Finding patterns in 2013 road accident data in United Kingdom

Humberto Gonzalez

June 25, 2015

1 Introduction

In this work we analize road accidents data in Great Britain using pattern mining techniques, specifically, associacion rules. The aim of this work is to find interesting associations in the data that may give us some clues about the factors that contribute to the accidents.

The data used comprises of the reports generated by police in the place of a reported road accident and icludes variabes describing the type of road, speed limit, weather and road conditions, traffic control signals and presence of junctions; the full description of the data can be found in [].

The *R* software with *arules* package [3] was used for all the analysis presented in this work. The *arules* package provides an interface between the *R* software and the *apriori* algorithm [2].

The results of the analysis are presented in three sections. In section 3 we analize the most frequent itemsets using three different criteria: the *support*, *lift* and the *Jaccard index*. In section 4 we try to find some interesting rules over the whole data set without doing any assumption on the data that may be given by an expert in the subject. Finally, in section 5 we try to discover the factors that contribute to higher severity in the accidents.

2 Preliminaries

The *apriori* algorithm was used to obtain both the frequent itemsets and the association rules. The parameters passed to the algorithm are adapted to the goals of each section. In this section we describe two measures that will be used throughout the work in order to produce meaningful results.

improvement in confidence [5] This rule will be used to remove redundant rules from the results obtained with *apriori*. To introduce this measure we first give the definition of redundant rule used in this work.

Definition 1. A rule $r : Z \to Y$ is said to be redundant if there exists a rule $s : X \to Y$ with $X \subset Z$, such that

$$|conf(r) - conf(s)| < \gamma \tag{1}$$

for some $\gamma > 0$. In other words, the extra elements in Z do not contribute significantly to the change in confidence.

From definition 1 we can see that if the rule $s^* = argmin_s |conf(r) - conf(s)|$ is such that $|conf(r) - conf(s^*)| < \gamma$, then the rule r is redundant and, therefore, it can be removed. We define this last absolute difference as the *improvement in confidence*.

Note that to obtain the *improvement in confidence* it is necessary to compute, for all rules $X \to Y$ in the results of the *apriori* algorithm, the confidence of the rules $Z \to Y$, where $Z \subset X$. This is a combinatorial problem which makes it unsuitable when a large number of rules is found with the *apriori*.

absolute difference of lift value to 1[2] This measure is used during the execution of the *apriori* algorithm to prune the search space, discarding those rules with $|1 - lift| < minDLift$, where $minDLift > 0$ is an user defined parameter. In this sense, $minDLift$ is to $lift$ what $minConf$ and $minSup$ are to $confidence$ and $support$. This is specially useful when we want to find rules which are scarce, i.e., rules with $minSup$ near zero, avoiding the computation of rules that are "close" to independence.

3 Frequent itemsets

In this section we present the most frequent itemsets in the Road Accident Data. The itemsets were computed using the *apriori* algorithm with minimum support $minSup = 0.1$ and itemsets maximum length $maxLen = 10$, obtaining 45,500 itemsets. The top 10 most frequent itemsets, i.e., those with maximum support are presented in table 1.

Itemset	**Support**
{Pedestrian_Crossing.Human_Control=none_50m}	0.9951608
{Carriageway_Hazards=none}	0.9824895
{Pedestrian_Crossing.Human_Control=none_50m, Carriageway_Hazards=none}	0.9777730
{Special_Conditions_at_Site=none}	0.9776071
{Pedestrian_Crossing.Human_Control=none_50m, Special_Conditions_at_Site=none}	0.9728761
{Special_Conditions_at_Site=none, Carriageway_Hazards=none}	0.9615246
{Pedestrian_Crossing.Human_Control=none_50m, Special_Conditions_at_Site=none, Carriageway_Hazards=none}	0.9569090
{Accident_Severity=slight}	0.8468773
{Accident_Severity=slight, Pedestrian_Crossing.Human_Control=none_50m}	0.8426367
{Accident_Severity=slight, Carriageway_Hazards=none}	0.8322876

Table 1: Top 10 most frequent itemsets ordered by *support*.

The *lift* and *Jaccard index* were computed on the previous 45,500 itemsets. The top frequent itemsets ordered by *lift* and *Jaccard index* are presented in tables 2[1] and 3, respectevely. In the case of the *Jaccard index* the itemsets having only one element were removed from the table as they have a value always equal to one.

itemset	**Support**	**Lift**	**Jaccard**
{Junction_Detail=not_junction_20m, X2nd_Road_Class=no_info, Junction_Control=no_info, Pedestrian_Crossing.Human_Control=none_50m, Pedestrian_Crossing.Physical_Facilities=none_50m, Urban_or_Rural_Area=rural}	0.2017453	11.23122	0.06013522
{Junction_Detail=not_junction_20m, X2nd_Road_Class=no_info, Junction_Control=no_info, Pedestrian_Crossing.Physical_Facilities=none_50m, Urban_or_Rural_Area=rural}	0.2019111	11.18606	0.08556645
{Junction_Detail=not_junction_20m, X2nd_Road_Class=no_info, Junction_Control=no_info, Pedestrian_Crossing.Human_Control=none_50m, Pedestrian_Crossing.Physical_Facilities=none_50m, Weather_Conditions=fine_no_high_winds, Road_Surface_Conditions=dry, Urban_or_Rural_Area=rural}	0.1138685	11.17481	0.02340488
{Junction_Detail=not_junction_20m, X2nd_Road_Class=no_info, Junction_Control=no_info, Pedestrian_Crossing.Physical_Facilities=none_50m, Weather_Conditions=fine_no_high_winds, Road_Surface_Conditions=dry, Urban_or_Rural_Area=rural}	0.1139838	11.13201	0.02945320
{Junction_Detail=not_junction_20m, X2nd_Road_Class=no_info, Junction_Control=no_info, Pedestrian_Crossing.Human_Control=none_50m, Pedestrian_Crossing.Physical_Facilities=none_50m, Weather_Conditions=fine_no_high_winds, Road_Surface_Conditions=dry, Special_Conditions_at_Site=none, Urban_or_Rural_Area=rural}	0.1101399	11.05649	0.01885064

Table 2: Top 5 most frequent itemsets ordered by *lift*.

[1]The complete table can be found in appendix A.

itemset	Support	Lift	Jaccard
{Junction_Detail=not_junction_20m, Junction_Control=no_info}	0.3943531	2.5293676	0.4993653
{X2nd_Road_Class=no_info, Junction_Control=no_info}	0.3949156	2.4789928	0.4946970
{Junction_Detail=not_junction_20m, X2nd_Road_Class=no_info}	0.3943675	2.4807574	0.4945242
{Pedestrian_Crossing.Human_Control=none_50m, Carriageway_Hazards=none}	0.9777730	1.0000387	0.4944114
{Pedestrian_Crossing.Human_Control=none_50m, Special_Conditions_at_Site=none}	0.9728761	0.9999998	0.4931528
{Special_Conditions_at_Site=none, Carriageway_Hazards=none}	0.9615246	1.0010785	0.4905496
{Accident_Severity=slight, Pedestrian_Crossing.Human_Control=none_50m}	0.8426367	0.9998310	0.4574480
{Accident_Severity=slight, Carriageway_Hazards=none}	0.8322876	1.0002879	0.4549594
{Accident_Severity=slight, Special_Conditions_at_Site=none}	0.8280759	1.0001965	0.4538684
{Pedestrian_Crossing.Human_Control=none_50m, Pedestrian_Crossing.Physical_Facilities=none_50m}	0.8121592	1.0024504	0.4488867
{Weather_Conditions=fine_no_high_winds, Road_Surface_Conditions=dry}	0.6744050	1.1888691	0.4465375

Table 3: Top 10 most frequent itemsets ordered by *Jaccard index*.

4 Association rules

To obtain the association rules in the dataset, *apriori* algorithm was parametrized with $minSup = 0.1$, $minConf = 0.8$ and limiting the number of items in the left hand side by 7, obtaining 118,397 rules. However, much of these rules are redundant, making it difficult to find the interesting ones. To remove the redundant rules we will use the *improvement in confidence* measure detailed in section 2, but, as the number of rules obtained by the *apriori* algorithm is large, we first need to reduce the size of the result set. To this purpose we will use the *Absolute Difference of Lift Value to 1* (section 2), removing all rules that are close to independence.

In this section a threshold for the *Absolute Difference of Lift Value to 1* of 0.1 was used reducing the number of rules obtained with *apriori* from 118,397 to 45,527. Then, the *improvement in confidence* measure was used to remove the rules which improvement was less than 0.05, reducing the total number of rules from 45,527 to 128. The results, ordered by different measures of interestingness, are presented below.

4.1 Most general and most specific rules

Here we define the most general rules to be those with higher support and which confidence is over the 0.8 threshold. Simmilarly, we define the most specific rules to be those with lower support, but wich confidence still over the 0.8 threshold. The top 10 most general and specific rules are presented in tables 4 and 5[2], respectevely.

lhs	rhs	sup	conf	lift
{Road_Surface_Conditions=dry}	{Weather_Conditions=fine_no_high_winds}	0.674	0.9627	1.1888
{Weather_Conditions=fine_no_high_winds}	{Road_Surface_Conditions=dry}	0.674	0.8328	1.1888
{Speed_limit=30}	{Urban_or_Rural_Area=urban}	0.5606	0.8687	1.3424
{Urban_or_Rural_Area=urban}	{Speed_limit=30}	0.5606	0.8663	1.3424
{Road_Type=single_carriageway, Urban_or_Rural_Area=urban}	{Speed_limit=30}	0.4667	0.9223	1.4292
{Speed_limit=30, Pedest_Cross.Physical_Facilities=none_50m}	{Road_Type=single_carriageway}	0.4121	0.8527	1.1278
{Junction_Control=no_info}	{X2nd_Road_Class=no_info}	0.3949	0.9991	2.4789
{X2nd_Road_Class=no_info}	{Junction_Control=no_info}	0.3949	0.9798	2.4789
{Junction_Detail=not_junction_20m}	{X2nd_Road_Class=no_info}	0.3943	0.9998	2.4807
{X2nd_Road_Class=no_info}	{Junction_Detail=not_junction_20m}	0.3943	0.9785	2.4807

Table 4: Top 10 most general rules

[2]The complete table can be found in the appendix B

lhs	rhs	sup	conf	lift
{X1st_Road_Class=A, Junction_Control=give_way_or_uncontrolled, Pedest_Cross.Physical_Facilities=none_50m, Weather_Conditions=fine_no_high_winds, Road_Surface_Conditions=dry, Carriageway_Hazards=none}	{Light_Conditions=daylight}	0.1016	0.8146	1.1111
{Weather_Conditions=raining_no_high_winds}	{Road_Surface_Conditions=wet_or_damp}	0.1028	0.9797	3.6979
{Number_of_Vehicles=2, X1st_Road_Class=unclassified, Weather_Conditions=fine_no_high_winds}	{Urban_or_Rural_Area=urban}	0.1045	0.8005	1.2371
{Number_of_Vehicles=1, Light_Conditions=daylight, Urban_or_Rural_Area=urban}	{Road_Surface_Conditions=dry}	0.1073	0.8011	1.1436
{Number_of_Vehicles=1, Speed_limit=30, Light_Conditions=daylight, Special_Conditions_at_Site=none}	{Road_Surface_Conditions=dry}	0.1094	0.8003	1.1425

Table 5: Top 5 most specific rules

4.2 Similar rules

Similar rules can be either quasi-conjunctions or quasi-equivalences. The quasi-conjunctions are the rules that satisfy both $X \to Y$ and $Y \to X$, while the quasi-equivalences are quasi-conjunctions that also satisfy $\neg Y \to \neg X$ and $\neg X \to \neg Y$. Because quasi-equivalences are also quasi-conjunctions we propose two different interestingness measure to compute them.

To obtain the quasi-conjunctions we use the *Kulczynski index*, defined as

$$Kulczynski(X \to Y) = \frac{1}{2}\left(conf(X \to Y) + conf(Y -> X)\right) = \frac{1}{2}\left(\frac{n_{xy}}{n_x} + \frac{n_{xy}}{n_y}\right). \qquad (2)$$

The rules with the highes *Kulczynski index* are presented in table 6.

lhs	rhs	sup	conf	Kulcz
{Junction_Detail=not_junction_20m}	{Junction_Control=no_info}	0.3943	0.9997	0.9987
{Junction_Control=no_info}	{X2nd_Road_Class=no_info}	0.3949	0.9991	0.9894
{Junction_Detail=not_junction_20m}	{X2nd_Road_Class=no_info}	0.3943	0.9998	0.9891
{Road_Surface_Conditions=dry}	{Weather_Conditions=fine_no_high_winds}	0.6744	0.9627	0.8977
{Speed_limit=30}	{Urban_or_Rural_Area=urban}	0.5606	0.8687	0.8675
{Road_Type=single_carriageway, Urban_or_Rural_Area=urban}	{Speed_limit=30}	0.4667	0.9223	0.8228
{X2nd_Road_Class=unclassified}	{Junction_Control=give_way_uncontrolled}	0.3593	0.8942	0.8107
{X2nd_Road_Class=unclassified, Pedest_Cross.Physical_Facilities=none_50m}	{Junction_Control=give_way_uncontrolled}	0.3042	0.9508	0.7833
{Road_Type=single_carriageway, Junction_Control=give_way_uncontrolled}	{X2nd_Road_Class=unclassified}	0.3049	0.8006	0.7797
{Junction_Detail=T_junction}	{Junction_Control=give_way_uncontrolled}	0.2815	0.8805	0.7251

Table 6: Top 10 similar rules computed with the *Kulczynski idex*.

To obtain the best quasi-equivalences we use the *causal support* (Sokhal-Michener index) defined as

$$causal(X \to Y) = Pr(X, Y) + Pr(\bar{X}, \bar{Y}) = \frac{n_{xy} + n_{\bar{x}\bar{y}}}{n}. \qquad (3)$$

The results for the best quasi-equivalences are presented in table 7.

If we analize the results given in tables 6 and 7 we can see that when there is no junction near the accident (Junction_Detail=not_junction_20m), then there cannot be information about the junction control (Junction_Control=no_info) nor the class of a second road (X2nd_Road_Class=no_info). More interesting is the fact that these implications are also quasi-equivalences, with a high *causal support*, which means that the police officers always fill-in the junction details in the accidents form whenever there is a junction. This can be interpreted as good training (or practices) of the police officers regarding the fill-in of the accident form. Also, some expected similarities can be

lhs	rhs	sup	conf	causal
{Junction_Detail=not_junction_20m}	{Junction_Control=no_info}	0.3943	0.9997	0.9989
{Junction_Control=no_info}	{X2nd_Road_Class=no_info}	0.3949	0.9991	0.9915
{Junction_Detail=not_junction_20m}	{X2nd_Road_Class=no_info}	0.3943	0.9998	0.9912
{Road_Surface_Conditions=dry}	{Weather_Conditions=fine_no_high_winds}	0.6744	0.9627	0.8385
{Weather_Conditions=raining_no_high_winds}	{Road_Surface_Conditions=wet_or_damp}	0.1028	0.9797	0.8358
{Speed_limit=30}	{Urban_or_Rural_Area=urban}	0.5606	0.8687	0.8288
{Road_Type=single_carriageway, Junction_Control=give_way_or_uncontrolled}	{X2nd_Road_Class=unclassified}	0.3049	0.8006	0.8271
{X2nd_Road_Class=unclassified}	{Junction_Control=give_way_uncontrolled}	0.3593	0.8942	0.8227
{X2nd_Road_Class=unclassified, Pedest_Cross.Physical_Facilities=none_50m}	{Junction_Control=give_way_uncontrolled}	0.3042	0.9508	0.7944
{Road_Type=single_carriageway, Urban_or_Rural_Area=urban}	{Speed_limit=30}	0.4667	0.9223	0.7821
{Speed_limit=60}	{Urban_or_Rural_Area=rural}	0.1385	0.9613	0.7801

Table 7: Top 10 quasi-equivalent rules computed with the *causal support*.

found as those between 1) the speed limit and the urban/rural zone and 2) the weather and the road conditions; in the last case, the equivalence may become higher if the no high winds and high winds were grouped.

4.3 Strong relations

A rule which shows a strong relation is a rule in which the ocurrence of the left hand side gives more information about the right hand side. In terms of probability this can be viewed as the dependence between random variables, i.e. $Pr(Y|X) \neq Pr(Y)$. Therefore, the farther the

$$lift(X \rightarrow Y) = \frac{Pr(Y|X)}{Pr(Y)}$$

deviates from 1, the more dependent the itemsets X and Y are.

In addition to the *lift* we can use statistical tests to tell wether the itemsets are independent or not. In this work we use the χ^2 to reject independence hypothesis and the *lift* as a measure of the strength of dependence. The results for the rules with higher *lift* are shown in table 8.

lhs	rhs	sup	conf	lift	χ^2
{Weather_Conditions=raining_no_high_winds}	{Road_Surface_Conditions=wet/damp}	0.1028	0.9797	3.6979	0
{Speed_limit=60}	{Urban_or_Rural_Area=rural}	0.1385	0.9613	2.7244	0
{Junction_Detail=not_junction_20m}	{Junction_Control=no_info}	0.3943	0.9997	2.5293	0
{Junction_Control=no_info}	{Junction_Detail=not_junction_20m}	0.3943	0.9976	2.5293	0
{Junction_Detail=not_junction_20m}	{X2nd_Road_Class=no_info}	0.3943	0.9998	2.4807	0
{X2nd_Road_Class=no_info}	{Junction_Detail=not_junction_20m}	0.3943	0.9785	2.4807	0
{Junction_Control=no_info}	{X2nd_Road_Class=no_info}	0.3949	0.9991	2.4789	0
{X2nd_Road_Class=no_info}	{Junction_Control=no_info}	0.3949	0.9798	2.4789	0
{X1st_Road_Class=unclassified, Junction_Control=give_way/uncontrolled}	{X2nd_Road_Class=unclassified}	0.1344	0.8698	2.1645	0
{Junction_Detail=T_junction, Junction_Control=give_way/uncontrolled}	{X2nd_Road_Class=unclassified}	0.2314	0.8219	2.0454	0

Table 8: Top 10 rules with higher *lift* and their corresponding p-value for the χ^2 test.

In table 8 we can observe that the p-value of the χ^2 test is 0 for all rules, hence, the null hypothesis of independence is rejected with confidence of 99.9%.

4.4 Prediction

A rule $X \rightarrow Y$ is good for prediction if on the one hand X and Y are dependent and, on the other hand, Y is likely to occur whenever X occurs. Because

$$conf(X \rightarrow Y) = Pr(Y|X)$$

we may use this measure as the prediction power of a rule.

As with strong relations in section 4.3, we use the χ^2 to test independence. We present the data ordered by confidence in the table 9.

lhs	rhs	sup	conf	lift	χ^2
{Junction_Detail=not_junction_20m}	{X2nd_Road_Class=no_info}	0.3943	0.9998	2.4807	0
{Junction_Detail=not_junction_20m}	{Junction_Control=no_info}	0.3943	0.9997	2.5293	0
{Junction_Control=no_info}	{X2nd_Road_Class=no_info}	0.3949	0.9991	2.4789	0
{Junction_Control=no_info}	{Junction_Detail=not_junction_20m}	0.3943	0.9976	2.5293	0
{Speed_limit=60}	{Pedest_Cross.Physical_Facilities=none_50m}	0.1424	0.9880	1.2136	0
{X2nd_Road_Class=no_info}	{Junction_Control=no_info}	0.3949	0.9798	2.4789	0
{Weather_Conditions=raining_no_high_winds}	{Road_Surface_Conditions=wet_or_damp}	0.1028	0.9797	3.6979	0
{X2nd_Road_Class=no_info}	{Junction_Detail=not_junction_20m}	0.3943	0.9785	2.4807	0
{Urban_or_Rural_Area=rural}	{Pedest_Cross.Physical_Facilities=none_50m}	0.3401	0.9639	1.1840	0
{Road_Surface_Conditions=dry}	{Weather_Conditions=fine_no_high_winds}	0.6744	0.9627	1.1888	0
{Speed_limit=60}	{Urban_or_Rural_Area=rural}	0.1385	0.9613	2.7244	0

Table 9: Top 10 rules with higher *conf* and their corresponding p-value for the χ^2 test.

As in section 4.3 we can observe that the p-value of the χ^2 test is 0 for all rules, hence, the null hypothesis of independence is rejected with confidence of 99.9% and we can assert that these rules are the best for prediction.

5 Factors that influence the severity of the accident

In this section we want to find the factors that make an accident with certain severity more likely to occur, in other words, we want to find X such that the change in the probability $Pr(Accident\_Severity = Serious \mid X)/Pr(Accident\_Severity = Serious)$ is "large". By noting that

$$\frac{Pr(Y \mid X)}{P(Y)} = \frac{conf(X \to Y)}{conf(\emptyset \to Y)} = lift(X \to Y), \tag{4}$$

the problem can be stated as finding the rules $X \to Accident\_Severity = Serious$ with high $lift$. In other words, we're more concerned with the change on the confidence than it its value.

As one may expect from a database of road accidents, the serious and fatal accidents are scarce: with support of 0.141526 and 0.011596, respectevely. In order to be able to construct rules that include this scarce items, the $minSup$ and $minConf$ in the *apriori* algorithm have to be small. This implies necessarily that the search space and the result set grow.

To find the rules with "high" lift, while reducing the search space and the number of results, we fix the consequent of the rule to be one of the three severities *fatal*, *serious* and use the *absolute difference of lift value to 1* measure detailed in section 2. Then, we use the *improvement in confidence* measure to discard the redundancy while maintaining the generality of the rules.

5.1 Accident_Severity=serious

To find the rules $X \to \{Accident\_Severity = serious\}$ we parametrize the *apriori* algorithm with $minSup = 0.001$, $minConf = 0.1$, $minDLift = 0.2$ and $maxLen = 6$. A total of 10,079 rules where found with *apriori* and after removing redundancy, with *improvement in confidence* greater than 0.025, 2,189 rules were left. The top rules ordered by $lift$ are presented in the table 10.

The interpretation of the rules in table 10 is straightforward, for example, for the first rule we can say that accidents with serious severity are 86% more likely to occur when there is only one vehicle involved, the speed limit is 60mph and the road suface is dry. This can be due to the fact that when the road conditions are fine, drivers tend to drive cautionless. Rule 8 is telling us that serious accidents happen with 39% more confidence when the speed limit is 60mph, which happens to be in rural zones (see equivalence in table 7); in this sense, rules 1 and 2 are telling the same story. One interesting rule is number 3, which appears to involve one vehicle and

rule	lhs	rhs	sup	conf	lift
1	{Number_of_Vehicles=1, Speed_limit=60, Road_Surface_Conditions=dry}	{Accident_Severity=serious}	0.00639	0.2635	1.8619
2	{Number_of_Vehicles=1, Road_Surface_Conditions=dry, Urban_or_Rural_Area=rural}	{Accident_Severity=serious}	0.01435	0.2430	1.7170
3	{Number_of_Vehicles=1, Pedes_Cross.Facilities=pelican/no_light}	{Accident_Severity=serious}	0.00525	0.2346	1.6578
4	{Road_Type=single_carriageway, Junction_Detail=not_junction_20m, Road_Surface_Conditions=dry, Urban_or_Rural_Area=rural}	{Accident_Severity=serious}	0.01914	0.2295	1.6222
5	{Road_Type=single_carriageway, Junction_Control=no_info, Road_Surface_Conditions=dry, Urban_or_Rural_Area=rural}	{Accident_Severity=serious}	0.01917	0.2295	1.6216
6	{Number_of_Vehicles=1}	{Accident_Severity=serious}	0.06101	0.1991	1.4071
7	{Light_Conditions=darkness-no_lighting}	{Accident_Severity=serious}	0.01040	0.1978	1.3976
8	{Speed_limit=60}	{Accident_Severity=serious}	0.02839	0.1969	1.3916
9	{X1st_Road_Class=B, X2nd_Road_Class=no_info}	{Accident_Severity=serious}	0.00993	0.1922	1.3587
10	{X1st_Road_Class=B, Junction_Detail=not_junction_20m}	{Accident_Severity=serious}	0.00968	0.1922	1.3581

Table 10: Top 10 rules with higher *lift* for the consequent $\{Accident\_Severity = serious\}$.

pedestrians with no crossing light[3].

In general, we can conlcude that factors like greater speed limit, road surface dry and bad lighting conditions contribute to higher severity in accidents.

5.2 Accident_Severity=fatal

The parameters used in this case were $minSup = 0.001$, $minConf = 0.01$, $minDLift = 0.2$ and $maxLen = 6$. Finding 7,766 rules from which 9 were left after removing the redundant ones with *improvement in confidence* greater than 0.01. The rules ordered by $lift$ are presented in the table 11.

rule	lhs	rhs	sup	conf	lift
1	{Number_of_Vehicles=2, X1st_Road_Class=A, Speed_limit=60, X2nd_Road_Class=no_info}	{Accident_Severity=fatal}	0.00100	0.0573	4.9447
2	{X1st_Road_Class=A, Light_Conditions=darkness-no_lighting}	{Accident_Severity=fatal}	0.00121	0.0514	4.4389
3	{X1st_Road_Class=A, Speed_limit=60, Junction_Detail=not_junction_20m}	{Accident_Severity=fatal}	0.00192	0.0458	3.9559
4	{Light_Conditions=darkness-no_lighting}	{Accident_Severity=fatal}	0.00208	0.0396	3.4161
5	{Speed_limit=60}	{Accident_Severity=fatal}	0.00432	0.0300	2.5882
6	{Number_of_Vehicles=1, Road_Type=dual_carriageway}	{Accident_Severity=fatal}	0.00100	0.0283	2.4406
7	{Speed_limit=70}	{Accident_Severity=fatal}	0.00157	0.0224	1.9385
8	{Urban_or_Rural_Area=rural}	{Accident_Severity=fatal}	0.00777	0.0220	1.8999
9	{}	{Accident_Severity=fatal}	0.01159	0.0115	1.0000

Table 11: Rules ordered by *lift* for the consequent $\{Accident\_Severity = serious\}$.

In table 11 we can see that the speed limit of 60mph increments the probability of fatal accidents in 250% (rule 5). As in previous section we can also see that darkness contribute to the increase in confidence of having fatal accidents. The presence of the $Number\_of\_Vehicles = 2$ (rule 1) tells us that accidents between vehicles increases the confidence of having a fatal accident[4]. Another difference between the factors that explain severity seriuos and fatal is the

[3]Further analysis like filtering all the rules with same consequent and in which the left hand side of the rule is a superset of $\{Number\_of\_Vehicles = 1, Pedes\_Cross.Facilities = pelican/no\_light\}$ could help to prove this hypothesis.

[4]It would be interesting to analyse the kind of vehicles involved in this accidents.

road type, being, in the former, single carriageway and, in the later, dual carriageway.

6 Conclusions

In sections 3 and 4 several rule interestingness measures were used to analyse the data without having any specific objective in mind, in other words, an exploratory analysis were conducted over the data. The majority of the rules obtained with this approach don't provide any knew knowledge —they describe relations that could be drawn from the common sense, for example $rain \rightarrow wet\_road$. Nevertheless, some interesting results —which we weren't looking at— could be found, as the ones described at the end of section 4.2.

On the other hand, in section 5, the objective of the analysis was clear: find the variables that contribute to the severity of the accidents. Following this last approach, we not only found the variables that contribute to the severity of the accidents, but we also found the factor (number) by which this variables influence the severity.

The difference between the results obtained with the exploratory analysis and the objective-driven analysis is that, in the later, the results are important for decision making, but not very surprising; while in the former, the results represent completely new knowledge. Hence, the relevance of the exploratory analysis and the rule interestingness measures that help in this task.

Another conclusion that can be drawn from this work is the importance of measures, as the *absolute difference of lift value to 1*, that help in reducing the search space for the *apriori* algorithm and the number of resulting rules, specially when the events that are being studied are scarce.

References

[1] J. Blanchard, F. Guillet, and P. Kuntz. Semantics-based classification of rule in- terestingness measures. *Post-Mining of As- sociation Rules: Techniques for Effective Knowledge Extraction,IGI Global*, pages 56–79, 2009.

[2] Borgelt. *Apriori documentation*. http://www.borgelt.net/doc/apriori/apriori.htmldiff, 2010.

[3] M. Hahsler, C. Buchta, B. Gruen, K. Hornik, and C. Borgelt. *arules: Mining Association Rules and Frequent Itemsets*. http://R-Forge.R-project.org/projects/arules/, 2014.

[4] M. Hahsler, B. Grun, K. Hornik, and C. Buchta. Introduction to arules a computational environment for mining association rules and frequent item sets in r. Technical report, Southern Methodist University, 2005.

[5] R. J. B. Jr., R. Agrawal, and D. Gunopulos. Constraint-based rule mining in large, dense databases. In *Proc. of the 15th Int Conf. on Data Engineering*, pages 188–197, 1999.

A Itemsets

itemset	Support	Lift	Jaccard
{Junction_Detail=not_junction_20m, X2nd_Road_Class=no_info, Junction_Control=no_info, Pedestrian_Crossing.Human_Control=none_50m, Pedestrian_Crossing.Physical_Facilities=none_50m, Urban_or_Rural_Area=rural}	0.2017453	11.23122	0.06013522
{Junction_Detail=not_junction_20m, X2nd_Road_Class=no_info, Junction_Control=no_info, Pedestrian_Crossing.Physical_Facilities=none_50m, Urban_or_Rural_Area=rural}	0.2019111	11.18606	0.08556645
{Junction_Detail=not_junction_20m, X2nd_Road_Class=no_info, Junction_Control=no_info, Pedestrian_Crossing.Human_Control=none_50m, Pedestrian_Crossing.Physical_Facilities=none_50m, Weather_Conditions=fine_no_high_winds, Road_Surface_Conditions=dry, Urban_or_Rural_Area=rural}	0.1138685	11.17481	0.02340488
{Junction_Detail=not_junction_20m, X2nd_Road_Class=no_info, Junction_Control=no_info, Pedestrian_Crossing.Physical_Facilities=none_50m, Weather_Conditions=fine_no_high_winds, Road_Surface_Conditions=dry, Urban_or_Rural_Area=rural}	0.1139838	11.13201	0.02945320
{Junction_Detail=not_junction_20m, X2nd_Road_Class=no_info, Junction_Control=no_info, Pedestrian_Crossing.Human_Control=none_50m, Pedestrian_Crossing.Physical_Facilities=none_50m, Weather_Conditions=fine_no_high_winds, Road_Surface_Conditions=dry, Special_Conditions_at_Site=none, Urban_or_Rural_Area=rural}	0.1101399	11.05649	0.01885064
{Junction_Detail=not_junction_20m, X2nd_Road_Class=no_info, Junction_Control=no_info, Pedestrian_Crossing.Physical_Facilities=none_50m, Weather_Conditions=fine_no_high_winds, Road_Surface_Conditions=dry, Special_Conditions_at_Site=none, Urban_or_Rural_Area=rural}	0.1102481	11.01379	0.02274279
{Junction_Detail=not_junction_20m, X2nd_Road_Class=no_info, Junction_Control=no_info, Pedestrian_Crossing.Human_Control=none_50m, Pedestrian_Crossing.Physical_Facilities=none_50m, Special_Conditions_at_Site=none, Urban_or_Rural_Area=rural}	0.1934083	11.01373	0.04464161
{Junction_Detail=not_junction_20m, X2nd_Road_Class=no_info, Junction_Control=no_info, Pedestrian_Crossing.Physical_Facilities=none_50m, Special_Conditions_at_Site=none, Urban_or_Rural_Area=rural}	0.1935670	10.96942	0.05800096
{Junction_Detail=not_junction_20m, X2nd_Road_Class=no_info, Junction_Control=no_info, Pedestrian_Crossing.Human_Control=none_50m, Pedestrian_Crossing.Physical_Facilities=none_50m, Carriageway_Hazards=none, Urban_or_Rural_Area=rural}	0.1933867	10.95777	0.04458637
{Junction_Detail=not_junction_20m, X2nd_Road_Class=no_info, Junction_Control=no_info, Pedestrian_Crossing.Human_Control=none_50m, Pedestrian_Crossing.Physical_Facilities=none_50m, Weather_Conditions=fine_no_high_winds, Road_Surface_Conditions=dry, Carriageway_Hazards=none, Urban_or_Rural_Area=rural}	0.1093755	10.92518	0.01870418

Table 12: Top 10 most frequent itemsets ordered by *lift*.

B Itemsets

lhs	rhs	sup	conf	lift
{X1st_Road_Class=A, Junction_Control=give_way_or_uncontrolled, Pedest_Cross.Physical_Facilities=none_50m, Weather_Conditions=fine_no_high_winds, Road_Surface_Conditions=dry, Carriageway_Hazards=none}	{Light_Conditions=daylight}	0.1016	0.8146	1.1111
{Weather_Conditions=raining_no_high_winds}	{Road_Surface_Conditions=wet_or_damp}	0.1028	0.9797	3.6979
{Number_of_Vehicles=2, X1st_Road_Class=unclassified, Weather_Conditions=fine_no_high_winds}	{Urban_or_Rural_Area=urban}	0.1045	0.8005	1.2371
{Number_of_Vehicles=1, Light_Conditions=daylight, Urban_or_Rural_Area=urban}	{Road_Surface_Conditions=dry}	0.1073	0.8011	1.1436
{Number_of_Vehicles=1, Speed_limit=30, Light_Conditions=daylight, Special_Conditions_at_Site=none}	{Road_Surface_Conditions=dry}	0.1094	0.8003	1.1425
{Accident_Severity=slight, Road_Type=single_carriageway, Junction_Detail=T_junction, Pedest_Cross.Physical_Facilities=none_50m, Weather_Conditions=fine_no_high_winds, Road_Surface_Conditions=dry}	{Light_Conditions=daylight}	0.1104	0.8146	1.1111
{Number_of_Vehicles=2, Accident_Severity=slight, Junction_Detail=T_junction, Road_Surface_Conditions=dry}	{Light_Conditions=daylight}	0.1113	0.8170	1.1144
{Number_of_Vehicles=2, Road_Surface_Conditions=dry, Urban_or_Rural_Area=rural}	{Light_Conditions=daylight}	0.1114	0.8494	1.1585
{Number_of_Vehicles=1, Pedest_Cross.Physical_Facilities=none_50m, Urban_or_Rural_Area=urban}	{Road_Type=single_carriageway}	0.1116	0.8409	1.1122
{Number_of_Vehicles=2, X1st_Road_Class=unclassified, Light_Conditions=daylight}	{Pedest_Cross.Physical_Facilities=none_50m}	0.1126	0.9057	1.1125

Table 13: Top 10 most specific rules

C R code

```
#######################################
# Data processing
#######################################
# Data loading
accident <- read.csv("../data/DfTRoadSafety_Accidents_2013.csv")
#vehicle <- read.csv("../data/DfTRoadSafety_Vehicles_2013.csv")

# Exploratory
summary(accident)

# Data selection
variables <- c("Number_of_Vehicles","Accident_Severity", "X1st_Road_Class", "Road_Type",
"Speed_limit", "Junction_Detail", "X2nd_Road_Class", "Junction_Control","Pedestrian_Crossing.Human_Control",
"Pedestrian_Crossing.Physical_Facilities", "Light_Conditions", "Weather_Conditions",
"Road_Surface_Conditions","Special_Conditions_at_Site", "Carriageway_Hazards",
"Urban_or_Rural_Area")

accident_var <- accident[,variables]
accident_var[] <- lapply(accident[,variables], as.factor)

summary(accident_var)
accident_trans <- as(accident_var, ""transactions")

#######################################
#ITEMSETS
#######################################

itemsets_l1 <- apriori(accident_trans,
                       parameter=list(supp=0.001,target=""frequent itemsets", maxlen=1))

itemsets <- apriori(accident_trans,
                                        parameter=list(supp = 0.1, target = ""frequent itemsets",maxlen=10))


quality(itemsets) <- cbind(quality(itemsets), lift=interestMeasure(itemsets,method=""lift",
transactions=accident_trans,reuse = TRUE))

# Jaccard index:
acum <- rep(0, length(itemsets))
for(i in 1:length(itemsets_l1)) {
        sup <- quality(itemsets_l1[i])$support
        acum <- acum + as.vector(is.subset(itemsets_l1[i], itemsets)) * sup
}
quality(itemsets) <- cbind(quality(itemsets), jaccard = quality(itemsets)$support / acum)

#
itemsets <- sort(itemsets, by="support")
inspect(head(itemsets, 20))
#
itemsets <- sort(itemsets, by="lift")
inspect(head(itemsets, 10))
#
itemsets <- sort(itemsets, by="jaccard")
inspect(head(itemsets, 50))
```

```
#######################################
#Rules
#######################################
rules <- apriori(accident_trans, parameter=list(supp = .1, conf=.8, target = ""rules",maxlen=7,ext=TRUE)),
arem="quot",minval=0.1))

ptm<-proc.time()
quality(rules)<-cbind(quality(rules),improvement=interestMeasure(rules,method="improvement", reuse = TRUE))
proc.time() - ptm

# rhs.support
quality(rules) <- cbind(quality(rules), rhs.support = support(rhs(rules), accident_trans))
# kulczynski
quality(rules) <- cbind(quality(rules), kulczynski = 1/2 * (quality(rules)$support /
quality(rules)$rhs.support + quality(rules)$support / quality(rules)$lhs.support) )
# causal
quality(rules) <- cbind(quality(rules), causal = 2 * quality(rules)$support + 1 -
quality(rules)$rhs.support - quality(rules)$lhs.support)

# add chi^2
quality(rules) <- cbind(quality(rules), chi2 = interestMeasure(rules, accident_trans,
 method = ""chiSquared",significance=TRUE,reuse=TRUE))
#add_ganascia
quality(rules)<-cbind(quality(rules),ganascia=2*quality(rules)$confidence-1)
#add_loevinger
loevinger<-((quality(rules)$confidence-support(rhs(rules),accident_trans))/
(1-support(rhs(rules),accident_trans))
quality(rules)<-cbind(quality(rules),loevinger=loevinger)

rr<-(quality(rules)$improvement)>0.05
rr[is.na(rr)]<-TRUE
rules_sub<-rules[rr]

rules_sub<-sort(rules_sub,by="confidence")
inspect(head(rules_sub,30))
rules_sub<-sort(rules_sub,by="support",decreasing=TRUE)
inspect(head(rules_sub,30))
rules_sub<-sort(rules_sub,by="support",decreasing=FALSE)
inspect(head(rules_sub,30))
rules_sub<-sort(rules_sub,by="kulczynski",decreasing=TRUE)
inspect(head(rules_sub,30))
rules_sub<-sort(rules_sub,by="causal",decreasing=TRUE)
inspect(head(rules_sub,30))
rules_sub<-sort(rules_sub,by="lift",decreasing=TRUE)
inspect(head(rules_sub,30))


quality(rules_sub)<-cbind(quality(rules_sub),loev_gan_diff=quality(rules_sub)$ganascia-
quality(rules_sub)$loevinger)
rules_sub<-sort(rules_sub,by="ganascia",decreasing=FALSE)
inspect(head(rules_sub,30))

inspect(head(sort(rules_sub,by="lift",decreasing=TRUE),300))


#######################################
#RULES_X->sev=serious
#######################################
rules<-apriori(accident_trans,
parameter=list(supp=0.005,conf=.1,target="rules", maxlen=6,
ext=TRUE, arem="quot", aval=TRUE, minval=0.2),
appearance = list( rhs = ""Accident_Severity=fatal",default="lhs"))

ptm<-proc.time()
quality(rules)<-cbind(quality(rules),improvement=interestMeasure(rules,
method="improvement", reuse = TRUE))
proc.time() - ptm

quality(rules) <- cbind(quality(rules), rhs.support = support(rhs(rules), accident_trans))

quality(rules) <- cbind(quality(rules), total_improvement = quality(rules)$confidence /
quality(rules)$rhs.support - 1)

quality(rules) <- cbind(quality(rules), abs_total_improvement = abs(quality(rules)$total_improvement))

#rules_sev3 <- rules
#save(rules, file="rules_sev_serious.RData")
#load(file="rules_sev_serious.RData")

# Remove subsets that don't contribute to confidence:
rr <-quality(rules)$improvement > 0.025
rr[is.na(rr)] <- TRUE
rules_sub <- rules[rr]
rules_sub <- sort(rules_sub, by="lift")
inspect(head(rules_sub, 20))

plot(rules_sub, method="grouped")
plot(head(rules_sub, 10), method="grouped", measure="diff")
plot(head(rules_sub, 10), method="graph", control=list(cex=.6), measure="lift")
#rules_df <- as(rules_sub, ""data.frame")
#ggplot(rules_df[1:20,c("rules","total_improvement")], aes(x=factor(rules), y=total_improvement*100)) +
geom_bar(stat = ""identity")


#######################################
#RULES_X->sev=fatal
#######################################
rules<-apriori(accident_trans,
parameter=list(supp=0.001,conf=.01,target="rules", maxlen=6,
ext=TRUE, arem="quot", aval=TRUE, minval=0.2),
appearance = list( rhs = ""Accident_Severity=fatal",default="lhs"))

ptm<-proc.time()
quality(rules)<-cbind(quality(rules),improvement=interestMeasure(rules,
method="improvement", reuse = TRUE))
proc.time() - ptm

quality(rules) <- cbind(quality(rules), rhs.support = support(rhs(rules), accident_trans))
```

```
quality(rules) <- cbind(quality(rules), total_improvement = quality(rules)$confidence /
quality(rules)$rhs.support - 1)

quality(rules) <- cbind(quality(rules), abs_total_improvement = abs(quality(rules)$total_improvement))

#rules_sev3 <- rules
load(file="rules_sev_fatal.RData")
#save(rules, file="rules_sev_fatal.RData")

rr <-quality(rules)$improvement > 0.01
rr[is.na(rr)] <- TRUE
rules_sub <- rules[rr]
rules_sub <- sort(rules_sub, by="lift")
inspect(head(rules_sub, 20))
```