IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Identification of Traffic Accident Patterns via Cluster Analysis and Test Scenario Development for Autonomous Vehicles

Emre Esenturk*[1], Albert Wallace[1], Siddartha Khastgir[1], Paul Jennings[1]
[1]WMG, University of Warwick, Coventry, CV4 7AL, UK

Corresponding author: Emre Esenturk (e-mail: e.esenturk.1@warwick.ac.uk).

The work presented in this paper has been carried under the Innovate UK and Centre for Connected and Autonomous Vehicles (CCAV) funded OmniCAV project (Grant No. 104529). This work is also supported by UKRI Future Leaders Fellowship (Grant MR/S035176/1). The authors would like to thank the WMG center of HVM Catapult and WMG, University of Warwick, UK, for providing the necessary infrastructure for conducting this study. WMG hosts one of the seven centers that together comprise the High Value Manufacturing Catapult in the UK.

ABSTRACT Increased safety is one of the main motivations for traffic research and planning. The arduous task has two components: (i) improving the existing traffic policies based on a good understanding of risk factors related to trends in traffic accidents, and (ii) underpinning the emerging technologies that will advance the safety of vehicles. For the latter route, the introduction of connected and automated vehicles (CAVs) is a promising option as CAVs can potentially reduce the number of accidents. However, to reap their benefits, they need to be introduced in a safe manner and tested for their ability to safely deal with risky scenarios. Unfortunately, the identification of such test scenarios remains a key challenge for the industry. This study contributes to increased safety by (i) analyzing UK's STATS19 accident data to identify patterns in past traffic accidents, and (ii) utilizing this information to systematically generate scenarios for CAV testing.

For task (i), the patterns in the accidents were identified in terms of static and time-dependent internal and external factors. For this purpose, the study employed a clustering algorithm, COOLCAT, which is particularly suitable for dealing with high-dimensional categorical data. Six different clusters emerged naturally as a result of the algorithm. To interpret the clusters, we applied a frequency analysis to each cluster. The frequency tests showed that in each cluster, certain distinct real-world situations were represented more significantly compared to the non-clustered reference case, which are the markers of each cluster. The second task (ii) complemented the first task by synthesizing the relationships between attributes. This was done by association rule mining using the market basket analysis approach. The method enabled us to develop, drawing from the characteristics of the clusters, non-trivial test scenarios that can be used in the testing of CAVs, especially in virtual testing.

INDEX TERMS Accident analysis, scenario development, cluster analysis, market basket analysis

I. INTRODUCTION

Over the past five years, more than a half million traffic accidents have been reported in the UK, distributed more or less evenly in each year [1] ("Road Safety Data - STATS19," 2020). Despite the traffic safety measures taken by the UK government, there has been a steady figure of over seventeen hundred on-road fatalities annually. In addition to the tragedy of losing loved ones, such accidents incur heavy costs to the economy overall, such as support services and healthcare systems. Clearly, as the first order of business, it is of prime importance to identify and analyze the factors leading to severe accidents in order to reduce the chances of occurrence. A promising and ambitious solution to reduction of traffic accidents is the introduction of Connected and Autonomous Vehicles (CAVs) which can significantly reduce the rate and severity of traffic accidents [2]-[4]. However, to reap the safety benefits of CAVs, it is essential to ensure that their introduction is done in a safe manner, and second, they are trusted, accepted, and used by the public. Establishing the capabilities and limitations of the CAVs and communicating them to the public is key to creating a state of "informed safety" which, in turn, leads to the development

of trust in CAVs [5]. However, owing to the increased complexity of CAVs [6], ensuring and evaluating their true capabilities and limitations remains a challenge [7]. It is suggested that to prove that CAVs are safer than human drivers, they need to be driven for over 11 billion miles [8]. This might seem to be an unrealistic proposition, but an alternate school of thought of Hazard Based Testing, that focuses on the quality of miles, suggests testing for *"how a system fails"* as compared to *"how a system works"* [9]. Understanding how a system may fail can be either done in a proactive manner (e.g., via safety assessments involving hazard identification) [10], or in a reactive manner (e.g., by analyzing road accident databases), [11]. While the former would be intrinsic to the system, the latter would yield extrinsic factors that may lead to hazards. Identifying extrinsic factors, even for normal, human-driven systems, requires a deep understanding of the relationships between them. Once such an understanding is achieved for human-driven systems, it can serve as a basis for developing tests and test scenarios to help train CAVs.

The goal of this study is to devise a systematic way that underpins the aforementioned reactive path by creating realistic real-world scenarios that are archetypal of high-risk traffic situations. This is a two-stage problem requiring one to develop an approach that is capable of (i) detecting patterns in a wealth of accident data and (ii) synthesizing scenarios based on the significant relationships within these patterns. In this study, improving on [12], we used a cluster analysis approach for stage (i) and association rule mining for stage (ii). We demonstrate our approach using the UK traffic accident database.

The approach presented in this study offers several prospects. First, cluster analysis can provide an efficient way to cast scattered accidents into natural groups which exhibit collective characteristics. These groups can sometimes be of very small sizes (or have very small sub-groups), which depict rare but distinct traffic situations that might be omitted using other traditional methods such as regression. Second, many existing traffic data analysis methods, a priori, categorize variables as dependent and independent. Our methodology does not require such assumptions and allows the extraction of naturally occurring relationships within the data (i.e., stage (ii)). Third, thanks to the particular clustering algorithm used in this study, streams of new incoming scenarios can be classified appropriately and efficiently, helping with maintenance of large databases.

Applying the suggested methodology, it was found that the accident dataset can be differentiated into six distinct clusters, each of which shows different characteristics. These are (i) fatal, late night, off-junction accidents on motorways with high-speed limit, (ii) two-wheeler (bicycles and motorbikes) accidents on minor roads at a junction while turning left or right, (iii) fatal, two-wheeler accidents on slip-roads connecting to major roads in foggy weather; (iv) off-junction accidents involving buses on unclassified roads; (v) accidents on private drives involving reversing and parked vehicles; and (vi) night accidents at multi-armed junctions of major roads with low speed limits involving buses and bicycles. Following the identification of these clusters, market basket analysis was applied to each cluster to ascertain the quantitative relationships between the in-cluster attributes, which can be regarded as proto scenarios. These rules are then combined to obtain scenarios that represent the corresponding clusters.

The remainder of this paper is organized as follows. Section 2 provides a brief review of the literature on accident data analysis concentrating on data mining methods. Section 3 provides an overview of the data format and how the data was processed into the form that was used in the study. Section 4 introduces our analysis method and the algorithms used. In Section 5, we present our findings. In Section 6, these findings are interpreted in the context of scenario generation and are utilized to systematically develop natural pre-crash exemplary scenarios. Finally, Section 7 concludes the paper.

II. BACKGROUND

A vast literature exists on traffic accidents and their relationships to surrounding conditions [13]. A commonly used approach for analysis is to formulate the relationships in a correlational setting using classical or contemporary techniques, including various types of regression models [14]-[20], [11], [57], [58]; Bayesian analysis [22]-[25]; neural-network models [21],[26]-[29].

An alternative approach is not to assume a pre-set relationship and let the data reveal itself. This provides more flexibility and fidelity for data mining methods. Following this spirit, in recent years, data mining strategies have attracted increased attention in safety research and automated driving systems (ADSs) such as association rule mining [30-32]; and decision trees [33-37].

One type of data mining strategy, which has been explored to a lesser extent (in the context of traffic accident data) is cluster analysis [38]. The crux of this technique is to group traffic accidents according to microscopically or macroscopically defined criteria, which allows for comparative examination of these groups [39]. Among the past studies, in [40] k-means clustering method was used to analyze accident hotspots whereas in [41] and [42] the same method was used to support the severity prediction of accidents. More recently, related k-means clustering methods were used by [12] for crash analysis at road junctions, by [43] for pedestrian pre-crash scenarios and by [44], [45] for the assessment of automated emergency braking systems in accidents.

To leverage the use of clustering methods, one needs to be mindful of the algorithms' data processing procedures. To this end. the first order of consideration is the suitability of the method for the data type under study. Most clustering

methods that have been employed in traffic research employed the k-means algorithm [46] and its variants, k-medoids [47] or k-modes [48]. While k-means is a popular solid clustering method it is not very suitable for categorical data as the mean of a categorical variable is not meaningful. On the other hand, k-medoids and k-modes can handle categorical data. However, they are known to suffer from poor performance when working with high dimensional data [12] and may not be the most ideal method if one intends to analyze datasets with a large number of attributes, which is one of the central aims of this study. This problem can be partially circumvented by reducing the dimensions (i.e., discounting certain variables with educated decisions/guesses), as was done in some recent works [12]. However, one should be wary of resorting to approaches such as handcrafted feature selection for cluster analysis as they may be prone to error or bias [53]. Considering that most traffic accident data, especially the UK STATS19 database, consist of attributes that are predominantly categorical, it is advisable to use an algorithm designed for categorical data clustering such as COOLCAT [49], ROCK [50], DBSCAN [51], and SQUEEZER [52], LIMBO [63].

A second point of consideration for deciding on an algorithm is the criterion for distinguishing clusters. Most clustering methods that have been employed in traffic research rely on distance-based algorithms using microscopic (local) criterion/basis for assignment to clusters such as DBSCAN and its many more recent variants [62]. However, employing an algorithm that works with criteria based on global properties (such as entropy) of the data groups can provide new insights to identify the trends in the data and is preferred in this study. Another issue to take into account is the speed. For instance, even though ROCK is a categorical clustering algorithm that utilises some level of nonlocal properties in its clustering procedure (forming clusters based on links instead of local distances) . However, due to it is agglomerative nature it is slow and not scalable. SQUEEZER on the other hand is fast, however, the clustering is very sensitive to ordering of the data, as the clusters are built incrementally from single element. Hence, considering these aspects, in this paper we use an entropy-based algorithm, COOLCAT, which is, by design suitable for categorical data clustering [49]. Moreover, COOLCAT can work with high-dimensional data without compromising on the quality. It distinguishes clusters based on the measure of entropy which is a global feature of the data. Also, COOLCAT is efficient and can handle streams of incoming data with ease. Furthermore, clustering with COOLCAT is relatively less data dependent since initial cluster seeds are independent of the order in the data. One downside of the COOLCAT is the initialization stage which has quadratic complexity which may increase the overall time cost. This is a price paid for requiring a more stable and consistent clustering which is a comparable cost to other similar clustering algorithms such as LIMBO.

While providing useful insight for understanding accident patterns, a cluster algorithm alone may not immediately convey a meaning to the clusters formed. In other words, one needs to understand what the produced clusters represent. For small clusters with a small number of attributes, this can be achieved by eyeballing the clusters. However, for clusters with a large number of data points and attributes, one needs a systematic way to interpret what each cluster signifies. Furthermore, even after a cluster obtains meaning in terms of its indicator attributes, this does not provide much clue on the relationship between these variables, which is crucial in understanding the development of individual scenarios. For this purpose, we propose a two-step procedure that identifies the key attributes that distinctively describe each cluster and then extracts the previously unknown relationships between the attributes within those clusters. The first step is to run comparative frequency tests between the clusters and the reference distribution of the attributes. The second step involves employing the association rule mining method (i.e., market basket analysis) on the distinguished attributes.

III. METHODOLOGY

A. FORM OF THE DATA AND PRE-PROCESSING

This study is based on an analysis of publicly available data collected from police reports in the UK [1]. Accidents from the 2016-2018 period were taken as the base data, which amounts to 389238 accidents in number. In its raw form, the data is stored in different files describing the accidents depending on the perspective of either common attributes (e.g., weather condition, light condition) or specific attributes (e.g., sex of the driver, vehicle type). Not all attributes recorded in the datasets were regarded as relevant for the analysis. For instance, the effects of cultural origin were discounted. Likewise, variables that were thought to be unimportant were disregarded, such as local authority district and police officer attendance. As the main goal of this study is scenario development, only those attributes (or variables) that have a direct influence on accidents were kept. After this, the data were reorganized from the perspective of the driver, which meant duplicating the common variables. Furthermore, only those accidents involving one vehicle or two vehicles with physical impact were considered. The reason for this is to keep the scope of the paper focused on test scenario generation for AVs. Since overwhelming majority of the traffic accidents involve one or two vehicles it was decided to restrict the analysis to such accident types.

Another important point is that most of the attributes recorded in the STATS19 database were categorical with many superfluous values. Therefore, certain variables are

restructured, for instance, by merging cases. An example of this is provided in the appendix. The full dictionary can be found in the STATS19 database [1]. Furthermore, for each accident with a missing value, a random value from the possible set of values from the respective category was assigned.

B. ODD AND BEHAVIOUR COMPETENCIES

As mentioned earlier, a major challenge in the CAV industry is the development of test scenarios. Considering the high demand in this domain, an established format for scenario description is instrumental for easy and standardized exchange of scenarios. This gave birth to the operation design (ODD) concept detailed in (BSI, 2020) and 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 consists of three main classes of descriptors: scenery (such as drivable are, junctions, physical structure, etc.), environmental conditions (such as weather and light conditions), and dynamic elements (such as traffic conditions and speed of the vehicle). As shown below, many of the attributes from the STATS19 dataset can be easily mapped onto the attributes in ODD. A complementary concept that is used in this paper (and included in STATS19 variables) is the "behavior competencies" (e.g., vehicle maneuver), which basically describes driving behavior [55]. Together, ODD and behavior competencies constitute the backbone for scenario development.

C. CRASH DATA VARIABLES

This study takes the perspective that the traffic accidents can be described solely in terms of the local effects, that is, factors and output that are immediately present at the time and location of the accident. Overall, 22 variables from the STATS19 database were selected to be used in the analysis: *Accident Severity, Skidding and Overturning, Time, 1st Road Class, Carriageway Hazards, 2nd Road Class, Speed Limit, Junction Detail, Junction Location, Light Conditions, Weather Conditions, Road Surface Conditions, Urban or Rural Area, Was Vehicle Left Hand Drive, Vehicle Type, Vehicle Maneuver, 1st Point of Impact, Did Vehicle Leave the Carriageway, Week or Weekend?, Pedestrian Crossing Facilities, Sex of the Driver, Age Band of the Driver*.

These variables were chosen because they either: provide information about the outcome of the accident e.g. *Accident Severity* and *Skidding and Overturning*, or provide information on the conditions around the accident e.g. *Light Conditions* and *Road Surface Conditions* or give details of the accident scenario e.g. *Vehicle Maneuver* and *Sex of Driver*. Variables that were superfluous like *local authority district* were removed.

Most variables included in the analysis are self-explanatory. We only describe the 1st *Road class* variable which shows the road type. This can come as *Motorway, A, B, C* or *unclassified road*. These are the standard UK road classes. *Motorways* and *A roads* are major roads while *B* and *C roads* are minor roads. *Unclassified* roads are roads that do not fit into the other classifications and are usually local roads intended for local traffic.

IV. DATA ANALYSIS

After cleaning and organizing the data, here we discuss the method of analysis. As noted previously, the rationale for using unsupervised learning approaches is that these techniques allow one to extract important information from the data without making any prior assumptions on the relationships between data attributes, which is a significant advantage.

We used a combination of complementary learning techniques. The first step involved clustering the data. Once this step is complete, the second step of the analysis is to understand what these clusters mean. The following subsections discuss these steps in detail.

A. CLUSTERING OF ACCIDENT DATA

This was the first step in the analysis. As mentioned earlier, clustering analysis has a long history, but its use in accident data is a relatively recent development. Therefore, although there are dozens of clustering algorithms available for general clustering purposes, the accident data under consideration are exclusively categorical and general-purpose clustering algorithms, such as k-means (which are designed for dealing with continuous variables), are less likely to yield high-quality clustering. Second, for the purposes of this study, we are more interested in differentiating clusters based on the global features of the attributes in each cluster, rather than individual similarity relationships between the data points in those clusters. The choice makes a marked difference in the type of algorithm to be used.

A.1. COOLCAT Categorical Clustering Algorithm

The COOLCAT algorithm was first proposed in [49]. It was designed specifically for categorical datasets. Unlike most other clustering algorithms (such as k-medoid and k-modes) that have been used in accident analysis research, COOLCAT is not based on a distance metric. Rather, central to COOLCAT is the concept of entropy, which is borrowed from physics and information theory and measures the disorder in a given system. Then, the goal of the algorithm is to group the data points of the system in clusters in a configuration that minimizes the average entropy. In this setting, entropy in a cluster can be quantified in terms of the normalized frequencies of the attributes within the cluster, treating each variable independently from each other. This crucial difference, that is, distinguishing clusters with respect

to globally defined differences instead of local metric distances, is one of the advantages of COOLCAT when dealing with categorical data and can help better describe the clusters in the interpretation stage. Another advantage of COOLCAT over more classical algorithms (such as k-means and k-medoids) is that COOLCAT performs incremental clustering and hence can handle streams of new incoming data without the need for clustering from scratch.

Given the number of clusters, the algorithm begins by forming cluster seeds that are chosen as the most different elements from each other in the dataset. Then, the remaining data points are assigned to the seed clusters one by one according to the average reduction in the entropy of the system. Once one iteration is completed, a portion of the data points may be redistributed among the clusters (provided that the new assignments decrease the overall entropy) to minimize path dependence effects.

B. INTERPRETATION OF CLUSTERS

The second step of the analysis focuses on ascertaining the meanings of the clusters formed by the clustering algorithm. This involves determining the significant variables that describe the clusters more distinctively and extract the a priori unknown relationships or rules between these significant variables.

B.1. Frequency Analysis for Identification of Significant Variables

Because the COOLCAT method is not metric-based, another approach for identifying the meaning of the clusters is needed. A frequency analysis was used to determine which variables appear significantly more than expected in each cluster compared with how frequently they are in the rest of the data. This is possible because the data is categorical and frequencies exist, whereas in continuous data, they would not.

Significant variables in each cluster were identified using the chi-square test. As the data is in binary form, for every data point, each variable has either a value of 1 if it was present in that accident or 0 if it was not. The chi-squared value for each variable is given by:

$$chi(\text{var}) = \frac{(O1(\text{var}) - E1\,\text{var}))^2}{2E1(\text{var})} + \frac{(O0(\text{var}) - E0\,\text{var}))^2}{2E0(\text{var})} \quad (1)$$

where var represents an arbitrary variable and

- *O1* –observed number of 1's in the cluster,
- *E1* – expected number of 1's in the cluster,
- *O0*–observed number of 0's in the cluster,
- *E0* – expected number of 0's in the cluster.

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 contains the full data (representing the distribution of the entire population). E1 is then given

$$E1(\text{var}) = cluster\_size \times \frac{frq(\text{var})}{N} \quad (2)$$

where *N* and *frq* are the total number of data points in the full data and the frequency of the variable in question, respectively. The significance of a variable is determined by whether the frequency of that variable significantly differs from the expected frequency (at a significance level *of p<0.05)* under the null hypothesis that it does not. After the significant variables are found, the index *relative frequency = observed/expected* is calculated to identify which variables are more overrepresented in the cluster. In the sequel, we require, for the relative frequency of a variable to be larger than a set threshold to be deemed as the signifier or indicator of a cluster (see Section 5).

B.2. Market Basket Analysis

Market Basket Analysis (MBA) (Agrawal, 1993) is a method that is mainly used on business transactional data to identify which 'products' are found together in 'customers purchases'. In general, the idea is to find association rules between variables that appear together unusually frequently. The first step in MBA is to find frequent itemsets using the Apriori algorithm. A *k* itemset is a subset of all possible variables of length *k*. For example, in a shopping context, an itemset could be {Bread, Milk, Eggs, Cheese}, while in a traffic accident context, the itemset would be {Motorbike, Entering Junction, Turning Left}. An itemset is said to be frequent if its support exceeds a given threshold. The support of an itemset *X* is given by the frequency of *X*, that is, the number of data points to which all members of the itemset belong to, divided by *N* the total number of data points, that is,

$$Support = \frac{frq(A \cup C)}{N} \quad (3)$$

The Support is essentially a measure of how rare an itemset is.

In the second step, once frequent itemsets are found, is to identify association rules within them. This is done by partitioning the itemset into two subsets, the antecedent and the consequent, which then gives the association rule antecedent → consequent. For example, an itemset *X={x1, x2, x3}* can be split into antecedent *A={x1, x2}* and consequent *C={x3}*, which would give the rule *A→C*.

Two metrics were used to identify the strength of the association: confidence and lift. Confidence is given by the frequency of the union of the antecedent and the consequent (the joint itemset), which corresponds to the intersection of the data points, divided by the frequency of the antecedent. i.e.,

$$Confidence = \frac{frq(A \cup C)}{frq(A)} \quad (4)$$

Intuitively, for rule $A \rightarrow C$, this is the probability that C occurs, given that A also occurs. The lift is given by the support of the entire itemset divided by the support of the antecedent multiplied by the support of the consequent.

$$Lift = \frac{Support(A \cup C)}{Support(A) \times Support(C)} \quad (5)$$

For association $A \rightarrow C$, this is a comparison between how often A and C actually appear together, with how often A and C would be expected to appear together if they were independent, based on their support within the dataset. If the lift is less than 1 it indicates that *A* is not strongly associated with *B* any more than it coincidentally appears together. On the other hand, if the lift is higher than one, then this indicates that, even if the rule has low confidence, the items appearing together are not coincidental. A summary of the concepts is given in figure 1.

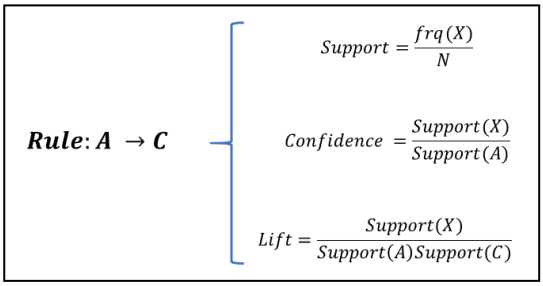

FIGURE 1. Relevant relations for Market Basket Analysis rules

V. RESULTS

In this section, we present the main findings of this study in two stages. First, the previously explained COOLCAT clustering method was applied to a sample of 20000 data points that were randomly selected from the collection of accident records. As COOLCAT is robust against high dimensionality, no attempt was made to reduce the number of attributes further. In the second stage, a combination of frequency analyses followed by MBA was carried out to extract the significant associations for each cluster which formed the scenarios obtained from those clusters. We report that the COOLCAT clustering algorithm was coded and executed in MATLAB 2019a while the MBA method was implemented in python 3.7 using the mlxtend package [59].

A. RESULTS FOR COOLCAT CLUSTERING

Here, we present the results of the clustering method. After the cleaning process, the data, which is entirely categorical, was converted into binary form (or business transaction form), where each category of a variable was treated as a new variable. The COOLCAT algorithm was applied to a random sample of 20,000 accidents that were selected from the reference list of 549,575 accidents that took place between 2016-2018.

For the differentiability and quality of the clusters, an assessment of the *goodness of clustering* needed to be performed in the post-clustering stage, as the total number of clusters is pre-specified in the COOLCAT algorithm. The ideal cluster number for a clustering is one of the topics that there is no scientific consensus as to which clustering is the best (simply because clustering assessments usually depend on the measure that one uses). Commonly used measures include average silhouette (AS) scores, Dunn index (DI), and the DB index, which are all based on distance functions imposed on the data. However, COOLCAT does not use a distance function for clustering, and distance-based assessments may not be ideal. Alternatively, one can use normalized mutual information (NMI), which is an information theoretic measure of the level of clustering. For the best clustering, we compared the scoring indices mentioned above, and the *majority rule* was applied to choose the ideal cluster number.

TABLE 1
CLUSTER QUALITY SCORES

Total Cluster Number	NMI	AS	DI	DB
2	0.21	0.19	2.37	0.8
3	0.26	0.14	2.00	0.88
4	0.28	0.10	1.74	0.92
5	0.31	0.05	1.75	0.93
6	0.33	0.07	1.76	0.92
7	0.34	0.05	1.63	0.92
8	0.35	0.03	1.62	0.92
9	0.34	0.05	1.75	0.91
10	0.35	0.05	1.40	0.91

Quality scores for varying total cluster numbers.

Table 1 shows that the NMI values tend to increase as the cluster number increases (with occasional drops). On the contrary, average Silhouette and Dunn scores tended to decrease with increasing cluster number (all computations were done with Hamming distance). It was observed that the DB score mostly stabilized after k>3 and was somewhat insensitive to the cluster numbers. In these respects $k = 2,3$ do significantly better in obtaining high AS and DI scores. However, NMI scores are very low for $k = 2,3$ (and AS has a theoretical bias towards configuration with low cluster numbers). For $k > 5$, the NMI scores were considerably higher compared to the case with $k<6$; however, the AS and DI scores were substantially low. Therefore, considering all aspects, the optimal cluster number was determined to be $k^* = 6$.

B. INTERPRETATION OF CLUSTERS

As discussed in the introduction, the advantage of the clustering algorithm is that it groups the data into distinct homogenous clusters without making any assumptions about the relationships among the variables. However, this does not inform us about what each cluster represents. Here, we systematically investigated and interpreted the clusters at varying levels of detail.

B.1. Frequency analysis of cluster attributes

The first level of analysis unveils which variables are over- or under-expressed in a particular cluster which are then interpreted as indicators of what that cluster *is* and what it *is not.* Here, the reference measure will be the entire data (all accidents between 2016-2018) which has its own distribution. Therefore, *significant* deviations from the reference distributions are interpreted as signifiers of the cluster under consideration. This deviation was assessed using the Chi-square test for each variable, as introduced in the previous section. The advantage of this approach is that it is free from human bias and provides a simple natural interpretation for each cluster if the clustering algorithm is capable of distinguishing data patterns from each other.

The frequencies of variables in the six clusters formed are compared to the reference frequencies (the whole data), and those variables that showed significant differences ($p<0.05$) were noted. To further strengthen the interpretation, only those variables (among the significant ones) that are over-expressed with at least 1.25 times more than the reference variables are designated as the cluster signifiers or indicators. Tables 2-4 show, for each cluster, the indicator variables and their relative frequencies (ratio of frequency of a variable within a cluster to the overall ratio of in the reference set). A thorough discussion of each cluster is provided in section 6.

B.2. Market Basket Analysis of Clusters with signifiers

The first-level investigation by frequency analysis is complemented by the second-level investigation, market basket analysis (MBA)- which runs on significant variables in each cluster. This is motivated by the idea that although the significant variables are clustered together, they are not necessarily directly linked to each other. MBA helps the variables that are strongly associated with each other to be more precisely identified and provides more arguments to make inferences on the signifiers. Note that it is possible to run the MBA on each cluster with the full set of variables, which has been adopted by some of the previous studies (Pande and Abdel-Aty, 2009). However, we believe that restricted MBA is more meaningful. This is because, on the theoretical side, one is really after those associations that are cluster specific, which describe, with more fidelity, the traffic scenarios that are more likely to occur in that particular cluster. In fact, this has been the whole point of the clustering method to start with, that is, a deeper and more focused analysis of patterns. On the practical side, narrowing down the number of variables significantly reduces the computational time, which will prove profitable if one tries to perform MBA on larger samples.

When applying the MBA, we adjusted the thresholds for the parameters depending on the cluster. The values for the minimal support, confidence, and lift for each cluster are presented in Table 3-8 along with the set of multi-item associations obtained from the Apriori algorithm. After testing, the threshold values of *support = 0.00001*, *confidence = 0.3*, and *lift = 1.5* were chosen. Such a low support threshold was used to allow almost all of the rarest variables to potentially appear in the output rules, as identifying edge cases is important in scenario testing. The confidence and lift thresholds were chosen as they provided a good number of strong rules. They also guarantee that for every rule, the consequent appears in at least one-third of the accidents in the cluster containing the antecedent (from the 0.3 confidence) and that the rule is observed over %50 percent more often than expected compared to random occurrence (from the lift value of 1.5).

To help give a high level understanding of the generated associations, a plot for each cluster was generated using the python package pyvis which shows the strongest links between variables. These are shown in the appendices (figures 8-13).

VI. DISCUSSION

A. UNDERSTANDING CLUSTERS WITH COOLCAT

TABLE 2.
SIGNIFIERS FOR CLUSTERS 1-2.

Cluster1	Rel. freq.	Cluster2	Rel. freq.
Serious	1.35	C	1.27
Fatal	3.37	Unclassified	1.49
Skidded/Jack-knifed	2.62	Unclassified2	2.19
Overturned	2.73	20mph	1.26
12am-3am	1.65	30mph	1.41
3am-6am	2.05	T or staggered junction	2.11
Motorway / A(M)	6.17	Private drive or entrance	2.09
Object on road	1.59	Entering junction	1.54
Pedestrian or animal on road	2.23	Clearing junction	1.52
Not a junction in 20m	2.65	Mid junction	1.73
50mph	2.56	Give way/ stop sign or uncontrolled	1.59
60mph	3.96	Bicycles	1.72
70mph	5.27	Motorbikes	1.36
Not a junction within 20m	2.67	Turning left	1.90
Not at or within 20 metres of junction	2.65	Turning right	1.98
Not at junction or within 20 metres	2.59		

Traffic light/ person	1.33
Darkness – no lights	3.80
High Winds	1.62
Goods	1.70
Changing lane to left	2.08
Changing lane to right	2.11
Overtaking moving vehicle - offside	1.56
Going ahead-bend	3.48
Nearside / nearside and rebounded	3.06
Offside / offside rebounded/crossed/etc	2.80
Wet/damp	1.37
Snow/Flood	2.46
Frost or ice	2.96
No Pedestrian Crossing	1.25

Significant variables for Clusters 1-2 and their relative frequencies with respect to the reference (unclustered) full data.

For Cluster 1, one reads from Table 2 that it is a *severe (i.e., serious and fatal)* accident cluster. It is also a *non-junction* cluster depicting accidents that took place on *motorways with high-speed limits* (50-70 mph) in *late night* in *dark places with no light.* These accidents in this cluster appear to involve *pedestrians* or *objects on the road*, *which* might be one of the reasons why *fatal and serious* accidents are over-expressed in this cluster. Adverse weather and road conditions such as h*igh winds, snowy weather, and frosty surfaces* seem to have played a role in drivers' loss of vehicle control and hit the *nearside* and *offside* of the road, causing such severe accidents. As this is a *non-junction* cluster with a high road *speed limit*, the related maneuvers are, expectedly, *overtaking* and *changing lanes.*

Cluster 2 (Table 3) significant variables suggest that this is a minor road cluster *(C roads and unclassified) at junctions* with *low-speed limit* (20-30 mph) involving more dominantly two-wheelers *(bikes and motorbikes)*. Being an *at-a-junction* cluster with two wheelers, the key maneuver types leading to accidents appear to be *left turns* and *right turns* (as one would expect).

TABLE 3.
SIGNIFIERS FOR CLUSTERS 3-4.

Cluster3	**Rel. freq.**	**Cluster4**	**Rel. freq.**
Fatal	1.48	Unclassified	1.44
Skidded/Jack-knifed	1.81	Not a junction in 20m	2.67
Overturned	1.67	30mph	1.27
A	1.37	Not a junction within 20m	2.69
Object on road	1.89	Not at or within 20 metres of junction	2.66
Pedestrian or animal on road	1.71	Not at junction or within 20 metres	2.56
Motorway / A(M)2	6.97	Traffic light/ person	1.34
A2	2.00	Buses/Trams	1.58
B2	2.16	Reversing	2.17
C2	2.57	Parked	2.95
40mph	2.18	Slowing or stopping	1.33
50mph	2.63	U-turn	2.03
60mph	2.62	Overtaking static vehicle - offside	1.72
70mph	1.59	Back	1.29
Slip road	5.44		
T or staggered junction	1.33		
Roundabout / mini-roundabout	2.45		
More than 4-arms / other junction	1.64		
Private drive or entrance	1.59		
Entering junction	1.71		
Clearing junction	1.58		
Mid junction	1.35		
Give way/ stop sign or uncontrolled	1.45		
Darkness – no lights	1.92		
High Winds	1.67		
Fog/Mist	2.19		
Motorbikes	1.25		
Turning left	1.80		
Changing lane to left	1.98		
Changing lane to right	1.64		
Going ahead-bend	1.49		
Nearside / nearside and rebounded	1.64		
Offside / offside rebounded/crossed/etc	1.90		
Old	1.26		
Wet/damp	1.31		
Snow/Flood	2.19		
Frost or ice	1.80		
Oil or mud	3.93		
Was_Vehicle_LHD?Yes	3.62		

Significant variables for Clusters 2-4 and their relative frequencies with respect to the reference data.

TABLE 4.
SIGNIFIERS FOR CLUSTERS 5-6.

Cluster5	**Rel. freq.**	**Cluster6**	**Rel. freq.**
A2	2.12	12am-3am	1.57
C2	1.35	3am-6am	1.69
Unclassified2	1.53	9pm-12am	1.50
T or staggered junction	1.58	A	1.61
Roundabout / mini-roundabout	2.02	A2	3.54
More than 4-arms / other junction	1.96	B2	3.27
Private drive or entrance	2.26	C2	3.15
Entering junction	2.35	20mph	1.31
Clearing junction	1.43	Crossroads	3.84
Give way/ stop sign or uncontrolled	1.29	Roundabout / mini-roundabout	1.80
Reversing	5.03	More than 4-arms / other junction	2.26
Parked	2.43	Clearing junction	1.65
Waiting	6.39	Mid junction	2.26
Slowing or stopping	3.87	Traffic light/ person	2.83
Moving off	2.52	Darkness - lights lit	1.66
Back	3.82	Bicycles	1.48
Female	1.39	Buses/Trams	2.33
		Moving off	1.42
		Turning left	1.33
		Turning right	1.48
		Nearside	1.27

	Pedestrian Crossing	3.40

Significant variables for Clusters 5-6 and their relative frequencies with respect to the reference data.

Cluster 3 is also a severe accident cluster indicated by the *fatal* accidents attribute. The main differences from Cluster 1 are that Cluster 3 is a *junction* cluster and the accidents in this cluster mostly occur on *A-roads* instead of m*otorways* which are important distinctions. Among the junctions, *slip roads* deserve special attention as they are highly over-expressed (*rel. freq.=5.44).* Adverse weather and road conditions also play a significant role in this cluster. Driving on high-speed *limit* roads under adverse weather with risky maneuver types at a junction (such as *changing lane to left, changing lane to right, going ahead with bend)* seem to have led to vehicles losing control *and leaving the carriageway (i.e., hitting the roadsides and getting rebounded)* which may be the reason behind *severe* outcomes. This cluster also has an interesting element, that is, accidents of *left-hand drive* (LHD) *vehicles* (European vehicles) which are generally ignored in most accident analyses due to being rare cases (but nevertheless important as we shall see later in this section).

Cluster 4 describes the *off-junction accidents* like Cluster 1. However, there are important differences. First, accidents in this cluster occur on roads with *slow speed limit.* Second, most of these accidents occur on *unclassified minor* roads where one can see parked or reversing vehicles. Interestingly the accidents frequently involve *buses* and *trams.*

Cluster 5 is another *junction* cluster but without adverse weather conditions. It predominantly involves *roundabouts, many-armed junctions* and *private drives*. The accidents in this cluster take place, mostly, at junction entrances. Interestingly, maneuvers which would normally be regarded as safe are substantially more expressed in this cluster such as *parked, reversing* and *waiting.* Therefore, a deeper analysis of this cluster can yield unexpected associations between these accident attributes.

Finally, Cluster 6 is a *night* cluster describing accidents that take place on *A-roads.* The difference from Cluster 3 is that the accidents happen at very *low speed limit* roads (20 mph). And differently from Clusters 2,3 and 5, in this cluster, *mid-junction* accidents are more prevalent in this cluster. Another specialty concerning this cluster is that this is the only cluster with *crossroads* type junction as significant. Curiously, bicycles and buses are more commonly represented in this cluster.

B. ASSOCIATION OF ATTRIBUTES IN CLUSTERS AND TEST SCENARIO GENERATION

As emphasized in the introduction, one of the main motivations of this study is the identification of the test scenarios (temporal/spatial conditions) for CAVs that are correlated with important outcomes. The clusters formed in Section 5, along with the significant variables identified, enabled us to find such conditions. In this section, we explicitly demonstrate how this is done using MBA. It should be noted at the outset that the MBA procedure is operated only on the *significant* variables of each cluster to extract the most relevant scenarios. This means that, based on the analysis of Section 5, no scenario will have the *Weekend or Weekday* and *Urban or Rural Area* variables as these variables were not found to be *significant.* Such information can either be deduced from the context or be generated randomly if they were to be included in a simulation.

In order to mine the most interesting associations the MBA parameters are taken according to the characteristics of each cluster (e.g. by varying the support threshold of variables in the respective clusters). Here we first display and discuss, in Tables 5-10, the top-ranking associations in terms of their confidence or lift values. For the purposes of scenario generation, the standard MBA procedure is modified considerably. First, repeating rules (from each cluster) are removed. Second, associations that are not mutually exclusive are combined in a consistent way to yield longer associations. The longer the association rule, the more detailed the concrete scenario. The rationale is that each independent rule depicts the strong tendency of a set of variables to appear together. A natural combination of such rules forms the conditions/characteristics (environment-related or driver-related) of a scenario. We note here that we do not require *an order or direction* for the associations of attributes that allow flexibility to focus on different accident settings. It should be emphasized that no hard rules (except for the requirement of a maneuver) are imposed to derive the scenarios; in principle, any compatible combination of rules and the attributes with high confidence and high lift could be a scenario candidate. Also, no claim is made on the presented exemplary scenarios being unique (they probably are not). Each exemplary scenario represents a non-trivial, interesting situation that is present in the respective cluster and leads to important consequences.

For each exemplary scenario a diagram was created using SUMO (Simulation of Urban Mobility) to aid with visualization [60].

For Cluster 1, we recall that this is a *serious* or *fatal* accident cluster on a *motorway* and *away from a junction* Understandably *lane-changing* maneuvers (rule #1) (to *right* or left) and *overtaking* combined with negative environmental conditions are associated with serious outcomes such as *leaving the carriageway* and *overturning* (rule #2) or *skidding/jack-knifing* (rule #3). From the association rules, it can also be inferred that *goods* vehicles are more at risk of getting involved in motorway accidents than other vehicles. Other rules can be interpreted in a similar manner (Table 5).

Exemplary scenario 1. A vehicle *overtakes another vehicle* that is *moving off* on a *motorway* with a *wet* surface. A possible *outcome* for this scenario is that it leads to an accident that causes *skidding and* rebounding from the nearside (rule #3) as shown in figure 2.

TABLE 5.
ASSOCIATION RULES FOR CLUSTER 1.

#	Ante.	Cons.	Sup.	Conf.	lift
1	Goods, Darkness – no lights, Changing lane to left	Motorway / A(M)	7.64E-05	0.66	20.46
2	Overturned, Serious, Motorway / A(M), Changing lane to left	Left Nearside	2.55E-05	0.64	8.51
3	Overtaking MV-off, Wet/damp, Motorway / A(M), Left Nearside	Skidded/Jack-knifed	2.55E-05	0.54	7.28
4	Goods, Changing lane to right	Motorway / A(M)	0.000828	0.43	13.46
5	Motorway / A(M), Right Offside, Overtaking MV-off	Wet/damp	4.19E-05	0.37	1.51
6	Going ahead-bend, Motorway / A(M), Darkness – no lights, Left Nearside	Skidded/Jack-knifed	1.64E-05	0.31	4.20

Association rules via application of MBA procedure on Cluster 1.

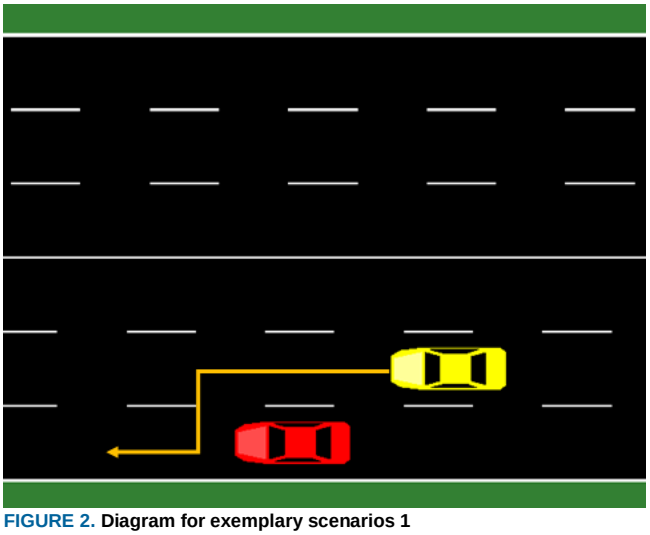

FIGURE 2. Diagram for exemplary scenarios 1

For Cluster 2, Table 6 lists some of the main associations. Cluster 2, being a two-wheeler cluster, comprises traffic situations for *bicycles* or *motorbikes*. Rule #1 indicates that accidents at *private drive or entrance*, when *clearing junction* to an *unclassified road* are strongly linked to *turning right* maneuvers. Rule #2 illustrates a scenario for *bicycles* on *unclassified roads*, but while *turning left* to an *unclassified road* clearing a junction. Both rules have high lifts.

Exemplary scenario 2. A bicycle on an *unclassified road* at a *T or staggered junction* makes a *left turn* and when about to *clearing the junction* gets into an accident (rule #2) as shown in figure 3.

TABLE 6.
ASSOCIATION RULES FOR CLUSTER 2

#	Ante.	Cons.	Sup.	Conf.	lift
1	Clearing junction, Private drive or entrance, Unclassified	Turning right	6.66 E-4	0.45	4.03
2	Turning left, T or staggered junction, Bicycles, Unclassified	Clearing junction	1.69 E-4	0.31	3.04

Association rules via application of MBA procedure on Cluster 2.

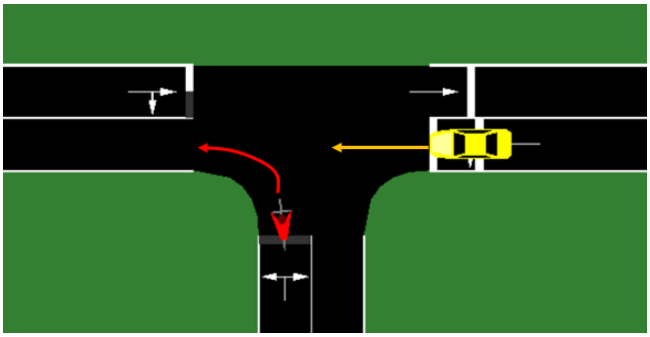

FIGURE 3. Diagram for exemplary scenarios 2

For Cluster 3, a number of interesting scenarios can be generated (Table 7). Again we describe the first few interesting rules (giving scenarios) and others can be interpreted in the same way. Rule #1 describes a situation in which vehicles are *going ahead and bending* at a *T or staggered junction* on an *A road* in a *windy* day gets into a crash and hit from the *nearside*. Such accidents are strongly linked to road surface being *wet/damp*. Rule #2 suggest that drivers should be careful at *T or staggered junction* as *going ahead and bending* to *clear junctions* on frosty/icy roads are strongly associated with accidents at such junctions.

Exemplary scenario 3. A vehicle during *high winds* at a *roundabout* of an *A road goes ahead and bend* in the *middle of the junction* and gets into an accident. The road surface was wet (rule #3) as shown in figure 4.

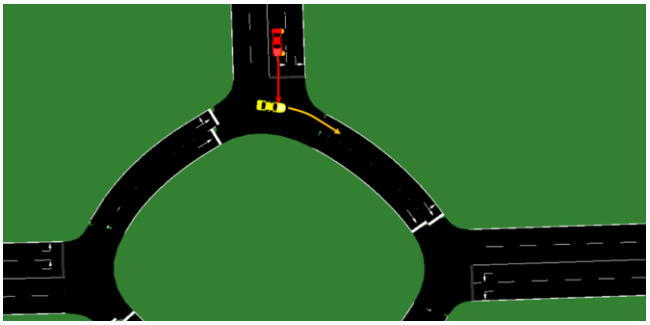

FIGURE 4. Diagram for exemplary scenario 3

TABLE 7.
ASSOCIATION RULES FOR CLUSTER 3

#	Ante.	Cons.	Sup.	Conf.	lift
1	Going ahead-bend, High Winds, T or staggered junction, A, Left Nearside	Wet/damp	2E-05	0.65	2.68
2	Clearing junction, Frost or ice, Going ahead-bend	T or staggered junction	8.92 E-05	0.63	1.96
3	Going ahead-bend, High Winds, Mid junction, Roundabout, A	Wet/damp	2.91 E-05	0.62	2.55
4	Motorbikes, Mid junction, Overturned, Going ahead-bend, Wet/damp, A	Roundabout	1.09 E-05	0.60	5.39
5	Turning left, Fog/Mist, Entering junction, Right Offside	T or staggered junction	1.09 E-05	0.60	1.87
6	Turning left, Motorbikes, Entering junction	T or staggered junction	0.00 0355	0.58	1.80
7	Skidded/Jack-knifed, Left Nearside ,Going ahead-bend, Wet/damp, Roundabout, A	Clearing junction	6.37 E-05	0.56	5.47
8	Clearing junction, T or staggered junction, High Winds, Going ahead-bend	Wet/damp	5.28 E-05	0.56	2.31
9	Turning left, More than 4-arms / other ,Wet/damp, A, Left Nearside	Clearing junction	1.09 E-05	0.55	5.28
10	Going ahead-bend, Wet/damp, LHD?Yes, Entering junction	T or staggered junction	2.37 E-05	0.54	1.69
11	Going ahead-bend, Slip road, Left Nearside	Wet/damp	0.00 0111	0.51	2.12
12	Skidded/Jack-knifed, Going ahead-bend, High Winds, Wet/damp, Roundabout, A	Clearing junction	1.27 E-05	0.50	4.84
	Changing lane to right, A, Mid junction	Roundabout	0.00 048	0.50	4.45
13	Entering junction, Darkness – no lights, T or staggered junction, Going ahead-bend, A	Wet/damp	9.46 E-05	0.49	2.01
14	Mid junction, A, Changing lane to left, Old	Roundabout	6.19 E-05	0.48	4.30
15	Clearing junction, Roundabout, Going ahead-bend, Left Nearside	A, Wet/damp	9.64 E-05	0.42	3.99
16	A, Changing lane to left, Skidded/Jack-knifed	Left Nearside	0.00 0247	0.42	5.56
17	Turning left, Right Offside, Skidded/Jack-knifed	T or staggered junction	0.00 0202	0.41	1.27
18	Clearing junction, Darkness – no lights, T or staggered junction, Left Nearside	Going ahead-bend	9.64 E-05	0.40	5.66
19	High Winds ,A, Mid junction, Going ahead-bend	T or staggered junction	5.28 E-05	0.39	1.20
20	Changing lane to left, Slip road, LHD?Yes	Overturned	9.1E -06	0.38	7.52
21	More than 4-arms / other ,Skidded/Jack-knifed, Entering junction, Going ahead-bend, Wet/damp, A	Left Nearside	1.46 E-05	0.38	5.09
22	Going ahead-bend, Fog/Mist, Roundabout	Wet/damp	3.46 E-05	0.36	1.49
23	Clearing junction, Skidded/Jack-knifed, Darkness – no lights, Going ahead-bend, Wet/damp	T or staggered junction, Left Nearside	2E-05	0.35	24.28
24	Turning left, Motorbikes, Wet/damp, Roundabout, Skidded/Jack-knifed	Clearing junction, A	1.27 E-05	0.35	7.23
25	Turning left, Right Offside, A	Clearing junction	0.00 0266	0.34	3.30
26	Overturned, T or staggered junction, Mid junction, Going ahead-bend	Left Nearside	9.64 E-05	0.34	4.54
27	Going ahead-bend, A, Mid junction, Left Nearside	Roundabout	0.00 0142	0.34	3.04
28	Motorbikes, Clearing junction, Skidded/Jack-knifed, Going ahead-bend, Wet/damp	A, Roundabout	1.46 E-05	0.33	5.05
29	Entering junction, Private drive or entrance ,Wet/damp, Going ahead-bend	Left Nearside	2.37 E-05	0.32	4.24
30	Clearing junction, T or staggered junction, Going ahead-bend, Left Nearside	Skidded/Jack-knifed	1.95 E-5	0.32	4.28
31	Mid junction, Skidded/Jack-knifed, Darkness – no lights, Going ahead-bend, A	T or staggered junction, Wet/damp	1.27 E-05	0.30	4.15

Association rules via application of MBA procedure on Cluster 3.

Cluster 4 describes accidents with *back* impact points which are generally found on minor roads (*unclassified)* of urban areas where vehicles often need to *reverse* their vehicles to park or to get into the road. The level of detail provided by this rule is low. As discussed earlier, in such cases, for scenario development, other relevant variables defining a scenario can be generated randomly.

TABLE 8.
ASSOCIATION RULES FOR CLUSTER 4

#	Ante.	Cons.	Sup.	Conf.	lift
1	Reversing	Back, Unclassified	0.009831	0.55	12.51

Association rules via application of MBA procedure on Cluster 4.

Exemplary scenario 4. A *vehicle reverses* on an *unclassified road* and gets hit from the *back* (rule #1) as shown in figure 5.

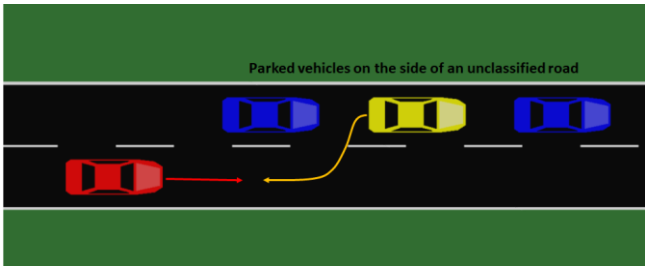

FIGURE 5. Diagram for exemplary scenario 4

Cluster 5 is a true junction cluster that mostly involves female drivers. Here we discuss the most strongly associated conditions. Rule #1 describes situations in which the vehicles are hit from the back while *moving off* and *entering the roundabout.* Rule #2 also describes an *entering junction situation* by *reversing at a private drive or entrance.*

TABLE 9.
ASSOCIATION RULES FOR CLUSTER 5

#	Ante.	Cons.	Sup.	Conf.	lift
1	Entering junction, Moving off, Back, Female	Roundabout	0.000446	0.48	4.35
2	Entering junction, Private drive or entrance ,Back, Female	Reversing	0.000313	0.35	19.6

Association rules via application of MBA procedure on Cluster 5.

Exemplary scenario 5. A vehicle reverses to *a private drive or entrance* and gets hit from the *back* while *entering the junction* (rule #2) as shown in figure 6.

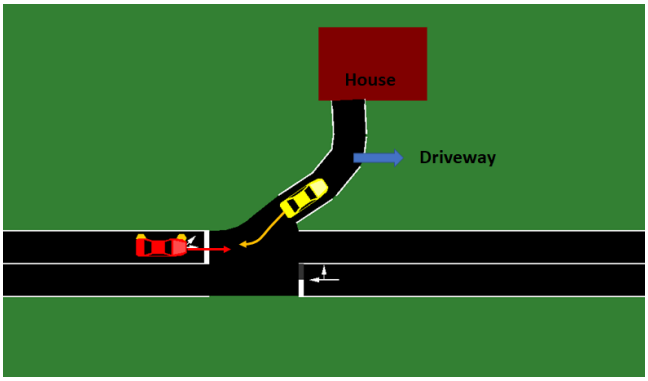

FIGURE 6. Diagram for exemplary scenario 7

Finally, a set of association rules, mined from Cluster 6 focusing on accidents involving buses/trams and bicycles, on crossroads are described in Table 9. Rule #1 indicates that of the accidents that involve *buses/trams* at *mid-junctions* that are trying *turn right*, a significant portion of them happen at crossroads. Also, when *buses/trams* that try *changing lane to left* end up, almost certainly, with crashes impacting on *nearside* (rule #2). Also, for *buses/trams* driving *at night*, *roundabouts* pose risks especially when *clearing junction* rule #3). Rule #4 depicts general situations linking *crossroad* accidents to *turning left* maneuvers which resulted in *nearside crashes* at *low speeds* at *mid-junctions*. On the other hand, *4-arm/other junction* accidents which involve *turning right* happen almost always when *clearing junctions (*rule #5). Furthermore, rule #7 suggests that cyclists who are *turning right* and clearing *junctions* are linked to accidents at *roundabouts.* Finally, rule #8 suggests that accidents in which *bicycles change lane left* almost certainly take place on crossroads.

TABLE 10.
ASSOCIATION RULES FOR CLUSTER 6

#	Ante.	Cons.	Sup.	Conf.	lift
1	Mid junction, Turning right, Buses/Trams	Crossroads, Traffic light/ person	1.0 E-3	0.75	2.15
2	Changing lane to left, Buses/Trams	Nearside	1.0 E-3	1.00	5.20
3	Roundabout / mini-roundabout, Darkness - lights lit, Buses/Trams	Clearing junction	1.0 E-3	0.75	4.39
4	Nearside, Mid junction, 20mph,Turning left	Crossroads	1.0 E-3	1.00	2.44
5	Nearside, Darkness - lights lit, Turning right, More than 4-arms / other junction	Clearing junction, Traffic light/ person	1.0 E-3	1.00	9.54
6	Clearing junction, More than 4-arms /	Nearside	1.1 E-3	0.6	3.12

	other junction, Buses/Trams				
7	Turning right, Bicycles, Clearing junction	Roundabout / mini-roundabout	1.0 E-3	0.60	3.00
8	Changing lane to left, Bicycles, Traffic light/ person	Crossroads	1.0 E-3	1.00	2.44

Association rules via application of MBA procedure on Cluster 6.

Exemplary scenario 6. A *bus* driving in *darkness with lights lit* makes a *right turn* on a junction *with more than 4-arms* and when *clearing* the junction gets into an accident (and *hit from nearside)* (rule #5) as shown in figure 6.

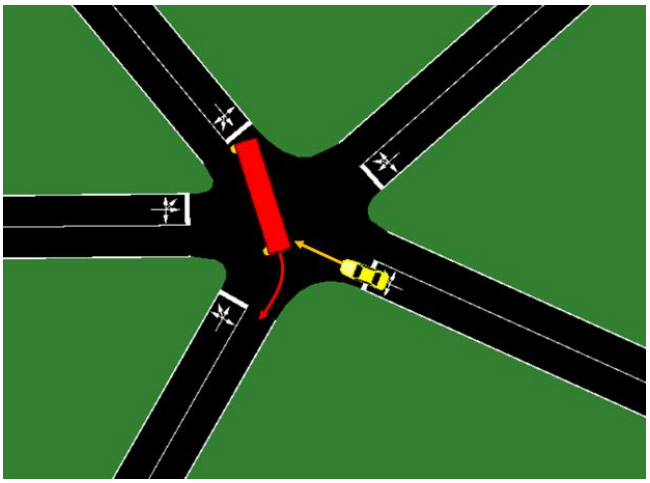

FIGURE 7. Diagram for exemplary scenario 6

VII. CONCLUSION

This study aimed to achieve two high-level objectives. The first objective was to underpin the research on safety analysis of traffic accidents by identifying patterns based on past accident records. This was performed using a cluster analysis method. This approach reveals the natural patterns in the data without making any prior modelling assumptions, which is advantageous considering the complexity of factors that can affect the outcomes. The second objective was to develop a method based on the information obtained from accident clusters, which will help design test case scenarios for AVs, thus filling an important gap in the industry. To achieve both objectives, several novel approaches were taken deepening some of the existing methods to obtain more useful results while considering possible future challenges in industrial applications (such as handling of continuously growing large datasets).

For the first objective, the COOLCAT clustering algorithm was used on the processed STATS19 dataset to determine the natural grouping of accidents. COOLCAT employs natural global clustering criteria (entropy) which suits particularly well to cluster noisy categorical data and is able to handle large dimensions with ease. To the best of our knowledge, this is the first application of the COOLCAT algorithm in traffic accident research. Using various cluster quality metrics, six clusters are obtained from the algorithm. The frequency tests conducted on each cluster indicated that Cluster 1 was described by *nighttime serious/fatal* accidents on *motorways away from the junctions*, which involved *changing lanes (right/left) and ended up with a skidding/overturning* vehicle*;* Cluster 2 was described by *minor road* accidents by *two-wheelers at junctions* on *low-speed limit roads* involving *right/left turns;* and Cluster 3 by *fatal/serious* accidents on *A roads* but *at junctions (*especially *slip roads)* by *left-hand driving vehicles.* Similarly, Cluster 4 can be represented by accidents on *unclassified roads* with *low-speed limits* (likely to be narrow street roads*) away from junctions* involving *U-turn* or *reversing* maneuvers, which often ended in hits from *the back;* and Cluster 5 depicts relatively more *minor* accidents *at junctions* with 'gentle' maneuvers such as *parked, waiting, and moving off.* Finally, Cluster 6 describes accidents at *junctions of* road A with *a low-speed limit where the main maneuver* types were *turning right/left* or *moving off.* The results suggest that particular care should be given in making policies/regulations for elements described in the clusters.

For the second objective, based on the information obtained from the clusters, the MBA methodology was applied for association rule mining. As the standard MBA produces repetitive rules (when ordering is not counted), which may only partially describe accidents, we extended the method considerably by systematically combining non-conflicting rules that provided much higher details for the test scenarios. As expected, scenarios obtained from this procedure reflect the characteristics of the cluster that they come from. Once the scenarios are obtained, they can be used in real or virtual environments for CAV training by varying the unspecified attributes as free variables. This will significantly speed up the training processes of CAVs, as they will be driven on quality miles rather than on random routes.

There are theoretical and practical implications of this work. First clustering, as a method for accident analysis, is underexploited. It can be used along with other existing methods (e.g., regression) and enhance them by homogenizing the data. Furthermore, data specific cluster models, such as COOLCAT can serve to better obtain higher quality results instead of more generic algorithms. On the practical front, the output of this work has immediate industrial applications. The proposed approach provides an a-to-z methodology to generate, in a nearly automated manner, high quality test scenarios that can be used in simulations by manufacturers. In fact, test scenarios obtained via the proposed method are now (after data formatting adjustments) deposited into the recently launched, world's largest scenario repository, SafetyPool™ [61].

There are also apparent limitations of this work, mostly due to the scope of the data that was used. The analysis can provide details to the extent that the data can provide, but not more. Although we tried to keep the number of attributes high, the real world contains conditions that may be important but not covered in the present data (such as the position of the sun and curvature of the road). In future studies, multiple data sources can be combined to provide a more detailed description of each accident, which will affect the formation of accident clusters and the association rules extracted from those clusters (i.e., more detailed test scenarios).

REFERENCES

[1] Road Safety Data - STATS19. (2020), UK Department for Transport: [Online] https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

[2] J. B. Cicchino, "Effectiveness of forward collision warning and autonomous emergency braking systems in reducing front-to-rear crash rates," *Acc. Anal. Prev.*, vol. *99*, pp.142, 2017.

[3] M. Guériau, R. Billot, N. E. El Faouzi, J. Monteil, F. Armetta, S. Hassas. "How to assess the benefits of connected vehicles? A simulation framework for the design of cooperative traffic management strategies," *Trans. Res. C*, vol. 67, pp. 266–279, 2016.

[4] C. Tingvall, "The Zero Vision: A Road Transport System Free from Serious Health Losses," In "*Transportation, Traffic Safety & Health: The New Mobility*, pp. 37–57, 1997.

[5] S. Khastgir, S. Birrell, G. Dhadyalla, P. Jennings, "Calibrating trust through knowledge: Introducing the concept of informed safety for automation in vehicles," *Trans. Res. C*, vol. 96, pp. 290–303, 2018.

[6] R. N. Charette. "This Car Runs on Code". *IEEE Spectrum*, 2009. [Online] http://www.real-programmer.com/interesting_things/IEEE%20SpectrumThisCarRunsOnCode.pdf

[7] S. Khastgir, S. Birrell, G. Dhadyalla, P. Jennings, "Identifying a gap in existing validation methodologies for intelligent automotive systems: Introducing the 3xD simulator," *Proc. of the IEEE Intel. Veh. Symp.* pp. 648–653, *2015*

[8] N. Kalra, & S. M. Paddock, "Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?" *Trans. Res A: Policy and Prac*, vol. 94, pp. 182–193, Dec. 2016.

[9] S. Khastgir, S. Birrell, G. Dhadyalla, P. Jennings, The Science of Testing: An Automotive Perspective. *SAE Technical Paper:* 2018-01-1070. DOI. https://doi.org/10.4271/2018-01-1070

[10] S. Khastgir, S. Brewerton, J. Thomas, P. Jennings, Systems "Approach to Creating Test Scenarios for Automated Driving Systems." *Reliab. Eng. & Syst. Safe.*, vol. 215, 2021, Art. No. 107610

[11] E. Esenturk, S. Khastgir, A. Wallace, P. Jennings, "Analyzing Real-World Accidents for Test Scenario Generation for Automated Vehicles", *IEEE Intell. Veh. Symp.* Jul 2021.

[12] P. Nitsche, P. Thomas, R. Stuetz, R. Welsh, "Precrash scenarios at road junctions: a clustering methods for car crash data," *Acc.t Anal. Prev.*, vol. 107, pp. 137-151, 2017.

[13] F.L. Mannering, C.R. Bhat, "Analytic methods in accident research research: Methodological frontier and future directions," *Anal. Meth. Accid. Res.* vol. 1, pp. 1–22, 2014.

[14] C. Caliendo, M. Guida, A. Parisi, "A crash-prediction model for multilane roads," *Acc. Anal. Prev.*, vol. *39* no. 4, pp. 657–670, 2007. DOI. https://doi.org/10.1016/j.aap.2006.10.012

[15] D. Lord, A. Manar, & A Vizioli, "Modeling crash-flow-density and crash-flow-V/C ratio relationships for rural and urban freeway segments," *Acc. Anal. Prev.*, vol. *37*, pp. 185–199, 2005.

[16] P.T. Savolainen, F.L. Mannering, D. Lord, M.A. Quddus, "The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives," *Accid. Anal. Prev.* vol. 43, pp. 1666–1676, 2011.

[17] R. Yu, M. Abdel-Aty, "Using hierarchical Bayesian binary probit models to analyze crash injury severity on high speed facilities with real-time traffic data," *Accid. Anal. Prev.* vol. 62, pp. 161–167, 2014. DOI:10.1016/j.aap.2013.08.009

[18] J. Ma, K.M. Kockelman,, P. Damien, "A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods," *Accid. Anal. Prev.* vol. 40 , pp. 964–975, 2008. DOI:10.1016/j.aap.2007.11.002

[19] C. Lee, M. Abdel-Aty, "Comprehensive analysis of vehicle–pedestrian crashes at intersections in Florida," *Acc. Anal. Prev.*, vol. 37, pp. 775-786, 2005

[20] M. Hossain, Y. Muromachi, "A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways," *Accid. Anal. Prev.* vol. 45, pp. 373–381, 2012. DOI:10.1016/j.aap.2011.08.004.

[21] Q. Zeng, H. Huang, X. Pei, S.C. Wong, "Modeling nonlinear relationship between crash frequency by severity and contributing factors by neural networks," *Anal. Methods Accid. Res.* vol. 10, pp. 12–25, 2016. DOI:10.1016/j.amar.2016.03.002

[22] M. Nowakowska, "Selected aspects of prior and likelihood information for a Bayesian classifier in a road safety analysis," *Acc. Anal. Prev.*, vol. 101, pp. 97–106, 2017.

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

[24] K. Xie, X. Wang, H. Huang, X. Chen, "Corridor-level signalized intersection safety analysis in Shanghai, China using Bayesian hierarchical models," *Acc. Anal. Prev.* vol. 50, pp. 25–33, 2013. DOI. https://doi.org/10.1016/j.aap.2012.10.003

[25] H. Huang, & M. Abdel-Aty, "Multilevel data and Bayesian analysis in traffic safety," *Acc. Anal. . Prev.*, vol. *42* no. 6, pp. 1556–1565, 2010.

[26] H. T. Abdelwahab, & M. A. Abdel-Aty. "Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections," *Trans. Res. Rec.*, vol. 1746, pp. 6–13, 2001.

[27] N. Formosa, M. Quddus, S. Ison, M. Abdel-Aty, J. Yuan. "Predicting real-time traffic conflicts using deep learning," *Acc. Anal. Prev.*, vol. 136, Art no. 105429, 2020.

[28] Y. C. Chiou, "An artificial neural network-based expert system for the appraisal of two-car crash accidents," *Acc. Anal. Prev.*, vol. 38, pp. 777–785, 2006.

[29] D. Delen, R. Sharda, & M. Bessonov, "Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks," *Acc. Anal. Prev.* vol. 38, pp. 434–444, 2006.

[30] S. Das, A. Dutta, R. Avelar, K. Dixon, X. Sun, M. Jalayer, "Supervised association rules mining on pedestrian crashes in urban areas: identifying patterns for appropriate countermeasures," *Int. J. Urban Sci.* vol. 23, pp. 38–40, 2019.

[31] A. Montella, "Identifying crash contributory factors at urban roundabouts and using association rules to explore their

relationships to different crash types," *Acc. Anal. Prev.* vol. 43, pp. 1451–1463, 2011, DOI. https://doi.org/10.1016/j.aap.2011.02.023
[32] A. Pande, M. Abdel-Aty. "A novel approach for analyzing severe crash patterns on multilane highways," *Accid. Anal. Prev.* vol. 56, pp. 95–10, 2009. DOI. https://doi.org/10.1016/j.aap.2009.06.003
[33] J. De Oña, G. López, J. Abellán, "Extracting decision rules from police accident reports through decision trees," *Accid. Anal. Prev.*, vol. 50, pp. 1151–1160 2013. DOI. https://doi.org/10.1016/j.aap.2012.09.006
[34] G. López, J. Abellán, A. Montella, J. De Oña, "Patterns of single-vehicle crashes on two-lane rural highways in Granada Province, Spain: in-depth analysis through decision rules," *Trans. Res. Rec*. vol. 2432, pp. 133–141, 2014. https://doi.org/10.3141/2432-16
[35] L. Y. Chang, W. C. Chen, "Data mining of tree-based models to analyze freeway accident frequency," *J. Safety Res*. vol. 36 no. 4, pp. 365–375, 2005.
[36] C. Lee, & X. Li, "Predicting driver injury severity in single-vehicle and two-vehicle crashes with boosted regression trees." *Trans. Res. Rec*, vol. 2514, pp. 138–148, 2015.
[37] A. Montella, M. Aria, A. D'Ambrosio, F. Mauriello, "Analysis of powered two wheeler crashes in Italy by classification trees and rules discovery," *Accid. Anal. Prev*. vol. 49, pp. 58–72, 2012.
[38] C. Aggarwal, C. Zhai, "A survey of text clustering algorithms. Mining Text Data." New York, NY, USA. Springer-Verlag: 2012. pp. 77–128
[39] S. Kumar, D. Toshniwal, "A data mining framework to analyze road accident data," *J. Big Data*, 2015, Art. no. 26
[40] T. K. Anderson, "Kernel density estimation and k-means, clustering to profile road accident hotspots," *Accid. Anal. Prev.*, vol. 41, 2009, pp. 359–364
[41] C. Zhang, J. N. Ivan, and T. Jonsson, "Collision type categorization based on crash causality and severity analysis," *86th Ann. Meet. Trans. Res. Board*, Washington, D.C., 2007.
[42] A. Iranitalab, and A. Khattak, "Comparison of four statistical and machine learning methods for crash severity prediction," *Acc. Anal. Prev*., vol. 108, pp. 27–36, 2017.
[43] Z. Ta., C. Yaoyue, X. Lingyun, H Wenhao, L. Pingfei, X. Jin, "Research of fatal car-to-pedestrian precrash scenarios for the testing of the active safety system in China," *Accid. Anal. Prev*., vol. 150, 2021, Art. No. 105857.
[44] J. Lenard, A. Badea-Romero, R. Danton, "Typical pedestrian accident scenarios for the development of autonomous emergency braking test protocols," *Accid. Anal. Prev*., vol. 73, pp. 73-80, 2014.
[45] B. Sui, N. Lubbe, J. Bargman, "A clustering approach to developing car to two-wheeler test scenarios for the assessment of aotimated emegerncy braking in China using in-deptsh Chinese crash data," *Accid. Anal. Prev*, vol. 131, 2019, Art. no. 105242.
[46] J. Mcqueen, "Some methods for classification and analysis of multivariate observations," *Proc. 5th Berkeley Symposium Math. Stat. Prob., Berkeley, CA, USA*, pp. 281–97, 1967.
[47] HS Park, CH Jun, "A simple and fast algorithm for k-medoids clustering," *Expert Syst. App.*, vol. 36 no. (2), pp., 3336–41, 2009.
[48] H. Huang, M. Abdel-Aty, "Multilevel data and Bayesian analysis in traffic safety," *Accid. Anal. Prev.*, vol. 42, pp. 1556–1565, 2010.
[49] D. Barbara, Y. Li, J. Couto, "COOLCAT: an entropy-based algorithm for categorical clustering," *Proc. 11th Int Conf Inf. Know. Manag.*, pp. 582-589, 2002.
[50] S. Guha, R. Rastogi, K. Shim, "ROCK: A robust clustering algorithm for categorical attributes," *Proc. 1999 Int. Conf. Data Eng.*, Sydney, Australia, pp. 512-521, Mar., 1999.
[51] Ester M, Kriegel H P, Sander J, Xu X. "A density-based algorithm for discovering clusters in large spatial databases." In *Proc. 1996 Int. Conf. Knowledge Discovery and Data Mining (KDD'96), Portland, Oregon, USA, Aug.*, 1996, pp. 226-231
[52] H. Zengyou, D. Shengchin, "Squeezer: An efficient algorithm for clustering categorical data," *J. Comp. Sci. Tech.*, vol. 17, 611-624, 2002.
[53] J. Zhao., J. Fang, Z. Ye, L. Zhang, "Large scale autonomous driving scenarios with self-supervised feature extraction", Arxiv
[54] BSI PAS 1883 2020: Operational Design Domain (ODD): taxonomy for automated driving systems (ADS). Specification, 2020, [Online]. https://www.bsigroup.com/en-GB/CAV/pas-1883/
[55] 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
[56] R. Agrawal, T. Imielinski, A. Swami, "Mining association rules between sets of items in large databases," *Proc. ACM SIGMOD*, pp. 207–216, 1993.
[57] F. G. Praticò, & M. Giunta, "Quantifying the effect of present, past and oncoming alignment on the operating speeds of a two-lane rural road," *Baltic J. of Road Brid. Eng*., vol. 7, 181–190. 2012. DOI:10.3846/bjrbe.2012.25
[58] A. Karimi, Ehsan K., "Investigating the effect of geometric parameters influencing safety promotion and accident reduction (Case study: Bojnurd-Golestan National Park road)," *Cogent Eng*, vol. 5, 1525813, DOI. 10.1080/23311916.2018.1525812
[59] S. Raschka, "MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack," *J. Open Sour. Soft.*, vol. 3, pp. 638, 2018, DOI. https://doi.org/10.21105/joss.00638
[60] P. A. Lopez, M. Behrisch; L. Bieker-Walz, J. Erdmann; Y. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel; P. Wagner, E. Wiessner, "Microscopic Traffic Simulation using SUMO," *2018 21st Int. Conf. Intel.Trans. Sys. (ITSC)*, 4-7 Nov, 2018
[61] Safety Pool™, online. https://www.safetypool.ai/
[62] K. Khan, S. Rehman, K. Aziz, S. Fong, S. Sarasvady, "DBSCAN, Past present and future," *The Fifth Inter. Conf. Appl. Digi. Info. and Web Techno.*, 1-19 Feb, 2014 DOI. 10.1109/ICADIWT.2014.6814687
[63] P. Andritsos, P. Tsaparas, R. J .Miller, K. C. Sevcik, LIMBO, "Scalable clustering of categorical data," *Adv. Data. Techno*., pp. 123-146, 2004

II. APPENDICES

A. PLOTS FOR ASSOCIATION RULES IN CLUSTERS

Below are the plots of association rules represented by arrows between variables along with their corresponding confidence values

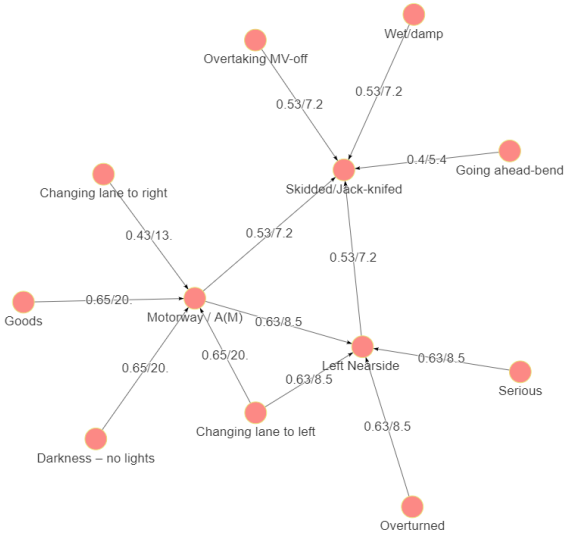

FIGURE 8. Association rules plot for Cluster 1

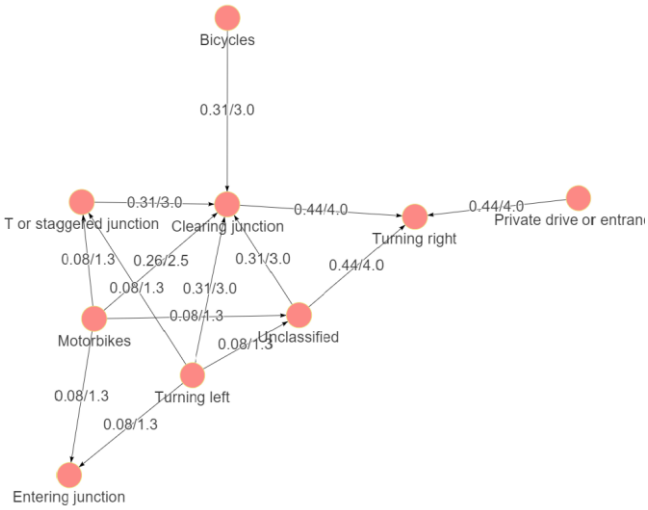

FIGURE 9. Association rules plot for Cluster 2

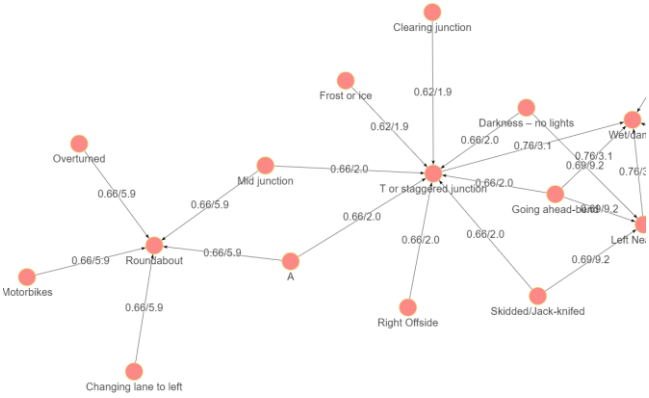

FIGURE 10. Association rules plot for Cluster 3

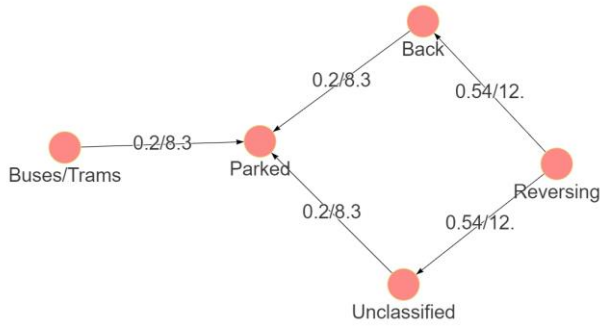

FIGURE 11. Association rules plot for Cluster 4

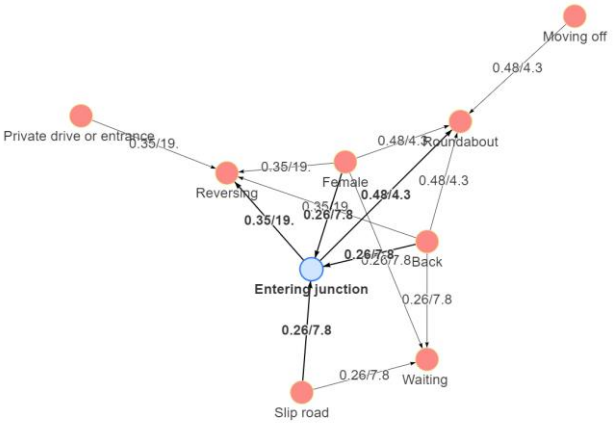

FIGURE 12. Association rules plot for Cluster 5

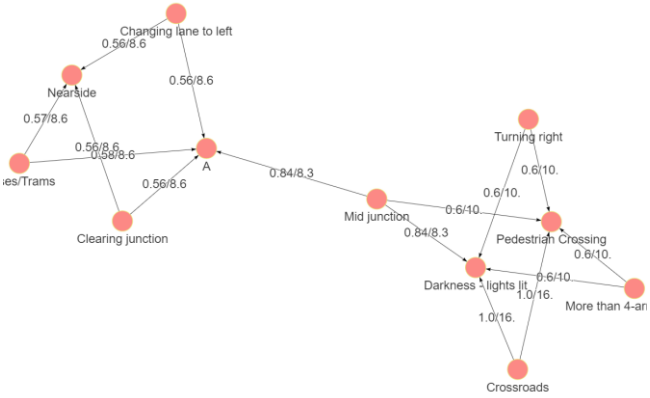

FIGURE 13. Association rules plot for Cluster 6

B. RESTRUCTURING OF STATS19 TRAFFIC VARIABLES

Here we provide an example of re-categorization of the data for the case of the traffic variable: *Vehicle types*. For the sake of simplicity of the analysis, the original categories (Table 3) of the raw data are restructured to give the new ones

TABLE 11.
ORIGINAL *VEHICLE TYPE* CATEGORIES OF STATS19

Code	Label
1	Pedal cycle
2	Motorcycle 50cc and under
3	Motorcycle 125cc and under
4	Motorcycle over 125cc and up to 500cc
5	Motorcycle over 500cc
8	Taxi/Private hire car
9	Car
10	Minibus (8 - 16 passenger seats)
11	Bus or coach (17 or more pass seats)
16	Ridden horse
17	Agricultural vehicle
18	Tram
19	Van / Goods 3.5 tonnes or under
20	Goods over 3.5t. and under 7.5t
21	Goods 7.5 tonnes and over
22	Mobility scooter
23	Electric motorcycle
90	Other vehicle
97	Motorcycle - unknown cc
98	Goods vehicle - unknown weight
-1	Data missing or out of range

TABLE 11.
RESTRUCTURED *VEHICLE TYPE* CATEGORIES

	Cleaned Vehicle Types
code	Label
1	Cars
2	Bikes
3	Buses/Trams
4	Horses/Tractors
5	Goods

EMRE ESENTURK completed his PhD at the University of Pittsburgh in mathematics.

Since his PhD, he has taken multiple research positions at the University of Warwick and University of Cambridge exploring areas of applied mathematics (PDEs, ODEs, integral equations) and applications of mathematics (materials science, atmospheric chemistry). In 2019 he joined Intelligent Vehicles Division at WMG, University of Warwick analyzing traffic accidents for safety analysis and scenario generation.

ALBERT G. WALLACE was born in Colchester, Essex, UK in 1998. He received a B.Sc. degree in Mathematics from the University of Warwick, Coventry, in 2019.

From 2019 to 2021, he was a Graduate Engineer with the Warwick Manufacturing Group Intelligent Vehicles Department. Since 2021, he has been a Research assistant with the Verification and Validation team at Warwick Manufacturing Group. His research interests include Connected Autonomous Vehicles, Unsupervised Machine learning and Regression.

SIDDARTHA KHASTGIR received his B.Tech + M.Tech, dual degree in mechanical engineering from Indian Institute of Technology (IIT) Kharagpur, India in 2011 and a Ph.D from University of Warwick, UK in 2019. He is the Head of Verification and Validation, Intelligent Vehicles at WMG, University of Warwick, UK. His research interests include trust in automation, system safety, and verification and validation of autonomous vehicles.

Dr Khastgir is a Chartered Engineer and a member of Institution of Mechanical Engineers (IMechE) UK. He has been named in the Forbes 30 Under 30 Europe 2018 – Industry list and was awarded the prestigious UKRI Future Leaders Fellowship in 2019.

PAUL JENINGS received the B.A. degree in physics from University of Oxford, Oxford, U.K., in 1985 and the Eng.D. degree from University of Warwick, Coventry, U.K., in 1996.

From 1985 to 1988 he was a Physicist at Rank Taylor Hobson. Since 1988 he has focused on industry-focused research for the Warwick Manufacturing Group (WMG), University of Warwick. His interests include connected and autonomous vehicles, testing, human factors, mobility and user engagement in product and environment design, focusing on automotive and healthcare applications. He is the Research Director at WMG, University of Warwick, UK.