Traffic Accidents Analysis

1 Introduction

The following report is an analysis of traffic accidents in the United Kingdom during the year of 2019, based on the freely available data published by the Department for (Transport 2021b; Transport 2021d; Transport 2021c). There are two primary concerns of traffic accidents, the frequency at which they occur and their severity. As such, the report will focus on those metrics.

The report is structured as follows. Firstly, a broad, frequentist approach to find common trends in the data, followed by the testing of three distinct hypotheses. Next, a probabilistic model was developed to predict the severity of accidents based on a subset of the predictors that are recorded via Transport (2021a). Finally, a selection of recommendations are proposed with reference to the analysis performed.

2 Analysis

2.1 Data Cleaning

The data employed in this analysis was for the most part clean and structured, except for a few key areas.

There were twenty-eight samples with missing coordinate data. However, these samples did have associated local district information. By use of the Beautiful Soup package (Richardson 2021), the coordinate data from all towns and cities in the United Kingdom were scraped from simple maps 2021 and imputed into the dataset where coordinate data was lacking.

A similar approach was taken to impute missing time values. In this case, the sunrise and sunset times for London were scraped from the web (sunrise and sunset 2021). 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

2.2.1 Geospaital Analysis

An initial clustering analysis based around eight 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. Based on this result, the hypothesis regarding accidents in rural locations was carried out.

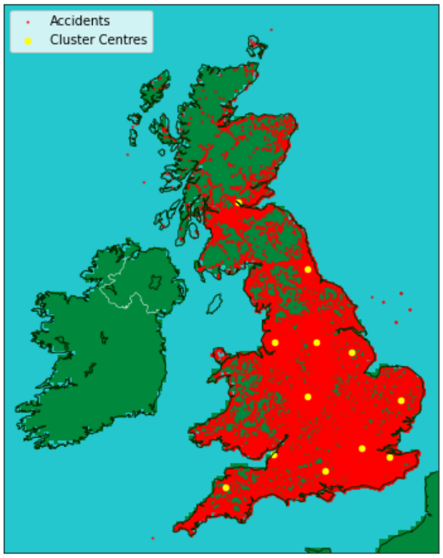


Figure 1: Geospatial clustering of accidents.

2.2.2 Temporal Analysis

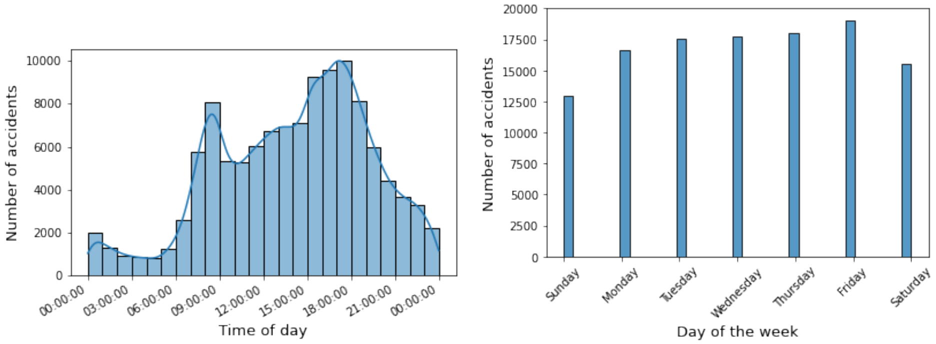
Considering the temporal dimension of the data, it was determined that there exist two primary peaks in accident frequency within the ranges 8am-10am, and then again from 3pm to 7pm, shown in figure 2(a).

These are the usual times at which people commute, and so such increases are to be expected.

Further, it can be seen from figure 2(b) 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.

2.2.3 Vehicle Types

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



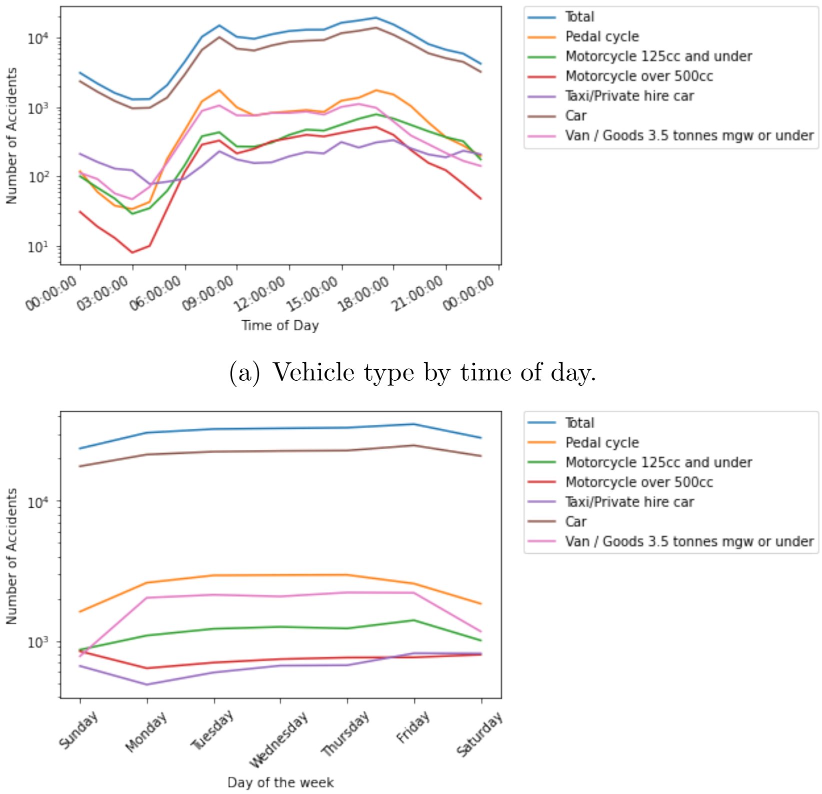
(a) Accidents by time of day.

(b) Accidents by day of the week.

Figure 2: Temporal analysis of accident frequency.

Both the absolute number of accidents, as well as the proportion of total accidents was considered on the temporal axis, as can be seen in figure 3.

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 time of day.

(b) Vehicle type by day of the week.

Figure 3: Type of vehicle involved in accidents.

As can be seen from 3(a), 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.

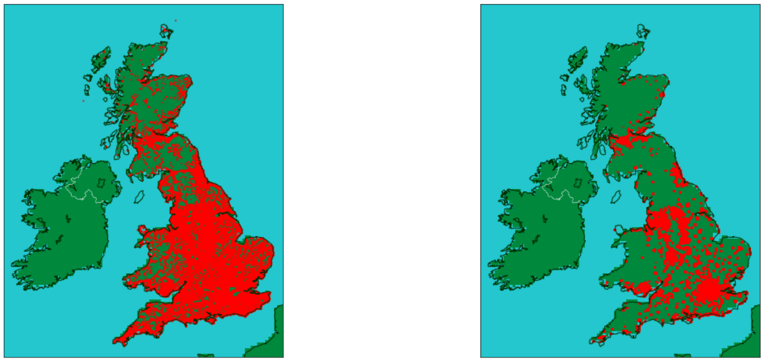
2.3 Hypothesis Testing

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 which type of vehicles are involved?
- 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?

2.3.1 Rural Accidents

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



(a) Accidents in rural areas.

(b) Accidents in urban areas.

Figure 4: Accidents by locality type.

As shown in figure 4, by plotting the geospatial data for the accident occurring in rural areas by people not living there, and comparing against the accidents in urban locations, 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.

Next, we determined the ratio of accidents in rural areas involving people not living in those areas to the total frequency of accidents, parametrised by vehicle type, figure 5.

Hence, we see that the most significant increase in accidents in rural areas exist for goods vehicles, motorcyclists of 500cc and higher, horse riders and mobility scooter riders. However, after taking the absolute magnitudes of these values into account, motorcyclists of all engine rating were the highest group at risk.

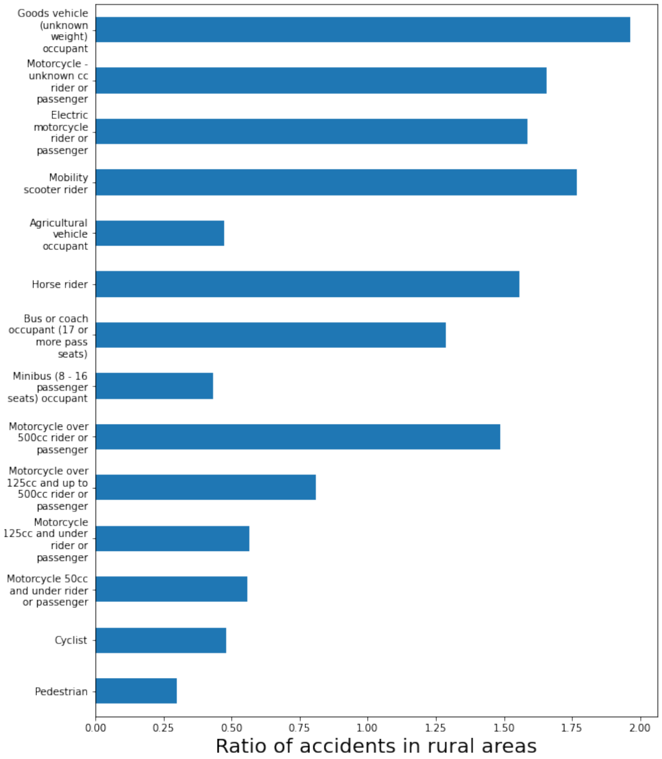
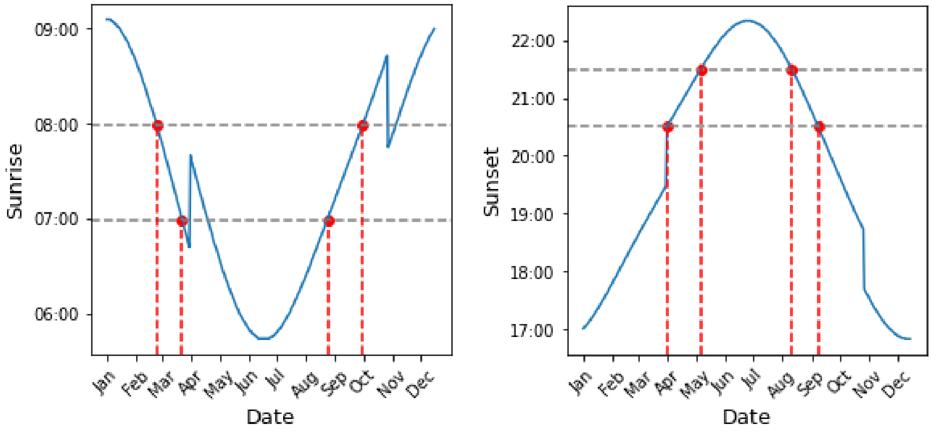


Figure 5: Ratio of accidents in rural areas to all areas by type of vehicle.

This could be due to motorcyclists choosing to ride in more rural areas for recreation, **CITE**, lending to a significant increase in accidents.

2.3.2 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 figure 6, the period tested for sunrise was between 7:00am and 8:00am, and for sunset between 8:30pm and 9:30pm.



(a) Sunrise times throughout the year.

(b) Sunset times throughout the year.

Figure 6: Sunrise and sunset throughout the year.

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 figure 6(a), 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 4938 accidents between 7:00am and 8:00am across the year, 2254 of which occurred during 155 days daylight, and 2684 occurring in 146 days of darkness. This leads to an average number of accidents per day in daylight of 14.54, and 18.38 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 3623 accidents occurred, 1054 of which were during 97 days of daylight, and the remaining 2569 accidents occurred during 204 days of darkness. The average number of accidents per day in daylight for this time period is 10.87, compared to 12.59 in darkness. This leads to a 16% increase in accidents during darkness during the hour of 8:30pm to 9:30pm.

2.3.3 Accidents at Premier League Football 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 five kilometres 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, the number of accidents was 1.30 standard deviations from the mean. This was taken to be significant and worthy of further investigation.

Taking the premier league fixtures of 2019 (fixturedownload.com 2022), along with the coordinates of every premier league club stadium (doogal 2022), the previous process was done iteratively for a total of 126 football matches during the year at 15 unique stadiums.

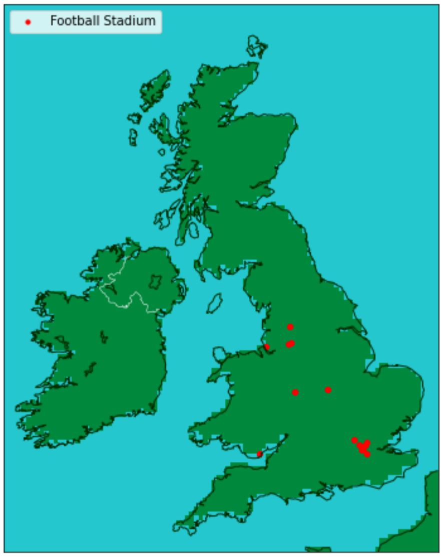


Figure 7: Premiere League footaball stadiums analysed.

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.

When summarised for every match, the z-score appears to be slightly below 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 table above (figure 8) shows the summary statistics for premier league football matches. The *Accidents 1* column shows the average number of accidents in the area of the stadium on days when there is a match, and the *Accidents 2* 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 was 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 developing such a 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 98,654 samples, 33.38% of the original merged data sets. A second assumption made was 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 an ANOVA hypothesis test (Sthle and Wold 1989).

The data was 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 CITE. This provided a balanced set on which the model could be trained.

A host of classification models were evaluated by cross-validation on a 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.

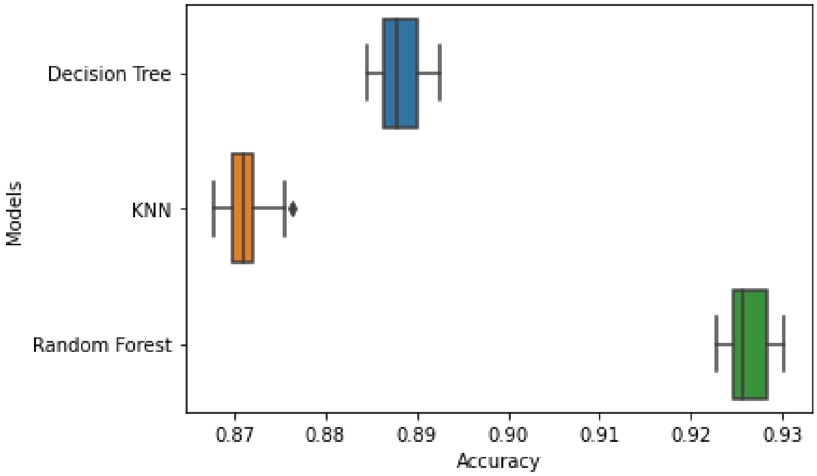


Figure 9: Accuracy of models evaluated using cross-validation.

As can be seen in figure 9, a stacked model achieved 95% accuracy during the cross-validation process.

However, it should be noted that when the model was later trained on the original dataset, that the accuracy regressed to 87%.

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, along the lines of **think bikee**.
- To consider better lighting conditions during the early hours of the morning and late at night depending on the time of year.
- To further investigate the reasons why there is a relative increase in accidents involving taxis during the early hours of the morning. This could suggest that overworking and tiredness is playing a key role.

References

Department for Transport (2021b). *Road Safety Data - Accidents 2019.* `https://data.dft.gov.uk/road-accidents-safety-data/dft-road-casualty-statistics-accident-2019.csv`.

Department for Transport (2021d). *Road Safety Data - Vehicles 2019.* `https://data.dft.gov.uk/road-accidents-safety-data/dft-road-casualty-statistics-vehicle-2019.csv`.

Department for Transport (2021c). *Road Safety Data - Casualties 2019.* `https://data.dft.gov.uk/road-accidents-safety-data/dft-road-casualty-statistics-casualty-2019.csv`.

Department for Transport (2021a). *Road accident and safety statistics: quality and methodology.* URL: `https://www.gov.uk/guidance/road-accident-and-safety-statistics-quality-and-methodology`.

Leonard Richardson (Apr. 20, 2021). *Beautiful Soup.* Version 4.11.0. URL: `https://www.crummy.com/software/BeautifulSoup/bs4/doc/#`.

simple maps (2021). *United Kingdom Cities Database*. URL: `https://simplemaps.com/data/gb-cities`.

sunrise and sunset (2021). *Sunrise and sunset London 2019*. URL: `https://www.sunrise-and-sunset.com/en/sun/united-kingdom/london/2019/\`.

fixturedownload.com (2022). *English Premier League 2019/20 fixture and results*. URL: `https://fixturedownload.com/results/epl-2019`.

doogal (2022). *UK football stadiums*. URL: `https://www.doogal.co.uk/FootballStadiums.php`.

Lars Sthle and Svante Wold (1989). "Analysis of variance (ANOVA)". In: *Chemometrics and Intelligent Laboratory Systems* 6.4, pp. 259–272. ISSN: 0169-7439. DOI: `https://doi.org/10.1016/0169-7439(89)80095-4`. URL: `https://www.sciencedirect.com/science/article/pii/0169743989800954`.