RAC Foundation

Mobility • Safety • Economy • Environment

RCIP Police Area Collision Profiles

Methodology

Dr. Craig Smith and Bruce Walton
Agilysis

March 2021

agilysis

The Royal Automobile Club Foundation for Motoring Ltd is a transport policy and research organisation which explores the economic, mobility, safety and environmental issues relating to roads and their users. The Foundation publishes independent and authoritative research with which it promotes informed debate and advocates policy in the interest of the responsible motorist.

RAC Foundation

89–91 Pall Mall

London

SW1Y 5HS

Tel no: 020 7747 3445

www.racfoundation.org

Registered Charity No. 1002705

RCIP Police Area Collision Profiles

Methodology

Dr. Craig Smith and Bruce Walton

Agilysis

March 2021

About the Authors

Dr. Craig Smith is a mathematician, with an established background in academic research. He studied for both his undergraduate master's degree and doctorate in mathematics at the University of Oxford. Since then, he has endeavoured to use his expertise to carry out robust and innovative analysis and research. As Agilysis' Data Scientist, Craig has extensive experience in handling a wide variety of data, and uses his background as a mathematician to explore the use of machine learning and artificial intelligence in advancing road safety research and unlocking the full potential of data.

Bruce Walton has been working with road safety data since 2002, coming from a background in analytical modelling, database design and IT training across several sectors. Since his appointment as project manager for the multi award-winning MAST Online project, Bruce has become recognised as expert in road casualty data, contributory factor analysis, resident risk, analytical architecture and enforcement data management. Bruce works with many road safety stakeholder organisations in the UK, and provides consultancy and training to international projects on road safety data architecture and reporting, such as the International Road Federation's 'World Road Statistics' programme. Bruce is also a member of the government's Standing Committee for Road Accident Statistics (SCRAS) in the UK.

Disclaimer

This report has been prepared for the RAC Foundation by Dr. Craig Smith and Bruce Walton of Agilysis Ltd. Any errors or omissions are the author's sole responsibility. The report content reflects the views of the authors and not necessarily those of the RAC Foundation.

Contents

List of Figures

List of Tables

List of Abbreviations

ATS	automatic traffic signal
CF	contributory factor
DfT	Department for Transport
IMD	Index of Multiple Deprivation
LSOA	Lower Layer Super Output Area
ONS	Office for National Statistics
RCIP	Road Collision Investigation Project

1 Introduction

This research paper forms part of the Road Collision Investigation Project (RCIP). The purpose of RCIP is to establish whether there is a business case for putting more resource into the investigation of road crashes – and, if there is, to establish how best to take this forward. The project, implemented by the RAC Foundation with government funding, began in the summer of 2018.[1]

1.1 Overview of RCIP

RCIP sets out to develop new approaches to harvesting and analysing data about the causes of road crashes from different sources, including information from police investigations beyond that captured in STATS19[2] returns. RCIP looks for patterns emerging from similar incidents in different places.

Trials have now been established in several police force areas, exploring a new and different approach which aims to identify and understand common themes and patterns that result in death and injury on the public highway. This insight is expected to shape future policymaking. The RCIP areas are as follows:

- Dorset, Devon and Cornwall (combining two police force areas);
- Humberside; and
- West Midlands.

1.2 Collision profiling research as part of RCIP

RCIP's aims include the following:

- the development of an appropriate analytical framework, grounded in systems thinking, for effective learning from road collision investigation;
- the identification and review of existing data from road collisions, the identification of additional sources of data, and testing of the extent to which fresh lessons can thus be learnt; the limitations of current data capture and analysis will be identified, as well as potential options for improvement; and
- the development and application of new analytical protocols for testing in a real-world setting involving two or more police constabularies, in partnership with Highways England and local highway authorities.

To address these aims, the RAC Foundation produced a research brief in February 2020 with the following objectives:

- to identify trends and/or patterns of collision frequency, severity and type for each police area over a selected time span;
- to provide as much relevant analytical input as is possible by area (e.g. demographics, road conditions, journey purpose, average speeds, traffic levels etc.);
- to illustrate patterns and trends using a visual presentation, where possible;
- to analyse how RCIP area performance (and components of that performance) compare to selected comparator police force areas as well as other relevant comparators (rural/urban, Index of Multiple Deprivation (IMD) areas, district / Lower Layer Super Output Area (LSOA));
- to point out, within each police force area, any geographical locations or area types that warrant further in-depth investigation by local RCIP investigators;
- to compose a summary report, to be used as a reference document by RCIP local Investigators; and

[1] www.racfoundation.org/collaborations/road-collision-investigation-project

[2] STATS19 is the national database consisting of data collected by a police officer when an injury road accident is reported to them.

- to propose the other types of road safety data analysis which might benefit (a) RCIP and/or (b) any future road collision investigation branch.

Agilysis successfully bid to undertake this research for three RCIP areas, with work commencing in April 2020. In May 2020, this was extended to cover the Metropolitan Police force area, which maintains an interest in the project, and has aspirations to provide further analytical support to the work.

As well as delivering a synthesis of the most significant findings of analysis carried out on data relating to each of the participating force areas individually, this analysis also considers lessons learnt from the innovative analytical techniques applied. The project has sought to apply deep learning models to road safety data to identify collision trends and types in a way which will provide value to the RCIP project, and this process has raised questions as well as provided answers. It is important that these questions are considered in order that improvements can be applied to analytical techniques in future.

1.3 Delivery

The research addresses these objectives by delivering four key outputs:

- comparator identification;
- trend analysis;
- collision type analysis; and
- synthesis.

This paper sets out to explain the methodology used to deliver each of these components. Details of output for each participating area, along with a synthesis of key findings, is set out in separate reports.

The comparator identification process, and the process used to arrive at it, is described in this paper. The outcome was as follows:

- Devon and Cornwall with Dorset was found to be most comparable to **Avon and Somerset**;
- Humberside was found to be most comparable to **West Mercia**; and
- West Midlands was found to be most comparable to **Greater Manchester**.

The trend analysis has been supplied to RCIP investigators primarily by means of online dashboards. The output from this analysis is extensive; individual reports containing synthesised key findings for each of the three areas have been published.

The collision type analysis used a novel machine-learning process to identify clusters of collisions with similar properties to each other. The methodology used is described in detail in this paper, with outputs for each area published separately. These reports also include collision type analysis output summarised in infographics; a sample infographic with an explanation of how to interpret it is provided in this paper.

To make the methodology as accessible as possible, the main body of this paper presents a non-technical summary of the approach used. Supporting information which is more technical and detailed in nature has been provided in appendices.

2 Overview of Methodology

The fundamental data used by this research is the body of STATS19 police reported injury collisions from 2009 to 2018. This dataset has been augmented with network data such as Department for Transport (DfT) traffic counts, and demographic data such as the Office for National Statistics (ONS) rurality and deprivation classifications.

In addition to reports such as this one, research output falls into three broad categories: identifying relevant comparator areas to the RCIP trial areas; providing comprehensive trend data on those areas and their comparators for the use of RCIP collision investigators; and producing an innovative collision type typology based on new analytical protocols to describe how collisions within and between areas compare with each other.

2.1 Comparator areas

At the outset of the project, suitable comparator areas were identified for each of the three areas. These comparators were selected because they exhibit meaningful resemblance to the RCIP areas, reflecting their distinct characters and providing useful context for understanding their collision characteristics.

The comparator identification process began from an objective base, with a neural network using unsupervised learning to cluster all police force areas of Britain into groups on the basis of their level of similarity. Details of the methodology used are provided in Appendix B.

The Metropolitan Police force area, which experiences a level of annual collisions over four times greater than the next largest area and is not readily comparable to any other force, was not considered suitable for inclusion in this methodology.

Scotland, with a single national force area covering a diverse nation, also challenged this methodology. It was decided for comparator purposes to subdivide Police Scotland along the boundaries of the former regional Scottish forces, grouped as follows:

- Police Scotland Strathclyde by itself (the most populous of the former force areas)
- Police Scotland North, comprising Grampian, Tayside and Northern
- Police Scotland South, comprising Lothian and Borders, Central, Fife, and Dumfries and Galloway

Input data for this comparative cluster analysis process included the following factors:

- network length by road class (classified into motorways, A roads and other roads);
- network length by rurality (urban, town and rural);
- traffic levels by road class;
- network density;
- population density; and
- deprivation by IMD (with separate variables weighting the highest and lowest quartiles).

To provide potential comparator areas which were similar in size to the RCIP areas, smaller police forces were then paired with their most similar larger areas. This provided areas as closely analogous as possible to the large areas formed by the RCIP groupings: West Midlands Police is the largest single force outside London, with 5,760 reported injury collisions annually on average, while Dorset Police and Devon and Cornwall Police combined have 5,707. Humberside is considerably smaller, with 2,498 collisions a year on average. A full list of the potential comparator area groupings identified, along with a description of how the pairing was achieved, is shown in Appendix B.

The cluster analysis process was then repeated using these 28 force area groupings to inform the final selection of comparators. The outcome of the second cluster analysis resulted in selection of the following comparator areas:

- Devon and Cornwall with Dorset was found to be most comparable to **Avon and Somerset**;
- Humberside was found to be most comparable to **West Mercia**; and

- West Midlands was found to be most comparable to **Greater Manchester**.

2.2 Input data

The input variables for the trend analysis and for identifying the collision types were designed with the intention of capturing the aspects of a collision most likely to be of interest to a collision investigator. This initial choice of variables, which was necessarily subjective, is likely to influence the output to a considerable degree.

These input variables are listed in detail in Appendix C, and are broadly classified into four types:

(1) collision variables, describing the overall nature of the event and its circumstances;
(2) vehicle variables, describing important features and behaviours of the involved vehicles and their drivers or riders;
(3) casualty variables, describing the characteristics and actions of persons injured; and
(4) contributory factor characteristics, identifying the presence of certain factors which were, in the contemporary opinion of an attending police officer, germane to the collision occurring.

The same variables that were used as input into the collision types analysis were also used for producing output for the comprehensive trend analysis. To illustrate this, Appendix C not only lists the variables but shows examples of how they were applied in the dashboards.

2.3 Collision type analysis

In accordance with RCIP's aim to develop new analytical protocols, an innovative approach was taken to identifying similar types of collision, by using a neural network then creating a novel approach to categorising and presenting the results. A deep autoencoder was trained to self-predict data on collisions from the past ten years for the whole of Britain, and separately for each of the RCIP areas and also for each comparator area. This model was then used as the basis for an intelligent clustering of collisions. Technical details of this process are included in Appendix A.

2.3.1 Analytical outcomes

The process of converting raw outputs from the neural network into meaningful analytical categories is an involved process. It is explained in detail here and will become clearer when actual sample outputs provided in this document, and complete outputs in each of the three profiles covering specific RCIP areas, are examined. The raw outputs of the autoencoders for each RCIP area consist of an assignment to each collision of a triple of numbers between 0 and 1. These 'latent variables' can be thought of as thought of as representing a point in three-dimensional space. The relative positions of these points are determined by the similarity of collisions based on input data; collisions that are similar are placed closer together, whilst collisions that are less similar are placed further apart. This allows the second half of the model to recreate the input data from the latent variables more efficiently.

Once hierarchical clustering of collisions, based on the relative position of the latent variables, has taken place, each collision is assigned to a cluster. This assignment partitions the overall collection of collisions in each RCIP area into a relatively small number of clusters (between 30 and 70) on the basis of their similarity to each other.

Analysis of the collisions in each cluster was then conducted to determine the features observed in the input data that best define their similarities. A description of each cluster is formed from a list of these features, which are shared by most of the collisions in the cluster. This description is then represented visually in the corresponding infographic.

Note that although these features are true of *most* collisions in a cluster, they may not all apply to *every* collision in that cluster. Likewise, a cluster may not contain all the collisions that fit the description assigned to that cluster, as a few such collisions may have other characteristics to which the model assigned greater weight and therefore may result in them being assigned to different clusters.

In addition to a description of the types of collision contained in each cluster, supplementary information is calculated to provide insights into how these types of collisions often take place. This information includes breakdowns by severity and road class, as well as the total number of collisions and the proportion that are fatal or serious.

2.3.2 Grouping

To organise clusters in a more useable way, a series of overarching groups were formed on the basis of some of the more prevalent features appearing in the descriptions of clusters. This process of assigning to groups was subjective and carried out for ease of presentation and interpretation, although assignment was nevertheless rooted in the outcome of the analysis.

Note that these groups do not form a partition of the clusters, and it is common for a cluster to naturally belong – in terms of its prevalent features – to more than one group. In these cases, clusters have been assigned to the first group in the hierarchy to which they could belong. There is a further element of subjectivity involved in forming such groups into a hierarchy of this kind, which should be noted. However, it is hoped that this process will provide useful insight into how collisions resemble each other in a way which may signify elevated risk, particularly for vulnerable road users and where associated with specific behaviours.

2.3.3 Siblings

Within each of these groups the clusters were classified again, into collections referred to as 'siblings'. The name was chosen to indicate that the groups of clusters have common features, and yet can also be placed into more detailed groups which have additional similarities as well as inheriting the overall characteristics of their group.

Sibling grouping is again based on the shared features of different clusters, in addition to the overarching similarity which assigned them to the same group. The approach used was initially to be as parsimonious as possible, and to group clusters which have only a small number of such defining features in common. These are then refined into gradually more specific sibling groups, containing clusters which 'inherit' all features of the broader group above them, as well as sharing additional features which describe in more detail the type of collisions they contain. This hierarchical approach provides insight into how a generalised group of collisions in fact often contains several related subordinate cluster families. Many such groups contain 'sibling' groups which both inherit characteristics of the wider family and also have unique distinguishing characteristics of their own.

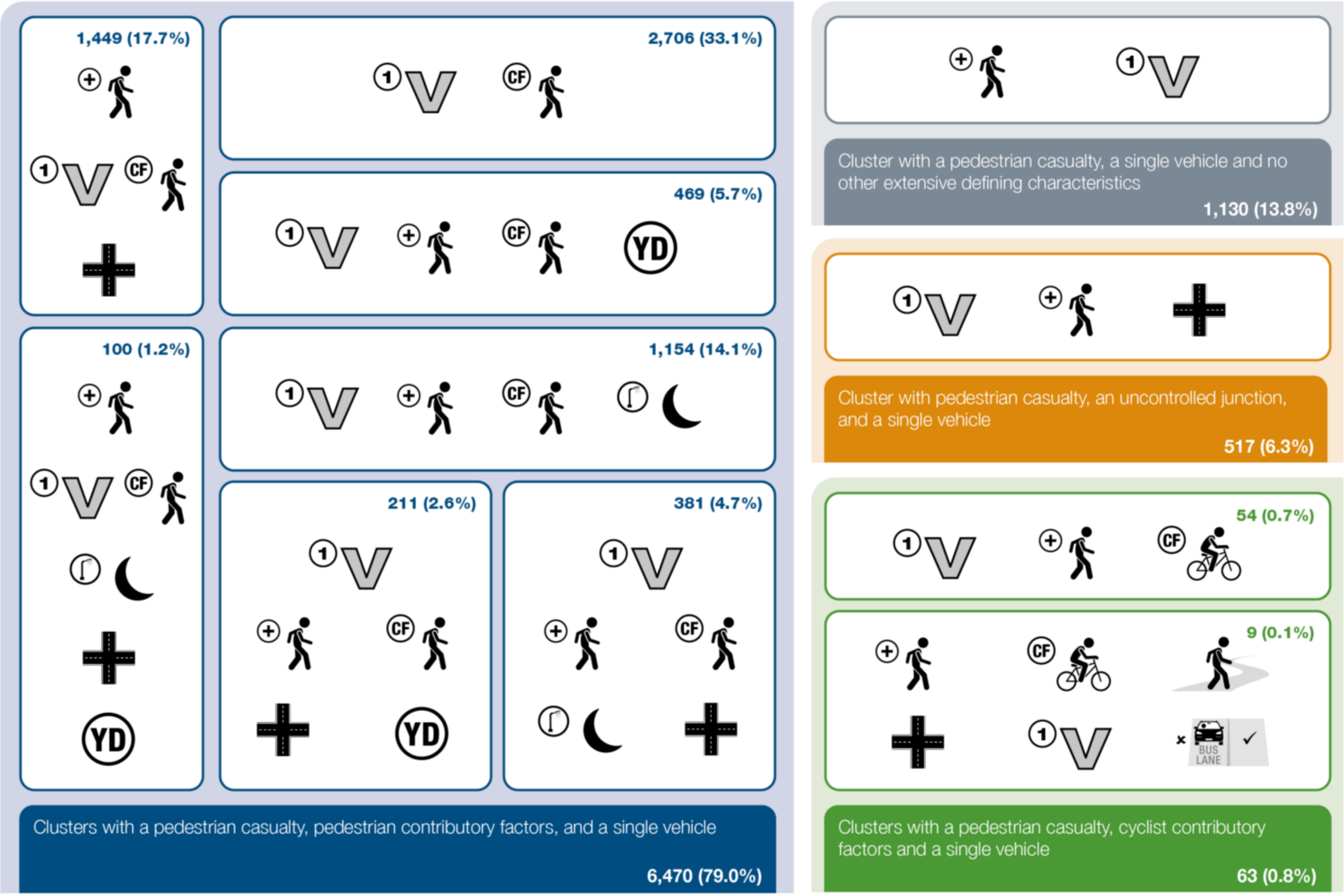
2.3.4 Depiction

To assist users of this research in understanding the output of this process, a series of infographics have been devised. A sample infographic is presented in Figure 2.1.

The clusters in each group are shown by the following diagrams. In each diagram:

- each coloured area shows a family of collisions within the group that have been grouped together based on similar characteristics;
- each of the inner boxes within that family represents sibling or 'Grandsibling' clusters that divide up these shared characteristics down to another level of separation; and
- all collision totals are additive, so percentages are based on the overall total for the entire group (and may not add up to 100% due to rounding).
- The key for the associated meanings represented by each infographic within the diagrams can be found in Appendix E.

Figure 2.1: Sample cluster family diagram


1,449 (17.7%)
2,706 (33.1%)
469 (5.7%)
100 (1.2%)
1,154 (14.1%)
211 (2.6%)
381 (4.7%)
Clusters with a pedestrian casualty, pedestrian contributory factors, and a single vehicle
6,470 (79.0%)
Cluster with a pedestrian casualty, a single vehicle and no other extensive defining characteristics
1,130 (13.8%)
Cluster with pedestrian casualty, an uncontrolled junction, and a single vehicle
517 (6.3%)
54 (0.7%)
9 (0.1%)
BUS LANE
Clusters with a pedestrian casualty, cyclist contributory factors and a single vehicle
63 (0.8%)

Source: Author's own

3 Future Recommendations

As is inevitable given the novel analytical approach taken by this research, some aspects of the methodology could usefully be refined for similar investigations in future. This section lists some thoughts which suggest themselves to the present authors.

- The present model was designed to cluster similar collisions on a single pass, but this does not take full account of the multivariate nature of the data. A similar methodology could be adopted in future to identify commonality at driver/rider or casualty level. This classification could then be fed back into a model to re-categorise specific collision groups from scratch. This approach should produce a more refined output, as classifications at lower levels would reduce 'noise' in the final output.
- The only sociodemographic data employed in the present model is from the IMD, but this is not the only possible approach to understanding the background of people involved in collisions. Commercial sociodemographic classifications which embrace a wider range of inputs may provide a richer typology for drivers, riders and casualties, potentially adding variables along the lines of 'affluent rural driver' and 'deprived suburban cyclist'. The value of adding such richer sociodemographic variables would increase in line with the size of the datasets, for example at regional or national level.
- Driver and casualty distance from home was not incorporated into the present analysis because neither of the proxy variables available within resource (crow-fly distance between home postcode and crash location, and commonality of local authority area) were considered sufficiently discriminating to provide quality input. However, there is potential for creating a more meaningful approach to the extent to which people involved in collisions come from the same communities. The ultimate objective would be to provide characterisations along the lines of 'goods vehicle driver far from home' and 'pedestrian casualty who lived locally'.
- The use of collision dynamics in the current study was the first time this derived variable has been used in research, and it is likely to benefit from further refinement. For instance, the 'other impact' dynamic was used frequently as a feature of clusters, but it is not a helpful identifier. This definition is likely to derive from anomalies in underlying data fields such as direction of travel, and it may be that a better approach to understanding these conflicts can be devised.
- There is a problem with rarely recorded features of collisions – for example, infrequently used contributory factors – being underused by the model. A future study could investigate the impact of using oversampling techniques to better balance the input variables, which may better highlight commonalities between collisions with rare features. This would entail designing a new model as it would be important to prevent such a method from detracting from identifying commonalities that apply to large numbers of collisions.
- All flow data was derived from DfT count point data, which has limited coverage outside major routes. It may be possible for future studies to acquire better data quality for weight of traffic on different roads by making use of telematics data. This could also include data on average vehicle speeds and even levels of speed compliance at crash locations.
- Future studies may obtain more detailed and focused typologies for specific groups of collisions by pre-filtering input data to focus exclusively on areas of interest. For example, collisions involving pedestrians and collisions involving specific driver demographics, such as young drivers, may benefit from this approach.
- There were several input variables that did not feature prominently in the final clusters, which intuitively might have been expected to have relevance. These include the time of day and day of the week when collisions occurred, and some contributory factor groups including speed choice, distraction, close following and fatigue. It would be helpful to re-evaluate how these variables are included in future machine-learning models, to elucidate if their absence in the present analysis is an unintended consequence of the methodology, or a genuine indication that they are not usually clustered with other collision features to any substantial degree.
- Occasionally, the model identified distinct clusters which have the same defining traits. These 'twin' clusters are an artefact of the iterative process which eventually results in cluster formation, in situations where

early iterations result in similar clusters beginning to form far apart in the latent space. It may prove possible to eliminate this phenomenon with refinements in model design and input variables.

Appendix A: Machine Learning

A.1 Deep Learning Models

Deep autoencoders take the high-dimensional input data containing information on collisions, vehicle involvement and involved casualties, encode this using the first half of the neural networks into a low-dimensional latent space, and then decode the data again into the original high-dimensional state. In theory, in order to have the most efficient autoencoders, similar collisions are placed closer together in the latent space, whilst dissimilar collisions are placed further apart.

Once collisions have been autoencoded into the latent space, and arranged by their levels of similarity, an agglomerative hierarchical clustering algorithm is applied. This process starts with all collisions as separate clusters, and iteratively joins the most similar clusters together until all collisions are in a single cluster.

Towards the end of this process, the silhouette value is measured at each stage; this measures each collision's cohesion (how similar it is to other collisions in the same cluster) relative to its separation (how similar it is to collisions in other clusters). This value is used to determine the stage at which clusters are taken from the hierarchical clustering algorithm, and hence influences the number of clusters to be analysed.

It is important to note that both the encoding and decoding halves of the neural network form non-linear functions, and so the best measure of distance within the latent space in the middle may not be the standard Euclidean 'crow-fly' metric. However, determining the ideal metric is an intractable problem, and so the Euclidean metric is the best compromise. This does mean that some caution should be exercised when using measures such as the silhouette value. It is also important to use best judgement, alongside the silhouette value, when deciding the number of clusters: too many clusters and the types of collisions they contain will become too specific to be useful; too few clusters and the types of collisions they contain will not have enough distinguishing features to be useful.

Another potential consequence of the non-Euclidean nature of the latent space is that the ideal sizes of clusters may not form at the same rate. As a consequence, some clusters in the final outputs may seem too large and somewhat lacking in distinguishing features, as if they were formed by combining clusters that were of a more useful size, whilst other clusters may seem too small and niche, as if they should ideally be combined with other clusters to become a more useful size. However, the arranging of clusters amongst their siblings in the infographics should help the reader to visualise which clusters could have been combined together into larger clusters.

Appendix B: Potential Comparator Areas

The following areas, comprising either single police force areas or combinations thereof, were identified as potential comparators for the RCIP areas. These groupings were primarily chosen owing to their proximity in the initial cluster analysis output. To avoid very small comparator areas which would not provide a sufficiently robust sample, every force with an annual average of fewer than 2,000 reported injury collisions was paired with the most similar larger force. Where the same force was most similar to more than one forces, priority was given to the highest mutual similarity. However, to prevent artificial groupings far larger than any actual forces, it was decided to avoid creating comparator groupings which had on average more collisions than the largest single potential comparator force (West Yorkshire).

For example: for both Bedfordshire and Northamptonshire, Cheshire was the most similar force; but while Northamptonshire was also the most similar force for Cheshire, Bedfordshire was only the third most similar force for Cheshire. Grouping all three forces would have created a potential comparator larger than the largest single comparator force (West Yorkshire). Therefore, Northamptonshire and Cheshire were paired with each other, while Bedfordshire was paired with its next most similar force (Derbyshire). This created two reasonably sized comparator areas which were still closely related by the cluster analysis output.

The following list shows all 28 potential comparator areas in descending order of average number of reported injury collisions annually,[3] with the final selected comparators shown in bold:

- West Yorkshire (the largest potential comparator, averaging 5,150 collisions annually)
- Thames Valley
- Kent
- Lancashire paired with Gwent
- Police Scotland South paired with Cumbria
- Hampshire
- North Yorkshire grouped with Wiltshire and Gloucestershire (the only grouping of three force areas)
- Northumbria paired with Durham
- **Greater Manchester** (averaging 4,314 collisions annually, rather smaller than West Midlands, which has 5,760 annually)
- Sussex
- Cheshire paired with Northamptonshire
- Derbyshire paired with Bedfordshire
- Essex
- Surrey
- Police Scotland Strathclyde
- Cambridgeshire paired with Warwickshire
- Lincolnshire paired with North Wales
- Norfolk paired with Suffolk
- Staffordshire paired with Cleveland
- **Avon and Somerset** (averaging 3,320 collisions annually, smaller than Devon and Cornwall with Dorset, which has 5,707 annually)
- Police Scotland North paired with Dyfed–Powys
- South Yorkshire
- Nottinghamshire
- Merseyside

[3] Total collisions were calculated over the period 2009 to 2018, and were based on data downloaded from MAST Online (see Appendix C).

- Hertfordshire
- **West Mercia** (averaging 2,448 collisions annually, a similar size to Humberside, which has 2,498 annually)
- South Wales
- Leicestershire (the smallest potential comparator, averaging 2,276 collisions annually)

Appendix C: Input Variables

C.1 Data Sources

The data used for the analyses described in this paper was drawn from government sources. Collision, vehicle and casualty data was taken from MAST Online[4] on the basis of police recorded injury collisions in Great Britain between 2009 and 2018 as supplied by DfT.[5] Deprivation statistics were obtained by matching postcodes to the IMD.[6] Traffic information was taken for individual count points,[7] spatially matched to road sections where collisions occurred, and then analysed to categorise roads by relative traffic quartiles wherever data was available.

For an explanation of how these input variables were applied during machine learning, see section 2.2 Input data above.

Table C.1: Collision input variables

Group	Title	Type	Definition	National model usage
101a	Severity_Fatal	Boolean	True: at least one casualty was killed	Used subtly
101b	Severity_Serious_Adjusted	Continuous	Probability that at least one casualty would have been classified as serious if injury-based reporting had been in place	Used subtly
102a	Junction_Controlled	Boolean	True: junction with ATS (automatic traffic signal) or authorised person	Used subtly
102b	Junction_Uncontrolled_Roundabout	Boolean	True: junction with roundabout or mini-roundabout	Used subtly
102c	Junction_Uncontrolled_Other	Boolean	True: junction with Give Way or Stop (not at roundabout)	Used extensively
103	Weather_Adverse	Boolean	True: any inclement weather conditions (rain, snow, fog, high winds or other)	Used subtly
104a	Date_PH	Boolean	True: was a weekday public holiday (Christmas, Easter or bank holiday)	Ignored as irrelevant
104b	Date_Weekend	Boolean	True: was a Saturday or Sunday	Ignored as irrelevant
105a	Time_Rush_AM_7to9	Boolean	True: was at or after 7 a.m. and before 9 a.m.	Used moderately
105b	Time_Night_7to7	Boolean	True: was at or after 7 p.m. and before 7 a.m. the following day	Used moderately
106a	Night_Streetlights	Boolean	True: was dark, and streetlights were present and lit	Used extensively
106b	Night_NoStreetlights	Boolean	True: was dark, and no lit streetlights were present	Used moderately

[4] MAST Online: see www.roadsafetyawards.com/winner/2010/mastonlinemastproject

[5] The publicly available version of this data can be downloaded from https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data. This source does not include contributory factors or directions of travel, which were supplied separately.

[6] http://imd-by-postcode.opendatacommunities.org/imd/2019

[7] https://roadtraffic.dft.gov.uk/#6/55.254/-6.053/basemap-regions-countpoints

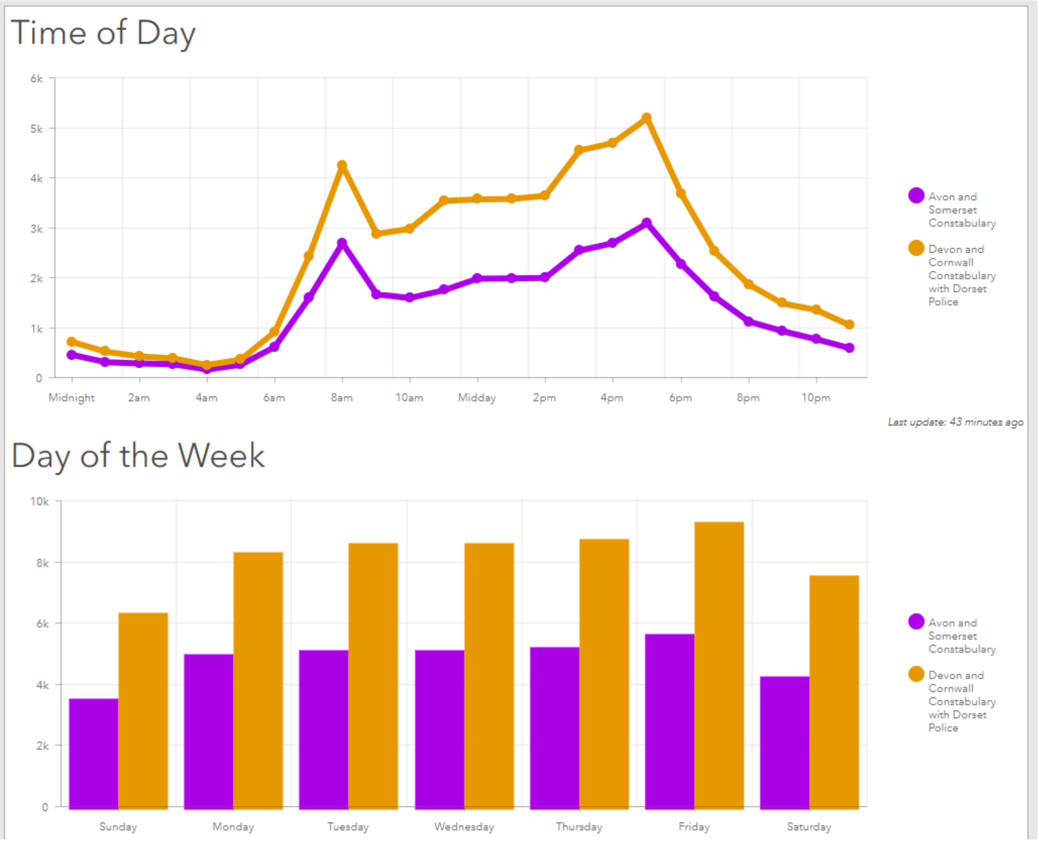
107	Vehicles_Single	Boolean	True: only one vehicle was involved	Used extensively
108	Population_Density_Raw	Continuous	Population per square km of LSOA / data zone in mid-2018	Ignored as irrelevant
109	Dynamics_HeadOn	Boolean	True: at least one vehicle had a front impact; *and* at least one other vehicle travelling in the opposite direction also had an impact	Used moderately
110	Dynamics_Shunt	Boolean	True: at least one vehicle had a rear impact; *and* at least one other vehicle travelling in the same direction also had an impact	Used extensively
111	Dynamics_SideImpact	Boolean	True: at least one vehicle had a side impact; *and* at least one other vehicle travelling in an adjacent direction[8] also had an impact	Used moderately
112	Dynamics_OtherImpact	Boolean	True: at least two vehicles had impacts	Used extensively
113	Vehicles_Count	Continuous	Number of vehicles involved	Used subtly
114	Casualties_Count	Continuous	Number of casualties resulting (of all severities)	Used subtly

Source: Author's ow

[8] Adjacent direction: from the side, across the driver's path

Figure C.1 shows some collision variables applied to trend analysis in an area dashboard.

Figure C.1: Collision variables in an area dashboard



Source: Author's own

Table C.2: Vehicle input variables

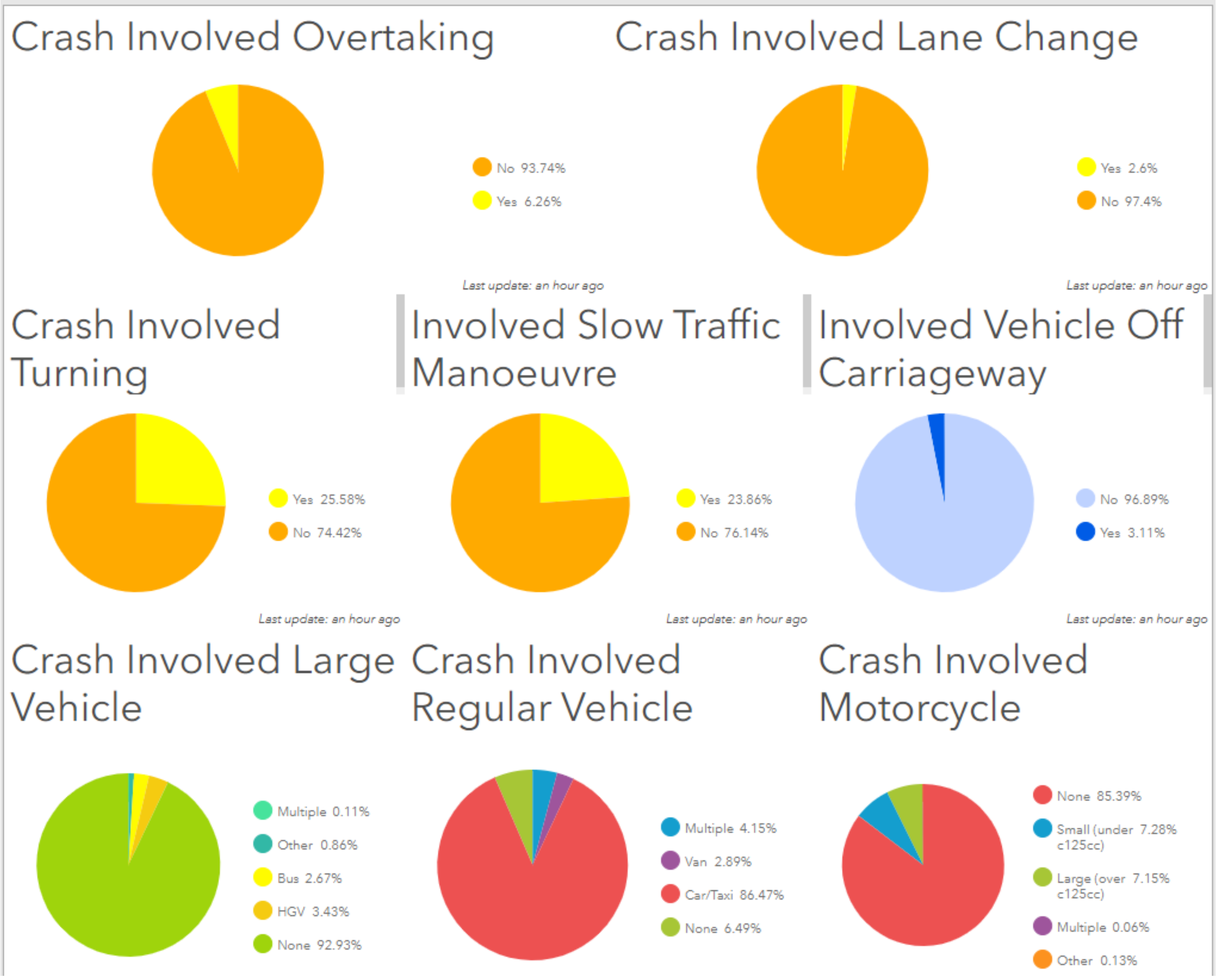
Group	Title	Type	Definition	Model usage
201a	Runoff_Nearside	Boolean	True: vehicle left carriageway to the nearside (whether rebounded or not)	Used extensively
201b	Runoff_Other	Boolean	True: vehicle left carriageway in any other fashion	Used moderately
202	Vehicle_HitRun	Boolean	True: vehicle was hit-and-run (excluding non-stop vehicles not hit)	Used subtly
203	Vehicle_NotInMainCway	Boolean	True: any vehicle on a footway; any vehicle on, entering or leaving a hard shoulder; a vehicle other than a bus in a bus lane or busway; or any vehicle other than a tram on a tram track	Ignored as irrelevant
204a	Vehicle_Overtaking	Boolean	True: vehicle was overtaking (offside or nearside)	Used subtly
204b	Vehicle_LeftTurn	Boolean	True: vehicle was turning left, or waiting to do so	Used moderately
204c	Vehicle_RightTurn	Boolean	True: vehicle was turning right, or waiting to do so	Used extensively
204d	Vehicle_SlowManeouvre	Boolean	True: vehicle was stopping, stationary or moving off	Used extensively

204e	Vehicle_LaneChange	Boolean	True: vehicle was changing lane (to left or right)	Used subtly
205a	Vehicle_Moped	Boolean	True: vehicle was a motorcycle with engine size 50cc or under	Ignored as irrelevant
205b	Vehicle_MC_MidSize	Boolean	True: vehicle was a motorcycle with engine size over 50cc up to 500cc (includes vehicles which were electric or of unknown engine size)	Used subtly
205c	Vehicle_MC_Large	Boolean	True: vehicle was a motorcycle with engine size over 500cc	Used subtly
205d	Vehicle_Large_GV_PSV	Boolean	True: vehicle was a bus, coach or tram; or a goods vehicle over 3.5 tonnes mgw or of unknown weight	Used extensively
206a	Driver_Young_Under25	Boolean	True: driver/rider of motor vehicle was aged 16–24 inclusive	Used moderately
206b	Driver_Old_70Plus	Boolean	True: driver/rider of motor vehicle was aged over 69	Used extensively
207	Driver_Deprived_BottomQuintile	Boolean	True: driver’s home postcode was in a LSOA classified by the ONS in the most deprived quintile of the Index of Multiple Deprivation	Used moderately
208	Driver_Working	Boolean	True: driver was recorded as working; and/or was driving a large vehicle; and/or was on a commuting journey in a taxi or light goods vehicle	Used extensively

Source: Author’s own

Figure C.2 shows some of these vehicle variables applied to trend analysis in an area dashboard.

Figure C.2: Vehicle variables in an area dashboard



Crash Involved Overtaking

Crash Involved Lane Change

Crash Involved Turning

Involved Slow Traffic Manoeuvre

Involved Vehicle Off Carriageway

Crash Involved Large Vehicle

Crash Involved Regular Vehicle

Crash Involved Motorcycle

Source: Author's own

Table C.3: Casualty input variables

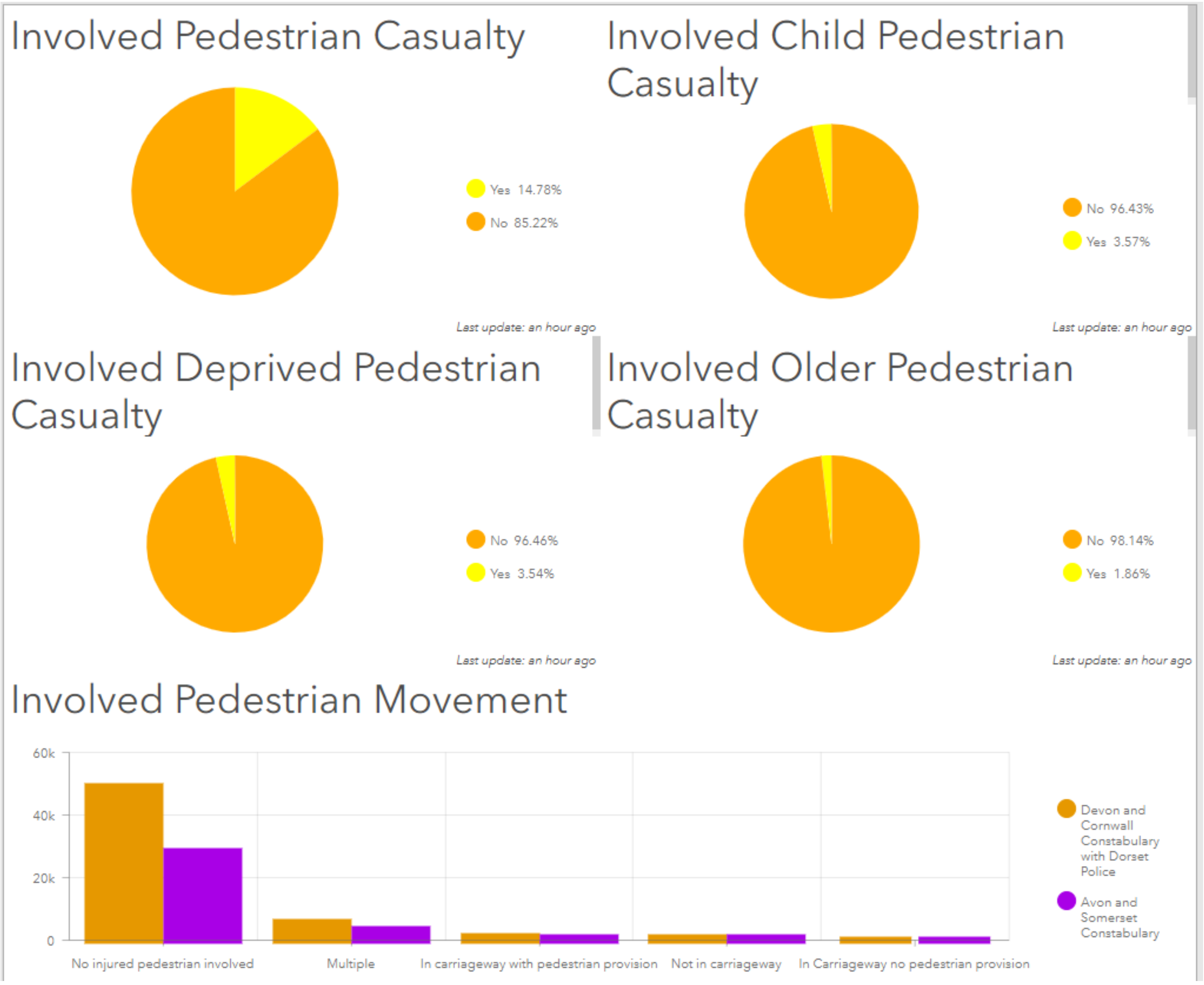
Group	Title	Type	Definition	Model usage
301a	Casualty_PCUser	Boolean	True: casualty was rider or pillion passenger on a cycle	Used extensively
301b	Casualty_HorseRider	Boolean	True: casualty was rider or pillion passenger on a horse	Ignored as irrelevant
301c	Casualty_MobilityScooterUser	Boolean	True: casualty was rider or pillion passenger on a mobility scooter	Ignored as irrelevant
302	Casualty_Pedestrian	Boolean	True: casualty was a pedestrian	Used extensively
303a	Casualty_ChildPedestrian_Under16	Boolean	True: casualty was a pedestrian aged under 16	Used extensively
303b	Casualty_OldPedestrian_70Plus	Boolean	True: casualty was a pedestrian aged over 69	Used moderately
304a	Casualty_Pedestrian_CrossingOrRefuge	Boolean	True: casualty was a pedestrian on a crossing, refuge or central island	Used moderately

304b	Casualty_Pedestrian_Footway	Boolean	True: casualty was a pedestrian on a footway	Used subtly
305	Casualty_Pedestrian_InCway_Masked	Boolean	True: casualty was a pedestrian anywhere in the carriageway who was masked by a stationary or parked vehicle	Used subtly

Source: Author's own

Figure C.3 shows some casualty variables applied to trend analysis in an area dashboard.

Figure C.3: Casualty variables in an area dashboard



Involved Pedestrian Casualty

Involved Child Pedestrian Casualty

Involved Deprived Pedestrian Casualty

Involved Older Pedestrian Casualty

Involved Pedestrian Movement

Source: Author's own

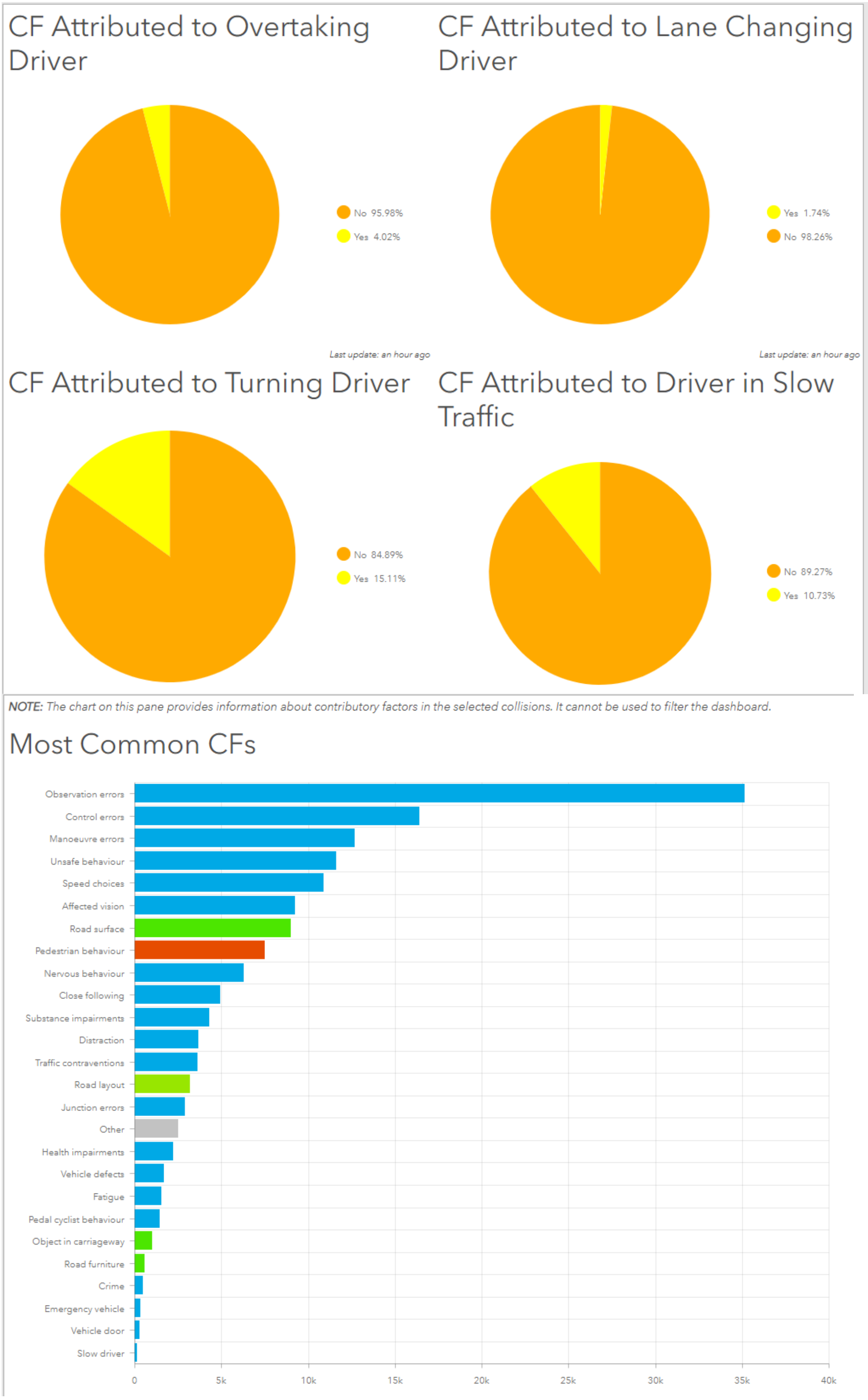
Table C.4: Contributory factor input variables

Group	Title	Type	Definition	Model usage
401a	Pedestrian_Casualty_Contributed	Boolean	True: any injured pedestrian or vehicle passenger had a pedestrian contributory factor (CF) assigned to them	Used extensively
401b	Pedestrian_Uninjured_Contributed	Boolean	True: any uninjured pedestrian had a pedestrian CF assigned to them	Ignored as irrelevant
402a	Driver_Contributed_Overtaking	Boolean	True: any overtaking driver or rider had any driver/rider CF assigned to them	Used subtly
402b	Driver_Contributed_Turning	Boolean	True: any turning driver or rider had any driver/rider CF assigned to them	Used extensively
402c	Driver_Contributed_LaneChange	Boolean	True: any lane-changing driver or rider had any driver/rider CF assigned to them	Used subtly
403a	Cyclist_Contributed	Boolean	True: any cyclist had any CF assigned to them	Used extensively
403b	P2W_Rider_Contributed	Boolean	True: any motorcyclist had any CF assigned to them	Used subtly
403c	Large_GV_PSV_Driver_Contributed	Boolean	True: any large vehicle driver had any CF assigned to them	Used extensively
404	Environmental_Factor_Contributed	Boolean	True: any participant had an environmental, vision-affected or other specific CF assigned to them	Used extensively
405	Vehicle_Factor_Contributed	Boolean	True: any driver or rider had a vehicle defect CF assigned to them	Ignored as irrelevant
406	Driver_Crime_Contributed	Boolean	True: any driver or rider had a crime-related CF assigned to them	Ignored as irrelevant
407	Driver_Intoxicated_Contributed	Boolean	True: any driver or rider had an intoxication CF assigned to them	Ignored as irrelevant
408	Driver_SpeedChoice_Contributed	Boolean	True: any driver or rider had a speed choice CF assigned to them	Used subtly
409	Driver_MobilePhone_Contributed	Boolean	True: any driver or rider had mobile phone CF assigned to them	Ignored as irrelevant
410	Driver_CloseFollowing_Contributed	Boolean	True: any driver or rider had close following CF assigned to them	Used subtly
411	Driver_Disobeyed_Contributed	Boolean	True: any driver or rider had any ‘disobeyed sign or marking’ CF assigned to them	Used subtly
412	Driver_Observation_Contributed	Boolean	True: any driver or rider had any observation CF assigned to them	Used extensively
413	Driver_Fatigue_Contributed	Boolean	True: any driver or rider had fatigue CF assigned to them	Ignored as irrelevant
414	Driver_Distracted_Contributed	Boolean	True: any driver or rider had any distraction CF assigned to them	Ignored as irrelevant
415	Driver_Careless_Contributed	Boolean	True: any driver or rider had aggressive and/or careless CF assigned to them	Ignored as irrelevant

Source: Author’s own

Figure C.4 shows some contributory factor (CF) variables applied to trend analysis in an area dashboard. The CF groupings on the right-hand pane in this illustration are categorised into those which refer to driver behaviour (in blue), pedestrian action (in red) and the road environment (in green).

Figure C.4: Contributory factor (CF) variables in an area dashboard



NOTE: The chart on this pane provides information about contributory factors in the selected collisions. It cannot be used to filter the dashboard.

Source: Author's own

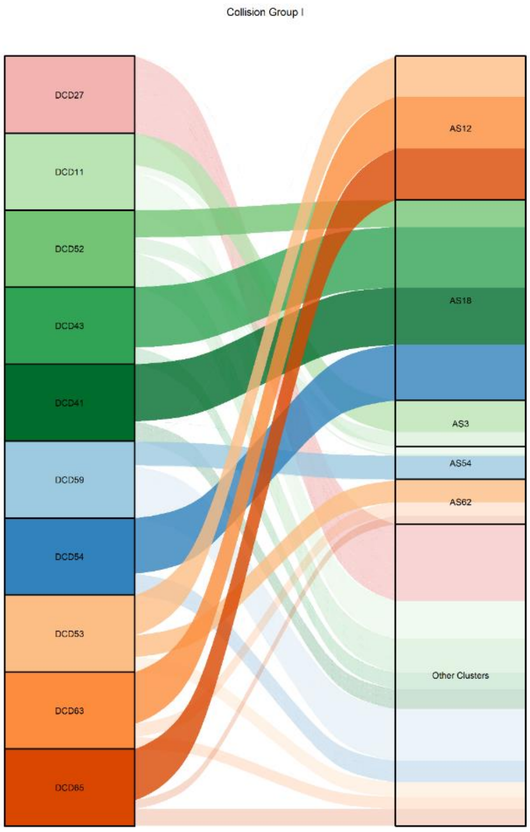
Appendix D: Projected Comparison Methodology

To provide insights into which types of collision are generally prevalent on all road networks and which are specific to an individual RCIP area, an attempt was made to compare clusters between RCIP areas and their chosen comparators. However, as clusters do not inherently have rigid definitions but instead exist as collections of collisions that are broadly similar, direct comparisons cannot be made.

To provide a best-fit comparison of clusters from an RCIP force area with clusters from the comparator area, the data held for collisions in each RCIP area were passed through the autoencoders of the models developed for the comparator areas. This provides each collision from the chosen RCIP area with latent variables which can be used to determine which cluster for the comparator area they would have most likely been assigned to. In this way, one can project each cluster for the chosen RCIP area onto one or more clusters for the comparator area and use this to determine which, if any, clusters for the comparator area are most similar.

This was carried out for each RCIP area, and Sankey diagrams were plotted to visualise these projections. For example, Figure D.1 shows the projection of clusters of collisions involving cyclist casualties from Devon and Cornwall Constabulary with Dorset Police onto the equivalent clusters for Avon & Somerset Constabulary. Each box on the left represents a cluster of collisions from Devon and Cornwall Constabulary with Dorset Police, and each box on the right represents a cluster from Avon and Somerset Constabulary. The flowing lines in the centre of the diagram show the movement of collisions from the clusters on the left to the clusters on the right under the projection. The thickness of the lines corresponds to the number of collisions in each movement.

Figure D.1: Sankey diagram visualising projection of DDC collisions involving cyclist casualties onto Avon & Somerset clusters



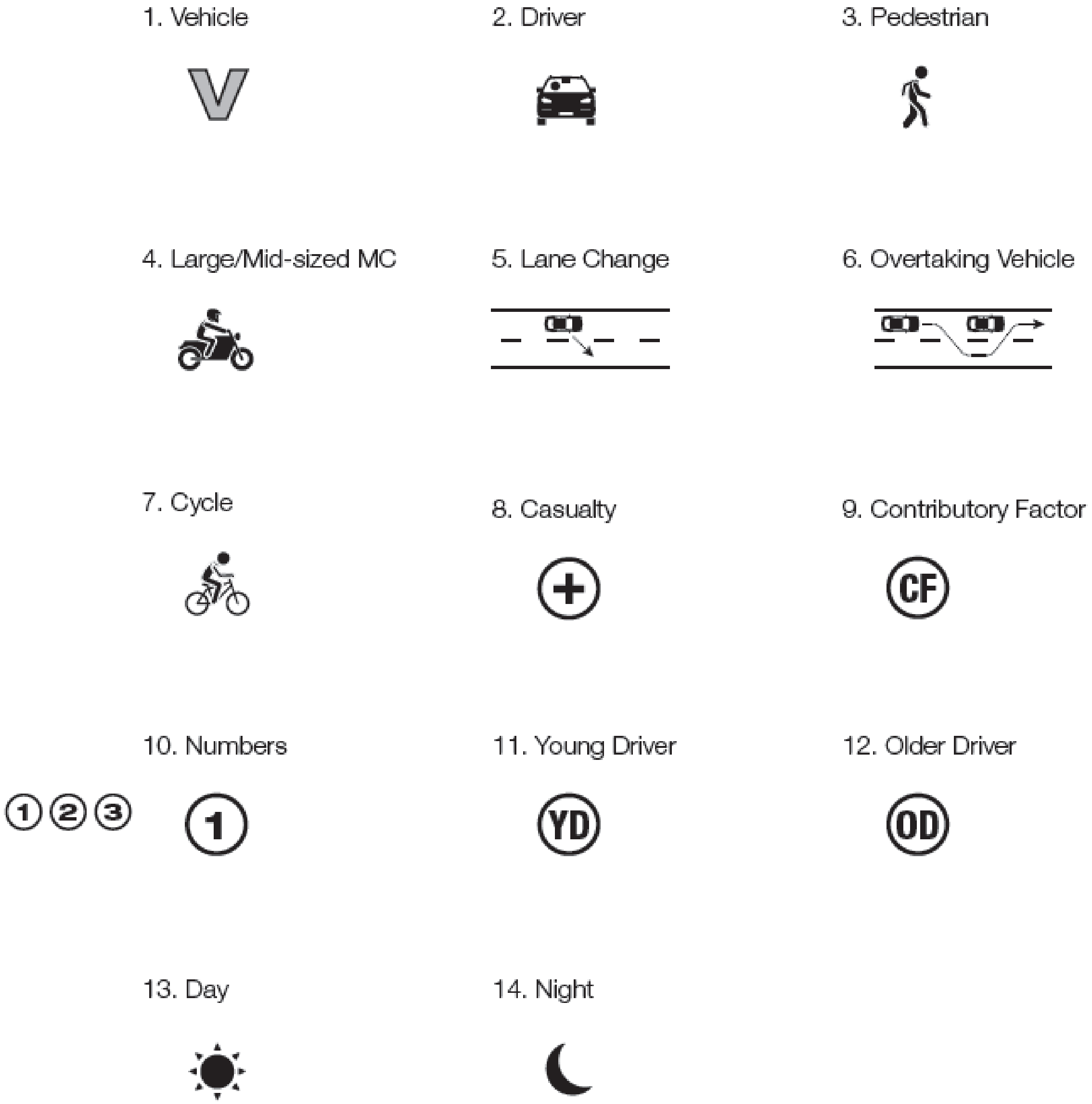
Source: Author's own

The complexity of such means that the insights that it might be possible to infer from them are limited. As a result, these diagrams have been excluded from this paper. However, where strong consonances between certain clusters in RCIP force areas and their comparator areas exist, this is noted in the relevant section of the area's profile.

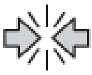
Appendix E: Infographics Key

Figure A.5 shows the icon definitions

Figure A.5: Icon directory (source: author’s own)



15. Other Impact



16. Driver Observation CF

17. Shunt

18. Uncontrolled Junction

19. Right Turn

20. Driver Turning CF

21. Environment CF

22. Runoff (nearside)

23. Young Driver

24. Older Driver

25. Night (streetlights)

26. Night (no streetlights)

27. Runoff (other)

28. Pedestrian footway

29. Slow Vehicle Manouever

30. Working Driver

31. Head On

32. Roundabout

33. Cyclist Casualty

34. Pedestrian Casualty

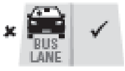
35. Deprived Driver

36. Hit and Run

37. LGV/PSV

38. Side Impact

39. Vehicle not in Carriageway



40. Pedestrian CF

41. Single Vehicle

42. Child Pedestrian Casualty

43. LGV/PSV CF

44. Driver Overtaking CF

45. P2W rider CF

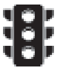
46. Large MC

47. Mid-sized MC

48. Moped

49. Left Turn

50. Controlled Junction



51. Pedestrian on Crossing

52. Adverse Weather

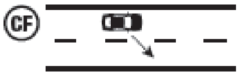
53. Weekend

54. Close following CF

55. No distinguishing Features

56. Speed Choice CF

57. Lane Change CF



58. Careless driver CF

59. More likely hit and run

60. Cyclist CF

61. Working Young or Old

RAC Foundation

Mobility • Safety • Economy • Environment

The Royal Automobile Club Foundation for Motoring Ltd is a transport policy and research organisation which explores the economic, mobility, safety and environmental issues relating to roads and their users. The Foundation publishes independent and authoritative research with which it promotes informed debate and advocates policy in the interest of the responsible motorist.

RAC Foundation

89–91 Pall Mall

London

SW1Y 5HS

Tel no: 020 7747 3445

www.racfoundation.org

Registered Charity No. 1002705