Seasonal Traffic Accident Analysis

John McCabe
School of Mathematics, Computer Science & Engineering
City University, London

Introduction

I noticed recently how difficult driving conditions become during the winter, this got me thinking; what is it that makes it so difficult? It seemed that it wasn't a single factor, but a combination of factors such as low sun and frost, rain and dark, increased traffic when in bad weather. An extension of that thought was to consider whether the causes of accidents vary seasonally and if that is the case is it possible to isolate those factors or combination of factors, ultimately to determine counter measures.

Data

I will be using multiple data sets.

1. UK Road Safety: Traffic Accidents and Vehicles[1]. This is downloaded from Kaggle and contains two data sets.
    a. **Accident Information:** Detailed accounts of UK traffic accidents. The includes details of geographical locations, weather conditions and day or night. The data also includes a host of other features including severity, speed, road conditions and how the accidents occurred. Single row per accident.
    b. Vehicle Information: Contains details of all the vehicles and passengers involved in the accident. Single row per vehicle, implication being that it is a one to many relationships between accident data and vehicle data.
2. UK Traffic Counts[2]. This is downloaded from Kaggle. There are several data sets available here, but I will be using:
    a. Raw-count-data-major-roads: Containing manual counts of vehicles (and type) that flowed past a given point on a given day, split by hour of day.
    b. **Raw-count-data-minor roads:** As per above but for minor roads.

The analysis of traffic data is beset with challenges as highlighted by Abdulhafedh[3], which although US focused appears relevant for UK accident analysis. The datasets have been chosen to address these challenges by marrying accident data with traffic count data. Intuitively, the number of accidents is related to the number of cars, but in order to dig into other causes it is necessary to be clear of the influence of traffic volume against other conditions.

Research questions

Are there seasonal factors that influence the number of traffic accidents? If so, is it possible to isolate the factors or combination of factors that are the root cause.

Analysis Plan

Data investigation

The first step will be to check the variables of the data sets. As I will be looking to align the data sets over space and time, these investigations will center around the timeframes and locations of the data points.

Taking the UK as whole would be very difficult due to local influences i.e. certain areas will have localized influences, London for instance will have very different challenges to a small village, therefore area selection is crucial. My starting point is to identify the most normal places in the UK [4], then to verify suitability, produce a map of each area with the accidents plotted. Ideally, I would be looking for some geographical dispersion and a small variation in overall size. Once the regions have been identified, the next step is to determine a suitable distance to extend to, with a view to incorporating as many accidents data points without extending into other regions. The maps in-line with the accident counts will determine this.

Alignment of traffic count data with accident data

The next step is to find a way to attach traffic volumes to the accident data. As accidents can occur any time and traffic counts are specific points in time at specific locations, I need to settle on a spatial-temporal measure that aligns the two data sets. This is to be determined by visualizing the counts and accident data by various time frames, then using a clustering algorithm to combine counts data at the same time within a small location. The idea being that if traffic volume will not vary much over a small distance. Once the counts data has been clustered, I will then run the same spatial-temporal function extending out from the count clusters to the accident data, effectively attaching a counts cluster (including both mean number of motorized vehicles and relative volume at that location) to each of the traffic accident data points. There will be several traffic accident points that will not have a counts data; these will need to be dropped.

Feature Analysis

I will need to categorize the traffic volume variable before analyzing the significance of each accident feature. I will use a simple measure of traffic/ accident density as the ‘number of accidents / mean count of vehicles’ (across all clusters). Plotting this calculated variable against each cluster will provide the intelligence need to establish suitable bin edges.

Once the traffic/accident measure is in place, I will use heatmaps to visualize each feature for each traffic/accident band to ascertain whether there is significant variation by month.

The final piece of analysis will be to run correspondence analysis, plotting the derived components by the traffic/accident band, again to determine whether there are any significant differences between the seasons.

Analytical Steps

Data Investigations

Area Investigations

The following towns were selected based on being the most normal towns in the UK.

- Didcot
- Bath Road area, Worchester
- Southwick, West Sussex
- East Leake, Nottinghamshire

After attempting to plot all accident data and creating an un-manageably large file, I took the decision to just plot 2016 accident data.

Reviewing the plots for the accidents with varying distances (10km, 20km and 30km) from the center of the town resulted in a choice of 30kms as the best distance choice. This was the distance that incorporated the most amount of accident data points, without encroaching (too much) onto metropolises such as Birmingham and London. Figure 1, shows how the accident data points varied, resulting in the selection of distance.

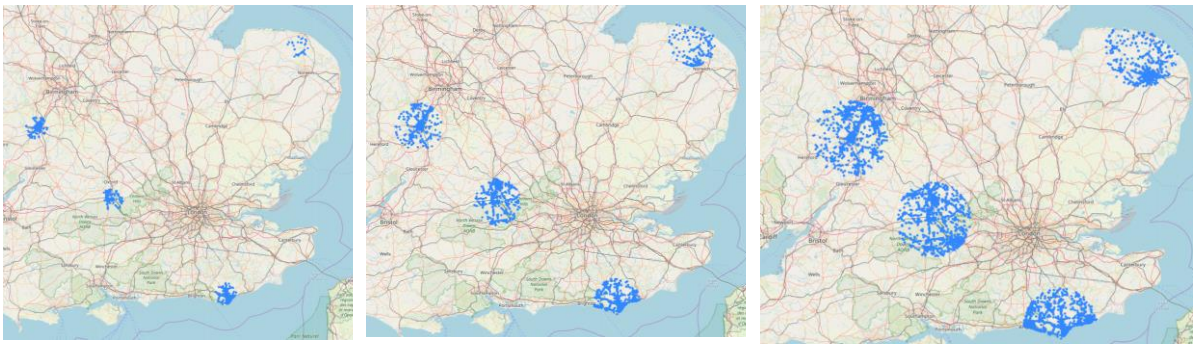

Figure 1. Plots of accident data points within varying distances from centres of the the most “normal” towns in the UK.

Time Investigations

The first step with the check the initial hypothesis of seasonal variation of the number of accidents that occur. In order to gauge whether this was correct, I created heatmaps, plotting month by year and the number of accidents. When reviewing it was clear that the years before 2010 were dominating. The following article [5]highlights the decline in traffic due to the financial crisis, road engineering improvements and improved car safety as the major factors. The pattern of both accidents and road deaths has remained relatively consistent since 2010 and as such I took the decision to remove data pre-2010 from my analysis. Once pre-2010 data was removed and the heatmaps re-run (figure 2), the heatmaps do indicate that there are indeed more accidents from September onwards in 3 of the 4 towns.

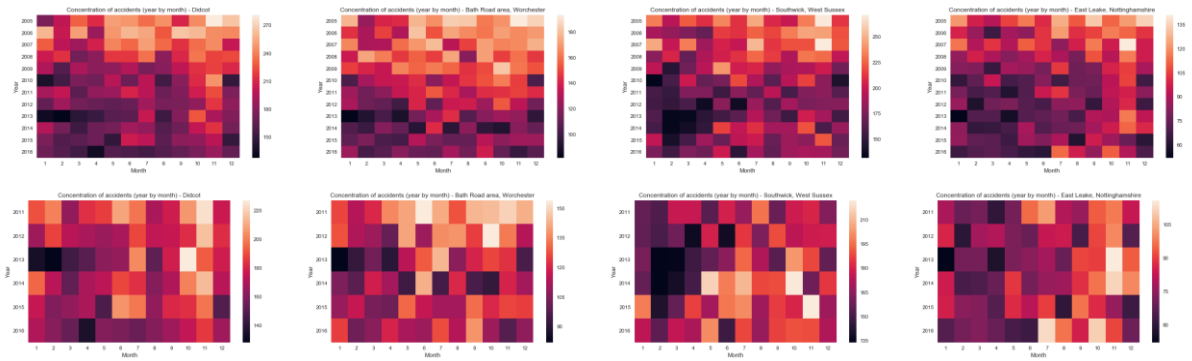

Figure 2. Top: Heatmap showing concentration of accidents from 2005 to 2016. 2005-2009, dominate. Bottom: Heatmaps showing 2010 onwards. With no years dominating it is possible to see that there is some seasonal variation in the number of accidents.

Reviewing the count data, highlighted that counts are only carried out at certain times of year and at certain times of the day. Figure 3 shows that counts only occur Monday to Friday between 7am and 7pm, in the following months; March, April, May, June, July, September and October.

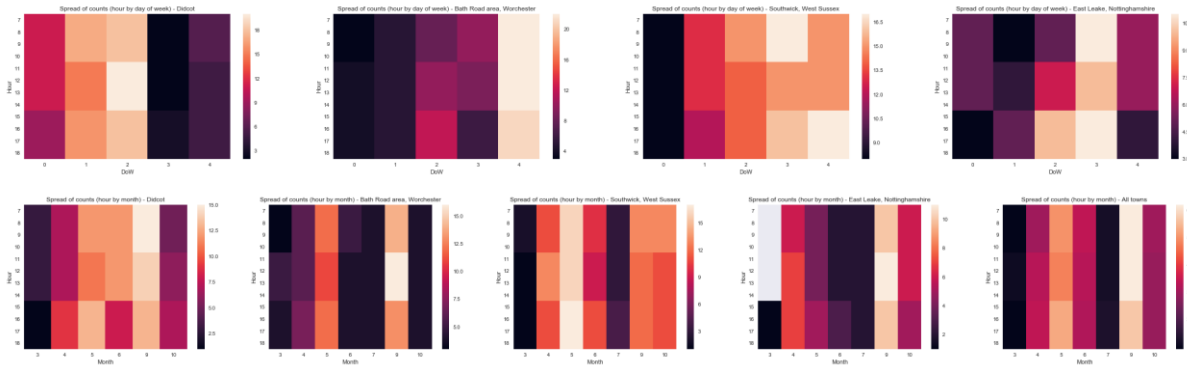

Figure 3. Top: Traffic counts only occur on certain days, at certain times in certain months.

This limits the scope of my investigations. Whereas I was keen to investigate the seasonal affects, I will now be limited to just investigating the periods where counts occur. This does raise a new question, why are counts limited to these periods and should they be?

Alignment of traffic count data with accident data

As I was focusing in on four locations, it was possible to drop all traffic counts that fell outside of the same distance from the center of each town as per the traffic accidents. Figure 4, shows the dispersion of traffic counts and traffic accidents, providing some intuition that the approach is feasible as there is clearly enough geographical overlap.

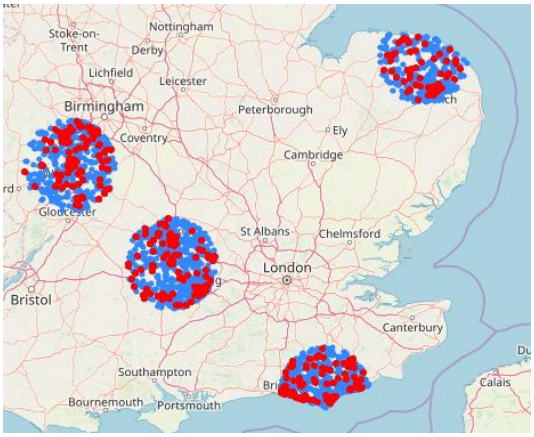

Figure 4. The blue dots represent traffic accidents whereas the red dots show where traffic counts have occurred in relation. This shows that there is clear overlap and as such we should be able to align traffic count data with traffic accident data.

Reviewing the map, also provides insight into the distance to stretch out from, from a traffic count to encompass traffic accident data. With the end-to end of each cluster being 60km, it would appear to be wise to try 2km-5km from each count point, which is likely to draw in most of the traffic accident data points.

Now that there is evidence of geographical overlap, I need to determine what timeframe to use for the space-time We have already seen from the accident data that there is “time of day” and “monthly” variation, so these will need to be fixed times that we need to look across.

Double checking the distribution of traffic volumes by “time of day”, “month” and “year”, shows that there are clearly varying volumes of traffic during the day (as expected). Figure 5 also show that there is considerable variation of traffic volumes on a monthly and annual basis. Closer inspection shows this is due to the number of traffic counts carried out rather than there being a clear pattern of traffic volumes over these periods

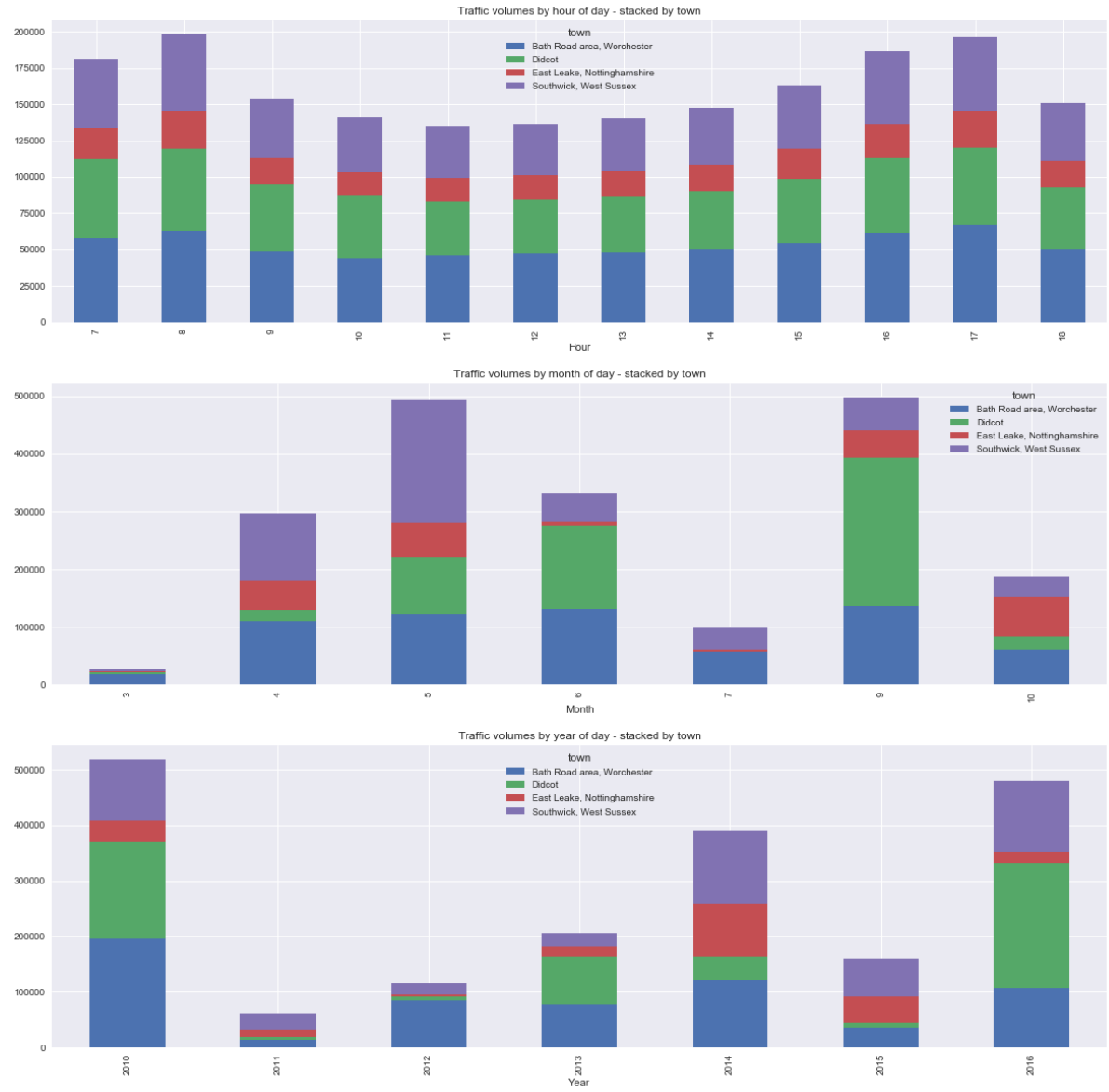

Figure 5. Summary of traffic volumes. The monthly and annual variations are due to number of counts, whereas the daily variation shows a clear variation of traffic volumes over a day.

Further analysis via DBSCAN clustering of the traffic counts data, with a distance variation of 0.2 km and a time distance of 1 day of the traffic count data (figure 6) clearly shows traffic counts last a whole day.

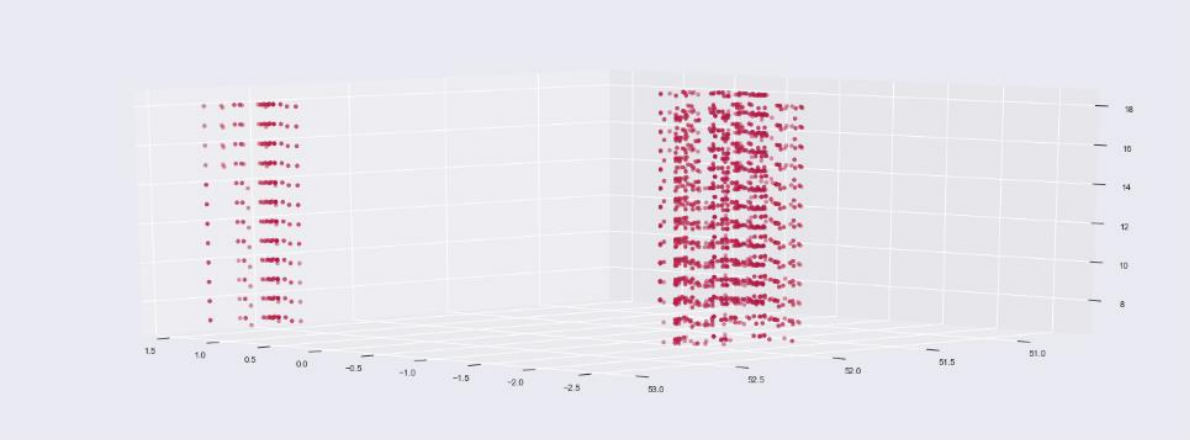

Figure 6. The columnar nature of the results of using DBSCAN clustering on the traffic counts data with max. distance 0.2km and max. time distance of 1 day shows duration of traffic counts.

As there may be multiple traffic counts that happen at the same time in the same area, an initial clustering exercise was undertaken to group these. The final decision was to use a geographical distance of 5km and a time distance of same hour, same month across days and years as the parameters for time-distance function applied to DBSCAN to initially cluster the traffic counts. Figure 7, shows the results of this clustering exercise.

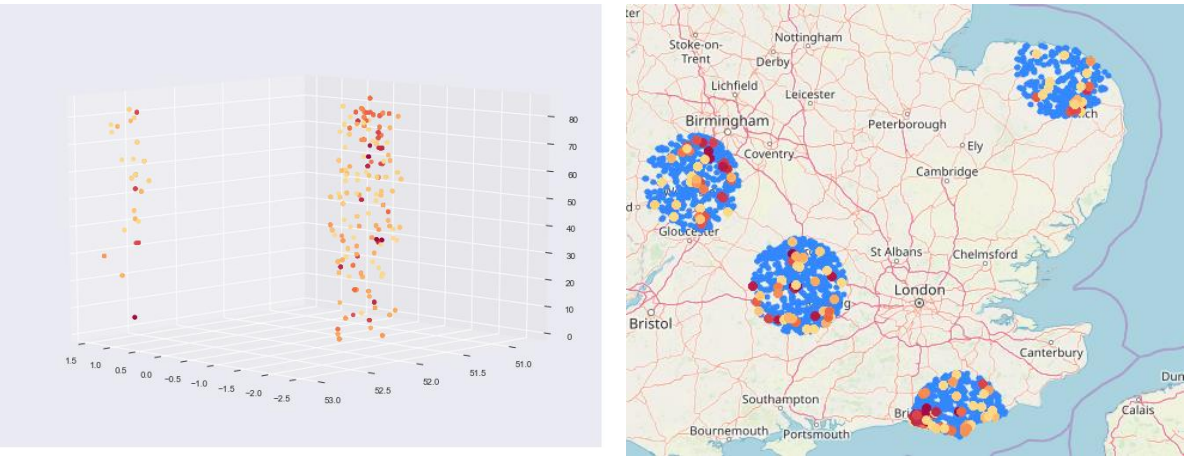

Figure 7. Left: Traffic count clusters in distributed by time and distance. Right: geographical view of clusters and accidents data.

The computation time required to match traffic accident data to the traffic count data wasn't available so to short-cut this, I took the approach of clustering the traffic accident data and then matching the clusters based on a mid-position. After running this there were naturally a number of traffic accident data points that didn't fall into a traffic count cluster and as such was dropped from the analysis. The results of this are shown in figure 8.

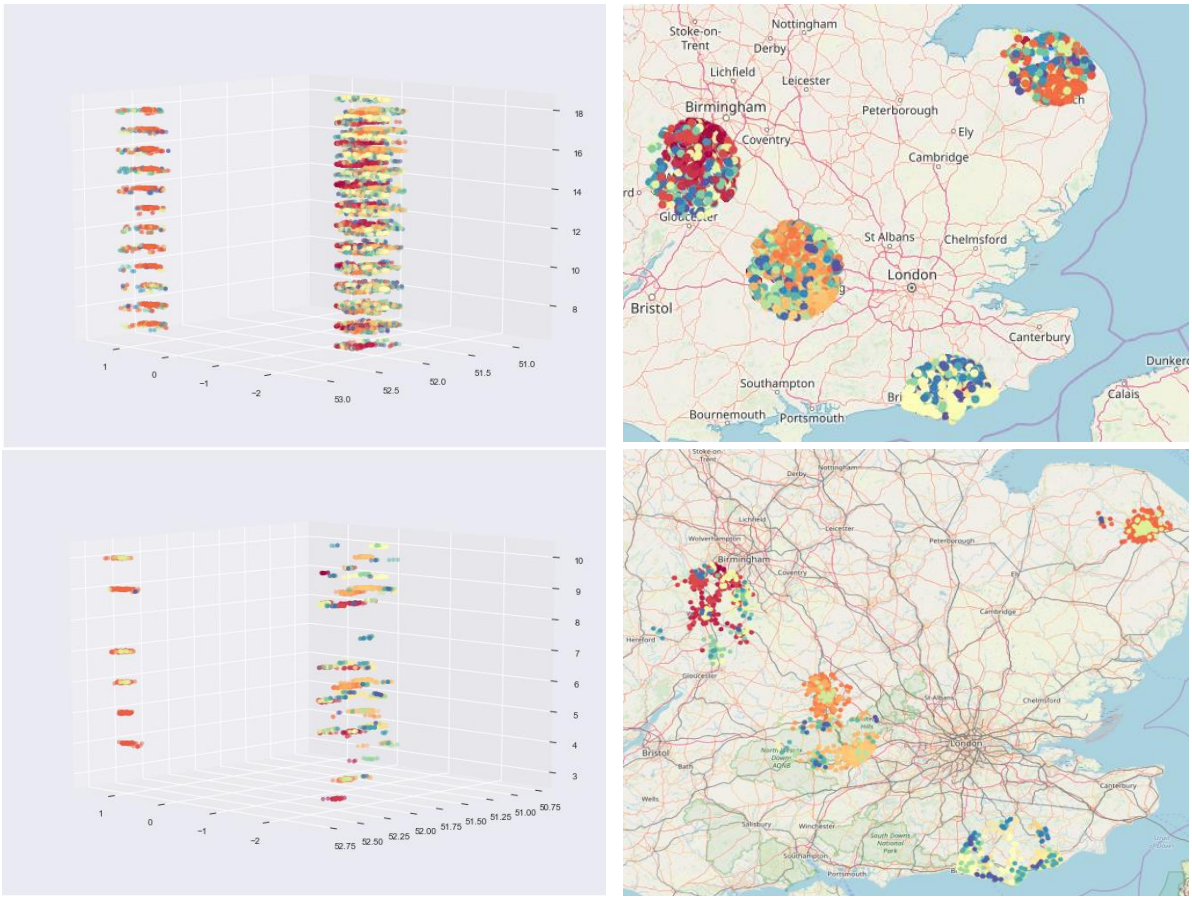

Figure 8. Top: Results of running DBSCAN on traffic accident data, with 5km distance measure, again same day, same hour but across years and days of week. Bottom: Results of traffic accident clusters matched to traffic count clusters.

Feature Analysis

With the number of vehicles passing the point of the accident, it was possible to create an accident by volume measure (# accidents / # vehicles). The next step was to categorize so that detailed analysis could be carried out on the attributes. I plotted the accident/volume data by cluster as shown in figure 9 to determine what suitable bin edges would be.

Figure 9. Top: Plot of number of accidents by number of vehicles in the vicinity. Bottom: Same plot, zoomed into 0-1 range due to dense cluster of points in that range.

I chose the ranges of 0-0.2, 0.2-0.5, 0.5+. The imbalanced ranges were designed to include enough data points for analysis.

Using heatmaps seemed a good way to determine if there were any discernable patterns between the low, medium and high accident counts. The first run of the heatmaps highlighted the fact that there is a dominant category in most of the attributes so to reduce the influence of these categories I used the percentages of the categories, restricted to 20%. The results of these are shown in figure 10.

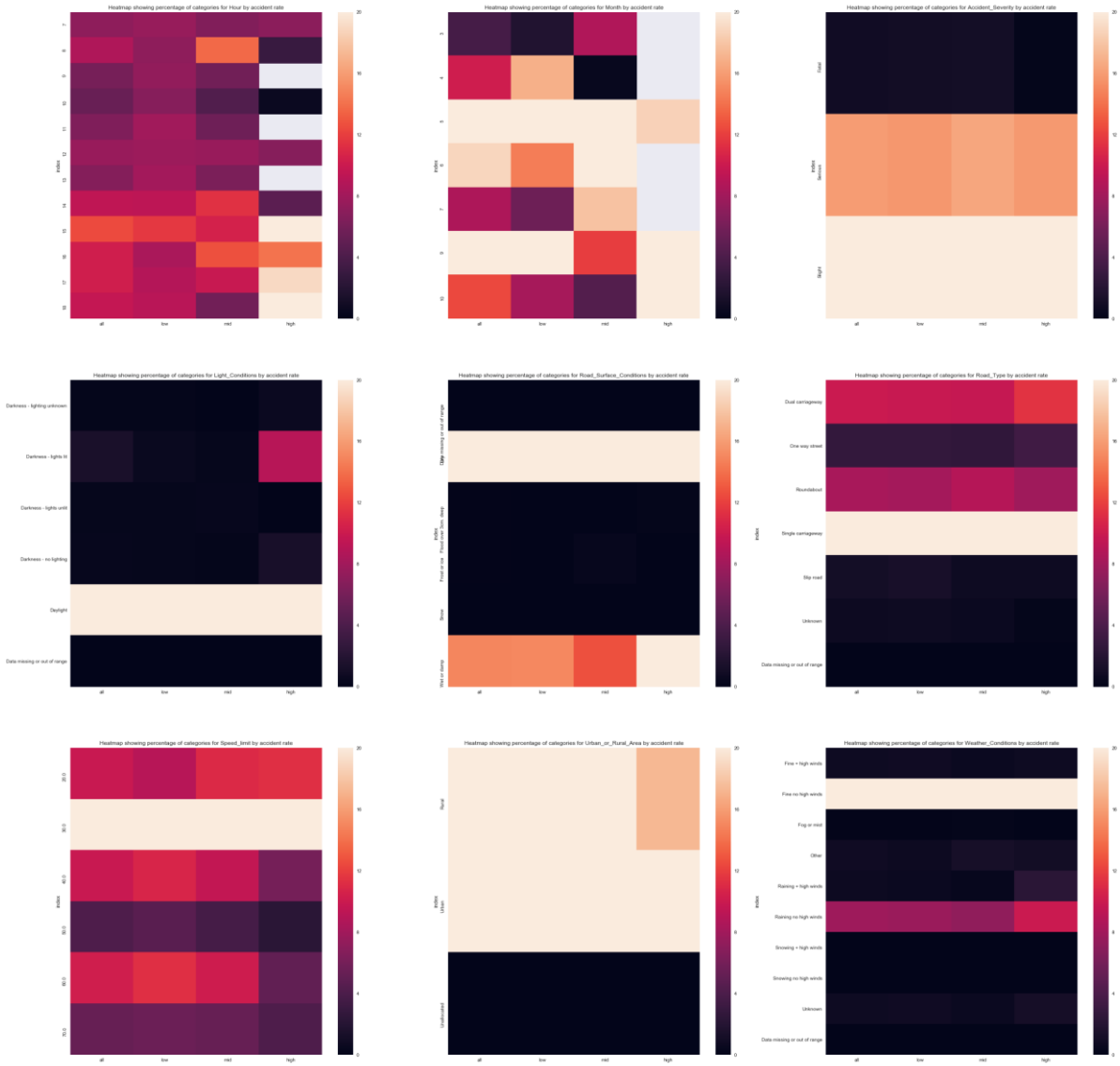

Figure 10. Heatmaps showing accident/ volume category (x-axis), key descriptive features from the traffic accident data (hour, month, accident severity, light conditions, road surface conditions, road type, speed limit, urban or rural area and weather conditions).

The final piece of analysis was to carry out correspondence analysis (CA) on the features. To effectively check for variations across seasons as per the goal and to mitigate the effects of traffic volume, CA was needed to be run on each attribute pair split by month and accident/ volume category. I have included a sample of the outputs in Figure 11 to illustrate.

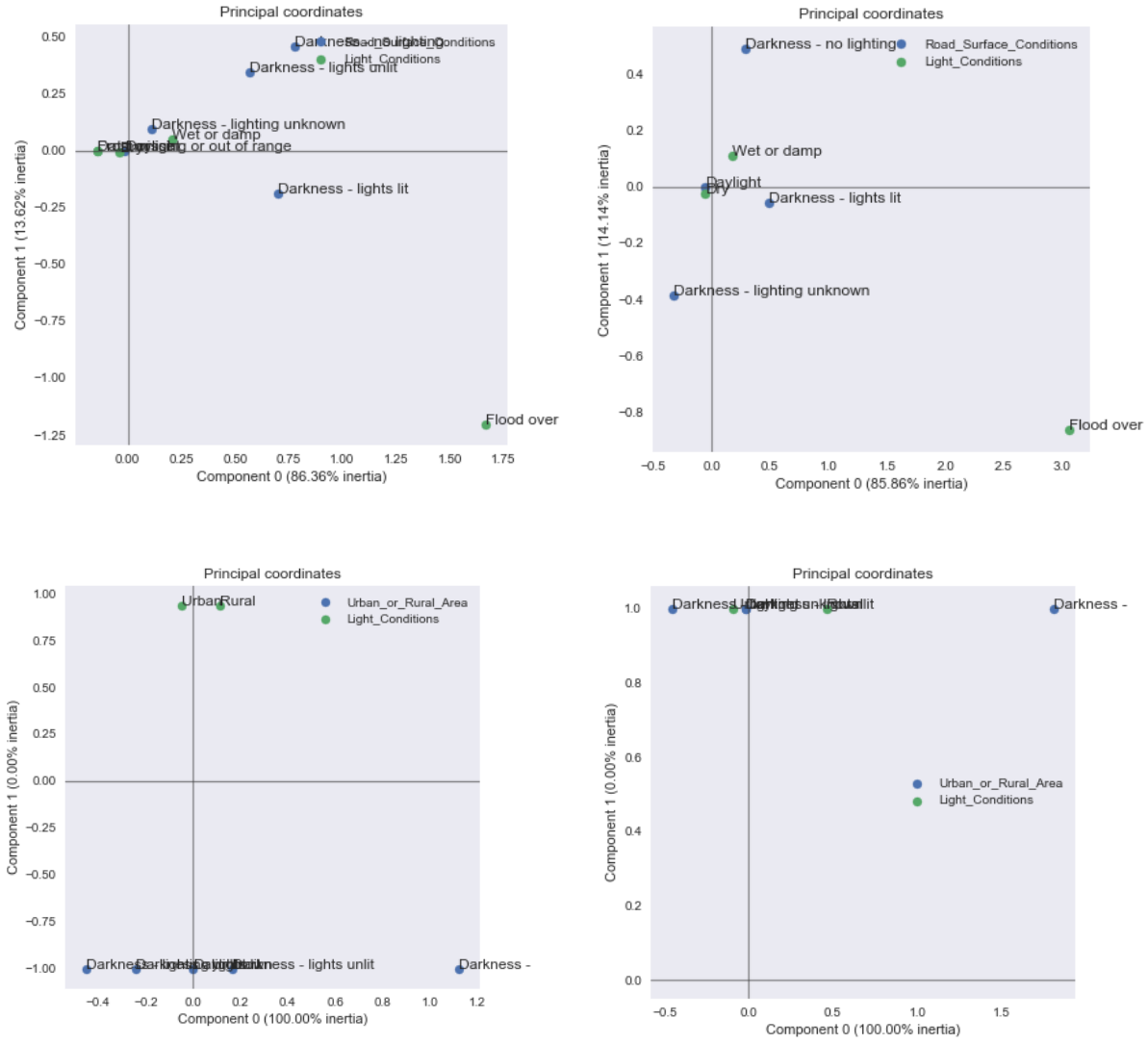

Figure 11. Left top and bottom: Principal component post CA showing correlations between attribute categories for all traffic accident data. Right top and bottom: As per left but filtered on accident/volume category = “high”.

Findings

Reviewing the heatmaps of attributes by accident/volume category for each category show:

Attribute	Low	Medium	High
Day	Very little variation across all hours	Small increase in the number of accidents in the evening.	Hotspots in the early evening, from 3pm to 6pm.
Month	Hotspots in May and September.		Hotspots in September and October.
Accident severity	Very few "Severe", dominated by "Slight" followed by "Serious" for all categories		
Light Conditions	Dominated by "Daylight"		Significant hotspot for "Darkness – lights lit" and "Darkness – no lighting"
Road surface conditions	Dominated by "Dry" followed by "Wet or damp".		Equally dominated by "Dry" and "Wet or damp"
Road type	Similar across all categories, dominated by "Single carriageway"		
Speed limit	Same pattern of 30mph dominating followed by 20mph, 40mph and 60mph.		Significantly different pattern to other two categories. 30mph dominates but has little concentration other than 20mph
Urban or rural	Equal "Urban" and "Rural"		"Rural" accounts for less than "Urban"
Weather conditions	Same pattern of "Fine no high winds" followed by "Raining no high winds"		Similar pattern, varies slightly by higher concentration of "Raining no high winds"

Reviewing the heatmaps seem to indicate that the "high" category has a different dispersion than low or medium. This would seem to indicate that high volume traffic conditions result in the recording of different circumstances and reasons for the accident than when traffic volumes are low and medium.

Due to the sheer volume of correspondence analysis plots it has been unable to review all the plots. Instead I have focused on the difference between month and all other attributes for just all traffic accidents and high volume, principally to help explain the variation seen from the heatmaps. After reviewing the CA plots side by side, there was no clear differences of correlation between all traffic accidents and those during busy periods or on busy roads.

Reflections

Implications

Ultimately this analysis has been able to answer the question posed. The results are not conclusive and although the heatmaps seem to indicate that there is variation in the factors that are recorded as conditions at the time of the accidents, the correspondence analysis was unable to uncover any clear difference of correlation between attributes of the accident and month the accident happened.

Effectiveness

As mentioned, the analysis was not as effective as hoped for. A potential reason for this was the relatively low number of accidents that was carried through to the final analysis. There were several reasons why the final sample was low, but the main reason was the limited reach of the traffic count data. As the analysis was focusing in on seasonal variation, it was always going to be a challenge once I had discovered that traffic counts are not carried out during the winter. It is clear from the accident data that there is an increase in accidents during this period and is also the time when the weather changes dramatically, which in turn impacts road conditions so was likely to have a big influence. Unfortunately, I was too far into the analysis to a) reverse out or b) find alternative data sources. The lack of traffic count data outside of the working day would seem to have less impact on the analysis but again would have impacted the quality of the analysis.

The visualisations worked effectively where they were used. In particular the identification of a big shift in the number of traffic accidents before and after 2010. This would not have been picked up without visualizing the data and as a result would have meant having out of date data used in the model.

One area that did appear to work well was the alignment of the two data sets, this was highlighted as a major challenge and running a clustering algorithm on first the traffic count data and then again on the accident data reduced the number of times the time, distance function had to be evaluated from millions to thousands. The visualisations worked well in parallel with the analytical modelling for selecting appropriate parameters to take forward and to verify the successful implementation.

The model could be improved, with more time and research into when how and where the traffic counts are carried out with regards to direction of travel would provide better results. My analysis has not considered direction of traffic and has generalised to the location.

Using heatmaps to determine variations of features was an effective way of distinguishing variations in concentrations of traffic accidents but did obscure the low volumes, but the low volumes would have also caused pronounced effects. Further investigation would be required to see whether there was real differences between low/ medium traffic volumes and high traffic volumes as pointed to by the heatmaps.

I believe the correspondence analysis would have been more effective if I had more time available. The CA plots I reviewed clearly showed how the features correlated. It is likely to have been more applicable with feature selection method early on in the process in order to reduce the final number of plots to review.

Generalisation

As referred to above, this analysis is severely limited due to the restrictions with the traffic count data, without this covering the key periods it is unlikely to be a generalised approach, certainly in identifying season variation.

That said, linking the traffic count data with the traffic accident data was successful and with that it may be effective in determining certain accident hotspots. Figure 9, the scatterplot of accidents/ number by cluster showed some extreme cases which would be worth further investigation.

---

[1] "UK Road Safety: Traffic Accidents and Vehicles." Accessed December 17, 2018. https://www.kaggle.com/tsiaras/uk-road-safety-accidents-and-vehicles.

[2] "UK Traffic Counts." Accessed December 17, 2018. https://www.kaggle.com/sohier/uk-traffic-counts.

[3] Abdulhafedh, Azad. "Road Traffic Crash Data: An Overview on Sources, Problems, and Collection Methods." Journal of Transportation Technologies 07 (March 13, 2017): 206. https://doi.org/10.4236/jtts.2017.72015.

[4] "'Most Normal' Town in England Unveiled," March 29, 2017, sec. Oxford. https://www.bbc.com/news/uk-england-oxfordshire-39428314.

[5] D'Urso, Joey, and Rachel Schraer. "Driving Home? 10 Things to Know about Roads," December 22, 2017, sec. UK. https://www.bbc.com/news/uk-42182497.