Application of Data Mining: Case of Road Accidents in the UK West Midlands Area

Mohamad Saraee, Jonathan Kerry , Michelle Lloyd and Christine Markey

School of Computing, Science and Engineering, Salford University, Salford, Manchester M5 4WT, UK
emails: m.saraee@salford.ac.uk, j.kerry@student.salford.ac.uk,
m.a.lloyd@student.salford.ac.uk, c.l.markey@student.salford.ac.uk

Abstract
We are aiming to use data mining techniques in the analysis of data recorded about road traffic accidents in the UK West Midlands Area in the year 2000. This data will then hopefully provide drivers with guidelines relating to what measures can be taken to help reduce the chances of them being injured in a road traffic accident. This analysis is therefore important in order to identify potential risks and circumstances which contribute to such accidents, and attempt to highlight measures which can be taken to minimise them. The data is currently in an unmanageable format which hinders the investigation of finding specific links between attributes. This means that no useful conclusions can be accurately drawn at present. We therefore intend to complete the analysis by using Envisioner software and classification techniques to determine attributes of high relevance within the data. Conclusions will then be drawn, helping to identify factors such as speed, weather and road conditions which contribute to an accident occurrence.

Keywords
Traffic Accidents, West Midlands, Injuries, Classification

1. Introduction

To identify important contributing factors to road accidents in Britain we have obtained a large dataset of every accident recorded in the West Midlands area in the Year 2000. The data is currently in an unsorted and unmanageable format and is stored all together in a Microsoft Access database table. Unfortunately with the data in its current format, no relevant points or conclusion can be drawn. It is hoped that by applying data mining processes and techniques to the dataset, relevant attributes and patterns can be established.

The main achievements of this work are as follows:

- greater awareness of the conditions affecting road traffic accidents.
- establishing which individuals are most likely to be involved in a road traffic accident.
- which vehicle types reduce the chance of fatality in an accident.
- justification of insurance premiums on age and gender.
- establish trends in the data to derive the above information.

The organisation of the papers is as follow: In section 2 we present a formal statement of the problem which describes the problem domain and the methods we intend to use to solve the problem. Next we present the relevant background research which was conducted to provide us with a firm theoretical base to the investigation. The methodology and techniques we used to interrogate the dataset are described in section 3 of the paper. Section 4 documents the experiments we conducted and the results we gained. Finally the conclusions to the problem are discussed in section 5.

2. Problem Statement

The purpose of this investigation is to reduce the number of road accidents in Britain by finding risks and circumstances which can be shown to be regular contributing factors to road accidents. If the major contributors can be established, these can be published to make people more aware of when they are potential at risk of an accident, and allow them to avoid these risks when possible. To establish the most common factors and circumstances related to road accidents we obtained a large dataset which records, in detail, every reported road accident that occurred in the West Midlands area in the year 2000. For each accident, 52 different attributes are recorded detailing information about all aspects of the crash such as where and when the accident occurred, how many cars and people were involved, the gravity of any injuries sustained, the make and model of the cars involved and any environmental factors such as weather, lighting and road conditions. A full list of all the attributes recorded can be found in Appendix B.

There are a total of 9297 different records in the dataset which calculates to 483444 pieces of information in either text or numerical formats. However the information in the dataset is relatively unsorted and is therefore in an unmanageable format, meaning no conclusions can be gained regarding road accidents.

We therefore intend to us the data mining software Envisioner and Microsoft Access to interrogate the dataset and extract any patterns, rules and constraints which can be used to establish the main contributing factors to road accidents. We plan to find attributes which have high relevance to road accidents and then construct association rules with which we can achieve a high level of support and therefore have a high level of confidence.

2.1 Related Work

There have been many studies carried out on the subject of road traffic accidents on Britain's roads. The motivation for these studies has covered a wide range of reasons. The medical profession has analysed the number of casualties admitted to hospital due to roads traffic accidents (RTA's) to determine the true extent of the effect of these accidents on Britain's health services. It is estimated that the annual cost to the NHS due to road traffic accidents is over £1billion per year [1]. This is a tremendous amount of money which could be used in many other medical areas. Indeed it has been noted by the government that accident prevention could save enough money to pay for over 230, 000 hip replacement operations. On a larger scale the World Health Organisation, in conjunction with the World Bank, concluded from their analysis that over 1million people die worldwide each year as a result of road traffic crashes and collisions and that by 2020 road traffic accidents could overtake HIV and Tuberculosis to rank third in the causes of premature death and disability around the world [1].

The Police force also has an interest in this area, carrying out studies related to various traffic offences which are committed, one of their top priorities being that of drink driving. In the West Midlands area 195 drivers out of 2,300 failed tests in 2001[2]. Although Police analysis's into the area have shown a drop in the number of drivers testing positive, it is a fact that drink driving causes an increased risk of road accidents and therefore a continued reduction is required if traffic accidents are to be reduced.

The Highways Agency estimate that ¼ of all road congestion is a result of a road traffic accident, congestion being another factor which has been considered and researched. In an attempt to combat this problem a 'new emergency service' [3] has recently begun trials in the West Midlands area with a priority to keep traffic moving as well as possible in the event of an accident. These new Traffic officers will help to improve services to motorists and minimise congestion caused by accidents [4].

Other studies conducted include those by environmental organisations to help reduce pollution by reducing traffic congestion and also studies carried out by the motor industry to determine and help improve vehicle safety.

From the studies conducted there have been several suggestions to aid the reduction of traffic accidents including increasing road safety awareness, traffic calming schemes and traffic monitoring to access the effectiveness of implemented techniques [5].

Prevention of motor accidents is undoubtedly the ultimate goal, prevention is after all better than cure, and the benefits of this achievement would be reaped by a wide range of people and organisations, not least those individuals involved.

In an attempt to draw conclusions to measures which could be taken to reach this goal data mining techniques have been used on the data to highlight data patterns and relevant areas of research. Research has shown that the data used in this work is indeed a good starting point for this analysis. The West Midlands area has the highest population (just over 2.6 million), the highest daily traffic flow on its major roads and also the highest number of cars registered in the UK [6]. These facts make it an excellent area to conduct this area of research and obtain realistic results.

Classification was the chosen data mining technique for this work. This analysis option allows previously unseen data trends to be found by grouping data into classes according to a common characteristic [6]. The technique is widely used in areas such as fraud detection, credit-risk applications and the medical profession for diagnosis [7]. The success of classification is due to its flexibility and capabilities surrounding the processing of large amounts of data [8]. Additionally, the method allows easily interpreted output in the form of graphical representation like decision tress, rather than complicated mathematical formula like that of the regression technique of data mining. This analysis method is therefore ideal for identification of data patterns which can then be used in a predictive manner [9] and satisfies the needs of this work.

3. Methodology

To familiarise ourselves with the dataset we began by compiling a set of questions which we felt were relevant to the study. These included queries such as what percentage of accidents resulted in a fatality, how many accidents involved children, what types are vehicles had the most accidents when travelling greater than 50 mph, which gender age group is more likely to have an accident and which four attributes contributed to accidents involving more than three vehicles. To find answers to these questions and progress the study further we used a mixture of Microsoft Excel, Access and Envisioner software. While Microsoft Excel and Access were used initially to carry out various queries on the data, Envisioner was used as the work progressed to carry out various classification techniques. These included association rules, relevance attributes and tree classifiers. This analysis aimed to highlight the data of the most importance in a road traffic accident and allow predictions to be made via the classification techniques. The results from this methodology

can be seen in the next section of the report.

4. Experimental Evaluation

The basic setup for the understanding of the data was to try and get a better appreciation of the content of the data and also a greater knowledge of the trends within it.

The first area to be evaluated is quite simply the percentage of fatalities within the data. From this we would be able to gauge the human cost in comparison to the total number of accidents. The results for this are achieved by running a filter on the data which restricts the severity type to 1 which symbolises a fatality. Please note that a fatality is recorded in the data and represents deaths which occurred on the scene or as a result of the accident within 30 days.

Similarly to the first area we also thought that it would be beneficial to look at what impact these accidents had on children. The data will be interrogated to see what percentage of the accidents involved 1 or more children. The result for this is gathered by using an SQL query over the data. Within this query the causality age was limited to greater than 0 and less than or equal to 16, which is the standard age range used in the definition of a child. Again, similar to the first area, this would increase our knowledge not only of the data but also of the severity of the actual accidents that had occurred.

It has also been decided that to try and silence the critics it would be of interest to see what gender age group was involved in the most accidents. It is not expected that the results from this will be able to provide any information to educate drivers, but it may be interesting to see if insurance companies are justified in charging higher premiums for different gender age groups. Again this result is derived from a series or queries, which restricted the driver ages into specific groups. These where then compare to the total number of drivers in that gender group in order to get a percentage.

To try and expand our knowledge the next step was to try and look at the data in more depth to see whether there are any trends in the types of vehicles that where involved in the accidents. Hopefully this would provide the type of information which could then be compared to registered vehicles within the area to see if there was a particular vehicle type which was involved in more accidents than others. The reason for looking at registered vehicles is to ensure that the high number for one vehicle type is not just because there are a larger volume in this particular area, and not because there is a possible safety problem. Again an SQL query is used to select only data which related to cars. This was then filtered by the different car makes so that the total number of accidents they where involved in can be derived.

The final area that will be looked at in this section will again hopefully increase knowledge of the subject, ready to complete a classification on the data. It was decided that it would be advantageous to look at a common occurrence on today's roads, multiple car accidents. For this the data will be interrogated to find the relevant attributes that contribute to accidents involving 3 or more vehicles. For this Envisioner will be used to look at the relevant attributes to the number of vehicles. However instead of running this over the whole data set though it will be executed over a query which will restrict the data to only have accidents involving the 3 or more vehicles as outlined previously. The standard setup within Envisioner will not be changed however there will be certain fields that will be excluded from the relevance. This includes reference information and also other information like cause or passenger details which have no direct affect on the accident.

From this the work will then turn its attention to totally using Envisioner and to performing a classification on the data. It has been decided that the classification should focus on the severity of the accidents, because fundamentally this is the most important part in the whole situation. To start the relevant attributes to the severity of the accident will be found. Again, similarly to previous times when the relevance of attributes has been found, certain attributes are not included because they have no relevance to the data. This includes the reference, causes and vehicle number for example.

From this association rules when then be located, which will hopefully help to meet the objectives of this work. For the association rules there are some more settings which will need to be set. This is the minimum support and confidence required for the rule to be included. It was decided that the minimum support for the rules should be 10% and that the minimum confidence should be 70%. Alike to the setup of the relevant attributes there are also fields which are not required to be included when the rules are derived. These included the reference, causes and vehicle number for example.

Finally a classification of all the data will be calculated based on the severity of the accident. The Entropy index will be used in the creation of the classifier and again attributes will be excluded that are deemed to be of no relevance.

4.1 Results

Within our initial research into the data the following results have been found. For the question relating to the total number of fatalities, it was found that 24 of the accidents resulted in a fatality. This is out of a possible 4505 total accidents which gives a percentage of fatalities as shown below:

$$(24 / 4505) \times 100 = \underline{0.53\%}$$

For the question relating to the number of accidents involving injury to children, it was found that 991 of the casualties where below the age of 17. This then allowed the percentage of accidents involving children to be calculated as follows:

$$\text{Casualty Age} <= 16 \text{ and } > 0 = 991$$

$$(991 / 4505) \times 100 = 21.99\%$$

For the question relating to the driver gender age group involved in the most number of accidents, the results found that males aged between 30 and 40 years of age where involved in significantly more road accidents than any other group. The full results of this are shown in Figure 1.

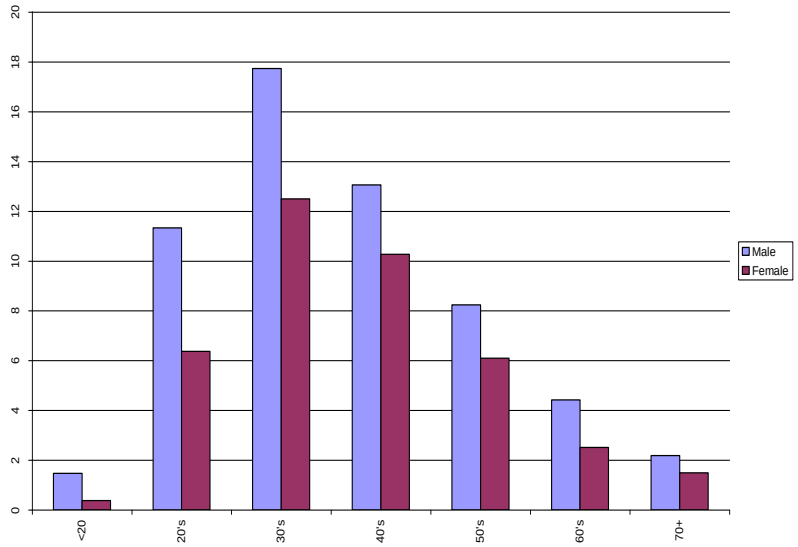

Figure 1: Driver gender age group involved in the most number of accidents

For the question relating to the types of vehicles which where involved in the accidents where the speed limit was greater than or equal to 50 mph, found that there was a total of 255 accidents, which involved 184 cars. The separation of these car makes is shown in Figure 2 below:

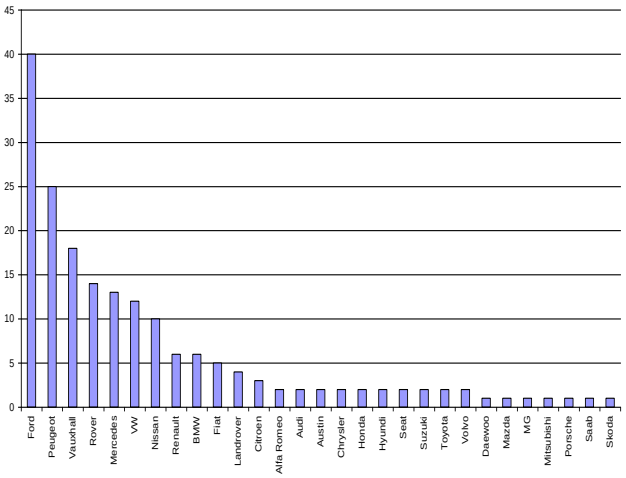

Figure 2: The separation of the car makers

For the question relating to the four attributes linked with accidents involving 3 or more vehicles, the relevant attributes found are shown in Figure 3.

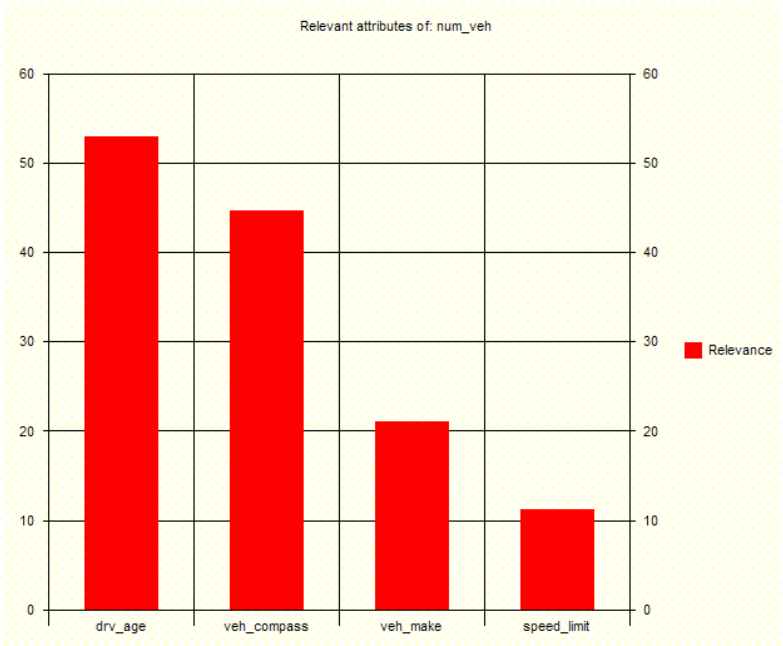

Figure 3: Relevant attributes linked with accidents involving 3 or more vehicles

As explained within the setup section the next stage of the work was to locate the relevant attributes for the severity of the accident. The result is shown in Figure 4.

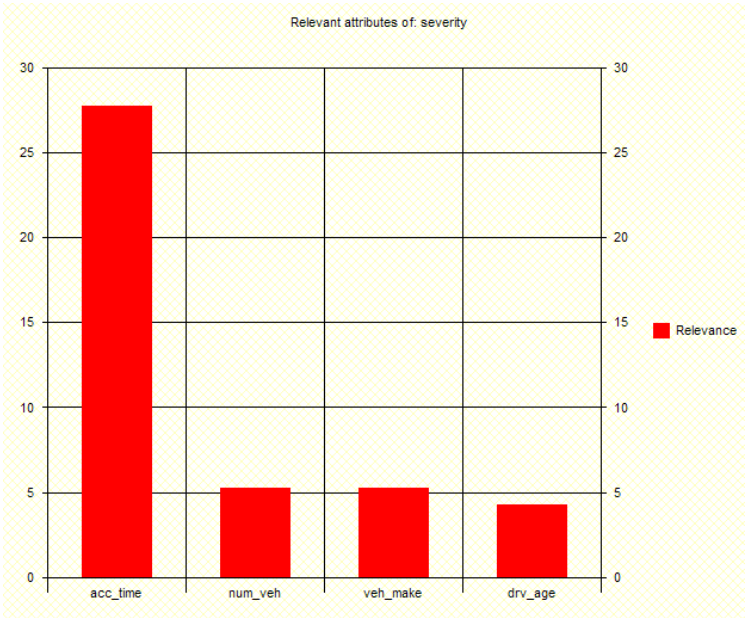

Figure 4: Relevant attributes for the severity of the accident.

When looking at the association rules that could be found within the data a couple stood out that where more significant in comparison to the others. These were:
IF vehicle = Ford
severity = 3 (minor injury)
IF driver age < 25
severity = 3 (minor injury)
IF number vehicles = 2 AND vehicle make = Rover OR Peugeot severity = 3 (minor injury)

The last section was to create a classifier for the data. The results of this can be viewed in Appendix A.

4.2 Discussion

From our initial research we have found a number of interesting facts, some of which where expected and some which where quite surprising.

From the 4505 individual accidents there were a total of 24 fatalities. Although it would be much better if this figure was 0, we feel that this level is reasonably acceptable considering the large volumes of traffic which are recorded for the West Midlands area.

The query looking at the number of accidents which involved injury to children returned a surprising high value. Approximately 22% of all the injured where up to and including 16 years or age. This is quite a large figure considering that no-one in this age range would be able to legally drive and therefore would either be injured as a passenger or a pedestrian. From reading further into this subject we have also found that the majority of injured within this category where pedestrians. This leads to the argument that to reduce this figure it would be necessary to better educate children, possibly within schools, about road safety, which would increase their awareness of road safety.
The age gender found to have had the most accidents was males aged 30 to 40. This is a surprising result as it is the general perception that males aged 17 to 25 are more likely to have an accident and are subsequently charged higher insurance premiums. Overall, it was found that there were 774 more accidents involving male drivers in comparison to those involving female drivers. This is surprising since it is stereotypically female drivers who are considered to be less competent than male drivers.

The car type found to be involved in the most accidents was Ford, by a considerable amount. This figure may be justified by the levels of car makes registered in the West Midlands area, i.e. the likeliness is that there will be a higher number of Ford's than other makes. Unfortunately, although the DVLA hold this information it is not published for public use and so no reliable connection can be drawn to support any argument these high numbers.

The final question within this section was to attain the relevant attributes which resulted in accidents involving 3 or more vehicles. Connections can be drawn from the results of this, especially in the fact that driver age and speed limit can be linked. It can be assumed that drivers or working age (20 to 60 for example) may be more likely to use a motorway, which accordingly means the speed limit will be increased, and so increase the possibility of multiple vehicles being involved in the accident.

As previously explained our focus moved from knowledge building to classifying the data, so that previously unseen data could be found within it. To start it was decided that the relevant attributes should be found in relation to the severity of the accident. As shown within the results section, the most relevant attribute was the accident time. This can be explained by the increased traffic flow during rush hour periods. More cars on the road increasing the chances of an accident occurring. This was also found within some of the research carried out on the subject area, and also that during weekends traffic levels are highest in the time between 10am and 12pm. Two attributes that we where surprised not to see in these results where weather and lighting, which are normally viewed as being large factors in many road traffic accidents. One reason for this might be because in adverse conditions drivers are more likely to con centrate harder therefore reducing the possibility of an accident.

With the relevant attributes found it was then possible to find association rules within the data, which could then be tested to see how accurate they where. As you can see from the rules outlined in the results section there are vehicles which can be assumed to be safer than others. In particular Ford and Vauxhall who both have high numbers of crashes recorded, yet have good records for injury severity. To test this, the test data shown in Figure 4 will be used:

Query1 : Select Query

reference	acc_date	acc_time	severity	num_cas	speed_limit	drv_age	breath_tst	veh_make
BRED13901	20000603	1815	3	1	30	26	2	FORD
BRED13211	20000524	910	3	1	30	58	5	FORD
BRED13222	20001201	2310	3	3	30	40	2	FORD
BRED13281	20000521	2020	3	2	30	31	2	FORD
BRED13311	20000428	1305	3	1	30	50	2	FORD
BRED13351	20000426	2229	3	1	30	40	5	FORD
BRED13352	20001204	145	3	2	40	19	2	FORD
BRED13371	20000526	1540	3	1	30	61	5	FORD
BRED13371	20000526	1545	3	1	30	37	2	FORD
BRED13411	20000427	2100	3	3	30	58	5	FORD
BRED13481	20000525	355	3	1	40	21	5	FORD
BRED13651	20000526	1115	3	1	30	39	2	FORD
BRED13751	20000508	654	3	1	40	35	2	FORD
BRED13751	20000529	2258	3	2	30	25	5	FORD

Figure 5: Test data used for association rules mining

It has also been noticed that drivers under 25 years of age have no recorded fatalities. This is quite surprising because these types of incidents generate quite a lot of press coverage and so it is commonly assumed that younger drivers are at

greater risk than older ones. Again to test the rule the following test data shown in Figure 6 were used:

Query1 : Select Query

reference	acc_date	acc_time	severity	num_cas	speed_limit	drv_age	breath_tst	veh_make	cas_age
BRED00071	20000102	425	3	1	40	21	6	ROVER	21
BRED00931	20000102	2345	3	5	30	22	2	VW	21
BRED01071	20000107	2325	3	1	30	18	2	VOLKSWAGEI	33
BRED01211	20000111	2204	2	3	30	19	2	DAIHATSU	24
BRED01381	20000109	1715	3	1	30	18	2	FORD	18
BRED01461	20000111	145	2	2	30	21	2	VW	21
BRED01621	20000109	2135	3	2	30	24	2	VAUXHALL	24
BRED01641	20000113	1650	3	1	30	22	2	VW	22
BRED01891	20000117	1055	3	1	30	21	5	VW	21
BRED01901	20000111	1855	2	3	40	21	3	FORD	21
BRED01921	20000109	1310	3	1	30	23	5	TOYOTA	23

Figure 6: Test data used for association rules mining

Finally, as previously outlined, the data has been classified (Appendix A) to create a graphical representation of the data. As with the previous analysis methods the make of the vehicles involved can be see to an important factor. Where accidents involved less than 1.50 vehicles, i.e. single vehicle incidents, there were particular vehicle makes that resulted in fatal or serious injures in 100% of the cases. These makes included DAF, Daihatsu and Hyundai, suggesting that either the drivers of these types of vehicles are more reckless than others or that the safety of these vehicles is not of a high standard in comparison to other vehicles.

It was also noticed that the car make of taxi, which we presume to include all cars used as taxis, was only involved in multiple vehicle accidents. This is consistent with the levels of taxi's operating in city centres where there is a higher volume of traffic.

Although the classification showed that breath tests were relevant in multiple vehicle accidents, no clear conclusions can be drawn regarding the outcome of the tests. However, we did find it interesting that this attribute was not found to be relevant when only a single vehicle was involved. It is commonly thought that drink driving contributes to a high number of this type of accident which this evidence does not support.

Finally, it was noticed that where a single vehicle was involved and there were multiple casualties, particular car makes resulted in fatal and serious injuries in 100% of cases. Theses makes include Audi, Daewoo and Mazda. This is surprising as these are commonly used as family vehicles and considered safe because of their size. Our research and the results of the classifier have shown that contrary to this smaller cars such as the Mini Cooper have better safety records.

5. Conclusion

In conclusion the data mining techniques used in this work have resulted in interesting findings from previously unmanageable data. Many of the findings disprove general perceptions of road traffic accidents, for example the surprising results regarding gender age group, driving conditions and car sizes. Other results have been as expected such as an increase in accidents during rush hour periods.

One of our main aims within this work was to help provide more awareness of the conditions regarding road traffic accidents and possible prevention methods. One solution that we have concluded from our results is that refresher courses should be provided for drivers every 10 – 15 years. This would hopefully help reduce the number of accidents within older age groups by giving them instruction on driving methods and techniques up to date with the current driving conditions.

It is also hoped that the results that we have found will be of some use in future investigations in this area. Hopefully research alike to this will help in the national campaign to reduce the number of road traffic accidents.

7. References

[1] - http://politics.guardian.co.uk/homeaffairs/story/0,11026,1187637,00.html
[2] - http://www.cwn.org.uk/999/west-midlands-police/2001/01/010109-drink-driving.htm

[3] - http://www.highways.gov.uk/news/misc/2004_04_26_darling.htm

[4] -http://www.smbc.sandwell.gov.uk/sandwelldirect/highwaysdirect/ safety/trafficcalming.htm

[5] - http://www.dft.gov.uk/stellent/groups/dft_transstats/documents/page/dft_transsats_505711.hcsp

[6] - www.mnmodel.dot.state.mn.us/glossary.html

[7] - http://www.dbmsmag.com/9608d53.html

[8] - http://databases.about.com/library/weekly/aa100700a.htm

[9] - http://moneycentral.msn.com/content/SavingDebt/P76537.asp