

Pattern Recognition Approach for Road Collision Hotspots Analysis: Case Study of Northern Ireland

Bronagh Coll
Queen's University, Belfast

Salissou Moutari
Queen's University, Belfast

Adele Marshall
Queen's University, Belfast

Abstract

In order to address road safety effectively, it is essential to understand all the factors, which attribute to the occurrence of a road collision. This is achieved through road safety assessment measures, which are primarily based on historical crash data. Recent advances in uncertain reasoning technology have led to the development of robust machine learning techniques, which are suitable for investigating road traffic collision data. These techniques include supervised learning (e.g. SVM) and unsupervised learning (e.g. Cluster Analysis). This study extends upon previous research work, carried out in Coll *et al.* [3], which proposed a non-linear aggregation framework for identifying temporal and spatial hotspots. The results from Coll *et al.* [3] identified Lisburn area as the hotspot, in terms of road safety, in Northern Ireland. This study aims to use Cluster Analysis, to investigate and highlight any hidden patterns associated with collisions that occurred in Lisburn area, which in turn, will provide more clarity in the causation factors so that appropriate countermeasures can be put in place.

Key Words: cluster analysis, hotspot, pattern, road safety.

1. Introduction

In recent years, there has been an increased focus on combating the rising problems in road safety at both a national and international level. In order to address road safety issues effectively, it is crucial to understand all the factors, which attribute to the occurrence of a road collision. This is achieved through road safety assessment measures, which are primarily based on historical crash data.

Many research efforts have been dedicated in applying statistical models to assess and identify any trends in the road collision data so that future predictions can be made. The most common models include generalised linear models, regression models, nonparametric statistical methods, etc. Most of these statistical models are essentially based on some assumptions, such as stationarity of the time series, known statistical distributions of road safety indicators, etc. One of the major limitation of statistical models is their tendency to concentrate on the means and miss the extremes; whereas, in reality road collision data may exhibit a highly fluctuated behaviour with extreme peaks. Recent advances in uncertain reasoning technology have led to the development of robust machine learning techniques, which do not require stationarity of the time series and therefore more appropriate for investigating road traffic collision data. These techniques include supervised learning (e.g. Support Vector Machines, Random Forest) and unsupervised learning (e.g. Cluster Analysis).

Given the various contributing factors in the occurrence of a road traffic accident, the assessment of indicators on an individual basis is both difficult and shortsighted. As such, for decision-making purposes, it seems logical that indicators should be aggregated into a single composite index, referred to as the road traffic Composite Safety Performance Index (CSPI).

CSPIs are most commonly employed in the analysis of the safety conditions of road traffic systems and continuous assessment of their performance, facilitating subsequent benchmarking of countries, regions (i.e. hotspot identification) etc. Coll *et al.* [3] developed an alternative non-linear aggregation approach for the estimation of a composite road safety index. The results from this study identified Lisburn as the hotspot area, in terms of road safety, in Northern Ireland in 2010. A snapshot of the results, shown in Figure 1, results indicate that out of the 29 policing areas in Northern Ireland, Lisburn was by far the most underperforming area.

The purpose of this study is to use an unsupervised learning technique, namely Cluster Analysis, to further investigate the pattern of collisions that occurred in this hotspot area, which in turn, will provide more clarity in the causation factors so that the appropriate countermeasures can be implemented.

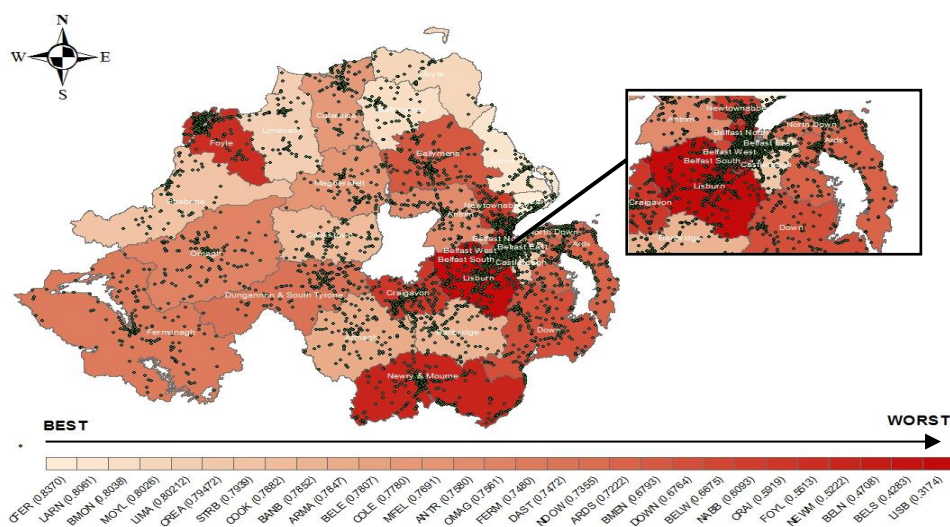


Figure 1 – Spatial representation for Northern Ireland: The darker the red the worse the overall road safety performance for the CSPI. The number of collisions (represented by the dots) is also plotted.

2. Data

The largest road collision database in the UK is known STATs19, which is collected and maintained by the police. The importance of the data collection and maintenance phases cannot be underestimated; it provides the basis upon which meaningful interpretations can be made. The quantity and quality of the information analysed is absolutely crucial in identifying key relationships in the road safety problem. Crash data contain road safety indicators, which are defined as a quantitative or qualitative measure derived by a series of observed facts relative to a particular collision [11]. This study is based on Northern Ireland road collision data stored in STATs19. The focus will be upon one particular policing area in Northern Ireland; Lisburn, which is located to the southwest of Belfast. This policing area was deemed to be the hotspot area, in terms of road safety, in Northern Ireland in 2010 according to results obtained in Coll *et al.* [3]. Thus, a more in-depth analysis of the causation factors in this troublesome area is required.

The STATs19 data consists of 82 variables, which are split into three main categories: collision, casualty and vehicle. Examples of some of the explanatory variables contained within each category include:

- Collision: the time of the collision, the location, the type of collision, the number of vehicles involved etc.

-
- Casualty: types of casualty, number of each type of casualty, age and sex of each casualty etc.
 - Vehicle: type of vehicle, impact on the vehicle, etc.

The STATs19 system uses three-category classification for the type of casualty: fatality seriously injured and slightly injured [5]. Damage only collisions, collisions reported after a 30-day period from the date of collision and collisions that have occurred in car parks are excluded from the data [5]. In this research the data for Lisburn area, in Northern Ireland, is extracted from the STATs19 database for the period of 2004 through to 2012. The study focuses on explanatory variables contained only within the collision category. Further analysis will explore attributes within the two remaining categories.

From 2004 to 2009, with the slight exception of 2007, the data exhibits completeness in the recorded collision variables. However, in recent dataset i.e. post 2009 there are several variables missing for the data. Cluster analysis is thus performed on the available, complete data for each year. From 2004 to 2009 variables that are analysed include carriageway type, day of the month, month of the year, day of week, number of vehicles involved, severity of the collision, speed limit, number of casualties, weather conditions, road surface conditions, junction control, junction detail. From 2010 onwards the latter four variables are excluded from the analysis, as they are not available in the datasets.

3. Methodology

The main objective of this study is to use an unsupervised learning technique, namely Cluster Analysis, to highlight any hidden patterns associated to the collisions data set for the Lisburn area in Northern Ireland. Cluster analysis is a prevalent unsupervised classification procedure, which involves grouping similar observations (i.e. data variables or feature vectors) into clusters, such that the observations in the same cluster render similar qualities and those in different clusters are classified as dissimilar to other clusters [6]. Unlike supervised classification where there exists a collection of labelled or pre-classified patterns; in Cluster Analysis the number of clusters and the data clusters are not known in advance. Therefore the category labels are data driven. In Figure 2, the data is grouped into 3 distinct clusters; where the similarity criterion is the distance between the points: two or more objects belong to the same cluster if they are “close,” according to a given distance measure and is commonly referred to as distance-based clustering.

Intra-cluster distance and Inter-cluster distance are the two aspects, which drive Cluster Analysis procedures. Intra-cluster distance determines how “close,” two data points are to each other [7]. It is commonly known as distance or similarity or measure or generally speaking proximity measures. Some of the most common distance measures include Euclidean distance, Minkowski distance and Manhattan distance, to name but a few. The choice of the distance measure depends upon the nature of data, as well as the objectives of the study. If the data is predominantly quantitative, it's most appropriate to use a distance measure such as the Euclidean distance and if the data is mainly qualitative, an index of dissimilarity is more suited [6]. Inter-cluster distance examines how “close” two clusters are to each other. This aspect is included in Cluster Analysis through linkage function or linkage criterion. Some of the most prevalent linkage criteria are as follows: single linkage, complete linkage, average linkage and Wards criterion. As each of the aforementioned linkage criteria has its own specific properties, they can yield different results when applied to the same dataset. Therefore, careful consideration must be afforded in order to select the most suitable linkage criteria for a dataset. It may be useful to apply several linkage methods and select the most desirable method.

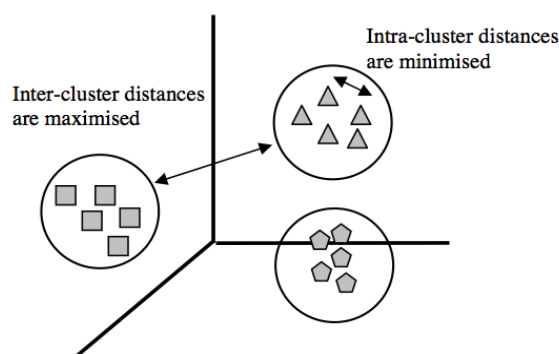


Figure 2: Two main aspects in Cluster Analysis the intra-cluster distances and inter-cluster distances.

The most commonly used procedures in Cluster Analysis are: Hierarchical Clustering and Partitioning Clustering. The basic idea behind hierarchical clustering is to form a binary tree of the data that successively merges similar collection of data points. A hierarchical method can be classified as either agglomerative or divisive. In the agglomerative algorithm or bottom up algorithm, an observation or a cluster of observations is merged into another cluster until all the clusters are merged into one single cluster. The divisive algorithm, also referred to as top-down approach begins with all the objects in one cluster. In each successive iteration; a cluster is divided into smaller clusters, until eventually each object is in its own distinct cluster. The results of the hierarchical approach can be visualised in a tree like diagram, called dendrogram, representing the arrangement of clusters. Partitioning clustering divides the observations into n clusters. This is achieved by starting with an initial partition or with cluster centres and then reallocating the observations according to some optimality criterion, thus allowing for poor initial partitions to be corrected at a later point in the process [6]. The partitioning algorithm is often run multiple times with different initial states and the best configuration obtained for all of the iterations is used as the output cluster [2]. Two of most prominent partitioning methods are the k -means algorithm and the k -medoids algorithm. In k -means clustering, each cluster is represented by its centre whereas in the k -medoids procedure each cluster is denoted by one of its components.

Cluster Analysis is pertinent in any domain, which aims to assess and identify trends in the data so that future predictions can be made, as well as data reduction and outlier detection. Its far reaching applicability is not only evident in areas such as pattern recognition, image processing and information retrieval, but also has a rich history in many other disciplines including geography, medicine, sociology, economics, archaeology, psychiatry, geology, and marketing, and road safety to name but a few. Research on clustering as a grouping technique is developing in the area of road safety and will be progressed further in this study. Studies in which Cluster Analysis has been used for road safety assessment include Anderson [1], Kanungo et al [4] Sohn *et al.* [5], Wagstaff *et al.* [9] etc.

In the next section, we will use Hierarchical clustering to investigate patterns within collisions in the dataset for Lisburn area. Partitioning clustering is beyond the scope of this study and will be considered in further work.

4. Discussion of Results

In this study, hierarchical clustering is applied to the 13 indicators, identified in Section 2 and further displayed in Figure 3 for Lisburn, from 2004 to 2009. In recent data (i.e. 2010 to 2012) several indicators namely weather conditions, road surface conditions, junction detail and junction control, were incomplete or completely missing from the data set and thus excluded in the analysis, during that period.

Before applying the hierarchical clustering algorithm the variables indicated in section 2.1 are normalised using function (1) below:

$$\frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

where x denotes each data point.

Figure 3 illustrates colour maps for the 13 indicators under investigation; these are fundamental in the interpretation of the clustering results. Each indicator is colour mapped according to a specific scale. For instance, the scale for days of the week goes from 1 to 7 (light to dark), with Monday labelled 1, Tuesday is labelled 2, and so on. Therefore, dark red would indicate a collision occurring at the weekend. The scale of the colour maps, with exception of severity, number of vehicles and casualties, are not ranked best to worst, but rather correspond to an attribute within that variable. Therefore, in these cases a darker colour of red does not symbolise a decrease in that road safety measure.

Results obtained for 2004, 2008 and 2012 are selected for discussion in this study. Figure 4 presents the results obtained for 2004, upon which several conclusions are drawn. Firstly, it is observed that the majority of collisions that occurred in Lisburn during 2004 were slightly injury collisions.

Furthermore, according to junction detail and junction control, over half of the recorded collisions happen at either give way or stop signs and at a staggered T junction, other junction, respectively. The number of vehicles involved in collisions is repeatedly between 1 and 3 vehicles with very few beyond 4. According to the carriageway type, very few collisions occurred on roundabouts, with the bulk of collisions taken place on dual carriageways and motorways.

Roughly speaking, there is a relativity even split in the speed limit of collisions. Day of week, day of month and month of the year demonstrate a spectra of red colours. In lay terms, this demonstrates that collisions are not specific to a certain day or week, or month. In relation to weather conditions, collisions mainly occurred in fine or rainy day, which is supported with the high volume of road surface conditions recorded as either dry or wet. It can be seen in Figure 4 that there are two distinct darker blocks. This indicates that the weather conditions are unknown which maybe a result of failings in recording by the police. For a closer examination of the results, a cluster is extracted from Figure 4 and shown in Figure 5. The dendrogram in Figure 5 shows 10 fatal and serious injury road collisions that happened during 2004. As indicated by junction control and junction detail, the selected collisions were not located at a junction and they all occurred in fine weather conditions. On the other hand, it can be observed that the fatal collisions occurred between 21h -00h in 50mph speed limit roads. Carriageway type indicates that the fatal collisions occurred on single carriageways. Further to this, the fatal collisions involved a single vehicle and had only one casualty. Regarding serious injury collisions, carriageway type indicates that they all occurred on either motorways or dual carriageways. The majority of the serious injury collisions involved a single vehicle.

Figure 6 illustrates the clustering results obtained for 2008. According to junction detail and junction control over a half of collisions during 2008 occurred at give way and stop signs, and at staggered T junctions or at another type of junction. In comparison with 2004, it is noted that for the carriageway type for 2008, the clusters are much darker than that for 2004. This indicates a shift in where collisions are occurring from 2004. The darker clusters on the carriageway type presented in 2008 indicate that the vast majority of collisions happened on slip roads. This information may prompt decisions makers to address the shift in collisions predominantly occurring on slip roads in Lisburn.

Finally, clustering results obtained for 2012 for Lisburn are shown in Figure 7. As in agreement with the other two years under study, the majority of collisions that occurred in Lisburn were slight injury collisions. Similar to 2008, the carriageway type indicates that the majority of collisions also occurred on slip roads. There is approximately an even split on collisions occurring on lower and higher speed limit roads. It can be observed from the time of day indicator that very few collisions occurred between 00h -6h.

The insights drawn from the results obtained in this study can aid decision makers in the identification and implementation of appropriate road safety interventions.

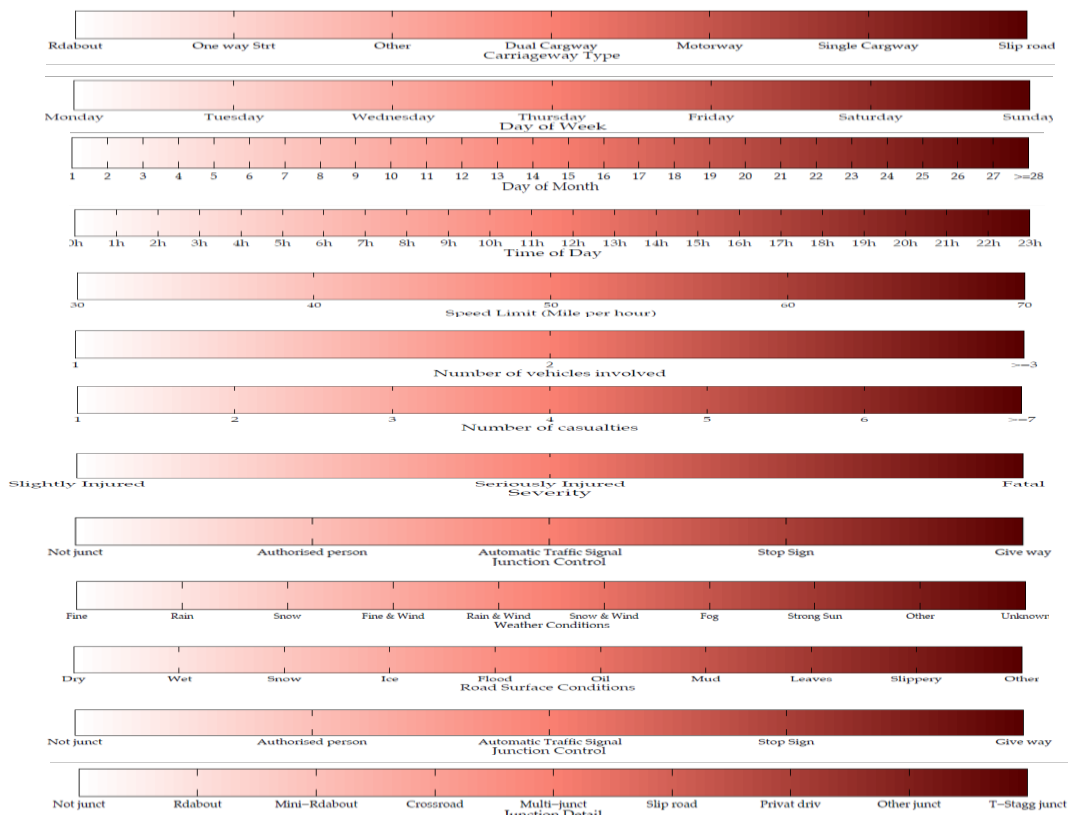


Figure 3 – Colour ramps for each indicator used in the cluster analysis

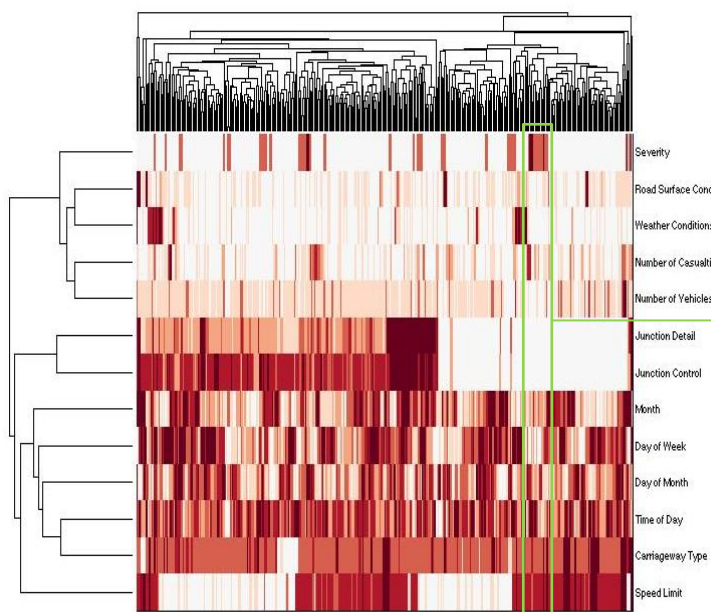


Figure 4 – Dendrogram for 2004

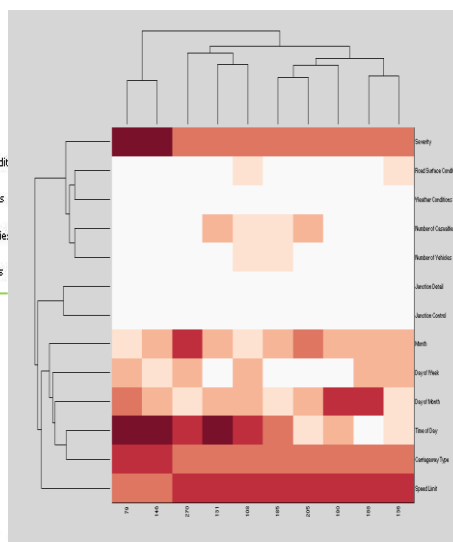


Figure 5 – Snapshot of clusters from the Dendrogram for 2004

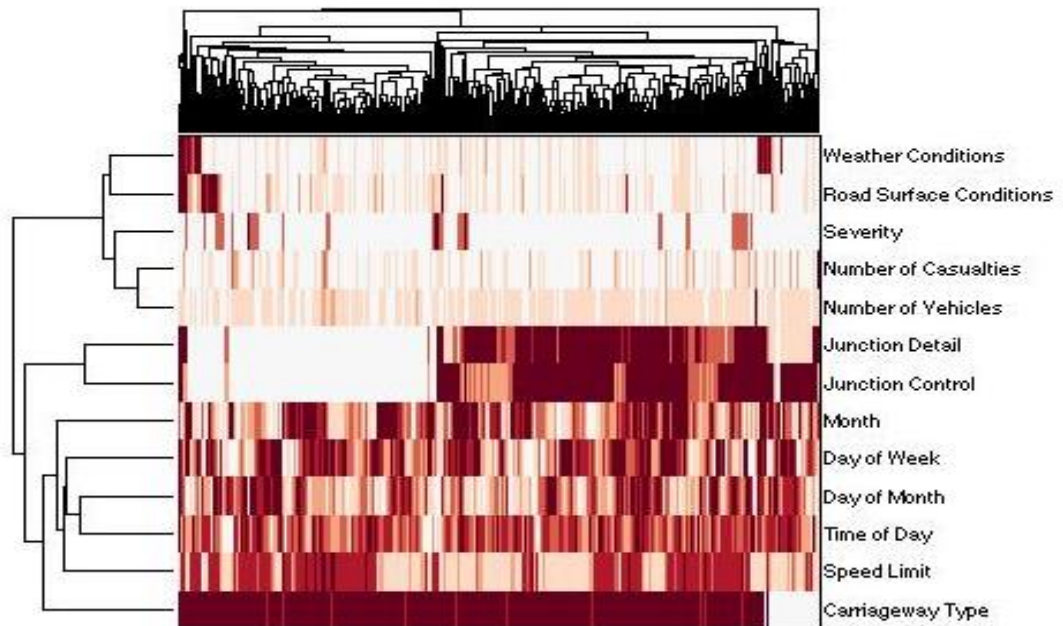


Figure 6 – Dendrogram for 2008

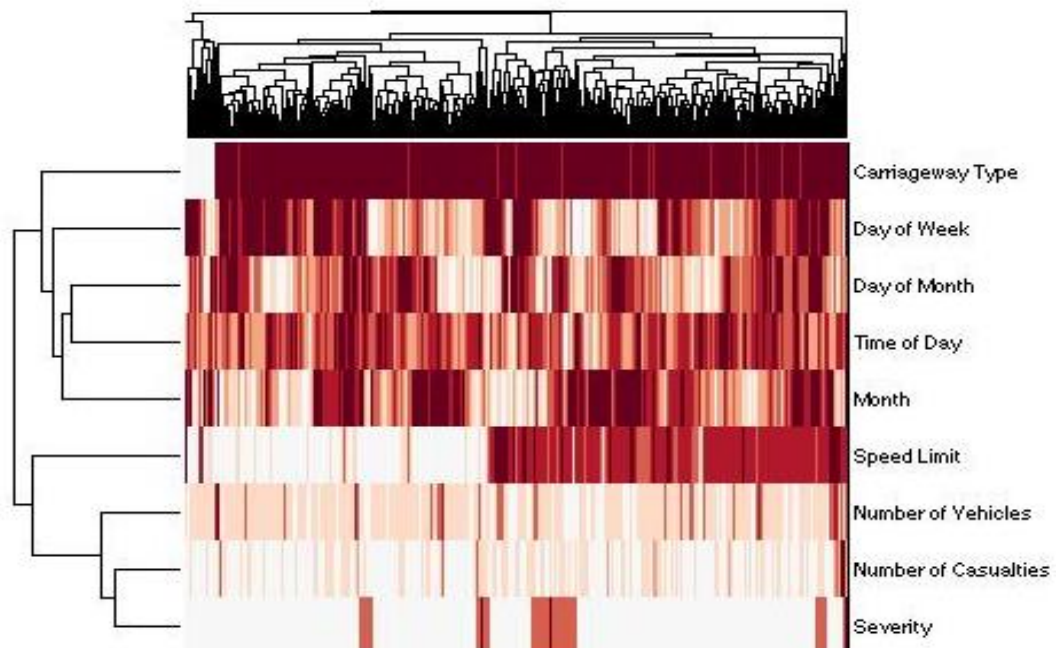


Figure 7 – Dendrogram for 2012

5. Conclusions and Further Work

This study uses hierarchical clustering for in-depth assessment of road collisions, which occurred in Lisburn area, the hotspot, in terms of road safety in Northern Ireland. The exploratory results obtained provide various insights, which could aid decision makers in the identification and implementation of appropriate road safety interventions to improve the road safety record of the area. Moreover, the scope of this research could be extended to investigate hazardous hotspots in Great Britain namely the Metropolitan Police district. Although the results attained in this study are fundamental tool in road safety exploratory

analysis providing meaningful taxonomies, it is not without limitations. The hierarchical algorithm is sequential and does not enable work to be undone. Furthermore, interpretation of the hierarchy can be complex or even confusing and is sensitive to noise and outliers. As such, comparative studies with other data mining methods namely principal component analysis could provide additional insights. Indeed, Cluster Analysis can be viewed as a fundamental tool for exploratory analysis, which often prompts further analysis and investigations.

Acknowledgements

The first author was funded by the Department of Employment and Learning DEL Northern Ireland. STAs19 data was supplied by the ESRC UK data archives at the University of Essex and also the Police Service of Northern Ireland (PSNI). Data was also provided by the Northern Ireland Statistics and Research Agency (NIRSA).

References

- [1] Anderson, T., (2008) Kernel density estimation and K-means clustering to profile road accident hotspots. *Accident Analysis and Prevention*, Vol. 14, No. 3, pp. 359 – 364.
- [2] Ayramo S., Karkkainen T., Introduction to partitioning based clustering methods with a robust example. Reports of the Department of Mathematical Information Technology Series C. Software and Computational Engineering. ISBN 951-39-2467.
- [3] Coll, B., Moutari, S., Marshall, A., (2013) Hotspot identification and ranking for road safety improvement: An alternative approach. *Accident Analysis and Prevention*, 2013, Vol. 59, pp. 604-617.
- [4] Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y. (2002) An efficient K-means clustering algorithm: analysis and implementation, *IEEE Transactions on pattern analysis and machine intelligence*, Vol 24, pp. 881- 892.
- [5] Lyons, R., Ward, H., Brunt, H., Macey, S., Thoreau, R., Bodger, O.G., Woodford, M. Using multiple datasets to understand trends in serious road traffic casualties, *Accident Analysis and Prevention*, Vol. 40, 2008, pp. 1406-1410.
- [6] Jain A.K., Murty, M.N., Flynn, P.J., (1999) Data Clustering: A review, *ACM Computing Surveys*, Vol. 3, No. 3, pp. 264-323.
- [7] Rencher, A.C, (2002) *Methods on Multivariate Analysis*, Second Edition, John Wiley and Sons Inc. Publication.
- [8] Sohn, S., Lee, S., (2002). Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea. ." *Safety Science*, Vol. 41, No. 1, pp. 1-14.
- [9] Wagstaff, K., Cardie, C., Rogers, S., Schroedl, S., (2001) Constrained K-means clustering with background knowledge, *Proceedings of the 18th International Conference on Machine Learning*, pp. 577-584.
- [10] Xie, Z. Yan, J. (2008) Kernel Density Estimation on traffic accidents in a network space, *Computers, Environment and Urban Systems*, Vol. 32, pp. 396 -406.