EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

Comparative analysis of regional variations in road traffic accident patterns with association rule mining

Albe Bing Zhe Chai[1,*], Bee Theng Lau[1], Mark Kit Tsun Tee[1] and Christopher McCarthy[2]

[1]Swinburne University of Technology Sarawak Campus, Kuching, Sarawak, Malaysia
[2]Swinburne University of Technology, Melbourne, Victoria, Australia

Abstract

INTRODUCTION: Road Traffic Accident (RTA) patterns discovery is vital to formulate mitigation strategies based on the characteristics of RTAs.
OBJECTIVES: Numerous studies have utilised the Apriori algorithm for RTA pattern discovery. Hence, this study aims to explore the applicability of FP-Growth algorithm to discover and compare the RTA patterns across several regions.
METHODS: Orange data mining toolkit is used to discover RTA patterns from the open-access RTA datasets from Ethiopia (12,317 samples), Finland (371,213 samples), Germany (50,119 samples), New Zealand (776,878 samples), the UK (1,048,575 samples), and the US (173,829 samples).
RESULTS: There are similarities and differences in RTA patterns among the six regions. The five common factors contributing to RTAs are road characteristics, type of road users or objects involved, environment, driver's profile, and characteristics of RTA location. These findings could be beneficial for the authorities to formulate strategies to reduce the risk of RTAs.
CONCLUSION: Discovery of RTA patterns in different regions is beneficial and future work is essential to discover the RTA patterns from different perspectives such as seasonal or periodical variations of RTA patterns.

Keywords: road traffic accident, knowledge discovery, pattern analysis, data mining, association rule mining

Received on 23 March 2023, accepted on 26 November 2023, published on 28 November 2023

doi: 10.4108/eetpht.9.3173

1. Introduction

Road Traffic Accident (RTA) is a global challenge that demands extensive research efforts to develop effective solutions for mitigating the increasing trend of RTAs. According to [1], the number of fatalities due to RTAs was estimated at 1.3 million people annually, while more than half of the fatalities are among vulnerable road users (pedestrians, bicyclists, motorcyclists). Moreover, RTA has climbed from being the 9th to the 8th leading cause of injury-related deaths globally [2]. It is projected to become the 5th leading cause by 2030 when no action is taken to halt the increasing trend [3]. From the RTA fatality rate perspective, differences are observed between the high-, middle- and low-income countries. Specifically, the average RTA fatality rate in low-income countries is three times higher than the high-income countries, with 27.5 and 8.3 deaths per 100,000 population respectively [2]. Therefore, it is essential to examine the situation of RTAs in different countries or regions to understand the RTA patterns and the significant factors that contribute to the occurrence of RTA.

Data mining utilizes machine-learning techniques to discover hidden characteristics in large datasets, such as modelling equations, trends, patterns, and relationships [4]. It can be performed using classification, regression, clustering, summarization, dependency modelling, association rule discovery, outlier detection, and episode discovery or prediction. Data mining can be used to mine meaningful

*Corresponding author. Email: abzchai@swinburne.edu.my; 104136569@students.swinburne.edu.my

information from RTA-related images and videos to explore RTA patterns. Association rule mining (ARM) is one of the promising data mining techniques that performs unsupervised pattern discovery on datasets. It is also relevant for RTA patterns analysis because the generated association rules can help to discover the hidden behaviours and relationships between the RTA attributes or factors.

Thus, this study aims to investigate the application of ARM in analysing RTA patterns across various regions. The objectives of this work can be outlined as follows.

(i) Develop association rules using the Frequent Pattern (FP) Growth algorithm and open-access RTA datasets.
(ii) Analyse RTA patterns in different regions based on the generated association rules, assisted by statistical visualizations such as scatter plots, pivot tables, and matrix plots.
(iii) Identify the factors that contributed to RTA based on the patterns identified from the ARM.

In general, this work is organized to have the following flow: Section 2 discusses related work from other researchers on RTA patterns discovery, followed by Section 3, which presents the methodology applied in this work. Then, the findings are presented and discussed in Section 4. Eventually, this work is summarised through the conclusion presented in Section 5.

2. Related Work

ARM is one of the data mining techniques that can be applied to analyse the relationships between the attributes in datasets. A typical application for this technique is the market-basket analysis to identify customer purchasing behaviours [5]. It is also relevant for RTA patterns discovery, where the frequent patterns and association rules computed by the ARM algorithm can be used to identify the characteristics of RTAs and formulate the appropriate countermeasures to prevent or mitigate future RTAs. For instance, [6] developed a framework that adopted the ARM algorithm to interpret the characteristics of RTA with different severity levels. The RTA dataset from Mujjafarnagar district, Uttar Pradesh, India, was clustered into three severity categories (fatal, major injury, and minor/no injury), and the Apriori algorithm in the Weka tool was applied to extract the association rules for each severity category. Another study conducted in India has developed a system to analyse the patterns and characteristics of RTAs with ARM [7]. In this study, various ARM algorithms (Apriori, Apriori TID, and Eclat) are investigated for their performance in generating association rules. Eventually, the Eclat algorithm was the most efficient among the three algorithms that generated the rules with the least amount of processing time. In addition, [8] proposed a framework to analyse the RTA dataset from the UK (2009 – 2016) through random clustering, decision tree, and ARM. Random clustering from Rapid Miner and the decision tree were implemented to cluster and classify the data based on three attributes (vehicles, weather, and lighting). Then, the Apriori algorithm was applied to generate the association rules used to discover the relationships between the dataset attributes. [9] also utilised the same UK traffic accident dataset (2005–2017) to investigate the characteristics and patterns of RTAs using the Apriori algorithm. The generated association rules are filtered with high lift and support and explored through visualizations such as matrix, scatter, and network plots.

Another work was performed by [10] that investigated the significant contributing factors towards RTA risk and severity level for the case of Elabuga town using the data between 2017 and 2018. The attributes were divided into accident characteristics, driver information, vehicle conditions, road conditions, and environmental conditions. The generated association rules were used to identify the combination of common factors that contributed to severe injuries or fatalities in RTAs at Elabuga town.

The RTA pattern analysis has also been done to investigate a specific category of crashes, such as fatal RTAs [11], [12]. Li, Shrestha, and Hu focused on the fatal RTAs in the US to compare the association rules generated by the Apriori algorithm with the output of the Naïve Bayes classifier and K-Means clustering [11]. On the other hand, [12] discovered the contributing factors of multi-fatality crashes in China using ARM and rules graph structures. The rules were evaluated with various visualizations, such as the directed rule network structures, matrix plots, scatter plots, and Venn diagrams. For the analysis of crashes with serious casualties in China, [13] evaluated the contributing factors and their relationships based on the association rules computed by the Apriori algorithm in R software. The dataset contained driver characteristics, vehicle conditions, road conditions, and environmental conditions.

Vehicle-pedestrian crashes are also an important category to analyse. [14] used the association rules to discover potential improvements or countermeasures focused on optimizing pedestrian safety. The Apriori algorithm was used to discover the patterns of vehicle-pedestrian crashes in Louisiana, and the significant patterns were analysed to formulate relevant strategies to raise awareness and mitigate future vehicle-pedestrian crashes. [15] also investigated the characteristics of pedestrian-involved crashes in Louisiana with RTA data between 2010 and 2019. The analysis focused on the RTA characteristics under different lighting conditions in which the Random Forest algorithm selected the significant variables, and the Apriori algorithm generated the association rules for the lighting conditions of daylight, dark with a streetlight, and dark without a streetlight. In Chennai, an urban Indian metropolitan, [16] examined the characteristics of vehicle-pedestrian crashes using the data obtained from the RADMS (road safety accident reporting database) for association rules generation. The association rules were classified into the fatal/grievous injury and simple injury/property damage only categories to analyse the RTA characteristics.

Another perspective is to analyse zone-specific RTA patterns. Weng et al. investigated the characteristics and contributing factors of work zone crash casualties at a work zone in Michigan with association rules [17]. A similar study

was found for the case in Nagpur city, Maharashtra, India, with the analysis of RTA characteristics performed on three different zones (high, medium, and low) in terms of the RTA occurrence risk [18]. The Apriori algorithm was used for each zone to identify the hidden RTA characteristics. Besides, driving behaviour is another crucial aspect that can contribute to RTAs when unsafe behaviours occur. [19] conducted a survey and collected 306 questionnaire data containing the driving behaviours of the respondents in Kuwait. The Apriori algorithm was applied to discover the hidden association rules that relate the driving behaviours with the RTAs in Kuwait.

The ARM technique can also be integrated into the prediction of RTAs, where there was a system proposed to determine the frequent patterns of RTAs as the references for the government or NGOs to formulate countermeasures and improve the road safety level at the major accident zones [20]. The frequent patterns of the previous RTAs were identified using the Apriori algorithm, while the Naïve Bayes classifier predicted the type of accident based on the generated association rules. Furthermore, [21] used the K-Means algorithm to categorize the RTA dataset for the Coimbatore city in India into three clusters based on the accident locations. The Apriori algorithm was applied to the three clusters to obtain the association rules and supplied to the Naïve Bayes algorithm for accident severity classification. A predictive model based on fuzzy logic was developed to predict the probability of an accident occurrence. This study was further improved by [22], which proposed a pattern-mining predictor system. The system continued the previous study by implementing an improved Apriori algorithm, which was more efficient in terms of processing time. In addition, further improvement was made with the proposal of road accident pattern (RAP) miner implementing the Scan Efficient Apriori (SEA) algorithm that performs a faster generation of the association rules [23].

A review of the state-of-the-art application of ARM showed that the amount of data available is a challenge to discovering the pattern and relationships between the attributes of RTAs. It was suggested that future research is necessary to include more data or useful attributes in mining the association rules for a more comprehensive interpretation of RTA patterns and relationships with contributing factors [8], [9], [11], [14], [18], [19]. In addition, the reviewed studies mostly applied the Apriori algorithm to generate the association rules. Thus, it is essential to discover the applicability of another algorithm, such as the Frequent Pattern (FP) Growth algorithm, to perform the ARM. The FP Growth algorithm was an alternative to the Apriori algorithm that uses the concept of tree construction (FP tree) to compute the association rules. It was discovered to produce the same set of association rules as the Apriori algorithm for the Groceries dataset [24]. Furthermore, [25] discovered that the FP Growth algorithm was more effective than the Apriori algorithm in terms of the time taken to generate the frequent patterns large dataset. Therefore, it is potential to investigate the generation of association rules for the RTA datasets using the FP Growth algorithm.

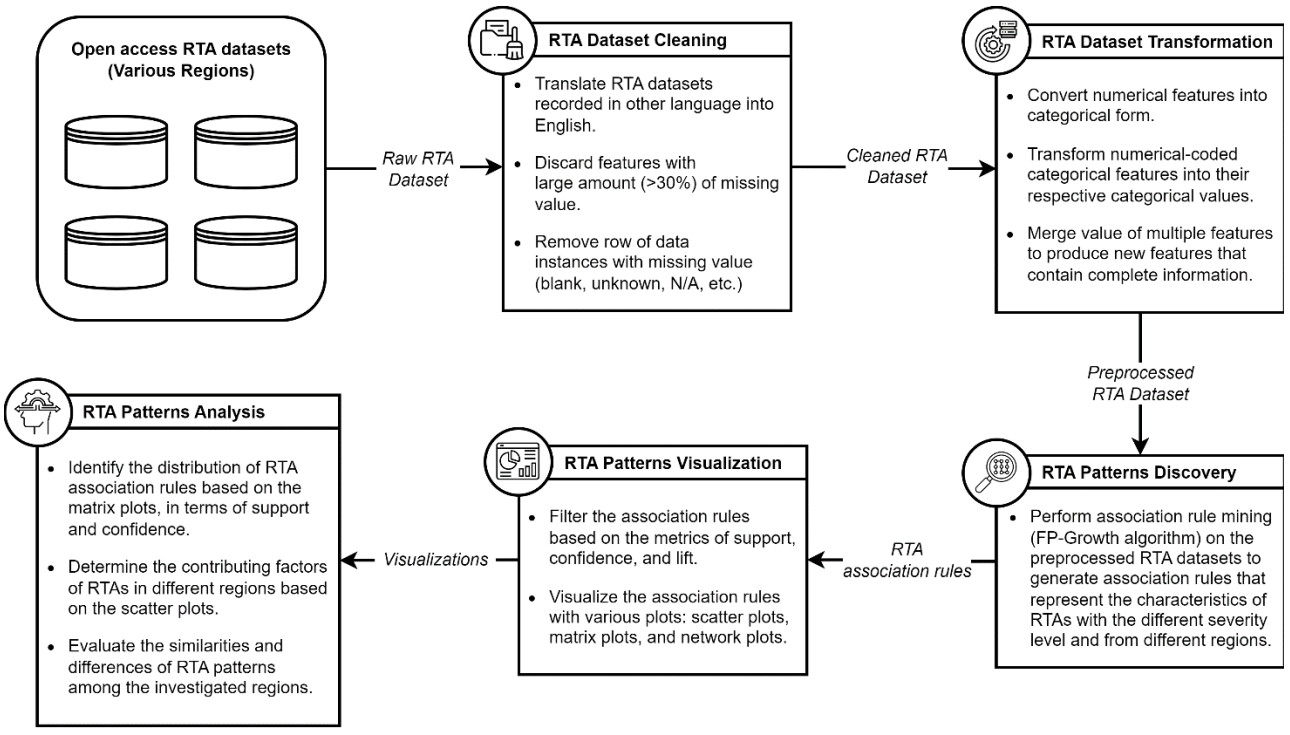

Figure 1. Workflow of RTA pattern analysis

Table 1. General information on the collected RTA dataset from different regions: Ethiopia, Finland, Germany, New Zealand, the UK, and the US.

No.	Source	Region	Duration	Sample Size
1	Bedane 2020	Ethiopia (Addis Ababa city)	2017 to 2020	12,317
2	National Open Data Portal Finland	Finland	2005 to 2021	371,213
3	Berlin Open Data Portal	Germany (Berlin city-state)	2018 to 2021	50,119
4	Waka Kotahi NZ Transport Agency Open Data Portal	New Zealand	2000 to 2021	776,878
5	UK Open Data Portal	UK	2010 to 2020	1,048,575
6	Fatality Analysis Reporting System (FARS)	US	2016 to 2020	173,829

3. Methodology

This study aims to collect RTA datasets from various regions through online open-access databases and apply ARM to analyse RTA patterns across the regions. The RTA pattern analysis conducted in this study can be summarized into five key stages: (1) RTA dataset cleaning, (2) RTA dataset transformation, (3) RTA patterns discovery, (4) RTA patterns visualizations, and (5) RTA pattern analysis. The flow of activities between the stages is illustrated in Figure 1. For this study, the investigation is conducted with the support of Python and the Orange software – an open-source toolkit for data visualization, machine learning, and data mining.

3.1 Data Collection

The initial step in this study involves data collection to acquire RTA datasets from different regions around the world. Due to the challenges associated with collecting RTA datasets physically, online sources are used for data collection. These collected RTA datasets comprise historical RTA records published on open data portals or obtained from published studies by other researchers. Eventually, a total of six RTA datasets were acquired (Table 1), which represent the regions of Ethiopia (Addis Ababa city), Finland, Germany (Berlin city-state), New Zealand, the United Kingdom (UK), and the United States (US). Additionally, these datasets contain records of RTAs that occurred between 2000 and 2021, spanning approximately 20 years. As for the sample size, the UK dataset has the largest number of instances, with 1,048,575 samples, followed by New Zealand with 776,878 samples, Finland with 317,213 samples, the US with 173,829 samples, Germany with 50,119 samples, and Ethiopia with 12,317 samples. The diversity in collected RTA datasets is believed to offer valuable insights for discovering RTA patterns.

3.2 Data Pre-processing

Before generating association rules for RTA pattern discovery, all collected RTA datasets are brought into the pre-processing stage so that the raw data is prepared into a suitable format for ARM. In this study, the raw RTA datasets are modified through data cleaning and transformation techniques.

The data cleaning technique involves filtering the raw RTA datasets to remove the columns with over 30% missing values. Moreover, filtering is applied to each of the instances (rows) for the datasets to eliminate rows with missing values. Data translation is another task performed to translate RTA datasets recorded in languages such as Finnish and German into English for easier interpretation.

Further pre-processing of the cleaned RTA datasets is performed with the data transformation technique. This includes converting numerical RTA features, such as speed limit, number of involved vehicles, and accident time into categorical form. Furthermore, the numerical-coded categorical features are transformed by converting their numerical codes into their respective categorical values (strings). The data transformation technique is also necessary to merge multiple features into a more comprehensive feature. For instance, merging the ‘weatherA’ and ‘weatherB’ columns helps to provide a complete representation of the weather conditions in the New Zealand dataset. Eventually, the data pre-processing stage produces clean and fully pre-processed RTA datasets for the six regions. These datasets are then forwarded to the ARM stage to compute the association rules for RTA pattern discovery across the six regions.

3.3 Association Rule Mining

The review in this paper reveals that previous investigations commonly employed the Apriori algorithm to discover RTA patterns. Therefore, this study aims to explore the generation

of association rules with the FP Growth algorithm, which is available in the Orange data mining toolkit. The FP Growth algorithm is associated with the concept of tree construction, where frequent itemsets are identified using a tree data structure known as the FP-tree [24]. For this study, all the data is divided into different clusters based on the severity level of the RTAs, and ARM is applied to each cluster to produce the association rules to investigate the RTA patterns under different severity levels. After association rules are computed with the FP Growth algorithm, their quality is to be assessed using three association rule metrics: support, confidence, and lift values. The equations for support, confidence, and lift are adapted based on the work from [14].

3.3.1 Support

For an association rule expressed as A → B, the itemset in the LHS, A, is referred to as the antecedent, while the itemset in the RHS, B, is called the consequent. The support of this rule can be measured as the proportion of instances in the dataset that satisfied the rules. Alternatively, support represents the probability for both the antecedent and consequent to co-exist in a dataset instance.

$$Support(A \rightarrow B) = P(A \cap B) = \frac{\#(A \cap B)}{N} \quad (1)$$

3.3.2 Confidence

The confidence of the association rule A → B reflects the accuracy of the rule. It describes the probability of the consequent being present in a data instance when the antecedent is present. A higher confidence value indicates that the consequent has a higher probability of appearing together with the antecedent. Thus, this refers to a more meaningful relationship because the rule is highly confident to represent patterns among the investigated data samples.

$$Confidence(A \rightarrow B) = \frac{Support(A \rightarrow B)}{Support(A)} = \frac{P(A \cap B)}{P(A)} = \frac{\#(A \cap B)}{\#A} \quad (2)$$

3.3.3 Lift

Lift for an association rule is defined as the ratio between the rule's confidence and expected confidence. The expected confidence refers to the probability of the consequence; hence, the lift of a rule is represented with equation (3). Moreover, the lift value of an association rule ranges from 0 to ∞ and helps to describe the relationship between the antecedent and consequent. When the value of lift is greater than 1, the rule is concluded to have a positive relationship between the antecedent and consequent. Conversely, a lift value of smaller than 1 indicates a negative relationship. The association rule is categorized as independent for a lift value equal to 1.

$$Lift(A \rightarrow B) = \frac{Confidence(A \rightarrow B)}{P(B)} = \frac{P(A \cap B)}{P(A) \times P(B)} \quad (3)$$

3.4 Road Traffic Accident Patterns Discovery and Visualisation

The final stage in this study is to perform the pattern analysis based on the association rules computed for each cluster of severity level. For the association rules, filtering is applied to select the rules that contain the severity level in either the antecedent or consequent column. Moreover, filtering association rules based on support, confidence, and lift values is essential to ensure that only relevant rules are selected for the analysis of RTA patterns. According to [12], threshold values for support, confidence, and lift are often subjectively chosen, typically between the range of 1 - 20% (support), 30 - 80% (confidence), and greater than 1.20 (lift). Following these guidelines, the association rules are filtered to have minimum support and confidence at 20% and 70%, respectively. For the lift value, the minimum threshold is set to 1.20. Nevertheless, there are cases where none of the rules satisfy the minimum lift threshold of 1.20. Consequently, further adjustments are made to gradually decrease the lift threshold until interesting associations are identified. This adjustment is also performed with the lift value always greater than 1.00 to ensure the positive correlation of the rules.

With the filtered association rules, the analysis is made with visualizations such as scatter plots, pivot tables, and matrix plots to illustrate the patterns of the resulting rules. The scatter plots are used to evaluate the distribution of the RTA association rules in terms of support and confidence. Apart from that, an association rule may be a subset of another rule and thus, rule grouping is performed to remove these subsets from the resulting rules. The top five rule clusters are then selected as the frequent patterns of RTAs and presented through pivot tables. Furthermore, the factors contributing to RTAs for each region are determined with the matrix plots that illustrate the percentage of the RTA features that existed in all the rules. The observations made from the scatter plots, pivot tables, and matrix plots are evaluated to highlight the similarities and differences in RTA patterns among the investigated regions.

4. Findings and Discussions

This work has used the open-access RTA datasets obtained from the open data portal in six regions: Ethiopia, Finland, Germany, New Zealand, the UK, and the US. The raw datasets collected are pre-processed and used to generate the association rules with the FP-Growth algorithm. The generated association rules are filtered based on the minimum metrics of support (20%), confidence (70%), and lift (1.2). Adjustments on the minimum value of lift are performed for some cases where no association rules are obtained with the initial minimum lift (1.2). With the resultant rules, investigations are conducted to analyse the general statistics of the rules, identify the RTA patterns in the six regions under different severity levels and determine the significant factors contributing to RTAs.

4.1 Generated Association Rules

This work has used the Orange data mining toolkit to perform the ARM. The minimum support and confidence values are set in the software and the maximum number of rules that can be generated per execution is capped at 100,000 rules. Filtering is also performed to select the rules that satisfy the minimum lift value applied to ensure the positive relationship represented by the rules. However, some adjustments are made for Germany (Minor severity), New Zealand (all severity), and the US (Fatal severity) due to the limited number of rules (or none) that can satisfy the minimum lift value of 1.2.

With the resultant association rules (Table 2), scatter plots have been produced to analyse the distribution of the rules in terms of support and lift values. The scatter plots are used to evaluate the characteristics of the rules among the different RTA severity levels in each region. In the case of the Ethiopia dataset, the association rules are distributed within the range of 0.20 and 0.40 in terms of support values. The lift values for the rules related to fatal RTAs are observed to concentrate between 1.20 to 1.40, while the rules for serious and slight injuries are concentrated between the lift values of 1.20 and 1.25.

For the Finland dataset, the association rules generated for all the severity levels are found to majorly concentrate between the support of 0.20 to 0.30. In addition, the lift value of 1.35 can be the middle point among the association rules of all severity levels in Finland, where the percentage of association rules with lift values above 1.35 is 55.9% (fatal), 57.1% (injury), and 53.7% (non-injury).

The resultant rules for Germany show that there are differences among the rules of all severity levels. The scatter plots illustrated that 60.4% from the fatal category have lift values of between 2.00 and 3.00, 4.2% from the serious category have lift values at approximately 4.50 (others at below 1.50) and 8.5% from the slight category have lift values around 1.22 (others at below 1.20).

Table 2. The minimum metric values applied for the generated association rules and the number of remaining rules after filtration.

Dataset	RTA Severity	Minimum Metric Value Applied			Number of Rules	
		Support	Confidence	Lift	Generated	Filtered
Ethiopia	Fatal	0.20	0.70	1.20	100,000	7,617
	Serious	0.20	0.70	1.20	100,000	491
	Slight	0.20	0.70	1.20	100,000	796
Finland	Fatal	0.20	0.70	1.20	3,902	143
	Injury	0.20	0.70	1.20	2,550	133
	Non-Injury	0.20	0.70	1.20	2,123	41
Germany	Fatal	0.20	0.70	1.20	2,589	439
	Serious	0.20	0.70	1.20	4,888	191
	Minor	0.20	0.70	1.15	7,121	177
New Zealand	Fatal	0.20	0.70	1.10	100,000	2,918
	Serious	0.20	0.70	1.05	100,000	2,928
	Minor	0.20	0.70	1.05	100,000	134
	Non-Injury	0.20	0.70	1.03	100,000	1,410
UK	Fatal	0.20	0.70	1.20	100,000	1,286
	Serious	0.20	0.70	1.20	100,000	1,256
	Slight	0.20	0.70	1.20	100,000	1,597
US	Fatal	0.20	0.70	1.13	100,000	3,584

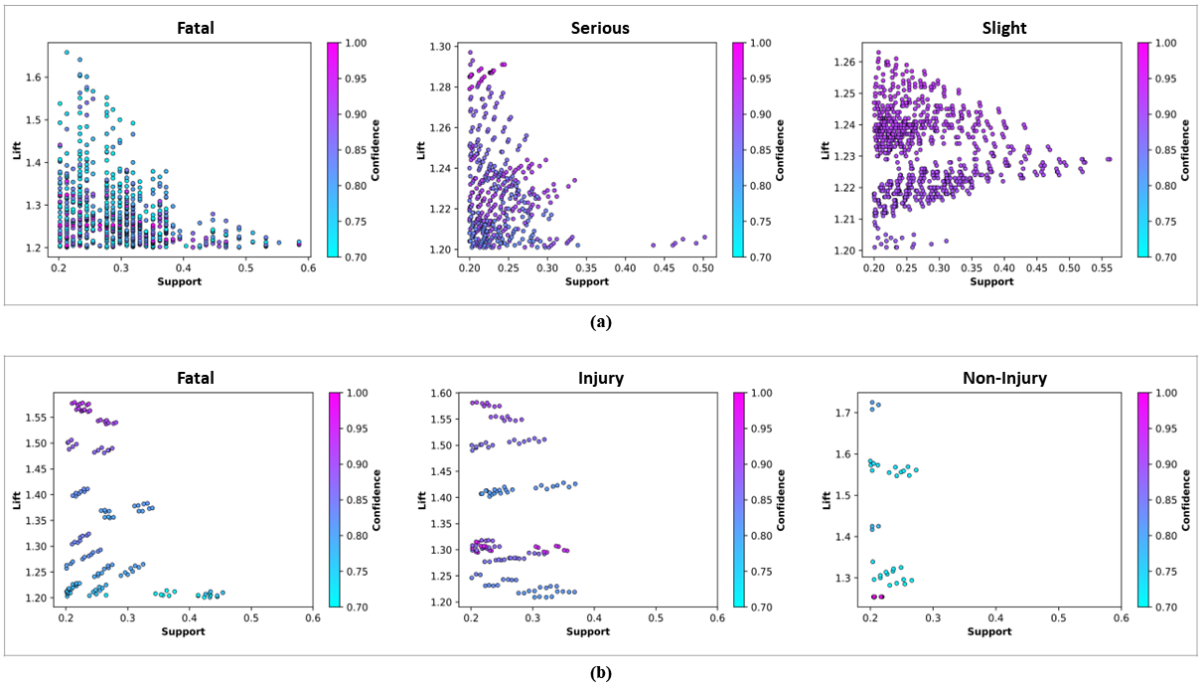

Figure 2. Distribution of the association rules generated for each severity level in different regions based on the support and lift values: **(a)** Ethiopia and **(b)** Finland.

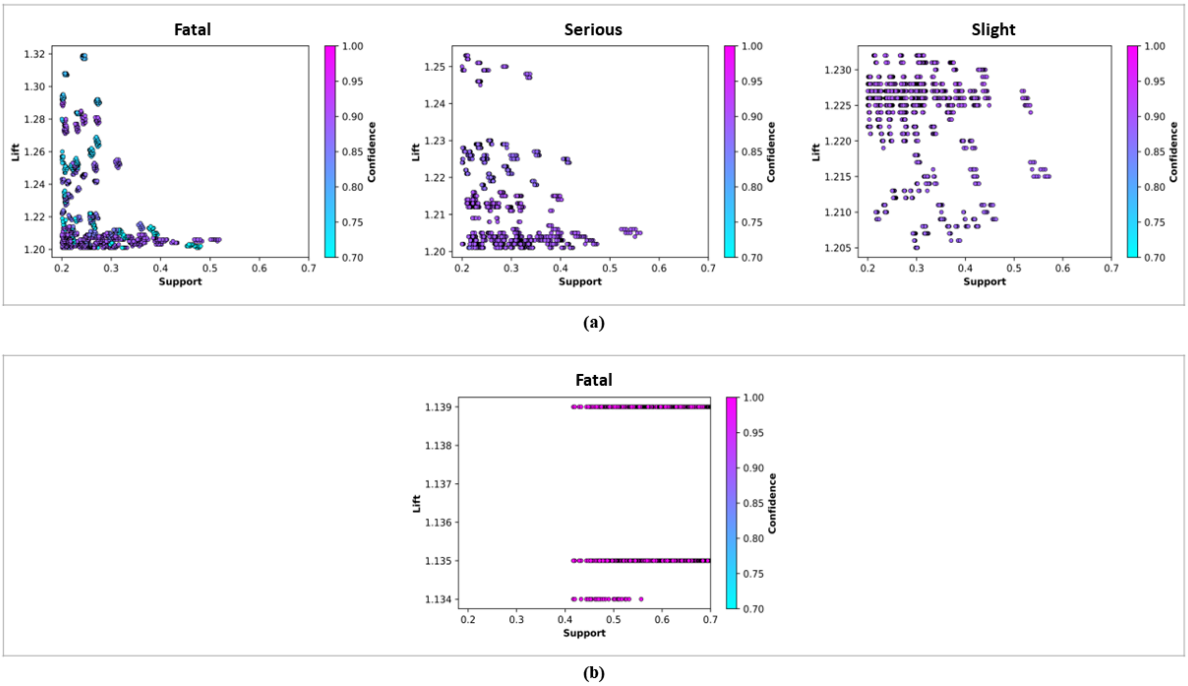

Figure 3. Distribution of the association rules generated for each severity level in different regions based on the support and lift values: **(a)** the UK and **(b)** the US.

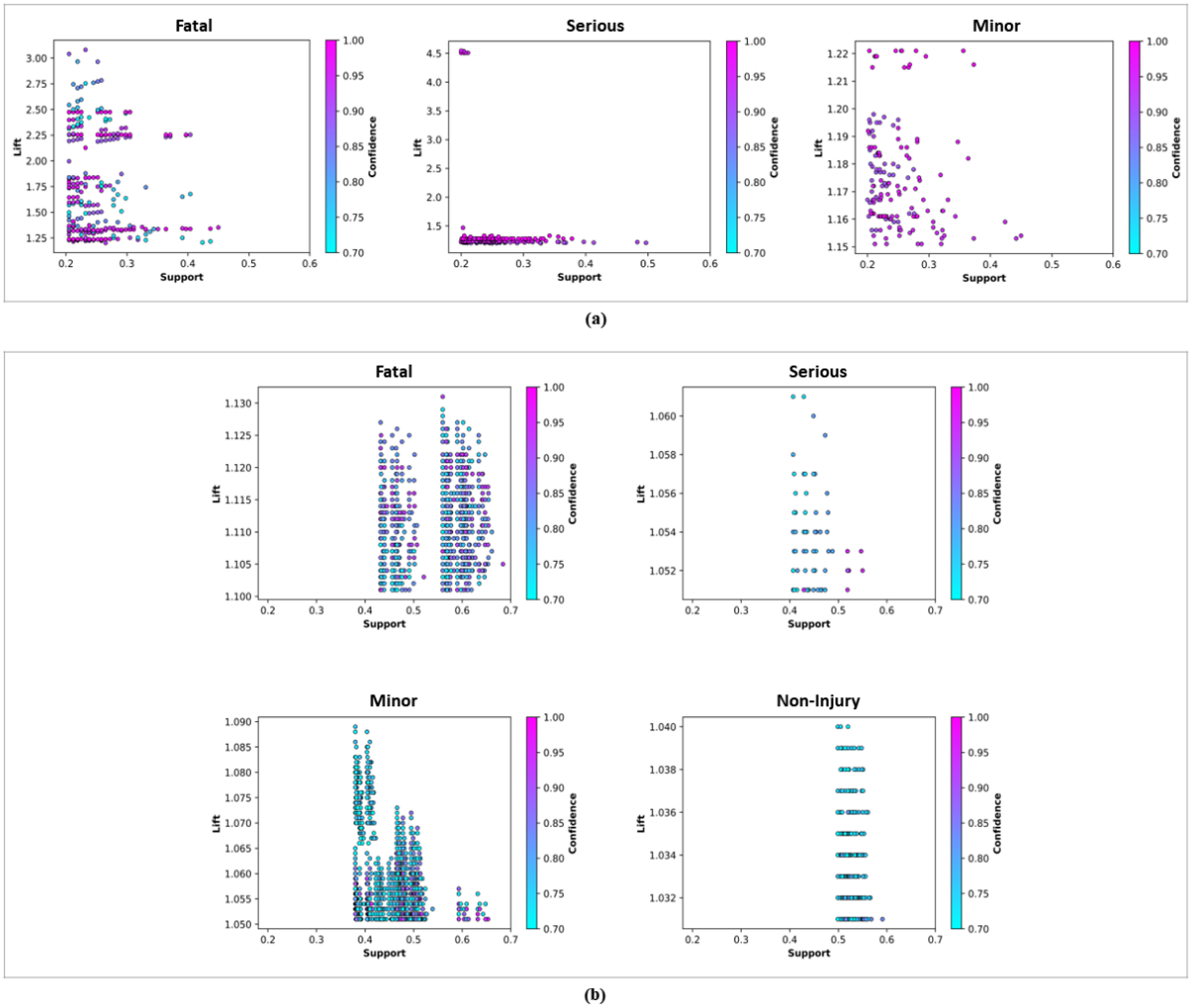

Figure 4. Distribution of the association rules generated for each severity level in different regions based on the support and lift values: **(a)** Germany and **(b)** New Zealand.

An interesting distribution of the association rules is observed for the New Zealand region, where all the rules are concentrated around higher support values (0.40 to 0.70) as compared to the other regions. However, the range of lift values is between 1.03 and 1.13, which is relatively low compared to the association rules from other regions.

For the US region, the dataset consisted of only the fatal RTAs and thus, analysis is performed only for the fatal category. The scatter plot produced for the rules generated for the US region observed that the rules can be grouped based on three lift values at 1.139, 1.135, and 1.134, respectively. This indicated that there is the possibility that the rules can be grouped into three major clusters and each of the rules is a subset of the clusters. The overall analysis of the association rules generated for each severity level in the six regions showed that there are differences between the rules in terms of support and lift values. All the rules from Ethiopia, Finland, Germany, and the UK regions are found to majorly distribute at lower support values ranging from 0.20 to 0.40. Conversely, the rules obtained for the New Zealand and the US regions are observed to scatter at a higher range of support between 0.40 to 0.70. This indicates that the generated rules are found to match 40% to 70% of the data instances in the datasets for New Zealand and the US regions. From the perspective of lift values, most of the rules are discovered to distribute between 1.20 and 2.00. However, there are exceptions for some cases where the lift values are not within the range identified. This can be seen from the RTAs with serious injuries in the Germany region with a higher lift value of around 4.50, which indicates a stronger relationship between the antecedent and consequent. Conversely, the lift values for the rules in New Zealand and the US regions are relatively low as compared to the others and this demonstrates weaker relationships between the itemset in the rules. The weaker relationships can be referred to as fewer data samples are expected to satisfy the rules when more data samples are collected in the future.

4.2 RTA Patterns in Different Regions

The analysis of RTA patterns is essential to explore the patterns of RTAs in different regions so that appropriate prevention or mitigation plans can be developed. This work has discovered the RTA patterns in the six regions based on the association rules generated with the open-access datasets. Findings from Figure 2, Figure 3, and Figure 4 revealed the possibilities of grouping the generated rules in which each of the rules can be a subset of another. Therefore, this work has grouped the rules and selected the top groups in terms of the lift value for comparisons (see Appendix A, Table 4 to Table 9).

In the case of Ethiopia, the fatal RTAs are discovered to involve drivers who mainly completed up to junior high school level. Moreover, asphalt roads are the main road type for RTAs to result in fatalities and this mostly involves the collision between 2 vehicles. A difference between the RTAs with fatalities and serious injuries can be the location, where fatal RTAs mostly occur on tangent roads while RTAs with serious injuries are found frequently at the Y-shape junction. The RTAs with serious and slight injuries are observed to be similar in that both mainly happened during the weekdays and daytime.

The patterns discovered for all the fatal, injury, and non-injury RTAs in Finland reflected the same characteristics, occurring on a dry road with a durable coating and mainly during the daytime with normal weather conditions. The association rules generated with the investigated RTA features do not show a differentiable pattern between the different RTA severity levels. However, this can indicate that RTAs in Finland mostly occur during the daytime with normal weather regardless of the severity levels. This can be a factor that increases human mobility and results in a higher RTA occurrence risk.

The computed association rules based on the dataset from Germany showed that RTA analysis can be performed based on the type of road users involved because the results confirmed that this can contribute to the severity level of RTAs. For instance, RTAs with fatalities and serious injuries in Germany are often found to involve vulnerable users (bicycle and pedestrian) and heavy vehicles. Conversely, minor injuries are often found in the RTAs that involve the collision between cars, without the presence of vulnerable users and heavy vehicles. An explanation for this finding is vulnerable users (pedestrians, bicyclists, and motorcyclists) do not have a protective shield when travelling on the roads, which leads to a higher risk of injuries in the RTAs. In addition, RTAs involving heavy vehicles are expected to create a larger impact on other vehicles during collisions and increase the risk of injuries. A common pattern among all the severity levels for RTAs in Germany is the higher occurrence during weekdays with daylight.

For New Zealand, the RTA patterns exhibit similar characteristics as those in Finland. There is no significant factor that can clearly distinguish between the RTA severity level of fatal, serious injuries, minor injuries, and non-injury. The RTAs in New Zealand are observed to majorly occur on the normal, 2-way sealed road and mainly between 4-wheeled vehicles regardless of the severity level. In addition, most of the RTAs are found to happen under fine weather conditions. These common factors are useful to describe the frequent patterns among the RTAs in New Zealand.

The RTAs in the UK region are found to have common factors among all the severity levels. For instance, all the RTAs mainly happen on give ways or uncontrolled roads which do not have authorized persons or physical facilities to control the pedestrian crossing. Furthermore, the results revealed that urban areas and single carriageways are also the hotspot locations of RTAs. An interesting pattern pointed out that the T or staggered junctions are the hotspots for fatal RTAs. This type of junction requires more attention from the drivers when passing the junction. RTAs in the UK region tend to result in slight injuries when involving two vehicles, and the speed limit at the site is 30 km/h. This showed that having a reasonable speed limit along the road could reduce the impact of collisions and hence, reduce the RTA severity.

For the US region, the investigated RTA dataset consists of only records for fatal RTAs and thus, the evaluation is only made on the fatal RTAs. The resultant association rules showed that fatal RTAs in the region are often not related to factors such as the involvement of young drivers, pedal cyclists, large trucks, and police pursuits. Moreover, drivers' behaviours such as distraction and drowsiness are also not the major factors contributing to fatal RTAs. The discovered patterns for fatal RTAs in the US region indicated that no significant factors were observed among all the investigated RTA features. Further exploration is needed by taking into consideration other factors such as light conditions, junction type, junction control, and road surface conditions.

4.3 Factors Contributing to RTA

The resultant association rules generated through this work can be used to analyse the factors contributing to RTAs in different regions. This is performed by evaluating the rules to determine the frequency of an RTA feature detected in the filtered rules. The matrix plots in Figure 5 and Figure 6 illustrate the frequency of different RTA features detected from the association rules among the six regions. With the matrix plots, the average percentage that each feature is present in an association rule has been calculated. The top factors contributing to RTA are listed in Table 3 as the overall RTA patterns in the six regions.

Based on the analysis of the RTA features, the results showed that Ethiopia, Finland, and the UK regions have similar factors contributing to their RTAs, which are the road characteristics (road geometrics and conditions), and the environmental factors (weather and light conditions). Furthermore, there are also interesting RTA features shown in some regions. The findings revealed that all the top factors contributing to RTA in Germany are from the category of involved road users and objects. This further confirmed the earlier findings on RTA patterns in the region, where the type of road users or objects involved majorly contributed to the RTA severity levels. For the New Zealand region, the top factors contributing to RTAs are found to be from the

category of road geometrics and conditions. This is consistent with the earlier discussion on the RTA patterns in New Zealand in which most of the RTAs occur on the normal, 2-way sealed roads. On the other hand, the top factors contributing to RTA in the US region are found to be mainly related to the category of drivers' profile (three out of the five top factors).

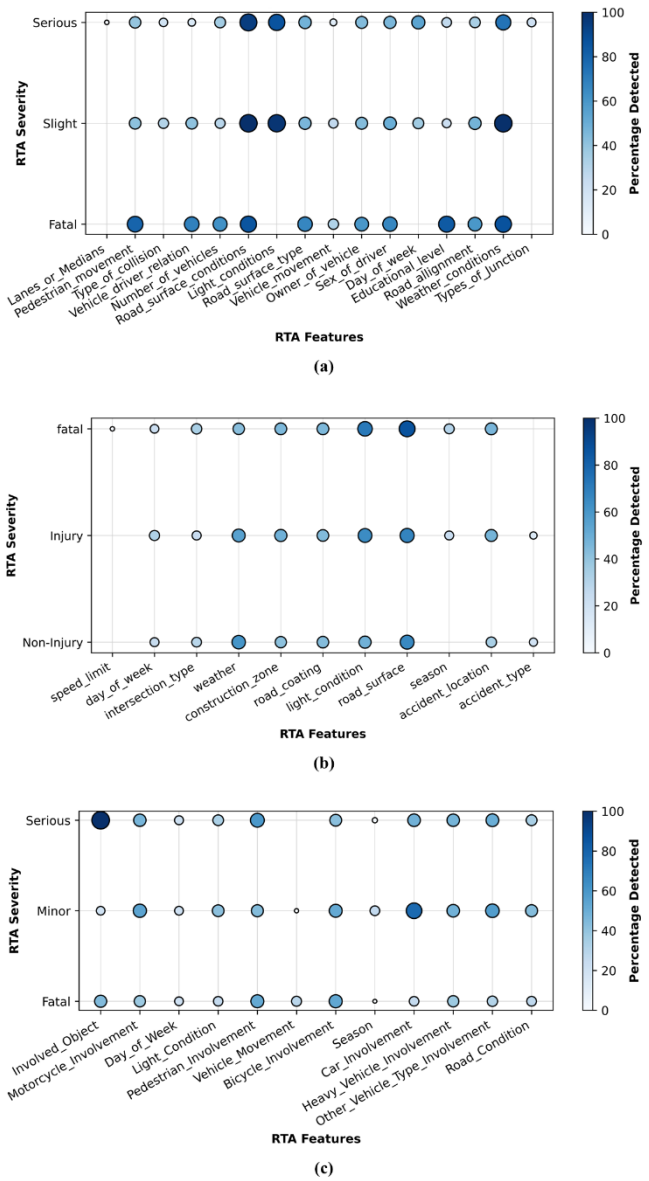

Figure 5. Percentage of RTA features to present in the RTA association rules for different regions: (a) Ethiopia, (b) Finland and (c) Germany.

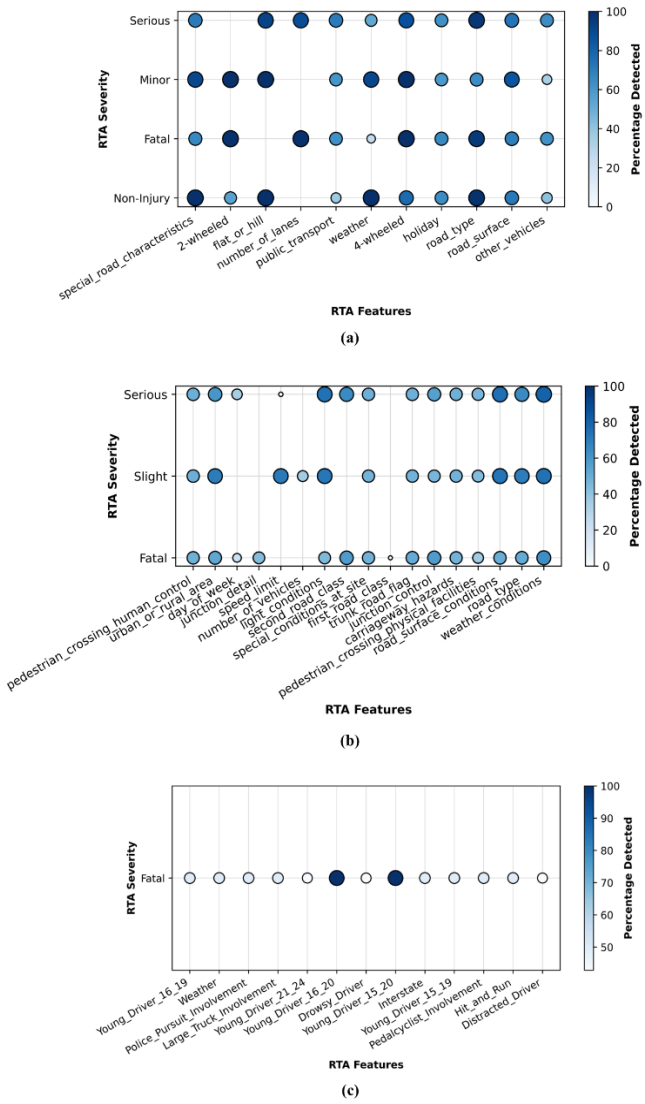

Figure 6. Percentage of RTA features to present in the RTA association rules for different regions: (a) New Zealand, (b) UK, and (c) US.

Comparisons among the six regions revealed that road characteristics and environmental factors are the common factors that can affect the driving quality of drivers. With this, the responsible authorities are recommended to include the design of road geometrics with all-around considerations, the usage of anti-skid materials for road surfaces to reduce the risk during severe weather, and the frequent maintenance of streetlights to always ensure their functionality. Other than the common factors contributing to RTA, there are also some factors such as the involvement of road users and driver's profile, which have shown different characteristics among the regions. The drivers' profile can be a subjective factor because drivers might show different behaviours in every driving trip, while the presence of road safety education in a region can affect the driving behaviours of a driver.

Table 3. Top factors contributing to RTAs for Ethiopia, Finland, Germany, New Zealand, the UK, and the US regions.

Region		Rank				
		1	2	3	4	5
Ethiopia	Feature	Road surface conditions	Weather conditions	Light conditions	Driver's gender	Road surface type
	Percentage (%)	93.43	87.00	61.40	54.36	53.84
Finland	Feature	Road surface	Light condition	Weather	Construction zone	Road coating
	Percentage (%)	73.41	61.82	53.09	45.26	44.57
Germany	Feature	Accident type involved object	Pedestrian involvement	Car involvement	Bicycle involvement	Motorcycle involvement
	Percentage (%)	55.00	51.97	50.61	49.16	46.65
New Zealand	Feature	4-wheeled	Road type	Special road characteristics	Road surface	Flat or hill
	Percentage (%)	91.29	89.00	82.18	75.51	73.23
UK	Feature	Weather conditions	Road surface conditions	Light conditions	Road type	Urban or rural area
	Percentage (%)	73.11	66.59	64.32	62.49	61.81
US	Feature	Young driver (15-20)	Young driver (16-20)	Hit and run	Interstate	Large truck involvement
	Percentage (%)	100.00	100.00	50.00	50.00	50.00

Overall, observation from Figure 7 showed that the RTA features from the category of road characteristics (30%) are the most significant factors among the six regions in terms of the average percentage detected from the association rules, followed by the type of road users and objects involved (23%), environmental factors (20%), drivers' profile (14%), and the characteristics of RTA location (13%). These RTA factors have the potential to be selected as the primary features for further research in the field of RTA analysis and prediction.

4.4 Limitations and Future Works

This study has successfully analysed the patterns and key factors contributing to RTA in various regions. However, some limitations have been identified to improve in the future research. Firstly, one common issue observed among the six RTA datasets investigated is the variation in RTA features recorded in these datasets. For example, the RTA features in the US dataset primarily related to driver behaviour, which is not widely recorded in other datasets. This introduces challenges for making in-depth comparisons among the six regions. Therefore, future research should consider the possibility of standardising the features for better comparison of RTA patterns across the regions. Moreover, it would be beneficial for future research to include more features for a more comprehensive analysis of RTA patterns and key factors.

Apart from that, this study conducted limited investigations into RTA patterns in developing countries, with Ethiopia being the only developing country investigated due to the limited availability of open-access RTA datasets from these countries. For a more inclusive pattern discovery, future research should collect more RTA datasets from developing countries. Furthermore, RTA pattern discovery is not limited solely to the severity level perspective but can be expanded to other dimensions. For instance, exploring seasonal variations in RTA patterns could help in formulating effective strategies to reduce RTA occurrences during different seasons. Finally, another potential area for future investigation is the integration of pattern analysis to support various RTA predictions such as the occurrence risk and severity level. This can be achieved by incorporating RTA

patterns analysis for feature selection or validation of the predicted output.

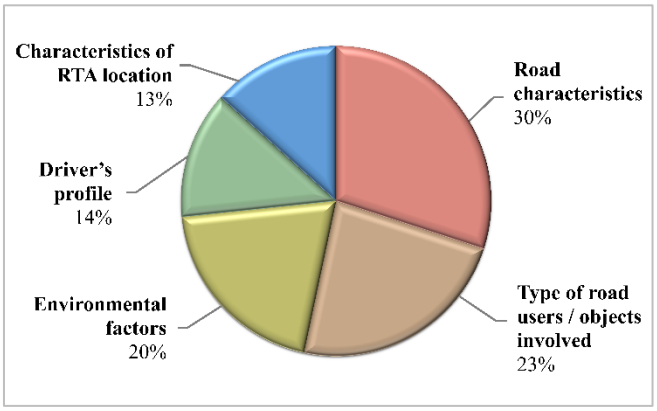

Figure 7. The overall distribution of the top factors contributing to RTA for Ethiopia, Finland, Germany, New Zealand, the UK, and the US with RTA data samples from 2000 to 2021.

5. Conclusion

The increasing trend of RTAs worldwide is a significant global challenge, with RTAs currently ranked as a top factor of injury-related fatalities. Therefore, research efforts are necessary to analyse the historical RTA records to determine the RTA patterns. This work has performed the RTA pattern analysis with ARM, a data mining technique to identify the patterns of RTAs in six regions including Ethiopia (Addis Ababa city), Finland, Germany (Berlin city-state), New Zealand, the UK, and the US. The open-access RTA datasets for the six regions were obtained and pre-processed before the stage of ARM. This included data cleaning and transformation techniques to remove the missing values and transform RTA features from numerical to categorical. This work also successfully generated the association rules with the FP-Growth algorithm based on the pre-processed open-access RTA datasets. The data instances were grouped based on the RTA severity levels (fatal, serious injuries, slight/minor injuries, and non-injury) and association rules were computed to represent the RTA patterns for each severity level. Filtering was applied with the minimum values of support (20%), confidence (70%), and lift (1.20) to ensure the quality of the rules. However, the generated association rules for Germany (minor injuries), New Zealand (all severity levels), and the US (fatal) regions could not satisfy the minimum lift value. Thus, adjustments were made with a lower lift value (above 1.00) for these cases to obtain useful rules for analysis.

RTA patterns in the six regions were evaluated with the top five association rules selected for each severity level. The findings from Ethiopia showed that road geometrics can contribute to the severity level of RTAs. RTAs that occurred at the Y-shape junction usually resulted in serious injuries, while fatalities were commonly observed for the collisions on tangent roads. This factor was also discovered with the dataset from the UK region, where there is an elevated risk of fatalities for the RTAs that happened at the T or staggered junctions. The analysis made with the dataset from Germany pointed out that the type of road users involved in RTA also contributes to the severity level. There is a higher risk of injuries and fatalities when the RTA involves vulnerable users and heavy vehicles. Another interesting pattern found is the effect of speed limit on the RTA severity level. Association rules produced with the dataset from the UK region showed that most of the RTAs with slight injuries are found to have a speed limit of around 30 km/h. There are also limitations to this study in that the results for Finland, New Zealand, and the US regions did not show a significant pattern to distinguish between the different severity levels of RTA. This showed the necessity to consider other prominent features in future studies to analyse the patterns of RTAs in these regions.

This work also analysed the significant factors contributing to RTAs in the six regions based on the percentage of a feature detected in the resultant association rules. These values were illustrated through matrix plots to determine the top factors. The findings showed that the top factors contributing to RTAs can be classified into five categories: road characteristics, type of road users or objects involved, environmental factors, driver's profile, and characteristics of RTA location. These factors can be considered significant for the six regions. Moreover, there are also differences between the top factors contributing to RTAs among the six regions. The findings with Germany dataset showed that the top factors focused on the involvement of various road users and objects, while road characteristics were found to be a major factor in New Zealand. In addition, the top factors contributing to RTAs in the US region are closely related to the driver's profile. These observations pointed out that there are other factors to consider for different regions other than the common factors found in this study. To conclude, this study has proven the effectiveness of ARM in analysing RTA trends and patterns. This allows the identification of key factors contributing to RTAs so that appropriate countermeasures are formulated to prevent RTA in the future.

References

[1] WHO. Road traffic injuries. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries (accessed October 6, 2022).

[2] WHO. Global Status Report on Road Safety (2018). https://www.who.int/publications/i/item/9789241565684 (accessed October 6, 2022).

[3] Onyemaechi NOC, Ofoma UR. The public health threat of road traffic accidents in Nigeria: A call to action. Annals of Medical and Health Sciences Research. 2017. https://doi.org/10.4103/amhsr.amhsr_452_15.

[4] Talia D, Trunfio P, Marozzo F. Data Analysis in the Cloud. Elsevier; 2015. Chapter 1, Introduction to Data Mining; pp 1-25.

[5] Kantardzic M. Data Mining: Concepts, Models, Methods, and Algorithms. 2nd ed. Wiley-IEEE Press; 2011. Chapter 8, Association Rules; pp 280-299. https://doi.org/10.1109/9780470544341.CH8

[6] Gupta M, Kumar Solanki V, Kumar Singh V. A Novel Framework to Use Association Rule Mining for classification of traffic accident severity. Ingeniería Solidaria. 2017; 13(21):37–44. https://doi.org/10.16925/in.v13i21.1726

[7] Arun V, Khan FN. Traffic Mishap Injury Severity: An Unsupervised Approach. 2020 IEEE International Conference for Innovation in Technology (INOCON); 2020 November 6-8; Bangluru, India. 2020. pp. 1-8. https://doi.org/10.1109/INOCON50539.2020.9298218

[8] Priya S, Agalya R. Association Rule Mining Approach to Analyze Road Accident Data. Proceedings of the 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT); 2018 March 1-3; Coimbatore, India. 2018. pp. 1-5. https://doi.org/10.1109/ICCTCT.2018.8550950

[9] Feng M, Zheng J, Ren J, Xi Y. Association rule mining for road traffic accident analysis: A case study from UK. Advances in Brain Inspired Cognitive Systems; In: Lecture Notes in Computer Science, 11691:520–529. https://doi.org/10.1007/978-3-030-39431-8_50

[10] Makarova I, Yakupova G, Buyvol P, Mukhametdinov E, Pashkevich A. Association rules to identify factors affecting risk and severity of road accidents. In: Proceedings of the 6th International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS 2020); 2020 May 2-4; Online. 2020. pp. 614-621. https://doi.org/10.5220/0009836506140621

[11] Li L, Shrestha S, Hu G. Analysis of road traffic fatal accidents using data mining techniques. In: Proceedings of 2017 15th IEEE/ACIS International Conference on Software Engineering Research, Management and Applications (SERA); 2017 June 7-9; London, UK. 2017 pp 363-370. https://doi.org/10.1109/SERA.2017.7965753

[12] Gu C, Xu J, Gao C, Mu M, E G, Ma Y. Multivariate analysis of roadway multi-fatality crashes using association rules mining and rules graph structures: A case study in China. PLOS ONE. 2022; 17(10):e0276817. https://doi.org/10.1371/journal.pone.0276817

[13] Xu C, Bao J, Wang C, Liu P. Association rule analysis of factors contributing to extraordinarily severe traffic crashes in China. Journal of Safety Research. 2018; 67:65-75. https://doi.org/10.1016/J.JSR.2018.09.013

[14] Das S, Dutta A, Avelar R, Dixon K, Sun X, Jalayer M. Supervised association rules mining on pedestrian crashes in urban areas: identifying patterns for appropriate countermeasures. International Journal of Urban Sciences. 2019; 23(1):30-48. https://doi.org/10.1080/12265934.2018.1431146

[15] Hossain A, Sun X, Thapa R, Codjoe J. Applying Association Rules Mining to Investigate Pedestrian Fatal and Injury Crash Patterns Under Different Lighting Conditions. Transportation Research Record. 2022; 2676(6):659-672, https://doi.org/10.1177/03611981221076120

[16] Sivasankaran SK, Natarajan P, Balasubramanian V. Identifying Patterns of Pedestrian Crashes in Urban Metropolitan Roads in India using Association Rule Mining. Transportation Research Procedia. 2020; 48:3496-3507. https://doi.org/10.1016/j.trpro.2020.08.102

[17] Weng J, Zhu JZ, Yan X, Liu Z. Investigation of work zone crash casualty patterns using association rules. Accident Analysis & Prevention. 2016; 92:43-52. https://doi.org/10.1016/J.AAP.2016.03.017

[18] Dalai B, Landge VS. Crash risk factor identification using association rules in Nagpur city, Maharashtra, India. Current Science. 2022; 123(6):781-790. https://doi.org/10.18520/cs/v123/i6/781-790

[19] Almutairi A, Alkandari D, Shummais L, Alajmi R, Toma T. Association Rule Mining for Driving Behaviors and Road Traffic Accidents in Kuwait. In: Proceedings of the 11th Annual International Conference on Industrial Engineering and Operations Management Singapore; 2021 March 7-11; Singapore. 2021. pp. 7550-7557. http://www.ieomsociety.org/singapore2021/papers/1301.pdf

[20] Nidhi R, Kanchana M. Analysis of Road Accidents Using Data Mining Techniques. International Journal of Engineering & Technology. 2018; 7(3):40-44. https://doi.org/10.14419/ijet.v7i3.10.15626

[21] Janani G, Devi NR. Road Traffic Accidents Analysis Using Data Mining Techniques. Journal of Information Technology and Applications (JITA). 2018; 14(2):84-91. https://doi.org/10.7251/JIT1702084J

[22] Joshi S, Alsadoon A, Senanayake SMNA, Prasad PWC, Yong SY, Elchouemi A, Vo TH. Pattern Mining Predictor System for Road Accidents. Communications in Computer and Information Science. 2020; 1287:605-615. https://doi.org/10.1007/978-3-030-63119-2_49

[23] Senanayake SMNA, Joshi S. A road accident pattern miner (RAP miner). Journal of Information and Telecommunication. 2021; 5(4):484-498. https://doi.org/10.1080/24751839.2021.1955533

[24] Pratama Y, Riziana AT, Saputri DY, Wahyudi R, Ismiyati R, Tahyudin I. A Comparative Analysis of Tertius, Apriori, and FP-Growth Algorithm in Groceries Dataset. 2022 1st International Conference on Smart Technology, Applied Informatics, and Engineering (APICS); 2022 August 23-24; Surakarta, Indonesia. 2022. pp. 65-69. https://doi.org/10.1109/APICS56469.2022.9918776

[25] Dharmarajan K, Dorairangaswamy MA. Analysis of FP-growth and Apriori algorithms on pattern discovery from weblog data. 2016 IEEE International Conference on Advances in Computer Applications (ICACA); 2016 October 24; Coimbatore, India. 2016. pp. 170-174. https://doi.org/10.1109/ICACA.2016.7887945

Appendix A. Top Association Rules for Different Regions

Table 4. Top groups of association rules generated for RTAs in Ethiopia: (a) Fatal, (b)Serious injuries, and (c) Slight injuries.

a) Fatal RTAs

RTA Features	Group				
	1	2	3	4	5
accident_severity	Fatal	Fatal	Fatal	Fatal	Fatal
educational_level	Junior high school	Junior high school	Junior high school	Junior high school	Junior high school
number_of_vehicles	2	2	2	2	-
owner_of_vehicle	Owner	Owner	Owner	-	Owner
pedestrian_movement	Not a Pedestrian	Not a Pedestrian	-	Not a Pedestrian	Not a Pedestrian
road_allignment	-	Tangent road	-	-	-
road_surface_conditions	Dry	Dry	Dry	Dry	Dry
road_surface_type	-	Asphalt roads	Asphalt roads	-	Asphalt roads
sex_of_driver	-	Male	Male	Male	Male
vehicle_driver_relation	-	Employee	-	-	Employee
vehicle_movement	Going straight	-	Going straight	Going straight	Going straight
weather_conditions	Normal	Normal	Normal	Normal	Normal

b) Serious Injuries RTAs

RTA Features	Group				
	1	2	3	4	5
accident_severity	Serious	Serious	Serious	Serious	Serious
day_of_week	Weekday	Weekday	-	-	-
light_conditions	Daylight	Daylight	Daylight	Daylight	Daylight
owner_of_vehicle	-	-	Owner	Owner	-
pedestrian_movement	-	-	-	Not a Pedestrian	-
road_surface_conditions	Dry	Dry	Dry	Dry	Dry
road_surface_type	-	-	Asphalt roads	-	-
sex_of_driver	Male	-	-	Male	-
types_of_junction	Y Shape	Y Shape	Y Shape	Y Shape	Y Shape
vehicle_driver_relation	-	-	-	-	Employee
weather_conditions	-	Normal	Normal	Normal	Normal

c) Slight Injuries RTAs

RTA Features	Group				
	1	2	3	4	5
accident_severity	Slight	Slight	Slight	Slight	Slight
day_of_week	Weekday	Weekday	Weekday	Weekday	Weekday
light_conditions	Daylight	Daylight	Daylight	Daylight	Daylight
number_of_vehicles	2	2	2	2	2
owner_of_vehicle	Owner	-	Owner	-	Owner
pedestrian_movement	-	-	-	Not a Pedestrian	Not a Pedestrian
road_allignment	Tangent road	Tangent road	-	Tangent road	-
road_surface_conditions	Dry	Dry	Dry	Dry	Dry
road_surface_type	-	Asphalt roads	Asphalt roads	-	-
sex_of_driver	Male	Male	Male	Male	Male
vehicle_driver_relation	Employee	Employee	Employee	Employee	Employee
weather_conditions	Normal	Normal	Normal	Normal	Normal

Table 5. Top groups of association rules generated for RTAs in Finland: (a) Fatal, (b) Injury, and (c) Non-injury.

a) Fatal RTAs

RTA Features	Group				
	1	2	3	4	5
severity	Fatal	Fatal	Fatal	Fatal	Fatal
accident_location	driveway	driveway	driveway	driveway	-
construction_zone	No	No	-	No	No
intersection_type	-	line accident	line accident	line accident	line accident
light_condition	daylight	-	daylight	daylight	daylight
road_coating	durable coating	durable coating	durable coating	-	durable coating
road_surface	dry	dry	dry	dry	dry
season	Summer	Summer	-	-	-
weather	-	-	clear	clear	clear

b) Injury RTAs

RTA Features	Group				
	1	2	3	4	5
severity	Injury	Injury	Injury	Injury	Injury
accident_location	driveway	driveway	-	driveway	-
construction_zone	No	No	No	-	No
day_of_week	-	-	Weekday	-	-
intersection_type	-	-	-	line accident	line accident
light_condition	daylight	daylight	daylight	daylight	daylight
road_coating	durable coating	durable coating	durable coating	-	-
road_surface	dry	dry	dry	dry	dry
season	Summer	-	-	-	-
weather	-	clear	clear	clear	clear

c) Non-Injury RTAs

RTA Features	Group				
	1	2	3	4	5
severity	Non-Injury	Non-Injury	Non-Injury	Non-Injury	Non-Injury
accident_location	-	-	-	-	driveway
construction_zone	-	No	-	-	-
day_of_week	-	-	-	Weekday	-
intersection_type	-	-	line accident	-	line accident
light_condition	daylight	daylight	-	-	-
road_coating	durable coating		durable coating	-	-
road_surface	dry	dry	dry	dry	dry
weather	clear	clear	clear	clear	clear

Table 6. Top groups of association rules generated for RTAs in Germany: (a) Fatal, (b)Serious injuries, and (c) Minor injuries

a) Fatal RTAs

RTA Features	Group				
	1	2	3	4	5
Severity	Fatal	Fatal	Fatal	Fatal	Fatal
Bicycle_Involvement	-	-	Yes	Yes	-
Car_Involvement	No	No	-	-	No
Heavy_Vehicle_Involvement	Yes	Yes	-	-	Yes
Light_Condition	Daylight	-	Daylight	-	Daylight
Motorcycle_Involvement	No	No	No	No	-
Other_Vehicle_Type_Involvement	No	No	-	No	No
Pedestrian_Involvement	-	-	No	No	-
Road_Condition	-	Dry	-	-	Dry

b) Serious Injuries RTAs

RTA Features	Group				
	1	2	3	4	5
Severity	Serious Injuries	Serious Injuries	Serious Injuries	Serious Injuries	Serious Injuries
Accident_Type_Involved_Object	Pedestrian	-	-	-	-
Bicycle_Involvement	-	No	No	No	No
Car_Involvement	-	-	Yes	Yes	Yes
Day_of_Week	-	-	Weekday	-	-
Heavy_Vehicle_Involvement	No	-	No	No	No
Light_Condition	-	-	-	-	Daylight
Motorcycle_Involvement	No	Yes	No	No	No
Other_Vehicle_Type_Involvement	-	-	No	No	No
Pedestrian_Involvement	Yes	-	-	-	-
Road_Condition	-	-	-	Dry	-

c) Minor Injuries RTAs

RTA Features	Group				
	1	2	3	4	5
Severity	Minor Injuries	Minor Injuries	Minor Injuries	Minor Injuries	Minor Injuries
Bicycle_Involvement	No	No	No	No	No
Car_Involvement	Yes	Yes	Yes	Yes	Yes
Day_of_Week	Weekday	-	-	Weekday	-
Heavy_Vehicle_Involvement	No	No	No	-	-
Light_Condition	-	Daylight	-	-	Daylight
Motorcycle_Involvement	No	No	No	No	No
Other_Vehicle_Type_Involvement	No	No	No	No	No
Pedestrian_Involvement	-	-	No	No	No
Road_Condition	-	-	Dry	-	-

Table 7. Top groups of association rules generated for RTAs in New Zealand: (a) Fatal, (b) Serious injuries, (c) Minor injuries and (d) Non-injury

a) Fatal RTAs

RTA Features	Group				
	1	2	3	4	
accident_severity	Fatal	Fatal	Fatal	Fatal	
2-wheeled	No	No	No	No	
4-wheeled	Yes	Yes	Yes	Yes	
holiday	None	None	None	None	
number_of_lanes	2	2	2	-	
other_vehicles	No	-	No	No	
public_transport	No	No	-	No	
road_surface	Sealed	Sealed	Sealed	Sealed	
road_type	2-way	2-way	2-way	2-way	
special_road_characteristics	None	None	None	None	
weather	-	Fine	Fine	Fine	

b) Serious Injuries RTAs

RTA Features	Group				
	1	2	3	4	5
accident_severity	Serious Injuries	Serious Injuries	Serious Injuries	Serious Injuries	Serious Injuries
4-wheeled	Yes	Yes	Yes	Yes	Yes
flat_or_hill	Flat	Flat	Flat	Flat	Flat
holiday	None	-	None	None	None
number_of_lanes	2	2	2	2	2
other_vehicles	-	No	No	No	No
public_transport	No	No	No	-	No
road_surface	Sealed	Sealed	-	Sealed	Sealed
road_type	2-way	2-way	2-way	2-way	2-way
special_road_characteristics	None	-	None	None	None
weather	Fine	Fine	Fine	Fine	-

c) Minor Injuries RTAs

RTA Features	Group			
	1	2	3	
accident_severity	Minor Injuries	Minor Injuries	Minor Injuries	
2-wheeled	No	No	No	
4-wheeled	Yes	Yes	Yes	
flat_or_hill	Flat	Flat	Flat	
holiday	None	None	None	
other_vehicles	-	No	No	
public_transport	No	-	No	
road_surface	Sealed	Sealed	Sealed	
road_type	2-way	2-way	2-way	
special_road_characteristics	None	None	None	
weather	Fine	Fine	-	

d) Non-Injury RTAs

RTA Features	Group			
	1	2	3	
accident_severity	Non-Injury	Non-Injury	Non-Injury	
2-wheeled	No	No	-	
4-wheeled	Yes	Yes	-	
flat_or_hill	Flat	Flat	Flat	
holiday	None	None	-	
other_vehicles	No	-	No	
public_transport	-	No	No	
road_surface	Sealed	Sealed	-	
road_type	2-way	2-way	2-way	
special_road_characteristics	None	None	None	
weather	Fine	Fine	Fine	

Table 8. Top groups of association rules generated for RTAs in the UK: (a) Fatal, (b)Serious injuries, and (c) Slight injuries

a) Fatal RTAs

RTA Features	Group				
	1	2	3	4	5
accident_severity	Fatal	Fatal	Fatal	Fatal	Fatal
carriageway_hazards	None	-	None	None	None
junction_control	Give way / Uncontrolled	Give way / Uncontrolled	Give way / Uncontrolled	Give way / Uncontrolled	Give way / Uncontrolled
junction_detail	T/staggered junction	T/staggered junction	T/staggered junction	T/staggered junction	T/staggered junction
pedestrian_crossing_human_control	None	None	None	-	None
pedestrian_crossing_physical_facilities	-	None	None	None	None
road_type	Single carriageway	Single carriageway	Single carriageway	Single carriageway	Single carriageway
second_road_class	Unclassified	-	-	-	-
special_conditions_at_site	None	None	None	None	-
trunk_road_flag	Non-Trunk	Non-Trunk	-	Non-Trunk	Non-Trunk
urban_or_rural_area	Urban	Urban	Urban	Urban	Urban
weather_conditions	Normal	-	-	-	-

b) Serious Injuries RTAs

RTA Features	Group				
	1	2	3	4	5
accident_severity	Serious Injuries	Serious Injuries	Serious Injuries	Serious Injuries	Serious Injuries
carriageway_hazards	None	None	None	None	-
day_of_week	-	Weekday		Weekday	Weekday
junction_control	Give way / Uncontrolled	Give way / Uncontrolled	Give way / Uncontrolled	Give way / Uncontrolled	Give way / Uncontrolled
light_conditions	Daylight	-	-	-	-
pedestrian_crossing_human_control	None	-	None	None	None
pedestrian_crossing_physical_facilities	None	None	None	None	None
road_surface_conditions	-	-	Dry	-	-
road_type	Single carriageway	Single carriageway	Single carriageway	Single carriageway	Single carriageway
second_road_class	Unclassified	Unclassified	Unclassified	Unclassified	Unclassified
special_conditions_at_site	None	None	None	-	None
trunk_road_flag	Non-Trunk	Non-Trunk	Non-Trunk	Non-Trunk	Non-Trunk
urban_or_rural_area	Urban	Urban	Urban	Urban	Urban
weather_conditions	Normal	Normal	Normal	Normal	Normal

c) Slight Injuries RTAs

RTA Features	Group				
	1	2	3	4	5
accident_severity	Slight Injuries	Slight Injuries	Slight Injuries	Slight Injuries	Slight Injuries
carriageway_hazards	None	-	None	None	None
junction_control	-	-	-	Give way / Uncontrolled	Give way / Uncontrolled
light_conditions	Daylight	Daylight	Daylight	-	Daylight
number_of_vehicles	2	2	2	-	-
pedestrian_crossing_human_control	None	None	None	None	None
pedestrian_crossing_physical_facilities	-	-	-	None	-
road_surface_conditions	Dry	Dry	Dry	Dry	Dry
road_type	Single carriageway	Single carriageway	Single carriageway	Single carriageway	Single carriageway
special_conditions_at_site	-	None	None	None	None
speed_limit	<30	<30	<30	<30	<30
trunk_road_flag	Non-Trunk	Non-Trunk	-	Non-Trunk	Non-Trunk
urban_or_rural_area	Urban	Urban	Urban	Urban	Urban
weather_conditions	Normal	Normal	Normal	Normal	Normal

Table 9. Top groups of association rules generated for fatal RTAs in the US

a) Fatal RTAs

RTA Features	Group		
	1	2	3
Severity	Fatal	Fatal	Fatal
Distracted_Driver	No	-	No
Drowsy_Driver	No	No	-
Hit_and_Run	No	No	No
Interstate	No	No	No
Large_Truck_Involvement	No	No	No
Pedalcyclist_Involvement	No	No	No
Police_Pursuit_Involvement	No	No	No
Weather	Clear	Clear	Clear
Young_Driver_15_19	No	No	No
Young_Driver_15_20	No	No	No
Young_Driver_16_19	No	No	No
Young_Driver_16_20	No	No	No
Young_Driver_21_24	-	No	No

Appendix B. Sources for Open-Access RTA Datasets

The open-access RTA datasets investigated in this work are accessible as follows.

Table 10. Sources for open-access RTA datasets.

No.	Region	Source URL
1	Ethiopia (Addis Ababa city)	https://doi.org/10.17632/xytv86278f.1
2	Finland	https://www.opendata.fi/data/en_GB/dataset/tieliikenneonnettomuudet
3	Germany (Berlin city-state)	https://daten.berlin.de/search/node/Strassenverkehrsunf%C3%A4lle%20nach%20Unfallort
4	New Zealand	https://opendata-nzta.opendata.arcgis.com/datasets/NZTA::crash-analysis-system-cas-data-1/about
5	UK	https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data
6	US	https://www.nhtsa.gov/file-downloads?p=nhtsa/downloads/FARS/