Traffic Accidents Analysis

1 Introduction

The following is report is a deep look into traffic accident data provided by the British government for the year of 2019. There report is concerned with the two primary concerns of traffic accidents, the frequency at which they occur and their severity. Firstly, a general wide-ranging analysis was done to determine common trends in the data, followed by the testing of two distinct hypotheses. Finally, a probabilistic model was developed to predict both the temporal and spatial likelihoods of an accident, as well as its relative severity.

2 Analysis

2.1 Data Cleaning

The dataset as provided by GOV is fairly error free except for a few areas. twenty eight samples with missing coordinate data. However, all of these samples had local district information. Coordinate data for all towns and cities in the United Kingdom was scraped from the REF website using the Beautiful Soup package REF, and imputed according to the local district feature.

A similar approach was taken to impute missing time values. In this case, the sunrise and sunset times for London were scraped from REF. According to the light conditions of the sample, the median time between sunrise and sunset on the day of the accident was imputed in the case of light, and the median time between sunset and sunrise in the case of darkness. The purpose of this was to ensure that imputed data does not contradict the obvious and empirically true covariance between the time of the day and light conditions.

2.2 Primary Analysis

2.3 Geospaital Analysis

An initial clustering analysis based around 8 cluster centres shows that generally the vast majority of accidents occur in densely populated areas, with significant levels in London, Birmingham, Manchester, Leeds, Newcastle and Edinburgh. However, it can also be seen that there is a significant number of accidents in Wales and in the South East, which would suggest that accidents also occur frequently in rural areas.

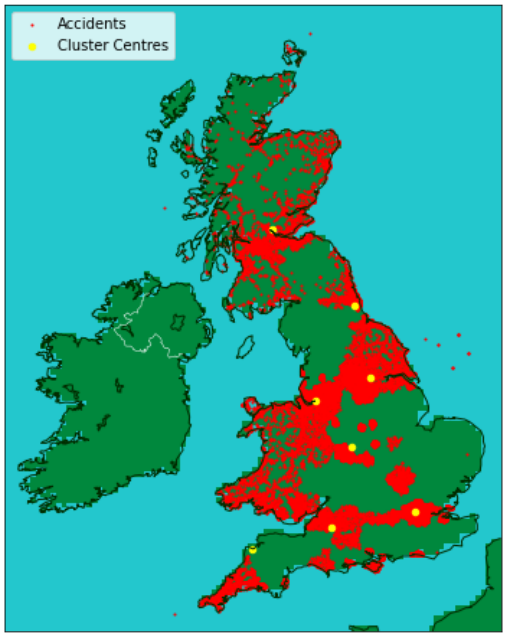
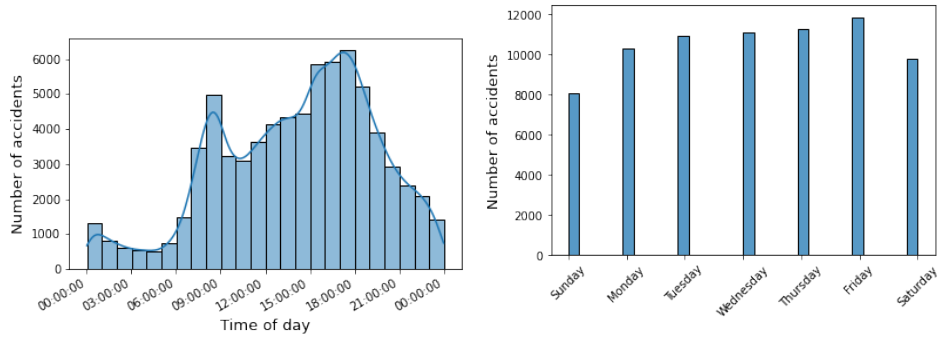


Figure 1: Geospatial clustering of accidents.

2.3.1 Temporal Analysis

An initial clustering analysis based around 8 cluster centres shows that generally the vast majority of accidents occur in densely populated areas, with significant levels in London, Birmingham, Manchester, Leeds, Newcastle and Edinburgh. However, it can also be seen that there is a significant number of accidents in Wales and in the South East, which would suggest that accidents also occur frequently in rural areas.

Considering the temporal dimension of the data, it was determined that there exist two primary SPIKES in accident frequency within the ranges 8am-10am, and then again from 3pm 7pm **figure ref**



(a) Accidents by time of day. (b) Accidents by day of the week.

Figure 2: Temporal analysis of accident frequency.

These are the times of commute, and so such increases are to be expected.

Further, it can be seen from **figure ref** that the frequency of traffic accidents is generally higher during the weekdays, with the maximum on Friday and the minimum on Sunday. This further supports the hypothesis that there is a significant increase in accidents during work-commute periods.

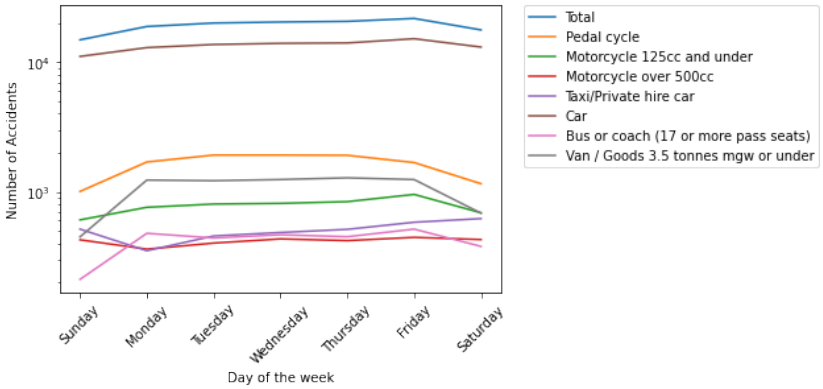
2.3.2 Vehicle Types

In the next stage of analysis, the aggregated accident data was subset according to the type of vehicle.

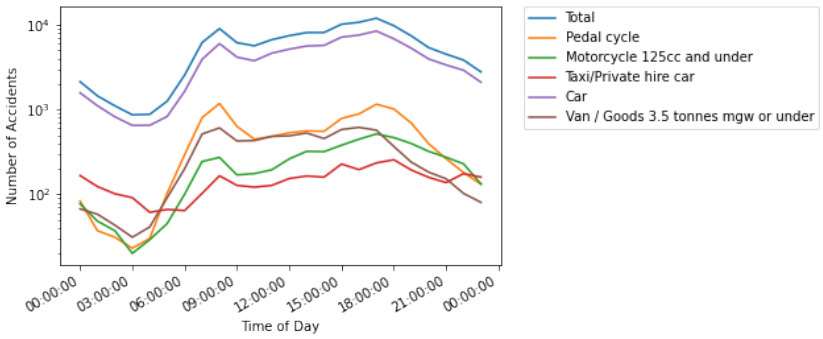
Both the absolute number of accidents, as well as the proportion of total accidents was considered on the temporal axis.

There is accident data for a total of twenty vehicle classifications. To simplify the analysis, only those contributing more than 2% of the total

number of accidents were considered.



(a) Vehicle type by day of the week.



(b) Vehicle type by time of day.

Figure 3: Type of vehicle involved in accidents.

As can be seen from **figure**, the distribution of accidents for each vehicle type generally follows the overall distribution for time of day. The one exception is for taxis and private car hire, which has a higher relative frequency of accidents between midnight and 9:00am than other vehicle types.

2.4 Hypothesis Testing

The following section describes three hypotheses that have been tested. The hypotheses are as follows:

- Is there a significant number of accidents in rural areas caused by or

involving drivers who do not live in the vicinity, and are they more likely to have more passengers?

- Is there more accidents at the same time of day during periods of the year after which the sun has gone down compared to when it is still light?
- Is there more accidents in the vicinity of football grounds on days when premier league football matches take place?

2.4.1 Rural Accidents

To test this hypothesis the data was subsetted for the conjunction of accidents taking place in rural areas and drivers who do not live in rural areas.

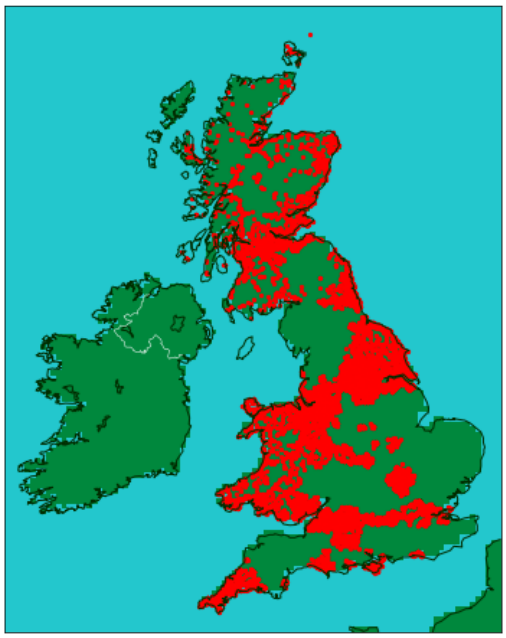


Figure 4: Accidents in rural areas caused by people not living in rural areas.

By plotting the geospatial data for this subset it can be seen that a significant number of accidents occurring in rural areas of Wales, the South East, North Yorkshire and Scotland are caused by people who do not live in such areas.

2.4.2 Sunrise & Sunset

- The hypothesis to be tested was that the ratio of average number of accidents per day in darkness would be higher than that in daylight for the same time delta over the entire year. - As can be seen in **ref fig**, the period tested for sunrise was between 7:00am and 8:00am, and for sunset between 8:30pm and 9:30pm. - To simplify the problem, The dates for which there were days of daylight and darkness in the specified time period were not included. - Graphically, this means, for example on **ref**, that only dates were analysed to the left and right of the outermost red dashed lines, as well as the dates included between the inner lines. - It was found that there was a total of 2988 accidents between 7:00am and 8:00am across the year, 1367 of which occurred during 155 days daylight, and 1621 occurring in 146 days of darkness. - This leads to an average number of accidents per day in daylight of 8.82, and 11.10 in darkness. - Hence, there is a 26% increase in accidents during darkness during the hours of 7:00am to 8:00am. - an equivalent analysis was done for the hours of 8:30pm to 9:30pm. - A total of 2143 accidents occurred, 424 of which were during 58 days of daylight, and the remaining 1719 accidents occurred during 204 days of darkness. - The average number of accidents per day in daylight for this time period is 7.31, compared to 8.43 in darkness. - This leads to a 15% increase in accidents during darkness during the hour of 8:30pm to 9:30pm

2.4.3 Accidents at Premier League Football Matches

A first hypothesis was tested to determine if there is a significant increase in traffic accidents surrounding football stadiums on the days of significant matches.

A test case was first carried out for the football match taking place at Old Trafford on Sunday 24th February 2019. Considering any accident within 5km of the stadium as being connected to the event, the accident count was compared against the number of accidents in this region every other Sunday during the year. On the day of the match, there were XNUM accidents in the region, 1.77 standard deviations above the mean. This was taken to be significant and worthy of further investigation.

Taking the premier league fixtures of 2019 REF, along with the coordinates of every premier league club stadium REF, the previous process was done iteratively for a total of 126 football matches during the year at 15

unique stadiums.

When summarised for every match, the z-score appears to be only 0.05 standard deviations from the average number of accidents. Although there is a significant difference in the number of accidents for different stadiums, on average the analysis implies that there is not a statistically significant rise in the number of accidents on the day of a football match compared to the typical same day of the week in the region.

The following table shows the summary statistics for premier league football matches. The second column shows the average number of accidents in the area of the stadium on days when there is a match, and the third column shows the average number of accidents on the same day of the week when there is not a match being played.

3 Predictive Model

- A statistical model has been developed in order to predict the conditions under which accidents are most likely to occur in, as well as severity of injuries sustained. - The purpose of the model is to be able to predict when an accident will occur, in order to aid in providing recommendations.

- The author assumes that the subset of all data samples that does not contain a single unknown value in the initial predictors of interest is sufficient for training the model. - After doing so there was still 39,462 samples, 36.27- A second assumption is that the nominal, ordinal and binary categorical predictors could all be treated equivalently during the feature selection process.

- In order to choose the most suitable features on which to train the model, all the features that seemed to be of value were joined into a single data set. - Categorical features were evaluated according to hypothesis test, and numerical features according to ANOVA.

- The data is heavily imbalanced on accident severity, on the order of 50:10:1 for slight, serious and fatal injuries respectively. - In order to accommodate for this imbalance, an auxiliary data set was produced by oversampling the minority class by using the SMOTE augmentation technique. - This provided a balanced set on which the model could be trained.

A host of classification models were evaluated by cross-validation on repeated stratified k-fold of the samples.

Decision tree based models were the by far the most accurate models employed, with the greatest accuracy coming from a stacked model.

4 Predictions

The author makes the following recommendations based on the analysis delivered.

- To increase awareness about the dangers of traffic accidents for bicyclers, targeting both the cyclist and the driver.
- To increase awareness of the dangers of high speed motorcycle use in rural areas, **think bike**.
- To consider better lighting conditions during the early hours of the morning and late at night depending on the sunset and sunrise cycle.

5 Conclusion

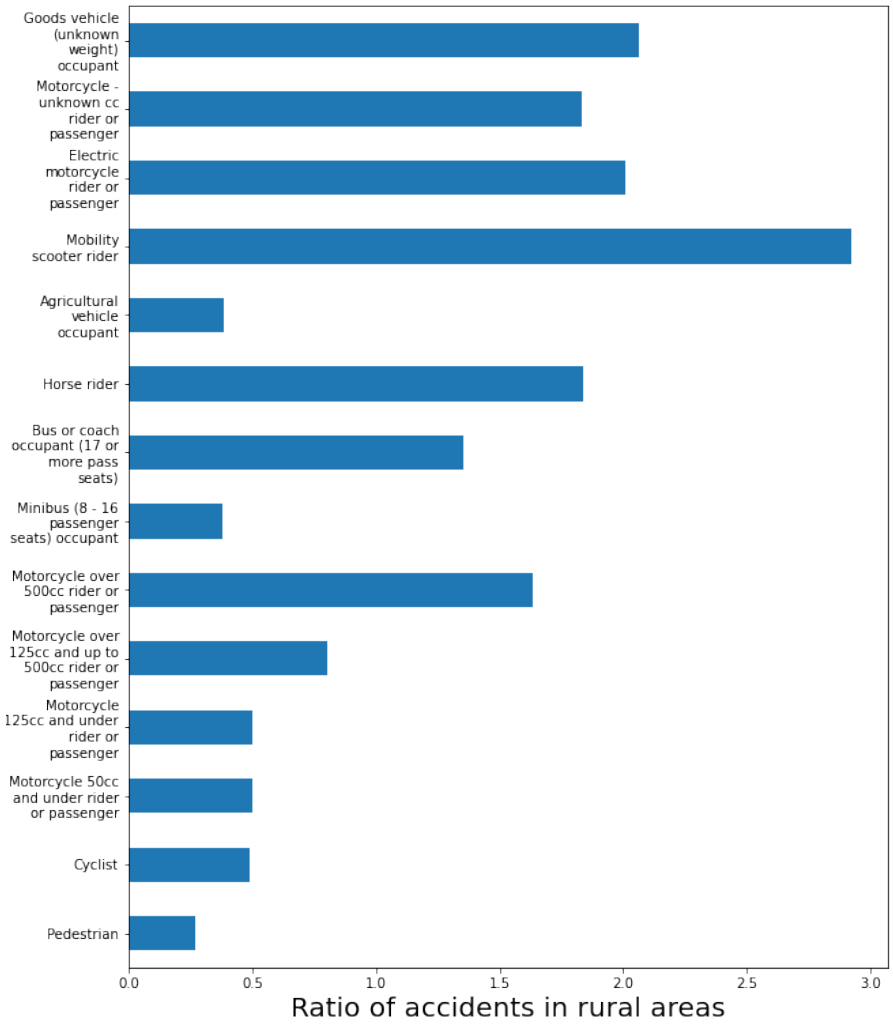
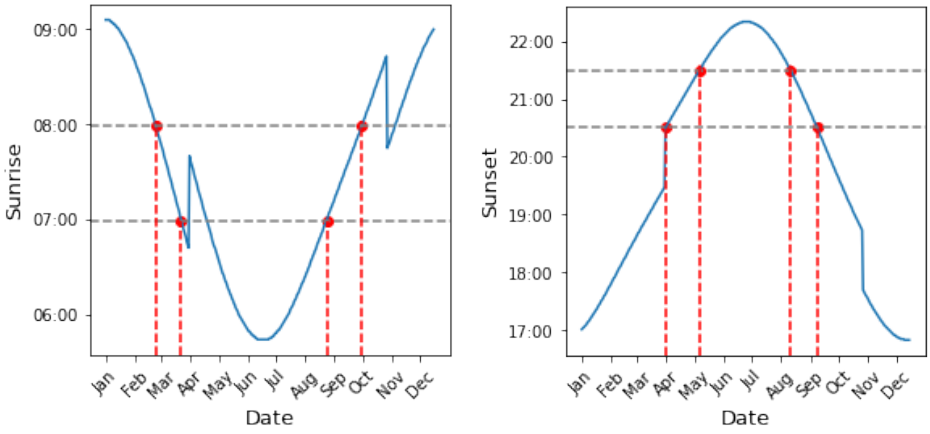

Goods vehicle (unknown weight) occupant
Motorcycle - unknown cc rider or passenger
Electric motorcycle rider or passenger
Mobility scooter rider
Agricultural vehicle occupant
Horse rider
Bus or coach occupant (17 or more pass seats)
Minibus (8 - 16 passenger seats) occupant
Motorcycle over 500cc rider or passenger
Motorcycle over 125cc and up to 500cc rider or passenger
Motorcycle 125cc and under rider or passenger
Motorcycle 50cc and under rider or passenger
Cyclist
Pedestrian
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Ratio of accidents in rural areas

Figure 5: .



(a) Sunrise times throughout the year. (b) Sunset times throughout the year.

Figure 6: Sunrise and sunset throughout the year.

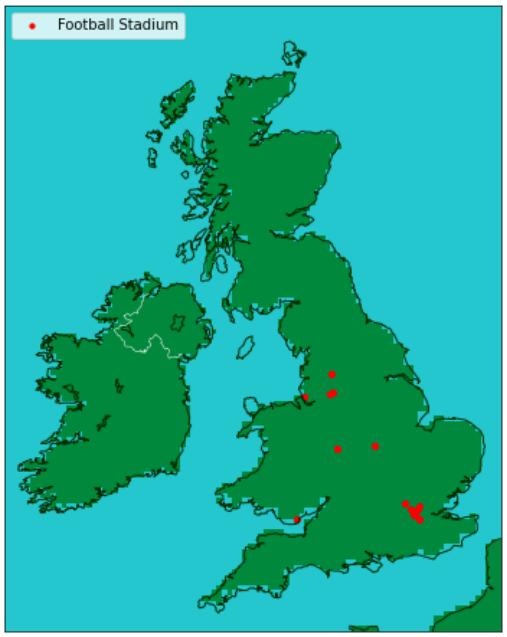


Figure 7: Premiere League footaball stadiums analysed.

Stadium	Accidents 1	Accidents 2	z-score
Anfield	3.0	3.6	-0.12
Cardiff City	1.9	1.8	0.47
Craven Cottage	16.1	16.4	-0.08
Emirates	21.6	21.5	-0.01
Etihad	3.8	5.2	-0.56
Goodison Park	3.3	3.5	0.01
King Power	1.7	1.9	0.34
Molineux	2.2	2.9	-0.29
Old Trafford	4.1	4.2	-0.04
Selhurst Park	14.2	11.7	0.63
Stamford Bridge	17.1	19.9	-0.47
Tottenham Hotspur	16.4	16.5	-0.0
Turf Moor	0.0	1.0	0.0
Vicarage Road	0.8	2.0	-0.23
Wembley	9.2	12.7	-0.86

Figure 8: Summary statistics for accidents surrounding Premiere League football stadiums.

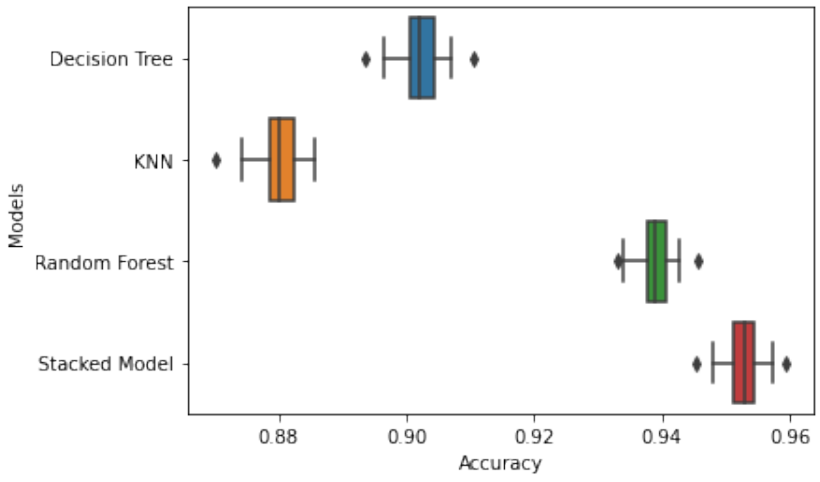


Figure 9: Accuracy of models evaluated using cross-validation.