

Contents

Data Set Description	1
Attribute Descriptions.....	3
Data Pre-processing	4
Step 1	4
Step 2	5
Step 3	6
Step 4	6
Step 5	6
Step 6	7
Step 7	7
Step 8	7
Step 9	7
Step 10	7
Data Analysis	8
Investigation of midday peak in accidents	9
Casualty Severity & Casualty Type Classification	12
Yearly distribution of casualties	14
Summary	15
Reflection	16
Bibliography	17

Table of Figures

Figure 1 - Initial Visualisation	8
Figure 2 - Number of Casualties by Age Range	9
Figure 3 – Midday Accident Peak Cluster Identification	10
Figure 4 - Slight Casualties Against Hour of Accident	11
Figure 5 - Casualty Type Classification Errors.....	13
Figure 6 - Month of Accident against Slight Casualties.....	14

Data Set Description

The data set selected concerns the road traffic accidents and casualties in Great Britain for the years 2006 to 2008 which is published by the Department for Transport and released under the OGL, Open Government License (1), license. The raw original data files used are available from the following location <http://data.gov.uk/dataset/gb-fatal-reported-road-accidents-06-08> (2). The data has been allocated a three star rating from data.gov.uk and is generally in a very clean format, a small level of modification however will be required to load the data into Weka for further analysis.

It is comprised of two data sheets, both in CSV format; the first sheet contains the majority of the detail and it describes the accident itself while the second sheet contains more specific fatal casualty details. The sheets are related together through the use of a *Fatal Accident Index*.

The first sheet, known as the *Fatal Accident Data* sheet, which has a total of 7981 data rows contains the following attributes;

- Fatal Accident Index
- Month of Accident
- Hour of Accident
- Longitude
- Latitude
- Pedestrian Casualties
- Pedal Cycles
- Motor Cycles
- Cars
- Buses or Coaches
- Vans
- HGVs
- Other Vehicles
- Total Vehicles Involved
- Fatal Casualties
- Serious Casualties
- Slight Casualties
- Total Number of Casualties

The *Fatal Casualty Data* sheet which consists of 8656 data rows contains the following casualty related information;

- Fatal Accident Index
- Fatal Casualty Type
- Fatal Casualty Age
- Fatal Casualty Sex

Attribute Descriptions

The fatal accident index is a string which uniquely identifies an individual accident, the first four characters of this string indicate the year of the accident in four character, YYYY, format. This fatal accident index is used to identify the all of the casualties for an accident in the casualties' data sheet.

The month of the accident field contains the month of the accident in textual format; possible values are January, February, March, April, May, June, July, August, September, October, November, December.

The hour of accident field contains an integer value which identifies an hour timeslot in which the accident occurred, the possible values range from 0 to 23. I.e. an accident which occurred at 0030 would be assigned a value of 0 as it occurred after 0000 but before 0100.

The longitude attribute contains the longitude at which the accident occurred.

The latitude attribute contains the latitude at which the accident occurred.

The pedestrian casualties attribute contains the number of pedestrian casualties that occurred as part of the accident in question.

The Pedal_Cycles attribute contains the number of pedal cyclists involved in the accident.

The motor_cycles attribute contains the number of motor cyclists that were involved in the accident.

The cars attribute contains the number of cars involved in the accident in question.

The buses_or_coaches attribute contains the number of buses/coaches that were involved in the accident.

The Vans attribute contains the number of vans involved in the accident.

The HGVs attribute contains the number of HGVs included in the accident in question.

The other_vehicles attribute contains the number of other vehicles involved in the accidents. A release from the Department of Transportation describes the types of vehicles that are included in this attribute. These include emergency service vehicles, refuse vehicles, agricultural vehicles, mobility scooters, motorised wheelchairs (3). Non-motorised vehicles are also including animal drawn vehicles, wheelchairs without motors and ridden horses (3)

The total_vehicles attribute contains the total number of all the vehicles that were involved in the accident.

The Fatal_Casualties attribute contains the total number of fatal casualties that occurred due to the accident.

The serious_casualties attribute contain the number of serious casualties that occurred due to the accident. A serious injury is one which causes a person to become an "in-patient" at a hospital or causes death 30 or more days after the accident (3). There are also a number of different injuries such as internal injuries or fractures that can cause a casualty to be identified as serious.

The `slight_casualties` attribute contains the number of slight accidents that occurred as a result of the accident. Slight casualties are casualties which do not require further hospital attention.

The `total_number_of_casualties` attribute contains the total number of casualties regardless of severity that have occurred due to the accident.

The `Fatal_Casualty_Type` attribute contains a textual description for the type of casualties that has occurred. This attribute comes from the second data sheet and there is an entry for each fatal casualty that occurs at a single accident. Examples of the data in this field are *Pedestrian*, *Motor_Cycle_Rider* and *Pedal_Cyclist*.

`Fatal_Casualty_gender` contains a textual value which states the gender of the casualty as either *Male* or *Female*, there is a single data record in the fatal casualty data sheet which contains a value of *Not_Reported* for this attribute.

The `Fatal_Casualty_Age` attribute contains the age of the fatal casualty in question. This attribute also contains *Not_Reported* entries however in this case there are eighteen records that contain a `Fatal_Casualty_Age` of *Not_Reported*. Ages range from 0, indicating less than one year old to ninety nine.

Data Pre-processing

While the data is originally in a relatively clean format a number of pre-processing tasks were required to get the data to be processed in Weka. An SQL database was used to perform these pre-processing tasks for simplicity and as it was a familiar environment to work within efficiently.

Step 1

The first step was to combine the two sheets into a single data source; as noted above the two sheets are linked together via the string identifier *fatal_accident_index*. It is possible for there to be a single accident entry which relates to multiple fatal casualties, to keep track of occurrences of this an additional two attributes were added to the final data source. First a string value which identified whether this accident was a duplicate of another accident in the data sheet, possible values being *True* and *False*. Secondly an integer which identifies how many other entries are in the table refers to the same accident. These two fields are titled *IsDuplicate* and *DuplicateCount* respectively. An additional unique identifier was added at this point named *InstanceID*, which while not useful when performing any investigation of the data in Weka aids performance in the SQL database.

The data from each sheet was initially imported into separate table before being brought together into a single table using the following query.

```
insert into AccidentCasualtyData

select a.Fatal_Accident_Index,

        a.Month_of_Accident,

        a.Hour_of_Accident,

        a.Longitude,

        a.Latitude,
```

```

a.Pedestrian_Casualties,

a.Pedal_Cycles,

a.Motor_Cycles,

a.Cars,

a.Buses_or_Coaches,

a.Vans,

a.HGVs,

a.Other_Vehicles,

a.Total_Vehicles_Involved,

a.Fatal_Casualties,

a.Serious_Casualties,

a.Slight_Casualties,

a.Total_Number_of_Casualties,

c.Fatal_Casualty_Type,

c.Fatal_Casualty_Sex,

c.Fatal_Casualty_Age,

(select case when count(cs.Fatal_Accident_Index)>1 then 'True'

else 'False'

end

from CasualtyData cs where cs.Fatal_Accident_Index = a.Fatal_Accident_Index),

(select count(cd.Fatal_Accident_Index)-1 from CasualtyData cd where cd.Fatal_Accident_Index = a.Fatal_Accident_Index) from
AccidentData a inner join CasualtyData c on c.Fatal_Accident_Index = a.Fatal_Accident_Index

```

Step 2

Next the records containing *Not_Reported* values were addressed. The following query was run against the newly created data to determine how many records containing missing age/gender data.

```

select count(instanceid) from AccidentCasualtyData where (Fatal_Casualty_Sex = 'Not_Reported' or
Fatal_Casualty_Age = 'Not_Reported')

```

From this query 19 records were identified as being affected, which was as expected from the initial count done against the data within the raw CSV sheets, a query was run to extract the *Fatal_Accident_Index* for each of these 19 entries. From this a further query was created to find the number of records in the database that these *Fatal_Accident_Index*'s refer to, this allowed a true view of how many entries in the combined data source were actually affected by this missing data.

The result of this query showed that 23 records in total were related to this missing data, the additional four records being entries which related to a different fatal casualty at the same accident as a fatal casualty which was missing data. As this was such a small number of the complete data set

these accidents were removed from the data to allow for easier processing. This left a total of 8633 records in the data set for further processing.

Step 3

Next an additional field called *Month_Number_Of_Accident* was added to hold a numeric representation of the month ranging from 1 for January to 12 for December.

Step 4

Double quotes surrounding the *Fatal_Accident_Index* were then removed using the SQL function REPLACE to replace the " character with nothing.

Step 5

At this point the first candidate arff file could be produced to import into Weka, the following header declaration was created.

```
@RELATION VAC
```

```
@ATTRIBUTE InstanceID numeric
```

```
@ATTRIBUTE Fatal_Accident_Index string
```

```
@ATTRIBUTE Month_of_Accident numeric
```

```
@ATTRIBUTE Hour_of_Accident numeric
```

```
@ATTRIBUTE Pedestrian_Casualties numeric
```

```
@ATTRIBUTE Pedal_Cycles numeric
```

```
@ATTRIBUTE Motor_Cycles numeric
```

```
@ATTRIBUTE Cars numeric
```

```
@ATTRIBUTE Buses_or_Coaches numeric
```

```
@ATTRIBUTE Vans numeric
```

```
@ATTRIBUTE HGVs numeric
```

```
@ATTRIBUTE Other_Vehicles numeric
```

```
@ATTRIBUTE Total_Vehicles_Involved numeric
```

```
@ATTRIBUTE Fatal_Casualties numeric
```

```
@ATTRIBUTE Slight_Casualties numeric
```

```
@ATTRIBUTE Total_Number_of_Casualties numeric
```

```
@ATTRIBUTE Fatal_Casualty_Type { 'Pedal_Cyclist', 'Other_Vehicle_Occupant', 'Motor_Cycle_Rider',  
'Car_Driver', 'Motor_Cycle_Passenger', 'HGV_Passenger', 'Car_Passenger', 'Bus_Driver', 'Pedestrian',  
'HGV_Driver', 'Van_Passenger', 'Bus_Passenger', 'Van_Driver' }
```

@ATTRIBUTE Fatal_Casualty_Sex { Male, Female }

@ATTRIBUTE IsDuplicate { True, False }

@ATTRIBUTE DuplicateCount numeric

@data

The *Fatal_Accident_Index* field, which is of no real interest, is the only attribute that will be unsuitable for the vast number of analysis with ever other attribute either being a numeric value or being categorised. The casualty type attribute has a large number of possible values however while the possible values could have been condensed slightly the option have a large amount of relevance with respect to the other attributes of each record so they were left as is.

When attempting to load the produced arff file in Weka a problem was identified with the Casualty type values, specifically Pedal Cyclist, which needed to be resolved before the data could be loaded.

Step 6

The value *Pedal Cyclist* was updated to *Pedal_Cyclist* in the produced SQL database by once more using the replace function provided in SQL to replace any spaces within the *Fatal_Casualty_Type* column with an underscore.

Step 7

A second attempt of loading data into Weka was then undertaken, the same headers were used as above with the notable change of *Pedal Cyclist* to *Pedal_Cyclist* for possible categories in the *Fatal_Casualty_Type* definition.

This arff file loaded successfully into Weka however upon further examination the age attribute had not been added to this file.

Step 8

Missing age attribute was added to the arff file with the following addition to the header definition added

@ATTRIBUTE Fatal_Casualty_Age numeric

The casualty age value was also initially left as numeric however it is quite possible that this may in future be sorted into bins to smooth the data to aid further processing with in Weka. This data was then loaded into Weka and was then ready for analysis to be performed upon it.

Step 9

The next task that was undertaken was to randomise the order of the data within the arff data sheet. This was performed using the *Randomise* functionality within Weka using an initial seed value of 42.

Step 10

The randomised data was then split into two separate data sheets to be used with any initial classification attempts. The first containing the training data, made up of 4136 records, that will be used to create the initial models while the second, comprised of 4317 entries, containing the data used to test the models created from the classification algorithms.

Data Analysis

The first step undertaken once the data was ready to be investigated within Weka was to use the built in functionality to visualise the data in simple bar charts, the InstanceId and *Fatal_Accident_Index* were omitted from this view as they are quite simply not relevant. Below is the initial visualisation created by Weka using the entire data set comprised of 8633 data records.

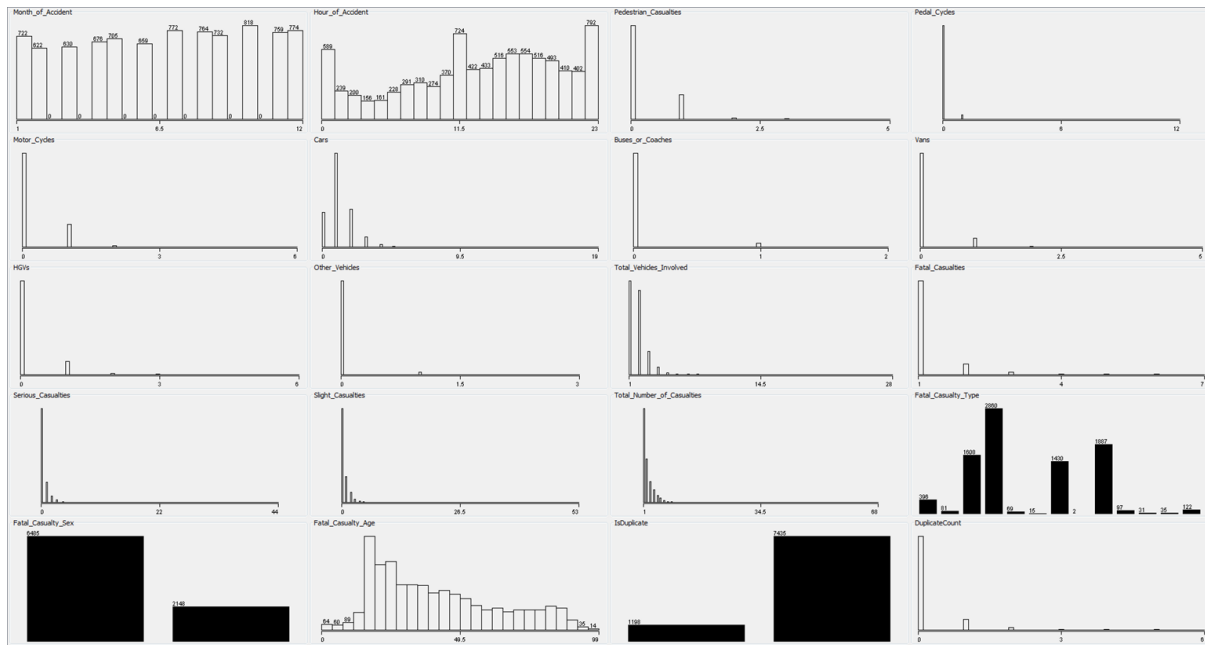


Figure 1 - Initial Visualisation

From this initial data visualisation there are a number of interesting features to investigate, most notably being the peak in accidents between 1100 and 1200 hours and the large peak in the fatal casualty age representation. It also allows a number of initial high level observations to be made such as the gender of a fatal casualty is much more likely to be male than female. There also does not appear to be any outliers or obviously incorrect data values within the pre-processed data that may skew future results.

Investigation of midday peak in accidents

As identified above an interesting peak in accidents occurs between 1100 and 1200 hours, this does not coincide with the traditional commute times to and from work/school, which would be expected between 0700 and 0900 hours for the morning commute and 1700 hours to 1900 hours for the afternoon travel at the end of the business and school day. There are small peaks during the timeslots however nothing that compares to the peak at 1100 hours.

One possible reason for this is that it coincides with the short break after the first period of lectures in school time tables and that there is therefore an influx of young drivers on the road for a brief period of time. It is also possible that the roads being quieter outside of rush hour leads to the fewer drivers moving at a much faster speed and therefore having a higher possibility of being involved in an accident.

The first step undertaken was to calculate the number of casualties across the range of ages within the dataset. As initially the only interest was in determining whether the simple hypothesis made had any grounding the ages were binned into equal width ranges for simplicity. As it is clear to see from the graph below even if there is a spike in casualties aged between 18 and 20 during this time it is far outweighed by any other single age range.

Number of Casualties by Age Range

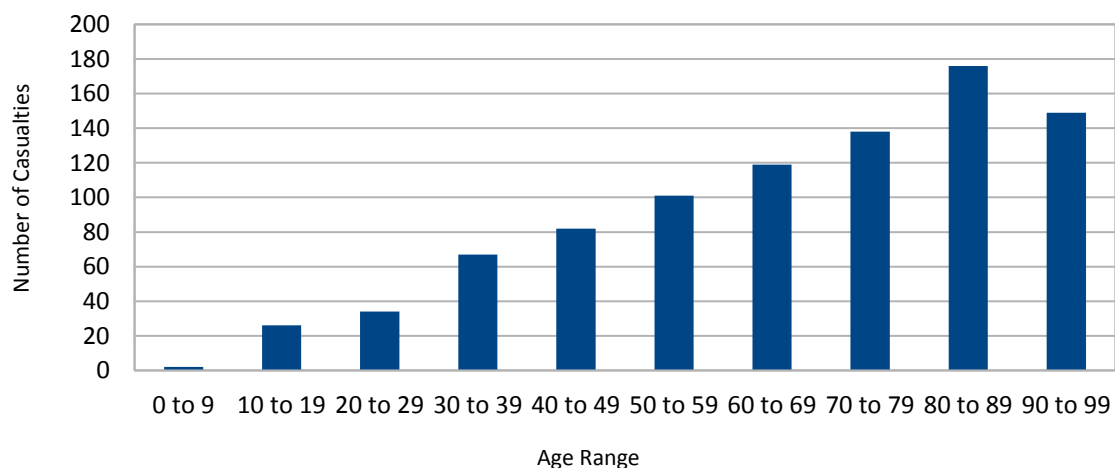


Figure 2 - Number of Casualties by Age Range

The next step undertaken to investigate this phenomenon was to attempt to determine an association between the age of the casualty and the hour of the day to. Both casualty age and hour of accident were converted to nominal types so these attributes could be processed using the Apriori algorithm to attempt to find an association rule that may support the above hypothesis. This however did not successfully find any associations between the casualty age and hour of accident across the entire data set suggesting the possibility that there is no strong correlation between the time of day at which the accident occurs and the age of casualty.

The next step taken to look at this behaviour was the initial attempts to look for any clustering of these accidents. This was initially done for a small selection of the attributes, casualty age, hour of incident, serious casualties, slight casualties and fatal casualties, using the SimpleKMeans algorithm which was selected as it would quickly allow clusters to be identified in the initial stages of investigation. It was expected that between two and four clusters would be identified so the algorithm was run and told to identify two, three, four and then five clusters. Four clusters appeared to give the most suitable output and the following image displays the clustering when viewing the time of the accident against the age of the casualty.

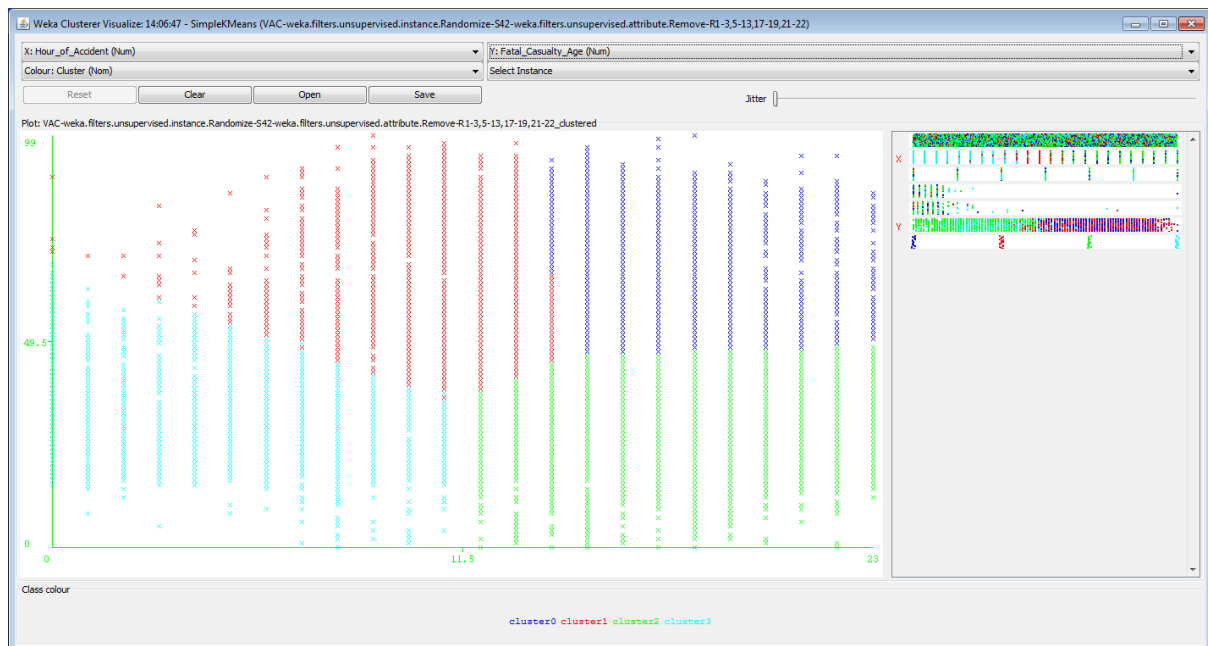


Figure 3 – Midday Accident Peak Cluster Identification

The clusters identified by the visualisation of the SimpleKMeans algorithm run are roughly representative of casualties between 0 and 50 in early morning to midday accidents, casualties between 50 and 99 in early morning to midday accidents, casualties between 0 and 50 in afternoon and evening accidents and finally casualties between 50 and 99 in afternoon and evening accidents.

The second cluster identified, in this case it has been assigned the colour red, appears to be the most prominent of the four classes during the 1100 hours to 1200 hours timeslot. As noted above when the clustering algorithm was run the number of casualties, categorised by severity, were also included to allow the ability to see if accidents which have been indicated as similar share any notable traits in the number of casualties or severities of casualties that occur.

There appears to be no significant differences between the casualties of accidents identified by cluster1 and casualties from any other accident. There is a single accident where a large number of slight casualties where is part of cluster1 however this number of casualties is much larger than every other accident within the same cluster and within the entire data set there is only one other accident which has a similar number of casualties.

Attempts were then made to produce a set of classification rules to investigate any determining criteria using the rule set generated by the RandomTree algorithm. Several attempts were made to create a suitable model using the hour of the accident, the number of vehicles involved, the number

of each type of casualty, the number of total casualties, the type of casualty and finally the age of the casualties and even when simplifying the number of attributes to be included in the model the rule sets were extremely complex, at minimum the size of the tree generated was 420 nodes and even with such a complex rule set the relative absolute error and root relative squared error were still extremely high, 86% and 91% respectively.

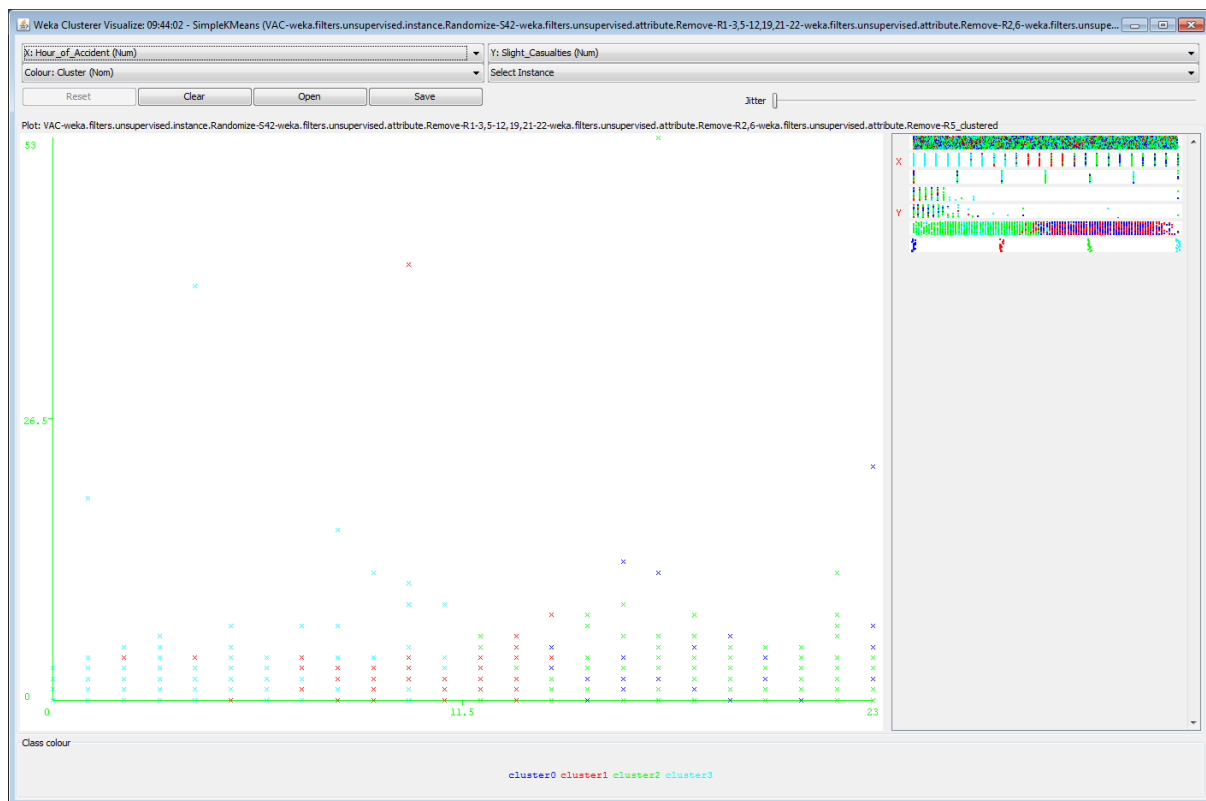


Figure 4 - Slight Casualties Against Hour of Accident

As this is only a single outlier it is not really telling of any trends and thus far accidents that occurred between 1100 hours and 1200 hours appear very similar to any other accidents.

To attempt to discover more about this phenomenon further clustering was attempted, this time with respect to the type of casualty and vehicles involved in the accident to see if commercial vehicles are frequently involved. From this further analysis the only point of note is that there may be a small increase the number of “other_vehicles” involved within the accident, as noted previously the “other_vehicles” is a place holder for vehicles such as emergency vehicles, motorised wheelchairs and horseback riders.

From the information gathered so far there is no conclusive reason as to why there is such a spike in accidents at this time or the factors that cause them, it's possible that roads being emptier at this time of the day allows traffic to flow at a faster speed on certain roads and therefore more accidents occur. Data regarding the types of road that these accidents occurs on is not included within this source however adding this information at a later date may allow for a greater understanding of this phenomenon to be built. With this additional data it may also be possible to produce a suitable classification model with a more reasonable level of error.

Casualty Severity & Casualty Type Classification

The general impression that is portrayed with respect to the number of casualties in an accident from the initial visualisation is that for the most part casualties of all severities are low with slight casualties being more prevalent than serious casualties which are in turn more prevalent than fatalities.

Looking at initial visualisations it appears the accidents that cause the most casualties are by far those that include buses however these types of accidents are much less frequent than others such as those involving cars. The distributions of the casualty severities between these types of accident however are very different; fatal casualties are much more likely in car crashes, both for drivers and passengers than in buses.

The first step undertaken was the generation of a suitable classification model utilizing the J48 algorithm within Weka. The first model generated using the base settings for the minimum number of objects within each class and the decision to at first produce a pruned tree. This model successfully classified 85.4% of the elements in the training set with a tree containing 101 nodes, 51 of which are leaf nodes. There was a reported 53% root relative squared error and relative absolute error of 28%, looking at the generated confusion matrix the model has failed completely to correctly classify motor cycle riders and bus drivers, however in the second case there is only a single instance which has most likely lead to this failure. There are other relatively large failures in classification, HGV drivers, bus drivers and bus passengers however in these cases the sample size is limited which may account for the problems encountered. There is a very notable misclassification of half of the total instances of car passengers casualties, this is interesting as unlike the other misclassification's noted previously there is a large sample size for this specific casualty type, 699 total with 331 incorrectly classified.

Another model was generated using the J48 algorithm, this time specified to produce an un-pruned tree to see if a more complex rule set will allow for greater accuracy in classification. The newly generated tree consists of 309 nodes and 155 leaf nodes; this is approximately three times as complex as the pruned tree model previously produced. This large increase in complexity was not accompanied by a large increase in accuracy; there was a two per cent increase in the accuracy of classification with small decreases in the relative absolute and root relative squared errors. The confusion matrix showed very similar results as in the first model with the only HGV drivers and bus passengers showing large improvements. This second model was then re-evaluated with the data from the test sheet previously created; the percentage of correctly classified instances was 84% which while not extremely high is roughly equal to the correct classification percentages found by the model against the training data.

This model will allow future accidents which contain incomplete detail to be classified; as noted above the current model only has a success rate of 86% however it is possible that with further processing and the addition of extra data attributes that this success rate could be increased.



Figure 5 - Casualty Type Classification Errors

Yearly distribution of casualties

Over the course of the year the pattern in which accidents occur the course of the year. The visualisation of month versus slight casualties shows that there is a spike in the second half of the year around the months of August, September and October. This may be due to the beginning of the school year where there are a large number of buses back on the road and rush hour traffic becomes more problematic. It would be expected for slight casualties to be slightly increased during these months while more serious and fatal casualty counts do not change. The visualisation below displays a graph showing the number of slight casualties on a monthly basis, there is also an additional peak during the month of May which coincides with the period of Easter holidays.

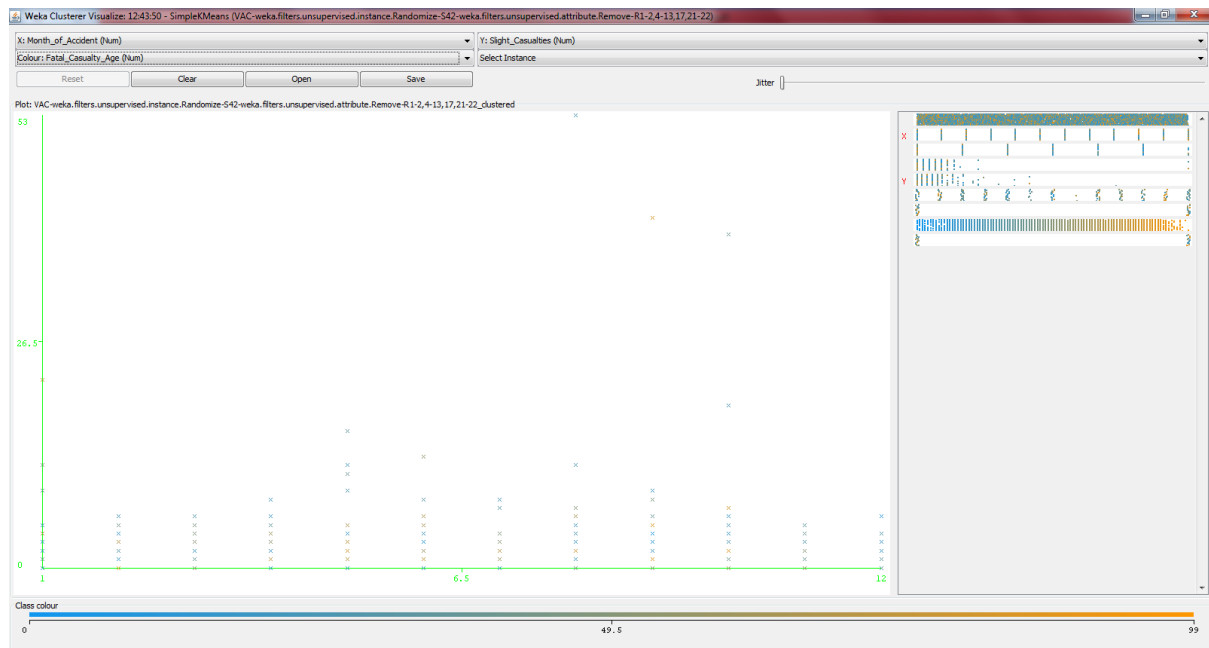


Figure 6 - Month of Accident against Slight Casualties

The distribution of fatal and serious casualties however are much more consistent, while there are some peaks the general level of both of these types are much more even in nature.

Summary

Further pre-processing could be undertaken in future to add a number of new additional data sources to the used data set to allow further information to be discovered. Notably the addition of road types would be extremely useful; additional attributes describing the classification of the road, i.e. A road, Motorway, Dual Carriage way and the level of congestion on the road at the time, or average detail for the road during different times of the day, would allow for further investigation into the causes of the midday spike in fatal accident rates. Details pertaining to weather conditions on the roads at the time of accidents would also be an extremely important and useful insight determining the true accident causes.

While no definitive reasoning has been discovered as to why fatal accident rates encounter a large spike between 1100 and 1200 hours several factors have been identified as to not being meaningful factors. The initial hypothesis was that it may be partially influenced by a break during school time tables and an influx of younger drivers on relatively empty roads for a brief period of time. This was shown not to be a factor in the increased accident rates and that the age of the casualties at this time are more likely to be older members of the public. With a number of extra attributes as described above this basic understanding could be used as a foundation to further investigate this phenomenon. In particular it could be interesting to review the levels of congestion on the roads at the time in questions to see if this is a contributing factor, it's possible that this casualty rate appears so high as it is in stark contrast to the earlier morning rush hour traffic which is extremely congested and therefore moving at a slower rate account for less serious injuries to be sustained due to the lower speed of any vehicles involved.

While the classification model created to predict the casualty type has a relatively high percentage of correct classifications the way in which incorrect classifications manifest themselves is very concerning. There is a rate of approximately fifty per cent of misclassification for the most common type of casualty type which due to the volume of data available for this casualty type in the data set would be expected to be one of the highest success rates. The model produced utilizing the J48 algorithm without taking advantage of pruning is extremely complex in terms of nodes when compared to the pruned version for very little additional accuracy. It's possible that this model could be used in conjunction with another model to provide classification predictions with a high level of certainty however it may be that with the additional information added as suggested above that an entirely more suitable model could be produced.

Reflection

The process has reinforced several key points;

- Domain knowledge is vital; while a small amount of knowledge regarding the domain from which the data was captured was held this was not enough. To begin to perform any analysis beyond the most basic a firm grounding of the subject area must be grasped. While a general logical approach and mind set are extremely important without this domain knowledge the task is extremely inefficient.
- Data set familiarity; Once more while it is easy to perform simple visualisations and the most trivial manipulations to make the most out of the data an extremely high level of familiarity is required. This can only come by spending a suitable amount of time with the data set and being very thorough during the pre-processing stage to ensure that the information is both complete and understood. It is also extremely important to ensure that the boundaries of the data are known to determine if and when it is suitable to draw in data from additional sources.

No interesting trends or results have been discovered as a result of this study with respect to the accident and casualty road traffic accidents data set selected. There has however been a greater appreciation and understanding of the problems presented by working with large volumes of data in an attempt to learn important historic lessons. It is much more representative of a scientific approach than many of the tasks that are encountered within the computing subject area with less of a binary success or failure end point.

If further work was to be performed to build upon what has been currently done a much larger amount of time would be allocated to pre-processing and simply looking at the properties of the data set. A number of additional data sources would be included to provide much greater information about each of the accidents to not only allow for more complex questions to be investigated but also to enable stronger models to be produced both for future predictive classification tasks and improved output from a range of clustering algorithms.

Bibliography

1. Open Government License. *Licensing for re-use - Open Government License*. [Online] The National Archives. <http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/>.
2. **Department of Transport**. Reported fatal personal injury road accident and casualty data: GB 2006-2008. *data.gov.uk*. [Online] Department of Transport, 19 5 2010. <http://data.gov.uk/dataset/gb-fatal-reported-road-accidents-06-08>.
3. —. *Reported Road Casualties in Great Britain: notes, definitions,*. [Document] s.l. : Department of Transport, 28 August 2013.
4. **Witten, Ian H, Frank, Eibe and Hall, Mark A**. *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques*. s.l. : Elsevier, 2001. 9780080890364.
5. **Linoff, Gordon S and Berry, Micheal J. A**. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. s.l. : John Wiley & Sons, 2011. 9781118087459.
6. **Han, Jiawei**. *Data Mining, Southeast Asia Edition: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems)*. s.l. : Morgan Kaufmann, 2006. 9780080475585.