Pre-crash scenarios at road junctions: A clustering method for car crash data

Philippe Nitsche[a], Pete Thomas[b], Rainer Stuetz[a], Ruth Welsh[b]

[a]*AIT Austrian Institute of Technology, Giefinggasse 2, 1210 Vienna, Austria*

[b]*Loughborough University, Epinal Way, Loughborough, LE11 3TU, UK*

Abstract

Given the recent advancements in autonomous driving functions, one of the main challenges is safe and efficient operation in complex traffic situations such as road junctions. There is a need for comprehensive testing, either in virtual simulation environments or on real-world test tracks. This paper presents a novel data analysis method including the preparation, analysis and visualization of car crash data, to identify the critical pre-crash scenarios at T- and four-legged junctions as a basis for testing the safety of automated driving systems. The presented method employs *k*-medoids to cluster historical junction crash data into distinct partitions and then applies the association rules algorithm to each cluster to specify the driving scenarios in more detail. The dataset used consists of 1056 junction crashes in the UK, which were exported from the in-depth "On-the-Spot" database. The study resulted in thirteen crash clusters for T-junctions, and six crash clusters for crossroads. Association rules revealed common crash characteristics, which were the basis for the scenario descriptions. The results support existing findings on road junction accidents and provide benchmark situations for safety performance tests in order to reduce the possible number parameter combinations.

Keywords:
Automated cars, Road safety, Intersections, Clustering, Car crashes, Pre-crash scenarios

1. Introduction

Over the past few years, automation of road vehicles has gained an increasing presence on the agendas of companies and public authorities, which have started to push Automated Driving Systems (ADS) into the forefront of research. On spots in a road network, where traffic conflicts are likely to occur, e.g. intersections, it must be ensured that automated vehicles can operate safely and efficiently, and even more important, that conventional vehicles driven by humans will have at least the same safety level as they have now. The technical reliability of ADS depends on the functionality under varying road infrastructure and transnational differences as well as on a safe interplay with traditional vehicles and vulnerable road users. Consequently, testing and validation procedures for those systems are paramount. There is a need for comprehensive testing, either in virtual simulation environments or on real-world test tracks. This leads to a challenge, namely to find the key driving situations to be evaluated. Since it is unrealistic

to cover all possible combinations of traffic situations and environment conditions, the most representative "benchmark" scenarios must be known.

As road intersections are locations, where the paths of multiple traffic participants are crossed, they are considered high-risk spots for safety researchers. For automated vehicles, road intersections of whatever type constitute a major point of interest along their routes due to the increased likelihood of conflicts with other road users. This paper presents a method to identify such conflict scenarios for the case of road junctions in the UK. It is important to note that the study excludes roundabouts and focuses on three-legged and four-legged intersections, both signalized and unsignalized. The study is based on 1056 junction crashes in the UK, which are initially partitioned by applying the k-medoids clustering method (Kaufman and Rousseeuw, 1990). As a second step, association rules (Agrawal et al., 1993) are computed to find associated crash attributes that ultimately build the scenario definition.

This paper is structured as follows: Section 2 gives an overview on relevant literature on road junction safety as well as clustering techniques. The proposed methodology is explained in section 3, followed by a description of the crash data used in this study (see Section 4). The cluster algorithm and association rule technique are given in the Sections 5 and 6, respectively. Section 7 comprises the results, before they are compared to existing findings and related to limitations and future work in Section 8. Finally, the paper is concluded in Section 9.

2. Background

2.1. Motivation and research objectives

Concerning road safety, it is still not clear what impact automated vehicles will have on crash risk, and what kinds of (new) risks they might cause. In particular, the safety risks coming with a mixed vehicle population, namely traffic with both driver-less and driver-operated vehicles are still subject to research. Although automated cars use sophisticated on-board sensors to recognize their environment, they have limitations, e.g. in challenging urban traffic situations, inclement weather conditions or when facing unexpected behaviour of traffic participants.

In Nitsche et al. (2014), an expert survey was conducted including questions on the role of road infrastructure, market readiness as well as to which extent certain factors influence the performance of selected automated driving functions on public roads. In summary, the main challenges found for ADS are complex urban environments, temporary work zones and poor visibility due to bad weather conditions. Road surface characteristics, road alignment and lighting were rated as minor influencing factors.

Three-legged and four-legged junctions are high-risk areas, which future automated cars should be capable to pass safely. Therefore, intersections play a particularly important role in testing assisted and automated driving. Automated vehicles should be capable of safely manoeuvring through an intersection and of avoiding or mitigating a collision. Intersection crash avoidance and mitigation systems (ICAMS) can be categorized into 1) infrastructure-only systems, such as active warning signs for drivers based on detected vehicles, 2) vehicle-based systems, including algorithms to predict and avoid collisions based on in-vehicle sensor data, 3) car-to-car systems based on vehicular communication and 4) cooperative infrastructure-to-vehicle communication systems (Mages, 2008). While the first system group is primarily made for human

drivers, automated vehicles mainly rely on vehicle-based systems, but may be assisted by cooperative systems.

The main research gap addressed by this work is that there are no standardized procedures for evaluating automated driving systems in junction environments. To this end, the research objective is to provide a set of pre-crash scenarios to understand typical high-risk situations at junctions. Due to a lack of accident data involving automated vehicles, a reasonable starting point is to analyse historical accidents with human drivers, assuming there is a certain overlap of crash risk. The study is preparatory research to a sub-microscopic simulation study, where virtual test drives will be conducted and ICAMS will be evaluated under varying conditions. The scenarios obtained in the underlying study will help to reduce the possible number of model parameter variations, such as vehicle trajectories, velocities, road and junction parameters etc.

2.2. Safety at road junctions

A query from the CARE crash database (ETSC, 2001) for the years 2003 to 2013 was analysed to get a picture about the intersection accident situation in the European Union. In general, it was found that every third road accident occurs at a junction. Four-legged intersections have the highest amount of both fatal and serious injuries with 43.9 and 43.2 percent, respectively. However, it must be noted that those percentages also depend on the exposure of different junction types, which has not been further analysed in this review. Due to the higher number of conflict points, four-legged junctions are generally unsafer than three-legged junctions (e.g. Bauer and Harwood, 1996; Harwood, 1995; David and Norman, 1975; Hanna et al., 1976). In this paper, safety-critical scenarios are obtained for three- and four-legged junctions, respectively, to further analyse this safety difference.

According to the CARE analysis, persons on pedal cycles and motorcycles were more often fatally injured at junctions than persons using other modes of transport. Every fourth fatally injured bicyclist was killed at a junction, while only every tenth fatally injured car occupant died due to a junction crash.

Van Maren (1980) reported that (multi-lane) unsignalized intersections have a lower number of crashes per million conflicts than signalized intersections. For signalized intersections, it was found that the dominant crash types are rear-end and head-on collisions (Polders et al., 2015; Obeng, 2007), however, Abdel-Aty et al. (2006) states that this also depends on the number of lanes and traffic volumes. In comparison to that, the majority of unsignalized intersection accidents are angle collisions (e.g. Molinero Martinez et al., 2008; Arndt, 2003; Layfield et al., 1996; Pickering and Hall, 1985). The most important variables affecting the safety of unsignalized intersections were studied by Haleem et al. (2010). Accordingly, these include the traffic volume on the major road and the existence of stop signs, and among the geometric characteristics, the configuration of the intersection, number of right and/or left turn lanes, median type on the major road, and left and right shoulder widths. In particular for angle crashes at unsignalized intersections, the factors were found to be traffic volume on the major road, the upstream distance to the nearest signalized intersection, the distance between successive unsignalized intersections, median type on the major approach, percentage of trucks on the major approach, size of the intersection and the geographic location within the state (Abdel-Aty and Haleem, 2011).

Several accident studies (Molinero Martinez et al., 2008; Lee et al., 2004; Najm et al., 2001) show that failure to yield right-of-way is the most dominant violation in crossing path scenarios.

This is followed by running a traffic signal or sign as one of the most frequent violations. Sandin (2009) concluded that the most common causation patterns include missed observation due to distraction or sight obstructions, which then led to no, late or premature action. Furthermore, a common causation was found to be incorrect prediction or faulty diagnosis, e.g. the drivers did not expect another vehicle to cross their path. Automated driving systems are expected to mostly solve the safety problems caused by those factors, e.g. through sensing and perception technologies. However, factors such as sight obstructions, unexpected road user behaviour and human error by other drivers still pose problems.

The method presented in this paper analyses historical accident data to understand the critical situations and factors at road junctions. Similar research has been conducted (e.g. Polders et al., 2015; Plavsic, 2010; Molinero Martinez et al., 2008; INTERSAFE, 2005; Wiltschko, 2004), however, the usage of k-medoids clustering and association rules in this context is novel.

2.3. Clustering accident data

In most cases, accident data as used in this study is of categorical nature, i.e. described by qualitative attributes (also called nominal attributes) of mainly arbitrary order. Although the categories can be coded as numbers, e.g. 1: female, 2: male, those numbers would not have mathematical meaning (e.g. Han et al., 2011; Lourenco et al., 2004). Therefore, dedicated statistical methods are necessary to analyse categorical data. Common clustering methods for categorical data are SQEEZER (He et al., 2002), ROCK (Guha et al., 1999), LIMBO (Andritsos et al., 2004), STIRR (Gibson et al., 1998), Link Clustering (Zengyou et al., 2005) or CACTUS (Ganti et al., 1999). Also, conventional clustering algorithms were modified to deal with categorical data, such as *k*-modes (Huang and Ng, 1999; Huang, 1997), *k*-histograms (Zengyou et al., 2003), *k*-medoids (Kaufman and Rousseeuw, 1990) or Generalized Self-Organizing Maps (Hsu, 2006), all of which have their advantages for different applications. Basically not a clustering method, but a popular classification algorithm for categorical data is Latent Class Analysis (Goodman, 1974), which is a model-based approach, assuming that a mixture of underlying probability distributions generates the data. Another approach is to use Multiple Correspondence Analysis (MCA, Lê et al., 2008) as a preprocessing step to transform the categorical variables to a continuous scale. Afterwards standard hierarchical or partitional clustering methods can be applied, usually only on the first principal components to reduce the dimensionality and stabilize the clustering by deleting the noise from the data.

As a popular and simple data mining technique, various researchers used association rules to discover patterns in their data (e.g. Weng et al., 2016; Kumar and Toshniwal, 2015; Montella, 2011; Mirabadi and Sharifian, 2010; Pande and Abdel-Aty, 2009). In this study, association rules are applied to clusters discovered by the *k*-medoids method to get more information on the underlying patterns of accident attributes, as explained in the following section.

3. Overall methodology

The methodology for evaluating the safety performance of assisted and automated driving systems is depicted in Figure 1. Depending on the objectives and contents of the test study, the target crash population and the safety performance indicators can be defined. This paper is devoted to the left half of the flow chart, with the objective to derive pre-crash scenarios for cars

at road junctions. Inspired by a study from Kumar and Toshniwal (2015), the idea was to initially partition the data by a clustering technique for categorical data, and then apply the association rule method on the data subsets to identify further parameters for the respective clusters.

The follow-up study will cover the right half of the chart, by evaluating the safety performance in a virtual simulation environment. The simulation models can be structured into 1) road environment models (including pavement, roadside and environmental conditions such as weather), 2) vehicle models (including sensor and control systems) and 3) driving (behaviour) models. Each of these model groups has numerous parameters to set, leading to a high number of possible combinations in the simulation runs. The method presented can aid engineers in parametrizing the models and to select the parameters that were found to be critical.

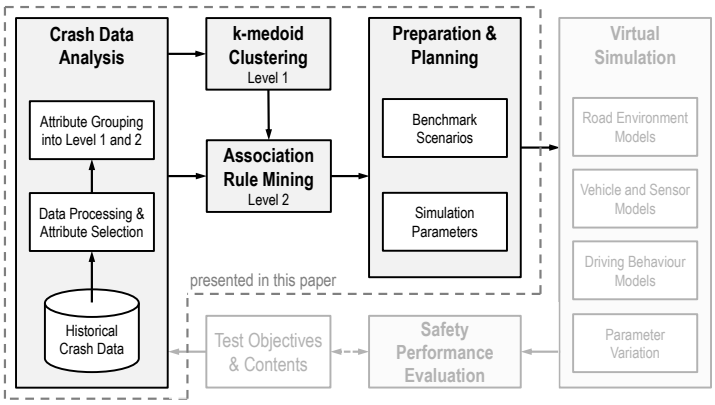

Figure 1: Overall methodology for evaluating the safety performance of assisted and automated driving systems

The crash data used and its processing steps are explained in Section 4, including the procedure of attribute selection, attribute coding and grouping into two levels. Level 1 is a reduced set of attributes describing the main collision parameters, for better partitioning and easier interpretation of the results, while Level 2 adds additional attributes describing the environment and causation factors. Level-1 data is used as input for the *k*-medoids clustering algorithm and level-2 data for finding association rules. The main reasons why this two-level approach was chosen are the following:

1. The *k*-medoids method achieved good clustering results on a smaller set of attributes. No clear partitioning was achieved when using all available attributes.
2. The results from applying the association rules on the whole dataset (without prior clustering) would be hard to interpret due to the high number of obtained rules. It must be noted that depending on the sample size and attribute dimensionality, millions of rules might be computed. This requires post-processing by applying dedicated algorithms or pruning techniques.

4. Data collection and processing

4.1. Background on the crash data used

The data used for this study stems from a project called OTS (On-The-Spot), which was commissioned by the UK Department for Transport and the Highways Agency (HA). It aimed to establish an in-depth research database of a representative sample of road accidents in the UK, to better understand the cause of accidents and injuries (Hill et al., 2001). Two crash investigation teams collected data from the years 1999 to 2010. One team was located at Loughborough University covering the South Nottinghamshire area in the East Midlands, and the other at the Transport Research Laboratories (TRL) covering the Thames Valley region.

The teams were responsible for collecting information at the scene of the accidents or, when the accidents already occurred, by liaison with emergency services, hospitals and local authorities. To arrive at the accident scene as quickly as possible, the teams had a direct link with the local police, and response vehicles driven by an OTS police officer were used (Cuerden et al., 2008). Data from both teams were collated into a single database that contains more than 2,000 variables.

4.2. Data collection

OTS is part of the RAIDS (Road accident in-depth studies) project, whose data query and export tool was used to download all necessary data elements including collisions with the following prerequisites:

- Junction type = “T or staggered junction”, “Crossroads”, “Multiple junction”, “Other junction” or “Using private drive or entrance”
- Police Accident Severity = “Fatal”, “Serious” or “Slight”[1]

As mentioned before, roundabouts were excluded from this study. The junction types included comprise signalized and unsignalized junctions of different shapes. This query resulted in 1056 crash cases from the OTS database, including more than 400 variables. However, it was decided to analyze the data on the car driver level, i.e. every sample corresponds to one driver involved in a crash, regardless if he/she was injured or not. This also means that every sample contains a car driven by the respective driver. Consequently, if two or more vehicles are involved in the same crash, the underlying crash and environment data is simply duplicated. Furthermore, there should be at least one car (including car-derived VANs, minibuses and SUVs) involved. This required a second query from the exported database as follows:

- Seating position of occupant = “Driver/Rider”
- At least 1 vehicle = “Car”
- Total number of vehicles ≥ 2

[1] It is important to mention that although the police reported a certain injury level, this might have been adapted by the crash investigation team based on more precise evidence.

This additional query resulted in an increased sample size of 1540, i.e. car drivers. The requirement of more than one vehicle means that single-car accidents were intended to be excluded, because speeding, fatigue or other human causation for single vehicle accidents are not relevant for the study. Another reason for having one record per driver is given by the background of the analysis, which focuses on safety risks involving automated vehicles instead of drivers. To this end, it is necessary to know the critical situations to be handled by drivers nowadays, as they are likely to happen to automated vehicles as well. Each sample is thus associated with an ego car, later denoted as car A, which collides with a secondary vehicle or road user, later denoted as B.

4.3. Attributes selection and coding

The number of variables was further reduced according to the following steps:

1. Include only variables that fit the scope of the study (see next section), e.g. not relevant were weekday or time of the crash, occupant data such as age or gender, vehicle damage or detailed injury data of different body parts.
2. Exclude variables with low variance, because they would fail to make a positive impact on model performance. In this study, all observations with more than 95 percent same values were excluded.
3. Group or combine highly correlated variables, e.g. OTS injury severity and police injury severity.
4. Exclude variables having unknown values in more than 30 percent of all samples.

Following this reduction process, the number of variables has been reduced to 41, which were grouped according to the original OTS data hierarchies “scene”, “vehicle” and “path”. The “scene” variables include general attributes about the crash, such as collision type and maximum injury of all involved persons. The “vehicle” variables are related to the pre-crash and collision circumstances from the perspective of the individual vehicle, i.e. driver, and includes for example the precipitating factor attributed to the vehicle, driver injury level or the pre-impact manoeuvre. The “path” variables describe the road environment, e.g. junction type, weather, traffic density or speed limit.

The original data contains variables in the following format: “Maximum injury level = Serious” from the four possible values uninjured, slight, serious and fatal. For the further calculations, all variables were converted to the binary-coded format. Consequently, this resulted in many more attributes, as each possible value was assigned to its own column, but it is a necessary step for applying most clustering algorithms.

The high number of attributes of the pre-processed OTS dataset made it necessary to further prepare the data for clustering. Usually, fewer attributes make it easier to interpret the clusters. Initial experiments with a varying number of attributes as input showed that the performance of the k-medoid method suffers from a higher dimensionality. Therefore, all attributes were divided into two levels as follows:

1. First level (5 variables, 25 attributes, see Table 1): This level of attributes was used as input for the k-medoids clustering. The idea is to derive clusters based on a set of main collision attributes first, before association rule mining is applied to each cluster with the second level attributes.

2. Second level (15 variables, 86 attributes, see Table 2): This level adds more detailed attributes on road infrastructure and accident causation to the level-1 attributes. They are intended to help tell a "story" describing each cluster by association rule mining.

Category	Short name	Description	Count	Freq.
Max. injury (of all persons involved in the crash)	MaxInj=Uninjured	No person injured (OTS injury level)	196	14.8%
	MaxInj=Slight	At least one person slightly injured (OTS injury level)	919	69.4%
	MaxInj=SeriousFatal	At least one person seriously or fatally injured (OTS injury level)	210	15.8%
Junction shape (attributed to the vehicle's path)	JctShp=X-minJoin	Road continues straight on with (minor) road joining from the left and right (crossroad)	224	16.9%
	JctShp=X-brkMaj	Road is temporarily broken by a (major) road passing across the vehicles path (Crossroad)	144	10.9%
	JctShp=NoJct	No junction present	20	1.5%
	JctShp=Other	Private drive, entrance or other junction type	7	0.5%
	JctShp=T-minLeft	Road continues straight on with (minor) road joining from the left	350	26.4%
	JctShp=T-minRight	Road continues straight on with an additional (minor) road joining from the right (T-Junction)	309	23.3%
	JctShp=T-termMaj	Road terminates with a (major) road passing across the vehicles path (T-Junction or accel. lane)	271	20.5%
First interaction (Road user type or object which the vehicle first interacted with)	1stIntAct=Car	Driver interacted with another car	987	74.5%
	1stIntAct=LGV-HGV	Driver interacted with a large or heavy goods vehicle	97	7.3%
	1stIntAct=PTW	Driver interacted with a powered two-wheeler (motorcycle or moped)	115	8.7%
	1stIntAct=Other	Driver interacted with another type of vehicle or object	37	2.8%
	1stIntAct=Cycle	Driver interacted with a bicyclist	50	3.8%
	1stIntAct=Pedestrian	Driver interacted with a pedestrian	39	2.9%
Manoeuvre (Action of the vehicle immediately before crash)	Manvr=GoingAheadOther	Driver was going straight ahead	781	58.9%
	Manvr=TurnL	Driver was turning left	59	4.5%
	Manvr=TurnR	Driver was turning right	79	6.0%
	Manvr=WaitTurnR	Driver was waiting to turn right	353	26.6%
	Manvr=Other	Driver was reversing, doing a u-turn, overtaking, undertaking, held up or waiting to turn left	53	4.0%
First point of impact (First point to come into contact with another vehicle, pedestrian or other object)	1stImpact=Back	First point of the impact was the car's back	126	9.5%
	1stImpact=Front	First point of the impact was the car's front	674	50.9%
	1stImpact=Nearside	First point of the impact was the car's nearside	218	16.5%
	1stImpact=Offside	First point of the impact was the car's offside	307	23.2%

Table 1: Crash attributes used for k-medoid clustering (level 1)

As described above, the second-level attributes deliver more information on the accident environment and causation. Most of the additional attribute groups in Table 2 are related to the vehicle's path describing the road layout, e.g. road type, speed limit or curvature. The attribute groups "collision code", "precipitating factor" and "driver injury" were added to the list to better understand the accident circumstances.

4.4. Further removal of unknowns

Samples with at least one unknown attribute value were removed as part of the data processing steps. This happened at two instances, namely 1) before computing the cluster with level-1 data and 2) before computing the rules with level-2 attributes for the data in each cluster. The first removal of unknowns resulted in a final sample size of $n = 1325$ for clustering, including $n = 930$ for T-junctions, $n = 368$ for crossroads and $n = 27$ for other or no junctions. The frequencies of the attributes are given on the right-hand side in Table 1. The second removal of unknowns was done on the extended level-2 dataset. Therefore, the final overall sample size ($n = 1070$) of the dataset used for the association rules is different to the clustering dataset (see Table 2).

5. Clustering of junction crashes

Due to different principles of clustering algorithms, one method might produce different clusters to another method. Hence, one has to choose the most appropriate method for the underlying dataset, taking into account the sample size, the number of attributes, the attribute types as well

Category	Short name	Description	Count	Rel. frequency
Collision type (The category letter of the UK STATS-19 collision code)	Coll=D-Cornering	Cornering (D)	16	1.5%
	Coll=H-CrossingNoTurns	Crossing (no turns) (H)	202	18.9%
	Coll=J-CrossingVehTurning	Crossing (vehicle turning) (J)	236	22.1%
	Coll=M-Manoeuvring	Manoeuvring (M)	104	9.7%
	Coll=Other	Other collision code	11	1.0%
	Coll=A-OvertakingLaneChange	Overtaking and lane change (A)	30	2.8%
	Coll=P-PedestrOther	Pedestrians Other (P)	25	2.3%
	Coll=F-RearEnd	Rear end (F)	188	17.6%
	Coll=L-RightTurnAgainst	Right turn against (L)	204	19.1%
	Coll=G-TurningVsSameDir	Turning versus same direction (G)	54	5.0%
Precipitating factor (The main cause of the crash, attributed to the respective occupant)	Prec=FailAvoidDriver	Driver failed to avoid object or vehicle on carriageway	64	6.0%
	Prec=FailAvoidOther	Other road user failed to avoid object or vehicle on carriageway	58	5.4%
	Prec=FailGiveWayDriver	Driver failed to give way	266	24.9%
	Prec=FailGiveWayOther	Other road user failed to give way	217	20.3%
	Prec=FailStopDriver	Driver failed to stop	84	7.9%
	Prec=FailStopOther	Other road user failed to stop	95	8.9%
	Prec=LossCntrDriver	Driver lost control of vehicle	23	2.1%
	Prec=LossCntrOther	Other road user lost control of vehicle	17	1.6%
	Prec=OtherDriver	Other precipitation by driver	27	2.5%
	Prec=OtherOther	Other precipitation by another road user	29	2.7%
	Prec=PedEnter	Pedestrian entered road without due care (driver not to blame)	17	1.6%
	Prec=PoorOvtkDriver	Inappropriate overtake by driver	7	0.7%
	Prec=PoorOvtkOther	Inappropriate overtake by other driver or rider	23	2.1%
	Prec=PoorMnvrDriver	Inappropriate turn or manoeuvre by driver	80	7.5%
	Prec=PoorMnvrOther	Inappropriate turn or manoeuvre by other driver or rider	63	5.9%
Driver injury (OTS injury level of the respective driver)	DrvInj=Uninjured	Driver suffered no injury	576	53.8%
	DrvInj=Slight	Driver was slightly injured	445	41.6%
	DrvInj=Serious	Driver was seriously injured	42	3.9%
	DrvInj=Fatal	Driver was fatally injured	7	0.7%
Area (around the crash location)	Area=Rural	Rural area (countryside, fields and only sparse housing)	368	34.4%
	Area=Urban	Urban area (at least one side of the road built up)	702	65.6%
Horizontal geometry (Qualitative assessment of curvature of road)	HorizGeom=Left	Left curve	22	2.1%
	HorizGeom=LeftSharp	Left sharp curve	4	0.4%
	HorizGeom=LeftSlight	Left slight curve	51	4.8%
	HorizGeom=Right	Right curve	25	2.3%
	HorizGeom=RightSharp	Right sharp curve	9	0.8%
	HorizGeom=RightSlight	Right slight curve	77	7.2%
	HorizGeom=Straight	Straight (no curve)	882	82.4%
Lighting (Light conditions at the time of the crash)	Light=DarkNSL	Darkness: no street lighting	50	4.7%
	Light=DarkSLUnk	Darkness: street lighting unknown	11	1.0%
	Light=DarkSL	Darkness: street lights lit	188	17.6%
	Light=DayNSL	Daylight: no streetlighting present	571	53.4%
	Light=DaySLUnk	Daylight: streetlighting unknown	243	22.7%
	Light=DaySL	Daylight: streetlights present	7	0.7%
Road type (on which the crash occurred)	RdType=DualCgw	Dual carriageway	161	15.0%
	RdType=OneWayStr	One way street	26	2.4%
	RdType=SingCgw	Single carriageway	883	82.5%
Speed limit (posted at the crash location)	SpdLim $\leq$ 20mph	20mph and less	1	0.1%
	SpdLim=30mph	30mph	584	54.6%
	SpdLim=40-50mph	40 or 50mph	270	25.2%
	SpdLim=60mph	60mph	159	14.9%
	SpdLim=70mph	70mph	56	5.2%
Surface (Road surface condition due to weather at the crash location)	Surf=Dry	Dry surface	673	62.9%
	Surf=Flood	Flooded surface	9	0.8%
	Surf=Icy	Icy surface	6	0.6%
	Surf=Snowy	Snowy surface	3	0.3%
	Surf=Wet	Wet surface	379	35.4%
Traffic control (Type of traffic control at the location of the crash)	TrfCtrl=None	No active or static yield instruction	582	54.4%
	TrfCtrl=GW	Static give-way instruction	245	22.9%
	TrfCtrl=Stop	Static stop instruction	14	1.3%
	TrfCtrl=Light	Traffic light control	229	21.4%

Table 2: Additional crash attributes used for association rule mining (level 2)

as the desired output of the study. The following sections address the clustering method chosen, which parameters were chosen and how it was applied to the OTS dataset.

5.1. The k-medoids method

The *k*-medoids method was chosen for the clustering, because it can cope with categorical data and is robust against outliers. It uses objects called medoids instead of centroids, as the popular *k*-means method does. Instead of using the mean as centre of the cluster, a member of the cluster is chosen as centre, whose average dissimilarity to all the objects in the cluster is minimal. In other words, the medoid is the most centrally located point in the cluster. Thus it is more robust to outliers, because it does not minimize a sum of squared Euclidean distances, as *k*-means does. Furthermore, *k*-medoids allows clustering categorical data, where a mean is impossible to define. For this reason, alternative dissimilarity measures can be applied, such as the "Hamming distance" (Hamming, 1950; Wegner, 1960) or the "Jaccard coefficient" (Jaccard, 1901).

One of the most powerful and commonly used algorithm for *k*-medoids is PAM (Partitioning Around Medoids) proposed by Kaufman and Rousseeuw (1990). It proceeds in two steps as follows:
Build step:

1. Choose *k* objects to become the medoids, or in case these objects were provided use them as the medoids
2. Calculate the dissimilarity matrix if it was not informed
3. Assign every object to its closest medoid

Swap step:

4. Within each cluster, each object is tested as a potential medoid by checking if the sum of within-cluster distances gets smaller using that object as the medoid. If so, the object is defined as a new medoid.
5. If at least one medoid has changed, go to (3), else end the algorithm.

The PAM algorithm works effectively for relatively small datasets such as the underlying OTS dataset. For larger datasets, alternative *k*-medoids algorithms should be used, such as CLARA (Clustering Large Applications, Kaufman and Rousseeuw, 1990).

5.2. Parameters used

The PAM algorithm was used, because it is most appropriate for the given sample size. The algorithm can produce better solutions than other *k*-medoids algorithms in some situations, but the computation times can be longer. The Hamming distance, originally used for the detection of errors in information transmission, was chosen as distance measure. It simply gives the number of mismatches between two vectors, thus it does not prefer 1s over 0s.

To study the separation of the resulting clusters, silhouette analysis (Rousseeuw, 1987) was used. Each cluster is represented by silhouette coefficients, which provide a measure of how close each point in one cluster is to points in the neighbouring clusters. Observations with silhouette coefficients near 1 are very well clustered. Small values indicate that the observation is close to

the decision boundary between to neighbouring clusters and observations with negative values are probably placed in the wrong cluster. The average silhouette width provides a measure for clustering validity, and is used to choose the most appropriate number of clusters.

The best number of clusters k was achieved by iteratively stepping from $k_{min} = 2$ to $k_{max} = 15$ clusters. Experiments with the dataset showed that a k_{max} greater than 15 does not result in any more change of the error function, as the curve flattens. The results from each k were compared to find the best k, i.e. the one with the lowest average silhouette value. Actually, finding the best k is one of the most debated problems in cluster analysis. In literature, various validity metrics can be found to compute the performance in partitioning, among which are the Akaikes Information Criterion (Akaike, 1974), the Bayesian Information Criterion (Schwarz, 1978), Calinski-Harabasz (Calinski and Harabasz, 1974) or Davies Bouldin index (Davies and Bouldin, 1979). For the scope of this study, it was sufficient to compare the silhouette values for graphical display for validating clusters. The entire clustering is displayed by combining the silhouettes into a single plot, as seen in Figure 2 (right) and Figure 3 (right) in a later section. The height of the silhouette represents the cluster size. For evaluating the best k, the average silhouette value of all objects within a cluster is calculated and compared to the others.

6. Specifying crash scenarios

As explained in the methodology section, the obtained clusters are further analysed by association rule mining, which was implemented in R by using the arules package (Hahsler et al., 2017, 2005). This section gives an overview on the principle of association rules and how the rules help to derive scenario parameters.

6.1. The association rules method

Association rule mining is a method to discover associations between attributes, also called "frequent itemset mining". A popular example of association rules is the market basket analysis, where retailers can get insights into which items are frequently purchased together so that marketing strategies and product shelving can be optimized. For example, if a customer buys "beer", then he/she often buys "crisps". This would be expressed as "beer $\rightarrow$ crisps", where the item "beer" is called the antecedent and the item "crisps" the consequent. One itemset I can contain multiple items. Applying the association rules terminology to the OTS dataset, then each sample is called a transaction $\{t_1, t_2, \ldots, t_n\} \in T$, and each attribute is an item $\{i_1, i_2, \ldots, i_m\} \in I$. An association rule can be written in the following mathematical form: $X \rightarrow Y$ where $X \subset I$, $Y \subset I$ and $X \cap Y = \emptyset$. Each rule is characterised by its support (see Equation 1) and its confidence (see Equation 2).

$$\text{supp}(X) = \frac{|\{t \in T; X \subseteq t\}|}{n} = P(X) \tag{1}$$

For itemsets, the support value gives the proportion of transactions t in the dataset, which contains the itemset X. For rules, the support is defined as the support of all items in the rule, i.e. $\text{supp}(X \rightarrow Y) = \text{supp}(X \cup Y) = P(X \wedge Y)$.

$$\text{conf}(X \rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)} = P(Y|X) \tag{2}$$

Equivalently, the confidence measures the strength of the rules and gives the conditional probability of the consequent Y given the antecendent X. In other words, it is the proportion of the transactions that contains X, which also contains Y. To explain the difference between the two measures, it is important to mention that two rules with flipped antecedent and consequent would both have the same support value. However, they would not have the same confidence, because the direction is taken into account.

The most common implementation was proposed by Agrawal et al. (1993), who called their method the Apriori algorithm. Accordingly, finding association rules involves two steps: 1) Find all frequent itemsets and 2) generate association rules from the frequent itemsets. The algorithm necessitates two parameters, namely a minimum support threshold, and a minimum confidence. By definition, if an itemset is below the minimum support threshold, then it is not frequent. If so, all its subsets must also be infrequent and can be pruned. In contrary, any subset of a frequent itemset must be frequent. By following this principle iteratively, the number of possible itemset configurations can be reduced tremendously with a simple algorithm.

The second step is to generate rules from the frequent itemsets found in Step 1. Here, the minimum confidence threshold comes into play: For each frequent itemset I, all nonempty subsets are generated. For every non-empty subset s of I, create the rule $s \rightarrow (I - s)$ if the minimum confidence for this rule is given. Since the rules are generated from frequent itemsets, each one also satisfies the minimum support. In this way, strong association rules can be found.

Depending on the data dimensionality, and on how low the minimum support and confidence thresholds have been set, the algorithm might produce millions of rules. Dedicated rule pruning and post-processing methods have been developed to find the rules of most interest. It was found that the confidence measure is a rather poor measure to discover the dependence of the consequent with respect to the antecedent (Guillaume et al., 1998; Silverstein et al., 1998). This paper uses a metric called lift (see Equation 3), also known as "interestingness".

$$\text{lift}(X \rightarrow Y) = \text{lift}(Y \rightarrow X) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X) \cdot \text{supp}(Y)} = \frac{P(X \wedge Y)}{P(X)P(Y)} \tag{3}$$

If the lift value is less than 1, then the occurrence of X is negatively correlated with the occurrence of Y, meaning that the occurrence of one likely leads to the absence of the other one. If the resulting value is greater than 1, then X and Y are positively correlated, meaning that the occurrence of one implies the occurrence of the other. If the lift equals 1, then X and Y are independent (Han et al., 2011). By setting an appropriate minimum lift value greater than 1, only high-lift rules can be extracted for interpretation.

6.2. Parameters used

The choice of the minimum support and confidence depends on the application and the expected outcome of the study. In theory, it is desirable to obtain rules with high support, high confidence and a lift value much greater than 1. The idea of this paper implies the analysis of certain accident situations and characteristics, which can be very rare (Montella et al., 2012).

After experimenting with different values, a minimum support of 0.03 was chosen, so that all itemsets occurring in less than 3 percent of the samples are disregarded. Choosing a lower threshold results in an increase of computation time and rules, which would all have to be interpreted. Choosing a higher support value might disregard relevant information about the clusters. There are different approaches in literature on the choice of a minimum confidence value. For example, Montella (2011) chose a threshold with conf=0.1 for their powered two-wheeler (PTW) study, which is much lower than usual. However, in this paper it is preferred to obtain rules, where the probability of the consequent given the antecedent is higher than 75 percent. Additionally, only rules with a lift>1.25 are considered for the results.

To further reduce the number of rules obtained, redundant rules were excluded according to the following procedure: A rule is redundant if a more general rule with the same or a higher lift exists. That is, a more specific rule is redundant if it is only equally or even less correlated than a more general rule. A rule is more general if it has the same consequent but one or more items removed from the antecedents. Formally, a rule $X \rightarrow Y$ is redundant if for $X' \subset X$: $\text{lift}(X' \rightarrow Y) \geq \text{lift}(X \rightarrow Y)$ (Hahsler et al., 2017).

7. Results

The crash dataset was divided into the two main junction types: 1) Three-legged T-junctions and 2) four-legged crossroads. For other types of junctions (e.g. private drives, pedestrian crossings), the sample size was too small (n=27) to compute clusters. This partitioning prior to clustering was done due to the scope of the study, namely to provide targeted scenarios and parameter variations for virtual vehicle simulations. The goal was not to find clusters characterized by junction types, but by driving situations, manoeuvres and injury outcome (see level-1 attributes). Furthermore, the number of intersection legs was found to be a significant variable to model intersection crashes (Abdel-Aty et al., 2006) and was used to group intersection crashes in various studies (e.g. Abdel-Aty and Haleem, 2011; Arndt, 2003; Persaud and Nguyen, 1998; Vogt and Bared, 1998).

7.1. Clusters found for T-junctions

The silhouette plot in Figure 2 (left) shows the average silhouette values (cluster validity) for all ks. In general, the higher the number of clusters the higher the silhouette values get. A higher number of clusters might be over-fitting and a lower number of clusters might be under-fitting. To find the best k, a compromise between cluster size and cluster validity had to be found. Association rules, which are computed for each cluster in the next step, were originally made for large-scale data. Hence, the goal was to avoid very small clusters, i.e. results with clusters containing less than 30 samples are disregarded (k=14 and k=15). Since k=13 has the highest average silhouette value with 0.383, the lowest number of samples that were allocated to the wrong cluster, and overall, the lowest percentage of clusters with negative silhouette values, it was chosen as most valid k.

Figure 2 (right) depicts the silhouette plot for each of the thirteen clusters, with one horizontal bar per sample within the cluster. Samples with a negative silhouette value might be assigned to the wrong cluster. However, the number of those samples is considerably low, expect for

cluster 4, where the average silhouette value suffers compared to the other clusters. Cluster 4 must therefore be treated carefully when interpreting the results.

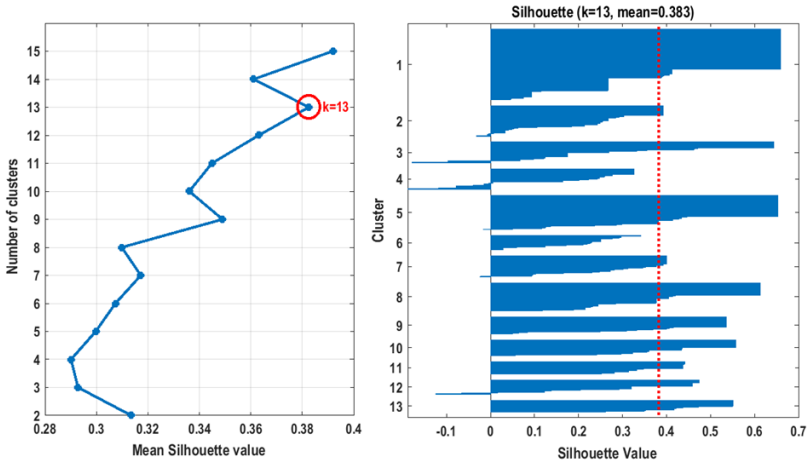

Figure 2: Mean silhouette values for all k's (left) and silhouette plot for k=13 (right) for T-junction clusters

The frequencies of each attribute within each cluster were compiled in a table to present the results at a glance (see Table 3). Cells shaded in grey indicate that the distribution of numbers for the given field is significantly different from the distribution in the whole population (χ^2-test with significance $\alpha = 0.05$) and that the particular number highlighted is over-represented. Due to values lower than 5 in the expected frequency table, the χ^2-test could not be applied to all observations.

Level-1 Attributes	**T-C1**	**T-C2**	**T-C3**	**T-C4**	**T-C5**	**T-C6**	**T-C7**	**T-C8**	**T-C9**	**T-C10**	**T-C11**	**T-C12**	**T-C13**
Sample size	**212**	**90**	**62**	**62**	**102**	**43**	**63**	**83**	**52**	**46**	**38**	**42**	**35**
MaxInj=Uninjured	0	11	7	8	15	9	63	7	0	0	4	2	5
MaxInj=Slight	212	69	52	42	78	30	0	68	45	0	29	0	25
MaxInj=SeriousFatal	0	10	3	12	9	4	0	8	7	46	5	40	5
JctShp=T-minLeft	195	0	1	0	0	0	58	3	51	41	0	1	0
JctShp=T-minRight	0	0	58	62	102	0	0	0	0	0	38	14	35
JctShp=T-termMaj	17	90	3	0	0	43	5	80	1	5	0	27	0
1stIntAct=Car	183	60	53	40	81	24	53	60	37	33	27	0	23
1stIntAct=LGV-HGV	18	6	5	4	5	2	7	2	5	5	4	3	2
1stIntAct=PTW	3	10	3	10	6	4	0	14	3	2	3	35	6
1stIntAct=Other	4	4	1	3	4	2	2	2	4	2	1	1	0
1stIntAct=Cycle	1	8	0	5	2	10	0	5	2	1	1	2	1
1stIntAct=Pedestrian	3	2	0	0	4	1	1	0	1	3	2	1	3
Manvr=GoingAheadOther	201	0	17	6	97	0	50	7	45	43	27	1	0
Manvr=Other	8	11	1	4	5	0	4	2	2	3	11	2	0
Manvr=TurnL	2	0	0	1	0	43	7	0	4	0	0	2	0
Manvr=TurnR	1	75	5	51	0	0	1	69	0	0	0	36	35
Manvr=WaitTurnR	0	4	39	0	0	0	1	5	1	0	0	1	0
1stImpact=Back	25	9	55	0	0	5	10	0	0	4	0	1	0
1stImpact=Front	162	68	1	0	102	18	35	0	0	37	0	11	35
1stImpact=Nearside	0	13	1	48	0	6	10	0	52	0	0	1	0
1stImpact=Offside	25	0	5	14	0	14	8	83	0	5	38	29	0

Table 3: Cluster results for T-junctions (k=13, n=930)

Cluster T-C1 is the largest cluster with a size of 212 crashes, from which all resulted in slight injury. More than 90 percent of the accidents occurred at T-junctions with a minor road

joining from the left. "1stImpact=Front" and "1stImpact=Back" are over-represented as well as "Manvr=GoingAheadOther". There is no clear indication on the collision type of this cluster, thus association rules are used for further analyses. The third largest **cluster T-C2** clearly groups collisions while turning, with a highly significant representativeness of frontal and nearside impacts, all of which occurring at roads terminated by a major road. Powered two-wheelers (PTW) and bicyclists have relatively high frequencies, but the car is still the dominant crash partner. **Cluster T-C3** with 62 samples represents car-to-car collisions at roads with minor roads joining from the right, mainly resulting in slight injury. Since there are mainly impacts on the back of the car, this cluster can be seen as rear-end crash group. **Cluster T-C4** occurred on a road with a minor road joining from the right, with nearside impacts in 77 percent of the cases and high frequencies for "Manvr=TurnR" and "1stIntAct=Car". The second largest cluster **Cluster T-C5** indicates rectangular collisions with another car crossing the cars trajectory from the right, although this assumption will be validated by association rule mining. **Cluster T-C6** is characterized by a left turn into a major road, which results in a collision mainly with another car. This cluster has a relatively high number of bicycle crashes (10). All 63 accidents in **Cluster T-C7** resulted in no injury for any of the participants. This is clearly a minor risk cluster mainly with cars and goods vehicles involved, with "Manvr=GoingAheadOther" having a high frequency. **Cluster T-C8** represents slight injury collisions with mainly other cars or PTW. Offside impacts were found over-represented, while turning right into a major road. **Cluster T-C9** involves nearside collisions only, which happened on a T junction with a minor road joining from the left, while the car was going straight. **Cluster T-C10** represents a group of high-risk collisions with serious or fatal injuries in all 46 cases. Front impacts are over-represented and "Manvr=GoingAheadOther" and "1stIntAct=Car" have high frequencies. Association rules will be used to analyse this cluster in more detail. In comparison to T-C9, **cluster T-C11** involves offside collisions only, which happened on a T junction with a minor road joining from the right, while the car was going straight or made another manoeuvre. Five of the 38 cases resulted in serious or fatal injury. **Cluster T-C12** is a PTW cluster, with 40 out of 42 collisions resulting in serious or fatal injury. In 85 percent of the cases, the car was turning right. Association rules will be used to analyse this cluster in more detail. The smallest **cluster T-C13** is characterized by right-turns into a minor road, with "1stImpact=Front" in all cases. Five of the 35 cases resulted in serious or fatal injury, which is most likely due to the six cases involving PTW. Association rules will be used to analyse this cluster in more detail.

7.2. Clusters found for four-legged junctions

For the crossroads dataset with 368 samples, k=6 was found to be most valid for separating the clusters, because it has a high average mean silhouette value of 0.395. The silhouette plot in Figure 3 (left) shows the average silhouette values for all ks. Although larger values were computed for higher ks (10-15), they were disregarded due to their small cluster sizes (<30) and possible overfitting. Figure 3 (right) depicts the silhouette plot for each of the six clusters, with one horizontal bar per sample within the cluster. The total mean silhouette value is higher and the number of samples with a negative value is lower compared to the T-junction dataset. This means that for the attributes and for the k chosen, the crossroads dataset seems to be better separated.

As for the T-junction dataset, the frequencies of each attribute within each cluster were compiled in a table to present the results at a glance (see Table 4). Cells shaded in grey indicate that the distribution of numbers for the given field is significantly different from the distribution in the

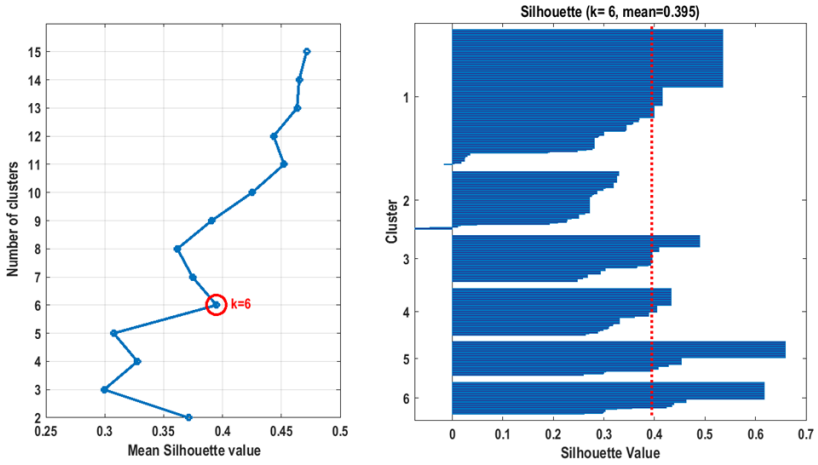

Figure 3: Mean silhouette values for all ks (left) and silhouette plot for k=6 (right) for four-legged junction clusters

whole population (χ^2-test with significance $\alpha = 0.05$) and that the particular number highlighted is over-represented.

Level-1 Attribute	**X-C1**	**X-C2**	**X-C3**	**X-C4**	**X-C5**	**X-C6**
Sample size	**142**	**60**	**48**	**49**	**35**	**34**
MaxInj=Uninjured	22	13	8	10	4	4
MaxInj=Slight	98	39	35	29	28	24
MaxInj=SeriousFatal	22	8	5	10	3	6
JctShp=X-minJoin	142	0	48	0	0	34
JctShp=X-brkMaj	0	60	0	49	35	0
1stIntAct=Car	118	44	38	39	30	28
1stIntAct=LGV-HGV	9	4	4	6	2	4
1stIntAct=PTW	3	7	1	3	1	0
1stIntAct=Other	3	1	1	0	0	2
1stIntAct=Cycle	2	2	4	1	1	0
1stIntAct=Pedestrian	7	2	0	0	1	0
Manvr=GoingAheadOther	116	32	25	35	25	29
Manvr=Other	4	0	0	0	0	1
Manvr=TurnL	5	9	2	2	1	1
Manvr=TurnR	15	19	21	12	9	3
Manvr=WaitTurnR	2	0	0	0	0	0
1stImpact=Back	12	5	0	0	0	0
1stImpact=Front	130	55	0	0	0	0
1stImpact=Nearside	0	0	48	0	35	0
1stImpact=Offside	0	0	0	49	0	34

Table 4: Cluster results for four-legged junctions (k=6, n=368)

Table 4 shows that the four-legged junction dataset is mainly separated by the type of junction and first point of impact. Experiments with varying parameters, such as initial medoid configuration or including the missing values did not result in different partitions. Including more attribute groups resulted in a decrease of the average silhouette value. For all clusters, the χ^2-test was not applied to the attribute groups "1stIntAct" and "Manvr" due to expected frequency values lower than 5. For the attribute group "1stImpact", only cluster X-C1 had sufficient frequency values for a χ^2-test. The distributions for injury level ("MaxInj") do not significantly differ in any cluster from the total population in their attribute group.

Cluster X-C1 is the largest cluster with 142 samples, which seems to mainly include rear-

end collisions, as the clusters X-C3 to X-C6 have no samples for "1stImpact=Back" and cluster 2 has only 5. **Cluster X-C2** groups situations on crossroads broken by a major road, with high numbers for turning left or right as well as "1stImpact=Front". Cars and PTWs were mostly involved. All situations in **Cluster X-C3** occurred on a road with minor roads joining from the left and right, and in all situations the car was hit on its nearside. All situations in **Cluster X-C4** occurred on a road broken by a major road passing the cars path, and in all situations the car was hit on its offside. All situations in **Cluster X-C5** occurred on a road broken by a major road passing the cars path, and in all situations the car was hit on its nearside, mainly by another car. As for the previous clusters, there is no statistical significance given for the manoeuvre, interaction or injury level distribution. The smallest **Cluster X-C6** represents collisions at roads with minor roads joining from left and right, where the car was hit on its offside, while going straight over the junction.

7.3. High-injury scenarios derived from association rules

For each identified cluster, association rules were computed using the parameters given in Section 6.2. In total, the analysis of each cluster resulted in 35 different crash scenarios comprising various parameters. Due to the high number obtained, not all of the rules for each cluster can be given in this paper. Therefore, only high-risk scenarios, which resulted in serious or fatal injury, are presented in this section, as they provide a set of safety-critical situations. More precisely, the further scenarios include crash situations from the T-junction clusters T-C4, T-C10, T-C12 and T-C13, and from the crossroads clusters X-C1, X-C2, X-C4 and X-C6. All rules obtained for each cluster are available as supplementary material to this paper.

As an example, Cluster T-C10 is selected for further explanation. Given the distributions in Table 3, the cluster can be described as follows: The car hits another car with its front resulting in serious or fatal injury, while going straight on a road with a minor road joining from the left.

A useful attribute to give a clearer indication about the crash circumstances is the collision type (indicated by letters A to Q in the OTS data specification, see Appendix A). For cluster T-C10, the collision types L ("Right Turn Against") and J ("Crossing with Vehicle Turning") were found to be the most frequent. Therefore, all rules containing those attributes within their items were further analysed to see which other attributes are associated with them.

Table 5 gives the 2-item and 3-item rules for T-C10 and collision type L, sorted by the five highest support values. The rules are sorted by the support to obtain the attributes that are often associated with each other. It can be seen that this collision type is associated with single carriageways (rule nr. 1) as well as with no traffic control ("TrfCtrl=None", see rule nr. 2, 4 and 11) and going straight ("Manvr=GoingAheadOther", see rule nr. 3). Another car as collision partner has already been defined by the cluster, but the rules reveal that "Coll=L_RightTurnAgainst" and "FirstIntAct=Car" are associated with dry surface (see rule nr. 5), uninjured driver of the ego car (see rule nr. 10), a fail to give way by the other car driver (see rule nr. 12), daylight (see rule nr. 13), 40-50 mph speed limit (see rule nr. 9) and urban area (see rule nr. 22).

Table 6 gives the 2-item and 3-item rules for T-C10 and collision type J, sorted by the five highest support values. It can be seen that this collision type is associated with a fail to give way by the other driver (see rule nr. 1). This combination is further associated with another car as collision partner (see rule nr. 5), no traffic control (see rule nr. 6), wet surface (see rule nr. 10), single carriageway (see rule nr. 11), rural area (see rule nr. 12), serious driver injury (see

Nr.	Antecedent	Consequent	Supp	Conf	Lift
1	Coll=L_RightTurnAgainst	RdType=SingCgw	0.237	0.818	1.413
2	Coll=L_RightTurnAgainst & TrfCtrl=None	RdType=SingCgw	0.237	1.000	1.727
3	Coll=L_RightTurnAgainst & Manvr=GoingAheadOther	RdType=SingCgw	0.237	0.900	1.555
4	Coll=L_RightTurnAgainst & RdType=SingCgw	TrfCtrl=None	0.237	1.000	1.357
5	Coll=L_RightTurnAgainst & Surf=Dry	FirstIntAct=Car	0.184	1.000	1.357
6	Coll=L_RightTurnAgainst & Surf=Dry	RdType=SingCgw	0.158	0.857	1.481
7	Coll=L_RightTurnAgainst & Area=Rural	RdType=SingCgw	0.158	0.857	1.481
8	Coll=L_RightTurnAgainst & DrvInj=Uninjured	RdType=SingCgw	0.132	1.000	1.727
9	Coll=L_RightTurnAgainst & SpdLim=40–50mph	FirstIntAct=Car	0.132	1.000	1.357
10	Coll=L_RightTurnAgainst & DrvInj=Uninjured	FirstIntAct=Car	0.132	1.000	1.357
11	Coll=L_RightTurnAgainst & DrvInj=Uninjured	TrfCtrl=None	0.132	1.000	1.357
12	Coll=L_RightTurnAgainst & Prec=FailGiveWayOther	FirstIntAct=Car	0.132	1.000	1.357
13	Coll=L_RightTurnAgainst & Light=DayNSL	FirstIntAct=Car	0.132	1.000	1.357
14	Coll=L_RightTurnAgainst & Light=DaySLUnk	RdType=SingCgw	0.105	1.000	1.727
15	Coll=L_RightTurnAgainst & SpdLim=40-50mph	Surf=Dry	0.105	0.800	1.448
16	Coll=L_RightTurnAgainst & DrvInj=Uninjured	Surf=Dry	0.105	0.800	1.448
17	Coll=L_RightTurnAgainst & Prec=FailGiveWayOther	Surf=Dry	0.105	0.800	1.448
18	Coll=L_RightTurnAgainst & Light=DayNSL	Surf=Dry	0.105	0.800	1.448
19	Coll=L_RightTurnAgainst & Light=DaySLUnk	TrfCtrl=None	0.105	1.000	1.357
20	Coll=L_RightTurnAgainst & Area=Urban	FirstIntAct=Car	0.105	1.000	1.357
21	Coll=L_RightTurnAgainst & Light=DaySLUnk	HorizGeom=Straight	0.105	1.000	1.267
22	Coll=L_RightTurnAgainst & Area=Urban	HorizGeom=Straight	0.105	1.000	1.267
23	Coll=L_RightTurnAgainst & Surf=Wet	HorizGeom=Straight	0.105	1.000	1.267

Table 5: Rules obtained for T-C10 with collision type L, sorted by the five highest support values

rule nr. 20) and 40–50 mph speed limit (see rule nr. 35). Taking a deeper look into the serious driver injuries, it can be noted that they are further associated with 40–50 mph speed limit (see rule nr. 27/28), wet surface (rule nr. 29) and single carriageway (rule nr. 30). However, this set of rules show that there is no clear indication on some attributes, such as the road type, as "RdType=DualCgw" is among the frequent items (see rules 42 to 45). Also, the driver can be uninjured or seriously injured or the area can be urban or rural. Those varying attributes could be used as varying parameter in the virtual simulation, while the others constitute the "static" environment and situation.

While the rules in the tables are relatively easy to interpret, this is no more the case with 4-, 5- or 6-item rules, also due to the high number of obtained rules. Therefore, each set of rules (comprising 2- to 6-item rules) was further visualized by directed graphs that were created from adjacency matrices of the associations found between all attributes. The graph was then reduced to the edges that direct to a certain consequent, represented by edge tables including source, target and weight of the edges. In this case, the targets (or consequents) were the collision types L (see Figure 4) and J (see Figure 5) and the sources were all remaining attributes. The weight or thickness of each edge represents the amount of associations identified between the respective antecedent node and the given consequent in the centre. In other words, nodes with thick edges indicate dominant crash attributes and thus define the scenario. For antecedent nodes that are not present in the graph, there were no associations found in the rules, thus they can be considered negligible for the respective scenario. Note that the graph does not reflect support, confidence or lift.

By visually inspecting the graphs and rules tables, the scenarios for this cluster can be described as follows (note that all crashes in the data occurred on UK roads with left-hand traffic):

Scenario T-10.1 (related to collision type L): Car *A* goes straight on a major road and hits another car *B* with its front, which is coming from the opposing direction and is turning right into a minor road. This happens on a single carriageway with a speed limit of 40 mph or 50 mph at an unsignalized junction, and is caused by *B* failing to give way. The surface is dry and *B* suffers serious or fatal injury.

Scenario T-10.2 (related to collision type J): Car *A* goes straight on a major road and hits

Nr.	Antecedent	Consequent	Supp	Conf	Lift
1	Coll=J_CrossingVehTurning	Prec=FailGiveWayOther	0.211	0.800	2.338
2	Coll=J_CrossingVehTurning	Light=DayNSL	0.211	0.800	1.520
3	Coll=J_CrossingVehTurning & Light=DayNSL	HorizGeom=Straight	0.211	1.000	1.267
4	Coll=J_CrossingVehTurning & HorizGeom=Straight	Light=DayNSL	0.211	1.000	1.900
5	Coll=J_CrossingVehTurning & FirstIntAct=Car	Prec=FailGiveWayOther	0.184	0.875	2.558
6	Coll=J_CrossingVehTurning & TrfCtrl=None	Prec=FailGiveWayOther	0.184	0.875	2.558
7	Coll=J_CrossingVehTurning & Area=Rural	Light=DayNSL	0.184	1.000	1.900
8	Coll=J_CrossingVehTurning & Area=Rural	HorizGeom=Straight	0.184	1.000	1.267
9	Coll=J_CrossingVehTurning & Prec=FailGiveWayOther	Surf=Wet	0.158	0.750	1.781
10	Coll=J_CrossingVehTurning & Surf=Wet	Prec=FailGiveWayOther	0.158	1.000	2.923
11	Coll=J_CrossingVehTurning & RdType=SingCgw	Prec=FailGiveWayOther	0.158	0.857	2.505
12	Coll=J_CrossingVehTurning & Area=Rural	Prec=FailGiveWayOther	0.158	0.857	2.505
13	Coll=J_CrossingVehTurning & Surf=Wet	Light=DayNSL	0.158	1.000	1.900
14	Coll=J_CrossingVehTurning & Light=DayNSL	Surf=Wet	0.158	0.750	1.781
15	Coll=J_CrossingVehTurning & Surf=Wet	Area=Rural	0.158	1.000	1.407
16	Coll=J_CrossingVehTurning & Area=Rural	Surf=Wet	0.158	0.857	2.036
17	Coll=J_CrossingVehTurning & Surf=Wet	HorizGeom=Straight	0.158	1.000	1.267
18	Coll=J_CrossingVehTurning & HorizGeom=Straight	Surf=Wet	0.158	0.750	1.781
19	Coll=J_CrossingVehTurning & TrfCtrl=None	RdType=SingCgw	0.158	0.750	1.295
20	Coll=J_CrossingVehTurning & DrvInj=Serious	Prec=FailGiveWayOther	0.105	1.000	2.923
21	Coll=J_CrossingVehTurning & SpdLim=30mph	Area=Urban	0.079	1.000	3.455
22	Coll=J_CrossingVehTurning & Area=Urban	SpdLim=30mph	0.079	1.000	3.800
23	Coll=J_CrossingVehTurning & SpdLim=30mph	Surf=Dry	0.079	1.000	1.810
24	Coll=J_CrossingVehTurning & Surf=Dry	SpdLim=30mph	0.079	0.750	2.850
25	Coll=J_CrossingVehTurning & SpdLim=30mph	RdType=SingCgw	0.079	1.000	1.727
26	Coll=J_CrossingVehTurning & SpdLim=30mph	TrfCtrl=None	0.079	1.000	1.357
27	Coll=J_CrossingVehTurning & DrvInj=Serious	SpdLim=40–50mph	0.079	0.750	2.375
28	Coll=J_CrossingVehTurning & SpdLim=40–50mph	DrvInj=Serious	0.079	1.000	3.455
29	Coll=J_CrossingVehTurning & DrvInj=Serious	Surf=Wet	0.079	0.750	1.781
30	Coll=J_CrossingVehTurning & DrvInj=Serious	RdType=SingCgw	0.079	0.750	1.295
31	Coll=J_CrossingVehTurning & Area=Urban	Surf=Dry	0.079	1.000	1.810
32	Coll=J_CrossingVehTurning & Surf=Dry	Area=Urban	0.079	0.750	2.591
33	Coll=J_CrossingVehTurning & Area=Urban	RdType=SingCgw	0.079	1.000	1.727
34	Coll=J_CrossingVehTurning & Area=Urban	TrfCtrl=None	0.079	1.000	1.357
35	Coll=J_CrossingVehTurning & SpdLim=40–50mph	Prec=FailGiveWayOther	0.079	1.000	2.923
36	Coll=J_CrossingVehTurning & SpdLim=40–50mph	Surf=Wet	0.079	1.000	2.375
37	Coll=J_CrossingVehTurning & SpdLim=40–50mph	Light=DayNSL	0.079	1.000	1.900
38	Coll=J_CrossingVehTurning & SpdLim=40–50mph	Area=Rural	0.079	1.000	1.407
39	Coll=J_CrossingVehTurning & SpdLim=40–50mph	HorizGeom=Straight	0.079	1.000	1.267
40	Coll=J_CrossingVehTurning & DrvInj=Uninjured	Light=DayNSL	0.079	1.000	1.900
41	Coll=J_CrossingVehTurning & DrvInj=Uninjured	HorizGeom=Straight	0.079	1.000	1.267
42	Coll=J_CrossingVehTurning & RdType=DualCgw	Light=DayNSL	0.079	1.000	1.900
43	Coll=J_CrossingVehTurning & RdType=DualCgw	Area=Rural	0.079	1.000	1.407
44	Coll=J_CrossingVehTurning & RdType=DualCgw	FirstIntAct=Car	0.079	1.000	1.357
45	Coll=J_CrossingVehTurning & RdType=DualCgw	HorizGeom=Straight	0.079	1.000	1.267
46	Coll=J_CrossingVehTurning & Surf=Dry	RdType=SingCgw	0.079	0.750	1.295

Table 6: Rules obtained for T-C10 with collision type J, sorted by the five highest support values

another car *B*, which is emerging from a minor road on the left with the intention to turn right. This happens on a single carriageway in a rural area with a speed limit of 40 mph or 50 mph at an unsignalized junction, and is caused by *B* failing to give way. The surface is wet and *A* suffers serious injury.

The same procedure was applied to the other clusters and their collision types. The Figures 6 and 7 illustrate all high-injury scenarios identified in a simplified manner to better understand the descriptions in the text. The red dots in the figures are the points of impact (i.e. front, offside or nearside). Surface conditions, area (rural,urban), speed limits, vehicle types and injury levels are not shown, but described in the following from the perspective of car *A*, i.e. the ego car associated with each sample.

Scenario T-4.1: Car *A* turns into a minor road and is hit by a PTW *B* on its nearside, which is going straight in the opposing direction. This happens on a single carriageway with 40–50 mph speed limit without active or static yield instruction and is caused by *A* failing to give way or manoeuvring inappropriately.

Scenario T-12.1: Car *A* turns right into a major road and is hit by a PTW *B* on the offside, which is going straight on the crossing path. This happens on a rural single carriageway controlled by a static give-way sign and is caused by *A* failing to give way. The surface is wet and *B* suffers serious or fatal injury.

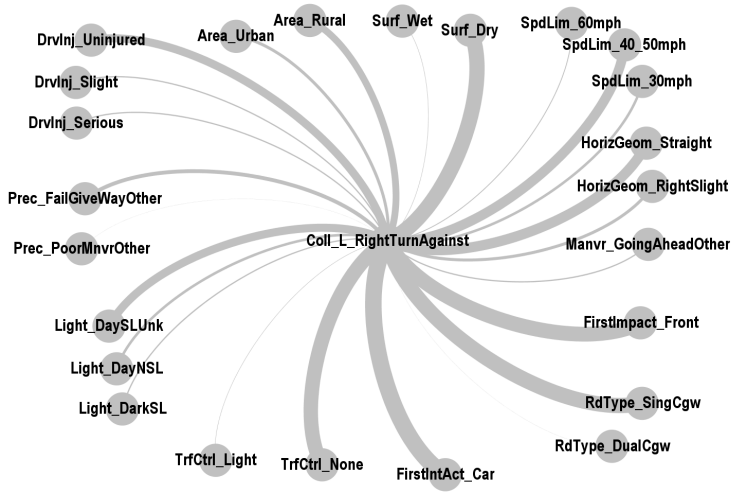

Figure 4: Weighted, directed graph obtained from all association rules for cluster T-C10 having collision type L as consequent

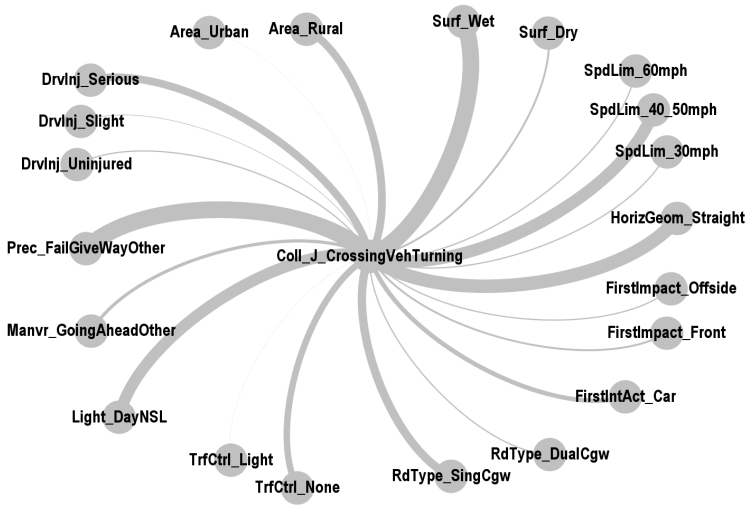

Figure 5: Weighted, directed graph obtained from all association rules for cluster T-C10 having collision type J as consequent

Scenario T-12.2: Car *A* turns right into a minor road and is hit on the offside by a PTW *B*, which is overtaking. This happens on an urban single carriageway with 30 mph speed limit without active or static yield instruction and is caused by an inappropriate overtake from *B*.

Scenario T-12.3: Car *A* turns left into a major road and is hit by a PTW *B* on its offside, which is going straight on the major road from the right. This happens on an urban single carriageway with 30 mph speed limit controlled by give-way signs and is caused by *A* failing to give way. *B*

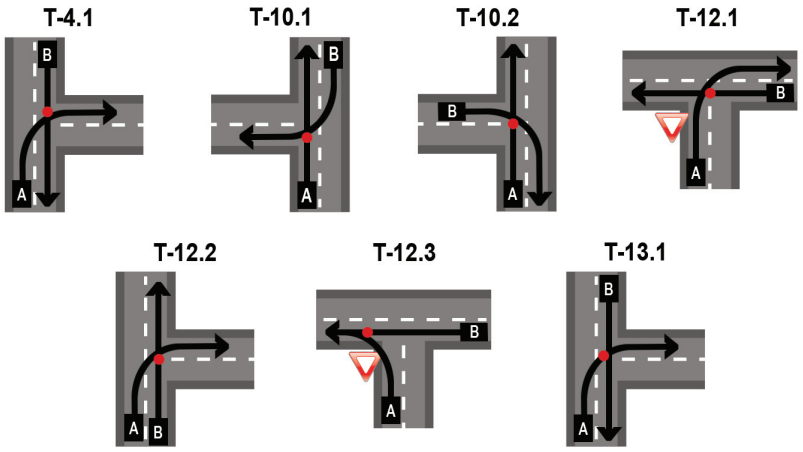

Figure 6: Simplified illustrations of all high-injury scenarios identified for three-legged junctions

suffers serious or fatal injury.

Scenario T-13.1: Car *A* turns into a minor road and hits a PTW *B* with its front, which is going straight in the opposing direction. This happens on a rural single carriageway with 30 to 50 mph speed limit without active or static yield instruction and is caused by *A* failing to give way or manoeuvring inappropriately. The surface is wet and *B* suffers serious or fatal injury.

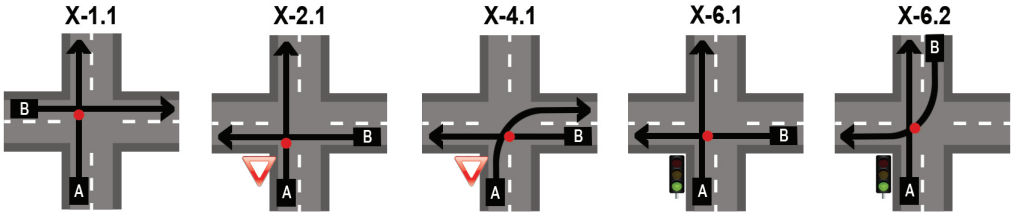

Figure 7: Simplified illustrations of all high-injury scenarios identified for four-legged junctions

Scenario X-1.1: Car *A* goes straight on a major road and hits another car *B* with its front, which is crossing the path from the left. This happens on a rural single carriageway with 60 mph speed limit without active or static yield instruction and is caused by *B* failing to give way.

Scenario X-2.1: Car *A* comes from a minor road and goes straight over a four-legged junction and hits another car or PTW *B* with its front, which crosses the path from the right. This happens on a rural road with 40–50 mph speed limit controlled by static give-way signs and is caused by *A* failing to give way.

Scenario X-4.1: Car *A* turns right into a major road and is hit by a car or LGV *B* on the offside, which is going straight on the major road from the right. This happens on a rural dual carriageway with 40–50 mph speed limit controlled by static give-way signs and is caused by *A* failing to give way. The surface is wet and *A* suffers serious or fatal injuries.

Scenario X-6.1: Car *A* goes straight on a major road and is hit by car *B* on the offside,

which comes from a minor road and crosses the path from the right. This happens on a single carriageway road with 30 mph speed limit controlled by traffic lights and is caused by *B* failing to give way. The surface is wet and *B* suffers serious or fatal injuries.

Scenario X-6.2: Car *A* goes straight on a major road and is hit by car *B* on its offside, which turns right from the opposing direction. This happens on a road with 60 mph speed limit controlled by traffic lights and is caused by *B* loosing control of the vehicle. *B* suffers serious or fatal injuries.

7.4. Comparison with high-frequency scenarios

This section compares the high-injury scenarios to the most frequent scenarios identified. Figure 8 and Figure 9 depict the top five high-frequency scenarios for three-legged and four-legged junctions, i.e. the scenarios with the highest number of crashes included. Table 7 shows the crash counts for each scenario including a short description. Some of the three-legged junction scenarios were combined due to their similarities. For example, T-2.1 and T-8.1, which were derived from two different clusters, were grouped. This was also done for the second and third most frequent scenarios for three-legged junctions. The count column in the table gives the number of crashes within the respective cluster that are allocated to the particular collision type. For example, the 44 samples for T-1.1 are the collisions of type F (rear-end) within cluster T-C1.

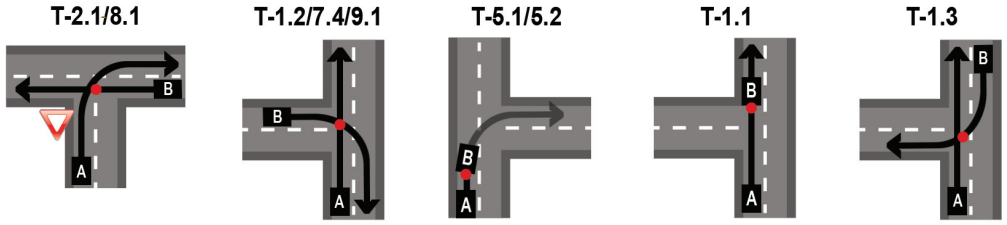

Figure 8: Simplified illustrations of the five most frequent scenarios identified for three-legged junctions

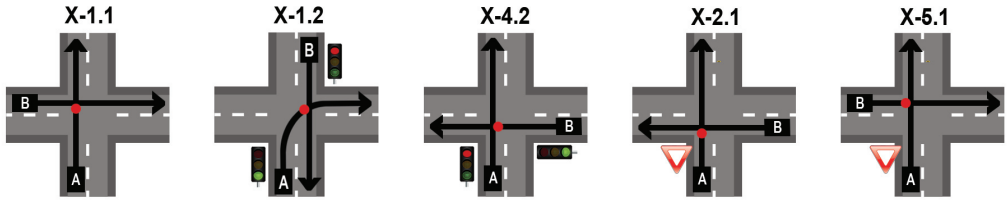

Figure 9: Simplified illustrations of the five most frequent scenarios identified for four-legged junctions

It can be observed that the top five most frequent scenarios at four-legged junctions do not include rear-end collisions. This finding corresponds to the crossing-path scenarios identified by Najm et al. (2001), which are primarily angle crashes. Furthermore, there is no particular scenario involving car-pedestrian or car-bicycle collisions only. This can be explained by the low number of pedestrians (2.4%) and cyclists (3.6%) as collision partners, compared to other cars (71.8%), motorcycles (9.7%) or goods vehicles (7.8%).

The low number of vulnerable road users is also a reason why the high-frequency scenarios at three-legged junctions do not include any of the high-injury scenarios. However, the three-legged junction scenarios include two rear-end collisions (T-5.1/5.2 and T-1.1), which are not included in the high-injury scenarios. This is due to the fact that the injury outcome was found to be lower for rear-end collisions than for angle collisions, which was also reported by Beck (2015).

Scenario	Count	Description
Three-legged junctions		
T-2.1/8.1	99	A turns right into major road and hits another car coming from the right. Road: Urban single carriageway. A fails to give way. Max. injury: Slight.
T-1.2/7.4/9.1	69	A goes straight and is hit by B turning right from a minor road on the left. Road: Urban, low-speed single carriageway. Traffic control: None. Max. injury: Slight.
T-5.1/5.2	55	A hits the rear of car B, which is waiting to turn right into a minor road. Road: Urban road. A fails to stop. Traffic control: None. Max. injury: Slight.
T-1.1	44	A hits the rear of car B travelling straight. Road: 70mph dual carriageway. A fails to avoid or stop. Traffic control: None. Max. injury: Slight.
T-1.3	42	A goes straight and is hit by B turning right into a minor road. Road: Single carriageway. Inappropriate manoeuvre from B. Light: Dark. Max. injury: Slight.
Four-legged junctions		
X-1.1	47	A goes straight on a major road and hits another car B crossing from the left. Road: Rural 60mph single carriageway.Traffic control: None. B fails to give way.
X-1.2	28	A turns right into a minor road and hits another car going straight in the opposing direction. Road: Urban road. B violates red light. Traffic control: Light. Max. injury: Slight.
X-4.2	24	A goes straight crossing a major road and is hit by another car B crossing from the right. Road: Urban single carriageway road. A violates the red light.
X-2.1	21	A goes straight and hits another car or PTW B crossing from the right. Road: Rural, 40–50 mph. Traffic control: Give-way sign. A fails to give way.
X-5.1	21	A goes straight crossing a major road and is hit by another car B going straight from the left. Road: 30mph single carriageway. A fails to give way. Traffic control: Give-way sign.

Table 7: High-frequency scenario descriptions

8. Discussion

8.1. Relation to existing findings

The pre-crash scenarios described above build the foundation for further research on testing assisted and automated vehicle technologies. This paper focussed on the scenarios with serious or fatal injury outcome, which were compared to the high-frequency scenarios. Although there is no doubt about the importance of vulnerable road user safety, neither the cluster analysis nor the association rule method resulted in a distinct pedestrian or cyclist scenario. Considering the frequency of certain crash types at junctions, car-pedestrian and car-cyclist collisions are discounted, which might not be true if injury frequencies were taken into account.

The method of clustering intersection crashes into distinct groups, including such a high number of variables as used in this study, is novel. Abdel-Aty et al. (2006) analysed numerous parameters to identify crash profiles for 45 different intersection configurations in Florida, however, this was made for different AADT values and numbers of lanes, which were not included in this study. Also, the objective of this study is different, because it aims at extracting relevant combinations of junction situations for simulation, while Abdel-Aty et al. (2006) provided crash profiles that assist in identifying intersections with specific problems. Therefore, the results cannot be directly compared.

Most existing research on intersection scenarios focussed on the classification of pre-crash manoeuvres, not combined with parameters about the road environment, collision partners, points of impact, injury types, causation factors and traffic control. Compared to literature, this study can be seen as more detailed in terms of crash circumstances. In the European INTERSAFE

project (INTERSAFE, 2005), intersection accidents were classified according to the pre-crash driving manoeuvres (in right-hand traffic). Twenty intersection situations were identified, from which the top five were: 1) *A* crossing path, with *B* coming from the left or right (which corresponds to the high-injury scenarios X-1.1, X-2.1 and X-6.1), 2) *A* turning left into the path of *B* coming from the left (see X-4.1), 3) *A* turning across the path of *B* coming from the opposite direction (see X-6.2, T-4.1, T-13.1), 4) *A* turning right into the path of *B* coming from the left (see T-12.3) and 5) *A* hitting the rear of *B* waiting to turn left (see the high-frequency scenarios T-1.1, T-1.2, T5.1).

The TRACE project identified six different scenarios at four-legged intersections from a statistical analysis of crashes in the European Union (Molinero Martinez et al., 2008). The scenario where *A* crosses the road and the trajectory of the opponent vehicle *B*, which is turning or going straight, is more frequent and more severe than any other. 70% of all intersection accidents belong to that scenario. This corresponds to the most frequent scenarios X-1.1, X-4.2, X-2.1 and X-5.1, from which X-1.1 was also found as one of the high-injury scenarios.

Of all intersection-related crashes analysed by Choi (2010), about 96 percent had critical reasons attributed to drivers, while critical reasons related to vehicle or environment were assigned in less than three percent of these crashes. Wiltschko (2004) concludes that ICAMS must be particularly designed to avoid red-light violations and fails to give way. This is also confirmed by this paper, since fails to give way are a precipitating factor in most scenarios.

8.2. Limitations and future work

At the moment, there are limited regulations on validating the reliability of highly automated road vehicles at junctions. This paper will contribute to the development of automated driving systems at junctions by providing evaluation scenarios for testing, taking into account the road and junction environment as well as the interplay with non-automated vehicles. Certain intersection layouts and design principles can facilitate a safe and reliable operation of automated vehicles, however, this study was done for the case where automated vehicles are expected to travel on existing roads without dedicated retrofitting.

A main limitation of this work is that the scenarios identified are based on human-related crash situations and do not necessarily reflect critical situations that come with sensor failure or misinterpretation of the automated driving control. Imagining that the ego car A operates automated, some scenarios such as rear-end crashes might be avoided by reliable environment perception and motion planning. Other scenarios comprise situations where human errors by other drivers or riders cause collisions, e.g. inappropriate overtakes, fail to stop or fail to give way. Future automated vehicles must also cope with the latter group of situations and must therefore be thoroughly tested, both in virtual environments and on public roads. Certainly, there may be different key testing scenarios depending on which issue is targeted. For example, targeting at maximum casualty reduction for vulnerable road users will require different testing measures than targeting at the vehicles' full functionality.

This study will be followed up by sub-microscopic simulation experiments conducted for the scenarios obtained, to evaluate the safety performance of ICAMS under varying conditions. The research further leads to recommendations on testing and validation procedures, with focus on virtual vehicle testing as a pre-stage or parallel activity to field operational tests on public roads, including static (e.g. road design and layout) and dynamic content (e.g. involved road users and

vehicles, their trajectories and behaviour).

9. Conclusions

This paper presents a novel approach on how to extract key pre-crash scenarios from accident data, which has been applied to three-legged and four-legged road junctions in the UK. The clustering method *k*-medoids was found to be most appropriate for the given dataset, since it is robust against outliers and can cope with categorical data. The study resulted in thirteen crash clusters for T-junctions and six crash clusters for four-legged junctions. Association rules were computed for each cluster and revealed associated crash characteristics, which were the basis for the scenario descriptions. Considering the clusters with high injury outcome, twelve pre-crash scenarios were identified, which constitute the core population of driving situations to be evaluated in virtual vehicle simulation. Failure to give way and inappropriate manoeuvres are among the main precipitating factors in the given dataset. In summary, the results support existing findings about junction safety and add further definition to the clusters identified. For example, as indicated in literature, higher injury levels coincide with powered two-wheelers involved as well as higher speed limits. The study is preparatory research to a sub-microscopic simulation study, where virtual test drives will be conducted and automated collision avoidance and mitigation systems will be evaluated under varying conditions. The scenarios obtained will help to reduce the possible number of model parameter variations, such as vehicle trajectories, velocities as well as road and junction parameters.

10. Acknowledgment

The research work for this paper was conducted in the scope of a joint PhD project between Loughborough University and the AIT Austrian Institute of Technology. The accident data in this study was acquired in cooperation with the UK Department for Transport that provided access to the RAIDS database. The Road Accident In Depth Studies (RAIDS) programme and associated database were commissioned by the United Kingdom Department for Transport in 2012 to consolidate data gathered from historic in depth collision investigation programmes dating back to the year 2000. Data collection is ongoing and since 2012, 1200 new cases have been investigated, the data is made available free of charge over the internet however conditional access is limited to those with a defined research need. For further information please contact RAIDS@dft.gov.uk.

References

Abdel-Aty, M. and Haleem, K. (2011). Analyzing angle crashes at unsignalized intersections using machine learning techniques. *Accident Analysis & Prevention*, 43(1):461–470.

Abdel-Aty, M., Lee, C., Wang, X., Nawathe, P., Keller, J., Kowdla, S., and Prasad, H. (2006). Identification of intersections' crash profiles/patterns. Final report, University of Central Florida.

Agrawal, R., Imielinski, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *Acm sigmod record*, volume 22, pages 207–216. ACM.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Andritsos, P., Tsaparas, P., Miller, R. J., and Sevcik, K. C. (2004). LIMBO: Scalable clustering of categorical data. In *EDBT*, pages 123–146. Springer.

Arndt, O. K. (2003). *Relationship Between Unsignalised Intersection Geometry and Accident Rates*. Dissertation, Queensland University of Technology.

Bauer, K. M. and Harwood, D. W. (1996). Statistical models of at-grade intersection accidents. FHWA-RD-96-125, Final Technical Report.

Beck, D. (2015). Investigation of Key Crash Types: Rear-end Crashes in Urban and Rural Environments. Research Report AP-R480-15, Austroads, Sydney, Australia. ISBN 978-1-925294-11-8.

Calinski, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, 3(1):1–27.

Choi, E.-H. (2010). Crash Factors in Intersection-Related Crashes: An On-Scene Perspective. NHTSA Technical Report DOT HS 811 366, U.S. Department of Transportation, National Highway Traffic Safety Administration, Washington, D.C., U.S.A.

Cuerden, R., Pittman, M., Dodson, E., and Hill, J. (2008). The UK On the Spot accident data collection study: Phase II report. Road Safety Research Report 73, Department for Transport, London.

David, N. A. and Norman, J. R. (1975). Motor vehicle accidents in relation to geometric and traffic features of highway intersections. FHWA-RD-76-128 Final Report.

Davies, D. L. and Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227.

ETSC (2001). EU Transport Accident, Incident and Casualty Databases - Current Status and Future Needs. Technical report, European Transport Safety Council, Brussels, Belgium.

Ganti, V., Gehrke, J., and Ramakrishnan, R. (1999). CACTUSclustering categorical data using summaries. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 73–83. ACM.

Gibson, D., Kleinberg, J. M., and Prabhakar, R. (1998). Clustering Categorical Data: An Approach Based on Dynamics Systems. In *Proceedings of the 24th VLDB Conference*, New York, NY, USA.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215–231.

Guha, S., Rastogi, R., and Shim, K. (1999). ROCK: A robust clustering algorithm for categorical attributes. In *Data Engineering, 1999. Proceedings., 15th International Conference on*, pages 512–521. IEEE.

Guillaume, S., Guillet, F., and Philipp, J. (1998). Improving the discovery of association rules with intensity of implication. In Åytkow, J. M. and Quafafou, M., editors, *Principles of Data Mining and Knowledge Discovery*, number 1510 in Lecture Notes in Computer Science, pages 318–327. Springer Berlin Heidelberg. DOI: 10.1007/BFb0094834.

Hahsler, M., Buchta, C., Gruen, B., and Hornik, K. (2017). *arules: Mining Association Rules and Frequent Itemsets*. R package version 1.5-2.

Hahsler, M., Gruen, B., and Hornik, K. (2005). arules – A computational environment for mining association rules and frequent item sets. *Journal of Statistical Software*, 14(15):1–25.

Haleem, K., Abdel-Aty, M., and Mackie, K. (2010). Using a reliability process to reduce uncertainty in predicting crashes at unsignalized intersections. *Accident Analysis & Prevention*, 42(2):654–666.

Hamming, R. W. (1950). Error detecting and error correcting codes. *The Bell System Technical Journal*, 29(2):147–160.

Han, J., Pei, J., and Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier. Google-Books-ID: pQws07tdpjoC.

Hanna, J. T., Flynn, T. E., and Tyler, W. L. (1976). Characteristics of intersection accidents in rural municipalities. *Transportation Research Record*, (601).

Harwood, D. W. (1995). Median Intersection Design. NCHRP Report 375, Transportation Research Board.

He, Z., Xu, X., and Deng, S. (2002). Squeezer: An efficient algorithm for clustering categorical data. *Journal of Computer Science and Technology*, 17(5):611–624.

Hill, J., Thomas, P., Smith, M., Byard, N., and Rillie, I. (2001). The methodology of on the spot accident investigations in the UK. In *Proceedings of 17th Conference on the Enhanced Safety of Vehicles (ESV)*, Amsterdam. National Highway Traffic Safety Administration, U.S. Department of Transportation.

Hsu, C.-C. (2006). Generalizing Self-Organizing Map for Categorical Data. *IEEE Transactions on Neural Networks*, 17(2):294–304.

Huang, Z. (1997). A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. In *SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, pages 1–8.

Huang, Z. and Ng, M. K. (1999). A fuzzy k-modes algorithm for clustering categorical data. *IEEE Transactions on Fuzzy Systems*, 7(4):446–452.

INTERSAFE (2005). Deliverable D40.4 - Requirements for intersection safety applications. Technical report.

Jaccard, P. (1901). tude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Socit Vaudoise des Sciences Naturelles*, 37:547–579.

Kaufman, L. and Rousseeuw, P. J., editors (1990). *Finding Groups in Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Kumar, S. and Toshniwal, D. (2015). A data mining framework to analyze road accident data. *Journal of Big Data*, 2(1).

Layfield, R. E., Summersgill, I., Hall, R. D., and Chatterjee, K. (1996). Accidents at urban priority crossroads and staggered junctions. TRL Technical Report 185, Transport Research Laboratory (TRL), Crowthorne, UK.

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software*, 25(1).

Lee, S. E., Knipling, R. R., DeHart, M. C., Perez, M. A., Holbrook, G. T., Brown, S. B., Stone, S. R., and Olson, R. L. (2004). Vehicle-based countermeasures for signal and stop sign violations. DOT HS 809 716.

Lourenco, F., Lobo, V., and Bacao, F. (2004). Binary-based similarity measures for categorical data and their application in Self-Organizing Maps.

Mages, M. A. (2008). *Top-Down-Funktionsentwicklung eines Einbiege- und Kreuzenassistenten*. Dissertation, Technische Universität Darmstadt.

Mirabadi, A. and Sharifian, S. (2010). Application of association rules in Iranian Railways (RAI) accident data analysis. *Safety Science*, 48(10):1427–1435.

Molinero Martinez, A., Carter, E., Naing, C. L., Simon, M. C., and Hermitte, T. (2008). Accident causation and pre-accidental driving situations: Part 1. Overview and general statistics. TRACE deliverable D2.1.

Montella, A. (2011). Identifying crash contributory factors at urban roundabouts and using association rules to explore their relationships to different crash types. *Accident Analysis & Prevention*, 43(4):1451–1463.

Montella, A., Aria, M., DAmbrosio, A., and Mauriello, F. (2012). Analysis of powered two-wheeler crashes in Italy by classification trees and rules discovery. *Accident Analysis & Prevention*, 49:58–72.

Najm, W. G., Smith, J. D., and Smith, D. L. (2001). Analysis of Crossing Path Crashes. DOT-VNTSC-NHTSA-01-03, U.S. Department of Transportation, National Highway Traffic Safety Administration.

Nitsche, P., Mocanu, I., and Reinthaler, M. (2014). Requirements on Tomorrows Road Infrastructure for Highly Automated Driving. In *The 3rd International Conference on Connected Vehicles & Expo (ICCVE 2014)*, Vienna, Austria.

Obeng, K. (2007). Some determinants of possible injuries in crashes at signalized intersections. *Journal of Safety Research*, 38(1):103–112.

Pande, A. and Abdel-Aty, M. (2009). Market basket analysis of crash data from large jurisdictions and its potential as a decision support tool. *Safety Science*, 47(1):145–154.

Persaud, B. and Nguyen, T. (1998). Disaggregate Safety Performance Models for Signalized Intersections on Ontario Provincial Roads. *Transportation Research Record: Journal of the Transportation Research Board*, 1635:113–120.

Pickering, D. and Hall, R. D. (1985). Accidents at rural T-junction. In *Planning & Transport Res & Comp, Sum Ann Mtg*.

Plavsic, M. (2010). *Analysis and Modeling of Driver Behavior for Assistance Systems at Road Intersections*. PhD thesis, TU München, München.

Polders, E., Daniels, S., Hermans, E., Brijs, T., and Wets, G. (2015). Crash Patterns at Signalized Intersections. *Transportation Research Record: Journal of the Transportation Research Board*, 2514:105–116.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Sandin, J. (2009). An analysis of common patterns in aggregated causation charts from intersection crashes. *Accident Analysis & Prevention*, 41(3):624–632.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464.

Silverstein, C., Brin, S., and Motwani, R. (1998). Beyond Market Baskets: Generalizing Association Rules to Dependence Rules. *Data Mining and Knowledge Discovery*, 2(1):39–68.

Van Maren, P. (1980). Correlation of Design and Control Characteristics with Accidents at Rural Multi-Lane Highway Intersections in Indiana: Interim Report. Interim report FHWA/IN/JHRP-77/22, Purdue University, Indiana State Highway Commission.

Vogt, A. and Bared, J. (1998). Accident Models for Two-Lane Rural Segments and Intersections. *Transportation Research Record: Journal of the Transportation Research Board*, 1635:18–29.

Wegner, P. (1960). A technique for counting ones in a binary computer. *Communications of the ACM*, 3(5):322.

Weng, J., Zhu, J.-Z., Yan, X., and Liu, Z. (2016). Investigation of work zone crash casualty patterns using association rules. *Accident Analysis & Prevention*, 92:43–52.

Wiltschko, T. (2004). *Sichere Information durch infrastrukturgestützte Fahrerassistenzsysteme zur Steigerung der Verkehrssicherheit an Strassenknotenpunkten*. Dissertation, Universität Stuttgart.

Zengyou, H., Xiaofei, X., and Shengchun, D. (2005). A Link Clustering Based Approach for Clustering Categorical Data. Technical report, China.

Zengyou, H., Xiaofei, X., Shengchun, D., and Bin, D. (2003). K-histograms: An efficient clustering algorithm for categorical dataset. Technical report, Department for Computer Science and Engineering, Harbin Institute of Technology, China.

Appendix A. Collision codes from STATS-19

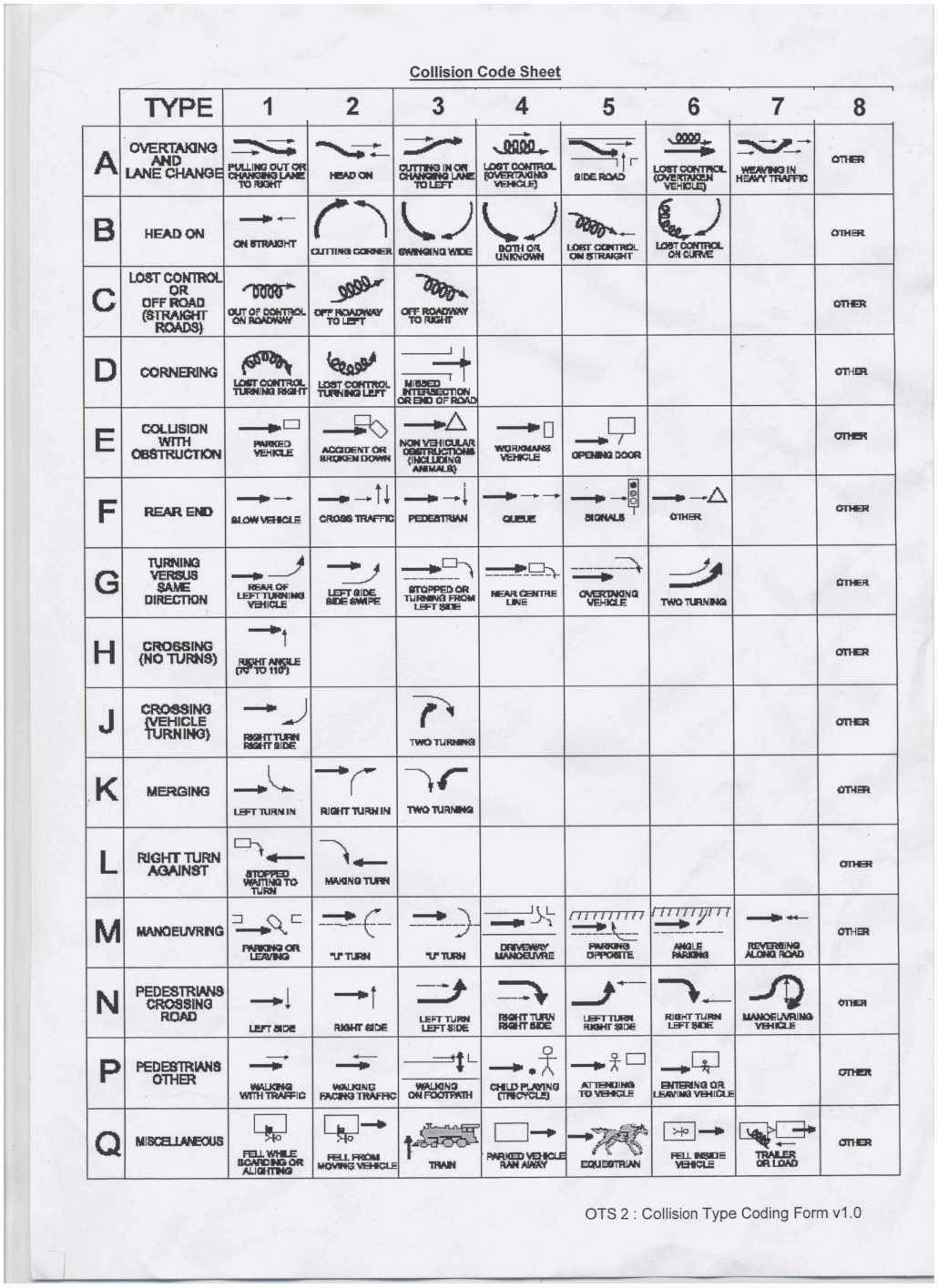

Figure A.10: Collision codes from STATS-19