Environmental Conditions and Road Traffic Collisions in the UK

Author: Toby Staines

1 Introduction

1.1 Domain Overview

Each year in the UK there are over 130,000 traffic collisions involving injury which are reported to the police, resulting in between 1,700 and 1,800 deaths [1]. Data on these incidents [2] is published annually by the Department for Transport, and has been recorded and published in a similar format since 1979, making it a rich data source for investigation.

A report is published alongside the data [1], which gives some interesting insights in to annual trends in accident and fatality rate, and breaks down the data by looking at specific road user types (e.g. Car divers, cyclists, pedestrians, etc.). It also considers the impact of possibly related factors, such as the weather, drink driving rates, and even GDP, on incident rates. However, the report has very little detail about regional trends, patterns across the year or across the day, or whether there are relationships between the physical environment and the rate of crashes. These are all areas which are ripe for investigation. If relationships are identified in these areas it could help to guide future infrastructure developments, leading to an improvement in UK road safety.

1.2 Report Aims

This report will investigate the following questions:

- Is there a relationship between type of road, or driving conditions, and the rate of traffic collisions and/or fatalities in the UK?
- How do the patterns of accident conditions vary across the country?

I will look to answer these questions by investigating the relationships between accident and fatality rates and:

- Road type
- Lighting conditions
- Weather conditions
- Road surface conditions

The expectation is that darkness, poor weather, and faster roads are all contributory factors to higher rates of accidents, and that this will be reflected in the geographic distribution of accident rates.

1.3 Data

The investigation will focus on the most recent data available, that for 2016. The data consists of three main tables (a full attribute list is provided in section 4.6):

1. Accidents – The table has 32 columns, detailing the location, time, date, lighting, weather, and road surface conditions, number of causalities, road type and other variables. Each observation represents one of 136,621 collisions involving injury reported to the police in 2016.

2. Casualties – Linked via 'Accident Index' to the Accidents table, the table has 16 columns, giving further detail on the casualties involved. There are 181,384 rows, each representing a single person injured in a collision.
3. Vehicles – This table gives details of the vehicles involved in collisions but is not used in this investigation.

To assist with this analysis, I have also obtained government estimates for distance travelled on different types of road in the UK in 2016 [3], the relative density of road traffic for each hour of each day of the week in 2016 [3], and the UK population, at Local Authority (LA) level, for 2016 [4]. A shapefile provided by the Office of National Statistics giving the geographical boundaries of each LA is also used for mapping purposes [5].

1.4 Plan

The analysis will begin by looking at the pattern of collisions throughout the year and throughout the day, in total, and on different road types. It will then investigate the data at LA level and attempt to identify correlations between environmental conditions and accident rates. If such correlations are identified, a model will be defined to describe the relationship, and deviations from this model investigated to identify regional patterns. As data on the occurrence of different environmental conditions is not directly available, proportions of accidents occurring in different conditions will be used for this purpose. I will look to identify clusters of LAs with a similar pattern of accident conditions to assess their spread across the country and how this compares to the model.

2 Analysis Process and Results

2.1 Data Preparation:

The categorical variables in the data are stored as a numeric code, with a separate spreadsheet detailing the meaning of each code for each variable. In order to translate the data into a readable format, it was loaded, along with the data dictionary, into Pandas Data Frames and the codes translated to values, before being exported to a new csv, which was then used for all remaining work.

On initial viewing the data appears to be almost totally complete, but upon closer inspection it was found that some of the values were equivalent to missing data (e.g. 'Unknown'). There were 3,999 records with data missing from key variables. These represented just under 3% of the data, so it was decided that they would have little impact on the overall results and they were removed.

In order to investigate incident rates by road type it was necessary to integrate estimates for distance travelled on different types of road [3]. The two data sets use different classifications for road type (detailed in section 4.4), so the more granular road types in the main data set were translated to those used in the distance travelled estimates.

2.2 The National Level:

2.2.1 Accident Distribution Across Lighting, Road and Weather Conditions:

All of the environmental conditions considered by this report showed a similar distribution (see Figure 1), with one category accounting for 70-85% of observations, a secondary category making up 10-25%, and the remainder covered by uncommon conditions. This is to be expected as, although data on weather conditions were not available, we know that in the UK it is generally dry and clear, but sometimes rainy, and more driving is done during daylight hours (see section 2.2.2).

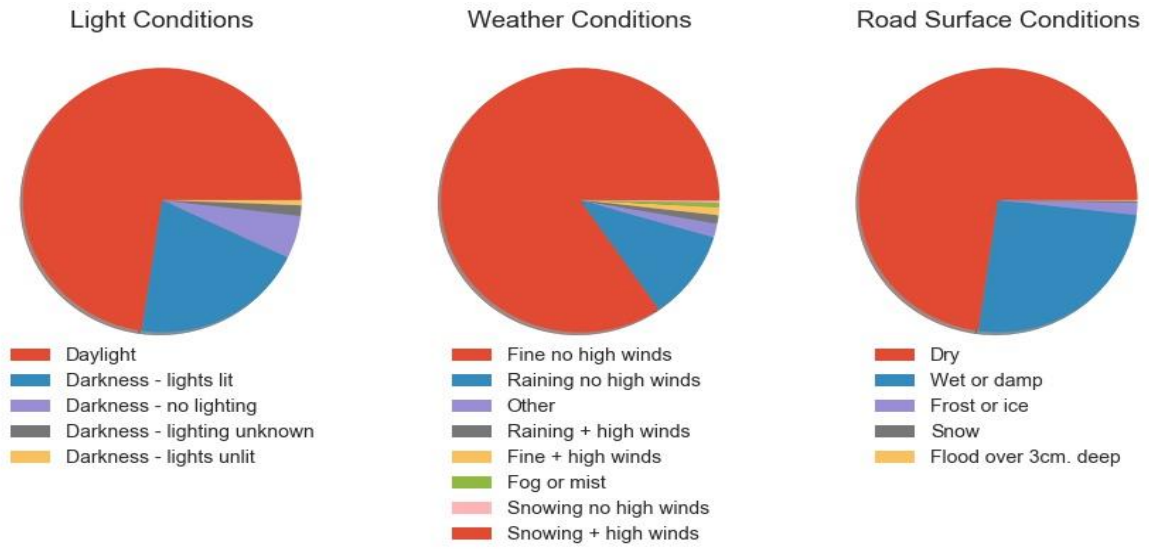

Figure 1: Occurrence of environmental conditions

2.2.2 Accident Distribution across Time:

Accidents were grouped by date and their distribution throughout the year visualised (Figure 2). The variance of incidents appears larger during the winter months, but the general trend shows little change across the year. The monthly mean number of accidents was 11,052, with a standard deviation of 533 (less than 5%). Taking in to account the change in traffic volume by month, the standard deviation is still only 8%. One point of interest is the three outliers in late December. These are the three days of the year with the lowest number of crashes: Christmas Day, Boxing Day, and New Year's Eve.

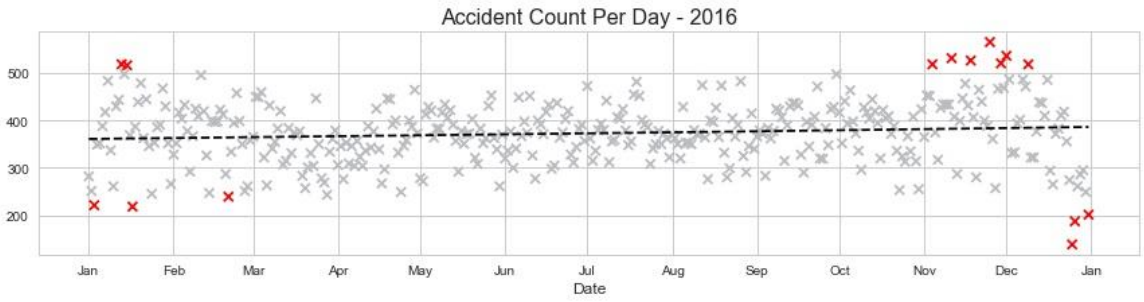

Figure 2: The distribution of road traffic accidents in the UK throughout 2016. Outliers (>2sd from mean) are marked in red.

Incidents were then grouped by the hour of the day in which they occurred (Figure 3, top). As one might expect, there is a peak in collisions around the times of the morning and afternoon rush hours, as far more driving is done at these times [3]. In order to investigate whether these times were actually the most dangerous, the number of crashes in each hour of the day were divided by the relative volume of traffic on the roads at that time, with the results shown in Figure 3 (bottom). The same was done for just the months of December and June, in order to see if there is an obvious difference between the darkest and lightest months. Figure 3 shows that car crashes are more likely at night, with a much smaller impact due to the morning rush hour than the raw numbers suggest, and no observable peak for the afternoon rush hour. However, it is not possible to draw a causal relationship between accident rate and light levels, as other factors, such as tiredness, are known to impact driver safety at night [6]. There is also little observable difference between December and June. December shows a higher peak at 1am, but this is a time at which it is dark year-round, so the peak must be due to other factors. Throughout the rest of the day the patterns are very similar.

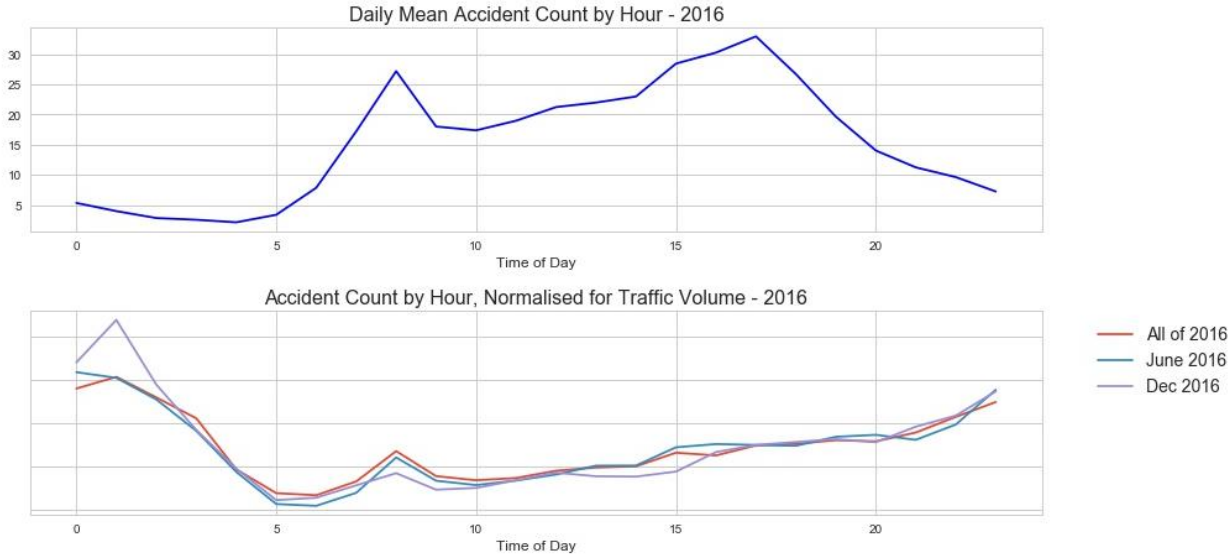

Figure 3: The distribution of road traffic accidents throughout the day in 2016

2.2.3 Accident Distribution Across Road Type:

Accidents, casualties, serious injuries, and fatalities were grouped by the type of road they occurred on and visualised in histograms, showing raw numbers and rate per distance travelled (Figure 4). Accidents, casualties and serious injuries (casualties and serious injuries shown in section 4.3) all show a similar pattern: in the raw numbers, minor urban roads show a higher number of crashes than urban A roads, and rural A roads show a higher number than minor rural roads. When taking in to account distance travelled, the pattern on both urban and rural roads is switched. The pattern of deaths is quite different from that of accidents, with the

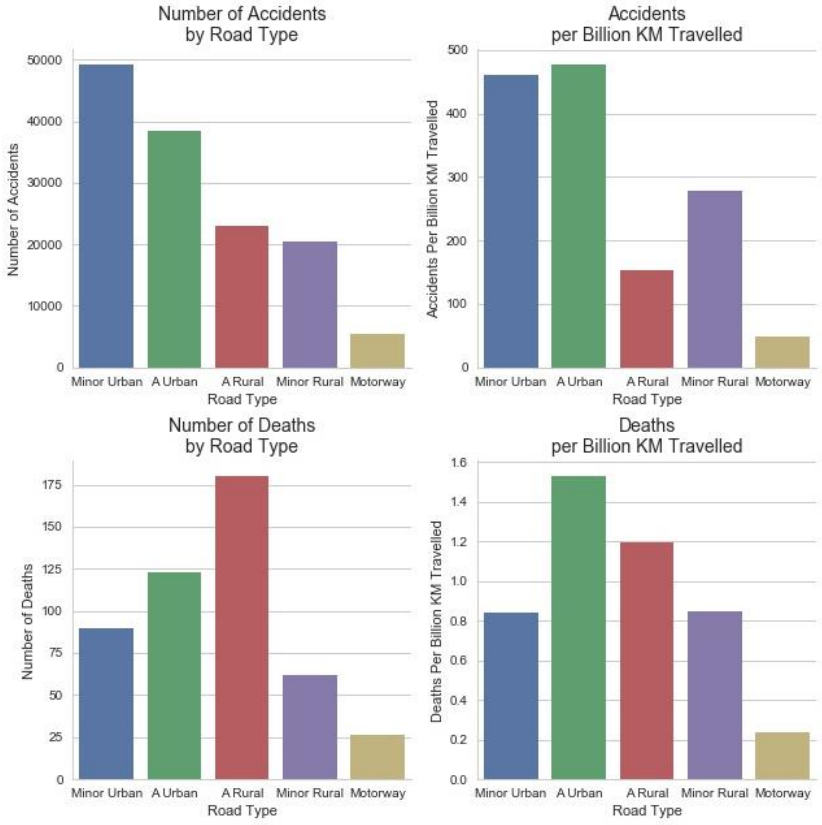

Figure 4: Accidents and deaths on different types of road in the UK in 2016

number of fatalities on rural A roads being far higher than that on either type of urban road. When taking distance travelled into account urban A roads have the highest rate of fatalities.

When viewing accident rates and road types by time, as in section 2.2.2, all road types showed a similar pattern. It was not possible to investigate the relationship between road type and accident rate at the LA level, as estimates of road use at this level are not available, but this is an area for further investigation if such data can be located.

2.3 Comparing Local Authorities:

The categorical variables 'Lighting Conditions', 'Weather Conditions', and 'Surface Conditions' were transformed in to dummy variables and grouped by LA, giving a count of how often each value occurred for each LA. This was the same level that population estimates and geographical boundaries were provided at, and the alternative LSOA level would have made the data too sparse, with 28,457 areas. Population estimates were also incorporated at this stage, enabling calculation of LA accident rates per 1000 inhabitants.

One extreme outlier LA, the City of London, was identified. Due to its low resident population but very busy nature, the City had an accident rate of 38.8 per 1000 inhabitants, over six times the rate of the next highest LA (Westminster, at 5.94, which is probably affected by similar bias, but to a lesser degree, which did not noticeably affect the results). The City was removed to avoid distorting subsequent results.

Dummy variable counts were translated into proportions of the total accident count for each LA. To account for the skewed nature of these variables (see section 2.2.1), enable easier comparison between variables and between LAs, the distribution of these proportions across LAs was rescaled between 0 and 1.

This process produced 18 new columns. Plotting each of these against the accident, casualty, serious injury, and death rates, and calculating Pearson's and Spearman's rank scores, showed that most of these combinations did not have a strong correlation.

Of the 72 comparisons made, the five giving an absolute correlation coefficient greater than 0.3 are listed in Table 1. Notably, none of the selected environmental conditions relate to lighting.

Table 1: Correlation between environmental conditions and accident or casualty rate

Environmental Condition	Dependent Variable	Pearson Correlation	Pearson P-value	Spearman's Rank	Spearman P-value
Weather Conditions: Raining no high winds	Accidents	-0.304351354	1.45E-09	-0.353806384	1.28E-12
Road Surface Conditions: Wet or damp	Accidents	-0.312038022	5.28E-10	-0.321076722	1.55E-10
Road Surface Conditions: Dry	Accidents	0.318076324	2.34E-10	0.311752501	5.49E-10
Weather Conditions: Fine no high winds	Accidents	0.294427866	5.13E-09	0.321802605	1.40E-10
Weather Conditions: Raining no high winds	Casualties	-0.277886431	3.78E-08	-0.315276032	3.42E-10

Having identified potential correlations, the next step was to build a regression model and see how this fitted across the country. A multivariate regression assumes no relation between the predicting variables, but it is obvious that there will be some in this case. The two Weather Conditions values and the two Road Surface Condition values are pairs originating from the same variables. In addition, a wet road surface is clearly linked to rain and a dry surface linked to fine weather. Figure 5 confirms that all four variables are strongly correlated. This allowed the use of a simple univariate regression model. 'Raining no high winds' was selected as it exhibited the strongest correlation to accident rate. **Error! Reference source not found.** shows their relationship. The resulting polynomial regression model (shown in red) was then subtracted from the true

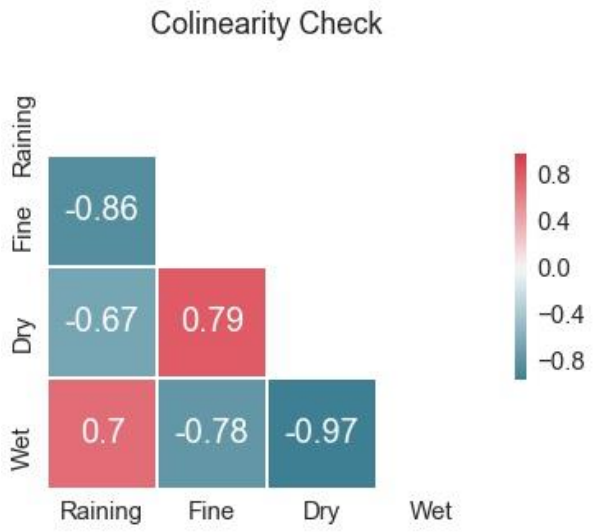

Figure 6: Confirming co-linearity of potential predictors

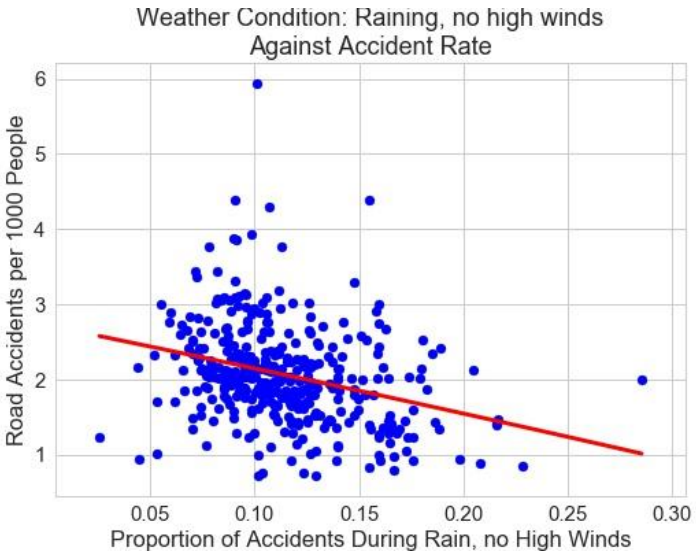

Figure 7: Raining, No High Winds against Accident Rate

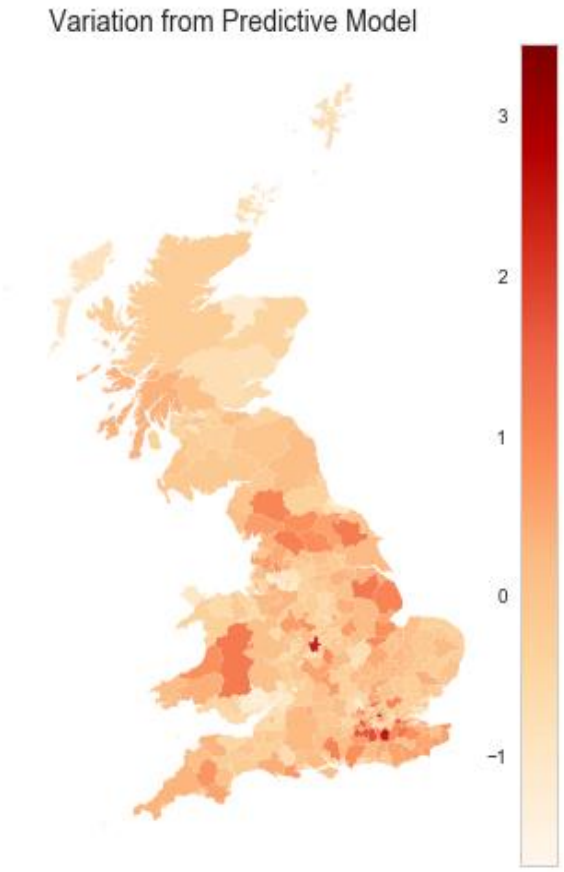

Figure 5: A map of the UK showing the difference in accident rate (per 1000 residents) from the rate predicted by a linear regression model based on rainy, windless weather conditions. High values indicate a higher than expected accident rate.

accident rate of each local authority and the residuals mapped in Figure 7. The model showed no obvious strong regional pattern, although higher than expected accident rates can be seen in the suburban areas around London and Birmingham.

With weather and surface conditions having been identified as a possibly contributory factor in accident rates I wanted to see if it was possible to identify groups of LAs with similar frequencies of these conditions. K-means clustering on rescaled road surface conditions did not provide good results. Clustering was also attempted with the variables rescaled to a uniform distribution, but this gave too much weight to the rarer conditions. Returning to the original proportions produced three clear clusters. The profile of the LA closest to the centroid of each cluster and the geographic distribution of clusters is shown in Figure 8.

2.4 Tools

Aside from some minor formatting changes to prepare data for loading, which were done in Excel, all other work was done using Python. Numpy and Pandas were used for the majority of data manipulation. Sci-kit Learn was used for normalising the data

(quantile_transform and MinMaxScaler), K-means clustering(KMeans) and for identifying the LA closest to the centre of each cluster(pairwise_distances_argmin_min). Matplotlib and Seaborn were used for graph based visualisations and Geopandas was used for geographic visualisations.

3 Discussion of Results

The analysis of road types detailed in section 2.2.3 did show that, with the exception of motorways, faster roads do tend to be more dangerous, in terms of the severity of accidents (as measured by number of fatalities), but not in terms of number of accidents, which are most frequent in urban areas. The government report accompanying the data [1] highlights that more fatalities take place on rural roads, and puts this down to the higher speeds and the fact that help is usually further away. However, this explanation does not fully hold when looking at deaths in relation to distance travelled, where urban A roads show the highest rate.

Overall, there are clear differences in collision and injury rates on different types of roads. The relatively higher rate of fatalities when compared to rate of accidents on A roads highlights the increased danger of higher speeds on these roads. At a time when police forces are disabling or removing speed cameras due to budget constraints [7], this work supports the case for reduction of speed limits and increased speed enforcement.

In relation to environmental conditions, the expectation was that driving in the dark and in poor weather was dangerous, and that this would be evidenced by a pattern of higher crash rates during the winter, and in areas which experience generally colder and wetter conditions (the north and west of the country), but this is not supported by the results.

There is no significant change in the rate of traffic accidents throughout the year; the monthly crash rate varies very little, and December and June have been shown to have very similar hourly patterns of incidents. The analysis also found no connection between the proportion of accidents occurring in different lighting conditions and the rate or severity of those accidents.

A relationship was identified between weather/road surface conditions and accident rate, but it was the reverse of that expected. LAs with a higher proportion of accidents involving wet conditions tended to have a

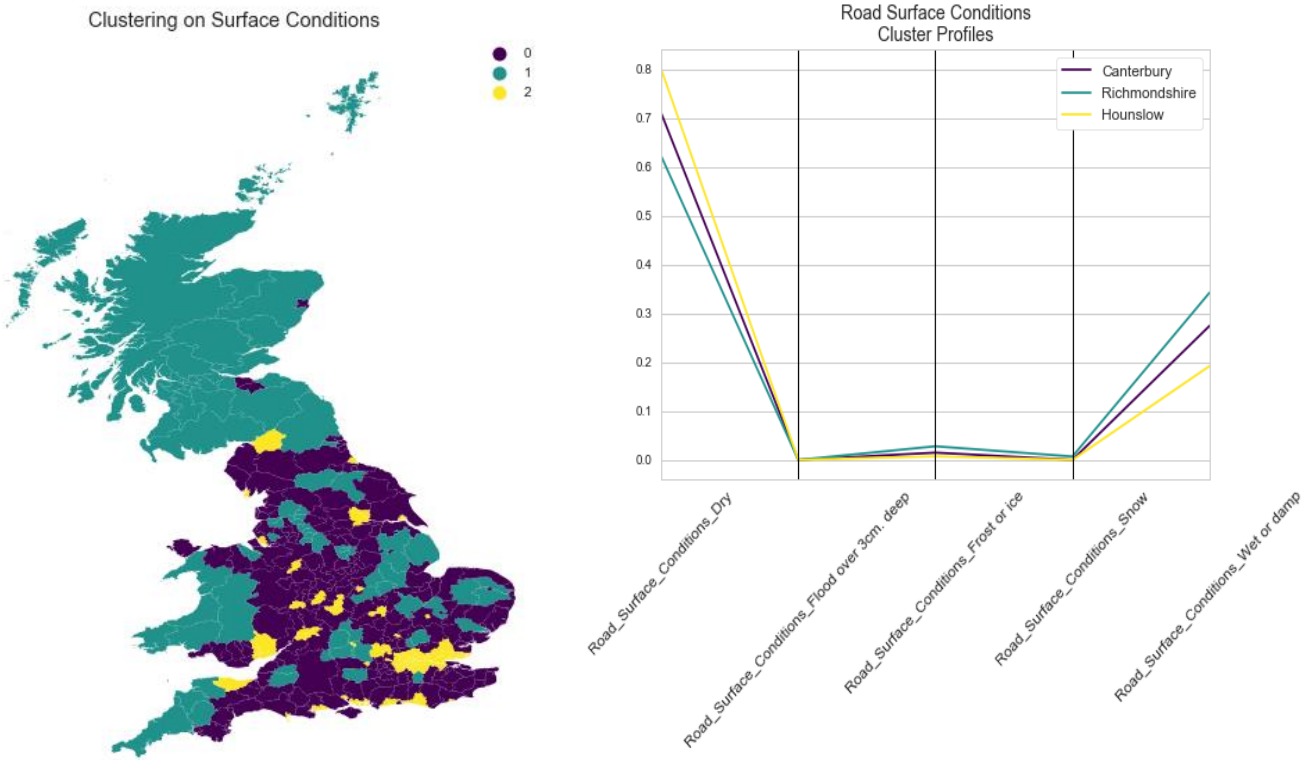

Figure 8: Clustering on Surface Conditions. Left: Geographic distribution of clusters. Right: Parallel coordinates plot showing the profile of the LAs closest to the cluster centroids.

lower overall rate of accidents. A possible reason for this could be that rather than causing more accidents, the poor weather reduces them, by discouraging driving, leading to fewer drivers, less congested roads and lower speeds. The relationship was not especially strong, but this could support the above hypothesis, as often a journey is required, regardless of the weather. Accurate daily weather records and traffic flow estimates at the LA level would allow greater investigation in this area.

It was possible to identify clusters of LAs with a similar profile of accident conditions (Figure 8) and these match expectations, following a clear geographic pattern of urban areas, the warmer, drier south and east of the country, and the colder, wetter north and west, along with some areas of higher altitude, such as The Pennines.

Although the results have been somewhat surprising the objectives of the investigation have been met, but there are still several areas of the data which could be investigated further. One area of possible inaccuracy in the analysis comes from the use of population estimates and the subsequent 'Accidents per 1000 people' measure. Although a good starting point, this does not take account of the fact that the victims of a crash will often not be resident in that LA. It assumes a relationship between the resident population of an area and those driving there, which is not necessarily valid. I would like to repeat the analysis using road traffic volumes at the LA level, if these were available, and compare the results to those of this piece of work.

4 Appendixes

4.1 Bibliography

[1] D. for T. UK Government, "Reported road casualties Great Britain, annual report: 2016 - GOV.UK." [Online]. Available: https://www.gov.uk/government/statistics/reported-road-casualties-great-britain-annual-report-2016. [Accessed: 30-Oct-2017].

[2] D. for T. UK Government, "Road Safety Data - Datasets," 2017. [Online]. Available: https://data.gov.uk/dataset/road-accidents-safety-data. [Accessed: 30-Oct-2017].

[3] D. for T. UK Government, "GB Road Traffic Counts - Datasets." [Online]. Available: https://data.gov.uk/dataset/gb-road-traffic-counts. [Accessed: 05-Nov-2017].

[4] O. for N. S. UK Government, "Population Estimates for UK, England and Wales, Scotland and Northern Ireland - Office for National Statistics." [Online]. Available: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland. [Accessed: 30-Nov-2017].

[5] O. for N. S. UK Government, "Local Authority Districts (December 2016) Full Clipped Boundaries in Great Britain - Datasets." [Online]. Available: https://data.gov.uk/dataset/local-authority-districts-december-2016-full-clipped-boundaries-in-great-britain2. [Accessed: 30-Nov-2017].

[6] R. Hunter, "The 24th Westminster Lecture on Transport Safety Staying Awake, Staying Alive: The problem of fatigue in the transport sector," *Parliam. Advis. Counc. Tansport Saf.*, 2013.

[7] M. Busby, "Only half of Britain's fixed speed cameras are active | UK news | The Guardian," *The Guardian*, 2017. [Online]. Available: https://www.theguardian.com/uk-news/2017/nov/04/only-half-of-britains-fixed-speed-camera-are-active. [Accessed: 04-Nov-2017].

4.2

4.3 Additional Figures

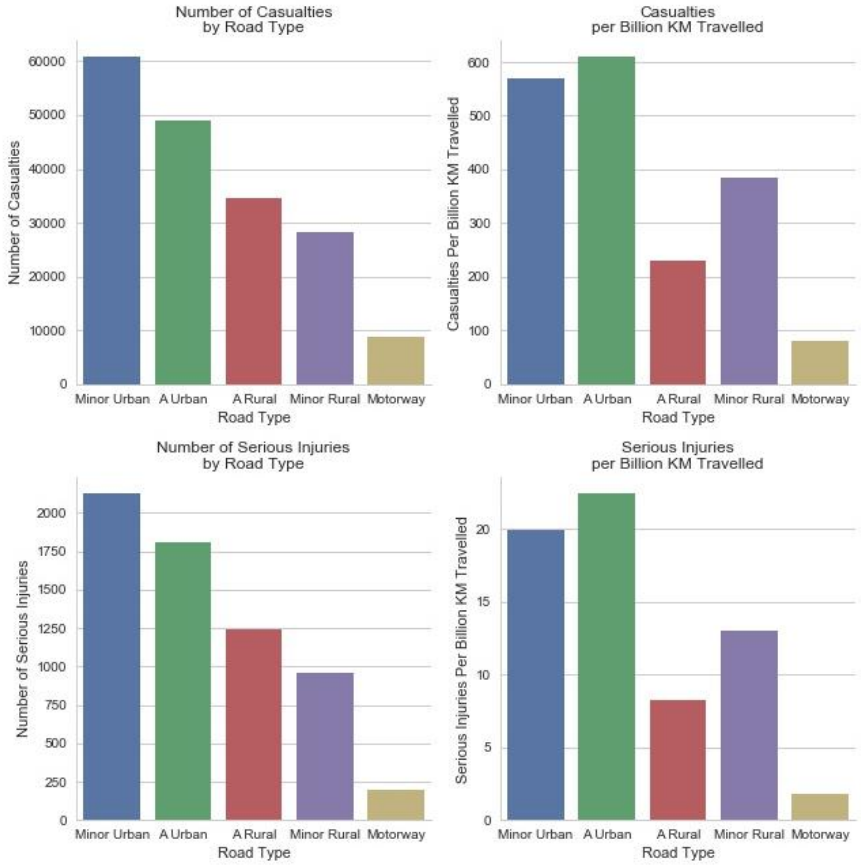

Figure 9: Casualties and serious injuries on different types of road in the UK in 2016

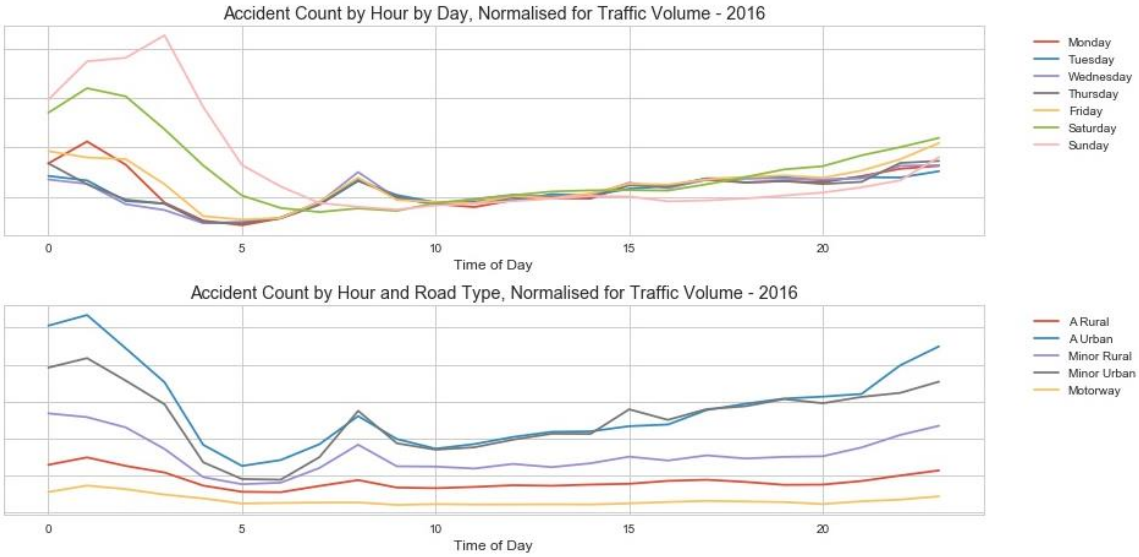

Figure 10: Top - Accident rate for each hour of each day of the week, normalised for traffic volume, in the UK in 2016. Bottom - Accident rate on different road types by hour for the UK in 2016, normalised for traffic volume

4.4 Road Type Categorisation

Table 2: This table shows the mapping used to merge traffic count estimates in to the main road safety data set. The road safety data set's classifications were updated as they were at a more granular level, so translation the other way was not possible.

Road Safety Data Values		Traffic Counts Values
A	Urban	A Urban
B	Urban	Minor Urban
C	Urban	Minor Urban
Unclassified	Urban	Minor Urban
Motorway	Urban	Motorway
A(M)	Urban	Motorway
A	Rural	A Rural
B	Rural	Minor Rural
C	Rural	Minor Rural
Unclassified	Rural	Minor Rural
Motorway	Rural	Motorway
A(M)	Rural	Motorway

4.5

4.6 Full List of Data Attributes

Table 3: This table provides a full list of all data attributes available in the 2016 Road Safety dataset.

Table	Columns
Accidents	Accident_Index
	Location_Easting_OSGR
	Location_Northing_OSGR

Table	Columns
	Longitude
	Latitude
	Police_Force
	Accident_Severity
	Number_of_Vehicles
	Number_of_Casualties
	Date
	Day_of_Week
	Time
	Local_Authority_(District)
	Local_Authority_(Highway)
	1st_Road_Class
	1st_Road_Number
	Road_Type
	Speed_limit
	Junction_Detail
	Junction_Control
	2nd_Road_Class
	2nd_Road_Number
	Pedestrian_Crossing-Human_Control
	Pedestrian_Crossing-Physical_Facilities
	Light_Conditions
	Weather_Conditions

Table	Columns
	Road_Surface_Conditions
	Special_Conditions_at_Site
	Carriageway_Hazards
	Urban_or_Rural_Area
	Did_Police_Officer_Attend_Scene_of_Accident
	LSOA_of_Accident_Location
Casualties	Accident_Index
	Vehicle_Reference
	Casualty_Reference
	Casualty_Class
	Sex_of_Casualty
	Age_of_Casualty
	Age_Band_of_Casualty
	Casualty_Severity
	Pedestrian_Location
	Pedestrian_Movement
	Car_Passenger
	Bus_or_Coach_Passenger
	Pedestrian_Road_Maintenance_Worker
	Casualty_Type
	Casualty_Home_Area_Type
	Casualty_IMD_Decile
Vehicles	Accident_Index

Table	Columns
	Vehicle_Reference
	Vehicle_Type
	Towing_and_Articulation
	Vehicle_Manoeuvre
	Vehicle_Location-Restricted_Lane
	Junction_Location
	Skidding_and_Overturning
	Hit_Object_in_Carriageway
	Vehicle_Leaving_Carriageway
	Hit_Object_off_Carriageway
	1st_Point_of_Impact
	Was_Vehicle_Left_Hand_Drive?
	Journey_Purpose_of_Driver
	Sex_of_Driver
	Age_of_Driver
	Age_Band_of_Driver
	Engine_Capacity_(CC)
	Propulsion_Code
	Age_of_Vehicle
	Driver_IMD_Decile
	Driver_Home_Area_Type
	Vehicle_IMD_Decile