Using k-means clustering to group locations in the UK according to the frequency of different road accident severities.

Road traffic accidents are currently one of the biggest concerns for global health organisations and national governments, particularly concerning the level of injury caused. Understanding the factors which contribute to accident severity and using these to predict accident severity for new crashes is hugely researched. Many studies have found that classifying accident severity is complex and not straight forward. As environmental factors appear to have some influence on accident severity, it is possible that environmental heterogeneity in the large areas used in many studies may contribute to accident severity in different ways. Using k-means clustering, this investigation set to cluster areas into groups according to their accident severity frequency using UK traffic accident data from 2016 to 2020. These clusters were then plotted on a map to identify areas visibly.

1. Introduction

The increasing numbers of motor vehicles across the globe has become a major concern for health organisations and national governments. In 2018, the global number of annual road traffic deaths had reached 1.35 million, with road traffic-related injuries being the leading cause of death for people aged between 5 and 29 years (World Health Organization, 2018). The World Health Organisation has declared that drastic action is required and in 2020, they set a target of halving road traffic deaths by 2030 as part of their resolution A/RES/74/299 "Improving global road safety". They aim to achieve this through improving road and vehicle design, laws and immediate emergency care for injured parties. Understanding what factors contribute to road accidents and the differentiation between those that cause accidents to be minor verses fatal can help to meet these goals by allowing more directed safety measures to be put in place.

Montella et al. (2012) defined a crash as “a rare, random, multi-factor event always preceded by a situation in which one or more road users fail to cope with the road environment”. Research into discovering and understanding these factors preceding a crash have been studied widely for a number of years. Initially statistical models were used to classify accident severity and determine important contributors to accident severity. Some of the most utilised methods included multinomial logit, ordered logit, ordered probit, binary logit, binary probit and nested logit (reviewed in detail by Savolainen et al. (2011)). However, statistical models are limited due to pre-defined underlying relationships between the dependent variable (accident severity) and the independent variables (potential accident risks) and model assumptions; for example, assuming independent variables are independent of one another (Chang and Wang, 2006). This has led to a rise in the popularity of applying machine learning methods, as these non-parametric models do not assume any relationship between variables or distribution (Wen et al., 2021). Machine learning tools are able to identify patterns and extract previously unknown information from large datasets. A variety of these techniques have been applied to road accident data and have been able to highlight a multitude of potential contributing factors. Environmental variables seem to consistently play a role in road accidents and in influencing crash severity, such as road condition, urban verses rural, weather, road type (Silva et al., 2020).

A potential issue with previous machine learning studies into accident severity is that many of these studies apply their models to road accident datasets across large areas which may have a lot of heterogeneity in environmental characteristics (Kumar and Toshniwal, 2016). It is possible that accidents and different accident severities may occur more or less frequently across small-scale areas (Kim and Yamashita, 2007, Puspitasari et al., 2020). The generalisation of these potential small-scale differences could mean that contributing factors are not picked up by models due to their effect on only a small concentration of accidents. Looking at the features in areas of high crash frequency or high fatal crash severity may reveal hidden information (Kumar and Toshniwal, 2016). K-means has been used as a method of grouping accident locations based on accident frequency previously. For example, Puspitasari et al. (2020) used k-means clustering to identify specific areas along the Jakarta Bogor Highway that had a higher frequency of accidents, allowing managers of the highway to target precisely where intervention is needed. These clusters can also improve other data mining model ability to extract information. Kumar and Toshniwal (2016) created clusters of areas with high-, medium- and low-frequency of accidents. They then applied association rules mining to extract differences in contributing factors for each group.

The aim of this study is to use k-means clustering with the aim of grouping locations based on the frequency of different accident severities. The data used is from the UK Department for Road Transport and includes details of UK road accidents reported by the police in the years 2016 to 2020. Two software for developing machine learning models are R and SAS Enterprise Miner (SAS EM). R is a free open-source software for statistical computing, with a variety of add-on packages which provide many built-in functions, including datamining techniques (R Core Team, 2021). SAS EM is a software which claims to be able to “streamline the data mining process to develop models quickly” (SAS Institute Inc., 2015). This tool aims to allow non-technical users with limited statistical skill to easily generate models and results which are easy to interpret. An additional aim of this study is to compare the results generated using k-means clustering by these two platforms.

2. Literature Review

When investigating and modelling traffic accident severity using data mining, a number of different methods have been applied due to their ability to outperform statistical methods. Li et al. (2012) compared the performance of support vector machine (SVM) models against an orbit probit model in predicting injury severity. Another comparative study also found that SVM models performed well, along with Random Forest, when compared to Mutinomial Logit and Nearest Neighbour Classification (Iranitalab and Khattak, 2017). A lot of the studies that use data mining techniques to try to classify accident severity using attributes associated with individual crashes. Often the accuracy of these models is fairly low, with some only having an accuracy of 50-60% (Silva et al., 2020). This suggests there is a lot of complexity with the interactions between factors that ultimately leads to the severity of an accident. Many studies use data from across large areas, such as whole countries and regions to create these models (Feng et al., 2019, Xu et al., 2018, Yu et al., 2019, Tavakoli Kashani et al., 2011). As environmental factors have often been found to have some effect on accident severity by many studies (Zeng and Huang, 2014, Pakgohar et al., 2011, Amiri et al., 2020, Delen et al., 2017), perhaps the heterogeneity across the large areas covered by these studies makes finding patterns in the factors which cause accident severity very difficult. It may be possible that accidents in different areas are associated with different factors which determine the accident severity outcome.

Many studies have used clustering to group road traffic accidents into those with similar characteristics. Kim and Yamashita (2007) used k-means clustering to identify areas with high numbers of pedestrian involved crashes, which they then said could be used to investigate socio-economic correlates with frequency of crashes involving pedestrians. Another study applied k-means clustering to group traffic accident locations into high-, medium- and low-frequency groups (Kumar and Toshniwal, 2016). They found that high-frequency locations tended to be highways or non-highway roads connecting cities. Accident attributes have also been used to cluster accidents, with one study identifying that the largest cluster seemed to support the hypothesis that younger people in urban areas being more likely to crash (Sinclair and Das, 2021). Almjewail et al. (2018) used k-means clustering to identify neighbourhoods in Saudi Arabia which had high accident frequency and found the common reasons for accidents seemed to be speeding and distracted driving. K-means clustering can also be used as a dimensionality reduction step before applying other techniques, such as association rule mining or random forest, improving the performance of the models (Yassin and Pooja, 2020, Nandurge and Dharwadkar, 2017).

3. Methodology

3.1 Dataset acquisition and preparation

The dataset used in this investigation was obtained from the UK’s Department for Road Transport government website. The dataset includes details of road accidents collected by the police using the STAT19 reporting system between 2016 and 2020 (Department for Public Transport, 2021). This reporting system is used to standardise the process of collecting details regarding a reported road accident involving at least one vehicle collision or a vehicle collision with a pedestrian. The initial dataset contains 597,973 recorded incidents, each with 36 attributes recorded. Each accident is given an accident severity level of either slight, severe or fatal (Department for Transport, 2021). A slight accident includes at least one person with a slight injury (an injury of minor character, such as bruises, cuts or sprains) but no further injury or death. A serious accident includes at least one person who is seriously injured (an injury requiring hospital in-patient treatment) but no one is killed. A fatal accident includes at least one death and potentially other injuries. The first step prior to modelling requires cleaning and preparation of the dataset into an appropriate format. This process was completed using R version 4.1.2 (R Core Team, 2021), which the following section will describe (full code can be found in Appendix A).

As the aim of this study is to cluster the frequency of different accident severities according to their location, an appropriate location variable needed to be selected. The dataset contained multiple variables which could locate an accident, including longitude and latitude, the northing and easting coordinates, the Lower Sayer Super Output Area (LSOA), the local authority district code and the local highway authority code. Due to high levels of heterogeneity in the longitude-latitude coordinates (513,168 and 536,231 distinct values respectively) and northing-easting coordinates (286,121 and 260,089 distinct values respectively) for each accident, these variables would have to be grouped to be able to calculate the frequency of accident severities for each area. As the other location variables essentially group the locations of accidents already, it makes sense to consider these instead. It would be preferable to use the LSOA of an accident, as this measurement is more fine scale and are roughly equal measurements according to population. However, the number of distinct LSOAs proved to be too large (34,146 distinct values) to cluster based on the computing power available. The next largest location variable, and so next most comprehensive, was the local authority district code with 383 distinct values.

The dataset was pivoted using the ‘tidyverse’ package (Wickham et al., 2019) [fig.1]. All variables, except the local authority district code and accident severity, were removed. Then accident severity was grouped by the local authority district code and the number of each accident severity according to the location was calculated. The dataset was then transformed using the ‘pivot\_wider’ function (Wickham et al., 2019). As the accident severities were provided in numeric format, these were transformed (1=’fatal’, 2=’serious’, 3=’slight’). The transformation of the data into its final format can be seen in figure 2.

> location\_df <- df %>%

+ select(accident\_severity, local\_authority\_ons\_district) %>%

+ group\_by(local\_authority\_ons\_district) %>% count(accident\_severity) %>%

+ pivot\_wider(names\_from=accident\_severity, values\_from=n) %>% rename(fatal = "1") %>%

+ rename(serious = "2") %>% rename(slight = "3") %>%

+ mutate(fatal = ifelse(is.na(fatal), 0, fatal),

+ serious = ifelse(is.na(serious), 0, serious),

+ slight = ifelse(is.na(slight), 0, slight)) %>%

+ as.data.frame()

**Figure 1:** R code used to transform dataset into appropriate form for clustering local authority districts by the severity of accidents.

**B**

**A**

|  |  |
| --- | --- |
| **Accident severity** | **Local authority district code** |
| 3 | E09000005 |
| 3 | E09000004 |
| 3 | E090000017 |
| 3 | E090000024 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Local authority district code** | **Fatal** | **Serious** | **Slight** |
| E06000001 | 11 | 115 | 381 |
| E06000002 | 8 | 158 | 883 |
| E06000003 | 14 | 157 | 487 |
| E06000004 | 18 | 215 | 849 |

**Figure 2: A)** Original dataset format. **B)** Dataset format used for clustering.

3.2 K-means clustering

Clustering is an unsupervised machine learning technique which groups data objects with the aim of data in each cluster being more similar to each other than to objects in other clusters. K-means clustering separates data into k numbers of clusters, with k being selected by the user (Tan et al., 2016). It is a partitional, prototype-based clustering method which means it divides data objects into separate groups with no overlapping and data objects are closer to the cluster centroid (average of all data objects within cluster) than to any other cluster centroid. To assign data objects to clusters, each point is initially assigned to one of k centroids and the centroid is updated based on the average of the points assigned. Data objects are then reassigned clusters if closer to any updated centroid than the one they are assigned to, and the centroids are updated to include new data points. This continues until no data points move. The distance between a point and centroid was calculated using Euclidean distance measurements (minimises the sum of squared error between a data object and its centroid).

3.2.1 R Implementation

The full code for this section can be found in Appendix B. The summary of the prepared dataset is provided in table 1. The difference between the maximum number of each accident severity is fairly large. This is because fatal accidents are rarer compared to slight accidents, so there is a lower frequency altogether. Additionally, the range between the smallest and largest frequency for each accident severity is also fairly large. This may affect the ability of the model to produce clusters which may have a higher frequency of fatal crashes relative to other areas and other accident severities. To improve this the data was normalised (1). Normalisation rescales the values and so makes the frequency of each accident severity more relative to each other and easier to compare.

**Table 1:** Summary table for fatal, serious and slight variables. Any decimal place is given to 1 d.p.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Accident Severity** | **Mean** | **Min** | **1st Quartile** | **Median** | **3rd Quartile** | **Max** |
| Fatal | 21.1 | 0 | 11 | 18 | 26 | 102 |
| Serious | 282.3 | 3 | 161 | 223 | 341.5 | 1912 |
| Slight | 1258 | 2 | 595 | 916 | 1424 | 10550 |

(1)

The R package used for the clustering process was ‘factoextra’ (Kassambara and Mundt, 2020). Clustering algorithms will produce clusters from datasets, even if the clusters produced are not significant and are meaningless. To test the clustering tendency of a dataset the Hopkin’s statistic was developed (Lawson and Jurs, 1990). This statistic compares the spread of the dataset points with a random scatter plot using Nearest Neighbours. If the dataset is similar to the random datapoints, then it likely has no clustering tendency, so will produce a low Hopkin’s statistic value. The more different (the closer to 1) the dataset is from the random dataset, the more clusterable the dataset is likely to be. To generate the Hopkin’s statistic the function ‘get\_clust\_tendency’ was used (Kassambara and Mundt, 2020). The value produced was 0.920, suggesting this dataset has a very high clustering tendency. K-means requires the number of clusters to be selected prior to clustering. To determine the number of clusters an elbow plot was produced. This plot demonstrates the increased level of information gained by adding clusters, demonstrated by the decrease in total within sum of squares. Adding too many clusters will decrease the amount of information gained as informative clusters become split unnecessarily. Therefore, the optimum number of clusters is chosen as the point just before there is a rapid decline in the amount of information gained, which is described as the ‘elbow’. In the plot produced the ‘elbow’ was not obvious so both 3 and 4 clusters used to decide which produced the best clusters [fig.3].

Chart, line chart

Description automatically generated

**Figure 3:** Elbow plot for dataset to determine optimal number of clusters.

The ‘kmeans’ function was used to conduct k-means clustering (Kassambara and Mundt, 2020). Then the function ‘fviz\_cluster’ was used to visualise the resulting clusters (Kassambara and Mundt, 2020). Figure 4 shows the result of setting k as 3 compared to setting k as 4. Using 4 clusters appeared to be more informative through more clearly identifying a group of outlying areas [fig.4B]. The areas in cluster 4 appear to be widely different, with groups of areas having higher values in dimension 1 and other groups having higher values in dimension 2. It was decided to remove cluster 4 to see if the remaining clusters cluster differently without the presence of outliers.

Graphical user interface, timeline

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

**Figure 4: A)** Plot of k-means produced clusters where k=3. **B)** Plot of k-means clusters where k=4. Colours of clusters indicate the separate cluster numbers. The point labels indicate the local authority district code for each point.

The dataset produced after removing cluster 4 contained 372 rows. The new elbow plot produced suggested that 4 clusters was the optimum [fig.5]. However, the clusters produced setting k=4 did not appear to group the points in the best way [fig.4A]. To see if adding a cluster was better able to group the areas, k=5 was used. Using 5 clusters appeared to produce clusters in which areas seemed to be more closely related when compared to using 4 clusters [fig.4A and fig.4B], so it was decided to keep this model.

Chart, line chart

Description automatically generated

**Figure 5:** Elbow plot for dataset with outliers removed to determine optimal number of clusters.

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

**Figure 6: A)** Plot of k-means produced clusters where k=4. **B)** Plot of k-means clusters where k=5. Colours of clusters indicate the separate cluster numbers. The point labels indicate the local authority district code for each point.

3.2.2 SAS Enterprise Miner Implementation

After importing the prepared dataset using the ‘File Import’ node, local authority code was given the role ID and the roles for fatal, serious and slight were set as the input. ‘StatExplore’ was applied to generate summary statistics. As with the R implementation, due to the large difference in the range of values between the accident severity and the large deviations for each one, the dataset was transformed using the ‘Transform Variables’ node. The method used was ‘Maximum Normal’ which selects the best transformation to maximise normality (SAS Institute Inc., 2017b).

**Table 2:** Summary table for fatal, serious and slight variables. Any decimal place is given to 1 d.p.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Accident Severity** | **Mean** | **Deviation** | **Minimum** | **Median** | **Maximum** |
| Fatal | 21.1 | 15.3 | 0 | 18 | 102 |
| Serious | 282.3 | 204.5 | 3 | 223 | 1912 |
| Slight | 1257.8 | 1117.3 | 2 | 916 | 10550 |

**Table 3:** Summary table for log-transformed fatal, serious and slight variables (3 s.f.).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Accident Severity**  **(log-transformed)** | **Mean** | **Deviation** | **Minimum** | **Median** | **Maximum** |
| Fatal | 0.182 | 0.113 | 0 | 0.163 | 1.54 |
| Serious | 0.133 | 0.0846 | 0 | 0.109 | 0.693 |
| Slight | 0.109 | 0.0853 | 0 | 0.0831 | 0.693 |

The ‘Cluster’ node in SAS EM has the option for SAS to automatically decide the appropriate number of clusters. SAS EM makes a preliminary clustering pass and then uses the multivariate means of the clusters as inputs for the second pass. The final number of clusters is the smallest number that satisfies four criteria. More detail on the methods used in the first pass, second pass and the four criteria can be found in the documentation (SAS Institute Inc., 2017a). However, this option produced 20 clusters so was not useful in this instance [fig.7A]. Instead, it is possible to set the maximum number of clusters manually, with the maximum possible number being 10. After trying multiple maximum cluster values, the number 3 was settled on [fig.7B]. The third cluster appeared to be made of outlier variables, with a root-mean-square standard deviation of 1.57 (3 s.f.) and the maximum distance from the cluster seed is 4.36 (3 s.f.), which is almost double the distance for the other clusters. Using the ‘Filter’ node, cluster 3 datapoints (7 in total) were removed.

Chart, pie chart

Description automatically generated Chart, pie chart

Description automatically generated

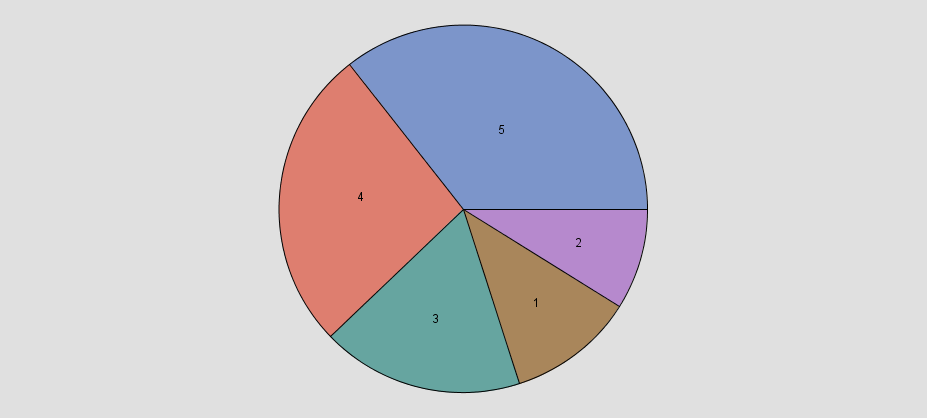
**B**

**A**

**Figure 7: A)** Segments produced by clustering using automatically generated number of clusters. **B)** Segments produced when the maximum number of clusters is set at 3. The numbers indicate the segment number.

Applying the ‘Cluster’ node again using the automatically generated number of clusters produced 20 clusters again [fig.8A]. So after manually selecting a number of clusters, 5 clusters appeared to generate the most informative and best distributed clusters [fig.8B], with the mean-root-square standard deviation ranging from 0.384 to 0.738 (3 s.f.) and the maximum distance from the seed cluster ranging from 1.45 to 2.42 (3 s.f.). The ‘Segment Profile’ node was then applied to understand the differences between the clusters. This will be discussed in section 4. The full pathway in SAS EM for the clustering analysis can be seen in Figure 9.

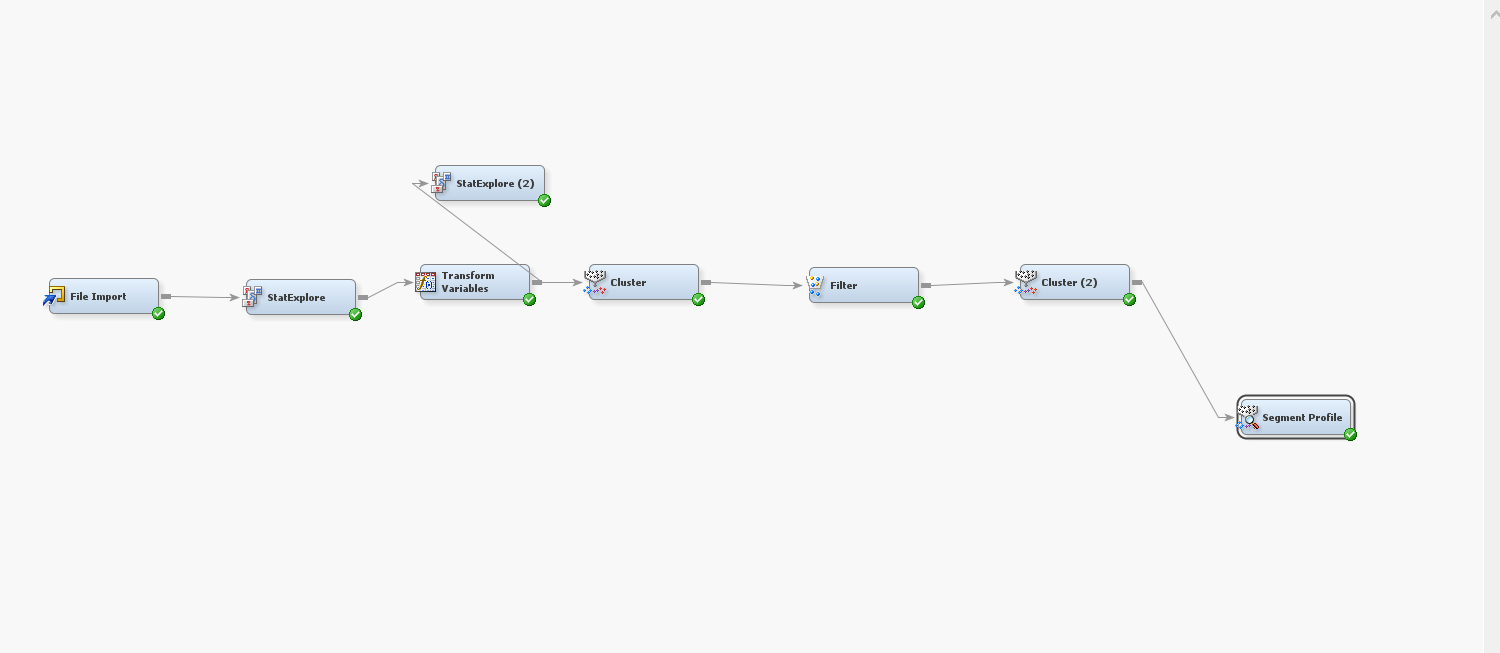
Chart, pie chart

Description automatically generated 

**A**

**B**

**Figure 8:** **A)** Segments produced by clustering using automatically generated number of clusters. **B)** Segments produced when the maximum number of clusters is set at 5. The numbers indicate the segment number.



**Figure 9:** The full path for clustering using SAS Enterprise Miner.

4. Results and discussion

K-means clustering was successfully able to cluster locations of accidents according to the frequency of different accident severities.

Starting with the clusters produced by R; cluster 1 seemed to contain locations with high accident frequency. Although, slight accidents had the highest value, serious and fatal accidents were still relatively high. On the other hand, cluster 2 and cluster 3 contained low accident frequency areas compared to the other clusters. For both of these, serious accidents were more prevalent. Cluster 2 seemed to have slightly more disparity between the severities, with fatal being very low. Cluster 3 has more equal prevalence of each severity. Both cluster 4 and cluster 5 contain areas with high prevalence of fatal accidents, with cluster 5 having a higher frequency of accidents.

**Table 4:** Cluster number with the corresponding count of locations within the cluster and the mean of the normalised frequencies for each accident severity (given to 3 s.f.).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Cluster** | **Count** | **Fatal** | **Serious** | **Slight** |
| 1 | 24 | 0.237 | 0.338 | 0.373 |
| 2 | 141 | 0.0948 | 0.0800 | 0.0598 |
| 3 | 51 | 0.162 | 0.189 | 0.181 |
| 4 | 112 | 0.228 | 0.112 | 0.0813 |
| 5 | 44 | 0.405 | 0.208 | 0.135 |

In SAS EM, each segment is the same as a cluster. Table 5 shows the worth of each accident severity for the clusters, which identifies which accident severity is more important in distinguishing the cluster. Figure 9 reveals the actual distribution of the frequencies of the different accident severities within each cluster against the population distribution. Fatal accidents have the highest worth for cluster 1, which makes sense considering cluster 1 contains areas with a high frequency of fatal accidents, shown by the small distribution on the right side of the graph [fig.9]. Areas in cluster 1 also seem to have higher frequencies of accidents altogether, with all accident severities having distributions concentrated around the higher frequencies compared to the population distribution. However, for cluster 2, fatal accident have negligible worth which is likely due to areas in this cluster seemingly having no fatal accidents but very high frequencies of slight and serious accidents. Cluster 3 is mostly distinguished by the slight accidents, with a slightly higher frequency of slight accidents compared to the slight and fatal. However, this difference is only slight, with the distribution of each accident severity being fairly close and mostly concentrated around lower frequencies. Cluster 4 is distinguished by the frequency of fatal accidents. These areas have relatively higher frequencies of fatal accidents and relatively low slight accident frequencies. However, the number of areas with the highest fatal accident frequencies in the distribution are fairly low and has a comparatively larger distribution of frequencies. Finally, the worth of each accident severity is relatively equal, and much higher than in the other clusters, for cluster 5. This is likely because areas in cluster 5 appear to have relatively low accident frequencies across all accident severities.

**Table 5:** Segment number with its corresponding percentage and count of locations and the corresponding worth of each accident severity (given to 3 s.f.).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Segment** | **Percentage** | **Count** | **Fatal** | **Serious** | **Slight** |
| 1 | 11.0 | 41 | 0.116 | 0.0560 | 0.0464 |
| 2 | 8.87 | 33 | 0.00874 | 0.101 | 0.153 |
| 3 | 18.0 | 67 | 0.0288 | 0.0956 | 0.114 |
| 4 | 26.6 | 99 | 0.174 | 0.0943 | 0.0722 |
| 5 | 35.5 | 132 | 0.223 | 0.275 | 0.212 |

Timeline, bar chart

Description automatically generatedTimeline

Description automatically generated

**Figure 9:** Histograms showing the relative difference in the distributions for each segment (1-5) and the population for each of the accident severities (fatal, serious, slight).

Looking at the clusters contained in the tables 4 and 5, it would appear that split of the number of locations in each cluster appear similar for the two software, although each cluster was assigned a different number. However, it is possible that the actual locations in each cluster completely differ between the two software. The locations in each cluster were compared to discover any potential overlap (table 6; code can be found in appendix). When merging the clusters produced in R with those produced in SAS EM by the local authority code, the datasets went from 372 rows each to 370 when combined. This is likely due to the removal of slightly different outliers by each model. According to table 6, it appears that R cluster 1 and SAS EM cluster 2 contain the same locations. R cluster 2 mostly overlaps with the locations in SAS EM cluster 5, with some from SAS EM cluster 3. R cluster 3 mostly contains the same locations as SAS EM cluster 3, with some from SAS EM cluster 2. R cluster 5 contains mostly areas from SAS EM cluster 1, with some overlap from SAS EM cluster 2 and 4. Finally, R cluster 4 contains the most overlap of SAS EM clusters; the most overlap occurs with SAS EM cluster 4, with some locations from SAS EM cluster 1, 3 and 5 also present.

**Table 6:** R clusters with the corresponding SAS EM clusters that contain the same local authority districts (the number of these shown in the count column).

|  |  |  |  |
| --- | --- | --- | --- |
| **R cluster** | **Count** | **SAS EM cluster** | **Count** |
| 1 | 22 | 2 | 22 |
| 2 | 141 | 3 | 18 |
| 5 | 123 |
| 3 | 51 | 1 | 2 |
| 2 | 10 |
| 3 | 39 |
| 4 | 112 | 1 | 1 |
| 3 | 10 |
| 4 | 92 |
| 5 | 9 |
| 5 | 44 | 1 | 36 |
| 2 | 1 |
| 4 | 7 |

Overall, the patterns of accident severity appear to be mostly consistent when the clusters are compared between the two software. This is especially true for R cluster 1 which contains the same areas as cluster 2 (with two removed from the original R cluster as outliers) and R cluster 2 which contains mostly cluster 5 with some overlap with cluster 3. Consistency across the models suggests that resulting clusters are fairly accurate and can be used for further investigation. That R cluster 1 and R cluster 2 are the most consistent is particularly useful, as these clusters represent the two extremes of accident frequencies, with R cluster 1 having particularly high relative frequencies of fatal accidents. Further investigations comparing what makes areas in R cluster 1 more prone to accidents and fatal accidents than those in R cluster 2 could reveal important influencing factors. This research could then improve physical safety measures and policies with improved understanding of the most influencing factors on road accidents.

One method of investigation can include plotting the clusters on a map to visually see where different clusters sit in the country and whether this provides any insight. Using R it is possible to do this using the package ‘ggmap’ (Kahle and Wickham, 2013). As the original dataset contains the latitude and longitude for each accident, the original dataset and the cluster dataset can be merged by the local authority district code. The accident clusters can then be plotted on a map using the latitude and longitude [fig.10]. This mapping makes it easier to visually identify where clusters are located in the country and whether these areas are linked in any way. R clusters 4 and 5 appear to have high incidences of fatal accidents and according to the map, these areas can be found across the country in large areas. However, R cluster 3, which appears to have a low incidence of fatal accidents comparatively, is located in very few small areas. That high accident fatalities are predominant across the country is alarming and so emphasises why investigating the reasons for this are important. Despite not attempted in this study, to understand potential differences in the areas categorised by each cluster, other factors could be plotted on the map to see if there are apparent relationships, such as urban verses rural. This can also be performed on a smaller scale, by filtering accident locations within a smaller radius and mapping those. This is something which can be seen in research by Sinclair and Das (2021).

Map, scatter chart

Description automatically generatedGraphical user interface, chart, scatter chart

Description automatically generated

**Figure 10:** Accidents mapped using their corresponding latitude and longitude. The colours represent the R cluster they belong to.

Another way the clusters produced by this analysis could be useful is as a prior step to other data mining steps. Many other studies have combined k-means clustering with other data mining methods in order to improve the findings of other models. (Kumar and Toshniwal, 2016) used k-means clustering to group locations based on accident frequency prior to using association rules mining to identify common characteristics among the accidents within the clusters. Sinclair and Das (2021) used a similar UK dataset to this study and was able to cluster numeric values, such as number of vehicles, number of casualities, vehicle age, vehicle engine capacity and age of driver. They then applied association rule mining and found evidence that suggested young people in urban areas were more likely to crash. Yassin and Pooja (2020) combined k-means and classification technique, random forest, to predict accident severity of road accidents in Ethiopia. They obtained an accuracy of 99.86% using this combined approach. Alkheder et al. (2017) improved their artificial neural network model’s ability to predict the different accident severities, particularly the more rare.

5. Conclusion

Future studies could involve utilising these clusters in association rule mining or classification techniques, such as decision trees, to see if any contributing factors differ between clusters and whether these differences correspond to differences in accident severity prevalence and overall accident frequencies. Additionally, there were many other variables included in the initial dataset used for this study, such as road type, speed limit, weather conditions, etc. It would be interesting to apply k-means clustering to this data and see whether different areas can be segregated according to differences in these factors.

6. References

ALKHEDER, S., TAAMNEH, M. & TAAMNEH, S. 2017. Severity prediction of traffic accident using an artificial neural network. *Journal of Forecasting,* 36**,** 100-108.

ALMJEWAIL, A., ALMJEWAIL, A., ALSENAYDI, S., ALSUDAIRY, H. & AL-TURAIKI, I. Analysis of traffic accident in Riyadh using clustering algorithms. 5th international symposium on data mining applications, 2018. Springer, 12-25.

AMIRI, A. M., SADRI, A., NADIMI, N. & SHAMS, M. 2020. A comparison between artificial neural network and hybrid intelligent genetic algorithm in predicting the severity of fixed object crashes among elderly drivers. *Accident Analysis & Prevention,* 138**,** 105468.

CHANG, L.-Y. & WANG, H.-W. 2006. Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accident Analysis & Prevention,* 38**,** 1019-1027.

DELEN, D., TOMAK, L., TOPUZ, K. & ERYARSOY, E. 2017. Investigating injury severity risk factors in automobile crashes with predictive analytics and sensitivity analysis methods. *Journal of Transport & Health,* 4**,** 118-131.

DEPARTMENT FOR PUBLIC TRANSPORT 2021. Road Safety - Accidents last 5 years.

DEPARTMENT FOR TRANSPORT 2021. Reported road casualties in Great Britain, provisional estimates: year ending June 2021.

FENG, M., ZHENG, J., REN, J. & XI, Y. Association Rule Mining for Road Traffic Accident Analysis: A Case Study from UK. International Conference on Brain Inspired Cognitive Systems, 2019. Springer, 520-529.

IRANITALAB, A. & KHATTAK, A. 2017. Comparison of four statistical and machine learning methods for crash severity prediction. *Accident Analysis & Prevention,* 108**,** 27-36.

KAHLE, D. & WICKHAM, H. 2013. ggmap: Spatial Visualization with ggplot2. *The R Journal,* 5**,** 144-161.

KASSAMBARA, A. & MUNDT, F. 2020. factoextra: Extract and Visualize the Results of Multivariate Data Analyses.

KIM, K. & YAMASHITA, E. Y. 2007. Using ak‐means clustering algorithm to examine patterns of pedestrian involved crashes in Honolulu, Hawaii. *Journal of advanced transportation,* 41**,** 69-89.

KUMAR, S. & TOSHNIWAL, D. 2016. A data mining approach to characterize road accident locations. *Journal of Modern Transportation,* 24**,** 62-72.

LAWSON, R. G. & JURS, P. C. 1990. New index for clustering tendency and its application to chemical problems. *Journal of chemical information and computer sciences,* 30**,** 36-41.

LI, Z., LIU, P., WANG, W. & XU, C. 2012. Using support vector machine models for crash injury severity analysis. *Accident Analysis & Prevention,* 45**,** 478-486.

MONTELLA, A., ARIA, M., D’AMBROSIO, A. & MAURIELLO, F. 2012. Analysis of powered two-wheeler crashes in Italy by classification trees and rules discovery. *Accident Analysis & Prevention,* 49**,** 58-72.

NANDURGE, P. A. & DHARWADKAR, N. V. Analyzing road accident data using machine learning paradigms. 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), 2017. IEEE, 604-610.

PAKGOHAR, A., TABRIZI, R. S., KHALILI, M. & ESMAEILI, A. 2011. The role of human factor in incidence and severity of road crashes based on the CART and LR regression: a data mining approach. *Procedia Computer Science,* 3**,** 764-769.

PUSPITASARI, D., WAHYUDI, M., RIZALDI, M., NURHADI, A. & RAMANDA, K. K-means algorithm for clustering the location of accident-prone on the highway. Journal of Physics: Conference Series, 2020. IOP Publishing, 012086.

R CORE TEAM 2021. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

SAS INSTITUTE INC. 2015. SAS® Enterprise Miner™ 14.1: Administration and

Configuration. Cary, NC: SAS Institute Inc.

SAS INSTITUTE INC. 2017a. *SAS® Enterprise Miner™ 14.3: Reference Help,* Cary, NC, SAS Institute Inc.

SAS INSTITUTE INC. 2017b. SAS® Enterprise Miner™: Tutorials and Examples. Cary, NC: SAS Institute Inc.

SAVOLAINEN, P. T., MANNERING, F. L., LORD, D. & QUDDUS, M. A. 2011. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. *Accident Analysis & Prevention,* 43**,** 1666-1676.

SILVA, P. B., ANDRADE, M. & FERREIRA, S. 2020. Machine learning applied to road safety modeling: A systematic literature review. *Journal of Traffic and Transportation Engineering (English Edition),* 7**,** 775-790.

SINCLAIR, C. & DAS, S. Traffic Accidents Analytics in UK Urban Areas using k-means Clustering for Geospatial Mapping. 2021 International Conference on Sustainable Energy and Future Electric Transportation (SEFET), 2021. IEEE, 1-7.

TAN, P.-N., STEINBACH, M. & KUMAR, V. 2016. *Introduction to data mining*, Pearson Education India.

TAVAKOLI KASHANI, A., SHARIAT-MOHAYMANY, A. & RANJBARI, A. 2011. A data mining approach to identify key factors of traffic injury severity. *PROMET-Traffic&Transportation,* 23**,** 11-17.

WEN, X., XIE, Y., JIANG, L., PU, Z. & GE, T. 2021. Applications of machine learning methods in traffic crash severity modelling: current status and future directions. *Transport reviews,* 41**,** 855-879.

WICKHAM, H., AVERICK, M., BRYAN, J., CHANG, W., D'AGOSTINO MCGOWAN , L., FRANÇOIS, R., GROLEMUND, G., HAYES, A., HENRY, L., HESTER, J., KUHN, M., LIN PEDERSEN, T., MILLER, E., MILTON BACHE, S., MÜLLER, K., OOMS, J., DAVID, R., SEIDEL, D. P., SPINU, V., TAKAHASHI, K., VAUGHAN, D., WILKE, C., WOO, K. & YUTANI, H. 2019. Welcome to the {tidyverse}. *Journal of Open Source Software,* 4**,** 1686.

WORLD HEALTH ORGANIZATION 2018. Global status report on road safety 2018. .

XU, C., BAO, J., WANG, C. & LIU, P. 2018. Association rule analysis of factors contributing to extraordinarily severe traffic crashes in China. *Journal of safety research,* 67**,** 65-75.

YASSIN, S. S. & POOJA 2020. Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach. *SN Applied Sciences,* 2.

YU, S., JIA, Y. & SUN, D. 2019. Identifying Factors that Influence the Patterns of Road Crashes Using Association Rules: A case Study from Wisconsin, United States. *Sustainability,* 11**,** 1925.

ZENG, Q. & HUANG, H. 2014. A stable and optimized neural network model for crash injury severity prediction. *Accident Analysis & Prevention,* 73**,** 351-358.

7. Appendix

**Appendix A**

R code used to prepare dataset.

##Data Preparation

library(tidyverse)

rm(list=ls())

#set working directory  
setwd("C:\\Users\\smann\\OneDrive\\University\\Data Science\\ASDM\\Project\\Road safety")

#import accident dataset

df<- read.csv("dft-road-casualty-statistics-accident-last-5-years.csv")

names(df)  
head(df)  
summary(df)  
str(df)  
  
#convert any missing data to NA

df[df==-1]<- NA

df[is.null(df)]<- NA

#count the number of distinct locations

n\_distinct(df$local\_authority\_ons\_district)

#create dataframe with location and accident severity frequencies

location\_df <- df %>%

select(accident\_severity, local\_authority\_ons\_district) %>%

group\_by(local\_authority\_ons\_district) %>% count(accident\_severity) %>%

pivot\_wider(names\_from=accident\_severity, values\_from=n) %>% rename(fatal = "1") %>%

rename(serious = "2") %>% rename(slight = "3") %>%

mutate(fatal = ifelse(is.na(fatal), 0, fatal),

serious = ifelse(is.na(serious), 0, serious),

slight = ifelse(is.na(slight), 0, slight)) %>%

as.data.frame()

#save prepared dataset

write.csv(location\_df, "location\_df.csv", row.names = FALSE)

**Appendix B**

R code for K-means clustering.

##K-means clustering

#data exploration

str(location\_df)

names(location\_df)

summary(location\_df)

#check for missing data  
colSums(is.na(location\_df))

head(location\_df)

length(location\_df[,1])

#make rownames the local authority district code  
rownames(location\_df) <- location\_df[,1]

location\_df[,1] <- NULL

#normalise function  
normalise <- function(df)  
{  
return(((df- min(df)) /(max(df)-min(df))\*(1-0))+0)  
}

#apply normalise function to dataset

lsoa<-rownames(location\_df)

norm<-as.data.frame(lapply(location\_df,normalise))

rownames(norm)<-lsoa

#get Hopkins stat  
tendency <- get\_clust\_tendency(norm, n = nrow(norm)-1, graph = FALSE)

tendency$hopkins\_stat

#elbow plot  
fviz\_nbclust(norm, kmeans, method = "wss")

#k-means clustering

library(factoextra)

set.seed(123)

km.fit <- kmeans(norm, 4)

km.fit$cluster

km.fit$size

#visualise clusters

fviz\_cluster(km.fit,norm)

#get means for clusters  
aggregate(norm, by=list(cluster=km.fit$cluster), mean)

#merge clusters numbers with normalised dataset

clusters <- as.data.frame(km.fit$cluster)

cl1 <- merge(norm, clusters, by=0 )

rownames(cl1) <- cl1[,1]

cl1[,1] <- NULL

#remove outliers (cluster 4)

cl1 <- cl1 %>% filter(`km.fit$cluster`!=4) %>%

select(-`km.fit$cluster`)

summary(cl1)

count(cl1$fatal)

#elbow plot  
fviz\_nbclust(cl1, kmeans, method = "wss")

#k-means clustering

set.seed(123)

km.fit <- kmeans(cl1, 5)

km.fit$cluster

km.fit$size

#visualise clusters

fviz\_cluster(km.fit,cl1)

#view cluster means

aggregate(cl1, by=list(cluster=km.fit$cluster), mean)

**Appendix C**

R code for producing map with clusters plotted.

#merge clusters with original dataset

clusters <- as.data.frame(km.fit$cluster)

clusters$local\_authority\_ons\_district <- row.names(clusters)

full <- merge(df, clusters, by="local\_authority\_ons\_district")

names(full)[ncol(full)] <- "cluster"

colSums(is.na(df))

str(full)

full$latitude <- as.numeric(full$latitude)

full$longitude <- as.numeric(full$longitude)

full$cluster <- as.factor(full$cluster)

library(ggmap)

# compute the bounding box using latitude and longitude

bc\_bbox <- make\_bbox(lat = latitude, lon = longitude, data = full)

bc\_bbox

# get maps from google

bc\_big <- get\_map(location = bc\_bbox, source = "google", maptype = "terrain")

# plot the points and color them by sector

ggmap(bc\_big) +

geom\_point(data = full, mapping = aes(x = longitude, y = latitude, color = cluster),

shape=".")

**Appendix D**

R code for comparing the locations in common between the clusters created by SAS and by R.

###compare SAS and R results

#import and isolate SAS cluster data

SAS\_df <- read.csv("sas-clusters\_TRAIN.csv")

SAS\_df <- SAS\_df %>% select(local\_authority\_ons\_district, X\_SEGMENT1\_)

#combine the R and SAS clusters into one dataset

compare <- merge(SAS\_df, clusters, by='local\_authority\_ons\_district')

names(compare)[ncol(compare)] <- "cluster"

#find locations in common

compare %>% filter(cluster==1) %>% count(X\_SEGMENT1\_)

compare %>% filter(cluster==2) %>% count(X\_SEGMENT1\_)

compare %>% filter(cluster==3) %>% count(X\_SEGMENT1\_)

compare %>% filter(cluster==4) %>% count(X\_SEGMENT1\_)

compare %>% filter(cluster==5) %>% count(X\_SEGMENT1\_)