square value of 1.165,52 and correlates with *severe casualties*. If 'no safety belt applied' is true, the characteristic showing the highest Chi-square value among the subset is 'alcohol'. The joint probability of a severe or fatal accident with 'no safety belt applied' and alcohol' is 2 % (total: 499). If 'alcohol' does not apply, the next decision node leads us to 'female driver'. The joint probability of a severe or fatal accident with 'no safety belt applied' and 'female driver' is 3 % (total: 579). Furthermore, the decision tree relates 'male driver' with 'age class 19 to 24'. The joint probability of this variable combination with *severe casualties* is 9 % (total: 1.865). If neither 'no safety belt applied' nor 'male driver' is valid, the decision tree leads us to the decision node 'age class 65+'. The joint probability for a severe or fatal accident and a driver above 65 years is 0,6 % (total: 118).

Compared with the outcomes from binary logistic regression, we can see that the driver-related decision tree does not integrate any characteristic that would not be part of the

regression model. However, we can see relations among the presented characteristics within the decision tree. On the other side, the logistic regression model provides information on the impact of a driver-related characteristic to increase *severe casualties*.

6.3 Decision tree with vehicle-related variables

The vehicle-related decision tree works with 32 vehicle-related characteristics representing dummy variables. The dichotomous target variable is *severe casualties* (0=no severe casualty, 1= severe casualty). Figure 46 shows the vehicle-related decision tree we generated with the CHAID-algorithm. The root node of the vehicle-related decision tree is 'airbag not deployed'. In the logistic regression model, 'airbag not deployed' also showed a high impact on *severe casualties*. Additionally, the vehicle-related decision tree selects engine power '24-90kW', and the vehicle colours green and grey.

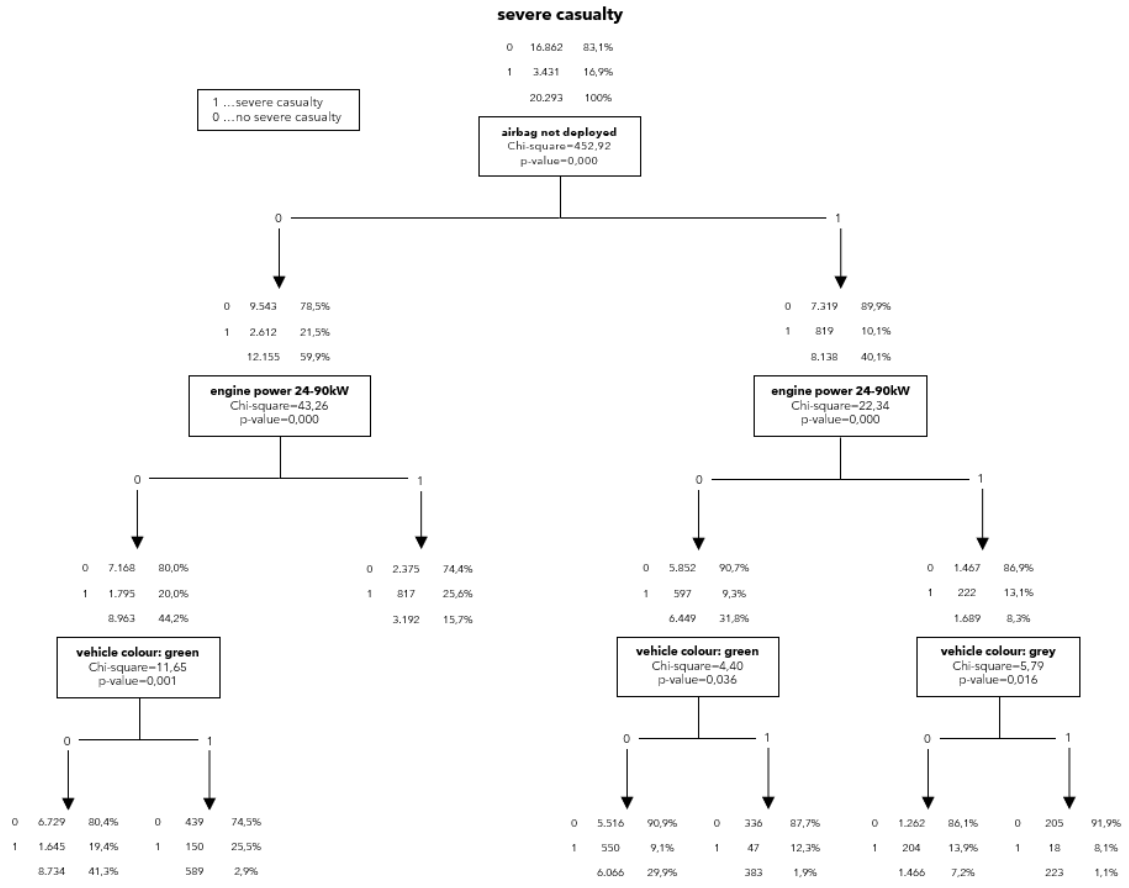

Figure 46: Vehicle-related decision tree. Input data: n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

6.4 Decision tree with roadway-related variables

The roadway-related decision tree works with 50 roadway-related characteristics representing dummy variables. The dichotomous target variable is *severe casualties* (0=no severe casualty, 1= severe casualty). Figure 47 shows the roadway-related decision tree we generated with the CHAID-algorithm.

The root node is 'wintry conditions'. The probability of a *severe casualty* in wintry conditions is 19 % (total: 3.771). If 'wintry conditions' apply, the tree shows a direct connection to 'other roads'. The resulting joint probability is 3 %. If 'other road' is not valid, the decision tree leads us to the decision node 'straight road'. Thus, the joint probability of a *severe casualty* with 'wintry conditions' and 'straight road' is 0,6 % (total: 112). If the root node 'wintry conditions' does not apply, the decision tree generates a link to 'wet road'. The joint probability of a *severe casualty* with 'wet roads' is 9,2 % (total: 1.865). If 'wet road' is valid, the decision tree links to 'curve'. The joint probability of a *severe casualty* and 'wet road' and 'curve' is 18 % (total: 362). If neither the root node 'wintry conditions' nor the decision node 'wet road' is true, the decision tree links to 'bridge'. The probability of a *severe casualty* and a bridge is 0,1 % (total: 23). Even if some characteristics occur very rarely (for example, 'bridge'), they show a relatively high probability for a *severe casualty* in case of their occurrence. For example, 20 % of accidents occurring on a bridge resulted in a severe or a fatal road traffic accident.

The roadway-related logistic regression model could not generate a significant model. Therefore, we cannot compare the outcomes of the decision tree with those of logistic regression.

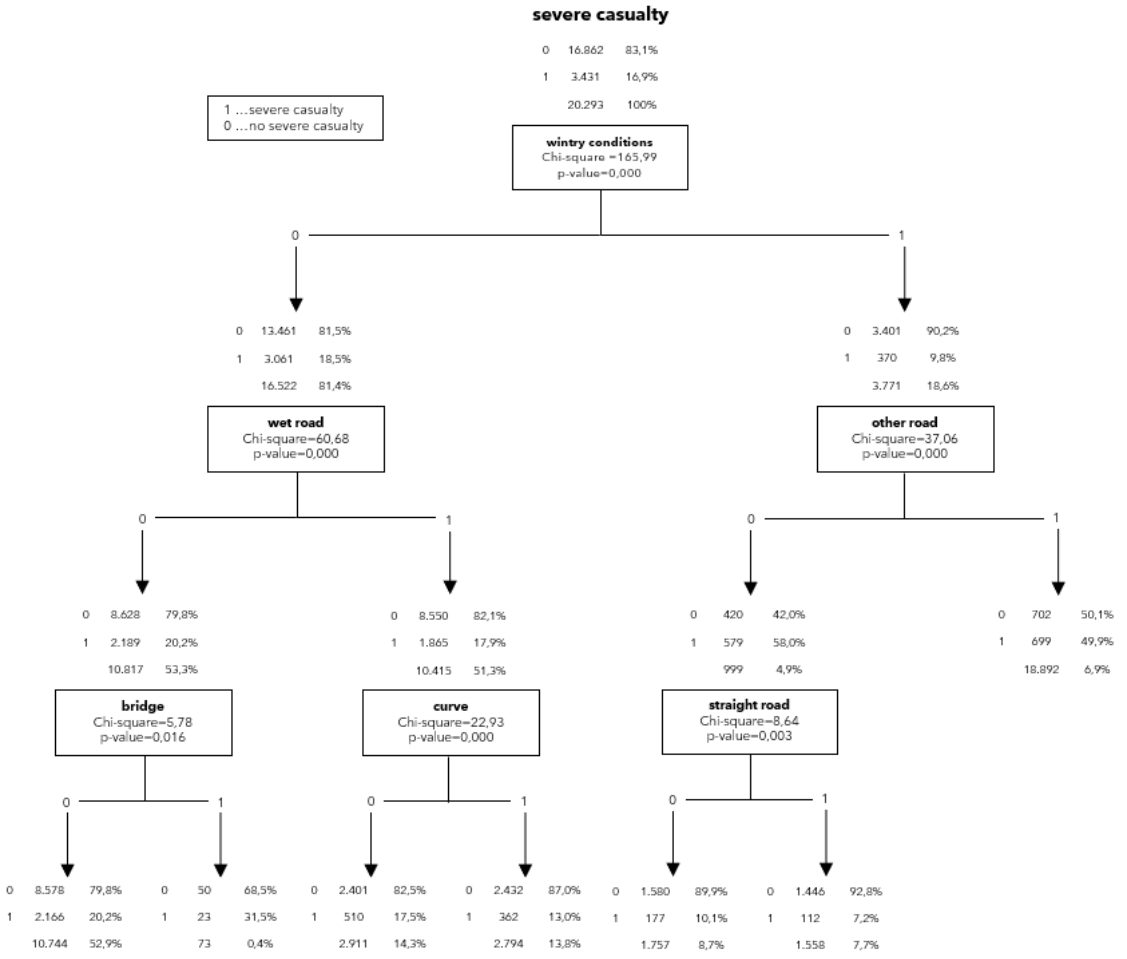

Figure 47: Roadway-related decision tree. Input data: n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

6.5 Decision tree with situation-related variables

The situation-related decision tree works with 22 characteristics representing dummy variables. The dichotomous target variable is *severe casualties* (0=no severe casualty, 1= severe casualty). Figure 48 shows the situation-related decision tree we generated with the CHAID-algorithm.

The situation-related decision tree shows ‘snow’ as the root node. The probability of a severe or fatal road traffic accident in snowy conditions is 1 % (total: 191). If ‘snow’ is true, the tree links to the weekdays ‘Monday to Thursday’. The joint probability of a severe or fatal accident in snowy conditions and on weekdays between Monday and Thursday is 0,5 %. If ‘Monday to Thursday’ is not valid, the decision tree leads us to the decision node ‘winter’. Thus, the joint probability of a severe or fatal accident in snowy conditions and winter is 0,3 % (total: 63). If the root node ‘snow’ does not apply, the decision tree leads us to the decision node ‘darkness’. The probability of a severe or fatal accident in ‘darkness’ is 10 % (total: 2.018). ‘Darkness’ shows a connection to ‘rain’. The joint probability of a severe or fatal road traffic accident in ‘darkness’ and ‘rain’ is 1 % (total: 275). If neither the root node ‘snow’ nor ‘darkness’ apply, the decision tree leads us to ‘winter’. The probability of a severe or fatal accident in winter is 2 % (total: 353).

Within the logistic regression model, ‘snow’, ‘winter’, ‘rain’, and ‘darkness’ represent variables with a high impact on *severe casualties*.

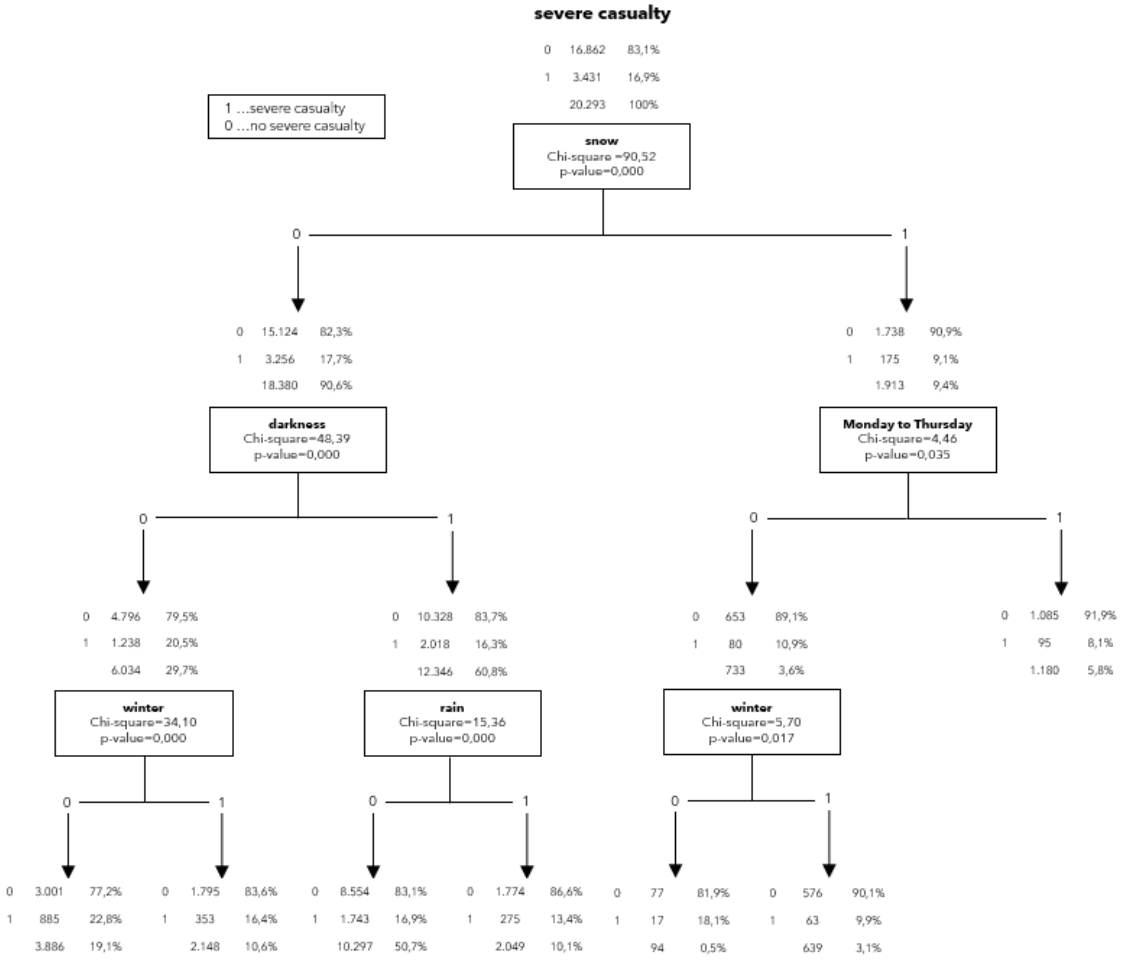

Figure 48: Situation-related decision tree. Input data: n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

6.6 Decision tree with all accident-related variables

For the decision tree, we include all accident-describing characteristics as dummy variables. Table 55 shows the characteristics to be integrated into the overall decision tree.

Category	*Type of variable*	*Characteristics*
Driver	Content variable	Sex: male, female Age class: 16 to 18, 19 to 24, 25 to 34, 35 to 44, 45 to 54, 55 to 64, 65+ Driving licence type: probationary driving licence Distraction: alcohol, distraction, fatigue Driving manoeuvre: speeding, skidding/drifting, hitting a tree, hitting an obstacle on the road, hitting a guard rail, sudden braking, hit and run Safety settings: no safety belt not applied
Vehicle	Content variable	Vehicle settings: airbag not deployed
Roadway	Content variable	Speed limit: driving ban, 70, 80, 100, 130 Road characteristics: intersection, curve, middle separation Road condition: wet road, wintry conditions, sand or grit on the road Road type: highway, expressway, regional road, other roads
Situation	Content variable	Weather conditions: rain, snow Light conditions: darkness, dusk or dawn Time: 0 to 6, 6 to 12, 12 to 18,18 to 24 Weekday: Monday to Thursday, Friday to Sunday Season: Winter, Spring, Summer, Autumn

Table 55: Accident characteristics to generate the overall decision tree.

The resulting overall decision tree shows a condensed picture of accident-describing characteristics. Within the overall decision tree (see figure 49), 'no safety belt applied' remains on the very top as the root node. The probability of a severe or fatal accident including a driver with 'no safety belt applied' is 3 % (total: 699). If 'no safety belt applied' is true, we can see a direct connection to 'alcohol'. The joint probability of a severe or fatal accident including the characteristics 'no safety belt applied' and 'alcohol' is 0,6 % (total: 120). So far, the overall decision tree shows the same picture as the driver-related decision tree. If 'alcohol' applies, the decision tree leads us to the decision node 'speed limit of 100 km/h'. The resulting joint probability of a severe or fatal casualty, 'no safety belt applied, 'alcohol' and 'speed limit of 100 km/h' is 0,2 % (total: 40). If 'alcohol' does not apply, the tree leads us to the decision node 'airbag not deployed'. The joint probability of a severe or fatal accident, 'no safety belt applied'

and ‘airbag not deployed’ is 0,8 % (total: 159). If the root node ‘no safety belt applied’ is not true, the next decision node represents ‘airbag not deployed’. Among the subset, the probability of a severe or fatal accident and ‘airbag not deployed’ is 3 % (total: 629). If ‘airbag not deployed’ is true, we can see a connection to ‘age class 65+’. The resulting joint probability of *severe casualty*, ‘airbag not deployed’ and ‘age class 65+’ is 0,5 % (total: 107). If neither ‘no safety belt applied’ nor ‘airbag not deployed’ is true, the decision tree leads us to the decision node ‘male driver’. In this case, the probability of a severe or fatal accident including a ‘male driver’ is 7 %.

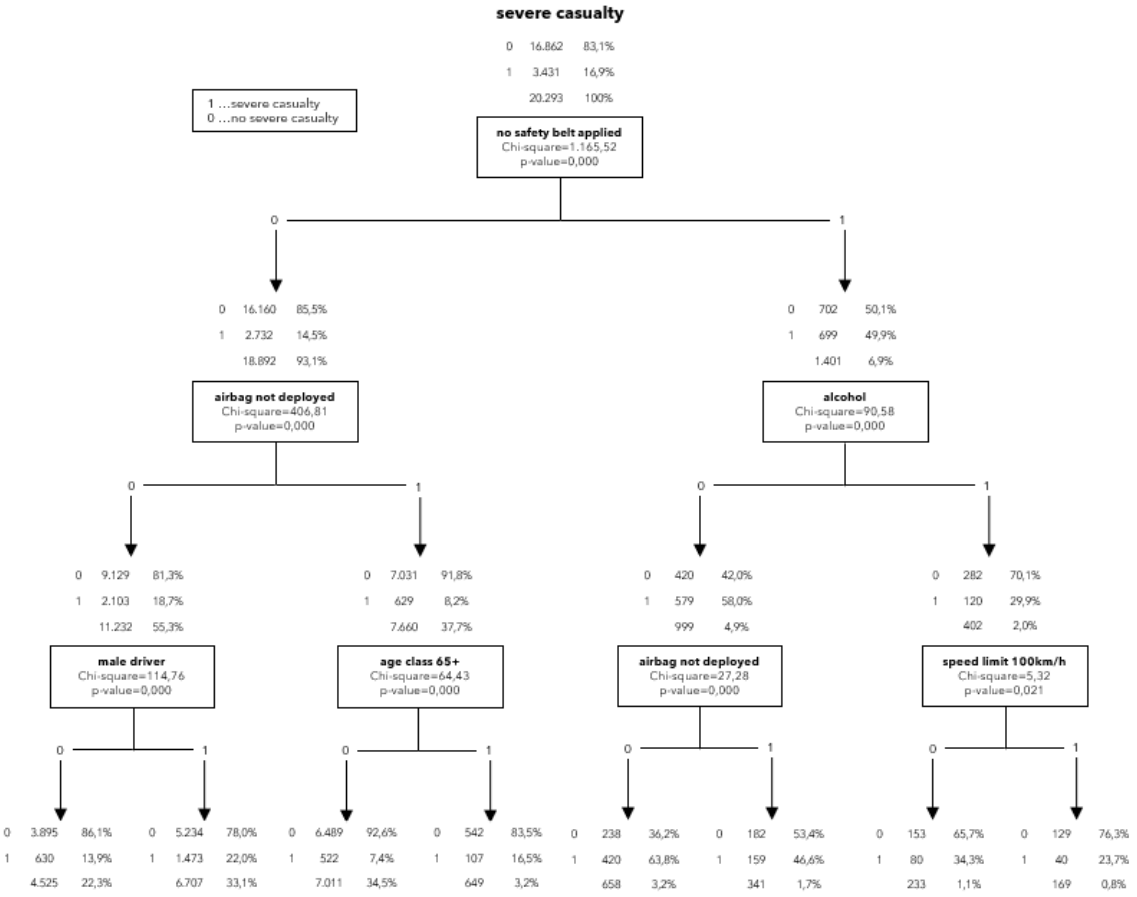

Figure 49: Decision tree generated with all accident-describing variables. Input data: n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

Within the overall logistic regression model, the characteristics ‘no safety belt applied’, ‘gallery’, ‘hitting an obstacle on the road’, ‘airbag not deployed’, ‘age class 16 to 18’, and ‘bridge’ showed the most significant impact on *severe casualties*. Thus, according to both approaches, the characteristics ‘no safety belt applied’ and ‘airbag not deployed’ represent two key characteristics to increase the risk of a *severe casualty*.

7. Road traffic accident data analysis IV: Bayesian networks

Given the random variables x_1, …, x_n with d values each, the associated probability results in a total of d^n values. Thus, the required storage capacity and the computational time for calculating probabilities grow exponentially with the number of variables. In some cases, probability values might be unknown, and their determination is time-consuming.

In practice, many application problems are highly structured or even overstructured, so that the distribution contains a lot of redundancy. The so-called Bayesian networks can reduce these (over)structured distributions. Bayesian networks are directed graphs where the nodes represent predictor variables (statements) and the edges represent the stochastic dependencies between the statements(Dörn, 2017, p. 149). In our case, we do not use Bayesian networks to determine unknown probabilities as we have an empirical and representative dataset on road traffic accidents and their outcomes (degree of injury). We generate Bayesian networks to detect causal relationships among accident-describing characteristics.

7.1 Generation of the Bayesian networks

The so-called Naïve Bayes Classifier is a simple and commonly used algorithm to generate a Bayesian network. Using conditional probabilities, the Naive Bayes Classifier determines which class an object belongs to with the most significant probability. Thus, the Naïve Bayes classifier is a simple probabilistic classifier based on the Bayes' Theorem.

The Bayes' Theorem determines the conditional probability as follows:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \times P(B \mid A)}{P(B)}$$

where:

$P(A \cap B)$	joint probability of A and B
$P(A \mid B)$	the conditional probability of the event A under the condition that B has occurred (a posteriori probability)
$P(B \mid A)$	the conditional probability of the event B under the condition that A has occurred
$P(A)$	the a priori probability of event A (prior of A)
$P(B)$	the a priori probability of event B (prior of A)

Conditional Probability

$P(A \mid B)$ stands for the conditional probability of A with given B, or the probability of A under the condition of B. Thus, in a random experiment, if event B is known, the possible outcomes of the experiment are reduced to B. The joint probability is the connected probability of two events.

Independence

The Naïve Bayes Classifier makes strong independence assumptions. This means that a particular feature of a class is independent of every other feature of the class.

Two events A and B are independent if both A and B have positive probabilities and if

$$P(A \mid B) = P(A) \text{ and } P(B \mid A) = P(B)$$

Example

We will explain the Bayesian network approach with the following example: When investigating historical single-vehicle road traffic accidents with single occupation and outside the built-up area, what is the probability that the accident results in a severe or fatal accident and that the driver is between 19 and 24 years old?

Between 2012-2019, 3.430 severe single-vehicle accidents with single occupation occurred outside the built-up area in Austria. 23,50% of these accidents involve 19 to 24 years old drivers (n = 806). In total, 20.293 severe single-vehicle accidents with single occupation occurred outside the built-up area in Austria within this period.

So, we have the following events:

SC: it is a severe casualty

AC: the driver is between 19 to 24 years old

The probability of a randomly selected accident to be a severe or fatal accident is:

$$\mathrm{P}\,(\mathrm{SC}) = \frac{3.430}{20.293} = 0{,}1690$$

The probability of a 19 to 24-year-old driver and a severe or fatal accident is 23,50 %, which corresponds to the following conditional probability:

$$\mathrm{P}\,(\mathrm{AC} \mid \mathrm{SC}) = 0{,}2350$$

With the following formula, we calculate the probability that the involved driver is between 19 and 24 years old:

$$P\,(AC \mid SC) = \frac{\mathrm{P}\,(\mathrm{AC} \cap \mathrm{SC})}{\mathrm{P}\,(\mathrm{SC})}$$

Let us put in the values:

$$0{,}2350 = \frac{\mathrm{P}\,(\mathrm{AC} \cap \mathrm{SC})}{0{,}1690}$$

Result:

$$\mathrm{P}\,(\mathrm{AC} \cap \mathrm{SC}) = 0{,}0397$$

The chance of a severe or fatal accident with a driver between 19 and 24 years old is 3,97%.

We will now generate a Bayesian network for each category (driver, vehicle, roadway, situation, and accident). Subsequently, we will create a Bayesian network including the variables from all categories.

Train-Test-Split

Since we do not build a prediction model, we do not use a train-test-split and use 100% of our observed data as input data.

Network Structure

We apply Tree Augmented Naïve Bayes (TAN) structure, a simple Bayesian network model. Compared to the classic Naïve Bayes model, it allows a predictor to depend on another predictor. TAN reveals several relationships between the accident-related variables.

Parameter learning

For parameter learning, we apply Bayes adjustment for small cell counts.

Predictor importance

The estimated Bayesian networks emit predictor importance for each accident-related characteristic. Predictor importance indicates the relative importance of a predictor variable for the model estimation. Predictor importance is the result of a variance-based sensitivity analysis. Since interactions exist among our input factors and our datasets consist of more than a few hundred entries, we do not estimate predictor importance. Otherwise, we calculate with too much of an inaccuracy risk. (Saltelli, Tarantola, Campolongo, and Ratto, 2004) The following chapters show the Bayesian networks for each category (driver, vehicle, roadway and situation) and the overall Bayesian network.

7.2 Bayesian network of driver-related variables

The generation of the driver-related Bayesian network results in integrating 21 driver-related characteristics. These characteristics represent dummy variables. Table 56 shows the role of the variables within the driver-related Bayesian network. The target variable is *severe casualties*.

Target variable	*Predictor variables*	
Degree of injury: • severe casualties	*Sex:* • male • female *Age class:* • 16 to 18 • 19 to 24 • 25 to 34 • 35 to 44 • 45 to 54 • 55 to 64 • 65+ *Impairment:* • alcohol • fatigue • distraction	*Driving manoeuvre:* • sudden braking • speeding • skidding/drifting • hitting a tree • hitting an obstacle on the road • hitting a guard rail • hit and run *Driving licence type:* • probationary driving licence *Safety settings:* • no safety belt applied

Table 56: Driver-related characteristics for the Bayesian network generation. The network includes 21 driver-related characteristics as dummy variables (0=characteristics not present, 1=characteristic is present) and the dichotomous target variable severe casualty (0=no severe casualty, 1=severe casualty).

The driver-related Bayesian network represents a TAN network and uses Bayes adjustment for small cell counts for parameter learning. Figure 50 illustrates the generated Bayesian network for driver-related characteristics. The network enlightens detected relationships among the characteristics and illustrates the joint probabilities of these relationships. The network detects a relationship among *severe casualties*, 'distraction', and 'hitting a tree'. It results in a joint probability of 0,46 %. The network also associates 'hitting an obstacle on the road' with 'distraction' (joint probability is 0,00 %). This value appears extremely low, but we must consider joint probability rather than conditional probability. When analysing the conditional probability of 'distraction', *severe casualty*, and 'hitting an obstacle on the road', we see that distraction shows a share of 50 % in the corresponding subset. 'Speeding' is associated with 'hitting a guard rail' (joint probability of 0,24 %), 'hitting a tree' (joint probability of 0,64 %), and 'fatigue' (joint probability of 0,03 %). 'Hit and run' shows a relationship with 'alcohol' (joint probability of 0,10 %). The driving manoeuvre 'sudden braking' links to 'skidding/drifting' (joint probability of 0,03 %). The network also associates 'skidding/drifting' with 'speeding' (joint probability of 0,52 %). Age classes '16 to 18' and '19 to 24' show a directed graph to

'probationary driving licence' with a joint probability of 0,45 % for age class '16 to 18' and 1,00 % for age class '19 to 24'. Looking at the network, we can see differences among female and male drivers: female drivers are associated with 'no safety belt applied' (joint probability of 0,52 %), whereas male drivers are associated with 'fatigue' (joint probability of 1,19 %) and 'alcohol' (joint probability of 2,08 %). However, figure 31 illustrates that male and female drivers hold more or less the same share for 'no safety belt applied'.

For traffic safety, the driver-related Bayesian network reveals interesting and logical relationships. For example, the network reveals different associations among female and male drivers. Detecting differences among both sex groups is essential for target-specific traffic safety work. For example, there exists an association between young drivers and 'probationary driving licence' and *severe casualties*. Chapter 8 will investigate variable combinations with young drivers and 'probationary driving licence' in detail. The illustrated network also suggests that driving behaviour such as 'speeding' leads to 'skidding/drifting', 'hitting a guard rail', 'hitting a tree', and, interestingly, 'fatigue'.

However, since the Bayesian network represents a condensed picture of possible variable combinations, many more variable combinations might exist than now detected in the Bayesian network. Therefore, we conduct a detailed analysis of all observed variable combinations and count their frequencies in chapter 8.

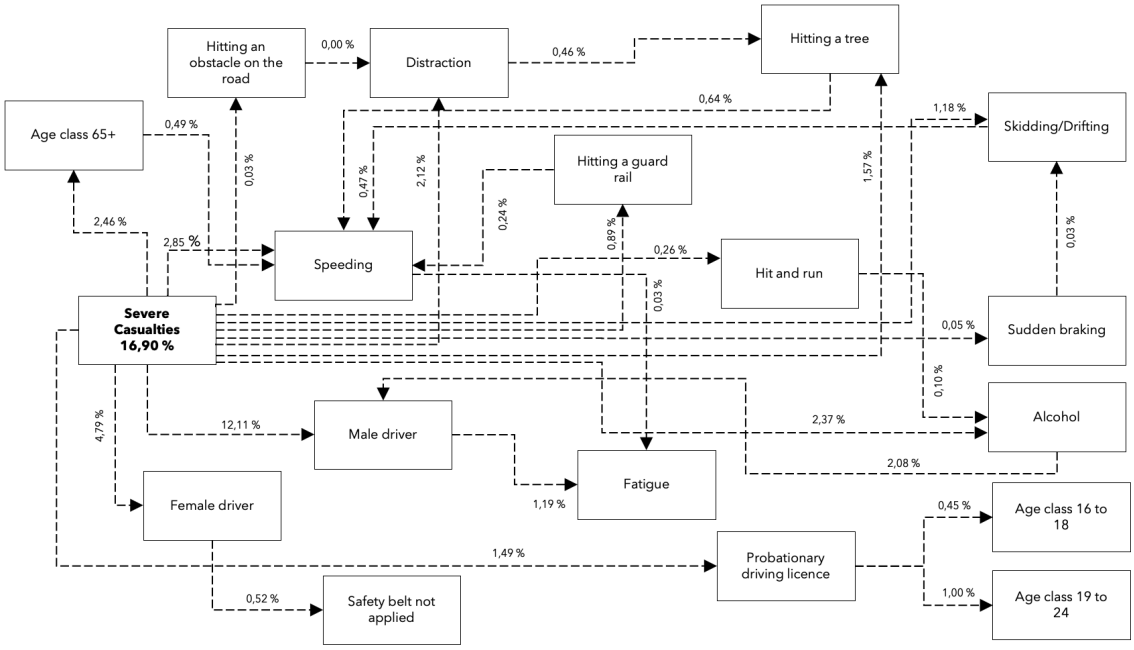

Figure 50: Driver-related Bayesian network. The network illustrates driver-related characteristics and their joint probabilities [%]. The root node *severe casualties* shows a probability or relative frequency of 16,90 % (3.430 of 20.293 single-vehicle accidents with a single occupation that occurred on the Austrian road network outside the built-up area between 2012 and 2019).

Another way of illustrating Bayesian networks is a tabular representation shown in table 57.

Predictor variable *P (SC)*		*First node (N1)* *P (SC ∩ N1)*		*Second node (N2)* *P (SC ∩ N1 ∩ N2)*
severe casualties	⟶	alcohol	⟶	male driver
16,90 %		2,37 %		2,08 %
severe casualties	⟶	male driver	⟶	fatigue
16,90 %		12,11 %		1,19 %
severe casualties	⟶	probationary driving licence	⟶	age class 19 to 24
16,90 %		1,49 %		1,00 %
severe casualties	⟶	hitting a tree	⟶	speeding
16,90 %		1,57 %		0,64 %
severe casualties	⟶	no safety belt applied	⟶	female driver
16,90 %		3,44 %		0,52 %
severe casualties	⟶	age class 65+	⟶	speeding
16,90 %		2,46 %		0,49 %
severe casualties	⟶	skidding/drifting	⟶	speeding
16,90 %		1,18 %		0,47 %
severe casualties	⟶	distraction	⟶	hitting a tree
16,90 %		2,12 %		0,46 %
severe casualties	⟶	probationary driving licence	⟶	age class 16 to 18
16,90 %		1,49 %		0,45 %
severe casualties	⟶	hitting the guard rail	⟶	speeding
16,90 %		0,89 %		0,24 %
severe casualties	⟶	hit and run	⟶	alcohol
16,90 %		0,26 %		0,10 %
severe casualties	⟶	sudden braking	⟶	skidding drifting
16,90 %		0,05 %		0,03 %
severe casualties	⟶	speeding	⟶	fatigue
16,90 %		2,85 %		0,03 %
severe casualties	⟶	hitting an obstacle on the road	⟶	distraction
16,90 %		0,03 %		0,00 %

Table 57: Tabular illustration of the driver-related Bayesian network. The table illustrates detected relationships among driver-related characteristics and their joint probabilities [%]. The network is based on 20.293 single-vehicle accidents with a single occupation that occurred on the Austrian road network outside the built-up area between 2012 and 2019.

7.3 Bayesian network of vehicle-related variables

The analysis of vehicle-related characteristics results in integrating 16 vehicle-related characteristics into the Bayesian network. These characteristics represent dummy variables, and the target variable is *severe casualties*. Table 58 shows the role of vehicle-related variables for the Bayesian network generation process.

Target variable	*Predictor variables*	
Degree of injury: • severe casualties	*Engine power (kW):* • *24-90* • *90-110* • *110+* *Kilometrage (km):* • *15.000-75.000* • *150.000-200.00* *Technical settings:* • airbag not deployed • insufficient vehicle security • vehicle fire	*Vehicle colour:* • *blue* • *brown* • *grey* • *green* • *red* • *black* • *silver* • *white*

Table 58: Role of variables within the vehicle-related Bayesian network. The network includes 16 vehicle-related characteristics as dummy variables (0=characteristics not present, 1=characteristic is present) and the dichotomous target variable severe casualty (0=no severe casualty, 1=severe casualty).

The vehicle-related Bayesian network uses a TAN structure and Bayes adjustment for small cell counts for parameter learning. The resulting network (see figure 51) shows a direct graph from 'insufficient vehicle safety' to 'airbag not deployed'.

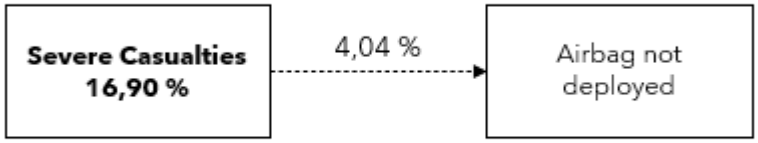

Figure 51: Vehicle-related Bayesian network. The network illustrates vehicle-related characteristics and their joint probabilities [%]. The root node *severe casualties* shows a probability or relative frequency of 16,90 % (3.430 of 20.293 single-vehicle accidents with a single occupation that occurred on the Austrian road network outside the built-up area between 2012 and 2019).

Table 59 shows the vehicle-related Bayesian network.

Predictor variable		*First node (N1)*
P (SC)		*P (SC ∩ N1)*
severe casualties	→	airbag not deployed
16,90 %		4,04 %

Table 59: Tabular illustration of the vehicle-related Bayesian network. The table illustrates detected relationships among vehicle-related characteristics and their joint probabilities [%]. The network is based on 20.293 single-vehicle accidents with a single occupation that occurred on the Austrian road network outside the built-up area between 2012 and 2019.

7.4 Bayesian network of roadway-related variables

The analysis of roadway-related characteristics integrates 17 characteristics into the Bayesian network. These characteristics represent dummy variables, and the target variable is *severe casualties*.

Table 60 shows the role of variables for the roadway-related Bayesian network generation.

Target variable	*Predictor variables*	
Degree of injury: • severe casualties	*Speed limit (km/h):* • driving ban • 50 • 70 • 80 • 100 • 130 *Road type:* • highway • expressway • country Road • other roads	*Road characteristics:* • intersection • curve • middle separation *Road surface condition:* • dry road • wet road • wintry conditions • sand or grit on the road

Table 60: Role of variables within the roadway-related Bayesian network. The network includes 17 roadway-related characteristics as dummy variables (0=characteristics not present, 1=characteristic is present) and the dichotomous target variable severe casualty (0=no severe casualty, 1=severe casualty).

We generate the Bayesian network using the TAN structure and Bayes adjustment for small cell counts for parameter learning. 'Wet roads' show a relation with curves, resulting in a joint probability of 1,78 %. 'Curves' are associated with 'highway' with a joint probability of 0,25 %. 'Highways' show directed graphs to 'middle separation' (joint probability of 0,43 %) and 'speed limit of 130 km/h' (joint probability of 1,32 %). *Severe casualties*, 'wintry conditions', and 'wet roads' result in a joint probability of 1,82 %. The network also detects a relation among 'intersections' and 'curves' with a joint probability of 0,02 %. The roadway-related Bayesian network suggests 'wet roads' to impact *severe casualties*. Also, three combinations exist that include 'highways' rather than the other road types (expressway, country road, or other roads).

On the other hand, in chapter 4.8, we could see that country roads are much higher in frequency among *severe casualties*. Chapters 8 and 9 will analyse road types in detail.

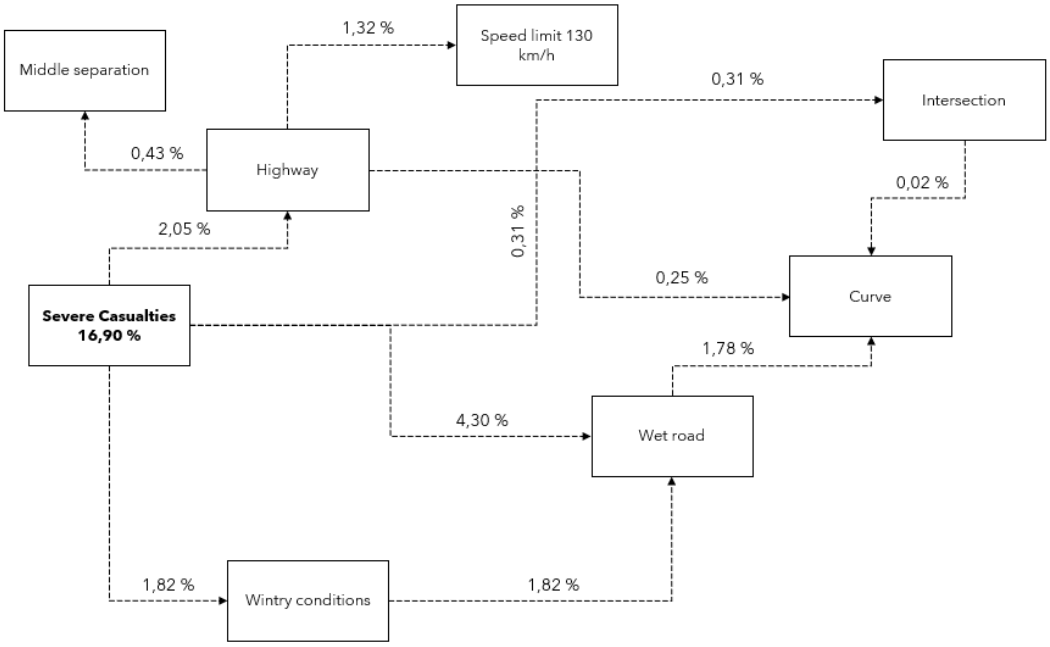

Figure 52: Roadway-related Bayesian network. The network illustrates roadway-related characteristics and their joint probabilities [%]. The root node severe casualties shows a probability or relative frequency of 16,90 % (3.430 of 20.293 single-vehicle accidents with a single occupation that occurred on the Austrian road network outside the built-up area between 2012 and 2019).

To have a more precise overview of the ranking of joint probabilities for each combination, we generate a tabular illustration of the roadway-related Bayesian network in table 61. The tabular network representation makes it easier to see that the triple combination of *severe casualties*, 'wintry conditions', and 'wet roads' result in a higher joint probability than the combination of *severe casualties*, 'wet road' and 'curve'.

Predictor variable *P (SC)*		*First node (N1)* *P (SC ∩ N1)*		*Second node (N2)* *P (SC ∩ N1 ∩ N2)*
severe casualties	→	wintry conditions	→	wet road
16,90 %		1,82 %		1,82 %
severe casualties	→	wet road	→	curve
16,90 %		4,30 %		1,78 %
severe casualties	→	highway	→	speed limit 130 km/h
16,90 %		2,05 %		1,32 %
severe casualties	→	highway	→	middle separation
16,90 %		2,05 %		0,43 %
severe casualties	→	highway	→	curve
16,90 %		2,05 %		0,25 %
severe casualties	→	intersection	→	curve
16,90 %		0,31 %		0,02 %

Table 61: Tabular illustration of the roadway-related Bayesian network. The table illustrates detected relationships among roadway-related characteristics and their joint probabilities [%]. The network is based on 20.293 single-vehicle accidents with a single occupation that occurred on the Austrian road network outside the built-up area between 2012 and 2019.

7.5 Bayesian network of situation-related variables

The analysis of situation-related characteristics integrates 14 characteristics into the Bayesian network. These characteristics represent dummy variables, and the target variable is *severe casualties*. Table 62 illustrates the role of variables to be integrated into the situation-related Bayesian network.

Target variable	*Predictor variables*	*Predictor variables*
Degree of injury: • severe casualties	*Daytime:* • 0 a.m. to 6 a.m. • 6 a.m. to 12 p.m. • 12 p.m. to 6. p.m. • 6 p.m. to 0 a.m. *Weekday:* • Monday to Thursday • Friday to Sunday *Season:* • spring • summer • autumn • winter	*Weather conditions:* • rain • snow *Light conditions:* • darkness • dusk or dawn

Table 62: Role of variables within the situation-related Bayesian network. The network includes 14 situation-related characteristics as dummy variables (0=characteristics not present, 1=characteristic is present) and the dichotomous target variable severe casualty (0=no severe casualty, 1=severe casualty).

The situation-related Bayesian network represents a TAN network and uses Bayes adjustment for small cell counts for parameter learning. The network shows relations between *severe casualties*, 'winter', and 'rain' (joint probability of 0,33 %). Also, it shows a relation between *severe casualties*, 'winter', and 'snow' (joint probability of 0,64 %). Considering the roadway-related Bayesian network where we could see a relation between *severe casualties*, 'wet road', and 'wintry conditions', we can assume that weather and wet road conditions have a considerable impact on *severe casualties*. Additionally, the network detects a relationship between 'summer' and 'darkness' (joint probability of 1,16 %) and weekdays from Friday to Sunday combined with time between '0 a.m. and 6 a.m.' (joint probability of 2,09 %).

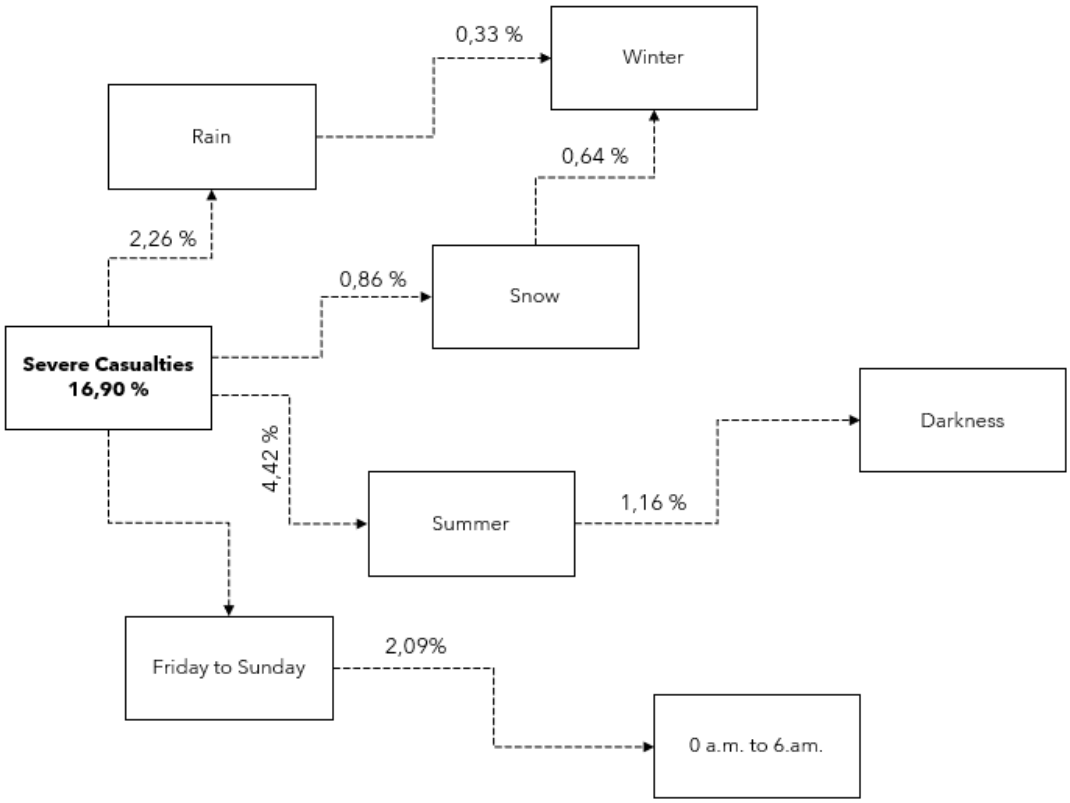

Figure 53: Situation-related Bayesian network. The network illustrates situation-related characteristics and their joint probabilities [%]. The root node *severe casualties* shows a probability or relative frequency of 16,90 % (3.430 of 20.293 single-vehicle accidents with a single occupation that occurred on the Austrian road network outside the built-up area between 2012 and 2019).

Predictor variable *P (SC)*		*First node (N1)* *P (SC ∩ N1)*		*Second node (N2)* *P (SC ∩ N1 ∩ N2)*
severe casualties	→	Friday to Sunday	→	0 a.m. to 6 a.m.
16,90 %		8,10 %		2,09 %
severe casualties	→	summer	→	darkness
16,90 %		4,42 %		1,16 %
severe casualties	→	snow	→	winter
16,90 %		0,86 %		0,64 %
severe casualties	→	rain	→	winter
16,90 %		2,26 %		0,33 %

Table 63: Tabular illustration of the situation-related Bayesian network. The table illustrates detected relationships among roadway-related characteristics and their joint probabilities [%]. The network is based on 20.293 single-vehicle accidents with a single occupation that occurred on the Austrian road network outside the built-up area between 2012 and 2019.

7.6 Bayesian network of all accident-related variables

We integrate 52 characteristics from the four categories driver, vehicle, roadway, and situation as dummy variables into the overall Bayesian network. Table 64 presents the role of the characteristics within the overall Bayesian network.

Category	*Variable*	*Characteristics*
Driver	Sex	male, female
	Age class	16 to 18, 19 to 24, 25 to 34, 35 to 44, 45 to 54, 55 to 64, 65+
	Driving licence type	probationary driving licence
	Distraction	alcohol, distraction, fatigue
	Driving manoeuvre	speeding, skidding/drifting, hitting a tree, hitting an obstacle on the road, hitting a guard rail, sudden braking, hit and run
	Safety settings	no safety belt not applied
Vehicle	Vehicle settings	airbag not deployed
Roadway	Speed limit	driving ban, 50, 70, 80, 100, 130
	Road characteristics	intersection, curve, middle separation
	Road condition	wet road, wintry conditions, sand or grit on the road
	Road type	highway, expressway, regional road, other roads
Situation	Weather conditions	rain, snow
	Light conditions	darkness, dusk or dawn
	Time:	0 to 6, 6 to 12, 12 to 6,6 to 0
	Weekday	Monday to Thursday, Friday to Sunday
	Season	winter, spring, summer, autumn

Table 64: Role of variables within the overall Bayesian network. The network includes 69 accident-describing characteristics as dummy variables (0=characteristics not present, 1=characteristic is present) and the dichotomous target variable severe casualty (0=no severe casualty, 1=severe casualty).

The overall network represents a TAN network and uses Bayes adjustment for small cell counts for parameter learning. This network becomes too crowded for graphical visualisation, so table 65 presents the Bayesian network with variable combinations ranked by joint probability. On the one hand, the network illustrates relations we have already seen in the previous chapters (e.g., *severe casualties*, 'male drivers', and 'alcohol' with a joint probability of 2,08 %). On the other hand, the network extends the knowledge about blackpatterns.

The situation-related Bayesian network detects a relationship between *severe casualties*, weekdays Monday to Friday and between '0 a.m. and 6 a.m.' with a joint probability of 2,09 %. Also, the overall Bayesian network shows a relationship between *severe casualties*, weekdays Friday to Sunday and 'alcohol' with a joint probability of 1,42 %. Subsequently, the network detects a relationship between 'alcohol' and 'darkness'. Therefore, we could articulate the assumption that nights during the weekend increase the number of drivers impaired by alcohol and thus the risk of *severe casualties*. Our logistics regression model (see chapter 5) also suggests that the weekdays from Monday to Thursday reduce the risk of *severe casualties*.

The overall network substantiates the impact of 'winter', 'wintry conditions', and 'wet roads' to increase the risk of *severe casualties* (as we have already seen in the situation-related Bayesian network). Table 65 now illustrates the overall Bayesian network.

Predictor variable		*First node (N1)*		*Second node (N2)*
P (SC)		*P (SC ∩ N1)*		*P (SC ∩ N1 ∩ N2)*
severe casualties	→	rain	→	wet road
16,90 %		2,26 %		2,15 %
severe casualties	→	Friday to Sunday	→	0 a.m. to 6 a.m.
16,90 %		8,10 %		2,09 %
severe casualties	→	alcohol	→	male driver
16,90 %		2,37 %		2,08 %
severe casualties	→	wintry conditions	→	wet road
16,90 %		1,82 %		1,82 %
severe casualties	→	curve	→	speeding
16,90 %		6,23 %		1,52 %

Table 65: Tabular illustration of the overall Bayesian network. The table illustrates detected relationships among accidents describing characteristics and joint probabilities [%]. The network is based on 20.293 single-vehicle accidents with a single occupation that occurred on the Austrian road network outside the built-up area between 2012 and 2019.

Predictor variable *P (SC)*		*First node (N1)* *P (SC ∩ N1)*		*Second node (N2)* *P (SC ∩ N1 ∩ N2)*
severe casualties	→	alcohol	→	darkness
16,90 %		2,37 %		1,51 %
severe casualties	→	Friday to Sunday	→	alcohol
16,90 %		8,10 %		1,42 %
severe casualties	→	highway	→	speed limit 130 km/h
16,90 %		2,05 %		1,32 %
severe casualties	→	probationary driving licence	→	age class 19 to 24
16,90 %		1,49 %		1,00 %
severe casualties	→	wintry conditions	→	female driver
16,90 %		1,82 %		0,75 %
severe casualties	→	snow	→	wintry conditions
16,90 %		0,86 %		0,72 %
severe casualties	→	hitting a tree	→	speeding
16,90 %		1,57 %		0,64 %
severe casualties	→	no safety belt applied	→	female driver
16,90 %		3,44 %		0,52 %
severe casualties	→	speeding	→	wintry conditions
16,90 %		2,85 %		0,49 %
severe casualties	→	skidding/drifting	→	speeding
16,90 %		1,18 %		0,47 %
severe casualties	→	probationary driving licence	→	age class 16 to 18
16,90 %		1,49 %		0,45 %
severe casualties	→	highway	→	middle separation
16,90 %		2,05 %		0,43 %
severe casualties	→	highway	→	curve
16,90 %		2,05 %		0,25 %

Continuation of table 65: Tabular illustration of the overall Bayesian network. The table illustrates detected relationships among accidents describing characteristics and joint probabilities [%]. The network is based on 20.293 single-vehicle accidents with a single occupation that occurred on the Austrian road network outside the built-up area between 2012 and 2019.

Predictor variable P (SC)		First node (N1) P (SC ∩ N1)		Second node (N2) P (SC ∩ N1 ∩ N2)
severe casualties	→	hitting a guard rail	→	middle separation
16,90 %		0,89 %		0,22 %
severe casualties	→	distraction	→	middle separation
16,90 %		2,12 %		0,21 %
severe casualties	→	hit and run	→	alcohol
16,90 %		0,26 %		0,10 %
severe casualties	→	wintry conditions	→	fatigue
16,90 %		1,82 %		0,09 %
severe casualties	→	sudden braking	→	skidding drifting
16,90 %		0,05 %		0,03 %
severe casualties	→	intersection	→	curve
16,90 %		0,31 %		0,02 %
severe casualties	→	hitting an obstacle on the road	→	middle separation
16,90 %		0,03 %		0,01 %

Continuation of table 65: Tabular illustration of the overall Bayesian network. The table illustrates detected relationships among accidents describing characteristics and joint probabilities [%]. The network is based on 20.293 single-vehicle accidents with a single occupation that occurred on the Austrian road network outside the built-up area between 2012 and 2019.

8. Road traffic accident data analysis V: Pattern recognition based on frequencies of variable combinations

Based on the maximum combination value we demonstrated in chapter 4, we now analyse the most frequent variable combinations (patterns) among the four categories driver, vehicle', roadway and situation. The calculation of the maximum combination value is based on a developed aggregation method (the PATTERMAX-method) that searches for identical variable combinations (blackpatterns) within the binary-structured road traffic accident database and counts their frequencies. Thus, we aggregate identical sequences of zeros and ones among our newly established binary road traffic accident database. A blackpattern is a variable combination that occurs more than ten times and includes two accident-describing characteristics. The maximum combination value indicates how often a specific variable combination (blackpattern) occurs within the historical database. Therefore, the blackpatterns shown in this chapter represent truly observed patterns. As we count the frequency of each detected blackpattern, we can calculate its relative frequency or joint probability.

8.1 Blackpatterns among driver-related variables

The blackpattern detection among driver-related variables works with 54 driver-related characteristics. The driver-related characteristics represent dummy variables. Thus, the PATERRMAX-method searches for identical sequences of zeros and ones within the characteristics and counts the frequencies of identical combinations. The advantage of this method is to identify and quantify patterns that were indeed observed between 2012 and 2019. Also, the method does not assume the relationship among the variables. As the method results in many detected blackpatterns, we only illustrate selected blackpatterns in this chapter. Before we do so, we explain how to read the following blackpattern tables.

How to read the blackpattern tables

The left side of the table shows the most frequent blackpatterns (variable combinations) among the entire road traffic accident sample (n=20.293). The right side of the table shows the most frequent blackpatterns (variable combinations) among *severe casualties* (n=3.431). Thus, the right represents a subset of the left side. The idea behind it is to see whether there are different blackpatterns within *severe casualties* compared to all casualties. The column 'Position' in the

middle of the table shows the numbers one to ten. One refers to the most frequently observed blackpattern. Consequently, the table shows the ten most common blackpatterns among the selected characteristics. The column 'Count' indicates how often the blackpattern occurred within our observation period (2012-2019).

We start with the most frequently observed blackpatterns among male and female drivers. Table 66 illustrates the ten most frequent blackpatterns for female drivers (ranked by the score in the middle of the table).

Count	Blackpatterns including female drivers and all casualties	Position	Blackpatterns including female drivers and severe casualties only	Count
1.132	female, safety belt applied, age class 19 to 24	**1**	female, severe casualty, safety belt applied, age class 19 to 24	92
830	female, safety belt applied, age class 25 to 34	**2**	female, severe casualty, safety belt applied, age class 25 to 34	62
432	female, safety belt applied, age class 35 to 44	**3**	female, severe casualty, safety belt applied, age class 45 to 54	56
391	female, safety belt applied, age class 19 to 24, **probationary driving licence**	**4**	female, severe casualty, safety belt applied, age class 65+	55
382	female, safety belt applied, age class 45 to 54	**5**	female, severe casualty, safety belt applied, age class 35 to 44	44
235	female, safety belt applied, age class 65+	**6**	female, severe casualty, safety belt applied, age class 55 to 64	40
212	female, safety belt applied, age class 55 to 64	**7**	female, severe casualty, safety belt applied, age class 19 to 24, **probationary driving licence**	32
162	female, safety belt applied, age class 16 to 18, **probationary driving licence**	**8**	female, severe casualty, safety belt applied, age class 16 to 18, **probationary driving licence**	19
154	female, safety belt applied, age class 16 to 18	**9**	female, severe casualty, safety belt applied, age class 65+, **fatigue**	14
131	female, safety belt applied, age class 19 to 24, **speeding**	**10**	female, severe casualty, safety belt applied, age class 19 to 24, **speeding**	11

Table 66: The ten most frequent blackpatterns among female drivers and other driver-related characteristics. The left side of the table represents the most frequent blackpatterns among the entire road traffic accident sample. The right side represents the most frequent blackpatterns among *severe casualties*. The right side is a subset of the left side. The column 'Count' indicates how often the blackpattern occurred within our observation period (2012-2019). n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

The ten most frequent combinations with 'female drivers' vary in age class but clearly show younger age classes at the top. However, the age class '65+' appears in the third position among *severe casualties*. All ten combinations include the characteristic 'safety belt applied'. The variable 'probationary driving licence' appears in the fourth and eighth most frequent combination among all casualties and the seventh and eighth most frequent combination among *severe casualties*. Interestingly, both sides include the variable 'speeding' in the tenth position for the same age class. Among *severe casualties*, we can see the variable 'fatigue' within the ninth most frequent combination, including age 'group 65+'. In addition to the detected patterns among female drivers, table 67 shows the same blackpattern scheme for 'male drivers'.

Count	Blackpatterns including female drivers and all casualties	Position	Blackpatterns including female drivers and severe casualties only	Count
817	male, safety belt applied, age class 19 to 24	**1**	male, severe casualty, safety belt applied, age class 25 to 34	149
689	male, safety belt applied, age class 25 to 34	**2**	male, severe casualty, safety belt applied, age class 19 to 24	137
377	male, safety belt applied, 35 to 44	**3**	male, severe casualty, safety belt applied, age class 65+	120
341	male, safety belt applied, age class 19 to 24, **probationary driving licence**	**4**	male, severe casualty, safety belt applied, age class 45 to 54	105
303	male, safety belt applied, age class 65+	**5**	male, severe casualty, safety belt applied, age class 35 to 44	85
281	male, safety belt applied, age class 45 to 54	**6**	male, severe casualty, **no safety belt applied**, age class 25 to 34	60
246	male, safety belt applied, age class 25 to 34, **alcohol**	**7**	male, severe casualty, safety belt applied, age class 55 to 64	58
199	male, safety belt applied, age class 19 to 24, **alcohol**	**8**	male, severe casualty, **no safety belt applied**, age class 19 to 24	45
189	male, safety belt applied, age class 55 to 64	**9**	male, severe casualty, safety belt applied, age class 25 to 34, **alcohol**	42
171	male, safety belt applied, age class 16 to 18, **probationary driving licence**	**10**	male, severe casualty, **no safety belt applied**, age class 45 to 54	39

Table 67: The ten most frequent driver-related variable combinations for male drivers. The left side of the table represents the most frequent blackpatterns among the entire road traffic accident sample. The right side represents the most frequent blackpatterns among *severe casualties*. The right side is a subset of the left side. The column 'Count' indicates how often the blackpattern occurred within our observation period (2012-2019). n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

Regarding the variables 'safety belt applied' and age classes, 'male drivers' show almost the same combinations as female drivers among all casualties. However, among *severe casualties* and 'male drivers', the variable 'no safety belt applied' occurs in the sixth, eighth and tenth place. In contrast to female drivers, the most frequent combinations among 'male drivers' include 'alcohol'. Also, the variables 'speeding' and 'fatigue' do not appear among the top ten combinations for male drivers (neither among the total sample nor among *severe casualties*). 'Probationary driving licence' appears within the fourth and tenth most frequent combination among all casualties, which is the case among female drivers in the fourth most frequent combination.

Comparing patterns among female and male drivers reveals gender-specific differences in impairment and driving behaviour. 'Probationary driving licence', 'speeding' and 'fatigue' appear among 'female drivers' and 'probationary driving licence' and 'alcohol' appear among 'male drivers'.

The tables above show that all age classes appear among the most frequent driver-related combinations. Therefore, we will not proceed with an in-depth investigation of patterns among different age classes but with patterns including the variable 'probationary driving licence'. Chapter 4 illustrated that the variable 'probationary driving licence' has a significant relationship with *severe casualties*. Because this variable is associated with younger age classes, we have a detailed look at the most frequent combinations including 'probationary driving licence' (see table 68).

Count	Blackpatterns including probationary driving licence and all casualties	Position	Blackpatterns including probationary driving licence and severe casualties only	Count
391	probationary driving licence, safety belt applied, female, age class 19 to 24	**1**	probationary driving licence, severe casualty, safety belt applied, male, age class 19 to 24	37
341	probationary driving licence, safety belt applied, male, age class 19 to 24	**2**	probationary driving licence, severe casualty, safety belt applied, female, age class 19 to 24	32
171	probationary driving licence, safety belt applied, male, age class 16 to 18	**3**	probationary driving licence, severe casualty, safety belt applied, female, age class 16 to 18	19
162	probationary driving licence, safety belt applied, female, age class 16 to 18	**4**	probationary driving licence, severe casualty, safety belt applied, male, age class 16 to 18	18
61	probationary driving licence, safety belt applied, male, age class 19 to 24, **alcohol**	**5**	probationary driving licence, severe casualty, safety belt applied, male, age class 19 to 24, **alcohol**	12
58	probationary driving licence, safety belt applied, female, age class 19 to 24, **speeding**	**6**	probationary driving licence, severe casualty, **no safety belt applied**, male, age class 19 to 24	9
46	probationary driving licence, safety belt applied, female, age class 16 to 18, **distraction**	**7**	probationary driving licence, severe casualty, safety belt applied, male, age class 19 to 24, **speeding**	7
43	probationary driving licence, safety belt applied, female, age class 19 to 24, **distraction**	**8**	probationary driving licence, severe casualty, safety belt applied, female, age class 19 to 24, **distraction**	6
43	probationary driving licence safety belt applied, male, age class 19 to 24, **speeding**	**9**	probationary driving licence, severe casualty, safety belt applied, male age, class19 to 24, **fatigue**	5
38	probationary driving licence, safety belt applied, female, age class 19 to 24, **skidding**	**10**	probationary driving licence, severe casualty, safety belt applied, male, age class 19 to 24, **distraction**	5

Table 68: The ten most frequent driver-related variable combinations including 'probationary driver's licence'. The left side of the table represents the most frequent blackpatterns among the entire road traffic accident sample. The right side represents the most frequent blackpatterns among *severe casualties*. The right side is a subset of the left side. The column 'Count' indicates how often the blackpattern occurred within our observation period (2012-2019). n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

Among the investigated road traffic accident sample, the characteristic 'probationary driving licence' occurs in combination with 'alcohol', 'speeding', 'distraction' and 'skidding/drifting'. In the event of *severe casualties*, the variable 'probationary driving licence' occurs in combination with 'alcohol', 'no safety belt applied', 'speeding', 'distraction' and 'fatigue'. The variable applies to age classes '16 to 18' and '19 to 24'.

The Austrian Road Safety Strategy declares 'alcohol' and 'distraction' as central challenges for reducing severe road traffic accidents. For the examined sample of single-vehicle accidents with single-occupancy, 'fatigue' represents an additional variable to be considered, especially among younger drivers. Since all three variables appear to play a relevant role in the examined sample, the following tables illustrate the top ten variable combinations for these three types of impairment (see table 69 to table 71).

Count	Blackpatterns including alcohol and all casualties	Position	Blackpatterns including alcohol and severe casualties only	Count
246	alcohol, safety belt applied, **male**, age class 25 to 34	**1**	alcohol, severe casualty, safety belt applied, **male**, age class 25 to 34	42
199	alcohol, safety belt applied, **male**, age class 19 to 24	**2**	alcohol, severe casualty, safety belt applied, **male**, age class 35 to 44	29
135	alcohol, safety belt applied, **male**, age class 35 to 44	**3**	alcohol, severe casualty, safety belt applied, **male**, age class 19 to 24	21
114	alcohol, safety belt applied, **male**, age class 45 to 54	**4**	alcohol, severe casualty, safety belt applied, **male**, age class 45 to 54	20
62	alcohol, safety belt applied, **male**, age class 55 to 64	**5**	alcohol, severe casualty, **no safety belt applied**, **male**, age class 25 to 34	13
61	alcohol, safety belt applied, **male**, age class 19 to 24, **probationary driving licence**	**6**	alcohol, severe casualty, safety belt applied, **male**, age class 19 to 24, **probationary driving licence**	12
35	alcohol, safety belt applied, **male**, age class 65+	**7**	alcohol, severe casualty, safety belt applied, **male**, age class 55 to 64	11
35	alcohol, safety belt applied, female, age class 35 to 44	**8**	alcohol, severe casualty, safety belt applied, **male**, age class 65+	10
34	alcohol, **no safety belt applied**, **male**, age class 25 to 34	**9**	alcohol, severe casualty, safety belt applied, female, age class 45 to 54	8
33	alcohol, safety belt applied, female, age class 45 to 54	**10**	alcohol, severe casualty, **no safety belt applied**, **male**, age class 45 to 54	8

Table 69: The ten frequent driver-related variable combinations including 'alcohol'. The left side of the table represents the most frequent blackpatterns among the entire road traffic accident sample. The right side represents the most frequent blackpatterns among *severe casualties*. The right side is a subset of the left side. The column 'Count' indicates how often the blackpattern occurred within our observation period (2012-2019). n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

Table 69 implies that the most frequent combinations with 'alcohol' impairment apply almost exclusively to 'male drivers'. The joint probability of a *severe casualty*, 'alcohol' impairment and 'male driver' is 2 % (total: 423). In contrast, the joint probability of a severe or fatal accident, impairment by alcohol and female driver is 0,2 % (total: 58). 'Probationary driving licence' and 'no safety belt applied' occur within both categories.

Table 70 continues with the most frequent combinations with 'distraction'.

Count	Blackpatterns including distraction and all casualties	Position	Blackpatterns including distraction and severe casualties	Count
93	distraction, safety belt applied, female, age class 19 to 24	**1**	distraction, severe casualty, safety belt applied, male, age class 65+	16
55	distraction, safety belt applied, female, age class 25 to 34	**2**	distraction, severe casualty, safety belt applied, male, age class 45 to 54	13
51	distraction, safety belt applied, male, age class 65+	**3**	distraction, severe casualty, safety belt applied, male, age class 25 to 34	12
50	distraction, safety belt applied, male, age class 19 to 24	**4**	distraction, severe casualty, safety belt applied, female, age class 65+	10
46	distraction, safety belt applied, female, age class 16 to 18, **probationary driving licence**	**5**	distraction, severe casualty, safety belt applied, female, age class 19 to 24	9
45	distraction, safety belt applied, male, age class 25 to 34	**6**	distraction, severe casualty, safety belt applied, male, age class 65+	7
43	distraction, safety belt applied, female, age class 19 to 24, **probationary driving licence**	**7**	distraction, severe casualty, safety belt applied, female, age class 25 to 34, **hitting obstacle next to the road**	7
41	distraction, safety belt applied, female, age class 65+	**8**	distraction, severe casualty, safety belt applied, male, age class 10 to 24	7
34	distraction, safety belt applied, male, age class 19 to 24, **probationary driving licence**	**9**	distraction, severe casualty, safety belt applied, female, age class 19 to 24, **probationary driving licence**	6
30	distraction, safety belt applied, female, age class 45 to 54	**10**	distraction, severe casualty, safety belt applied, male, age class 19 to 24, **hitting a tree**	6

Table 70: The ten most frequent driver-related variable combinations including 'distraction'. The left side of the table represents the most frequent blackpatterns among the entire road traffic accident sample. The right side represents the most frequent blackpatterns among *severe casualties*. The right side is a subset of the left side. The column 'Count' indicates how often the blackpattern occurred within our observation period (2012-2019). n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

Figure 31 shows that among *severe casualties*, 'female drivers' hold a higher share of 'distraction' than 'male drivers'. Interestingly, among *severe casualties*, the variables 'distraction' and 'hitting an obstacle next to road' appear in seventh place and the variables 'distraction' and 'hitting a tree' in tenth place. Both associations were also identified by the Bayesian networks (chapter 7.2). One advantage of the developed PATTERMAX-method is identifying logical links within the datasets. The ability of this method to do so becomes evident with this example.

Table 71 shows the results for the third most frequent type of impairment, 'fatigue'.

Count	Blackpatterns including fatigue and all casualties	Position	Blackpatterns including fatigue and severe casualties only	Count
134	fatigue, safety belt applied, male, age class 19 to 24	**1**	fatigue, severe casualty, safety belt applied, male, age class 65+	26
86	fatigue, safety belt applied, male, age class 25 to 34	**2**	fatigue, severe casualty, safety belt applied, male, age class 19 to 24	22
82	fatigue, safety belt applied, male, age class 65+	**3**	fatigue, severe casualty, safety belt applied, male, age class 25 to 34	20
59	fatigue, safety belt applied, male, age class 35 to 44	**4**	fatigue, severe casualty, safety belt applied, male, age class 55 to 64	19
49	fatigue, safety belt applied, male, age class 45 to 54	**5**	fatigue, severe casualty, safety belt applied, male, age class 35 to 44	19
42	fatigue, safety belt applied, male, age class 55 to 64	**6**	fatigue, severe casualty, safety belt applied, female, age class 65+	14
38	fatigue, safety belt applied, female, age class 19 to 24	**7**	fatigue, severe casualty, safety belt applied, male, age class 45 to 54	10
37	fatigue, safety belt applied, female, age class 65+	**8**	fatigue, severe casualty, safety belt applied, female, age class 55 to 64	9
35	fatigue, safety belt applied, male, age class 19 to 24, **probationary driving licence**	**9**	fatigue, severe casualty, safety belt applied, female, age class 19 to 24	8
35	fatigue, safety belt applied, female, age class 45 to 54	**10**	fatigue, severe casualty, safety belt applied, female, age class 25 to 34	6

Table 71: The ten most frequent driver-related variable combinations including 'fatigue'. The left side of the table represents the most frequent blackpatterns among the entire road traffic accident sample. The right side represents the most frequent blackpatterns among *severe casualties*. The right side is a subset of the left side. The column 'Count' indicates how often the blackpattern occurred within our observation period (2012-2019). n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

The combination of 'fatigue' and male drivers turns out to be more common than the combination of 'fatigue' and female drivers. As shown in figure 31, *severe casualties* and 'male drivers' hold a higher share of 'fatigue' than *severe casualties* and 'female drivers'. Young 'male drivers' (age classes '19 to 24' and '25 to 34') appear on top when analysing all casualties, whereas 'male drivers' over 65 appear on top when analysing *severe casualties*. Besides differences in age classes, no special variable is combined with 'fatigue', except for 'probationary driving licence' in ninth place among all casualties.

We continue with the detailed investigation of driver-related variables and illustrate the most frequent combinations regarding driving manoeuvres. Chapter 4 revealed that the variables 'speeding' and 'skidding/drifting' are associated with a high maximum combination value. Table 72 illustrates the most frequent variable combinations with 'speeding'.

Count	Blackpatterns including speeding and all casualties	Position	Blackpatterns including speeding and severe casualties only	Count
131	speeding, safety belt applied, female, age class 19 to 24	**1**	speeding, severe casualty, safety belt applied, male, age class 25 to 34	16
100	speeding, safety belt applied, male, age class 19 to 24	**2**	speeding, severe casualty, safety belt applied, male, age class 45 to 54	12
86	speeding, safety belt applied, female, age class 25 to 34	**3**	speeding, severe casualty, safety belt applied, female, age class 19 to 24	11
74	speeding, safety belt applied, male, age class 25 to 34	**4**	speeding, severe casualty, safety belt applied, male, age class 25 to 34, **hitting a tree**	10
59	speeding, safety belt applied, female, age class 35 to 44	**5**	speeding, severe casualty, safety belt applied, female, age class 45 to 54	10
58	speeding, safety belt applied, female, age class 19 to 24, **probationary driving licence**	**6**	speeding, severe casualty, safety belt applied, male, age class 19 to 24	10
54	speeding, safety belt applied, female, age class 19 to 24, **skidding**	**7**	speeding, severe casualty, safety belt applied, male, age class 25 to 34, **hitting an obstacle next to the road**	8
43	speeding, safety belt applied, male, age class 19 to 24, **probationary driving licence**	**8**	speeding, severe casualty, safety belt applied, male, age class 19 to 24, **probationary driving licence**	7
40	speeding, safety belt applied, female, age class 25 to 34, **skidding**	**9**	speeding, severe casualty, safety belt applied, male, age class 19 to 24, **hitting an obstacle next to the road**	7
40	speeding, safety belt applied, female, age class 45 to 54	**10**	speeding, severe casualty, safety belt applied, male, age class 35 to 44	6

Table 72: The ten most frequent driver-related variable combinations including 'speeding'. The left side of the table represents the most frequent blackpatterns among the entire road traffic accident sample. The right side represents the most frequent blackpatterns among *severe casualties*. The right side is a subset of the left side. The column 'Count' indicates how often the blackpattern occurred within our observation period (2012-2019). n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

Generally, 'speeding' is associated with younger age classes. Among all casualties, 'speeding' co-occurs with 'probationary driving licence' and 'age class 19 to 24' for female and male drivers. Also, 'speeding', 'female drivers', and 'skidding' appear combinedly among all casualties. Among *severe casualties*, 'speeding' results in combinations including 'male drivers' and 'age class 25 to 34' as well as the variables 'hitting a tree' and 'hitting an obstacle next to the road'. For 'male drivers' and 'age class19 to 24' and *severe casualties*, 'speeding' co-occurs with the variables 'probationary driving licence' and 'hitting an obstacle next to the road'.

Table 73 illustrates the most frequent variable combinations with 'skidding/drifting'.

Count	Blackpatterns including skidding/drifting and all casualties	Position	Blackpatterns including skidding/drifting and severe casualties only	Count
80	skidding/drifting, safety belt applied, female, age class 19 to 24	**1**	skidding/drifting, severe casualty, female, age class 19 to 24, safety belt applied	6
55	skidding/drifting, safety belt applied, female, age class 25 to 34	**2**	skidding/drifting, severe casualty, female, age class 45 to 54, safety belt applied	5
54	skidding/drifting, safety belt applied, female, age class 19 to 24, **speeding**	**3**	skidding/drifting, severe casualty, male, age class 19 to 24, safety belt applied	5
40	skidding/drifting, safety belt applied, female, age class 25 to 34, **speeding**	**4**	skidding/drifting, severe casualty, female, age class 65+, safety belt applied, **speeding**	4
38	skidding/drifting, safety belt applied, female, age class 19 to 24, **probationary driving licence**	**5**	skidding/drifting, severe casualty, female, age class 25 to 34, safety belt applied, **speeding**	4
38	skidding/drifting, safety belt applied, male, age class 19 to 24	**6**	skidding/drifting, severe casualty, male, age class 19 to 24, safety belt applied, **speeding**	4
34	skidding/drifting, safety belt applied, female, age class 35 to 44	**7**	skidding/drifting, severe casualty, male, age class 55 to 64, safety belt applied	4
32	skidding/drifting, safety belt applied, male, age class 19 to 24, **speeding**	**8**	skidding/drifting, severe casualty, female, age class 35 to 44, safety belt applied	4
32	skidding/drifting, safety belt applied, female, age class 45 to 54	**9**	skidding/drifting, severe casualty, male, age class 35 to 44, safety belt applied, **alcohol**	4
25	skidding/drifting, safety belt applied, male, age class 19 to 24, **probationary driving licence**, **speeding**	**10**	skidding/drifting, severe casualty, female, age class 19 to 24, safety belt applied, **probationary driving licence**	3

Table 73: The ten most frequent driver-related variable combinations including 'skidding/drifting'. The left side of the table represents the most frequent blackpatterns among the entire road traffic accident sample. The right side represents the most frequent blackpatterns among *severe casualties*. The right side is a subset of the left side. The column 'Count' indicates how often the blackpattern occurred within our observation period (2012-2019). n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

Figure 31 shows that female drivers hold a higher share in 'skidding/drifting' than 'male drivers' in the event of a severe road traffic accident. We can see that the two variables 'speeding' and 'probationary driving licence' are combined with 'skidding/drifting'. We can see 'alcohol' impairment combined with 'skidding/drifting' on the ninth position within *severe casualties*.

The pattern detection among driver-related variables concludes with investigating co-occurring variables regarding safety settings. Chapter 4 reveals that *severe casualties* show a relatively high joint probability with 'no safety belt applied'. Therefore, we illustrate the most frequent blackpatterns with 'no safety belt applied' in table 74.

Count	Blackpatterns including no safety belt applied and all casualties	Position	Blackpatterns including no safety belt applied and severe casualties only	Count
34	no safety belt applied, male, age class 25 to 34, **alcohol**	**1**	no safety belt applied, severe casualty, male, age class 25 to 34	60
30	no safety belt applied, male, age class 25 to 34	**2**	no safety belt applied, severe casualty, male, age class 19 to 24	45
29	no safety belt applied, male, age class 19 to 24	**3**	no safety belt applied, severe casualty, male, age class 45 to 54	39
28	no safety belt applied, male, age class 19 to 24, **alcohol**	**4**	no safety belt applied, severe casualty, male, age class 35 to 44	30
22	no safety belt applied, male, age class 35 to 44, **alcohol**	**5**	no safety belt applied, severe casualty, male, age class 65+	29
22	no safety belt applied, male, age class 35 to 44	**6**	no safety belt applied, severe casualty, male, age class 55 to 64	18
19	no safety belt applied, male, age class 65+	**7**	no safety belt applied, severe casualty, male, age class 25 to 34, **alcohol**	13
17	no safety belt applied, female, age class 25 to 34	**8**	no safety belt applied, severe casualty, female, age class 19 to 24	10
15	no safety belt applied, male, age class 45 to 54	**9**	no safety belt applied, severe casualty, male, age class 19 to 24, **probationary driving licence**	9
12	no safety belt applied, male, age class 55 to 64	**10**	no safety belt applied, severe casualty, female, age class 65+	9

Table 74: Most frequent driver-related variable combinations including 'no safety belt applied'. The left side of the table represents the most frequent blackpatterns among the entire road traffic accident sample. The right side represents the most frequent blackpatterns among *severe casualties*. The right side is a subset of the left side. The column 'Count' indicates how often the blackpattern occurred within our observation period (2012-2019). n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

As table 74 shows, 'no safety belt applied' is the only variable with more blackpattern among *severe casualties* (60 combinations) than all casualties (34 combinations). Among all casualties, 'no safety belt applied' appears together with 'alcohol' on the very top. Among *severe casualties*, this combination appears within the seventh most frequent combination.

8.2 Blackpatterns among vehicle-related variables

In this chapter, we explore recurring blackpatterns among vehicle-related variables. Vehicle-related variables include engine power, kilometrage, vehicle colour, and safety settings. The characteristic 'airbag not deployed' shows a relatively high joint probability with *severe casualties* (see chapter 4.7*)*. Thus, we are curious to see whether this characteristic appears among the ten most frequent blackpatterns.

How to read the blackpattern tables

The left side of the table shows the most frequent blackpatterns (variable combinations) among the entire road traffic accident sample (n=20.293). The right side of the table shows the most frequent blackpatterns (variable combinations) among *severe casualties* (n=3.431). Thus, the right represents a subset of the left side. The idea behind it is to see whether there are different blackpatterns within *severe casualties* compared to all casualties—the column 'Position' in the middle of the table shows the numbers one to ten. One refers to the most frequently observed blackpattern. Consequently, the table shows the ten most common blackpatterns among the selected characteristics. The column 'Count' indicates how often the blackpattern occurred within our observation period (2012-2019).

Table 75 shows the most frequent vehicle-related blackpatterns. The blackpatterns include engine power, vehicle colour, kilometrage and safety settings. However, kilometrage does not appear within the top ten combinations. The characteristic 'airbag not deployed' is present in both categories but is ranked higher among all casualties (in the third to sixth most frequent combinations and the eighth and ninth combination). Among *severe casualties*, 'airbag not deployed' is present in the seventh and tenth most frequent combination. In total, 'airbag not deployed' occurred 819 times within *severe casualties* in the period under review. 'Airbag not deployed' occurred 7.318 times among accidents with slight injuries.

Count	Blackpatterns including vehicle-related variables and all casualties	Position	Blackpatterns including vehicle-related variables and severe casualties only	Count
975	airbag deployed, 24-90 kW	**1**	severe casualty, airbag deployed, 24-90 kW	247
958	airbag deployed, 24-90 kW, black	**2**	severe casualty, airbag deployed, 24-90 kW, black	222
939	**airbag not deployed**, 24-90 kW	**3**	severe casualty, airbag deployed, 24-90 kW, blue	213
890	**airbag not deployed**, 24-90 kW, black	**4**	severe casualty, airbag deployed, 24-90 kW, grey	190
868	**airbag not deployed**, 24-90 kW, blue	**5**	severe casualty, airbag deployed, 24-90 kW, red	142
805	airbag deployed, 24-90 kW, blue	**6**	severe casualty, airbag deployed, 24-90 kW, white	122
770	airbag deployed, 24-90 kW, green	**7**	severe casualty, **airbag not deployed**, 24-90 kW	95
641	**airbag not deployed**, 24-90 kW, grey	**8**	severe casualty, airbag deployed, 110 kW, black	90
602	**airbag not deployed**, 24-90 kW, red	**9**	severe casualty, airbag deployed, 24-90 kW, green	87
533	airbag deployed, 24-90 kW, red	**10**	severe casualty, **airbag not deployed**, 24-90 kW, blue	81

Table 75: The ten frequent vehicle-related variable combinations. The left side of the table represents the most frequent blackpatterns among the entire road traffic accident sample. The right side represents the most frequent blackpatterns among *severe casualties*. The right side is a subset of the left side. The column 'Count' indicates how often the blackpattern occurred within our observation period (2012-2019). n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

8.3 Blackpatterns among roadway-related variables

After investigating blackpatterns among driver- and vehicle-related variables, we will now analyse roadway-related variables. Roadway-related variables include road characteristics (straight road, curve, intersection, middle separation) and road surface condition (dry road, wet road, wintry conditions, sand or grit on the road).

How to read the blackpattern tables

The left side of the table shows the most frequent blackpatterns (variable combinations) among the entire road traffic accident sample (n=20.293). The right side of the table shows the most frequent blackpatterns (variable combinations) among *severe casualties* only (n=3.431). Thus, the right represents a subset of the left side. The idea behind it is to see whether there are different blackpatterns within *severe casualties* compared to all casualties. The column 'Position' in the middle of the table shows the numbers one to ten. One refers to the most frequently observed blackpattern. Consequently, the table shows the ten most common blackpatterns among the selected characteristics. The column 'Count' indicates how often the blackpattern occurred within our observation period (2012-2019). Table 76 shows the most frequent roadway-related variable combinations. Among all casualties, the combination of 'wet road' and 'curve' occurs more often than 'wet road' and 'straight road'. Most accidents occur on 'country roads' within a '100 km/h speed limit'.

Count	Blackpatterns including roadway-related variables and all casualties	Position	Blackpatterns including roadway-related variables and severe casualties only	Count
2.279	speed limit 100 km/h, regional road, dry road, straight road	**1**	severe casualty, speed limit 100 km/h, regional road, dry road, straight road	615
1.467	speed limit 100 km/h, regional road, dry road, curve	**2**	severe casualty, speed limit 100 km/h, regional road, dry road, curve	391
1.243	speed limit 100 km/h, regional road, wet road, curve	**3**	severe casualty, speed limit 100 km/h, regional road, wet road, straight road	255
996	speed limit 100 km/h, regional road, wet road, straight road	**4**	severe casualty, speed limit 100 km/h, regional road, wet road, curve	195
954	speed limit 100 km/h, regional road, wintry conditions, curve	**5**	severe casualty, speed limit 130 km/h, highway, dry road, straight road	132
900	speed limit 100 km/h, regional road, wintry conditions, straight road	**6**	severe casualty, speed limit 100 km/h, regional road, wintry conditions, straight road	116
529	speed limit 130 km/h, highway, dry road, straight road	**7**	severe casualty, driving ban, regional road, dry road, straight road	83
356	driving ban, regional road, dry road, straight road	**8**	severe casualty, speed limit 70 km/h, regional road, dry road, straight road	77
335	speed limit 70km/h, regional road, dry road, straight road	**9**	severe casualty, speed limit 100 km/h, another road, dry road, straight road	74
275	speed limit 130 km/h, highway, wet road, straight road	**10**	severe casualty, speed limit 100 km/h, regional road, wintry conditions, curve	70

Table 76: Most frequent roadway-related variable combinations. The left side of the table represents the most frequent blackpatterns among the entire road traffic accident sample. The right side represents the most frequent blackpatterns among *severe casualties*. The right side is a subset of the left side. The column 'Count' indicates how often the blackpattern occurred within our observation period (2012-2019). n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

8.4 Blackpatterns among situation-related variables

In addition to roadway-related variables, the analysis of situation-related variables explores blackpatterns consisting of weather conditions and light conditions. Weather conditions comprise 'clear or overcast weather', 'rain' and 'snow' while light conditions comprise 'daylight', 'darkness' and 'dusk or dawn'. As we can see in chapter 4, these situation-related variables result in a relatively high maximum combination value.

How to read the blackpattern tables

The left side of the table shows the most frequent blackpatterns (variable combinations) among the entire road traffic accident sample (n=20.293). The right side of the table shows the most frequent blackpatterns (variable combinations) among *severe casualties* (n=3.431). Thus, the right represents a subset of the left side. The idea behind it is to see whether there are different blackpatterns within *severe casualties* compared to all casualties. The column 'Position' in the middle of the table shows the numbers one to ten. One refers to the most frequently observed blackpattern. Consequently, the table shows the ten most common blackpatterns among the selected characteristics. The column 'Count' indicates how often the blackpattern occurred within our observation period (2012-2019).

Table 77 represents the most frequent variable combinations for situation-related variables. Within both categories, most accidents occur in 'daylight' and during 'clear or overcast weather'.

Count	Blackpatterns including situation-related variables and all casualties	Position	Blackpatterns including situation-related variables and severe casualties only	Count
591	6 p.m.to 12 a.m., Monday to Thursday, Winter, daylight, clear or overcast weather	**1**	severe casualty, 12 p.m. to 6 p.m., Monday to Thursday, Summer, daylight, clear or overcast weather	134
587	12 a.m. to 6 p.m., Monday to Thursday, Summer, daylight, clear or overcast weather	**2**	severe casualty, 6 p.m. to 12 a.m., Monday to Thursday, Summer, daylight, clear or overcast weather	101
453	6 a.m. to 12 a.m., Monday to Thursday, Summer, daylight, clear or overcast weather	**3**	severe casualty, 12 p.m. to 6 p.m., Friday to Sunday, Summer, daylight, clear or overcast weather	94
442	12 a.m. to 6 p.m., Monday to Thursday, Summer, daylight, clear or overcast weather	**4**	severe casualty, 6 p.m. to 12 a.m., Monday to Thursday, Winter, daylight, clear or overcast weather	93
439	12 a.m. to 6 p.m., Friday to Sunday, Summer, daylight, clear or overcast weather	**5**	severe casualty, 12 p.m. to 6 p.m., Monday to Thursday, Autumn, daylight, clear or overcast weather	89
408	6 p.m.to 12 a.m., Monday to Thursday, Winter, daylight, clear or overcast weather	**6**	severe casualty, 6 p.m. to 12 a.m., Monday to Thursday, Spring, daylight, clear or overcast weather	85
404	6 p.m. to 12 a.m., Monday to Thursday, Spring, daylight, clear or overcast weather	**7**	severe casualty, 0 a.m. to 6 a.m., Friday to Sunday, Autumn, darkness, clear or overcast weather	85
400	6 p.m. to 12 a.m., Monday to Thursday, Autumn, daylight, clear or overcast weather	**8**	severe casualty, 12 p.m. to 6 p.m., Friday to Sunday, Spring, daylight, clear or overcast weather	82
383	12 p.m. to 6 p.m., Monday to Thursday, Autumn, daylight, clear or overcast weather	**9**	severe casualty, 6 p.m. to 12 a.m., Friday to Thursday, Summer, daylight, clear or overcast weather	81
375	6 p.m. to 0 a.m., Monday to Thursday, Winter, darkness, clear or overcast weather	**10**	severe casualty, 6 p.m. to 12 a.m., Friday to Sunday, Summer, daylight, clear or overcast weather	

Table 77: The then most frequent situation-related variable combinations. The left side of the table represents the most frequent blackpatterns among the entire road traffic accident sample. The right side represents the most frequent blackpatterns among severe casualties. The right side is a subset of the left side. The column 'Count' indicates how often the blackpattern occurred within our observation period (2012-2019). n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

8.5 Blackpatterns among all accident-related variables

As a final investigation with the PATTERMAX-method, we detect the most frequent variable combinations (blackpatterns) among all accident-related variables. The variable characteristics represent dummy variables. Thus, a blackpattern represent an identical sequence of zeros and ones. The advantage of this method is to identify and quantify patterns that were indeed observed between 2012 and 2019. Also, the method does not assume the relationship among the variables. As the method results in many detected blackpatterns, we only illustrate selected blackpatterns in this chapter. Before we do so, we explain how to read the blackpattern tables.

How to read the blackpattern tables

The left side of the table shows the most frequent blackpatterns (variable combinations) among the entire road traffic accident sample (n=20.293). The right side of the table shows the most frequent blackpatterns (variable combinations) among *severe casualties* (n=3.431). Thus, the right represents a subset of the left side. The idea behind it is to see whether there are different blackpatterns within *severe casualties* compared to all casualties. The column 'Position' in the middle of the table shows the numbers one to ten. One refers to the most frequently observed blackpattern. Consequently, the table shows the ten most common blackpatterns among the selected characteristics. The column 'Count' indicates how often the blackpattern occurred within our observation period (2012-2019).

Count	Blackpatterns including all content variables and all casualties	Position	Blackpatterns including all content variables and severe casualties only	Count
984	dry road, clear or overcast weather, daylight	**1**	severe casualty, dry road, clear or overcast weather, daylight	269
673	dry road, clear or overcast weather, daylight, **airbag not deployed**	**2**	severe casualty, dry road, clear or overcast weather, **darkness**	104
427	dry road, clear or overcast weather, daylight, **curve**	**3**	severe casualty, dry road, clear or overcast weather, daylight, **curve**	104
306	dry road, clear or overcast weather, daylight, **airbag not deployed**, **curve**	**4**	severe casualty, dry road, clear or overcast weather, daylight, **airbag not deployed**	83
256	dry road, clear or overcast weather, **darkness**	**5**	severe casualty, dry road, clear or overcast weather, daylight, **fatigue**	76
245	dry road, clear or overcast weather, daylight, **fatigue**	**6**	severe casualty, dry road, clear or overcast weather, **darkness**, **curve**	57
205	dry road, clear or overcast weather, **darkness**, **alcohol**	**7**	severe casualty, dry road, clear or overcast weather, daylight, **no safety belt applied**	56
191	**wet road**, **rain**, daylight	**8**	severe casualty, dry road, clear or overcast weather, daylight, **distraction**	43
181	**wet road**, **rain**, daylight, **airbag not deployed**, **curve**	**9**	severe casualty, **wet road, rain,** daylight	42
178	dry road, clear or overcast weather, **darkness**, **airbag not deployed**, **curve**	**10**	severe casualty, dry road, clear or overcast weather, darkness, **no safety belt applied**	37

Table 78: The top ten combinations among all accident-related variables. The left side of the table represents the most frequent blackpatterns among the entire road traffic accident sample. The right side represents the most frequent blackpatterns among *severe casualties*. The right side is a subset of the left side. The column 'Count' indicates how often the blackpattern occurred within our observation period (2012-2019). n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

Comparing this table with the outcomes of binary logistic regression (chapter 5), decision trees (chapter 6) and Bayesian networks (chapter 7), we can see similar characteristics among the detected blackpatterns. The characteristics 'airbag not deployed' and 'no safety belt applied' appear among all the outcomes and are also part of the detected blackpatterns. The characteristics 'curve', 'wet road' and 'darkness' appear among all casualties and *severe casualties*. Regarding impairments, 'alcohol' and 'fatigue' appear among all casualties, whereas 'fatigue' and 'distraction' appear among *severe casualties*.

Thus far, we have only presented a selection of all blackpatterns. A disadvantage of the PATTERMAX-method is the generation of many blackpatterns (i.e., all blackpatterns observed between 2012 and 2019). The frequency of the blackpatterns is one way to classify their relevance. However, in the next step, we want to statistically examine all detected blackpatterns and present those significantly correlated with the target variable *severe casualties*.

9. Road traffic accident analysis part VI: Pattern significance

Another key advantage of the PATTERMAX-approach is the assignability of each detected blackpattern to the recorded accident. A blackpattern corresponds to an added column within the binary accident database where zero refers to 'blackpattern does not apply', and one refers to 'blackpattern applies'. This way, we know the exact distribution of the detected blackpattern among the historical road traffic accidents. Consequently, we can statistically evaluate whether the detected blackpattern significantly correlates with *severe casualties*. Therefore, the final analysis foresees the statistical evaluation of our detected blackpatterns. This chapter illustrates all blackpatterns showing a significant relationship with *severe casualties*. We apply Fisher's exact test to estimate this relationship and calculate the Phi coefficient. Also, we illustrate the frequency of each blackpattern showing a significant relationship with the target variable *severe casualties*.

9.1 Driver-related blackpattern significance

The PATTERMAX-method detects 529 driver-related blackpatterns (variable combinations including at least two accident-describing characteristics). 37 % (total: 197) of the detected blackpatterns occur at least ten times. We evaluate the relationship of these blackpatterns with the target variable *severe casualties* by applying Fisher's exact test. The resulting p-value indicates a significant relationship between a blackpattern and *severe casualties*. Also, we calculate the Phi coefficient to estimate the strength of the relationship. Of 197 driver-related blackpatterns occurring at least ten times between 2012 and 2019, 88 blackpatterns show a significant relationship with *severe casualties*. Table 79 shows an excerpt of these significant driver-related blackpatterns. The selection is based on the resulting Phi coefficient. A positive Phi coefficient indicates that the blackpattern occurs comparatively often among *severe casualties*. A negative Phi coefficient does not mean that the blackpattern does not occur among *severe casualties*. It comparatively occurs more often among casualties with slight injuries. Thus, table 79 illustrates blackpatterns showing a significant relationship with *severe casualties* and a positive Phi coefficient.

The driver-related blackpattern evaluation reveals that blackpatterns that correlate with *severe casualties* primarily include 'male drivers'. Table 79 only shows two significant blackpatterns with 'female drivers'. Both patterns include young 'female drivers' with the characteristic 'no safety belt applied'. Young 'male drivers' with the characteristic 'no safety belt applied' also significantly correlate *with severe casualties*. The characteristic 'fatigue' appears among

blackpatterns with 'male drivers' older than 35 years. Also, blackpatterns with 'male drivers' between the age classes '19 to 24', '25 to 34', and '55-64', and the characteristic 'speeding' show a significant relationship with *severe casualties*. However, 'no safety belt applied' clearly is the most common characteristic in the below shown blackpatterns.

Driver-related blackpatterns	Fisher's exact test p	Phi coefficient ϕ	Blackpattern Frequency n
male, age 25 to 34, speeding	0,022	0,016	190
male, age 65+, fatigue	0,004	0,02	147
male, age 65+ distraction	0,014	0,018	127
male, 25 to 34, no safety belt applied	0,000	0,092	110
male, age 35 to 44, fatigue	0,046	0,015	106
male, 19 to 24, no safety belt applied	0,000	0,073	87
male, age 55 to 64, fatigue	0,001	0,023	86
male, age 65+, no safety belt applied	0,000	0,068	83
male, age 45 to 54, distraction	0,004	0,02	79
male, 25 to 34, alcohol, no safety belt applied	0,028	0,016	74
male, age 45 to 54, no safety belt applied	0,000	0,08	71
male, age 35 to 44, no safety belt applied	0,000	0,071	64
Male, age 55 to 64, speeding	0,009	0,02	46
male, age 55 to 64, no safety belt applied	0,000	0,058	37
male, age 19 to 24, speeding, no safety belt applied	0,000	0,055	37
male, age 45 to 54, alcohol, no safety belt applied	0,021	0,018	29
female, 25 to 34, no safety belt applied	0,004	0,022	28
female, 19 to 24, no safety belt applied	0,002	0,024	26

Table 79: Driver-related blackpatterns showing a significant relationship with the target variables *severe casualties* and a positive Phi coefficient. n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

9.2 Vehicle-related blackpattern significance

The PATTERMAX-method detects 440 vehicle-related blackpatterns, consisting of at least two vehicle-related characteristics. Of these 440 blackpatterns, 136 (31 %) occurred at least ten times between 2012 and 2019. We evaluate the relationship of these blackpatterns with the target variable *severe casualties* by applying Fisher's exact test. The resulting p-value indicates a significant relationship between a blackpattern and *severe casualties*. Also, we calculate the Phi coefficient to estimate the strength of the relationship. A positive Phi coefficient indicates that the blackpattern occurs comparatively often among *severe casualties*. A negative Phi coefficient does not mean that the blackpattern does not occur among *severe casualties*. It comparatively occurs often among casualties with slight injuries. In total, 30 vehicle-related blackpatterns show a significant relationship with *severe casualties*. The statistical evaluation of vehicle-related blackpatterns shows that none of the significant blackpatterns receives a positive Phi coefficient. Since we only illustrate blackpatterns with a significant relationship and a positive Phi coefficient in this chapter, we cannot illustrate any vehicle-related blackpatterns.

9.3 Roadway-related blackpattern significance

The PATTERMAX-method detects 394 roadway-related blackpatterns, of which 129 (33 %) occur at least ten times between 2012 and 2019. A blackpattern is a variable combination including at least two roadway-related characteristics. We evaluate the relationship of these blackpatterns with the target variable *severe casualties* by applying Fisher's exact test. The resulting p-value indicates a significant relationship between a blackpattern and *severe casualties*. Also, we calculate the Phi coefficient to estimate the strength of the relationship. A positive Phi coefficient indicates that the blackpattern occurs comparatively often among *severe casualties*. A negative Phi coefficient does not mean that the blackpattern does not occur among *severe casualties*. It occurs comparatively often among casualties with slight injuries. In total, 27 roadway-related blackpatterns significantly relate to the target variable *severe casualties*. Thus, table 80 illustrates blackpatterns showing a significant relationship with *severe casualties* and a positive Phi coefficient.

Roadway-related blackpatterns	Fisher's exact test p	Phi coefficient ϕ	Blackpattern Frequency n
speed limit 100km/h, country road	0,000	0,046	2.922
speed limit 100km/h, country road, curve	0,000	0,036	1.914
speed limit 100km/h, country road, wet road	0,003	0,021	1.253
speed limit 100km/h, other road	0,001	0,024	340
speed limit 100km/h, other road, curve	0,022	0,016	318
driving ban, highway	0,014	0,017	166
speed limit 100km/h, other road, wet road	0,031	0,016	162
driving ban, other road	0,000	0,033	130
speed limit 100km/h, highway, tunnel	0,022	0,017	18
speed limit 100km/h, country road, bridge	0,007	0,022	15
speed limit 80 km/h, country road, tunnel	0,005	0,023	11

Table 80: Roadway-related blackpatterns showing a significant relationship with the target variables *severe casualties* and a positive Phi coefficient. n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

The most common variable combination among significant roadway-related blackpatterns with a positive Phi coefficient is 'speed limit 100km/h' and 'country road'. The road characteristic 'curve' and the road condition 'wet road' are significant blackpatterns showing a relatively high frequency. This picture matches the outcomes of our previous analyses. The Bayesian networks (chapter 7.4) and the decision trees (chapter (6.4) also highlight 'wet road' and 'curves' as characteristics correlating with *severe casualties*.

9.4 Situation-related blackpattern significance

The PATTERMAX-method detects 1.250 situation-related blackpatterns consisting of at least two situation-related characteristics. 297 (24 %) of these blackpatterns occurred at least ten times between 2012 and 2019. We evaluate the relationship of these blackpatterns with the target variable *severe casualties* by applying Fisher's exact test. The resulting p-value indicates a significant relationship between a blackpattern and *severe casualties*. Also, we calculate the Phi coefficient to estimate the strength of the relationship. A positive Phi coefficient indicates that the blackpattern comparatively occurs more often among *severe casualties*. A negative Phi coefficient does not mean that the blackpattern does not occur among *severe casualties*. It occurs comparatively often among casualties with slight injuries. In total, 36 situation-related blackpatterns show a significant relationship with the target variable *severe casualties*. Table 81 illustrates blackpatterns with a positive Phi coefficient and a significant relationship with the target variable *severe casualties* (estimated with Fisher's exact test) and the strength of this relationship (estimated with the Phi coefficient).

Blackpatterns with a timeframe between '0 a.m. to 6 a.m.', weekday Friday to Sunday, 'darkness' and 'right drift' represent the most frequent and significant situation-related blackpatterns with a positive Phi Coefficient. The characteristic 'darkness' appears among all the below shown situation-related blackpatterns except for one. The situation-related logistic regression (chapter 5.5) also suggests that severe or fatal accidents decrease from Monday to Thursday on weekdays. Regarding time, '6 p.m. to 0 a.m.' frequently appears among the illustrated blackpatterns. Among meteorological seasons, no particularly conspicuous season shows up. As we can see in chapter 4, accidents are almost equally distributed over the meteorological seasons.

Situation-related blackpatterns	Fisher's exact test p	Phi coefficient ϕ	Blackpattern Frequency n
0 a.m. to 6.a.m., Friday to Sunday, autumn, darkness, right drift	0,004	0,021	177
0 a.m. to 6.a.m., Friday to Sunday, winter, darkness, right drift	0,014	0,017	166
0 a.m. to 6.a.m., Friday to Sunday, spring, darkness, right drift	0,010	0,019	139
6 p.m. to 0 a.m., Monday to Thursday, spring, darkness, right drift	0,006	0,02	135
0 a.m. to 6.a.m., Friday to Sunday, autumn, darkness, left drift	0,001	0,023	129
0 a.m. to 6.a.m., Friday to Sunday, winter, darkness, left drift	0,029	0,016	128
6 p.m. to 0 a.m., Monday to Thursday, autumn, darkness, left drift	0,003	0,22	125
0 a.m. to 6.a.m., Friday to Sunday, summer, darkness, left drift	0,000	0,025	103
0 a.m. to 6.a.m., Friday to Sunday, spring, darkness, left drift	0,019	0,017	100
0 a.m. to 6.a.m., Monday to Thursday, autumn, darkness, right drift	0,001	0,024	98
0 a.m. to 6.a.m., Monday to Thursday, darkness, right drift	0,023	0,016	83
6 p.m. to 0 a.m., Monday to Thursday, spring, darkness, left drift	0,022	0,017	82
6 p.m. to 0 a.m., Friday to Sunday, darkness, left drift	0,001	0,025	74
0 a.m. to 6.a.m., Monday to Thursday, summer, darkness, left drift	0,000	0,03	65
6 p.m. to 0 a.m., Friday to Sunday, summer, darkness, left drift	0,001	0,026	64
6 p.m. to 0 a.m., Friday to Sunday, daylight, left drift	0,016	0,018	45
0 a.m. to 6.a.m., Monday to Thursday, autumn, rain, darkness, left drift	0,011	0,02	30
12 p.m. to 6 p.m., Monday to Thursday, winter, dusk or dawn, left drift	0,300	0,017	26
6 a.m. to 12 a.m., Monday to Thursday, summer, daylight	0,027	0,018	25
6 p.m. to 0 a.m., Monday to Thursday, autumn, dusk/dawn, right drift	0,023	0,17	22
6 p.m. to 0 a.m., Friday to Sunday, spring, rain, darkness, right drift	0,007	0,021	22
6 a.m. to 12 p.m., Friday to Sunday, winter, rain, daylight, right drift	0,022	0,017	18

Table 81: Situation-related blackpatterns showing a significant relationship with the target variables *severe casualties* and a positive Phi coefficient. n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

9.5 Overall blackpattern significance

The PATTERMAX-method detects 12.705 blackpatterns that include at least two accident-related characteristics from all four categories (driver, vehicle, situation, and roadway). 144 (1 %) of these blackpatterns occurred at least ten times between 2012 and 2019. We evaluate the relationship of these blackpatterns with the target variable *severe casualties* by applying Fisher's exact test. The resulting p-value indicates a significant relationship between a blackpattern and the target variable *severe casualties*. Also, we calculate the Phi coefficient to estimate the strength of the relationship. A positive Phi coefficient indicates that the blackpattern occurs comparatively often among *severe casualties*. A negative Phi coefficient does not mean that the blackpattern does not occur among *severe casualties*. It occurs comparatively often among casualties with slight injuries. 11 overall detected blackpatterns significantly correlate with the target variable *severe casualties*. Table 82 illustrates blackpatterns with a positive Phi coefficient and a significant relation with the target variable *severe casualties*.

As within the evaluation of driver-related blackpatterns, most of the illustrated significant blackpatterns include 'male drivers' (except for one blackpattern). Similar to roadway-related blackpatterns, the characteristics 'speed limit 100 km/h' and 'country road' frequently occur among the significant overall blackpatterns. Also, the variables 'curve' and 'wet appear in the most frequent and significant overall blackpatterns with a positive Phi coefficient. Regarding road types, road conditions and road characteristics, this underpins the results of the previous analyses. Also, 'male drivers' and 'fatigue' appear in the below-shown patterns. This combination corresponds to the results of chapter 4, where 'male drivers' hold a higher share among *severe casualties* than female drivers. Furthermore, 'right drift' occurs twice in combination with 'highway' and 'speed limit 130 km/h'.

Overall blackpatterns	Fisher's exact test p	Phi coefficient ϕ	Blackpattern Frequency n
speed limit 130km/h, highway, right drift, male driver	0,001	0,027	44
speed limit 100km/h, country road, left drift, male driver	0,000	0,032	41
speed limit 100km/h, country road, curve, left drift, male driver	0,011	0,020	30
country road, right drift, female driver	0,042	0,015	28
speed limit 100km/h, country road, left drift, male driver, fatigue	0,001	0,028	20
speed limit 130km/h, highway, drifting right, male driver, fatigue	0,040	0,015	16
speed limit 100km/h, country road, wet road, age 25-34, right drift, male driver	0,001	0,027	12
speed limit 100km/h, country road, left drift, male driver, no safety belt applied	0,000	0,031	10
speed limit 100km/h, country road, darkness, right drift, male driver	0,003	0,026	10
speed limit 80km/h, country road, right drift, male driver	0,016	0,020	10

Table 82: Overall blackpatterns showing a significant relationship with the target variables *severe casualties,* and a positive Phi coefficient. n=20.293 single-vehicle accidents with single occupation and personal injury occurring outside the built-up area on the Austrian road network (3.431 are *severe casualties*).

10. Discussion and Outlook

The main finding of this thesis is that the examination of recorded accident circumstances reveals blackpatterns (i.e., recurring combinations of accident-related characteristics) that are significantly associated with severe and fatal road traffic accidents.

10.1 Content-related discussion

Key insights for driver-related characteristics

In the event of a severe road traffic accident, ‘no safety belt applied’ occurs more often than ‘distraction, ‘fatigue’, or ‘alcohol’. The driver-related characteristic ‘no safety belt applied’ is the most noticeable factor in increasing the risk of a *severe casualty*. ‘No safety belt applied’ shows the highest Phi coefficient among all accident-describing characteristics. Furthermore, it results in the highest odds ratios compared to all accident-related characteristics in binary logistic regression. In addition, among driver-related characteristics, the most frequent blackpattern with ‘no safety belt applied’ and *severe casualties* appears more often than the most frequent blackpattern with ‘no safety belt applied’ and accidents with slight injuries.

Additionally, when analysing *severe casualties*, the relative frequency of male drivers not applying a safety belt is twice as high as female drivers. The analysis shows that ‘no safety belt applied’ is associated with all age groups. Therefore, young drivers also show significant blackpatterns with *severe casualties* and 'no safety belt applied'.

Furthermore, the most frequent blackpatterns with ‘no safety belt applied’ are also associated with ‘alcohol’. The analysis of driver-related accident characteristics reveals gender-specific differences, especially for different types of impairment. Male drivers steer three times more often than female drivers regarding' alcohol'. The characteristics ‘speeding’ and ‘fatigue’ do not show gender-specific differences. ‘Distraction’ shows a higher relative frequency among *severe casualties* and female drivers. Considering the co-occurring variables between *severe casualties* and male or female drivers, *severe casualties* with male drivers occur with the variable ‘no safety belt applied’ and 'alcohol’. Thus, there exist different blackpatterns among female and male drivers in the case of a severe or fatal road traffic accident that may be of interest for future road safety work. The Austrian Road Safety Strategy declares ‘alcohol’ and ‘distraction’ as central challenges for reducing severe road traffic accidents. For the examined

sample of single-vehicle accidents with single-occupancy and personal injury, 'fatigue' represents an additional variable to be considered, especially among younger drivers.

Among single-vehicle accidents outside the built-up area with a single occupation and personal injury, the relative frequency to observe a male driver in a *severe casualty* is 12 %. In comparison, the relative frequency to observe a female driver in a *severe casualty* is 5 %. In fact, among the illustrated blackpatterns showing a significant relationship with *severe casualties*, most blackpatterns include male drivers.

In addition, the comparison of blackpatterns among female and male drivers reveals specific differences in driving behaviour and driving licence type. Among *severe casualties*, the relative frequency of female drivers owning a 'probationary driving licence' is higher than for male drivers. The most frequent blackpatterns with *severe casualties* and female drivers include 'probationary driving licence', 'fatigue', and 'speeding'. The blackpattern 'probationary driving licence' and 'alcohol' appears more often among male drivers. Figure 31 shows that female drivers hold a higher share in 'skidding/drifting' than male drivers in the event of a severe or fatal road traffic accident. The blackpatterns illustrated in chapter 8 show that the two variables 'speeding' and 'probationary driving licence' are combined with 'skidding/drifting', especially among female drivers.

Equally important, when comparing age classes, age class '19 to 24' and age class '25 to 34' show the highest relative frequencies among *severe casualties*. Also, the characteristic 'probationary driving licence' shows a relatively high number among *severe casualties*.

Key insights for vehicle-related characteristics

The vehicle-related characteristic 'airbag not deployed' shows a significant relationship with *severe casualties* when analysing it individually. 'Airbag not deployed' substantially impacts *severe casualties* among all the applied pattern recognition methods. The logistic regression model assigns the characteristic 'airbag not deployed' with a fundamentally high impact to increase the risk of observing a severe or fatal road traffic accident. Among *severe casualties*, a blackpattern including 'airbag not deployed' appears in fourth place.

Key insights for roadway-related characteristics

The univariate analysis and the pattern recognition approaches show that the most *severe casualties* occur on 'country roads' within a 'speed limit of 100 km/h'. Regarding roadway-related characteristics, the characteristics 'curve', 'middle separation', 'intersection', 'tunnel' and 'bridge' correlate with *severe casualties*. Some of these characteristics might rarely occur among the entire dataset. Still, if they occur, the outcome of the accident has a relatively high probability of resulting in a severe or fatal road traffic accident. 'Wet roads' and 'wintry

conditions' are the two road surface characteristics that significantly correlate with *severe casualties*. Also, both characteristics are present among all pattern recognition approaches. The Bayesian networks and decision trees highlight 'wet road' and 'curves' with *severe casualties*. Among the detected blackpatterns, the combination of 'wet road' and 'curve' occurs more often than 'wet road' and 'straight road'. Thus, the road characteristic 'curve' and the road condition 'wet road' result in a significant blackpattern with a relatively high frequency.

Key insights for situation-related characteristics

The analysis of situation-related variables reveals that the variables 'darkness' and 'Monday to Friday' hold a higher share among *severe casualties* than among accidents with slight injuries (see figure 44). Regarding daytime, blackpatterns with a timeframe between '0 a.m. to 6 a.m.', weekday Friday to Sunday, 'darkness' and 'right drift' correlate significantly with *severe casualties*. The situation-related logistic regression (chapter 5.5) also suggests that severe or fatal accidents decrease from Monday to Thursday on weekdays.

Also, the characteristics 'rain' and 'snow' hold a relatively high share among the number of severe or fatal road traffic accidents. As we can see in the presented blackpatterns, the most *severe casualties* occurred on 'country roads' within a 'speed limit of 100 km/h' combined with the characteristic 'wet road'. Additional blackpatterns showing a significant relationship with *severe casualties* are

- 'highway', 'speed limit of 130 km/h', 'wet roads', and
- 'highway', 'speed limit of 130 km/h', and 'wintry conditions'.

Based on these results, it may be helpful to evaluate carefully whether further mandatory speed reductions in case of poor road and weather conditions could significantly reduce severe or fatal road traffic accidents. To evaluate the impact of such a measure, the generation of an accident prediction model could be considered. The estimated 95 % confidence intervals may serve as input variables to develop an accident prediction model.

It remains to be mentioned that an extension of the so-called 'Road Safety Inspections' (Nadler, Nadler, and Strnad, 2016) towards 'country roads' is currently being considered and discussed. The recognized blackpatterns underpin the significance of this project.

10.2 Methodological discussion

Contingency tables and conditional and joint probabilities are simple and yet powerful tools to explore the data and to get a first impression of a variable's impact on *severe casualties*. For example, male drivers share 57 % and female drivers 43 % within the investigated road traffic accidents. The probability of a *severe casualty* is 12 % for a male driver and 5 % for a female driver. We can see that this distribution does not correspond to the initial distribution among all accidents. Also, conditional probability is an integral part of the presented pattern recognition methods, especially for the Bayesian networks and the PATTERMAX-method. As we can see in chapter 4, multiple variables show a significant relationship with *severe casualties*. However, when looking at the resulting Phi coefficient, we can see that the maximum Phi coefficient is .240 for the variable 'no safety belt applied', representing a weak relationship. Given the assumption that road traffic accidents do not present monocausal events, we consider the maximum combination value a more reliable measure to facilitate the pattern recognition process. It analyses all variables and reveals how often a specific variable combination (blackpattern) appears within the dataset. The calculation of the maximum combination value is aligned with the developed aggregation method (PATTERMAX-method). The PATTERMAX-method searches for identical variable combinations (blackpatterns) within the binary-structured road traffic accident database and counts their frequencies. At this point, we can calculate the respective conditional probability of the detected blackpattern among *severe casualties*.

Binomial logistic regression has proven to be a powerful tool to estimate a variable's impact on *severe casualties* compared to all accident-describing characteristics. For this study, the benefit of binary logistic regression is twofold:

- it helps us to exclude characteristics having no significant relationship with our target variable *severe casualties;*
- it helps us estimate the impact of an accident-related characteristic on *severe casualties* compared to all investigated accident characteristics.

For the subsequent pattern recognition procedures, this is essential information to identify blackpatterns that

- exclusively include accident-related characteristics having a significant relationship with *severe casualties;*
- can be assessed because of knowing the impact of each included accident-related characteristic on *severe casualties*.

Decision trees represent an efficient way to depict 'the big picture' off accident-describing characteristics that influence *severe casualties*. However, they do not provide in-depth knowledge about accident circumstances and substantial differences among male and female drivers or between road types or weekdays. Nevertheless, we consider decision trees a helpful

tool to estimate which variables impact *severe casualties*. Also, decision trees represent a first step towards pattern detection as they associate accident-related characteristics with each other.

After that, we conclude that the Bayesian network approach is suitable for visualising more detailed and relevant combinations of accident-related characteristics. Furthermore, the Bayesian network approach automatically estimates the joint probability of a detected sequence of characteristics among all casualties. Nonetheless, the PATTERMAX-approach queries blackpatterns in even greater detail. Within our newly established binary road traffic accident dataset, the PATTERMAX-method searches for identical sequences of zeros and ones, counts their frequencies and calculates their conditional probabilities. A blackpattern is a variable combination that occurs more than ten times and includes two accident-describing characteristics.

What is more, the PATTERMAX-approach exclusively identifies variable combinations that indeed occurred within the recorded accident data while the Bayesian network approach generalises certain relationships among accident-related variables. Thus, the PATTERMAX-approach allows us to gain better insights into valid and recurring blackpatterns (i.e., recurring combinations of accident-related characteristics). Another key advantage of the PATTERMAX-approach is assigning a detected blackpattern to the recorded accident. A blackpattern corresponds to an added column within the binary accident database where zero refers to 'blackpattern does not apply', and one refers to 'blackpattern applies'. This way, we know the exact distribution of the detected blackpattern among the historical road traffic accidents. This way, we can statistically evaluate whether the detected blackpattern shows a significant relationship with *severe casualties*.

To sum up, we recommend the application of binomial logistic regression and the PATTERMAX-method to gain in-depth knowledge about recurring accident blackpatterns and the impact of accident-describing characteristics to increase *severe casualties*.

10.3 Limitations and disclaimer

Superordinate framework conditions (e.g., traffic policy, StVO, safety culture, etc.) and how police officers record accidents strongly influence accident data quality. For example, assessing the alleged main cause of the accident represents a subjective assessment by the police officer who fills out the accident data sheet on site. Depending on how differently police officers may be trained on accident surveys, there always exists a so-called evaluation bias going along with road traffic accident records. This thesis does not examine these superordinate parameters. The focus is placed exclusively on examining the officially available traffic accident data. Since the alleged main cause of the accident is of higher meaning than all the other entries in the datasheet (i.e., all the other accident-describing variables), this thesis

does not analyse the alleged main causes of accidents. It strictly focuses on the investigation of the accompanying circumstances.

Therefore, the interpretation of the identified accident patterns requires caution.

It is to emphasize that this thesis presents a pattern recognition method based on recorded (historical) road traffic accidents. The target is to reveal evidence-based patterns (i.e., recurring accident conditions) from historical road traffic accidents records. The results of this thesis will allow us to say how likely it is to find a specific pattern in a single-vehicle accident with a single occupation. The thesis does not focus on building a prediction model for road traffic accidents. It may, however, be possible to estimate the probability of death or severe injury in the event of a single-vehicle accident based on the proposed model.

The thesis exclusively relates to the analysis of historic accident data and the question of whether recurring patterns (variable combinations) underly these data. It does not include any data on traffic performance. Therefore, it is not possible to deduce how likely it is for a particular accident or a specific accident pattern to occur. However, the work makes it possible to say that if an accident occurs, with what probability will it show a particular pattern (based on historical road traffic accidents only).

There are further points to consider when using or interpreting the data:

- The focus is on pattern recognition based on recorded (i.e., historic) road traffic accidents.

- Traffic performance is not part of this thesis. Therefore, it is impossible to conclude how likely one specific accident or accident pattern is to occur.

- Conditional and joint probabilities describe how likely it is for a variable to appear among *severe casualties*. It does not indicate its occurrence throughout traffic performance (i.e., a 16 % probability for male drivers does not suggest that 16 % of male drivers have an accident. It simply demonstrates that male drivers appear in 16 % of all historic traffic accidents).

- The accident-related variables do not occur with equal frequency (i.e., their actual frequency is unknown). A direct comparison of the variables is not valid (i.e., an 80 km/h speed limit on regional roads is less common than a 100 km/h speed limit. Thus, comparing the occurrences of severe casualties within both speed limits and concluding that fatal road accidents occur more often within 100 km/h speed limits is invalid. A valid comparison requires additional knowledge on the respective speed limits within the road network and traffic performance within the speed limits, which is not considered or estimated in this work). It is, however, valid to descriptively compare the underlying patterns within an 80 km/h speed limit and a 100 km/h

speed limit and to see whether there appear different patterns within different speed limits.

- The prediction of future accidents is not part of this thesis. Still, we will suggest a bootstrapping resampling method to build a robust parameter estimation with 95 % confidence intervals.

10.4 Outlook

In general, there exist three fields of application for future work.

- Expansion of the PATTERMAX-method and binomial logistic regression on further accident types: This thesis illustrates a pattern recognition approach for road traffic accident data using single-vehicle accidents with single-occupancy. In Austria, there exist ten different types of accidents. Further work may foresee the application of the established methods on the remaining types of accidents.

- Developing an accident prediction model: We suggest generating a detailed investigation based on accident-describing variables. In particular, the 95 % confidence intervals showing the probability range of a characteristic among *severe casualties* may be of interest to set up an accident prediction model. A prediction model can estimate the impact of selected measures on accident occurrences. Such a model could serve as a decision-making tool by estimating the statistically valid effect of different measures to reduce severe and fatal road traffic accidents.

- Consideration of traffic performance: Further research may also integrate traffic performance as an additional accident-describing variable. To estimate the explanatory power of traffic performance on accident frequency and degree of injury, sufficient data quality on traffic performance must be available. Furthermore, the analysis with traffic performance should examine all accident types and not only a selected accident sample.

11. Summary

Chapter one provides an overview of the thesis context, research gap, research questions and associated targets, and the scientific classification of the thesis.

Chapter two represents a theoretical chapter where we dive into road traffic accident data (i.e., uncertainty, noise and bias, rare events, heterogeneity, and over-dispersion). Also, we discuss pattern recognition methods within this chapter.

Chapter three looks at the existing accident types, and we present the reasons for choosing one specific accident type on which we test and run the pattern recognition approach. Moreover, we discuss the characteristics of the existing road traffic accident database and point out the reasons for the data reprocessing task. This reprocessing task leads to developing a binary database that includes more than 150 accident-related variables. Next, we categorise these accident-related variables into the following scheme: driver-related variables, vehicle-related variables, roadway-related variables, and situation-related variables. The third chapter concludes with the definition of the dependent variable.

After the three introductory chapters, we jump into analysis part I in chapter four. This chapter presents each accident-related characteristic in detail with the help of descriptive statistics. First, we show how often a variable occurs among all accidents (severe and fatal accidents and accidents with slight injuries). Second, we only show how often a variable occurs among severe and fatal accidents. Based on the contingencies, we calculate the probability for a severe or fatal road accident given the respective accident-related variable.

Additionally, we apply Fisher's exact test to determine a possible relationship between an accident-related variable and the dependent variable (severe and fatal road accidents). Fisher's exact test shows whether there is a significant relationship between the two variables and outputs the Phi coefficient to determine the strength of the relationship. Also, we generate a robust parameter estimation (95% confidence intervals showing the likelihood of a variable and a severe or fatal accident to occur) by applying a bootstrap resampling method on the newly established accident database. Moreover, we calculate a so-called maximum combination value as the first value towards blackpattern detection. This value tells us how often a specific variable co-occurs with (an)other accident-related variable(s).

Chapter five uses binomial logistic regression to estimate each variable's impact on severe road traffic accidents with an odds ratio (i.e., the strength of the relationship between an accident-related variable and the target variable *severe casualties* (i.e., severe or fatal accidents) compared to all observed variables). By knowing which variable appears to increase

the risk of a severe road traffic accident, we can assess the overall impact of the detected blackpatterns.

Furthermore, we grow decision trees using the CHAID-algorithm in chapter six. Decision trees generate a generalized tree-like structure of variable combinations that appear to increase the probability of a severe road traffic accident. At this point, binomial logistic regression and decision trees help us identify variables that aggravate an accident outcome and the respective degree of injury. However, since we are interested in gaining in-depth knowledge of recurring variable combinations (blackpatterns), we zoom deeper into the underlying data structures.

Consequently, we apply an explorative Bayesian network paradigm in chapter seven. Also, we apply a developed pattern detection method based on the frequency of variable combinations and joint probabilities (PATTERMAX-method) in chapter eight.

In chapter nine, the pattern recognition process concludes with a statistical evaluation of whether the detected blackpatterns show a significant relationship with the target variable *severe casualties*. Like the beginning, so the end, and we calculate Fisher's exact test and the Phi coefficient.

To conclude, we highlight the most aggravating accident-related variables and blackpatterns in chapter ten. Also, we compare the applied pattern recognition methods. The discussion highlights the advantages and the limitations of the PATTERMAX-method combined with binomial logistic regression to gain in-depth knowledge about accident circumstances. The combined application of both methods enables a precise detection and comparison of blackpatterns. For example, do accident patterns among female drivers differ from accident patterns among male drivers? Do accident patterns on regional roads within an 80 km/h speed limit differ from those on a 100 km/h speed limit? Additionally, the combined approach enables the assessment of the detected blackpatterns with the help of an odds ratio.

Within the research outlook, we propose to expand the PATTERMAX-approach in combination with binomial logistic regression on other accident types. The newly established accident database might also serve as a reliable source for accident prediction. The estimated 95% confidence intervals may represent input variables for a prediction model.

12. References

Aga, M., Woldemanuel, B., & Tadesse, M. (2021). *Statistical modeling of numbers of human deaths per road traffic accident in the Oromia region, Ethiopia*. doi:10.1371/journal.pone.0251492.

Ahmed, L. (2017). Using Logistic Regression in Determining the Effective Variables in Traffic Accidents. Applied Mathematical Sciences 11(42).

Al Musawi, A. (2018). *Introduction to Machine Learning*. Retrieved 04 12, 2021, from https://www.researchgate.net/project/Introduction-to-Machine-Learning.

Alavi, S., Mohammadi, M., Souri, H., Kalhori, S., Jannatifard, F., & Sepahodi, G. (2017). *Personality, Driving Behavior and Mental Disorders Factors as Predictors of Road Traffic Accidents Based on Logistic Regression*. Iranian Journal of Medical Sciences 42(1).

Almamook, R., Keneth, K., Abdulbaset, A., & Alkasisbeh, M. (2019). *Comparison of Machine Learning Algorithms for Predicting Traffic Accident Severity*. IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT). doi:10.1109/JEEIT.2019.8717393.

Assi, K., Rahman, S., Mansoor, U., & Ratrout, N. (2020). *Predicting Crash Injury Severity with Machine Learning Algorithm Synergized with Clustering Technique: A Promising Protocol*. International Journal of Environmental Research and Public Health 17(15). doi:10.3390/ijerph17155497.

Ayman, E., & Ali, S. (2019). *Anomaly Detection Methods for Categorical Data: A Review*. Association for Computing Machinery. New York (USA): ACM Computing Survey 52(2). doi:10.1145/3312739.

Basu, S., & Saha, P. (2017). *Regression Models of Highway Traffic Crashes: A Review of Recent Research and Future Research Needs*. 10th International Scientific Conference Transbaltica 2017: Transportation Science and Technology . Procedia Engineering 187. doi:10.1016/j.proeng.2017.04.350.

Bayes, T. (1763). *An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter*

to John Canton, A. M. F. R. S. The Royal Society. doi:10.1098/rstl.1763.0053.

Berrar, D. (2019). *Introduction to the Non-Parametric Bootstrap.* Encyclopedia of Bioinformatics and Computational Biology. doi:10.1016/B978-0-12-809633-8.20350-6.

Bundesministerium für Inneres (BMI) I. (2020). *Straßenverkehrstote in Österreich*. Retrieved July 7, 2021, from bmi.gv.at: https://www.bmi.gv.at/202/Verkehrsangelegenheiten/files/Jahresvergleich_1983_2020.pdf.

Bundesministerium für Inneres (BMI) II. (2020). *Verkehrsstatistik 2020*. Retrieved July 7, 2021, from bmi.gv.at: https://www.bmi.gv.at/202/Verkehrsangelegenheiten/unfallstatistik_vorjahr.aspx.

Cerema-ONISR. (2020). *Final evaluation of 80km/h speed limit on single cariageway roads outside the built-up areas.* Retrieved from https://www.onisr.securite-routiere.gouv.fr/en/knowledge-centre/evaluation/evaluation-of-the-measures/80-kmh-speed-limit-on-rural-single-carriageways.

Cerwenka, P., Hauger, G., Hörl, B., & Klamer, M. (2007). Handbuch der Verkehrssystemplanung. Österreichischer Kunst- und Kulturverlag, Wien.

Chong, M., Abraham, A., & Paprzycki, M. (2005). *Traffic Accident Analysis Using Machine Learning Paradigms.* Informatica (Slovenia) 29.

Ciaburro, G. (2017). *MATLAB for Machine Learning: Functions, algorithms, and use cases.* Packt Publishing.

Clay, F. (2016, May). *Getting started with Negative Binomial Regression Modeling*. (University of Virginia Library, Editor) Retrieved August 31, 2021, from data.library.virginia.edu: https://data.library.virginia.edu/getting-started-with-negative-binomial-regression-modeling/.

Cui, S., Ling, P., & Zhu, H. (2018). *Plant Pest Detection Using an Artificial Nose System: A Review.* Sensors. doi:10.3390/s18020378.

Da Cruz Figueira, A., Pitombo, C., Meira, P., De Oliveira, S., & Camargo Larocca, A. (2017). *Identification of rules induced through decision tree algorithm for detection of traffic accidents with victims: A study case from Brazil.* Case Studies on Transport Policy 5(2). doi:10.1016/j.cstp.2017.02.004.

Das, S. (2014). *Investigating the Pattern of Traffic Crashes Under Rainy Weather by Association Rules in Data Mining.* 93rd TRB Annual Meetings for Presentation and Publication under Safety 44.

Dörn, S. (2017). *Bayes-Netze.* Programmieren für Ingenieure und Naturwissenschaftler. Springer Verlag.

Efron, B. (1979). *Bootstrap methods: another look at the jackknife.* The Annals of Statistics 7(1). doi:10.1214/aos/1176344552.

European Commission. (2019, December 18). *Archive:Road safety statistics - characteristics at national and regional level*. Retrieved Agust 8, 2021, from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Road_safety_statistics_-_characteristics_at_national_and_regional_level&oldid=463733.

European Commission. (2020). *Handbook on the external costs of transport.* Version 2019 – 1.1, CE Delft, Directorate-General for Mobility and Transport.

Feng, M., Zheng, J., Ren, J., & Xi, Y. (2020). *Association Rule Mining for Road Traffic Accident Analysis: A Case Study from the UK.* Advances in Brain Inspired Cognitive Systems. doi:10.1007/978-3-030-39431-8_50.

Gao, Z., Pan, R., Yu, R., & Wang, X. (2018). *Research on Automated Modeling Algorithm Using Association Rules for Traffic Accidents.* 2018 IEEE International Conference on Big Data and Smart Computing (BigComp). doi:10.1109/BigComp.2018.00027.

García de Soto, B., Bumbacher, A., Deublein, M., & Adey, B. (2018). *Predicting Road Traffic Accidents using Artificial Neural Network Models.* Infrastructure Asset Management 5(4). doi:10.1680/jinam.17.00028.

Getahun, W., & Dejen, M. (2020). *The Application of Count Regression Models on Traffic Accidents in Case of Addis Ababa, Ethiopia.* Wollo University. Abyssinia Journal of Science and Technology 5(1). doi:10.20372/ajst.2020.5.1.55.

Gonzãlez-Manteiga, W., Sánchez, J., & Romo, J. (1994). *The bootstrap - A review.* Computational Statistics.

Gutierrez-Osorio, C., & Pedraza, C. (2020). *Modern data sources and techniques for analysis and forecast of road accidents: A review.* Journal of Traffic and Transportation Engineering (English Edition) 7(4). doi:10.1016/j.jtte.2020.05.002.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*. Waltham (USA): Morgan Kaufmann Publishers.

Hayatu, H., Abdullahi, M., Ahmad, B., Ali, Y., & Mohammed, U. (2020). *Feature Relevance Analysis and Classification of Kaduna State Road Traffic Accident Data using Machine Learning Techniques.* Data Mining Projects.

Helwig, N. (2017). *Clustering Methods*. Retrieved August 18, 2021, from http://users.stat.umn.edu/~helwig/notes/cluster-Notes.pdf.

Herry Conuslt, & KFV. (2017). *Unfallkostenrechnung Straße 2017 (UKR 2017).* Forschungsarbeiten des österreichischen Verkehrssicherheitsfonds, Wien: Bundesministerium für Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie (BMK).

Hosmer, W., & Lemeshow, S. (2000). *Applied logistic regression (2nd ed).* New York: John Wiley & Sons.

Jung, K., Lee, J., Gupta, V., & Cho, G. (2019). *Comparison of Bootstrap Confidence Interval Methods for GSCA Using a Monte Carlo Simulation*. Frontiers in Psychology 10. doi:10.3389/fpsyg.2019.02215.

KFV, & FGM. (2021). *Österreichische Verkehrssicherheitsstrategie 2021–2030*. Wien: Bundesministerium für Klimaschutz.

Kovera, A. (2017). *Machine Learning with Clustering: A Visual Guide for Beginners with Examples in Python 3.* Platform, CreateSpace Independent Publishing.

Krishnaveni, S., & Hemalath, M. (2011). *A Perspective Analysis of Traffic Accident using Data Mining Techniques.* International Journal of Computer Applications 23(7). doi:10.5120/2896-3788.

Kumar, S., & Toshniwal, D. (2016). *A data mining approach to characterize road accident locations.* Journal of Modern Transportation 24(1). doi:10.1007/s40534-016-0095-5.

Kumar, S., & Toshniwal, D. (2017). *A comparative analysis of heterogeneity in road accident data using data mining techniques.* Evolving Systems 8(2). Berling-Heidelberg: Springer Verlag. doi:10.1007/s12530-016-9165-5.

Labib, M., Rifat, A., Hossain, M., Das, A., & Nawrine, F. (2019). *Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning in Bangladesh.* 7th International Conference on Smart Computing & Communications (ICSCC). doi:10.1109/ICSCC.2019.8843640.

Lee, J., Yoon, T., Kwon, S., & Lee, J. (2019). *Model Evaluation for Forecasting Traffic Accident Severity in Rainy Seasons Using Machine Learning Algorithms: Seoul City Study*. Applied Science 10(1). doi:10.3390/app10010129.

Ma, W., & Yuan, Z. (2018). *Analysis and Comparison of Traffic Accident Regression Prediction Model.* 3rd International Conference on Electromechanical Control Technology and Transportation (ICECTT 2018).

Mackie, J. (1965). *Causes and Conditions.* American Philosophical Quarterly 2(4). University of Illinois Press on behalf of the North American Philosophical Publications. Retrieved from https://www.jstor.org/stable/20009173.

Mauro, R., De Luca, M., & Dell'Acqua, G. (2013). *Using a K-Means Clustering Algorithm to Examine Patterns of Vehicle Crashes in Before-After Analysis.* Canadian Center of Science and Education. Canadian Center of Science and Education. doi:10.5539/mas.v7n10p11.

Mehta, C., & Patel, N. (1983). *A Network Algorithm for Performing Fisher's Exact Test in r × c Contingency Tables.* Journal of the American Statistical Association 78(382). doi:10.2307/2288652.

Montella, A. (2011). *Identifying crash contributory factors at urban roundabouts and using association rules to explore their relationships to different crash types.* Accident Analysis and Prevention 43. doi:10.1016/j.aap.2011.02.023.

Nadler, F., Nadler, B., & Strnad, B. (2016). *ROAD SAFETY INSPECTION (RSI): Handbuch zur Durchführung von RSI.* Retrieved from https://www.bmk.gv.at/themen/verkehr/strasse/verkehrssicherheit/vsf/forschungsarbeiten/38_ris.html.

Nicholson, A., & Wong, Y. (1993). *Are accidents Poisson distributed? A statistical test.* USA: Accidents, Analysis & Pevention 25 (1). doi:10.1016/0001-4575(93)90100-B.

Nwankwo, C., & Godwin, N. (2015). *Statistical Model Of Road Traffic Crashes Data In.* International Journal of Scientific & Technology Research 4(9).

ÖAMTC. (2021). *Tempolimits auf Europas Straßen*. Retrieved July 15, 2021, from https://www.oeamtc.at/thema/reiseplanung/tempolimits-auf-europas-strassen-16186154.

Oberösterreichisches Rauminformationssystem. (2021). *DORIS interMAP - Unfallhäufungen*. (L. Oberösterreich, Editor) Retrieved July 8, 2021, from doris.at: https://wo.doris.at/weboffice/synserver?project=weboffice&client=core&user=guest&view=unfaelle.

Pan, C. (2016). *Deep Learning Fundamentals: An Introduction for Beginners*. CreateSpace Independent Publishing Platform.

Pei, X., Sze, N., Wong, S., & Yao, D. (2016). *Bootstrap resampling approach to disaggregate analysis of road crashes in Hong Kong.* Accident Analysis & Prevention 95(B). doi:10.1016/j.aap.2015.06.007.

Pradhan, B., & Sameen, M. (2019). *Review of Traffic Accident Predictions with Neural Networks.* Laser Scanning Systems in Highway and Safety Assessment . Springer Verlag. doi:10.1007/978-3-030-10374-3_8.

Prasejito, J., & Musa, Z. (2016). *Modelling Zero-inflated Regression of Road Accidents at Johor Federea Road F001.* Geotechnics, Infrastructure and Geomatic Engineering. The 3rd International Conference on Civil and Environmental Engineering for Sustainability (IConCEES 2015). doi:10.1051/matecconf/20164703001.

Priya, S., & Agalya, R. (2018). *Association Rule Mining Approach to Analyze Road Accident Data.* 018 International Conference on Current Trends towards Converging Technologies (ICCTCT). doi:10.1109/ICCTCT.2018.8550950.

RVS 02.02.21. (2015). *Richtlinien für das Verkehrswesen: Verkehrssicherheitsuntersuchungen.* Wien: BMK, GZ.BMK-300.041/0031-IV/ST-ALG/2015, Österreichische Forschungsgesellschaft Straße-Schiene-Verkehr (FSV).

Saharan, S., & Baragona, R. (2017). *A cluster analysis on road traffic accidents using genetic algorithms.* The 4th International Conference on Mathematical Sciences, AIP Conf. Proc. 1830, 030002-1–030002-6. doi:10.1063/1.4980927.

Saltelli, A., Tarantola, S., Campolongo, F., & Ratto, M. (2004). *Sensitivity Analysis in Practice. A Guide to Assessing Scientific Methods.* New York: John Wiley & Sons.

Schlögl, M., & Stütz, R. (2019). *Methodological considerations with data uncertainty in road safety analysis.* Accident Analysis & Prevention. doi:10.1016/j.aap.2017.02.001.

Schölkopf, B., & Smola, A. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* Cambridge, Massachusetts, USA: Massachusetts Institute of Technology.

Shanthi, S., & Ramani, G. (2012). *Feature Relevance Analysis and Classification of Road Traffic Accident Data through Data Mining Techniques.* San Francisco (USA): Proceedings of the World Congress on Engineering and Computer Science 2012 Vol I.

Statistik Austria. (2015, May 26). *Standard-Dokumentationen (Definitionen, Erläuterungen, Methoden, Qualität) zur Statistik der Straßeverkehrsunfälle ab 2012*. (B. S. Austria, Editor) Retrieved July 15, 2021, from statistik.at: https://www.statistik.at/web_de/dokumentationen/index.html.

Statistik Austria. (2019). *Straßenverkehrsunfälle 2019.* Wien: Bundesanstalt Statistik Österreich.

Statistik Austria. (2020). *Straßenverkehrsunfälle 2020 mit Personenschäden*. Retrieved July 8, 2021, from statistik.at: https://www.statistik.at/atlas/verkehrsunfall/.

Statistik Austria. (2021). *Standard-Dokumentation Metainformationen (Definitionen, Erläuterungen, Methoden, Qualität) zur Statistik der Straßenverkehrsunfälle.* Direktion Raumwirtschaft, Bereich Kraftfahrzeuge, Straßenverkehrssicherheit, Wien.

Tingvall, C., & Haworth, N. (1999). Vision Zero - An ethical approach to safety and mobility. *6th ITE International Conference Road Safety & Traffic Enforcement: Beyond 2000.* Melbourne.

Wegman, F., & Elsenaar, P. (1997). *Sustainable solution to improve road safety in the Netherlands. A 'polder model' for a considerably safer road traffic system.* Boston, USA: Contribution to the 67th Annual Meeting of the Institut of Transportation Engineers.

Weng, J., Zhu, J., Yan, X., & Liu, Z. (2016). *Investigation of work zone crash casualty patterns using association rules.* Accident Analysis and Prevention 92. doi:10.1016/j.aap.2016.03.017.

Wiharto, W., & Suryani, E. (2020). *The Comparison of Clustering Algorithms K-Means and Fuzzy C-Means for Segmentation Retinal Blood Vessels.* Acta Inform Med 2020 MAR 28(1). doi:10.5455/aim.2020.28.42-47.

Woolf, P. (2021, March). *Bayesian Network Theory*. Retrieved July 18, 2021, from libretexts.org: https://eng.libretexts.org/Bookshelves/Industrial_and_Systems_Engineering/Book%3A_Chemical_Process_Dynamics_and_Controls_(Woolf)/13%3A_Statistics_and_Probability_Background/13.05%3A_Bayesian_network_theory.

Yang, S., & Berdine, G. (2015). *The Negative Binomial regression*. (The Southwest Respiratory&Critical Care Chronicles, Editor) doi:10.12746/SWRCCC.V3I10.200.

Yu, R., Wang, G., Zheng, J., & Wang, H. (2013). *Urban Road Traffic Condition Pattern Recognition Based on Support Vector Machine.* Journal of Transportation Systems Engineering and Information Technology 13(1). doi:10.1016/S1570-6672(13)60097-5.

Zhou, X., Lu, P., Zheng, Z., Tolliver, D., & Keramati, A. (2020). *Accident Prediction Accuracy Assessment for Highway-Rail Grade Crossings Using Random Forest Algorithm Compared with Decision Tree.* Reliability Engineering & System Safety 200. doi:10.1016/j.ress.2020.106931.

Zong, F., Xu, H., & Zhang, H. (2013). *Prediction for Traffic Accident Severity: Comparing the Bayesian Network and Regression Models.* Mathematical Problems in Engineering. Application of Discrete Mathematics in Urban Transportation System Analysis. doi:10.1155/2013/475194.

13. Figures

14. Tables