ARTICLE IN PRESS

International Journal of Transportation Science and Technology xxx (xxxx) xxx

Contents lists available at ScienceDirect

International Journal of Transportation Science and Technology

journal homepage: www.elsevier.com/locate/ijtst

A data mining approach for traffic accidents, pattern extraction and test scenario generation for autonomous vehicles

Emre Esenturk *, Daniel Turley, Albert Wallace, Siddartha Khastgir, Paul Jennings

WMG, University of Warwick, Coventry, UK

ARTICLE INFO

Article history:
Received 12 November 2021
Received in revised form 1 October 2022
Accepted 9 October 2022
Available online xxxx

Keywords:
Accident analysis
Scenario development
Cluster analysis
ROCK algorithm
Market basket analysis
Geometric analysis of clusters

ABSTRACT

To effectively fight against traffic accidents, it is of great importance to analyse and understand the conditions that are linked with accidents. Such an analysis can serve as the basis to (i) develop *reactive* measures by finding the links between the pre-accident conditions (ii) devise *proactive* strategies that will prevent the occurrence of accidents by making the vehicles safer. This paper contributes to advancement of both approaches. For (i), one needs to identify the patterns in accidents. For (ii), introduction of Connected and Automated Vehicles (CAVs) is a promising solution. However CAVs need to be tested under numerous traffic scenarios to prove their safety before their deployment on public roads. This necessitates a great demand for high quality test scenarios for CAVs.

This paper achieves two goals. First, it analyses the past traffic accidents (UK's STATS19 database) to identify trends in the heterogeneous accident data and unravel the relationships between pre-accident conditions. This is done using a clustering algorithm (ROCK). Seven distinct large clusters emerge as a result. Each of these clusters are then further analysed for their meaning using the frequency analysis and geometric analysis. Secondly the paper underpins the proactive route (ii) by systematically developing, using the information in each cluster, test-case scenarios for CAVs which reflect the risk-prone conditions of the respective clusters. This is done using a data mining method (Market Basket algorithm) and further geometric interpretation of clusters. This way explicit scenarios are developed carrying the characteristics of the clusters that they come from.

1. Introduction

According to (Road Safety Data - STATS19, 2020), despite all efforts by the UK government and introduction of safety systems, between the years 2015–2021, the number of fatal and serious accidents on the UK road network consistently exceeded 1500 deaths and caused over 25,000 serious injury cases annually. Besides the heavy toll on the lives and wellbeing of the public traffic accidents incur huge costs to the nation's economy. An early report (UK Department of Transport, 2012) on the economic burden of traffic accidents on public states that, on average, *serious* accidents (is defined later) cost more than 200 k GBP (per accident) while fatal accidents cost close to 1.9 m GBP (per accident). These figures take into account

Peer review under responsibility of Tongji University and Tongji University Press.

* Corresponding author.
E-mail address: emre.esenturk@ndm.ox.ac.uk (E. Esenturk).

https://doi.org/10.1016/j.ijtst.2022.10.002
2046-0430/© 2022 Tongji University and Tongji University Press. Publishing Services by Elsevier B.V.

three main sources of loss: lost output (direct economic costs of loss output), medical and ambulance costs, and human costs (which reflect pain, grief and suffering). These tragic findings call for thorough understanding of the causes and characteristics of the accidents as well as effective ways to fight against their occurrence by minimising relevant factors.

Investigations on the sources of the accidents suggest that a large majority of them depend on both situational factors such as adverse weather conditions and human factors such as being distracted at the wheel or reacting to a situation too late. Hence, to prevent the occurrence of future traffic similar accidents, a two-pronged strategy is appropriate (i) by systematically identifying situational factors and trends in the data to develop counter measures (ii) by eliminating or at least minimising human driving using advanced technology. For (i), a careful analysis of the past accident records can be useful to decipher the hidden relationships between accident conditions and patterns. For (ii), advanced driver assistance systems (e.g. autonomous vehicles or connected vehicles) can be employed to mitigate hazards. In particular, one of the most promising candidate is connected autonomous vehicles or CAVs. However, for the deployment of CAVs, they must first be rigorously tested accross multitude of hazardous situations. To test CAVs in every possible variation of every possible situation would be laborious, superfluous, expensive and impractical (Kalra and Paddock, 2016). For this reason, alternative approaches have been developed which focus on the quality of scenarios/test miles driven (Feng et al., 2020), as opposed to merely the quantity; or study how a system fails (Delecki et al., 2022), instead of a system works. One approach that analyses how a system fails is the Hazard Based Testing (HBT) (Khastgir et al., 2018a). Test scenario generation as part of HBT can be done in two ways: (1) knowledge-based methods (e.g. using STPA (Systems-Theoretic Process Analysis), a hazard analysis technique (Leveson, 2012)) (Khastgir et al., 2018b), (Khastgir et al., 2021); (2) data-based strategies which can be executed as part of testing strategy (Xizhe et al., 2021). In this paper we explore the latter, data-based test scenario generation method.

Thus, this paper has two objectives underpinning the strategy outlined above. The first objective involves analysis of past accident data without enforcing any pre-set model or relationship structure between accident variables. This will allow the data to reveal itself in its naturality without any researcher bias and help discover the patterns in it. The second objective is, drawing from the built knowledge, to produce high quality test scenarios that will be used to train CAVs, equipping them with the necessary ''knowledge" before they are deployed on the roads. This way the safety of CAVs can be assured which will minimise the human factor from accidents. For the first objective, the sporadic data is differentiated into relatively homogenous clusters using a robust clustering algorithm (ROCK) (Guha et al. 1999) after some key modifications. To the best of our knowledge this is the first application of the ROCK algorithm in the context of accident analysis. Following this the clustered data is inspected more deeply to capture its intrinsic characteristics. This is done using frequency and geometric analyses. Such analyses can be useful, for instance, for policymakers to devise counter measures to mitigate the occurrence of high risk situations. For the second objective, specific data mining methods such as market basket analysis (Agrawal et al., 1993) and centroid data neighbourhood analysis are employed to produce a list of collection of attributes that describe accidents of particular types. Such collection of attributes carry, to the extent that the data can provide, characteristics of clusters that they come from and can be taken as the high-fidelity representations of the certain classes of accidents clusters they come from.

2. Background

The nature of traffic accidents is heterogeneous. Accidents, by definition are not meant to occur and fortunately they are relatively rare. Nonetheless, one can still identify features which are common between accidents. Knowledge of these similarities can be exploited to ascertain what factors lead to accidents. There is a sizeable literature on the relationship between the factors and the outcomes of accidents such as frequency or severity of accidents (Lord and Mannering 2010), (Mannering and Bhat, 2014). A number of statistical methods have been employed to investigate these relationships. To this end, much attention was given to supervised learning techniques. A non-exhaustive list of relatively recent studies are as follows.

For crash frequency prediction various type of regression models have been applied such as extensions of Poisson regression (Lord et al., 2005), (Lord et al., 2010), negative binomial regression (Caliendo et al., 2007), Tobit models (Anastasopoulos et al., 2008), multilevel Bayesian models (Xie et al., 2013), (Nowakowska, 2017). For severity prediction, logit/probit models, (Lee and Abdel-Aty, 2005), modified logit models (Esenturk et al., 2021), ordered and nested logit models (Yasmin and Eluru, 2013), (Yasmin et al., 2014), Bayesian logit models (Chen et al., 2016), (Yu and Abdel-Aty, 2014), (Ma et al., 2008), bivariate probit models (Lee and Abdel-Aty, 2008), neural-network models. For a review we refer the reader to (Savolainen et al., 2011). It is worth mentioning that, while generalised linear regression type models have made significant contributions to safety research, over the last years there has been a gradual shift towards their more advanced versions which allow variability in model parameters (instead of fixed parameters that the classical models assume). Such modern approaches have the advantage of accounting for unobserved heterogeneity in parameters providing better predictive versions of classical counterparts such as random parameter logit models (Milton et al., 2008; Eluru, 2008; Kim et al., 2010; Anastasopoulos et al., 2011), random parameter probit models (Paleti et al., 2010; Russo et al., 2014), random multinomial logit models (Hossain and Muromachi, 2012), multi-level Bayesian models (Huang and Abdel-Aty, 2010; Yu and Abdel-Aty, 2014), random parameter count models (Anastasopoulos and Mannering, 2009). Also, another shortcoming of the classical models has been their inability to accurately capture the nonlinear relationship between the variables (noting that they were generally semi-linear). This led to adoption of more advanced machine learning (ML) methods in recent years which have become prevalent predictive tools in safety research domain (Lord and Mannering, 2010). These ML type models include and neural-network models (Abdelwahab and Abdel-Aty, 2001; Formosa et al., 2020), (Chiou, 2006; Delen et al., 2006; Zeng et al., 2016), support vector machine models (Li et al., 2008; Li et al., 2012; Yu and Abdel-Aty, 2014).

In this strand of research, as a general goal, a deeper understanding of the specific types of accidents were sought such as pedestrian accidents, (Corbett & Morrongiello, 2017; Lee & Abdel-Aty, 2005), (Al-Ghamdi, 2002; Lefler & Gabler, 2004), bicycle accidents and near misses (Lee et al., 2017; Poulos et al., 2017) etc.). In a similar spirit, investigating the relationship between the accident types and intersection characteristics (Lee & Abdel-Aty, 2005), the specific causal factors related to the road (e.g. road structure (Meuleners et al., 2020) or weather conditions (e.g. fog (Al-Ghamdi, 2007)) have also received attention.

In this work we investigate crash patterns using an unsupervised learning algorithm. We adopt a two-stage strategy (briefly discussed in the introduction) where the first stage of the investigation is the 'analysis' or breaking down of the data using an unsupervised learning method, cluster analysis in particular. Having the data in the form of clusters, as opposed to a single heterogeneous dataset, makes discovering the relationships between variables (i.e., what features of accidents tend to imply others) significantly easier.

Cluster analysis is a classical learning method (Aggarwal & Zhai, 2012), but its use in the context of analysis of traffic accidents is relatively more recent (Kumar and Toshniwal, 2015). The advantage of this approach for this paper is that the patterns in the accident data emerge naturally rather than being imposed on the data by a pre-assumed model, as would be the case in most regression models. Some recent applications of clustering include crash analysis for road junctions (Nitsche et al., 2017), pedestrian crashes (Sun et al., 2019), (Tan et al., 2021), severity analysis (Esenturk et al., 2022) and investigation of automatic braking systems (Lenard et al., 2014), (Sui et al., 2019). Clustering analysis has also been used for large truck crash analysis (Rahimi et al., 2019) and classifying driving styles of travellers (Mohammadnzar et al., 2021).

The present paper contributes to the literature in multiple ways. First, it furthers the application of clustering approach in the accident analysis domain using a more suitable clustering algorithm. In particular, considering that the variables that describe the accident scenarios are of categorical form (as in many other databases) and that most accident datasets are notoriously noisy, which obscure the relationship between variables, a non-local clustering algorithm is more appropriate. Indeed, most traffic research studies to date have used k-means clustering (Nitsche et al., 2017), (Iranitalab and Khattak, 2017), or its variants such as k-medoids (Park and Jun, 2009). While k-means is a popularly used clustering algorithm for most applications it is not very suitable for categorical data as the central concept of k-means clustering, the average distance, loses its meaning for categorical data (decimal distances do not make sense). On the other hand, the more suitable candidates k-medoids or k-modes (which is more suitable for categorical data) suffer from the curse of dimensionality. For this reason, clustering algorithms in which microscopic distances are the base criterion for assigning the points to clusters are less well suited for accident data analysis. Instead, algorithms which are based on non-local properties beyond microscopic distances should be preferred. In this paper, a modified version of the ROCK (RObust Clustering with Links) algorithm (Guha et al., 1999) was used which clusters the data points based on their 'regional' properties, i.e., "links", rather than the pairwise distances. When dealing with categorical variables and a high number of dimensions (i.e., larger than 100) the ROCK algorithm is known to produce higher quality robust clusters.

Besides ROCK, there are other clustering algorithms that make use of predefined global properties (such as entropy or cluster width (e.g. COOLCAT, (Barbará et al., 2002), SQUEEZER (He et al., 2002), CLOPE (Yang et al., 2002)). While these algorithms generally work well with categorical data, they also have certain drawbacks. For instance, SQUEEZER is sensitive to the initial ordering of the data while CLOPE requires parameter fine tuning to detect the optimal clustering and its speed heavily depends on the tuning parameter which, sometimes, can lead to very slow clustering. The COOLCAT algorithm, on the other hand, requires (as in k-modes and k-medoids) pre-specification of the number of clusters which is apriori unknown. As for ROCK, it has no sensitivity issues (unlike SQUEEZER) as the algorithm goes through the entire dataset at each iteration (albeit at the cost of reduced speed). Moreover, ROCK, with the modifications proposed here, does not require parameter tuning which is definitely an advantage over CLOPE. Furthermore, the modified ROCK algorithm naturally terminates at a final cluster distribution. The final cluster number, with this approach, is not imposed apriori but depends on the structural parameters derived from the data. Hence, it saves one from the additional task of finding the best cluster number.

A second contribution of this work is that, instead of focusing on particular type of pre-crash situations (e.g. road junctions (Nitsche et al., 2017), pedestrian accidents (Lenard et al., 2014)), the present study aims at developing comprehensive set of associations (or rules) which yield quality test scenarios (derived from these 'rules'). These associations take into account a large number of factors from the major classes of explanatory attributes (physical, temporal, vehicular and human related) and are built on the top of clustering. Here association rules mining constitutes the 'synthesis' stage of our investigation where, for each cluster, the identified features are put together in a systematic and meaningful way. This defines, at the high level, the ingredients for representative scenarios inheriting the essential qualities of their respective clusters. More specifically, for this synthesis (of scenario) stage, the market basket analysis (MBA), was employed. This method, like clustering, does not assume any pre-defined relationship among the variables and extract association rules between the variables using the Apriori algorithm (Agrawal et al., 1993). This algorithm provides a quantitative measure for linking the accident variables which are otherwise rare or seem unrelated. It was successfully applied to a traffic accident dataset to derive rules for non-intersection crashes (Pande and Abdel-Aty, 2009), pedestrian crashes, (Das et al, 2019a) and rainy weather crashes (Das et al, 2019b). In (Nitsche et al., 2017) it was used alongside clustering for road junction accidents. In this work we use this method to extract a broader classes of traffic situations including much larger set of environmental factors. In the context of automated driving systems (ADS) categorization (Ulbrich et al., 2015), these scenarios correspond to 'logical scenarios'. Logical scenarios are the representations which use a state space and parameters for each entity in a scenario. For the scenarios derived in this paper, most parameters take categorical values (e.g., *road surface conditions* being *wet/damp; vehicle manoeuvre* being *turn left*). However, as the produced scenarios lack the detailed temporal description or specific numerical values due to the format of the original data (police reports) they may not be classified as fully concrete scenarios. Nevertheless, by randomly generating values for a small set of parameters fully concrete scenarios can easily be obtained. To give an example, *weather conditions*, in Stats19 data, is a categorical variable. If a constructed logical scenario had the (categorical) value of *fine/no wind* this would correspond to, in the ODD framework, to a *wind speed* in the range of [0–3.3 m/s]. To get a concrete scenario one would randomly generate wind speed that falls in this range.

A third contribution of the paper is that we also made use of the geometric information obtained from clusters which can be regarded as a method by itself. For the accident 'analysis' identification stage, in addition to using standard statistical tests (chi-squared), the centroid neighbourhood of a cluster was treated as the 'core' of the cluster from which a set of cluster defining variables can be deduced. For the synthesis of scenarios those that carry the characteristics of the significant or important variables were taken as the candidate scenarios.

In the broader context, concerning the research on test-scenario generation, much effort has been put, in recent years, for creating libraries with critical, high quality scenarios for CAV simulation and testing in order to reduce the need for on-road tests which are inefficient. However, deeming the scenarios as 'critical' is not straightforward task and various approaches have been developed such as designing a criticality measure (Feng et al., 2020), scenario randomisation (Khastgir et al., 2017), similarity analysis (Hemmati et al., 2010). The contribution of the present study is more similar, in spirit, to similarity analysis and criticality measure design in the sense that the clustering stage ensures the diversity/dissimilarity of scenarios while the thresholds set for the association rules mining give a kind of criticality measure for the exposure frequency and rarity of the scenarios.

Finally, in recent years, researchers began investigating, besides the physical factors (ambient physical conditions), the effects of socio-cultural origin on accidents. There is growing evidence that driving style is influenced by cultural factors (Factor et al., 2007). However, these effects can be considered inter-country level, that is, varies depending on the country legislations and informal rules observed by the majority of drivers in those countries (Sagberg et al., 2015). Since our dataset is based on UK accidents only, the accidents can be assumed to be relatively homogenous in terms of their societal influence across UK. For this reason, in this study, we have not specifically investigated socio-cultural effects.

3. Data format and its processing

3.1. Data collection

The dataset used in this study is known as "Road Safety Data - STATS19, 2020" which is a publicly available dataset updated yearly by the UK Department of Transport. It contains information about traffic accidents that are reported to the police, this includes the circumstances surrounding the accident, the vehicles involved and the subsequent casualties.

In this article, the years 2016–2018 are analysed with a reported 389238 accidents occurring in this time frame. Initially the accident data is stored in two separate files. The first covers the circumstances or conditions around which each accident takes place (sometimes referred to as the scenery elements) such *as road class, weather conditions*, and is denoted by AccD. The second dataset, which is denoted by VehD, covers information on each individual vehicle and driver that was involved in an accident such as vehicle type, vehicle manoeuvre, age band of the driver (vehicle and driver characteristics). The number of entries in VehD and AccD differ as there is usually more than one vehicle involved in each accident. This means, to include details of both the scenery and the vehicles together in the same analysis, the cardinalities of AccD and VehD need to be matched allowing them to be horizontally concatenated into one larger set. The match was done, for every point in VehD, by finding the corresponding accident in AccD and copying the variables over. This meant that the scenery variables were duplicated for each vehicle in an accident. Following this, any accidents involving more than two vehicles were removed. To justify this step the implicit assumption taken was that each vehicle acts independently (sometimes called the vehicular chaos assumption (Helbing, 2001). This is commonly used in traffic flow models and is valid for dilute traffic. Furthermore, any accidents that did not result in a physical impact on vehicles were also removed. This leaves the final cardinality of the data set to be 549,575 points which is the number of vehicles which got involved in a single vehicle accident or 2-vehicle accident that took place between 2016 and 2018 and resulted in physical impact on the vehicles.

Secondly, in this study, one takes the view that an accident is described sufficiently by the conditions that are present at the time and location of the incident. Any effects which are of societal and cultural origin have been discounted. Furthermore, it was decided at the outset that some categories from the raw data were not to be included (e.g. Local Authority District, Police Officer Attendance, Towing and Articulation, Vehicle Location etc.) as they were not deemed to be of crucial importance.

3.2. Form of the data and Pre-processing

The original data was recorded in unordered categorical form and was labelled in a subjective manner with many categories that were irrelevant to the analysis. Before beginning the analysis, a systematic cleaning of the data was carried out.

This involved merging some of the superfluous categories of similar kind (e.g. Minibus and Bus or Coach) or removing those which have zero frequency (unobserved) over the three years 2016–2018. Indeed, as far as the clustering and market basket analysis methods are concerned, unobserved categorical values have no effect on the results. Furthermore, for any incomplete data points, rather than discarding the whole accident, a categorical value was randomly generated based on the frequency distribution of the categories and assigned to the corresponding attribute. The original glossary for the raw data can be found in (''Road Safety Data - STATS19'', 2020).

Finally, at the pre-processing stage the data was kept in the categorical form. Once the pre-processing was complete, the categorical data was cast into binary (transaction) form which was more convenient for the purpose of analyses of Section 5.

3.3. Operational design domain for CAVs and behaviour competencies

In the CAV domain, the concept of Operational Design Domain (ODD) (BSI Pas 1883, 2020), (SAE J3016, 2018) is fundamental to evaluating and understanding the safety of the CAVs. ODD is defined as *''Operating conditions under which a given driving automation system or feature thereof is specifically designed to function, including, but not limited to, environmental, geographical, and time-of-day restrictions, and/or the requisite presence or absence of certain traffic or roadway characteristics*'' ODD is classified into three main attributes: scenery, environmental conditions and dynamic elements. The variables in Stats19 dataset can be easily matched to the ODD attributes, providing an insight into the types of conditions that lead to accidents and allowing us to test the CAV in these particular conditions to establish if it could handle the same scenario on the road. Another concept in the CAV domain, to define behaviour, is the concept of ''behaviour competencies'' (Thorn et al., 2018). Test scenarios for a CAV need to be a function of the CAV's ODD and its behaviour competencies. The Stats19 dataset also provides information about manoeuvres which can be mapped to behaviour competencies for a CAV. Together, a scenario can be constructed up to some number of secondary variables which are treated, in the ODD framework, as free parameters.

3.4. Crash data variables

Overall, 23 traffic variables were identified, which include ODD and behaviour competency variables: *Accident Severity, Skidding & Overturning, Day of Week, Time of Day, 1st Road Class (ODD), Carriageway Hazards (ODD), 2nd Road Class (ODD), Speed Limit (ODD), Junction Detail (ODD), Junction Location (ODD), Junction Control, Light Conditions (ODD), Weather Conditions (ODD), Urban or Rural area (ODD), Was Vehicle Left Hand Drive?, Vehicle Type, Vehicle Manoeuvre, 1st Point of Impact, Did Vehicle Leave the Carriageway, Sex of the Driver, Age Band of the Driver, Road Surface Condition, Weekend or Weekday?, Pedestrian Crossing Physical Facilities (ODD).* A list of all variables with their subcategories are found in Table 6.

4. Methodology and data analysis

In this section the two-stage methodology of analysing accidents and synthesising scenarios is detailed. The specific methods and the rationale behind them are discussed. For the data analysis step the ROCK algorithm is used to cluster the data followed by the frequency and centroid neighbourhood analyses to interpret and make sense of the clusters obtained. For the synthesis step the market basket method is applied to significant variables (defined later) on each cluster.

4.1. Clustering of accident data and the ROCK algorithm

Clustering is a well-studied data mining technique (Everitt et al., 2001) that attempts to group the data points in a given dataset. The underlying principle is that members of a cluster should be similar to each other and at the same time dissimilar to the rest of the data in some well-defined and meaningful way. There is a plethora of clustering algorithms. However, not all clustering methods are appropriate for the specific problem under consideration. Therefore, before choosing a specific method, it is worth contemplating on the format of the data in order not to end up with poorly defined clusters.

Firstly, as noted in Section 3, the processed Stats19 data is entirely categorical. This means popular clustering algorithms that are designed for continuous variables, such as k-means, are not ideal. This is partly because standard Euclidean or taxicab distance metrics are not meaningful for categorical data which is non-ordered. Even if one uses a distance measure that suits the categorical form of the data such as Jaccard distance or Hamming distance, one should be wary of the fact that the numerical operations that involve decimal values have no meaning for categorical data. For instance, for the k-means algorithm, if applied to categorical data, the centre of the cluster may not exist (decimal value). Scenarios could be placed in a cluster due to being 'similar' to the centroid but may have nothing in common with the other accidents in the cluster.

Secondly, the accident data is highly noisy. This blurs the clustering. For this reason, methods which give higher resolution should take precedence over the ones that offer high speed (but lower resolution). For noisy data clustering, algorithms that are based on ''microscopic'' properties of data points (e.g., k-modes or k-medoids using pairwise similarity) are certainly valid to use, however, they may produce low quality clusters. Furthermore, these algorithms are known to suffer from the curse of dimensionality since a single distant data point can skew the clustering. Instead, the problem of high noise can be dealt more effectively by using an algorithm that uses regional or global properties of the data to form the clusters such as ROCK (Guha et al, 1999).

A third consideration is the final number of clusters as the outcome of clustering operation. Most cluster analysis techniques work by prespecifying a number of clusters to be found, then allocating and reallocating data points to these clusters based on some algorithm (e.g. k-modes, COOLCAT). While this strategy is widely used and acceptable it assumes some apriori knowledge of the underlying accident data. It may lead to overclustering or underclustering. Nevertheless, there are methods to mitigate this issue and determine the optimal number of clusters in the data. For instance, one may run the clustering algorithm for different pre-specified cluster numbers and then apply some clustering quality metric to each case to determine what cluster number is 'best'. On the other hand, it would be preferable if the clustering algorithm stopped at a point which was naturally determined by the properties of the dataset itself.

Considering these issues and our data, it was decided a modified version of the ROCK Algorithm (Guha et al, 1999) was suitable to cluster our data. An additional advantage was that ROCK is also known to work well with high dimensional data and is robust against outliers, hence particularly suitable to the current data with 23 attributes with 101 values. ROCK is an agglomerative hierarchical clustering algorithm and is different to many clustering algorithms in that it does not directly use the similarity between pairs of points to make decisions about merging. Instead, in ROCK, the novel concept of links is introduced, where the linkage between any pair of points is defined as the number of neighbours they have in common. Then, merging decisions are taken based on the number of crosslinks between clusters. Two points are neighbours if their similarity is above a certain threshold θ. The motivation for clustering using links relies on that, in not well separated data such ours, an accident could be quite similar to another accident whilst actually belonging to a different cluster. However, it is unlikely that two points which share a large number of neighbours belong to different clusters.

In the implementation of the algorithm, as the data metric, the Jaccard measure was used to measure the similarity between accidents which is given by, for any data points.p_1, p_2

$$J(p_1, p_2) = \frac{|p_1 \cap p_2|}{|p_1 \cup p_2|},$$

that is, the similarity measure is just the size of the intersection set divided by the size of the union set. The Jaccard distance d_J is defined by $d_J(p_1, p_2) = 1 - J(p_1, p_2)$.

The algorithm begins with every point as its own single cluster. At each scan the two clusters which maximises the following goodness criteria is merged.

$$Goodness = \frac{link(C_i, C_j)}{(n_i + n_j)^{1+2f(\theta)} - {n_i}^{1+2f(\theta)} - {n_j}^{1+2f(\theta)}},$$

where ${n_i}^{f(\theta)}$ is the expected number of a cluster of size n_i, $link(C_i, C_j)$ is the number of actual crosslinks between the clusters C_i and C_j and denominator denotes the expected number of crosslinks. So, roughly, *goodness* is a normalised measure of how many shared neighbours (links) two clusters have. The denominator is needed to treat the large and small clusters 'fairly'. Indeed, a cluster pair should not get a higher goodness score just because they are large in size (and hence large in total number of links). The specific form of the denominator comes from the basic estimate of number of links ${n_i}^{1+2f(\theta)}$ in a given set of size n_i. The difference of terms in the denominator is to account only for the cross-cluster links but not the links within the clusters. The exponent $f(\theta)$ measures the dependence of expected number of links in a set given the size of the set and is based on the preset similarity threshold. The exact form of the function is not known, but it should be a decreasing function of θ. In (Guha et al., 1999), it is suggested that the function $f(\theta) = \frac{1-\theta}{1+\theta}$ generally works well.

When applying the ROCK algorithm to our problem, considerable modifications needed to be made. Firstly, from the numerical experiments carried out, it was observed that with the suggested exponent function $\frac{1-\theta}{1+\theta}$, the output clusters were unstable and also dependent on sample size, which is problematic. To remedy this, the θ dependence of f was derived empirically from the data. As can be seen from Fig. 1 the forms of the empirical f and the heuristic f are markedly different.

Secondly, the original algorithm stops at a preset cluster number. So, another modification to the algorithm was to not pre-specify the number of clusters but instead to find a natural stopping criteron. It was decided that a natural point for the algorithm to stop is when the Goodness measure is less than one, or equivalently, when the number of links after merging two clusters would be less than the expected number of links, i.e., the point where merging does not improve clustering. Finally, for the value of the similarity threshold θ, the original algorithm does not provide or suggest a specific range. Considering the sporadic nature of traffic accidents, a natural threshold value for the current problem was considered to be - the median of distances. One way to think about this was that pairs of points whose similarity were above median similarity were considered neighbours. Curiously, this also (roughly) corresponds to the case in which two data points are regarded as neighbours if they share more variables than they differ.

4.2 . Methodology for analysis of individual clusters

Once the optimal clustering is found, then the next task is to interpret the meanings of the individual clusters. This includes identifying those variables which best represent each cluster and finding the relationships between these significant variables to understand why they have emerged together in those specific clusters.

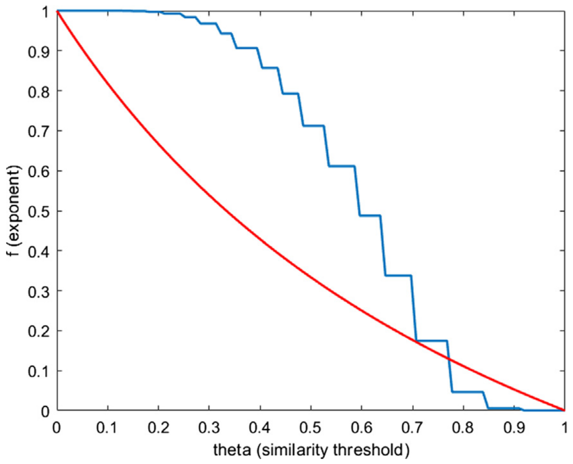

Fig. 1. Plot of $f(\theta)$: heuristic, $\frac{1-\theta}{1+\theta}$ (red); empiric (blue).

4.2.1 . Chi squared test for identification of significant variables

Because the data is categorical in nature and the variables within the cluster have frequencies, it is possible to analyse these frequencies to determine which variables are observed significantly more than expected in each cluster using how common they are in the full data over the three years as the comparison.

To find these significant variables, a chi-squared test is used. Since the data is in binary form, for every variable there are four different quantities needed: the observed number of 1′s in the cluster (O1), expected number of 1′s in the cluster (E1), observed number of 0′s in the cluster (O0) expected number of 0′s in the cluster (E0). The expected number of 1′s is given by the size of the cluster multiplied by the frequency of the variable in a comparison set divided by the size of the comparison set. This comparison set in our case was the full data. E0 is then given by clustersize-E1:

$$E1(var) = clustersize \times \frac{frq(var)}{N}$$

where N is the size of the full data and $frq(var))$ denotes the frequency of the variable var. The chi squared value for each variable is given by:

$$chi(var) = \frac{(O1(var) - E1(var))^2}{E1(var)} + \frac{(O0(var) - E0(var))^2}{E0(var)}$$

Any variables with a chi score over a critical threshold (given by a contingency table with p = 0.05) are deemed to be significant. After the significant variables are found, any variables with $relativefrequency := observed/expected$ less than 1.25 are omitted to leave only the variables which are highly overrepresented in the cluster.

4.2.2. Analysis of neighbourhood of centroids

To further interpret the clusters, an analysis of the centroids was carried out. In ROCK the centroid is simply the data point in the cluster that has the smallest average distance to all other points. This single point however may not be representative of the cluster as they become larger, due to the averaging effect over many points. To counter this a neighbourhood around the centroid is analysed. This allows the “essence” of the cluster to be analysed whilst incorporating the frequencies around the centroid. All the points within the median distance were chosen to be in the neighbourhood. The frequency of each variable in the neighbourhood was then found and normalised with respect to the cluster size.

The motivation of this analysis was to find the variables which represent each cluster. This was done by finding the variables that appeared in the cluster with over 1.5 times the percentage frequency of any other cluster. Variables that appeared within clusters with over 50 % frequency are also selected.

4.3. Methodology for synthesis of scenarios

4.3.1 . Market basket analysis for scenario generation

Market Basket analysis (MBA) was originally formulated (Agrawal, 1993) for use on transactional data to find which products customers commonly bought together, which would allow for more efficient planning and advertising around these products. In general, the aim is to find what are known as association rules between those variables that appear together unusually frequently. In the context of scenario development, transactional data refers to the set of conditions that are present (=1) or non-present (=0) for a given accident. It is then important to identify when particular set of conditions appear together in certain class of accidents. MBA helps to find such collections of risk generating conditions which, for our purposes, correspond to risky pre-crash scenarios.

Central to MBA is the concept of a k itemset. A k itemset is a subset of possible variables of length k. For example, in the shopping domain a 4- itemset could be {Bread, Milk, Eggs, Cheese} whilst in general a k-itemset is $\{x_1, x_2, \ldots, x_k\}$ which is a subset of some variables X. The apriori algorithm is used to find all frequent itemsets. An itemset is said to be frequent if it has a support that is over a minimum threshold. The support is given by the number of data points that all members of the itemset belong to, divided by the total number of data points. The support is a measure of how rare the itemset is and it is defined by.

$$Support = \frac{frq(X)}{N}$$

The next step is to find association rules within these frequent itemsets. To do this the item set is partitioned into two separate subsets, which are known as the antecedent and the consequent. To give a general example: The itemset X = $\{x_1, x_2, \ldots, x_k\}$ can be split into the antecedent $A = \{x_1, x_2, x_3\}$ and consequent $C = \{x_4, x_5, \ldots, x_k\}$ which would then give the association $A \rightarrow C$.

In order to quantify the strength of these rules two metrics are used, the confidence and lift. The confidence is given by the frequency of the whole original itemset divided by the frequency of the antecedent, i.e.,

$$Confidence = \frac{Support(X)}{Support(A)}$$

Intuitively if we have a confidence of 0.8 then for every accident with the antecedent present 80 % of them also contain the consequent. The lift, on the other hand, is calculated by the support of the whole itemset divided by the support of the antecedent multiplied by the support of the consequent.

$$Lift = \frac{Support(X)}{Support(A)Support(C)}$$

This means that if the lift is greater than 1 then, even if the rule has quite low confidence, it indicates that A is strongly associated with C relative to how often we'd expect them to appear together based on how common the individual variables are in the set.

4.3.2. Finding key scenarios in the neighbourhood of centroids

The centroids and the key variables around the centroids can be used for scenario generation purposes also. It was already noted that, due to the random nature of the data, not all data points in a given cluster carry the signature of that cluster. To eliminate such points we focus on a close neighbourhood of the centroid (the closest 10 % points of the full cluster) and select those data points of which more than half of the defining variables overlap with the important variables found for that cluster. Thus, only points with substantial resemblance to the cluster qualify as exemplary scenarios from the respective clusters.

5. Results

As outlined in the introduction and described in Section 4 the study is performed in two steps. Firstly, the data is broken down into relatively more homogenous clusters. Frequency analysis and geometric analysis are applied to interpret the clusters. Then, in the second stage a modified market basket analysis and geometric neighbourhood of the centroids of the clusters are used to extract pre-crash scenarios from each cluster.

5.1. Clustering of accident data

The purpose of this investigation was to abstract from real world data common occurrences and co-occurrences in road traffic accidents. In order to make claims about what relationships (between variables) can be deduced from the real-world data, it was considered important that any relationship should arise naturally from the data with as minimal prior assumptions as possible. Hence, it was decided not to impose any type of clustering property (e.g. final number of clusters) which would contain an intrinsic bias.

To test clustering quality, a combination of internal validation methods was employed for the clustering output which gave a quantitative measure to how well clustered the data was. The measures which estimated how well the data was separated into clusters were the Dunn Index and the Davies-Bouldin Index, whilst the Silhouette measure gave an average value to how well suited each point was to its respective cluster.

From a total of 549,575 accidents a random sample of 20,000 accidents was generated. Such sampling is typical when using the ROCK algorithm for large datasets (as the computational time scales with N^3, N being the sample size). The 1st step in implementing ROCK was to provide the algorithm with a similarity threshold. In the present study, although the ideal similarity threshold was derived directly from the data, additional numerical experiments were run to demonstrate that our ansatz is valid (see Table 1). The ideal value for similarity threshold was 0.36 which is just slightly above the median

Table 1
Quality Measures and final number of clusters varying with theta threshold similarity.$\theta = 0.36$ scores highest on Dunn and Silhouette so is the optimal theta to cluster this dataset.

Theta	0.35	0.36	0.4
Number of Clusters	9	26	76
Dunn	0.7601	0.8325	0.7642
DB	1.8416	1.5842	1.3694
Silhouette	0.0541	0.0799	0.0548

similarity of the data (as the ansatz from the previous section suggested). It is also the value, for the Jaccard measure, that, two data points share more attributes than they differ (12 attributes out 23). We note here that similar experiments were conducted with samples of smaller sizes (e.g. 5000, 10000) which yielded similar results and ensured that 20,000 data points give a good representation of the full set.

Using the empirically derived *f*, the ROCK algorithm yielded 26 clusters whose size distribution is shown in Table 2. Clusters 18, 19, 22, 23, 24, 25 and 26 were seen as large enough (the minimal cluster size was taken as 100) to be able to perform statistical analysis on. The remaining clusters, which make up a minority of the total data, may still be important as the seeds of some rare features. However, the application of statistical methods to these clusters might be misleading considering the rarity of some variables.

5.2. Interpretations of Clusters: Frequency analysis

Frequency analysis, as shown in Section 4, was performed to determine what variables more ‘significantly’ represents a given cluster. The idea was to compare the frequency of variables between each cluster and a reference data (entire dataset for years 2016–2018). Table 6 (see Appendix) lists the significant variables for the 7 largest clusters along with the variables’ relative frequencies with respect to the entire (processed) data (i.e., $f_{rel} = \frac{f_{clus}}{f_{ref}}$). For brevity, the following terminology is adopted: the variables which are significantly more frequent (compared to the reference set) are called *overrepresented* and those variables which are significantly less frequent are called *underrepresented*. Knowing which variables are overrepresented or underrepresented in a cluster, can allow us to determine the general characteristics and properties of that cluster.

It can be seen from Table 6 that Cluster 18 is a relatively small cluster (cluster size = 117) that has a significant overrepresentation of *fatal* accidents (6.92). This relatively large figure is partly due to *fatal* accidents being rare (around 1 in 100). Other overrepresented variables in the cluster help explain why this cluster might be much more likely to be fatal. For instance, both *skidded* and *overturned* are overrepresented here, with relative frequencies 4.97 and 8.19 respectively. Cluster 18 also appears to be a night time cluster with *9 pm-12am*, *12am-3am* and *3am-6am* all being overrepresented, as well as *darkness – lights lit* and *darkness – no lights* at the same time. This cluster also tends to involve *pedestrians* and *objects on the road*. Furthermore, it describes accidents that took place *off the junction*. In the UK database, *off-junction* accidents are those that take place more than a 20 m away from a junction and any type of intersection is denoted as a junction. A second observation related to Cluster 18 is that the drivers are generally either *very young* (age:0–20 years) or *old* (age:56 +), as both are overrepresented at the ratios of 2.38 and 1.57 respectively. By comparison, Cluster 19 is characterised by *serious* but not *fatal* accidents. It also contains a more than expected number of both *skidded* and *overturned variables.* However, this cluster seems to be a *junction* cluster whereas cluster 18 was not, which appears to be the main difference. The common junction types for this cluster have many arms, *crossroads* and *more-than-4-arms* in particular. Cluster 22 also comprises *serious* and *fatal* accidents at junctions (unlike Cluster 18) but on *minor roads* (unlike Cluster 19). The junction type for Cluster 22 is predominantly *roundabouts* involving *agricultural vehicles*. A common feature in Clusters 18, 19 and 22 is that *left hand drive vehicles* (abbreviated as LHD, are the vehicles in which the driver’s seat is on the left) are highly overrepresented (rel. freq. > 3.0). Considering that, in all of these clusters, severe accidents (*fatal* and *serious)* are overrepresented there might be a link between the accident *severity* and *left-handedness* of vehicles. Furthermore, *motorbikes* and *very young* are both overrepresented in all of these clusters which suggest a connection.

Clusters 23–26 are comparatively much larger clusters with at least 2000 data points in each. Cluster 23 is a definite *motorway* cluster depicting accidents that mostly happened on road segments with high *speed limit* and *away from junctions.*

Table 2
Distribution of clusters produced by the ROCK algorithm.

Cluster Id	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Cluster Size	1	1	2	1	2	2	3	2	7	1	5	15	9	12	2	23	3
Cluster Id	18	19	20	21	22	23	24	25	26								
Cluster Size	117	138	38	31	472	2225	5070	2678	9140								

Combining these with the facts that most accidents happened late night make it not a surprise *serious* and *fatal* accidents are overrepresented in this cluster. As it is an *off-junction* cluster the typical manoeuvres in this cluster are *changing lanes* (right or left) and *overtaking.* Cluster 24 is similar to Cluster 23 in that it is also mostly an *off-junction* cluster on a *motorway*, however, describe peculiar cases near the traffic lights where vehicles can be in *static* position (e.g. parked). Cluster 25 reflects accidents that are at major roads (*A-roads* and *Motorways),* intersections (e.g. *crossroads, roundabouts* and with low speed limit). The distinguishing manoeuvres are waiting and *moving off* which are compatible with the low *speed limit* variables (*20mph, 40 mph*). Cluster 26 describes accidents also at intersections of roads with low *speed limits.* The *high winds* in this cluster seem to have adverse effects on large vehicles such as *buses/trams.*

Another interesting result of the analysis is that mid-day accidents do not appear as significant in any of the clusters. This is partly due to the type of analysis conducted here (characterising the clusters based on their differences) and our choice of cut-off value for the minimal relative frequency of variables in clusters (set to 1.25 in our analysis) to deem them as significant. In particular, since *midday* is the most common time for accidents in general, each cluster contains a considerable proportion of such accidents and therefore midday periods do not appear as significant variables. However, should our cluster characterisation method put more emphasis on the presence of a variable in a cluster (rather than its relative difference) then such variables could come out as significant variables of the particular cluster in question (see Section 5.3).

5.3. Interpretations of Clusters: Geometric analysis

In Section 5.2 the significant variables that ‘‘describe” clusters were found using frequency tests with respect to the background non-clustered (entire) data. A more direct way of finding the key variables that illustrates each cluster is via geometric interpretation of each cluster. Since the clustering algorithm (ROCK) is metric based one can determine the centroid and its close neighbourhood which can be viewed as the ‘‘essence” of the cluster. Although the individual centroids may not entirely reflect the true meaning of the cluster due to randomness in the data its close neighbourhood does. Hence, as a general method we focus on the closest 50 % percent of data points (of the respective clusters) and select those points which have at least 50 % percent frequency in the neighbourhood of centroids or which are at least 50 % more frequently expressed than in all other clusters. This selection process gives a set of key variables for each cluster which partly overlap but are also different from the significant variables obtained via the frequency analysis. Such difference is natural and expected as *there is no single way of interpreting the clusters*. Table 3 below illustrates the key variables obtained from this geometric interpretation.

It is seen that the geometric interpretation of Cluster 18 is mostly aligned with the frequency interpretation in that Cluster 18 is a *night time, fatal* cluster describing accidents that take place *away from junctions*. Both interpretations suggest that the accidents are affected by adverse weather conditions such as *wet/damp, fog/mist, frost/ice.* The two methods differ in *left-hand drive* vehicles. This is because presence of such vehicles are very rare and therefore not detected by the geometric method but is caught by the frequency analysis as features in accidents in a cluster are selected in comparison to the entire (processed) data. So both methods are complementary to each other. For Cluster 19, as in frequency analysis, geometric analysis suggests that it is a low *speed limit* cluster *at a junction,* where *overtaking* and *changing lanes* are the important variables. What is different from the frequency analysis is that variables such as *daylight, cars/taxis* have a more important role in this interpretation since they are more commonly appearing variables. The other clusters can be interpreted in the same way.

6. Test scenario generation

Here it is shown how the clusters can be further explained and processed to identify trends in the data and then synthesise test scenarios which are reflective of the accident-prone conditions that are of interest. Two different scenario production methodologies are used based on the two different scenario identification strategies of Section 5. These are the Market Basket Analysis (frequency based) and centroid neighbourhood method (geometry based).

6.1. Market basket analysis of the clustered data

After significant variables for each cluster is obtained Market Basket Analysis (MBA) is performed on the clusters including the significant variables in the analysis. The general methodology was presented in Section 4.3. Here, additional operations to standard MBA routines are discussed for scenario generation purposes.

Based on the strategy outlined in Section 4.3, hundreds of rules were initially derived for each cluster, each with a pre-set minimal support, confidence and lift levels. There is some level of arbitrariness in the threshold values and different studies use different thresholds which may vary depending on the specific data used (Pande and Abdel-Aty, 2009). What matters, as a rule of thumb, is that confidence values should be reasonably larger than 0 (usually above 10 % and ideally closer to 1) and lift values should be reasonably larger than 1 (usually above 1.1 but can be lower depending on the study (Das et al., 2019)). It should be noted that, in a cluster, if there were no rules made up of only significant variables (which met the thresholds of confidence and lift) then other variables that already exist in that cluster were also used to derive additional rules that meet the thresholds. Specifically, in this study, the lower thresholds for confidence levels were set as 0.5 (50 %) across clusters. For the lift, the threshold value was taken as 2 in all cases except for Cluster 24 which was set to 1.1. For the majority of the

Table 3
Important variables for each cluster obtained via geometric analysis.

Categorical Variables	***Cluster 18***	***Cluster 19***	***Cluster 22***	***Cluster 23***	***Cluster 24***	***Cluster 25***	***Cluster 26***
Accident Severity	Fatal	Slight	Serious		Slight	Slight	Slight
Skidding & Overturning	Overturned	Skidded/Jack-knifed, Overturned	No skidding	No skidding	No skidding	No skidding	No skidding
Time of Day	12am-3am, 3am-6am, 9 pm-12am	12am-3am, 3am-6am, 9 pm-12am		12 pm-3 pm			
1st Road Class	A road	A road	A road	Motorway / A(M), A road	Unclassified	A road	Unclassified
Carriageway Hazards	None	Object on road, Pedestrian or animal on road	None	None	None		None
2nd Road Class		A2, B2, C2	Motorway / A(M)2				
Speed Limit		20mph, 30 mph		70mph	30mph	30mph	30mph
Junction Detail	Not a junction in 20 m	Crossroads, More than 4-arms / other junction	Slip road, Private drive or entrance, T or staggered junction	Not a junction in 20 m	Not a junction in 20 m	Crossroads	T or staggered junction
Junction Location		Clearing junction				Mid junction	Entering junction
Junction Control		Traffic Light/person	Give way/stop sign/ uncontrolled			Traffic Light/ person	Give way/stop sign/uncontrolled
Light Conditions	Darkness – no lights	Daylight		Daylight	Daylight	Daylight	Daylight
Weather Conditions	High Winds, Fog/Mist No wind	No wind	No wind	No wind	No wind	No wind	No wind
Urban or Rural Area	Urban	Rural	Rural	Urban	Rural	Rural	Urban
Vehicle Type	Car/Taxis	Cars/Taxis	Agricultural Vehicles, Cars/Taxis	Goods vehicle	Cars/Taxis	Cars/Taxis	Cars/Taxis
Vehicle Manoeuvre	Going ahead-bend	Slowing or stopping, Changing lane to left, Changing lane to right, Overtaking static vehicle offside, Overtaking nearside, Overtaking moving vehicle offside, Reversing,		Going ahead	Going ahead	Going ahead	Going ahead
1st Point of Impact	Offside		Front	Front	Front	Front	Front
Did vehicle leave Carriageway	Nearside / nearside and rebounded			Did not leave the carriageway	Did not leave the carriageway	Did not leave the carriageway	Did not leave the carriageway
Sex of Driver	Male	Male	Male	Male	Male	Male	Male
Age band of Driver	Very young, Young	Mid aged	Young		Mid aged	Mid aged	Mid aged
Road Surface Conditions	Snow/Flood, Frost or ice, Wet/damp	Wet/damp	Wet/damp	Wet/damp	Dry	Dry	Dry
Weekend or Weekday?	Weekend	Weekday	Weekday	Weekend	Weekday	Weekday	Weekday
Vehicle Left Hand Drive?	LHD?No	LHD?Yes	LHD?No	LHD?No	LHD?No	LHD?No	LHD?No
Pedestrian Crossing	No Pedestrian Crossing	Pedestrian Crossing	No Pedestrian Crossing	No Pedestrian Crossing	No Pedestrian Crossing	Pedestrian Crossing	Pedestrian Crossing

association rules the lift values were much higher than 1.1 which indicated non-trivial associations. The reason why the lift threshold was taken on the lower end in Cluster 24 was that, overall, the lift values in this cluster were low. As for the support threshold, the minimal support across all clusters were set to 0.00001.

From the initial MBA output, it was seen that the basic rule lists had many repeating cases, that is, often a given rule was a subset of another rule. To simplify the rule lists, the first additional operation carried out was to filter out all the rules which are already subsets of larger rules were, leaving only the non-repeating rules in each cluster. This significantly reduced the number of outputted rules in each cluster.

While the filtered rules list allows one to interpret, for each cluster, the links between the significant variables, a single rule may not provide a concrete scenario by itself due to lack of detail in the links produced. For instance, a valid rule derived from MBA may not have any vehicle information but may only list some environmental conditions. This is certainly important knowledge for safety analysis but not so adequate for test-scenario generation purposes since a critical component of scenarios are the vehicle characteristics (e.g. *vehicle manoeuvre*). Therefore, a 2nd step is needed where we consistently combined the (filtered) rules which yielded scenario-like detailed combinations. The idea is that, since each rule shows a chain of conditions which are tightly connected it is plausible to view an accident as a situation where a combination of these chain of risk-bearing conditions taking place together which describes a high-risk scenario with substantial more detail. Clearly, for a real accident, the chain of conditions (i.e., rules) cannot be conflicting.

Table 4
List of consistent rules with high detail obtained from MBA.

Cl. #	antecedents	consequents	Supp.	Conf.	lift
18.1	Pedestrian or animal on road, Going ahead-bend, A, Pedestrian Crossing, Darkness - lights lit, Weekend, Overturned	Oil or mud, Back, 12am-3am, High Winds	1.82E-06	1	274787.5
18.2	Pedestrian or animal on road, Going ahead-bend, High Winds, A, Pedestrian Crossing, 12am-3am	Darkness - lights lit, Weekend, Back, Overturned	1.82E-06	1	3330.8
18.3	Right Offside, Weekend, C, 12am-3am, High Winds, Overturned	Going ahead-bend, Serious, Wet/damp	1.82E-06	1	184.5
18.4	Object on road, LHD?Yes, Skidded/Jack-knifed, Back, 12am-3am, Pedestrian Crossing	Darkness - lights lit, A, Wet/damp	1.82E-06	1	27.3
19.1	Turning left, Darkness - lights lit, Weekend, Pedestrian Crossing,12am-3am, A, Crossroads	Nearside	1.27E-05	0.47	3.09
22.1	Left Nearside,Skidded/Jack-knifed, Darkness – no lights, B,T or staggered junction, Entering junction	Going ahead-bend	1.82E-05	0.53	7.413815
23.1	Pedestrian or animal on road, Darkness – no lights, Skidded/Jack-knifed, Going ahead-bend,12am-3am	Left Nearside	1.27E-05	0.7	9.355606
23.2	Left Nearside,Going ahead-bend, Motorway / A(M), Serious, Wet/damp, Very young,12am-3am	Darkness - lights lit, Overturned	1.82E-06	1	104.3
23.3	Offside, Right Offside, Frost or ice, Goods, Going ahead-bend,12am-3am	Darkness – no lights, Overturned	3.64E-06	1	122.8
23.4	Offside, Very young, Left Nearside, Skidded/Jack-knifed, Fatal, Motorbikes	Going ahead-bend	1.82E-06	1	14.1
23.5	Offside, Skidded/Jack-knifed, Darkness - lights lit,3am-6am, Very young, Changing lane to left	Wet/damp, Right Offside, Motorway / A(M),	1.82E-06	1	2486.8
23.6	Darkness - lights lit,3am-6am, Goods, Changing lane to right, Wet/damp	Motorway / A(M),Offside	3.64E-06	1	220.8
23.7	Darkness – no lights, Weekend, Changing lane to left, Goods, High Winds	Motorway / A(M),Wet/damp	1.82E-06	1	1140
23.8	Darkness - lights lit, Goods, Changing lane to right, Motorway / A(M), Wet/damp	Offside	2.37E-05	0.65	4.106303
23.9	Offside, Skidded/Jack-knifed, Darkness - lights lit,3am-6am, Very young, Changing lane to left	Wet/damp, Right Offside, Motorway / A(M),	1.82E-06	1	2486.8
23.10	Slowing or stopping, Darkness – no lights, Weekend, Motorway / A(M)	Back	4.91E-05	0.794118	5.410165
23.11	Darkness – no lights, Very young, Overtaking moving vehicle-off, Skidded/Jack-knifed	Wet/damp	2E-05	0.52381	2.170854
24.1	Old, Unclassified, Back, Dry, Slowing or stopping, 3 pm-6 pm	Daylight	4.91E-05	0.931034	1.28
25.1	Back, Roundabout, 6 pm-9 pm, A, Motorbikes, Wet/damp, Pedestrian Crossing, Weekend, Entering junction	Darkness - lights lit, Rural, Waiting	1.82E-06	1	170.5
26.1	Moving off, Private drive or entrance	Entering junction	0.001226	0.706499	2.28
26.2	Moving off, Unclassified, Crossroads	Entering junction	0.001456	0.531915	1.714549
26.3	Moving off, T or staggered junction, Unclassified	Entering junction	0.003892	0.628378	2.03
26.4	Mid junction, Turning left, Unclassified	T or staggered junction	0.002325	0.696458	2.17
26.5	Turning left, Unclassified, Entering junction	T or staggered junction	0.003961	0.674203	2.10
26.6	Mid junction, Unclassified, Turning right	T or staggered junction	0.007386	0.634021	1.98
26.7	Turning left, Clearing junction, Unclassified	T or staggered junction	0.002618	0.63225	1.97
26.8	Bicycles, Roundabout, Turning right	Mid junction	0.001006	0.632	3.00
26.9	Clearing junction, Unclassified, Turning right, Bicycles	T or staggered junction	0.000266	0.610879	1.903519
26.10	Turning left, Unclassified, Bicycles, More than 4-arms / other	Entering junction	2.91E-05	0.592593	1.910134
26.11	Turning left, Clearing junction, Bicycles, T or staggered junction	Unclassified	0.000169	0.574074	1.675928

Utilizing this idea, the rules in the filtered rule list (for each cluster) were combined such that no two conflicting variables from the same category are combined (e.g., *dry* surface and *wet* surface). This substantially reduces the size of the rules, and each row in the combined rule list provided significantly more information. In particular, nearly all combined rules show definitive manoeuvre information about the vehicle and the manoeuvre outcomes. We interpret the new combined rules as possible scenarios as they are a collection of accident-prone conditions in a way reflecting the patterns existing in their respective clusters. A point of care is that, as multiple rules are combined, the confidence and lift values change also. If the constituent rules were independent of each other the resultant confidence and lift values would be the product of constituent rules. However, this is often not the case. For this reason, in the following tables, the confidence and lift values needed to be calculated from scratch based on the entire reference data to increase the support. We note that each rule derived is given in terms of the variables in the respective clusters. Then, for scenario generation purposes the other variables can be treated as free parameters which can be randomly varied. An important point in interpreting the results is that, for scenario generation purposes no specific attention is paid to the order of the rules. Simply, for any rule A → B it is assumed that the antecedent and consequents together forms the itemset collection {A,B}, i.e., the scenario.

Table 4 lists the collection of rules obtained from all clusters alongside the support, confidence and lift values respectively. The support is an indicator of the percentage frequency of the rule as a whole, whilst the confidence is the probability of the consequent given the antecedent and the lift is the probability of the rule as a whole divided by the product of the probabilities of its constituent parts. Recalling from Tables 6 that Cluster 18 is a *night* cluster with adverse road surface effects the rules obtained from Cluster 18 naturally reflect these (rules #18.1–4). Also, as this cluster is a *non-junction* cluster, the manoeuvres mostly involve *going-ahead* type manoeuvres(instead of junction type manoeuvres such as *turn right/left).* Rule #19.1 from Cluster 19, describes a serious potential accident at the *crossroads* on an *A-road* where a vehicle tries to *turn left* where there is a *pedestrian crossing.* Rule #22.1 provides an example of a vehicle *going ahead with bend* while *entering* a *T or staggered junction* on a *B-road* which ended up *skidding* and *leaving the carriageway from nearside.*

Similarly, rules #23.1–4 from Cluster 23 describe scenarios at junctions with *going ahead-bend* manoeuvres on *motorways, A-roads* all in *late night* which led to severe outcomes such as *overturning* of the vehicles or vehicle losing control and *left carriageway* ending in *serious* injury or *death*. Some of these rules did not specify the junction type which can be randomly selected for concrete scenario generation. Rules #23.5–9 describe *changing lane to left or right* manoeuvres on *Motorways* with *wet/damp surface.* Rules #23.10 demonstrates type of accidents which may have been caused by sudden *slowing and stopping* of vehicles often resulting in crashes from *back.* Another type of accident scenario (∼23.11) which is linked to Cluster 23 involves a *very young* driver *overtaking a moving vehicle at night with no light and getting jack-knifed*. Rule #24.1 illus-

Table 5
List of exemplary test scenarios from Cluster 2 via centroid neighbourhood method.

Ex. Scenario 1	Ex. Scenario 2	Ex. Scenario 3	Ex. Scenario 4	Ex. Scenario 5	Ex. Scenario 6	Ex. Scenario 7
Slight	Slight	Slight	Slight	Slight	Slight	Slight
Skidded/Jack-knifed	Skidded/Jack-knifed	Overturned	Skidded/Jack-knifed	Skidded/Jack-knifed	No Skid	Overturned
6 pm-9 pm	9 pm-12am	3 pm-6 pm	9am-12 pm	3 pm-6 pm	6 pm-9 pm	6am-9am
A	A	A	A	Unclassified	A	A
None	None	None	None	None	None	None
A2	A2	Unclassified2	C2	Unclassified2	A2	Unclassified2
30mph	30mph	30mph	40mph	30mph	20mph	30mph
T or staggered junction	Crossroads	T or staggered junction	Crossroads	Crossroads	More than 4-arms / other junction	Slip road
Entering junction	Entering junction	Entering junction	Entering junction	Clearing junction	Clearing junction	Entering junction
Traffic light/ person	Traffic light/ person	Traffic light/ person	Traffic light/ person	Traffic light/ person	Traffic light/ person	Traffic light/ person
Daylight	Darkness - lights lit	Daylight	Daylight	Darkness - lights lit	Daylight	Daylight
No wind	No wind	No wind	No wind	No wind	No wind	No wind
Rural	Rural	Rural	Urban	Rural	Rural	Rural
Cars/Taxis	Bicycles	Cars/Taxis	Cars/Taxis	Cars/Taxis	Cars/Taxis	Cars/Taxis
Going ahead-bend	Slowing or stopping	Slowing or stopping	Slowing or stopping	Slowing or stopping	Going ahead-bend	Going ahead-bend
Front	Front	Front	Front	Front	Offside	Offside
Nearside / nearside and rebounded	Nearside / nearside and rebounded	Nearside / nearside and rebounded	Did not leave carriageway	Nearside / nearside and rebounded	Nearside / nearside and rebounded	Offside / offside rebounded/crossed/ etc
Male	Male	Male	Male	Male	Male	Male
Mid aged	Mid aged	Mid aged	Mid aged	Mid aged	Mid aged	Old
Wet/damp	Dry	Dry	Wet/damp	Wet/damp	Dry	Dry
Weekday	Weekday	Weekend	Weekday	Weekday	Weekend	Weekday
LHD?No	LHD?No	LHD?Yes	LHD?No	LHD?No	LHD?Yes	LHD?No
Pedestrian Crossing	Pedestrian Crossing	Pedestrian Crossing	Pedestrian Crossing	Pedestrian Crossing	Pedestrian Crossing	Pedestrian Crossing

trates a common type of accident scenario in which a sudden *slowing or stopping* manoeuvre causes a crash. A more passive type of junction accident scenario is depicted in rule #25.1 where a vehicle is hit from the *back* while *waiting* at the *entrance* of a *roundabout.* As for Cluster 26, rules 26.1–3 illustrate accident-prone *moving off* situations at *private drive or entrances, crossroads* and *T or staggered* junctions. Rules #26.4–7 describe accidents that involve *turning right and turning left* at *T or staggered junctions* on *unclassified roads.* Finally, rules #26.8–11 describe accidents with *turning right* or *turning left* manoeuvres by *bicycles* at the *crossroads* or *T or staggered junctions.*

6.2. Scenario production via neighbourhood analysis of the clustered data

Section 5.3 identifies, for each cluster, the important variables of the respective cluster from a geometric point of view. For the scenario candidates we concentrate on the very close neighbourhood of the centroid and make a selection from the data points with top 10 % similarity to the centroid. As there may be occasional cases which happen to be in the neighbourhood by chance, we further require that only those points of which more than 50 % of its variables overlap with the important variables found in Sec 5.3. An example of the accidents that meet these criteria are shown in Table 5. In total 7 scenarios were obtained from Cluster 22. The other scenarios from the rest of the clusters can be obtained in a similar way.

6.3. Exploitation of the developed scenarios in CAV research

As discussed in Section 1 and 2 this study took a data driven approach for scenario generation. Section 5 and Section 6 provide explicit *logical scenarios* as the backbone for the simulation based Verification & Validation workflow. Once the logical scenario creation is complete, such as the scenarios in Tables 4 and 5, they are passed to a toolchain which generates concrete scenarios from these logical ones via constraint randomisation as briefly discussed in the Section 2 (Xizhe et al., 2021). The concrete scenarios contain the ODD attributes with specified ranges which follow the established standards (ISO22737). After this stage the scenarios are stored in a scenario database. As part of the OmniCAV project the scenarios developed through this procedure have been hosted in the Safety^{TM} Pool Database. Once they are in the database these scenarios can be retrieved by scenario selector via API calls from the database and be used as testcases by the stakeholders depending on their needs.

7. Conclusion

In this paper, we have analysed past traffic accident records from the UK STAT19 database with two high level aims (i) detecting trends in terms of the variables (or the attributes) that are present at the time and place of accidents and the relationships between the variables (ii) developing, based on these obtained patterns, a systematic way that allows one to extract high-risk pre-crash test scenarios from the data.

To achieve the first part of these goals, we took the approach of not using any pre-set models and tried to find the patterns that naturally exist in the data. For (i), keeping in mind the categorical nature of the dataset, the ROCK algorithm was employed to split the data into 26 natural clusters some of which had only few data points. Seven large clusters whose sizes were above 100 data points were analysed in detail to ascertain relationships that are specific to these clusters. To this end, the Chi-square test was employed to understand the meaning of the clusters based on the frequencies of attributes in each cluster. The application of this test yielded a number of significant variables which help characterize the clusters. In addition, the concept of a geometric neighbourhood of the centroid was utilized to give an interpretation of what each cluster meant by also providing a set of variables, for each cluster, which portrayed the essential attributes.

For our second main goal (ii), using a data-based approach, we developed two complementary ways of producing risky scenarios. The first method is the market basket analysis which enables one to derive quantitative relationships between the attributes of clustered data. Going beyond the standard rule generation procedure we combined large numbers of rules in a consistent way to deduce refined combinations of rules that define the test scenarios desired. A second method was based on exploiting the geometric interpretation of clusters which can be described using the essential variables that were obtained for each cluster by the analysis of the neighbourhood of each cluster. Furthering this reasoning, we were able to search and select, for each cluster, a number of data points which were close enough to the centroid of the clusters and which also substantially overlapped with the essential variables of the respective cluster. By mapping the variables of each cluster to ODD attributes and behaviour competencies, the selected data points provide the test scenarios for autonomous vehicle testing.

Appendix

Here we present the table of significant variables obtained in Section 5. (See Table 6).

Table 6
List of significant variables and their relative frequencies for Clusters 18, 19 and 22. In below A, B, C, denote the UK road type (major roads, connecting roads, minor roads respectively). The speed values cited below are the posted limit on the roads that the accidents took place Road types followed by the number “2″ indicates the type of the (second) intersecting road in those accidents that took place at junctions.

Variables (categorical form)		Cl 18 sig. var. & rel. freq.	Cl19 sig. var. & rel. freq.	Cl 22 sig. var. & rel. freq.	Cl 23 sig. var. & rel. freq.	Cl 24 sig. var. & rel. freq.	Cl 25 sig. var. & rel. freq.	Cl 26 sig. var. & rel. freq.
Severity	**Slight**							
	Serious	2.05	1.69	2.15	1.81			
	Fatal	6.92		1.90	3.84			
Skidding & Overturning	**No Skid**							
	Skidded/Jack-knifed	4.97	3.72	3.15	3.37			
	Overturned	8.19	4.25	3.85	3.43			
Time of day	**12am-3am**	5.00	2.74	2.11	2.24		1.34	
	3am-6am	5.05	1.60	1.98	2.10		1.49	
	6am-9am							
	9am-12 pm							
	12 pm-3 pm							
	3 pm-6 pm							
	6 pm-9 pm			1.42			1.26	
	9 pm-12am	2.71	2.20	1.76	1.90		1.45	
1st road class	**Motorway / A(M)**				4.08	1.57	1.42	
	A	1.45	1.62					
	B			1.51				
	C	1.76		1.81				
	Unclassified							
Carriageway Hazards	**None**							
	Object on road	10.59	6.74	2.30	2.19			
	Pedestrian or animal on road				2.41			
2nd Road Class	**Not a junction in 20 m**	2.45			2.67	2.68		
	Motorway / A(M)2			5.23			1.84	
	A2		4.02	1.85			2.93	
	B2		3.30	2.90			2.73	
	C2		3.29	3.53			2.73	1.33
	Unclassified2							
Speed Limit	**20mph**	1.75	2.86	1.64			1.49	1.49
	30mph							
	40mph		2.19	1.85			1.50	1.25
	50mph	2.80		2.24	1.91			
	60mph	2.00		1.78	2.51			
	70mph			1.43	3.31			
Junction Detail								
	Slip road			5.75				
	T or staggered junction							
	Crossroads		4.21	1.29			3.24	1.44
	Roundabout / mini-roundabout			2.53			1.50	1.41
	More than 4-arms / other junction		3.24	1.52			2.41	1.84
	Private drive or entrance							
Junction Location								
	Entering junction			1.27			1.38	2.93
	Clearing junction		3.30	2.75			1.73	2.73
	Mid junction			1.46			1.83	2.73
Junction Control								
	Traffic light/ person	(2.04)	3.53		1.43	1.32	2.45	1.48
	Give way/ stop sign or uncontrolled			1.32				

(*continued on next page*)

Table 6 (*continued*)

Variables (categorical form)		Cl 18 sig. var. & rel. freq.	Cl19 sig. var. & rel. freq.	Cl 22 sig. var. & rel. freq.	Cl 23 sig. var. & rel. freq.	Cl 24 sig. var. & rel. freq.	Cl 25 sig. var. & rel. freq.	Cl 26 sig. var. & rel. freq.
Light Conditions	**Daylight**							
	Darkness - lights lit	1.76	2.21	1.84			1.70	
	Darkness – no lights	3.96		2.46	3.82			1.49
Weather Conditions	**Fine/No wind**							
	High winds	6.40	3.02	3.11	2.43		1.26	3.24
	Fog/Mist	6.14	3.47	3.25				
Urban or Rural	**Urban**							
	Rural		1.27	1.34				
Vehicle Type	**Cars/Taxis**							
	Motorbikes	(1.63)	2.27	2.23	1.39			
	Buses/Trams		2.70				1.79	1.50
	Goods				1.61			
	Agricultural vehicles			3.64	2.06			
Vehicle Manoeuvre	**Reversing**					1.59		
	Parked				2.00	1.81		
	Waiting						1.50	2.40
	Moving off						1.38	1.38
	Slowing / stopping				1.28	1.31		
	Turning left			1.57			1.52	1.73
	Turning right or U						1.39	1.82
	Changing lane left		5.06		1.62		1.42	2.44
	Changing lane right	(4.86)		2.01	1.84			
	Overtaking moving vehicle - offside		3.22		1.73			
	Overtaking static vehicle - offside					1.28		
	Overtaking nearside	3.84	5.70					
	Going ahead-bend			1.94	3.28			
	Going ahead							
Point of Impact	**Back**	(1.46)					1.29	1.70
	Nearside		1.44	3.65				
	Front							
	Offside	1.62	1.42					
Did vehicle leave the carriageway	**Did not leave carriageway**							
	Nearside / nearside and rebounded	6.85	2.32	3.65	3.46			
	Offside / offside rebounded/crossed/etc	6.20	3.91	3.00	3.05			
Sex of the driver	**Male**							
	Female							
Age band of the driver	**Very young**	2.38	1.57	1.37	1.44			
	Young							
	Mid aged							
	Old	1.58						
Road surface	**Dry**							
	Wet/damp	2.09	1.62	1.84	1.79			
	Snow/Flood	8.32			2.75			
	Frost or ice	5.78	3.92	2.72	3.28			
	Oil or mud	17.91	15.19					
Weekday or Weekend	**Weekday**							
	Weekend	2.28	1.42	2.01	1.53			
Was_Vehicle_LHD?	**Was_Vehicle_LHD?No**							
	Was_Vehicle_LHD?Yes	9.61	9.26	3.36			1.37	1.26
Pedestrian Crossing	**No Pedestrian Crossing**							
	Pedestrian Crossing	1.39	4.42	1.31			3.13	1.78

References

Abdelwahab, H.T., Abdel-Aty, M.A., 2001. Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. Transp. Res. Rec. 1746, 6–13. https://doi.org/10.3141/1746-02.

Aggarwal, C., Zhai, C., 2012. A survey of text clustering algorithms. Mining Text Data. Springer-Verlag:, New York, NY, USA, pp. 77–128.

Agrawal, R., Imielinski, T., Swami, A., 1993. Mining association rules between sets of items in large databases. Proceedings of the ACM SIGMOD 207–216.

Al-Ghamdi, A.S., 2002. Pedestrian-vehicle crashes and analytical techniques for stratified contingency tables. Accid. Anal. Prev. 34 (2), 205–214. https://doi.org/10.1016/S0001-4575(01)00015-X.

Anastasopoulos, P., Mannering, F., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accid. Anal. Prev. 41 (1), 153–159.

Anastasopoulos, P., Tarko, A., Mannering, F., 2008. Tobit analysis of vehicle accident rates on interstate highways. Accid. Anal. Prev. 40 (2), 768–775.

Barbará B., Li Y., Couto J. (2002), "COOLCAT: An entropy-based algorithm for categorical clustering," in Proc. 11th Int. Conf. Inf. Knowl. Manage., 582–589.

BSI PAS 1883 2020: Operational Design Domain (ODD): taxonomy for automated driving systems (ADS). Specification, 2020 https://www.bsigroup.com/en-GB/CAV/pas-1883/

Caliendo, C., Guida, M., Parisi, A., 2007. A crash-prediction model for multilane roads. Accid. Anal. Prev. 39 (4), 657–670. https://doi.org/10.1016/j.aap.2006.10.012.

Chen, C., Zhang, G., Liu, X.C., Ci, Y., Huang, H., Ma, J., Guan, H., 2016. Driver injury severity outcome analysis in rural interstate highway crashes: a two-level Bayesian logistic regression interpretation. Accid. Anal. Prev. 97, 69–78. https://doi.org/10.1016/j.aap.2016.07.031.

Chiou, Y.C., 2006. An artificial neural network-based expert system for the appraisal of two-car crash accidents. Accid. Anal. Prev. 38 (4), 777–785. https://doi.org/10.1016/j.aap.2006.02.006.

Corbett, M.R., Morrongiello, B.A., 2017. Examining how different measurement approaches impact safety outcomes in child pedestrian research: Implications for research and prevention. Accid. Anal. Prev. 106 (August 2016), 297–304. https://doi.org/10.1016/j.aap.2017.06.002.

Das, S., Dutta, A., Avelar, R., Dixon, K., Sun, X., Jalayer, M., 2019a. Supervised association rules mining on pedestrian crashes in urban areas: identifying patterns for appropriate countermeasures. Int. J. Urban Sci. 23, 38–40. https://doi.org/10.1080/12265934.2018.1431146.

Das, S., Dutta, A., Sun, X., 2019b. Patterns of rainy weather crashes: Applying rules mining. Journal of Transportation Safety & Security 12, 1083–1105.

Delecki, H., Itkina, M., Lange, B., Senanayake, R., Kochenderfer, M., 2022. How Do We Fail? Stress Testing Perception in Autonomous Vehicles https://arxiv.org/abs/2203.14155.

Delen, D., Sharda, R., Bessonov, M., 2006. Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. Accid. Anal. Prev. 38 (3), 434–444. https://doi.org/10.1016/j.aap.2005.06.024.

Esenturk E., Khastgir S., Wallace A., Jennings P., Analyzing real-world accidents for test scenario generation for automated vehicles, IEEE Intelligent Vehicles 2021 Symposium Proceedings, July 11-17 (2021).

Eluru, N., Bhat, C., Hensher, D., 2008. A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. Accid. Anal. Prev. 40, 1033–1054.

Esenturk, E., Wallace, A., Khastgir, S., Jennings, P., 2022. Identification of Traffic Accident Patterns via Cluster Analysis and Test Scenario Development for Autonomous Vehicles. IEEE Access 10, 6660–6675.

Everitt, B., Landau, S., Leese, M., 2001. *Cluster Analysis*. London: Arnold. https://doi.org/10.1016/S0306-4379(00)00022-3.

Factor, R., Mahalel, D., Yair, G., 2007. The social accident: A theoretical model and a research agenda for studying the influence of social and cultural characteristics on motor vehicle accidents. Accid. Anal. Prev. 39, 914–921.

Feng, S., Feng, Y., Yu, C., Zhang, Y., Liu, H.X., 2020. Testing Scenario Library Generation for Connected and Automated Vehicles, Part I: Methodology. IEEE Tran. Intel. Trans, Sys.

Formosa, N., Quddus, M., Ison, S., Abdel-Aty, M., & Yuan, J. (2020). Predicting real-time traffic conflicts using deep learning. *Accid. Anal. Prev.*, *136* (December 2019). https://doi.org/10.1016/j.aap.2019.105429.

Guha S, Rastogi R, Shim K. (1999) ROCK: A robust clustering algorithm for categorical attributes. InProc. 1999 Int. Conf. Data Engineering, Sydney, Australia, Mar., 512-521.

Ghamdi, A.S., 2007. Experimental evaluation of fog warning system. Accid. Anal. Prev. 39.

He, Z., Xu, X., Deng, S., 2002. 'Squeezer: An efficient algorithm for clustering categorical data'. J. Comput. Sci. Technol. 17, 611–624.

Helbing, D., 2001. Traffic and Related many particle systems. Review of Modern Physics 73, 1067–1138.

Hemmati H., Arcuri A., Briand L. (2010). Reducing the cost of modelbased testing through test case diversity," in *Proc. IFIP Int. Conf. Test. Softw. Syst.* Berlin, Germany: Springer, 2010, 63–78.

Hossain, M., Muromachi, Y., 2012. A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways. Accid. Anal. Prev. 45, 373–381. https://doi.org/10.1016/j.aap.2011.08.004.

Huang, H., Abdel-Aty, M., 2010. Multilevel data and Bayesian analysis in traffic safety. Accid. Anal. Prev. 42 (6), 1556–1565. https://doi.org/10.1016/j.aap.2010.03.013.

Iranitalab, A., Khattak, A., 2017. Comparison of four statistical and machine learning methods for crash severity prediction. Acc. Anal. Prev. 108, 27–36.

Kalra, N., Paddock, S.M., 2016. Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?. Transportation Research Part A: Policy and Practice 94 (December), 182–193. https://doi.org/10.1016/j.tra.2016.09.010.

Khastgir, S., Birrell, S., Dhadyalla, G., Jennings, P., 2018. Calibrating trust through knowledge: Introducing the concept of informed safety for automation in vehicles. Transportation Research Part C: Emerging Technologies 96, 290–303. https://doi.org/10.1016/j.trc.2018.07.001.

Khastgir, S., Brewerton, S., Thomas, J., Jennings, P., 2021. Systems Approach to Creating Test Scenarios for Automated Driving Systems. Reliab. Eng. Syst. Saf. 107610.

Khastgir S., Dhadyalla G., Birrell S., Redmond S., Addinall R., Jennings P. (2017) "Test scenario generation for driving simulators using constrained randomization technique," SAE Tech. Paper 2017-01-1672, https://www.sciencedirect.com/science/article/pii/S0951832021001551.

Khastgir, S., Birrell, S., Dhadyalla, G., & Jennings, P. (2018a). The Science of Testing: An Automotive Perspective. *SAE Technical Paper: 2018-01-1070*. https://doi.org/10.4271/2018-01-1070.

Kim, J.K., Ulfarsson, G., Shankar, V., Mannering, F., 2010. A note on modeling pedestrian injury severity in motor vehicle crashes with the mixed logit model.. Accid. Anal. Prev. 42.

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

Lee, C., Abdel-Aty, M., 2005. Comprehensive analysis of vehicle-pedestrian crashes at intersections in Florida. Accid. Anal. Prev. 37 (4), 775–786. https://doi.org/10.1016/j.aap.2005.03.019.

Lee, J., Abdel-Aty, M., Cai, Q., 2017. Intersection crash prediction modeling with macro-level data from various geographic units. Accid. Anal. Prev. 102, 213–226. https://doi.org/10.1016/j.aap.2017.03.009.

Lefler, D.E., Gabler, H.C., 2004. The fatality and injury risk of light truck impacts with pedestrians in the United States. Accid. Anal. Prev. 36 (2), 295–304. https://doi.org/10.1016/S0001-4575(03)00007-1.

Lenard, J., Badea-Romero, A., Danton, R., 2014. Typical pedestrian accident scenarios for the development of autonomous emergency braking test protocols. Accid. Anal. Prev. 73, 73–80.

Leveson, N., 2012. Engineering a Safer WorldSystems Thinking Applied to Safety https://library.oapen.org/viewer/web/viewer.html?file=/bitstream/handle/20.500.12657/26043/1004042.pdf?sequence=1&isAllowed=y.

Li, Z., Liu, P., Wang, W., Xu, C., 2012. Using support vector machine models for crash injury severity analysis. Accid. Anal. Prev. 45, 478–486.

Li, X., Lord, D., Zhang, Y., Xie, Y., 2008. Predicting motor vehicle crashes using Support Vector Machine models. Accid. Anal. Prev. 40, 1611–1618.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. Transportation Research Part A: Policy and Practice 44 (5), 291–305. https://doi.org/10.1016/j.tra.2010.02.001.

Lloyd, D., Wilson, D., Tuddenham, F., Goodman, G., Bhagat, A., 2012. Reported Road Casualties Great Britain: 2012 https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/269601/rrcgb-2012-complete.pdf.

Lord, D., Manar, A., Vizioli, A., 2005. Modeling crash-flow-density and crash-flow-V/C ratio relationships for rural and urban freeway segments. Accid. Anal. Prev. 37 (1), 185–199. https://doi.org/10.1016/j.aap.2004.07.003.

Lord, D., Washington, S., Ivan, J., 2005. Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accid. Anal. Prev. 37 (1), 35–46.

Ma, J., Kockelman, K.M., Damien, P., 2008. A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. Accid. Anal. Prev. 40 (3), 964–975. https://doi.org/10.1016/j.aap.2007.11.002.

Mannering, F., Bhat, R.C., 2014. Analytical methods in accident research: Methodological frontier and future directions. Analytical methods in accident research 1, 1–22.

Meuleners, L.B., Fraser, M., Johnson, M., Stevenson, M., Rose, G., Oxley, J., 2020. Characteristics of the road infrastructure and injurious cyclist crashes resulting in a hospitalisation. Accid. Anal. Prev. 136 (August 2019). https://doi.org/10.1016/j.aap.2019.105407.

Milton, J., Shankar, V., Mannering, F., 2008. Highway accident severities and the mixed logit model: An exploratory empirical analysis. Accid. Anal. Prev. 40, 260–266.

Mohammadnzar, A., Arvin, R., Khattak, A., 2021. Classifying travellers' driving style using basic safety messages generated by connected vehicles: Application of unsupervised machine learning. Transportation Research C 122, 102917.

Nitsche, P., Thomas, P., Stuetz, R., Welsh, R., 2017. Precrash scenarios at road junctions: a clustering methods for car crash data. Accident Analyis and Prevention 107, 137–151.

Nowakowska, M., 2017. Selected aspects of prior and likelihood information for a Bayesian classifier in a road safety analysis. Accid. Anal. Prev. 101, 97–106. https://doi.org/10.1016/j.aap.2017.01.009.

Paleti, R., Eluru, N., Bhat, C., 2010. Examining the influence of aggressive driving behavior on driver injury severity in traffic crashes.. Accid. Anal. Prev. 42.

Pande, A., Abdel-Aty, M., 2009. A novel approach for analyzing severe crash patterns on multilane highways. Accid. Anal. Prev. 56, 95–102. https://doi.org/10.1016/j.aap. 2009.06.003.

Park, H.S., Jun, C., 2009. A simple and fast algorithm for K-medoids clustering. Expert Syst. Appl. 36.

Poulos, R.G., Hatfield, J., Rissel, C., Flack, L.K., Shaw, L., Grzebieta, R., McIntosh, A.S., 2017. Near miss experiences of transport and recreational cyclists in New South Wales, Australia. Findings from a prospective cohort study. Accid. Anal. Prev. 101, 143–153. https://doi.org/10.1016/j.aap.2017.01.020.

Rahimi, A., Azimi, G., Asgai, H., Jin, X., 2019. Clustering approach toward large truck crash analysis. Transportation Reseearch Record 2673, 73–85. https://doi.org/10.1177/0361198119839347.

Road Safety Data - STATS19. (2020). Retrieved April 2, 2020, from UK Department for Transport website: https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data.

Russo, B.J., Savolainen, P.T., Schneider, V.H., Anastasopoulos, P.C., 2014. Comparison of factors affecting injury severity in angle collisions by fault status using a random parameters bivariate ordered probit model. Anal. Methods Accid. 2, 21–29.

SAE J3016, 2018. Taxonomy and Definitions related to driving automation systems for on-road motor vehicles https://www.sae.org/standards/content/j3016_201806/.

Sagberg, F., Selpi, S., Piccinini, G.F.B., Engstrom, J.A., 2015. Review of Research on Driving Styles and Road Safety. Hum. Factors 57, 1248–1275.

Savolainen, P.T., Mannering, F.L., Lord, D., Quddus, M.A., 2011. The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. Accid. Anal. Prev. 43 (5), 1666–1676. https://doi.org/10.1016/j.aap.2011.03.025.

Shamsunnahar, Y., Eluru, N., 2013. Evaluating alternate discrete outcome frameworks for modeling crash injury severity. Accident Analysis & Prevention 59, 506.

Shamsunnahar, Y., Eluru, N., Bhat, C., Tay, R., 2014. A latent segmentation based generalized ordered logit model to examine factors influencing driver injury severity. Anal. Methods Accid. Res. 1, 23–38.

Sui, B, Lubbe N., Bargman J.,, 2019. A clusterng approach to developing car to two-wheeler test scenarios for the assessment of aotimated emegerncy braking in China using in-deptsh Chinese crash data 131, 105242.

Sun, M., Sun, X., Shan, D., 2019. Pedestrian crash analysis with latent class clustering method, Accid. Anal. Prev., 124, 50–57 US Department of Transportation. A framework for automated driving systems testable cases and scenarios https://www.nhtsa.gov/sites/nhtsa.gov/files/documents/13882-automateddrivingsystems_092618_v1a_tag.pdf.

S. Ulbrich, T. Menzel, A. Reschka, F. Schuldt, M. Maurer (2015), ''Defining and Substantiating the Terms Scene , Situation , and Scenario for Automated Driving," 2015.

tan, Z., che, Y., hu, W., li, P., xu, J., 2021. Research of fatal car-to-**pedestrian** precrash scenarios for the testing of the active safety system in China. Accident Analysis & Prevention 150.

Thorn, E., Kimmel, S., Chaka, M., 2018. A framework for automated driving system testable cases and scenarios. Washington, DC: National Highway Traffic Safety Administration Report No. DOT HS 812 623.

Xie, K., Wang, X., Huang, H., Chen, X., 2013. Corridor-level signalized intersection safety analysis in Shanghai, China using Bayesian hierarchical models. Accid. Anal. Prev. 50, 25–33. https://doi.org/10.1016/j.aap.2012.10.003.

Xizhe, Z., Khastgir, S., Asgari, H., Jennings, P.A., 2021. Test framework for automatic test case generation and execution aimed at developing trustworthy AVs from both verifiability and certifiability aspects. In: The 24th IEEE International Conference on Intelligent Transportation Systems (ITSC 2021), Indianapolis, IN, United States, 19-22 Sep 2021.

Yang Y., Guan X., You J. (2002), CLOPE: a fast and effective clustering algorithm for transactional data, Proceedings of the 8th International Conference on Knowledge discovery and data mining for transactional data, 682-687.Yasmin, S., Eluru, N., 2013. Evaluating alternate discrete outcome frameworks for modeling crash injury severity. Accid. Anal. Prev. 59, 506–521.

Yu, R., Abdel-Aty, M., 2014. Using hierarchical Bayesian binary probit models to analyze crash injury severity on high speed facilities with real-time traffic data. Accid. Anal. Prev. 62, 161–167. https://doi.org/10.1016/j.aap.2013.08.009.

Zeng, Q., Huang, H., Pei, X., Wong, S.C., 2016. Modeling nonlinear relationship between crash frequency by severity and contributing factors by neural networks. Anal. Methods Accid. Res. 10, 12–25. https://doi.org/10.1016/j.amar.2016.03.002.