RAC Foundation

Mobility • Safety • Economy • Environment

Rural road segmentation research

Dr Craig Smith, Suzanne Coles MIHE and Bruce Walton

March 2023

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

RAC Foundation

Mobility • Safety • Economy • Environment

Rural road segmentation research

Dr Craig Smith, Suzanne Coles MIHE and Bruce Walton

March 2023

About the Authors

Suzanne Coles has developed her knowledge and understanding of road casualty data over 17 years working for Local Government. As a Member of the Institute of Highway Engineers (MIHE) and with a master's degree in Transport Planning & Engineering she has expertise in identifying and delivering road safety engineering schemes that seek to reduce road casualties and encourage the safe and effective movement of traffic. In addition, Suzanne has worked on process improvements and infrastructure upgrades to enhance network management for road users. Suzanne combines her skills of analysis and interpretation with her insight from working in local government to contribute to the development and delivery of reliable and effective road safety solutions.

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 an 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.

Acknowledgements

UK Centre for Ecology and Hydrology (UKCEH)
Somerset County Council
East Riding of Yorkshire Council
Norfolk County Council
Road Safety Foundation
British Cycling
Department for Transport
Kent County Council
Surrey County Council
West Sussex County Council
East Sussex County Council

Contents

List of Figures

List of Tables

Foreword

Two things those of us working in road safety know are that in this country we are very fortunate to have amongst the best road casualty data in the world to work from and that it is our rural roads that pose amongst the greatest road safety risks. We also know that our road safety data isn't perfect – there's always room for improvement – and so we need to reflect from time-to-time on whether we really understand what that data is telling us.

This report is the result of one such reflection, specifically posing the question: is the 'rural' designation helpful or potentially misleading? I found myself asking what image comes to mind for a 'rural road'; the answer is a narrow country lane, bounded by hedgerows, twisty in places, hence hard for drivers to anticipate what lies ahead. I tried that question on others working in the sector and got similar answers back. But in the world of road casualty data the term 'rural road' actually encompasses far more than country lanes, including many stretches of dual carriageways – 'rural' in this context simply means 'not urban'.

Having for many years rather loosely referred to 'rural road risk' myself I began to worry that in doing so I might have been misleading others about the real nature of that risk and, by turn, what best might be done by way of mitigation and where best to concentrate our efforts. The latest annual casualty data helps illustrate the point.

In 2021, there were 1,558 reported road deaths in Great Britain, 981 of them on so-called rural roads. However, 649 (66%) of the 981 rural road deaths occurred not on the twisting country lanes of our imaginations but on motorways (86 deaths) and A roads (563 deaths).

Looking more broadly at all those killed or seriously injured on rural roads, 55% (6,458 out of 11,693) of casualties died or were hurt on motorways or A roads.

That is why we commissioned expert road safety analysts Agilysis to explore whether the road safety community might be better served by having a more sophisticated – a more accurate – categorisation. Not an over-complicated typology, but something that practitioners would find informative in their work and useful in briefing others. Something that would, at minimum, recognise that the road-related risk of a fast-moving stretch of dual carriageway, albeit far from a built-up area, is materially different from that posed by a narrow, lightly trafficked single-track lane.

The results of the Agilysis analysis are set out in this report. I think Agilysis are on to something, but in many senses the real work starts from here, because what's needed is an informed debate about whether having such a categorisation is genuinely useful, if so whether the Agilysis framework is right or needs some refinement and, once refined, how best might this approach be adopted. I suspect it's a debate somewhat overdue, but it's one the Foundation is keen to foster.

Steve Gooding

Steve Gooding

Director, RAC Foundation

Executive Summary

For many years the Department for Transport have applied a binary rurality categorisation to Britain's roads, which assigns each police recorded injury collision to either an 'urban' or a 'rural' road. This study examines the diverse collection of roads presently categorised as 'rural', with the goal of proposing a new typological approach which describes rural highways based on road user experience.

The project set out to:

- Assemble data for the highway characteristics of 'rural roads' as presently defined.
- Examine how those characteristics could generate more informative definitions.
- Arrange these definitions into a formal taxonomy with reproducible outputs.
- Evaluate the taxonomy and seek stakeholder feedback.

A random sample of 'rural road' routes was selected, stratified by road class, with total sample length weighted by relative traffic volume according to national statistics. The sample included 434 routes covering a total 1,563km of Britain's 'rural road' network.

The project began by using machine learning to cluster all sampled routes using input data which covered the physical characteristics of the carriageway, the usage of the road, and the local environment. The output from this model grouped the sample routes into eleven clusters, further arranged into seven superclusters.

The machine learning output was then subjected to qualitative review to verify the allocation of routes to clusters, based on human assessment of their attributes and physical characteristics. Following this review, the taxonomy was refined to ten clusters, which were grouped into four superclusters.

A small number of routes from across the whole sample were held back from the qualitative review for use as a blind sample to test the machine learning outputs further. The results of this test were strong, showing high correlation between the model and qualitative analysis assignments.

The four superclusters defined in the final version of the taxonomy were:

- ***Principal Roads***, split into two clusters *Regional Distributors* and *Main Roads*
- ***Country Roads***, split into three clusters *Primary Routes, Secondary Routes*, and *Tertiary Routes*
- ***Neighbourhood Roads***, split into two clusters *Residential Distributors* and *Village High Streets*
- ***Winding Roads***, split into three clusters *Country Lanes, Hill Passes* and *Remote Roads*.

STATS19 police reported collision data was intentionally not used as input data. Instead, it was used to test the taxonomy further, by identifying any substantive differences in risk posed to road users on different categories of rural road. This analysis, carried out using reported injury collisions on the sample routes from 2015 to 2019, sought to determine whether patterns existed between collisions within the same super cluster, or between super clusters. It revealed some clear patterns which are consistent with the relative properties of the clusters.

A project Development Day was held when expert stakeholders were invited to scrutinise the proposed taxonomy, offer feedback, and consider the most constructive ways forward. Stakeholders generally agreed that moving away from a binary taxonomy of urban and rural roads would be more reflective of network reality and therefore could promote improved understanding and decision making. The taxonomy would provide better insight into high harm routes, make it easier to identify comparable roads within and between authorities, and provide potential collateral benefits for other highway authority activities such as speed limit reviews, maintenance, and planning applications.

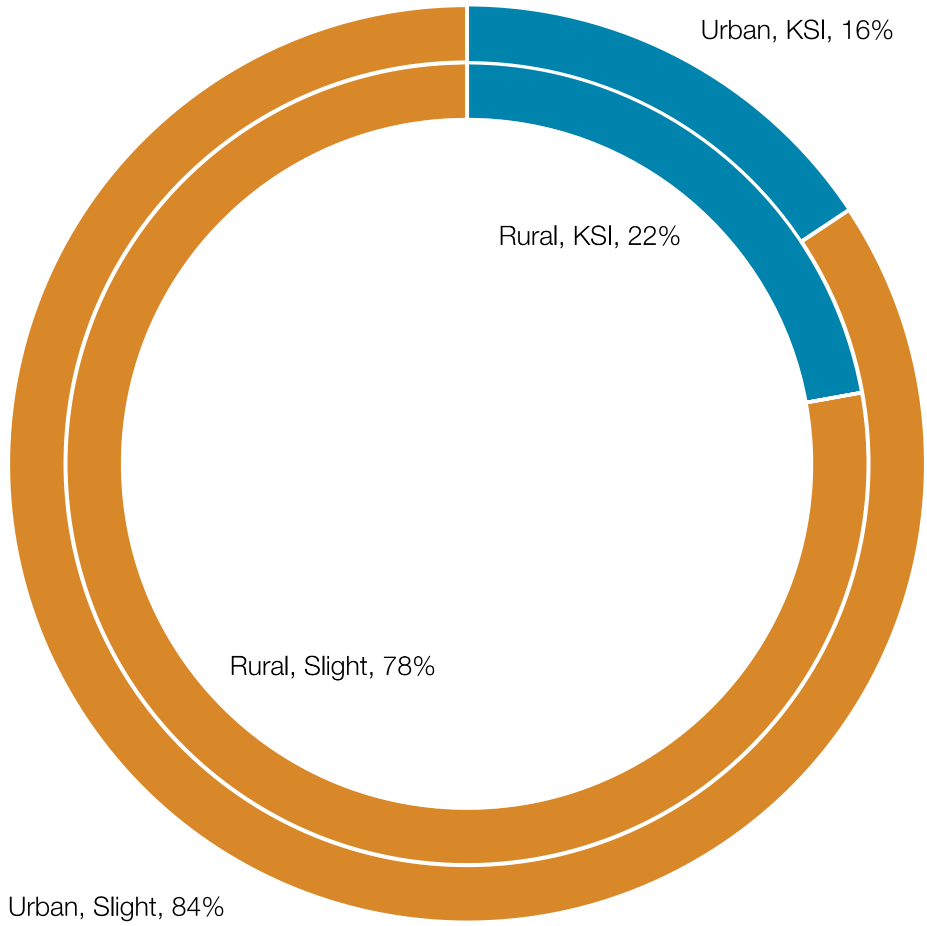
To achieve these potential benefits identified by this study, both stakeholders and the research team felt that the proposed taxonomy now requires additional validation based on a larger and more contiguous rural network than a nationwide random sample. Accordingly, this report proposes a methodology for a data driven process to apply it, using a decision tree. This could be used by a future project to validate the taxonomy and refine it as necessary, with the ultimate objective of making the benefits identified by stakeholders available nationally.

1. Project background and objectives

1.1 Background – the current definition of rural roads

For decades, the Department for Transport (DfT) has assigned rurality to police-recorded injury collisions on Britain's roads. This binary classification characterises each collision as occurring on either an 'urban' or a 'rural' road. Consequently, official figures appear to suggest that the road network can meaningfully be divided into two broadly consistent categories (see Figure 1.1).

Figure 1.1: Reported collisions on 'urban' and 'rural' roads 2012–2021 [source: MAST Online, www.roadsafetyanalysis.org]



However, this classification is not based upon any characteristics of the highway itself. Rather, it is predicated on the statistical category of areas through which the road passes, which is ultimately derived from census returns. Consequently, the phrase 'rural roads' does not precisely describe the category; a phrase such as 'roads which pass through rural communities' would describe them more accurately.

Nevertheless, the term 'rural roads' has stuck, and there is consequent dissonance between the apparent meaning of the classification and its actual provenance. This dissonance has led to a lack of accurate technical appreciation among users of the data of what it really represents. For instance, non-expert decision makers may easily misapprehend which roads are being discussed. Misunderstandings of this kind can result in poor decisions; for example, in selecting appropriate interventions to address road risk. Even highly knowledgeable data users are prone to being misled. Any classification that places much of the M6 into the same category as single-track roads winding up mountainsides is likely to be problematic as a basis for robust analysis.

This study sets out to examine the diverse array of roads presently described as 'rural', with the goal of proposing an alternative typological approach that would be more descriptive of the highway as road users experience it and consequently would render the categorisation far less misleading. The following sections describe the basis for the present categorisation, which differs slightly between countries of the UK.

1.1.1 England and Wales

For roads in England and Wales, DfT uses a polygon layer that defines built-up areas, and rural areas are simply all parts of the country that lie outside these polygons. Built-up areas were stored as rasterised shapes dissolved from urban polygons with a 50 m generalisation. The definition of urban polygons were based on the ONS rurality classification[1] of census Output Areas, which considers an area to be 'built up' if it forms part of a contiguous conurbation with a population of at least 10,000.

Consequently, the phrase 'rural road' in England and Wales presently means any road that lies at least 50 m outside Output Areas in a town or city with a population of more than 10,000.

1.1.2 Scotland

For roads in Scotland, the Scottish Government applies different criteria in defining this statistic, as it is a devolved matter.[2] The classification is based on Data Zones, which are geographic areas of approximately equal population aggregated from Census Areas. It treats small settlements differently to the equivalent ONS classification: conurbations of between 3,000 and 10,000 people are described as *Small Towns*, and roads in these areas are considered 'urban'.

Consequently, the phrase 'rural road' in Scotland presently means any road that lies outside Data Zones comprising a town or city with a population of more than 3,000.

1.1.3 Issues

In summary, there are several issues that make it challenging to apply the present definition to practical applications:

- The lack of correlation between the category as defined and actual road-user experience, with consequent difficulties in the applied study of road risk and future danger;
- Potentially misleading dissonance between the title of the classification and its actual provenance, leading to understandable confusion about its meaning; and
- Dissimilarities in definition for various parts of the UK, leading to diminished comparability.

This study sets out to propose an alternative methodology for defining a 'rural road', based upon its own nature rather than the statistical geographic units that contain it.

1 https://www.ons.gov.uk/methodology/geography/geographicalproducts/ruralurbanclassifications/2011ruralurbanclassification
2 https://statistics.gov.scot/resource?uri=http%3A%2F%2Fstatistics.gov.scot%2Fdata%2Furban-rural-classification

1.2 Project objectives and outline

The project set out to achieve the following objectives:

a. Assemble data for the engineering, environmental and experiential characteristics of highways within the present definition of 'rural roads';
b. Examine how those characteristics could be used to generate more informative definitions;
c. Strive to arrange these definitions into a formal taxonomy, with reproducible outputs; and
d. Evaluate the taxonomy by applying it in collision analysis and seeking stakeholder feedback.

1.2.1 Project principles

To ensure that the output would be as robust and reproducible as possible, the following overarching principles were established from the outset:

- **All input data to be derived from consistent and nationally available sources**. This principle was adopted to eliminate the danger that local variations in information availability, configuration or quality could create apparent differences between roads, which were in fact mere artefacts of input data.
- **No collision data to be used as input to the definition process**. Since one of the project objectives was to generate a taxonomy fit for use in collision analysis, using collisions as input could create a 'circular reference' issue; applying the output definitions to collisions may simply reflect the same criteria that were selected for input.
- **The project design would incorporate quantitative and qualitative components**. While the taxonomy should be informed by objective facts about the highways as far as possible, an amount of human judgement would also be indispensable, to incorporate professional expertise and reflect the project objective of providing insight into subjective road-user experiences.
- **Explore stakeholder opinions and identify potential applications**. It was recognised that feedback from national and local stakeholders would be of value in developing the taxonomy and identifying ways in which it could be applied.

1.2.2 Project design

Following project initiation and the establishment of these guiding objectives and principles, the following phases were identified. The outcomes of each of these are described in detail below.

1. **Defining routes and sampling**. Identify the national network of rural routes by applying existing definitions to road sections, and then select from this network a random sample that is broadly representative of rural roads in Britain.
2. **Initial phase – machine learning**. Devise and apply a machine-learning algorithm for multivariate analysis of the relevant properties of this sample, and including representing engineering, environmental and usage factors.

3. **Qualitative analysis**. Perform a thorough qualitative analysis of output clusters from the machine-learning model, to establish how effectively they matched the subjective perception of human experts.
4. **Outputs**. Propose a taxonomy for rural roads based on these outputs, and then propose a methodology by which this could readily be reproduced on a more extensive network, taking stakeholder feedback into account.

2. Defining routes and sampling

2.1 Scope

Roads were considered in scope for this project if they satisfied the definitions outlined in section 1.1. In England and Wales, this included any road that did not pass through a built-up area, as defined by the DfT polygon layer. In Scotland, this included any road that did not pass through a Data Zone classified as a Town (conurbation of between 3,000 and 10,000 population) or Urban area (conurbation of over 10,000 population). Motorways were excluded, as were major junctions and slip roads.

2.2 Segments to routes

A bespoke national route network was used for this project, to ensure that the roads sampled were substantial stretches of road. Details of how this route network was constructed are outlined in appendix 7.3 (Technical appendix on road network construction).

A pool of routes was created from which a random sample was drawn. This included all routes that were within scope of the project and were at least 1 km in length. This predominantly excluded small residential streets and routes where road names changed over short distances.

A random sample of roads was selected, the size of which was dictated by the available budget for procuring data. This sample was stratified by road class, with the total sample lengths weighted by the relative rural traffic volume that these classes of roads carry nationally. According to the DfT's TRA0202,[3] this consists of approximately 50% A-road traffic, 25% B-road traffic and 25% traffic on unclassified roads.

The result was a sample of 434 routes covering 1,563 km of road, of which 784 km was A-roads (136 routes), 385 km was B-roads (76 routes) and 394 km was unclassified road (222 routes). Figure 2.1 shows a map of this sample.

Figure 2.1: Random sample of rural routes used in this study

Credits: Esri, Maxar, Earthstar Geographics and the GIS User Community

3 https://www.gov.uk/government/statistical-data-sets/road-traffic-statistics-tra

3. Initial phase – machine learning

3.1 Model design

The initial approach of this project was to cluster rural roads based on a variety of data covering the physical characteristics of the carriageway, the usage of the road and the local environment. This would classify all sampled roads as one of a fixed number of rural road types. A mixture of qualitative and quantitative analysis would then determine the commonalities in these rural road types, from which a taxonomy would be determined.

Clustering relies on the proximity of objects being clustered, with nearby objects belonging to the same cluster as each other. In this case, the measure of proximity should reflect the similarity of roads with respect to the data available. It would have been possible to cluster roads based on the data using a standard metric for distance between points in the input data. However, this distance metric would depend heavily on the choice of input variables and their distributions, as well as their relationship to each other. Consequently, the output from that approach would be vulnerable to unintentional bias from the input selection. Instead, this project sought to construct a more natural distance metric using *neural networks* and *deep autoencoders*.

Neural networks provide a way of approximating numerical functions up to a minimal degree of error. They combine linear transformations of the data with non-linear perturbations, giving them flexibility to approximate most functions while also being computable. Example data with known values is used to train the neural network, providing an approximation that minimises error for the data provided.

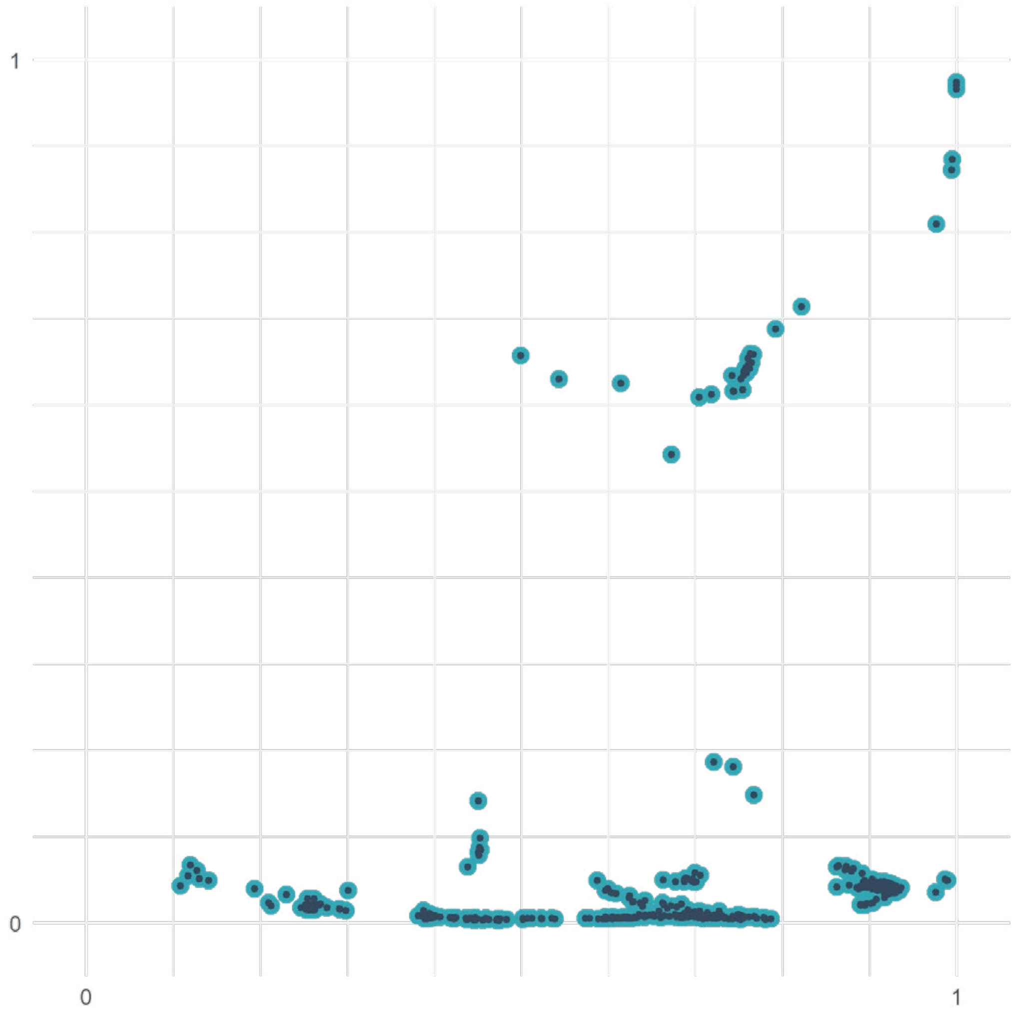
By combining two neural networks – with the first taking many variables as input and outputting only a few, and the second taking fewer variables and outputting more – one can create a new neural network with a *bottleneck* in the middle (see appendix 7.4, Figure 7.37). This combined neural network is trained to approximate a function whose outputs are the same as the inputs. The original two neural networks then have the following properties:

- The first neural network takes many variables and summarises them into fewer.
- The second neural network takes fewer summary variables and attempts to reconstruct the original variables.

On the data used to train the neural network, the error introduced by first summarising the data and then reconstructing it is minimal. Neural networks of this kind are often referred to as *deep autoencoders*. For these to be as efficient as possible, the summary variables produced in the bottleneck are similar for data points that are alike across the wide range of input variables but they are dissimilar for data points that differ noticeably. It is these summary variables that are used to cluster the sampled roads. More details on the use of deep autoencoders can be found in appendix 7.4.

The term *latent variable* is often used to describe the summary variables produced in the bottleneck of deep autoencoders. The set of possible latent variables is referred to as the *latent space*. In this project, a two-dimensional latent space was used, resulting in two latent variables which can be plotted on a graph. Figure 3.1 shows a representation of this space, with each point representing a rural road from the sample.

Figure 3.1: Autoencoder latent space



An *agglomerative clustering algorithm* was applied to data points in the latent space provided by the deep autoencoder. This algorithm combines data points into a number of clusters. A *gap statistic*, which measures the fit of clusters relative to the expected fit of random groups of points, was calculated and used to determine the optimal number of clusters in the data. This was determined to be seven superclusters, with the option of further subdividing into 11 clusters. More information on both the clustering algorithm and the gap statistic can be found in appendix 7.4, and details on the resulting clusters and superclusters can be found in section 3.4.

3.2 Model inputs

A series of model inputs were collected from a variety of sources to cover the properties of the roads, their usage and the environment in which they exist. It is important to note that collision data from STATS19 was not used as an input to determine this taxonomy of roads. Instead, STATS19 was used to determine whether substantial differences exist in the types of risk posed to road users on these different categories of road (see section 5.2).

3.2.1 Road properties

To encode the physical characteristics of carriageway, the following data points were used:

- The average and minimum width of the road;
- The average gradient of the road;
- The existence of carriageway separation;
- The sinuosity and straightness of the road; and
- The speed limit of the road.

Road class (e.g. motorway, A-road, B-road, unclassified) was not used, as this is mainly an artifact of British network history and not necessarily indicative of a road's characteristics. There are questions over the extent to which speed limits are intrinsic characteristics of the road, reflective of the safe speed for vehicles, and this will vary nationally. Nonetheless, speed limit is the closest widely available proxy for the safe speed of the road.

3.2.2 Road use

To capture the extent and nature of road-user activity, the following data points were provided to the model:

- Vehicle speeds from in-vehicle telematics:
  - Peak average vehicle speeds;
  - Off-peak average vehicle speeds;
  - Peak 85^{th} percentile speeds; and
  - Off-peak 85^{th} percentile speeds.
- The number of annual telematics counts, as a proxy for annual average daily flow.
- The travel times, in minutes, to the nearest town and urban area.

Although speed data is available at several other time periods, peak and off-peak time periods were chosen to represent the speed profile on the sampled roads as these would have the largest samples of probe data, giving more robust speed values. More information on this data is contained in section 3.3.2.

3.2.3 Environment

As characteristics of the area the road passes through, the following data points were used:

- The surrounding network density.
- The surrounding population density.
- The proportion of the surrounding area that is:
  - Arable land;
  - Broadleaf woodland;
  - Coniferous woodland;
  - Improved grassland;
  - Semi-natural grassland;
  - Built-up areas and gardens;
  - Mountain, heath, or bog; and
  - Coastal or water.

The land-use data was taken from a UK Centre for Ecology and Hydrology (UKCEH) dataset, details of which are contained in section 3.3.4. This data was included to provide insights into both the view from the road and the profile of hazards alongside the carriageway.

3.3 Input data acquisition and issues

The bespoke route network that was outlined in section 2.2, which is based on Ordnance Survey's Open Roads, was used as a unifying network in this project. Data from a variety of sources were matched spatially to this network.

3.3.1 Road width and gradient

Ordnance Survey's MasterMap Highways Network[4] was procured, along with the attribute data associated with it. This data was procured in individual kilometre OSGB grid squares, each of which was priced depending on the density of roads in the square. Despite the sample of rural roads running predominantly through areas of low network density, most routes either start or finish in areas with higher network density, resulting in the acquisition of some more expensive tiles of data. This was the main limitation on the size of the sample. The chosen sample of roads resulted in the procurement of eight high-density squares, 771 medium-density squares and 2,477 low-density squares.

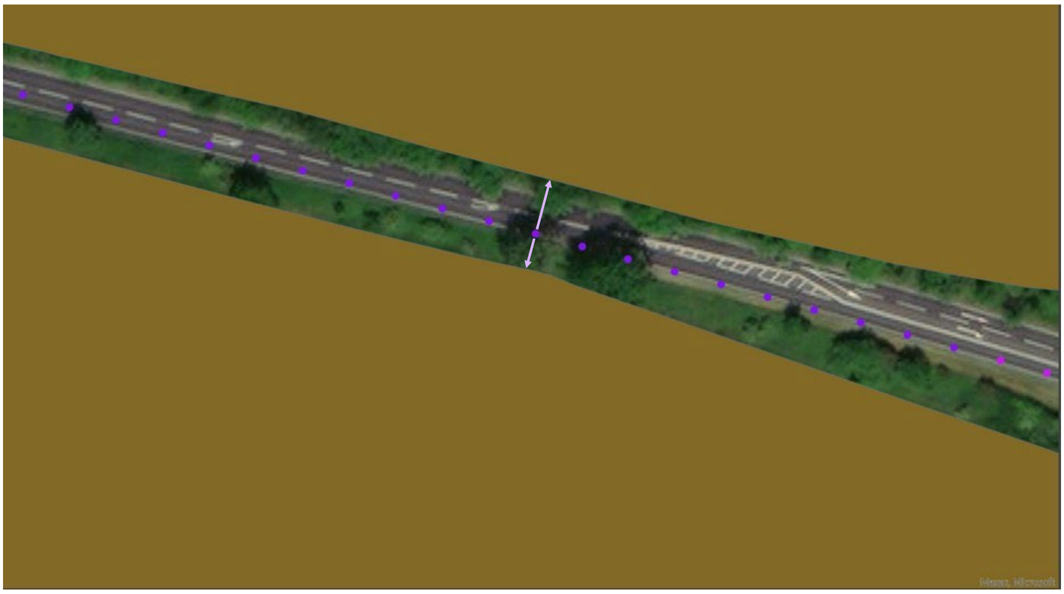
From this data, the average gradient along the road was calculated using road length and changes in elevation along the road. The form of way was also used to determine whether the roads had divided carriageways.

Road width, both average and minimal, was also obtained from this data. However, coverage of these metrics was limited. Of the 434 routes in the sample, only 216 had road-width data available. Alternative approaches were used to estimate road width for the remaining 218 routes.

An initial attempt at estimating road width was made using a combination of OpenStreetMap[5] land-use data and land-registry data[6] either side of the road to determine where the carriageway ended. At regular sample points along the road, road width was calculated as the total distance from the road centre line to the nearest spatial polygon along a perpendicular bearing from these two datasets on either side of the road. These values were then used as a basis for estimating minimum and average road widths, as described below. Figure 3.2 shows a map of the road extent, the sample points along the road centre line and the surrounding spatial polygons.

4 https://www.ordnancesurvey.co.uk/business-government/products/mastermap-highways
5 https://www.openstreetmap.org/
6 https://use-land-property-data.service.gov.uk/

Figure 3.2: Map of road-width estimation



Credits: Esri, Maxar, Earthstar Geographics and the GIS User Community

Distances were measured to a maximum of 100 m either side of the road. Any pairs of measurements for which one side of the centre line was greater than 5 m wider than the other side, which would imply a centre-line displacement of over 2.5 m, were removed as anomalous. Measurements were also considered anomalous, and therefore removed, if they were at least one standard deviation away from the average along the route.

Figure 3.3 shows that, for some roads, very few of these estimates were available for sample points along the road. Points are coloured white where estimated road widths could not be calculated due to gaps in OpenStreetMap and the land-registry data. It is likely that better coverage would be available from using Ordnance Survey's MasterMap Topography Layer[7] for polygons surrounding road extents, although this would require additional expenditure on data procurement which lies outside the scope of the current project. Average and minimum road widths were only calculated on roads for which over 10% of the sample points achieved width measurements.

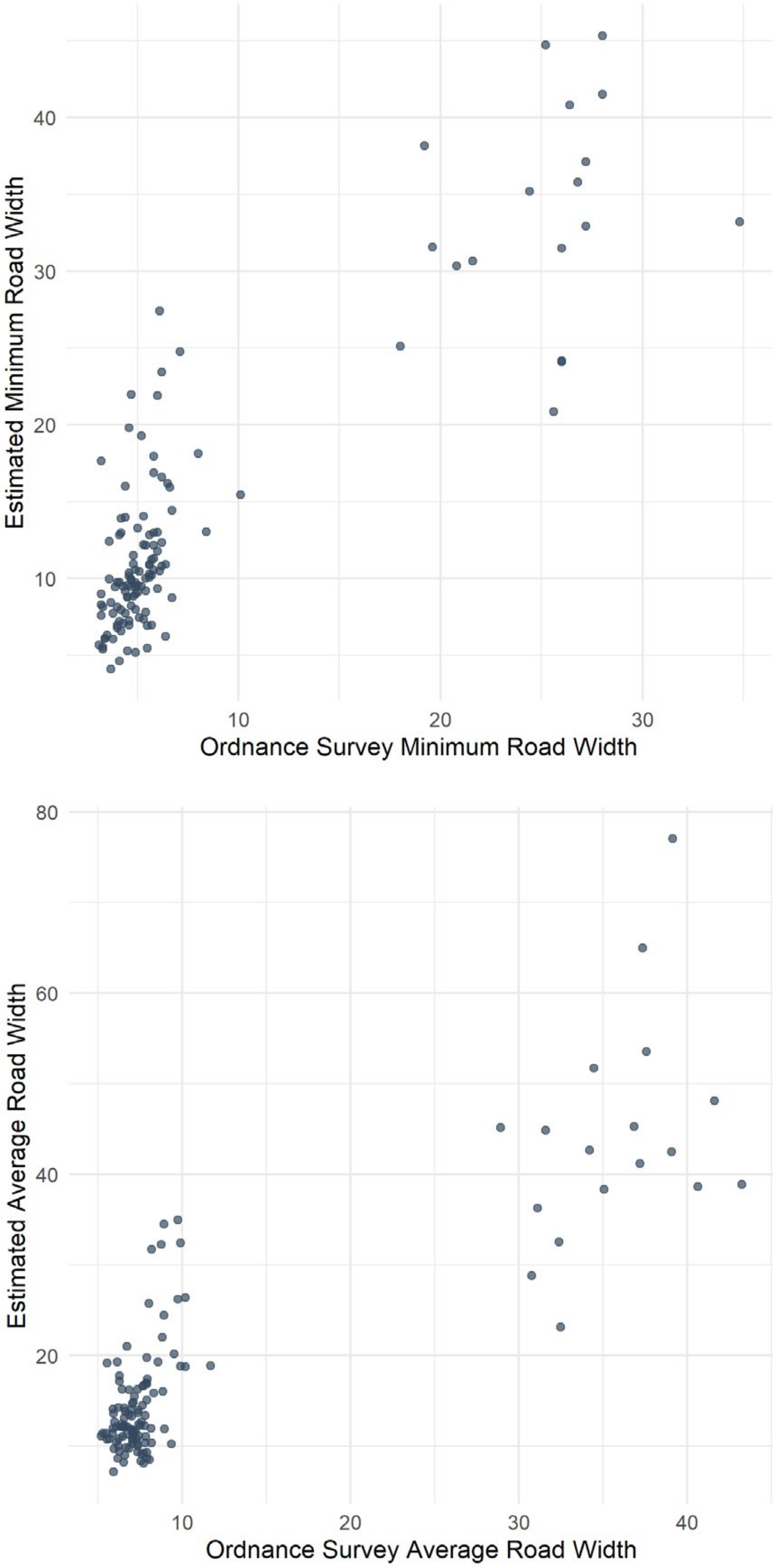
7 https://www.ordnancesurvey.co.uk/business-government/products/mastermap-topography

Figure 3.3: Map of coverage of estimated road width

Credits: Esri, Maxar, Earthstar Geographics and the GIS User Community

This measurement is far from a perfect measure of road width, as it consistently overestimates the extent of the road. This is evident from Figure 3.2. However, the estimated measure did correlate reasonably well with the Ordnance Survey road width where there was overlap, with Pearson's correlation coefficients of 0.87 and 0.86, and R-squared values of 0.75 and 0.74 for average and minimum widths, respectively. Figure 3.4 shows a comparison of road widths from the two sources. Using the sample of roads for which both Ordnance Survey road width and this estimated road width was available, a pair of linear adjustment models were fit and used to rescale the measurement, so the resulting road-width values align more satisfactorily to the Ordnance Survey average and minimum road widths.

Figure 3.4: Comparison of Ordnance Survey road widths with estimated widths



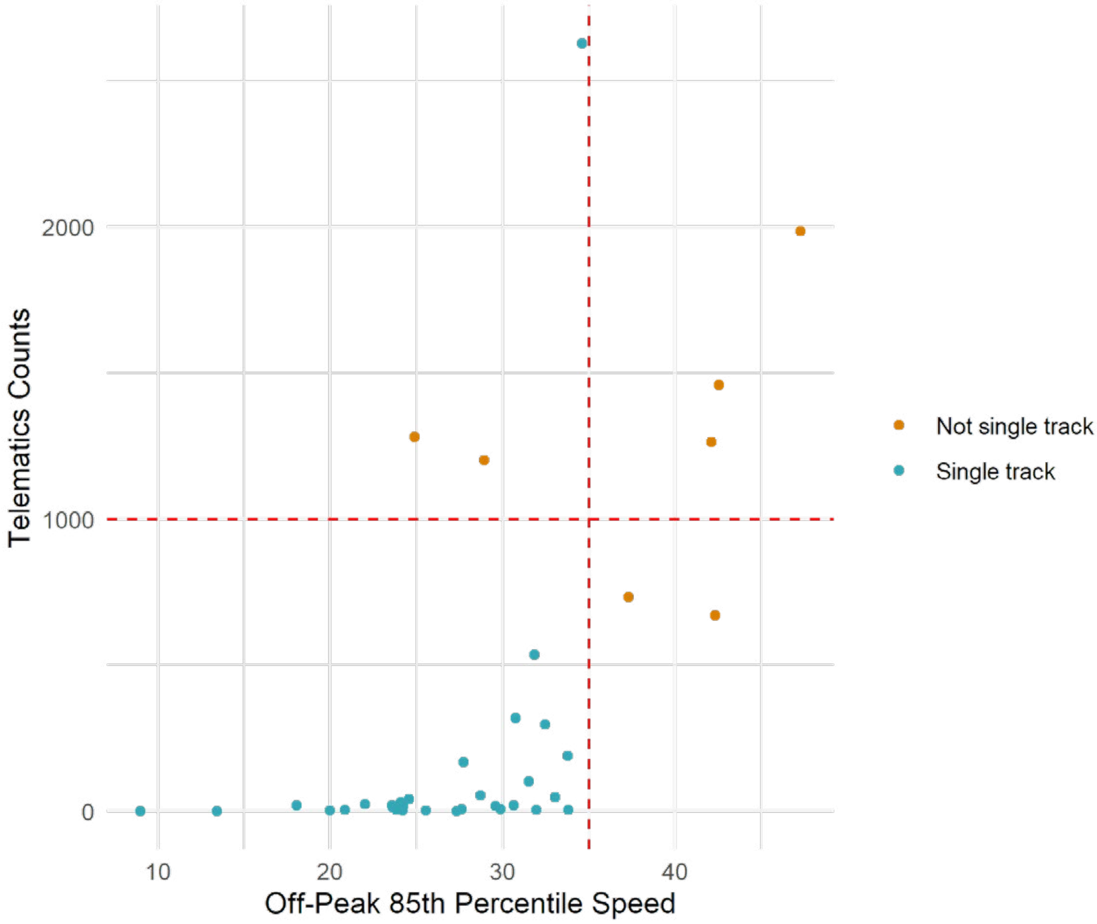
It is important to note that it was not possible to use this method of estimating road width for all roads. Due to limitations in the coverage of OpenStreetMap and land-registry data, road widths for only 308 of the 434 sample roads could be obtained. This provided estimated minimum and average road widths for 181 of the 218 routes for which Ordnance Survey did not have road widths, leaving 37 routes with no available road width data at all.

So that no data was missing when the deep autoencoder was trained and the clustering algorithm was carried out, a nominal road width was assigned to the remaining 37 routes. The intention of this was to distinguish single-track roads with passing spaces from wider roads where traffic can pass freely. For this, telematics-based speed and flow data was used, about which more information is provided in section 3.3.2. Single-track roads are more likely to have suppressed high-end speeds and so a threshold was set for off-peak 85th percentile speeds of 35 mph, above which roads are unlikely to be single track. To distinguish the effect on speed suppression of congestion from that of having a narrow carriageway, a threshold of traffic flow was also set. Roads having more than 1,000 vehicle telematics counts across the year (broadly equivalent to 800 daily traffic counts when compared to manual surveys) are also unlikely to be single track.

As a result, roads with off-peak 85th percentile speeds below 35 mph and with fewer than 1,000 annual vehicle telematics counts were given a narrow road width of 2.5 m, both average and minimum. The remaining roads were assigned the average values from the known average and minimum road widths, of 6.9 m and 4.3 m, respectively.

The remaining 37 routes were manually inspected and those that were single-track roads with passing places were labelled. Figure 3.5 shows a comparison of off-peak 85th percentile speeds and telematics counts for these routes. This demonstrates that only one single-track road does not fit within the thresholds set out above.

Figure 3.5: Off-peak 85*th* percentile speeds by telematics counts for roads with no width data



3.3.2 Traffic speed and flow

Telematics-based speed data was obtained from Ordnance Survey,[8] attached to the MasterMap Highways network. As well as average speeds by time of day, high-end (85th percentile) speeds were also obtained. Telematics sample counts were also used as a proxy for traffic flow. These telematics counts correlate well when compared against DfT's count point data. Each telematics count corresponds to approximately 0.8 average daily vehicles.

The telematics speed data is available in six different time periods:

- AM peak (7:00 to 10:00, weekdays).
- Off-peak (10:00 to 16:00, weekdays).
- PM peak (16:00 to 19:00, weekdays).
- Evening (19:00 to 00:00, every day).
- Weekend (7:00 to 19:00, weekends).
- Night (00:00 to 4:00, every day).

8 https://www.ordnancesurvey.co.uk/business-government/products/mastermap-highways-speed-data

Peak speeds, taken as the average of the AM peak and PM peak, and off-peak speeds were selected to represent the speed distributions at the sampled roads. Although all time periods could have been used, low-flow rural roads usually have fewer probe data points from which averages and 85th percentiles can be drawn, which reduces the reliability of this data during evenings, nights and weekends.

Although average speeds were available for all roads in the sample, some roads had too few probe samples to calculate robust 85th percentile speeds. Of the 434 routes in the sample, 88 did not have peak 85th percentile speed data available and 53 did not have off-peak 85th percentile speeds. In these cases, missing values were set as the average for all roads in the sample with the same speed limit and with the same type of carriageway separation.

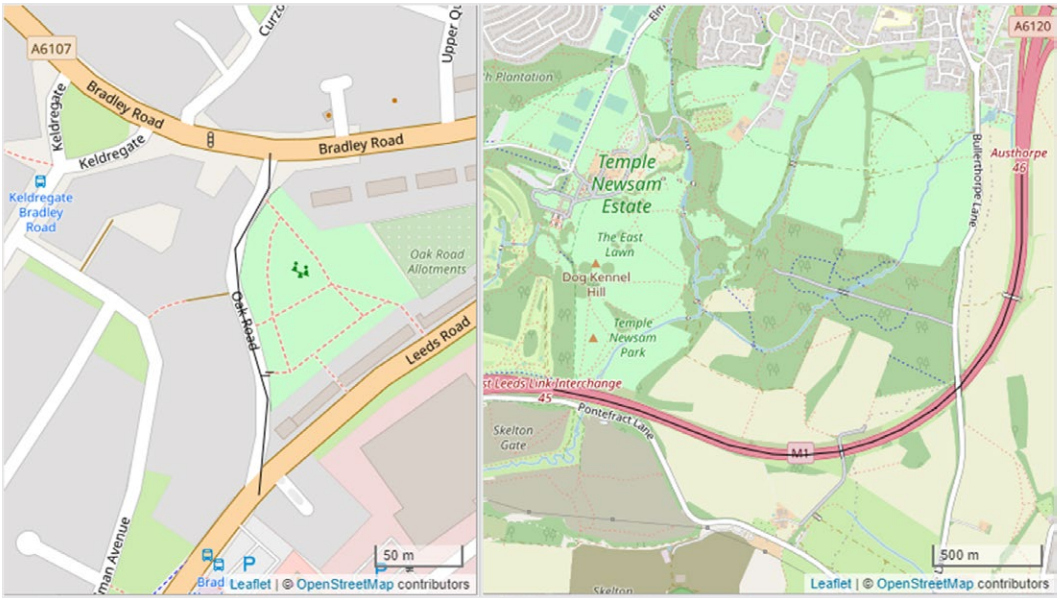
3.3.3 Network geometry

Two measures were used to assess the frequency, severity and variation of curvature along the routes. Firstly, sinuosity, as defined in Bovet & Benhamou's (1988) paper,[9] measures the variation in turning angles along the route. This gives a measure of how often a route changes direction and how severe these changes are. Secondly, straightness measures the relative difference between the travelled distance along the road and the straight-line distance between the endpoints. This gives a measure of the presence of curvature, regardless of the turning angles involved.

Although sinuosity and the lack of straightness are often correlated, it is possible to have sinuous roads that are mostly straight, and roads that are neither straight nor sinuous. Oak Road, on the left of Figure 3.6, is straight as the straight-line distance between endpoints is similar to the driven length of the road but it is also sinuous as it repeatedly bends in different directions. However, the section of the M1 along the south-east of Leeds, as shown on the right of Figure 3.6, is both not straight and not very sinuous, as there is little variation in the turning angle along the road.

9 Bovet, P. & Benhamou, S. (1988). *Spatial analysis of animals' movements using a correlated random walk model.* Journal of Theoretical Biology, 131(4): 419–433.

Figure 3.6: Straightness and sinuosity road samples



3.3.4 Land use

As a proxy for roadside topography, a land-use classification dataset was used to determine the roadside environment. This was taken from the UKCEH Land Cover[10] map. This is a vector spatial layer, segmenting the UK into 21 land-use classes, which have been summarised into the following eight classes for this project:

- Arable land.
- Broadleaf woodland.
- Coniferous woodland.
- Improved grassland.
- Semi-natural grassland.
- Built-up areas and gardens.
- Mountain, heath or bog.
- Coastal or water.

Each route was buffered by 25 m either side of the road, and the proportion of this surrounding area that was in each of the above classes was calculated as a separate variable.

The UKCEH Land Cover data was used under an Innovation Licence for determining the feasibility of creating a rural road taxonomy and to explore the role land-use data could play in establishing this taxonomy. The authors gratefully acknowledge the assistance and support from UKCEH, which made this possible. Wider development of this taxonomy into a usable dataset would likely require the establishment of an alternative licence arrangement.

Alternative land-use datasets may also be available from the European Space Agency[11] (ESA) using Sentinel-1 and Sentinel-2 imagery, from Ordnance Survey's MasterMap Topography Layer,[12] or from OpenStreetMap.[13] These sources vary in cost, granularity and accuracy.

10 https://www.ceh.ac.uk/data/ukceh-land-cover-maps
11 https://esa-worldcover.org/
12 https://www.ordnancesurvey.co.uk/business-government/products/mastermap-topography
13 https://osmlanduse.org/

3.3.5 Local environment

To capture the local environment around the road, both network density and population density were calculated. A 200-m buffer was taken around the sample routes, within which the total road length and total population were calculated, per square kilometre. Road length was calculated spatially from Ordnance Survey's MasterMap Highways network. Population data was taken for 2020 from WorldPop Constrained Population Counts.[14] These population counts were modelled to a 100-m resolution using a combination of census data, covariate data and building footprints.

3.3.6 Remoteness

Travel times to the nearest urban areas or towns were included as two separate measures of remoteness. For this purpose, urban areas were defined to be Low Level Super Output Areas (LSOAs) or Scottish Data Zones that form part of conurbations of over 10,000 people, while towns were those in settlements of over 3,000 people, including urban areas. An open-source routing engine, OSRM,[15] was used to determine the shortest travel time from any point along the route to any urban area or town, in minutes.

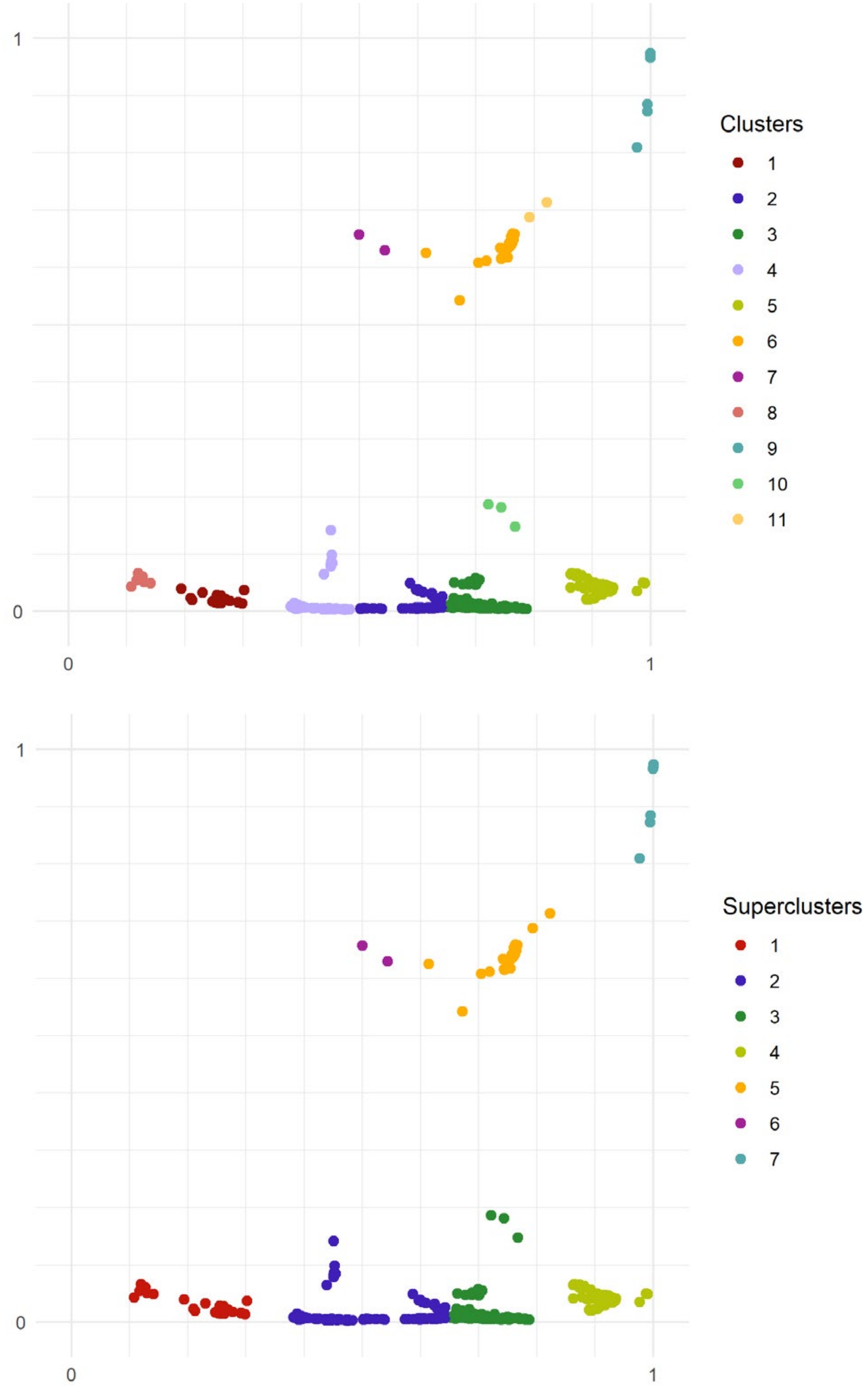
3.4 Cluster outputs

3.4.1 Clustering in latent space

As described in section 3.1, clustering on the latent space resulting from the autoencoder gave seven superclusters, which can be broken down further into 11 individual clusters. Figure 3.7 shows the distributions of these clusters and superclusters in the latent space.

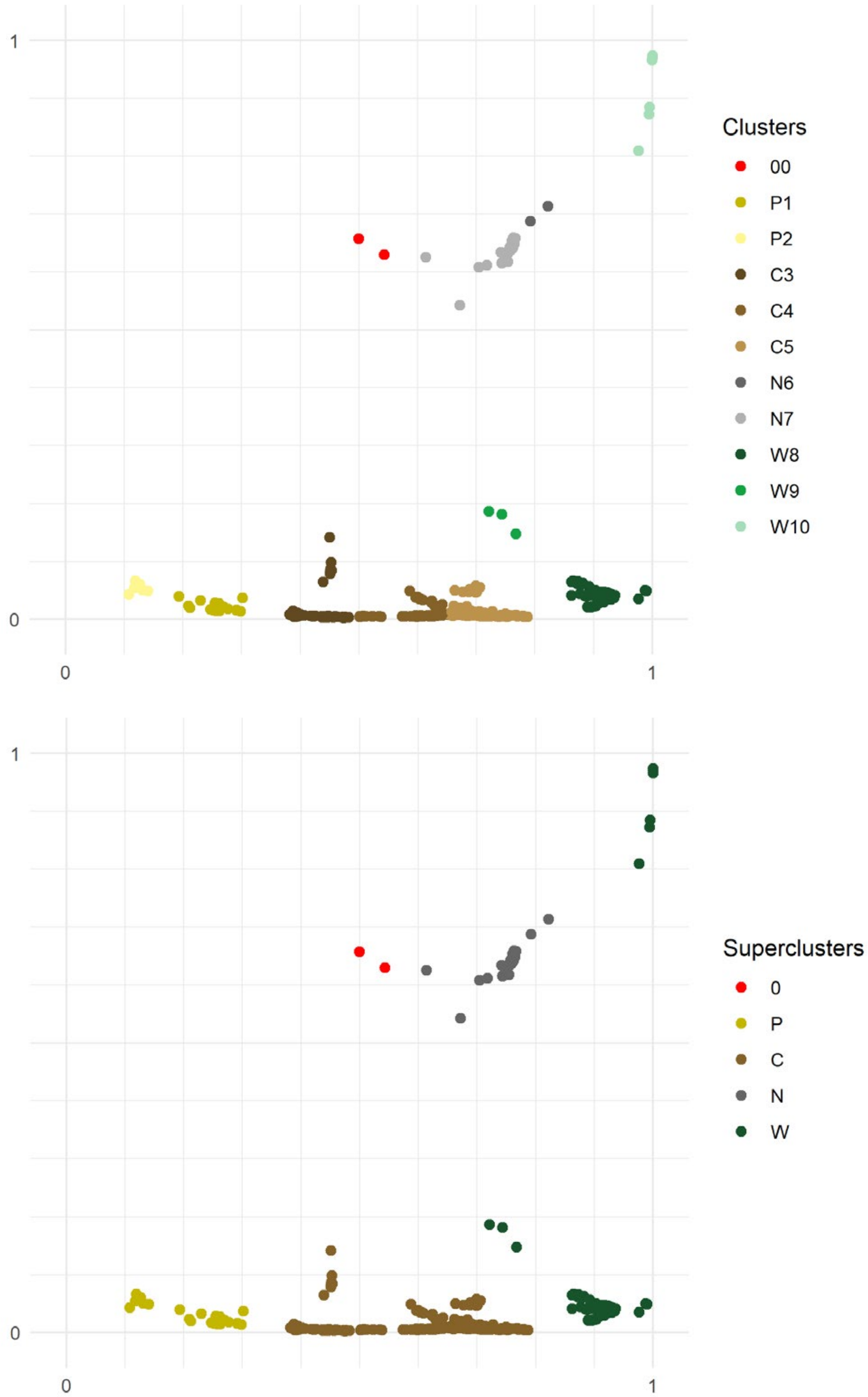
14 https://hub.worldpop.org/geodata/listing?id=78
15 https://project-osrm.org/

Figure 3.7: Clusters and superclusters in the latent space



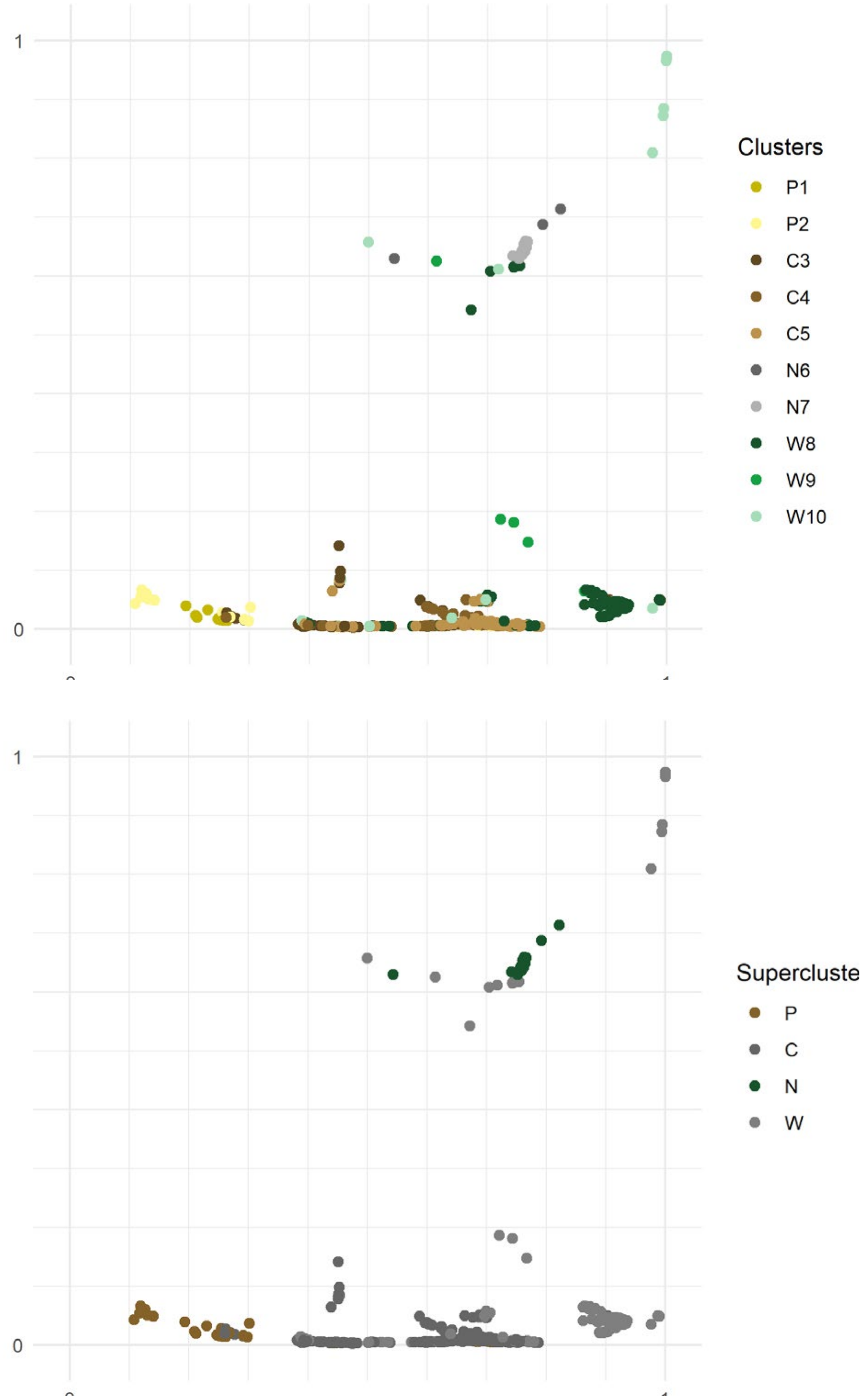
Following a review of the individual clusters and how they combine to form superclusters, the definitions of superclusters were adjusted to form more intuitively cohesive groups. This involved removing supercluster 1, as it contained only two sections of road, absorbing supercluster 7 into supercluster 4, and redistributing supercluster 3 between superclusters 2 and 4. This was carried out based purely on the analysis of measures in section 3.4.2, without looking at the effects in the latent space. Figure 3.8 shows the result of this manual adjustment in the latent space, with clusters and superclusters renamed to better illustrate the combinations of clusters into superclusters.

Figure 3.8: Clusters and superclusters in the latent space following manual adjustment



Following the manual qualitative analysis detailed in section 4, some clusters were manually reassigned to the clusters and superclusters to which they appeared better suited. This reassignment of the latent space is discussed in more detail in section 4.1.2, but it is shown here in Figure 3.9 for comparison.

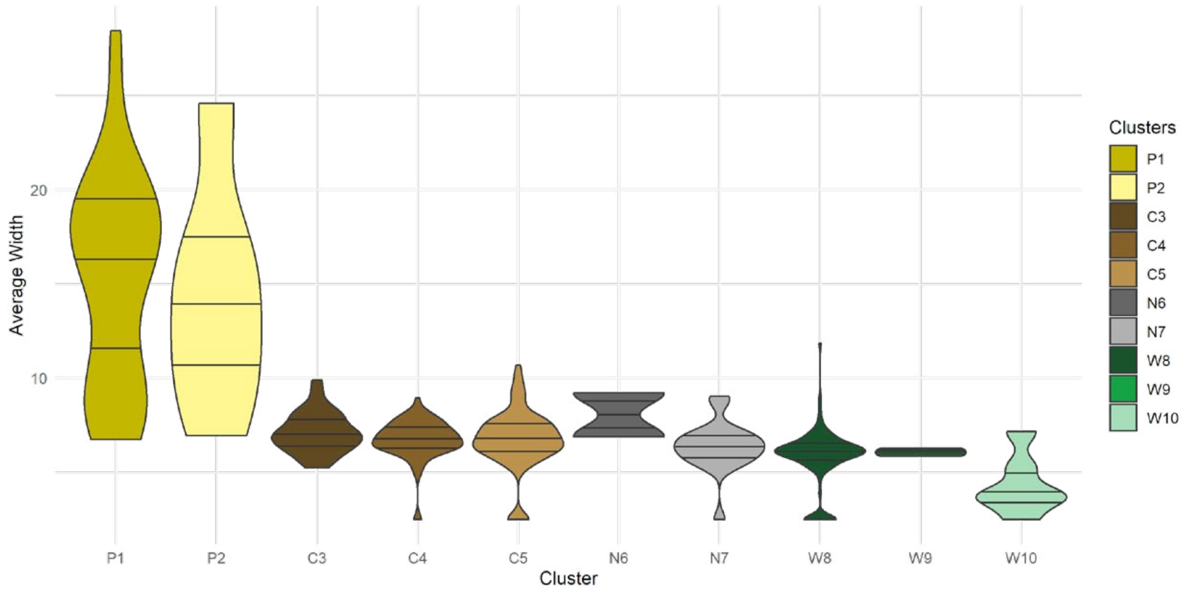
Figure 3.9: Clusters and superclusters in the latent space following manual reassignment



3.4.2 Individual measure outputs

Figure 3.10 shows violin plots of the distributions of average widths across the rural road clusters. One plot for each data input used in the clustering algorithm can be found in appendix 7.1 (Figure 7.1 to Figure 7.24). These vertical histograms highlight the similarities and differences across the clusters. For example, Figure 3.10 shows that roads in clusters P1 and P2 tend to be very wide, while W10 contains roads that are notably narrow.

Figure 3.10: Average width distributions for rural road clusters



Each cluster was analysed in the context of the other clusters using these distributions, and the characteristics that were diagnostic of each cluster were determined. From these, brief text descriptions were written to summarise the unique characteristics of each cluster. These can be found in section 5.1 at the levels of both clusters and superclusters. These provide generalised descriptions of the kinds of roads that appear in each cluster.

It was found that some features were not necessarily diagnostic for any clusters. This might be because there were no commonalities in this data between roads in the same cluster, or because it provides no additional support in discerning different clusters because of correlation with another feature. The profiles of peak average speed and off-peak average speed in Figure 3.11 and Figure 3.12 are similar, for example, and therefore off-peak average speed provides no additional diagnostic information once peak average speeds have been considered. A complete set of similar charts for all input variables is presented in appendix 7.1.

Figure 3.11: Peak average speed distributions for rural road clusters

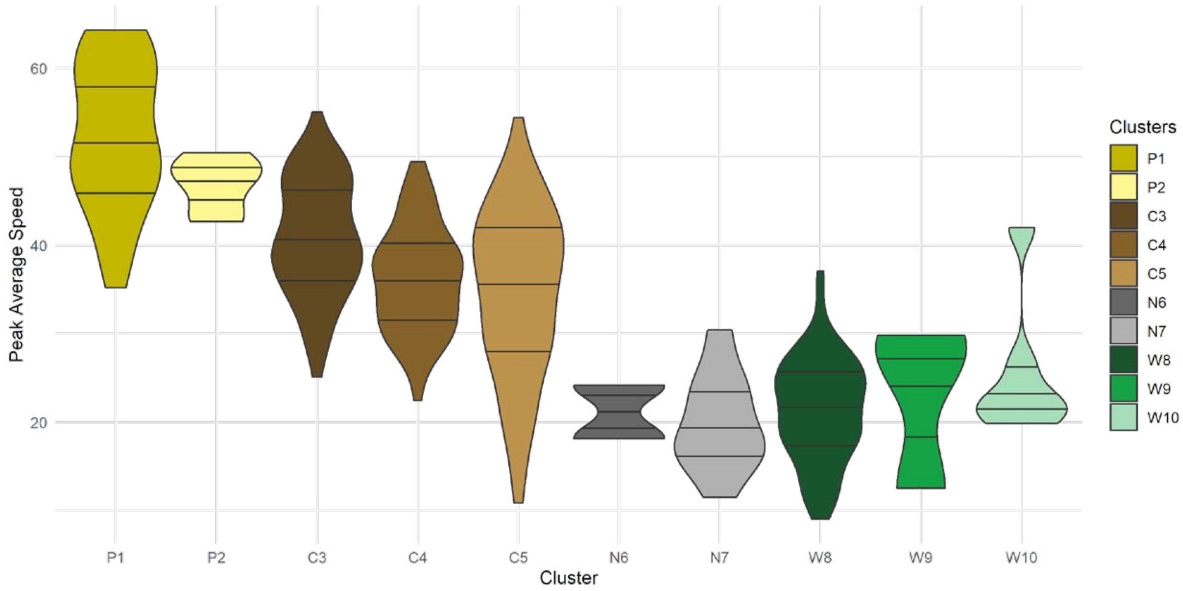
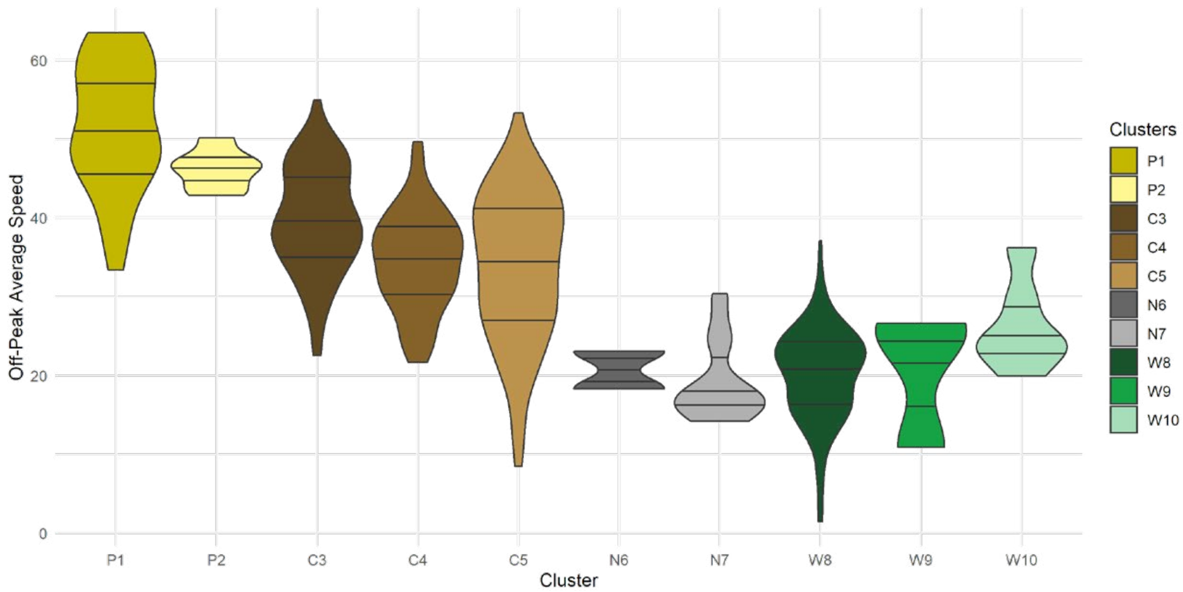


Figure 3.12: Off-peak average speed distributions for rural road clusters



4. Qualitative analysis

4.1 Testing descriptions compared to samples

Having obtained the cluster outputs generated in the latent space, as described in section 3.4, the machine-learning outputs were subjected to qualitative analysis to verify the allocation of routes to each cluster based on their attributes and physical characteristics. This was undertaken by examining the description of each cluster, reviewing the properties of each route within a cluster and subjectively assessing its match to the cluster description. Thereafter, routes were viewed using street-level imagery to consider their physical appearance relative to other routes within the same cluster. To achieve maximum coverage, Google Streetview, Mapillary and Geograph were used as image sources during this review process. All images reproduced in this report are taken from public domain sources.

This analysis identified not only how well each route matched the cluster description but also the cohesiveness of the cluster, i.e. whether there were stronger similarities or greater variability between route characteristics.

Reviewing the properties of each route also enabled verification of those attributes that could be seen emerging in the violin plots mentioned in section 3.4.2 as key defining variables of a cluster, while also confirming which variables offered minimal contributions in determining cluster assignment.

4.1.1 Similarities that were not input metrics

In some instances, viewing the physical characteristics of a route revealed similarities that were not captured in the cluster description because they were not input variables used by the machine-learning model. Variables including the use of routes by non-motorised vulnerable road users (VRUs) and the presence and amount of highway features – such as vehicle restraint systems (VRS), traffic signs and road markings – were noted for consideration in future applications of a taxonomy. However, their use as input metrics relies on the existence of robust and comprehensive datasets with national coverage, which at present are not available. Local data may be available from some highway authorities, but the use of data that is inconsistently available would not enable development of a uniform, and therefore comparable, taxonomy that can be applied across the country.

Although the presence of VRUs and highway features were not available as direct inputs for this research, the model may have been influenced to some extent by implicit correlation between them and other variables. For example, higher levels of VRU traffic are implied by higher population densities. If this taxonomy is remodelled in future, there may be additional opportunities to add other nationally available data sources to enhance this effect; for example, bus route maps as a proxy for the presence of pedestrians. However, further investigation would be required to assess the robustness of this approach.

4.1.2 Refinement

As a result of the qualitative analysis, some routes were reassigned to different clusters based on their properties and appearance, to ensure best fit and maximum cohesiveness within clusters. While some variation was inevitable at cluster level, the qualitative analysis confirmed greater cohesiveness at the supercluster level. Having adjusted the numbers of superclusters and clusters following initial review of the latent space outputs, illustrated by Figure 3.9, the seven superclusters were further refined to four superclusters, and one cluster containing just two sample routes with little cohesiveness was removed and those routes were reassigned.

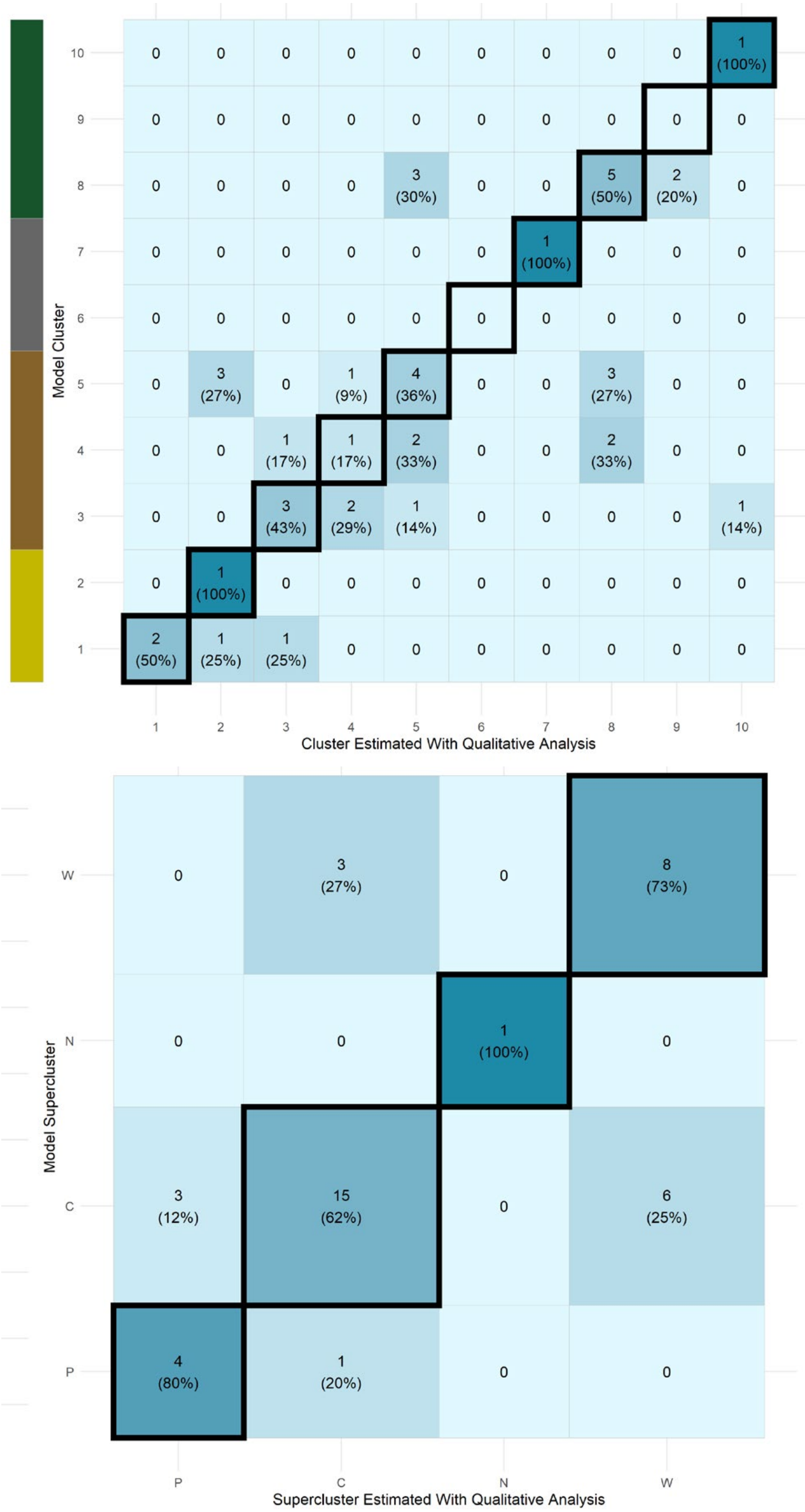
Refinement of the supercluster and cluster taxonomy also enabled refinement of the descriptive names assigned to each, to provide a clearer and more concise understanding of the routes in each category. The outcome of this reassignment is shown in Figure 3.9.

4.2 Validation

4.2.1 Blind sample testing

A small number of routes from across the whole sample were held back from the qualitative analysis stage and used as a blind sample to test the machine-learning outputs further. Again, the route properties and physical appearance were reviewed, and each route within the blind sample was subjectively assigned to a supercluster and cluster. This was then checked against the machine-learning outputs. As shown in Figure 4.1, the results of this were process were generally positive, demonstrating high correlation between the AI model output and blinded qualitative analysis assignments.

Figure 4.1: Blind testing results



4.2.2 Difference between supercluster and cluster levels

The results in Figure 4.1 do, however, show that there is greater variability between the machine-learning and qualitative analysis assignment of routes at cluster level than for superclusters. This indicates that a degree of overlap exists in route characteristics between clusters, which can present challenges when categorising a route. This is particularly evident between clusters C5 and W8, where correlations between the machine-learning and qualitative analysis outputs are weakest. Exploring and understanding these similarities and differences is necessary to further improve the taxonomy.

Furthermore, the absence of sample routes from supercluster 6 within the blind sample highlights a limitation of the overall sample in providing an equal distribution of routes across all clusters. This is a consequence of budgetary limitations on sample size, as well as the methodology described in section 2.2.

5. Outputs

5.1 Proposed taxonomy

The resulting output from the machine-learning and qualitative analysis stages reveals a two-level classification model of superclusters and clusters. Table 5.1 outlines the supercluster level of the taxonomy, detailing four superclusters with a descriptive summary of each, with representative but non-exhaustive references to the key variables used in determining each supercluster.

Table 5.1: Supercluster taxonomy

Reference	Name	Description
Supercluster P	Principal Roads	These roads are wide and generally also straight and flat. They have high speed limits and fast-moving traffic, so are unlikely to be used extensively by VRUs. These roads usually run through populated areas close to towns.
Supercluster C	Country Roads	These are narrower single-carriageway roads that are not particularly winding or steep. They are moderately trafficked and often run through improved grassland.
Supercluster N	Neighbourhood Roads	These are mostly 30 mph limit, single-carriageway roads with moderate traffic speeds, which are in or very close to towns. They run through areas of high population density and are usually in built-up areas.
Supercluster W	Winding Roads	Narrow single carriageways, unclassified and sometimes single track. Mostly 60 mph limits, but low speeds and little traffic. They are sometimes quite close to towns, where VRUs are likely to use the carriageway.

5.1.1 Principal Roads

The first of the superclusters is P, which contains two clusters termed 'Regional Distributors' and 'Main Roads'. These clusters are described in Table 5.2. In the descriptions used in Table 5.2, the term 'VRU' specifically refers to non-motorised vulnerable road users, which for the purposes of this report constitute pedestrians, cyclists and horse-riders.

Table 5.2: Supercluster P

Reference	Name	Description
Cluster P1	Regional Distributors	These roads have speed limits of 60 mph or 70 mph, are often dual carriageways, and exhibit high average and 85th percentile speeds. They have fairly high traffic flows and are very unlikely to be used extensively by VRUs. They are usually close to towns and quite close to cities; surrounding areas have high network and population densities. The surrounding land is most likely to be arable or broadleaf woodland.
Cluster P2	Main Roads	These roads mostly have speed limits of 60 mph with some carriageway separation, and moderate average and 85th percentile speeds. They have moderate traffic flows and are unlikely to be used extensively by VRUs. They are usually close to a town but can be distant from urban centres, sometimes going through areas of low population density.

Figure 5.1 illustrates a typical Regional Distributor road, for which the characteristics include a central median between opposing flows of traffic, a high speed limit and operating speeds, together with high traffic volumes. While often hidden from view by roadside vegetative screening, these roads are close to areas of high network and population density.

Figure 5.1: Regional Distributor example

Figure 5.2: Main Road example

The second cluster, Main Roads, are often close to towns but may also pass through areas of low population density. They too have high operating speeds and speed limits, and are generally relatively wide compared to routes in other superclusters. However, unlike Regional Distributors, they have no physical separation between opposing flows; instead, traffic may be separated by a painted central median, as illustrated in Figure 5.2.

5.1.2 Country Roads

Supercluster C has three clusters, as outlined in Table 5.3. There are many similarities between these three clusters: all routes are single carriageway of average straightness, typically carry above-average flows and traverse areas of wide-ranging gradient. An example of the first cluster – Primary Routes – is shown in Figure 5.3 and includes roads that can be distant from urban areas (up to 60 minutes' travel away) and those that are generally subject to higher speed limits, but where road width is visibly narrower than for Supercluster P.

Table 5.3: Supercluster C

Reference	Name	Description
Cluster C3	Primary Routes	These are often classified roads, of average straightness and mostly have 60 mph limits. They are often quite far from towns and sometimes very distant from urban areas. They run through areas of low network and population densities. They usually have highway features such as road markings, signs and roadside kerbs.
Cluster C4	Secondary Routes	These roads are straighter than average, with speed limits ranging from 30 mph to 60 mph. They are often close to towns and not too far from an urban area. They traverse areas of variable network density, but generally those with above-average population density. They sometimes have defined highway features such as centre lines and edge of carriageway lines.
Cluster C5	Tertiary Routes	These are narrower single carriageways, including some single-track roads. They are not winding and have average straightness but are sometimes steep. They are mostly 60 mph roads, although some have 30 mph or 50 mph limits, and traffic flow varies widely. They are generally quite close to towns or cities, and surrounding areas vary greatly in network and population densities. The local land use is usually arable land. Road markings and other highway features are rare, and VRUs are likely to use the carriageway.

Figure 5.3: Primary Route example

There are subtle variations in the characteristics between cluster C3 (Primary Routes) and cluster C4 (Secondary Routes, see Figure 5.4), with the latter being subject to different speed limits between 30 mph and 60 mph (national speed limit), having fewer highway features such as vehicle restraint systems or kerbed carriageway edges, and often being located closer to towns.

Figure 5.4: Secondary Route example

The third cluster within Supercluster C is Tertiary Routes. These routes display further variation in characteristics, with a narrowing of the road width and a reduction in highway features, often with loss of the carriageway centre line. These routes remain close to town or urban areas, but local population density can vary and, as a result, traffic flows are also more variable.

As might be expected on routes such as that shown in Figure 5.5, vehicle drivers and riders are more likely to share these routes with non-motorised road users.

Figure 5.5: Tertiary Route example

5.1.3 Neighbourhood Roads

The Neighbourhood Roads supercluster contains two clusters covering 'rural' routes that pass through towns and built-up areas (see Table 5.4). While adjacent land use may vary between each cluster, with residential land use surrounding Residential Distributors and Village High Streets typically comprising mixed-use (residential/commercial/education) development, routes in both clusters are generally subject to a maximum 30 mph speed limit and have dedicated roadside facilities for non-motorised VRUs.

Table 5.4: Supercluster N

Reference	Name	Description
Cluster N6	Residential Distributors	These mostly unclassified roads are of generally average width, with some bottlenecks. They are fairly straight and not steep. They have moderate average and 85th percentile speeds, and traffic levels are around average. They pass through towns and are very close to urban areas, so population and network densities are very high, and the area is very likely to be built-up. These roads usually have roadside facilities for VRUs.
Cluster N7	Village High Streets	These roads are narrow, more sinuous and less straight than average. They have low average and 85th percentile speeds and average traffic. They are generally close to towns but far from urban areas. Local population density is high and surrounding network density is average. The surrounding area is likely to contain built-up land or grassland.

An example of a Residential Distributor is shown in Figure 5.6 and a Village High Street is shown in Figure 5.7.

Figure 5.6: Residential Distributor example

Figure 5.7: Village High Street example

5.1.4 Winding Roads

The fourth supercluster contains three clusters and includes rural routes that are the narrowest and most remote in nature (see Table 5.5). Country Lanes, such as the example shown in Figure 5.8, are often unclassified single-track roads with high speed limits but lower operating speeds. The gradient of the environment through which they traverse can be variable and associated with this there can be greater sinuosity in the route's path.

Table 5.5: Supercluster W

Reference	Name	Description
Cluster W8	Country Lanes	These roads are winding and not straight, with a variety of gradients. Speed limits are mostly 60 mph with some 30 mph, but these roads have fairly low average and 85th percentile speeds and very little traffic. They are often close to towns and not too far from an urban area. Surrounding areas have average network density but low population density, and are likely to be arable or grassland.
Cluster W9	Hill Passes	These unclassified roads are winding and steep, with low average and 85th percentile speeds and variable flows. They are generally quite close to towns or cities, but surrounding areas have low network and population densities. The local land use is usually broadleaf woodland.
Cluster W10	Remote Roads	These are very narrow single carriageways that are more sinuous and less straight than average but usually flat. They are mostly 60 mph roads, with low average and 85th percentile speeds and low traffic flow. They are very far away from towns and can be hours away from the nearest urban area. They go through areas of very low network and population densities. The surrounding land use is most likely to be grassland or water.

Figure 5.8: Country Lane example

The second cluster within this supercluster are routes termed 'Hill Passes' – a name that reflects routes that traverse the steepest of gradients in the environs in which they pass, as shown in Figure 5.9. These routes also tend to have the narrowest width of all rural roads.

Figure 5.9: Hill Pass example

The final category of rural routes in this taxonomy is cluster W10 (Remote Roads). In contrast to Hill Passes, these routes are generally flat, often adjacent to coastal waters, but share the remote nature of this supercluster in having long travel times from towns, and travel times of up to several hours to urban areas. An example of such a road is shown in Figure 5.10.

Figure 5.10: Remote Roads example

Additional photos of example routes from each cluster are available in appendix 7.2.

5.2 A case study – collision analysis

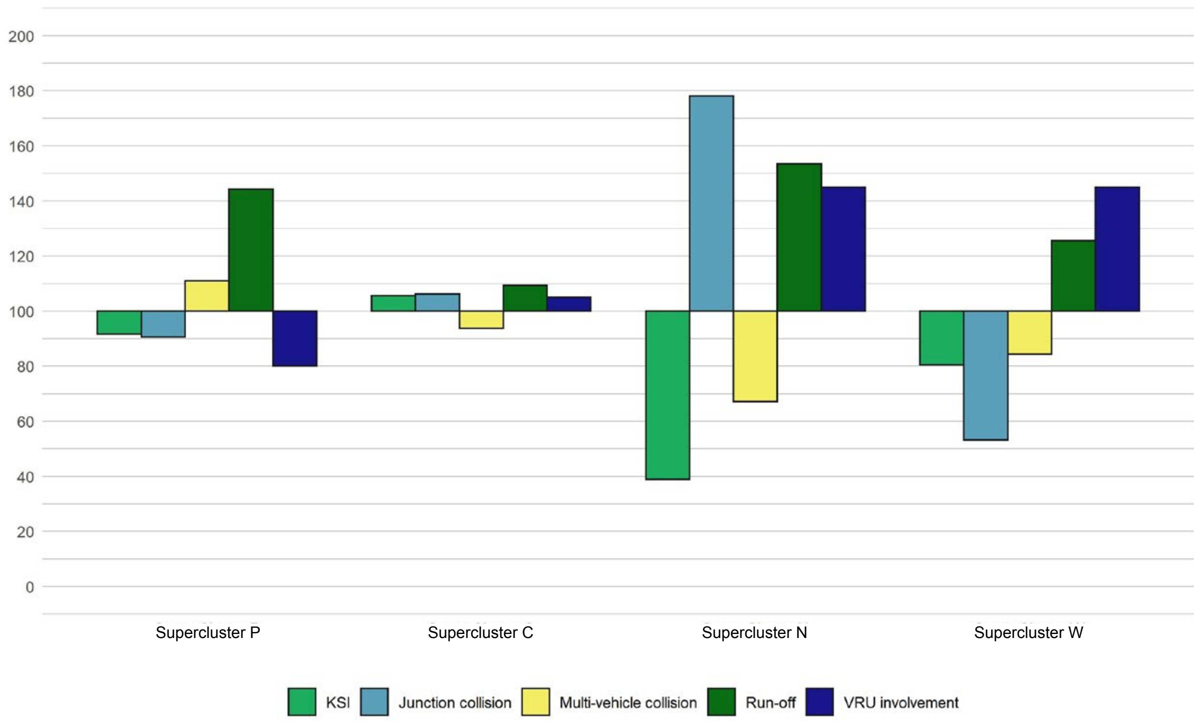
As referenced in section 3.2, STATS19 police-reported collision data was not used as an input variable. Instead, it was used to further test the taxonomy and determine any substantial differences in the type of risk posed to road users on different categories of road. Analysis was carried out on injury collisions reported across the sample roads during the five-year period of 2015–2019. The analysis also sought to determine if patterns might exist between collisions occurring on routes both within the same supercluster and between superclusters.

The variables considered in this analysis were concentrated on those that best characterise the individual nature of each collision; for example, vehicle involvement, vehicle location, VRU involvement and collision dynamics. The analysis did not look at variables that were likely to be reflected by input variables used by the machine-learning model, such as road type and speed limit, nor those that were not persistently present, such as weather conditions.

As mentioned earlier, the limited sample size has resulted in unequal distribution of routes across superclusters. This limitation is again reflected in the distribution of collisions in each supercluster, which ranged from only seven collisions in supercluster N up to 942 collisions in supercluster C. Therefore, the significance of the patterns among collisions in supercluster N is low.

Despite this, analysis of the collision data on routes within each supercluster against the overall sample revealed some clear patterns that are consistent with the relative properties of those superclusters. As shown in Figure 5.11, where supercluster variations are shown against a whole sample baseline index of 100, VRU involvement is under-represented compared with the whole sample index in supercluster P (Principal Roads), but over-represented in supercluster N (Neighbourhood Roads). Equally, the occurrence of collisions at junctions on Neighbourhood Roads is over-represented compared with the whole sample index, but under-represented on Winding Roads. These patterns are consistent with the relative route characteristics of less VRU presence on Principal Roads and high network density in the built-up areas of Neighbourhood Roads.

Figure 5.11: Collision analysis of superclusters against the whole sample



The collision patterns on routes within each supercluster were looked at in more detail and further trends have been identified.

5.2.1 Collision analysis results – Principal Roads

Table 5.6 shows the predominant outcomes of the collision analysis for supercluster P (Principal Roads). The analysis revealed that a 'typical' collision might involve multiple vehicles on a link between junctions, because of a slow or stopping manoeuvre by one or more of the vehicles. As a result, most such collisions result in vehicle impact from the rear. While collisions involving VRUs is unlikely, the VRU most likely to be injured is a motorcyclist.

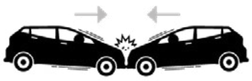
Table 5.6: Supercluster P collision patterns (corresponding percentages from whole sample shown in square brackets)

	72% not at a junction [68%]
	71% involved multiple vehicles [64%]
	53% resulted in rear impact [35%]
	30% involved a slow or stopping manoeuvre [19%]
	61% of single-vehicle collisions involved run-off to nearside [57%]
	16% involved a VRU [20%], of which 76% involved motorcycles [14%]

5.2.2 Collision analysis results – Country Roads

There are several similarities between collisions occurring on routes in supercluster C and those occurring on routes in supercluster P, as shown in Table 5.7. Collisions on Country Roads typically involved multiple vehicles and occurred away from a junction. However, more collisions on these roads involved head-on impact between vehicles and over a third of those in the sample occurred on a bend. The likelihood of a collision involving a VRU is slightly higher in supercluster C and, although still representing less than a quarter of the collision sample, they are again most likely to involve a motorcyclist.

Table 5.7: Supercluster C collision patterns (whole sample figures in brackets)

	66% not at a junction [68%]
	60% involved multiple vehicles [64%]
	31% resulted in head-on impact [25%]
	37% involved at least one vehicle negotiating a bend [29%]
	56% of single-vehicle collisions involved run-off to nearside [57%]
	21% involved a VRU [20%], of which 69% involved a motorcycle [14%]

5.2.3 Collision analysis results – Neighbourhood Roads

The routes in supercluster N had only seven reported injury collisions in the five-year analysis period and all occurred on roads within cluster N7 (Village High Streets). Therefore, it is difficult to ascertain any strong patterns in the collisions occurring in this group. However, based on what was analysed and as shown in Table 5.8, most collisions occurred at junctions involving multiple vehicles and nearly a third involved a VRU.

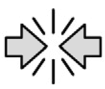
Table 5.8: Supercluster N collision patterns (whole sample figures in brackets)

	57% at a junction [32%]
	57% involved multiple vehicles [64%]
	29% involved a VRU [20%]

5.2.4 Collision analysis results – Winding Roads

As found with Neighbourhood Roads, not all clusters in supercluster W had routes on which collisions were reported during the five-year timeframe. The results shown in Table 5.9 are therefore only applicable to routes in cluster W8 (Country Lanes). Few collisions in this cluster occurred at a junction, but more collisions involved single vehicles, with over half of the collisions in the sample occurring on a bend. Where vehicles left the carriageway as a result of an incident, 80% collided with a roadside object such as a tree, and pedal cyclists are over-represented in this supercluster compared with the sample as a whole.

Table 5.9: Supercluster W collision patterns (whole sample figures in brackets)

	83% not at a junction [68%]
	54% involved multiple vehicles [64%]
	39% resulted in head-on impact [25%]
	52% involved at least one vehicle negotiating a bend [29%]
	54% of single-vehicle collisions involved run-off other [43%]
	80% of vehicles leaving the carriageway hit a roadside object [17%]
	29% involved a VRU [20%], of which 47% involved pedal cyclists [3%]

5.2.5 Comparison of collision trends between initial machine-learning clusters and refined clusters

Having identified that there are patterns in the collisions occurring on routes across superclusters, the impact of the qualitative analysis and resulting reassignment of routes between superclusters was examined to consider if this affected the collision analysis. Focusing on seven key collision characteristics, Table 5.10 outlines the differences in the percentages of each collision characteristic between the pure machine-learning output and the machine-learning output with qualitative refinement. The respective percentages for each collision variable for both the pure machine-learning outputs and refined outcomes are available in appendix 7.5.

As shown in Table 5.10, there was no change in the collision characteristics for supercluster N; this is most likely due to the small sample size.

Table 5.10: Results from collision analysis, comparing raw machine-learning clusters with those after qualitative process

Collision characteristic	Supercluster P	Supercluster C	Supercluster N	Supercluster W
% KSI	–2%	+1%	-	–2%
% Junction collisions	+10%	–3%	-	–5%
% Single vehicle collisions	–6%	+4%	-	–5%
% Multiple vehicle collisions	+6%	–4%	-	+5%
Predominant point of impact	–4%	+1%	-	+1%
% Run-off	–7%	+4%	-	–8%
% VRU involvement	–3%	+3%	-	+1%

5.2.6 Route risk

Having examined the typical collision types in which road users might be involved on any of the routes within each supercluster, the collective risk and individual risk were also calculated for each supercluster.

The collective risk to users of each supercluster is illustrated by the average collision density per year (collisions per kilometre). To put this into context, a collision density index has also been calculated. This compares the rates from each supercluster with roads across the entire sample. Index values are 100-based; therefore, as shown in Table 5.11, the annual average number of collisions per kilometre of route in supercluster P is nearly three times the whole sample average, while the collision density for Winding Roads is less than a quarter of that for the whole sample.

Table 5.11: Supercluster collective risk

	Collision density	Collision density index
Supercluster P	0.555	289
Supercluster C	0.190	99
Supercluster N	0.062	37
Supercluster W	0.028	15

The risk to individual users can also be determined using the collision rate, which takes traffic volume into account (collisions per million vehicle kilometres), and a collision rate index, which enables relative comparison between superclusters. Table 5.12 shows that the average collision rate per kilometre per year is less than one across all superclusters. However, the indexes reveal that the risk to individual users of routes in supercluster C of being involved in a collision is almost 50% higher than that of the whole sample, while individual risk in supercluster W is over twice that of all routes in the study.

Table 5.12: Supercluster individual risk

	Collision rate	**Collision rate index**
Supercluster P	0.064	60
Supercluster C	0.159	148
Supercluster N	0.116	109
Supercluster W	0.247	231

The sample size used in determining risk using the collision density and collision rates outlined above should be considered. Given the small number of collisions on routes in some superclusters, a measure of risk based on a larger and more balanced sample would offer a more robust measure of risk.

5.3 Validation with stakeholders

From the outset, the project team envisaged that any recategorisation of rural roads would require stakeholder engagement and input. Specifically, highway authorities with extensive rural networks, and government departments and agencies should scrutinise any proposed taxonomy and offer feedback.

To facilitate this process, a Development Day was held when relevant expert stakeholders were invited to examine and discuss progress to date and consider the most constructive ways forward. The following sections summarise the key points raised in response to this exercise.

5.3.1 Strategic engagement

There was a general feeling among stakeholders that moving away from a binary taxonomy of urban and rural roads would be more reflective of network reality, and could therefore promote better insight and decision-making. At a national level, a more detailed categorisation would enable presentation of national statistics in a more meaningful context, and ministerial briefings could be better informed. Work on policy and development of education, training and publicity interventions, such as “Think” campaigns, would also benefit from refinements in rural road categorisation. Unavoidably, by its very nature national government is more removed from reality on the ground than are highway authorities, so refinements in road taxonomy could be of particular benefit at national level.

For highway authorities, improvements could lead to more meaningful engagement with government; for instance, it could lead to enhanced conversations around prioritising routes for community recreation and wellbeing reasons. Also, gathering data to support funding applications would be easier and less subjective.

5.3.2 Insight into road risk and danger

In addition to the general policy benefits of improving the categorisation of rural roads, stakeholders identified specific benefits of analytical approaches to casualty data.

A network segmentation more closely related to the characteristics of roads will provide better insight into high harm routes and make it easier to identify comparable roads. This should facilitate mass action treatments, delivering intervention efficiently across many similar routes. Comparative methodologies could also be applied to roads in other authority areas, which would expedite evaluation of interventions and sharing of best practice.

Including the taxonomy in publication of national-level casualty data would validate and support such local initiatives. To make this possible, an important eventual goal would be to identify an open data route to applying and updating the taxonomy. The approach suggested in section 6 signposts a direction of travel towards this objective.

5.3.3 Other applications for highway authorities

Although this research was originally motivated primarily by a desire to improve insight into road risk, it became apparent during consultation that there were also potential collateral benefits for other aspects of highway authority activities. One frequently mentioned application was providing enhanced input into speed-limit review processes. Decisions about prioritising routes for active travel could also become easier.

More efficient planning of routine activities, such as maintenance and winter weather interventions, is another area where the taxonomy could have a positive impact. For example, it may provide improved understanding of danger for road workers. There is also potential for improved prioritisation and planning of new works. The taxonomy could also be used to support authority consideration of planning applications and HGV-routing decisions, both of which can have major implications for traffic on rural roads.

5.3.4 Stakeholder concerns

There was concern expressed about some aspects of the proposed taxonomy. Firstly, it seemed undesirable to stakeholders to introduce a detailed taxonomy for rural roads without also doing so for roads in urban areas, which carry a higher proportion of traffic. While taking this forward lies outside the scope of the present commission, the researchers feel that the general approach applied in this project would also lend itself to applications for urban roads and would welcome the opportunity to expand the principle of detailed taxonomies based on highway properties to urban roads.

Secondly, stakeholders also felt that the taxonomy required validation over a larger and more contiguous network than the nationwide random sample used in this research. Another related issue raised was the lack of sufficiently large samples when applying the taxonomy to collision analysis. However, highway authorities have already expressed an interest in collaborating with the research team in a future project to apply the taxonomy more widely. It is hoped that a viable opportunity to attempt the application of the taxonomy will present itself in the future.

Thirdly, when considering how the taxonomy should be applied in practice, some stakeholders were concerned that the machine-learning model used in the initial stages of developing the taxonomy felt too much like a 'black box', which did not lend itself to transparent scrutiny and would be hard to explain to third parties. In response to this concern, and based on stakeholder input during the Development Day, the research team proposes in section 6 an outline schema for generating a transparent 'decision tree' for applying the taxonomy. While machine learning provided a valuable tool for initial phases of taxonomy development, it does not represent a feasible way of taking the project forward due to the lack of transparency and consistency that may result.

Finally, stakeholders provided feedback on the names used to describe the clusters of roads presented at the Development Day. It was felt that well-selected names would be crucial to using the taxonomy successfully; equally, poorly chosen names could generate confusion of the same kind as the existing 'rural road' categorisation. Accordingly, the names and descriptions presented at the Development Day have been reviewed, and the versions presented in this report have been significantly altered in response to this concern.

6. The next steps – a proposed decision-tree algorithm

To support these stakeholder aspirations and resolve concerns, the research team believes it is desirable to establish a decision tree, which could be employed programmatically, in a systematic and consistent manner, across large networks or even the entire country. This section introduces a draft of how such a decision tree might look. Naturally, this approach would require validation by means of application to a substantial contiguous network; a process that doubtless would lead to refinements of the method. Table 6.1 lays out a series of logical steps that could form the basis of an objective algorithm for applying the taxonomy across a network. The final determination of the cluster that would be applied to each route is shown in bold. The decision tree also appears in graphical format in Figure 6.1.

In this draft decision tree, the steps describe quantitative metrics such as road width with imprecise terms, using phrases such as ‘low’ and ‘very high’, or variables like ‘x%’ and ‘y daily flow’, rather than exact values. This is intentional;

these approximations appear in italics in Table 6.1 to accentuate this point. A definitive version of the decision tree would include absolute benchmarks to apply at each step of the process, but at this stage of development, promulgating exact values would be overly speculative; it is preferable to establish these values rigorously by evaluating them against reality. One of the most important purposes of any future research would be to assess outcomes for a range of values against real data, with the goal of establishing definitive absolute measures.

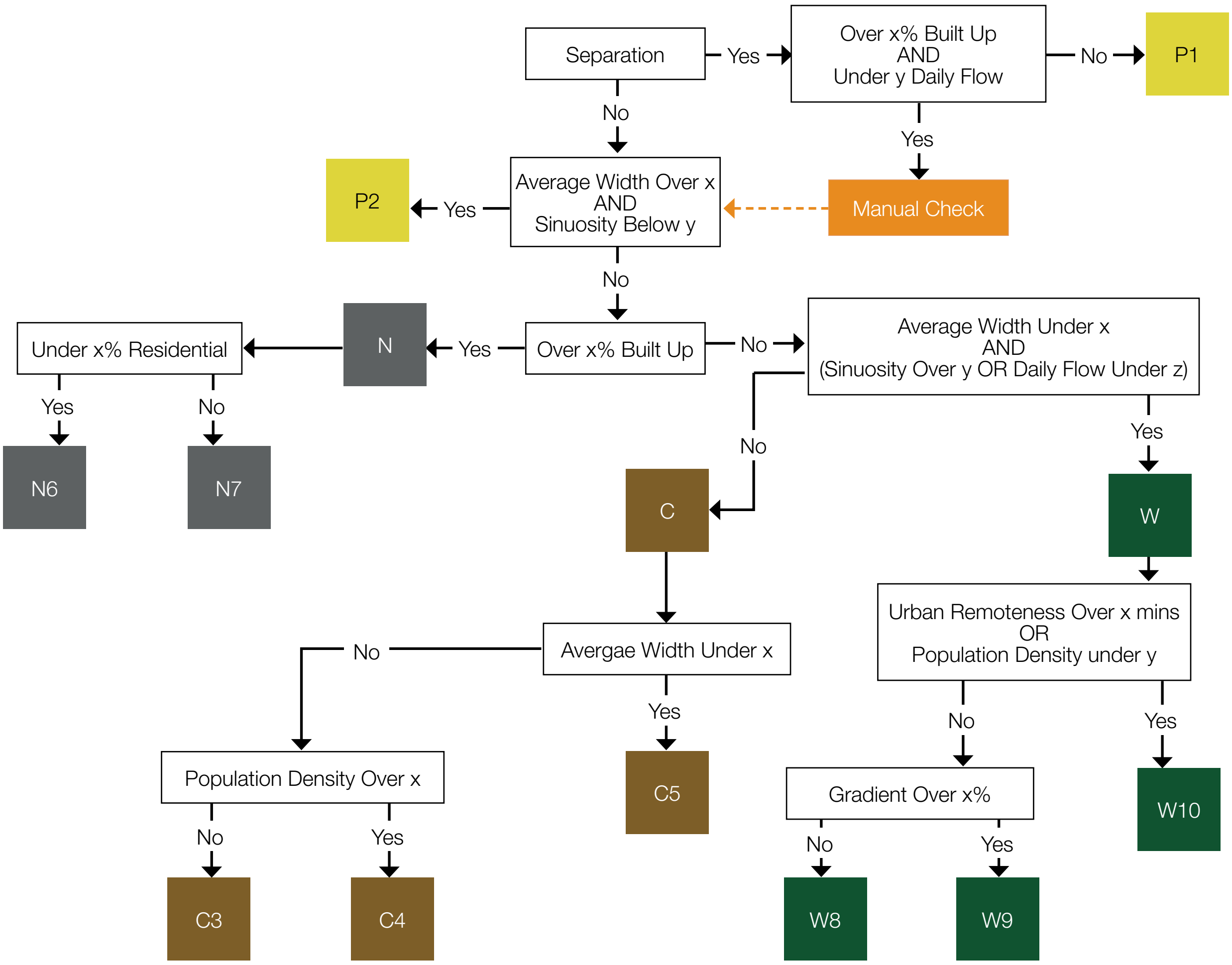
It is unavoidable that applying absolute criteria to continuous data throws up anomalous results in borderline cases, as evaluating this taxonomy with expert stakeholders amply demonstrated. To acknowledge this, the current draft decision tree anticipates the need for a degree of manual intervention to minimise the impact of this phenomenon. For example, the decision tree shown in Figure 6.1 acknowledges the probability that in a small but non-zero number of routes, a central median is present in a road through a residential area, which should be characterised as N7 but could be misleadingly classified as P1 because of carriageway separation.

Manual mediation of this kind, implemented by means of dip checks or similar techniques, should be kept to a minimum to ensure that implementation of the taxonomy remains as objective as possible. However, at this stage it is wiser to acknowledge the probability that a degree of human judgement will be required, for the sake of the integrity of the overall taxonomy. Table 6.1 shows instances of this underlined.

Table 6.1: Draft decision-tree schema for rural roads taxonomy

Step	Description
1	Distinguish Principal Routes
1.1	**P1** for carriageway separation AND *little* built-up land use AND *very high* flow
1.2	<u>Manual check</u> for residential routes with carriageway separation
1.3	**P2** for no carriageway separation AND *fairly* wide AND straight (*low* sinuosity)
2	Distinguish Neighbourhood Roads by *high* percentage of built-up land use
2.1	**N7** for *higher* percentage of residential land use
2.2	**N6** for *lower* percentage of residential land use
3	Distinguish Winding Roads as *very* narrow AND EITHER *very* sinuous OR *very low* flow
3.1	**W10** for roads *very* remote from urban areas OR *very low* population density
3.2	**W9** for roads with *steeper* gradients, otherwise **W8**
4	Remaining roads are categorised as Country Roads
4.1	**C5** for *narrow* roads
4.2	**C4** for *higher* population density, otherwise **C3**
4.3	<u>Manual check</u> for overlap between C5 and W8

Figure 6.1: Draft decision-tree schema for rural roads taxonomy



7. Appendices

7.1 Individual measure outputs

Figure 7.1: Average width distributions for rural road clusters

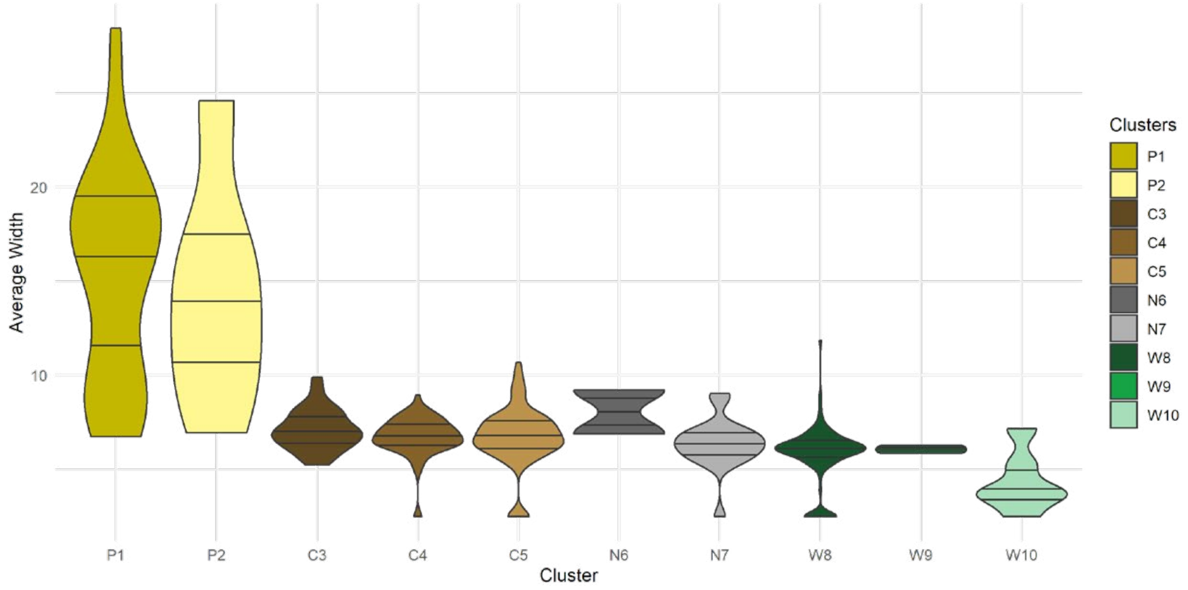


Figure 7.2: Minimum width distributions for rural road clusters

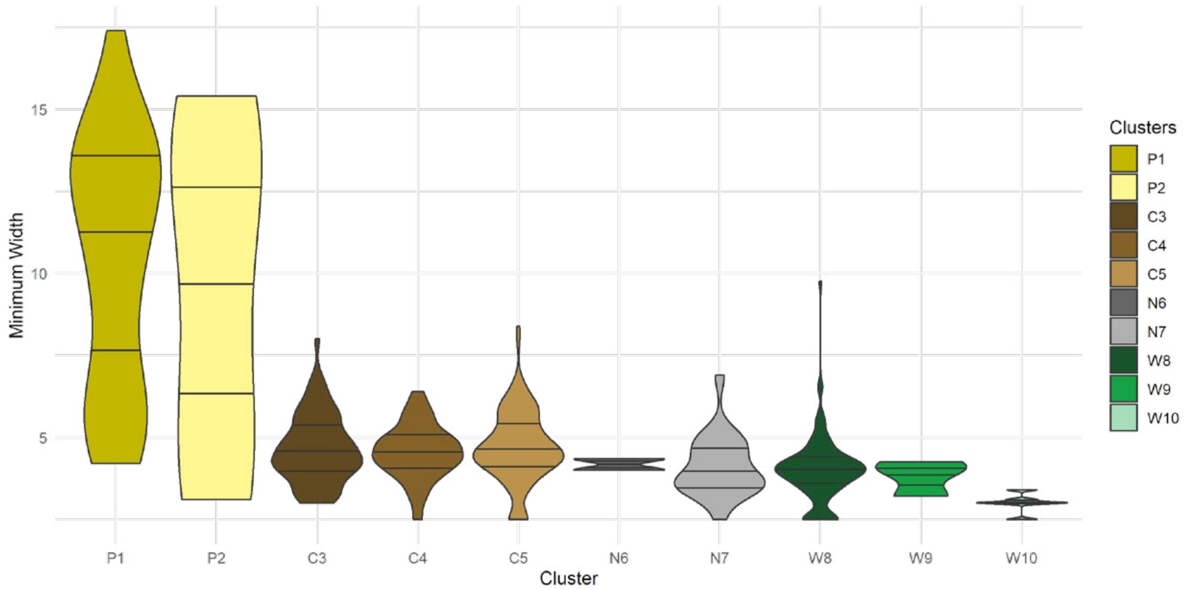


Figure 7.3: Gradient distributions for rural road clusters

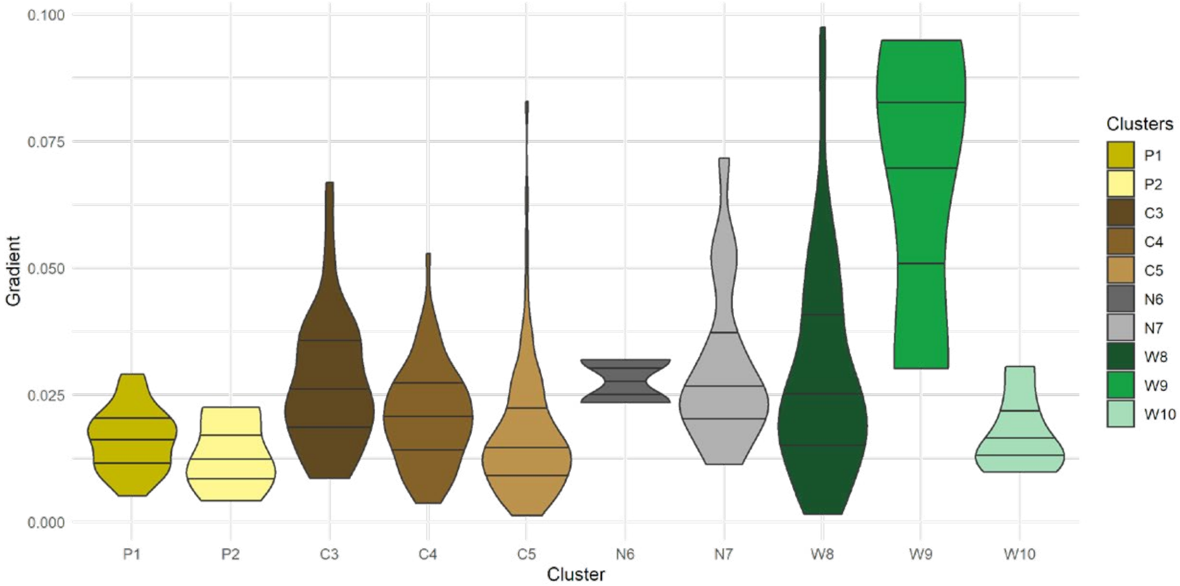


Figure 7.4: Carriageway separation distributions for rural road clusters

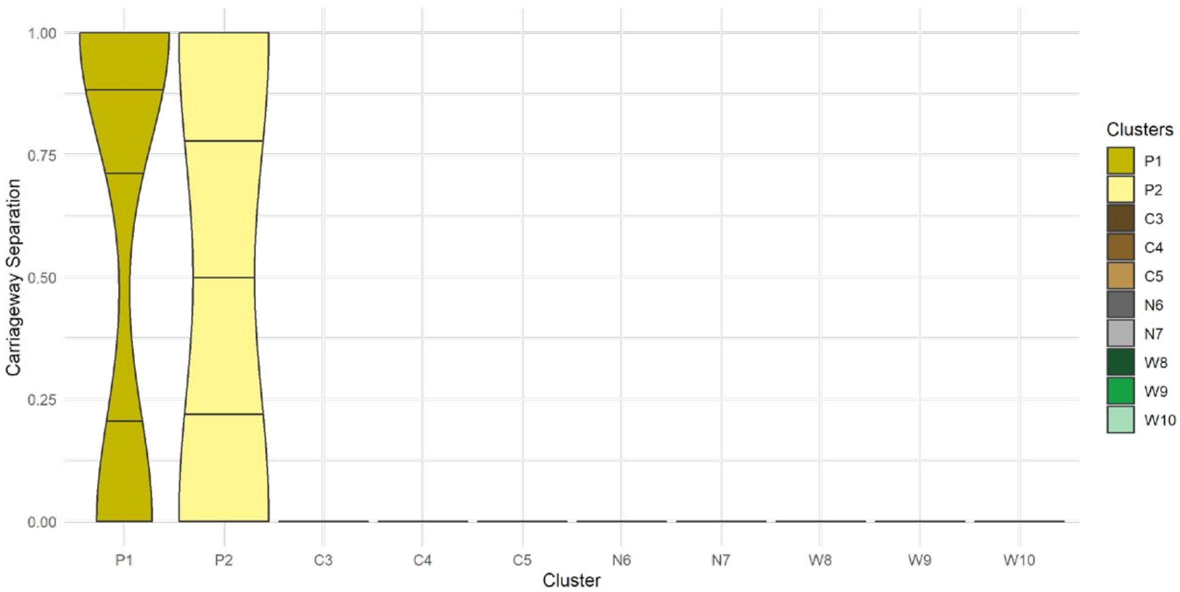


Figure 7.5: Sinuosity distributions for rural road clusters

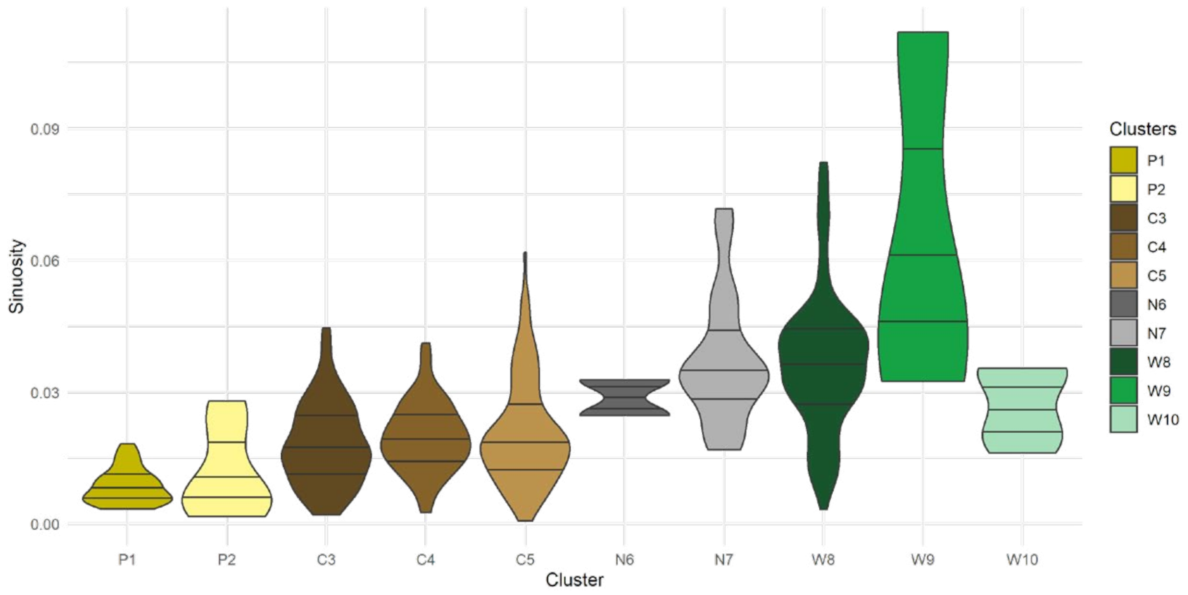


Figure 7.6: Straightness distributions for rural road clusters

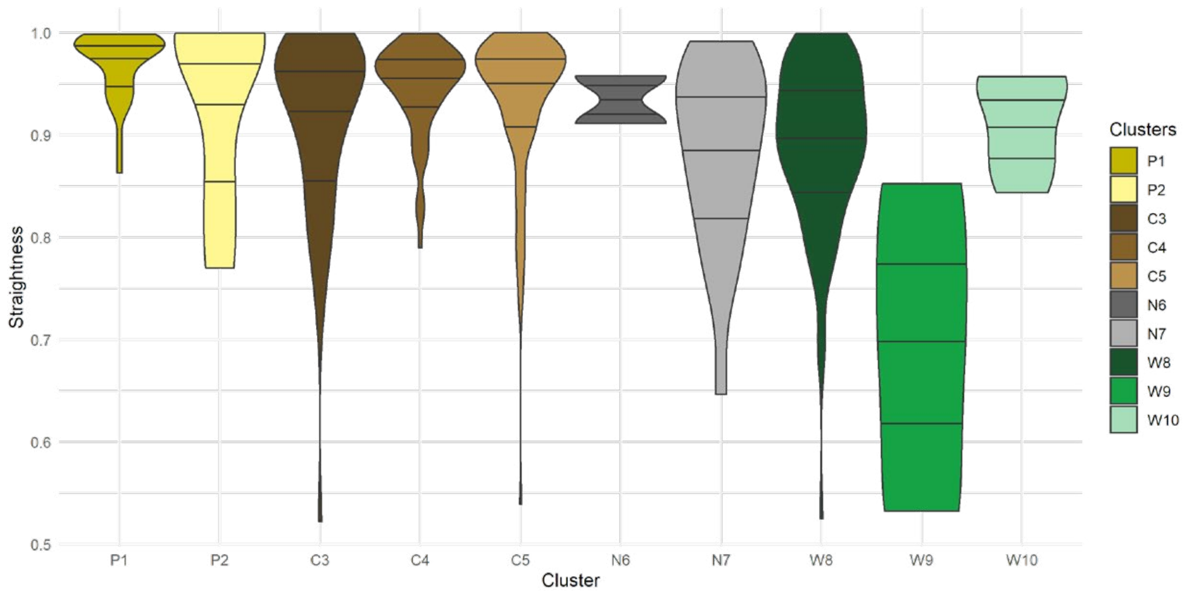


Figure 7.7: Speed limit distributions for rural road clusters

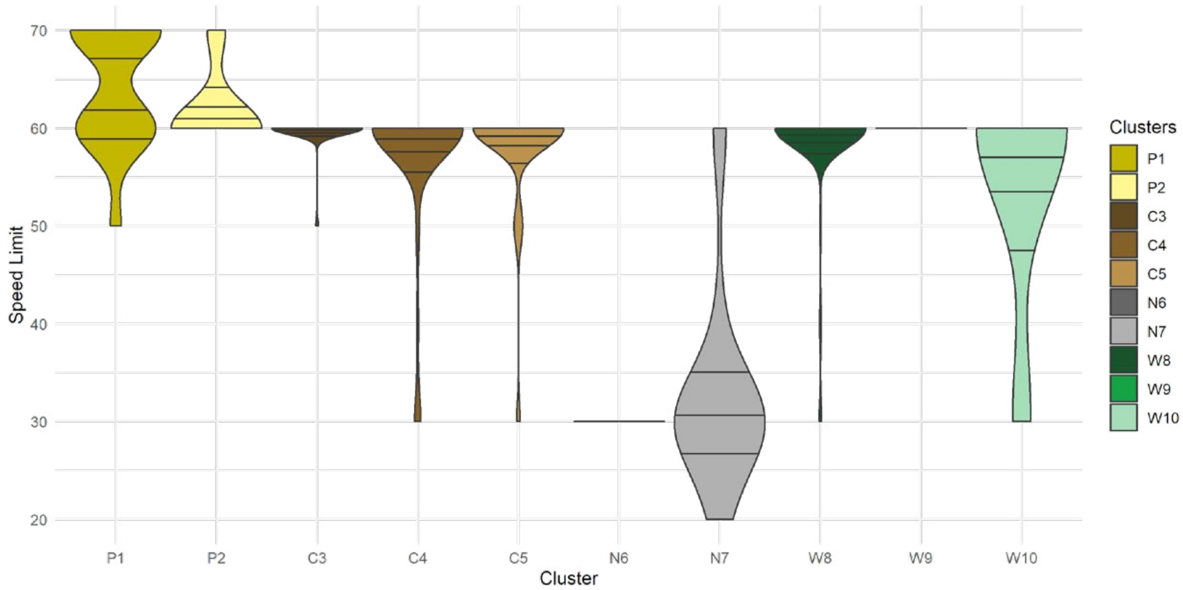


Figure 7.8: Peak average speed distributions for rural road clusters

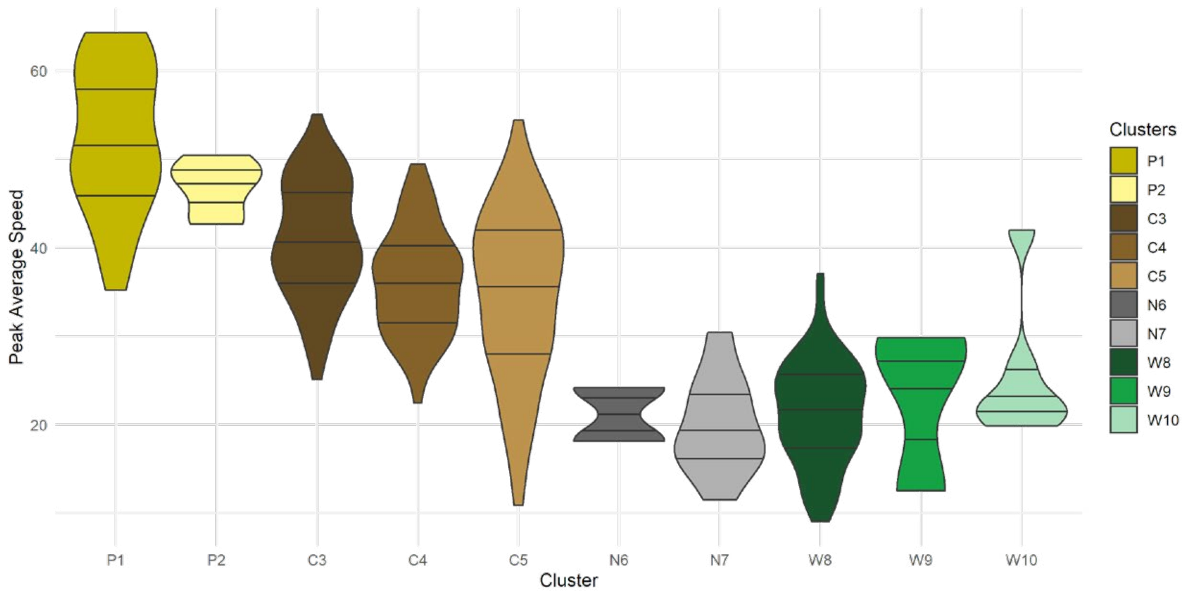


Figure 7.9: Off-peak average speed distributions for rural road clusters

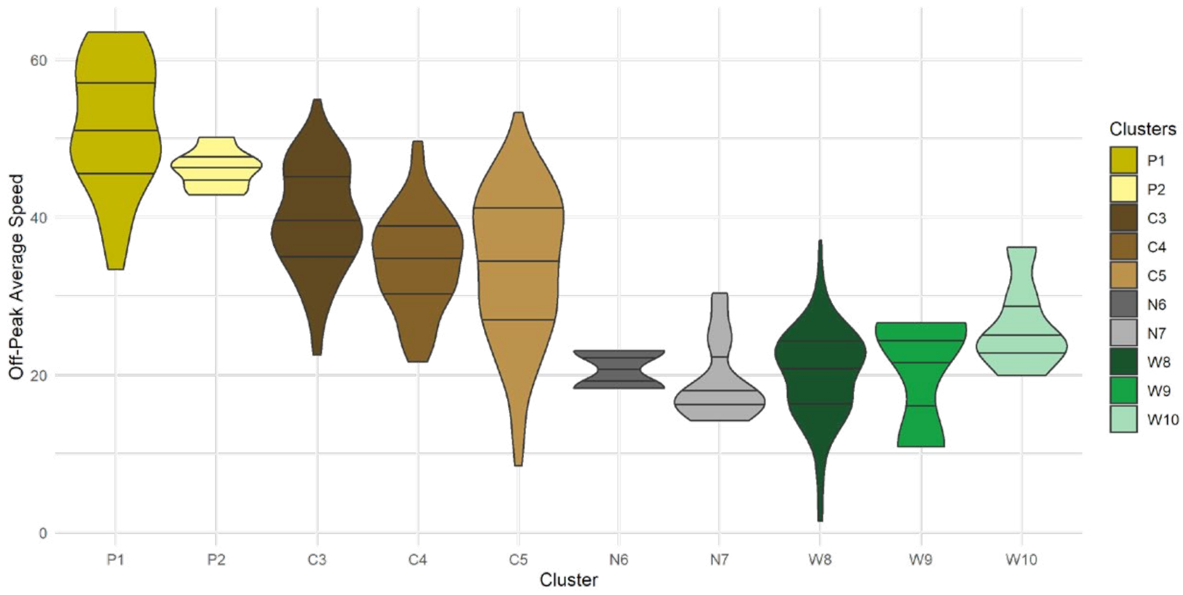


Figure 7.10: Peak 85th percentile distributions for rural road clusters

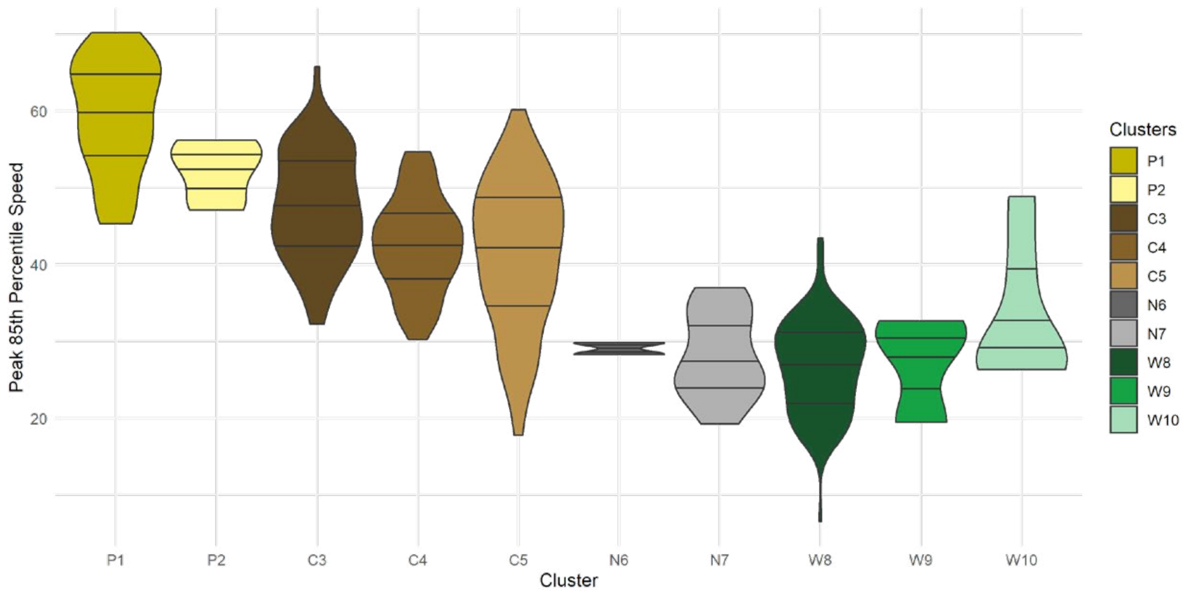


Figure 7.11: Off-peak 85th percentile distributions for rural road clusters

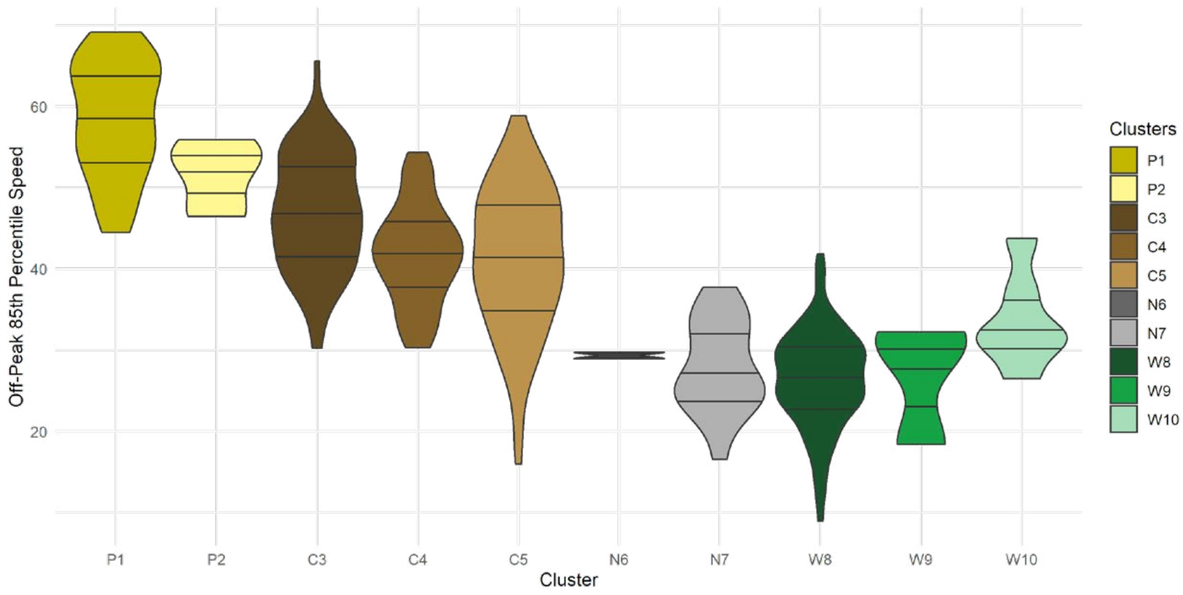


Figure 7.12: Telematics count distributions for rural road clusters

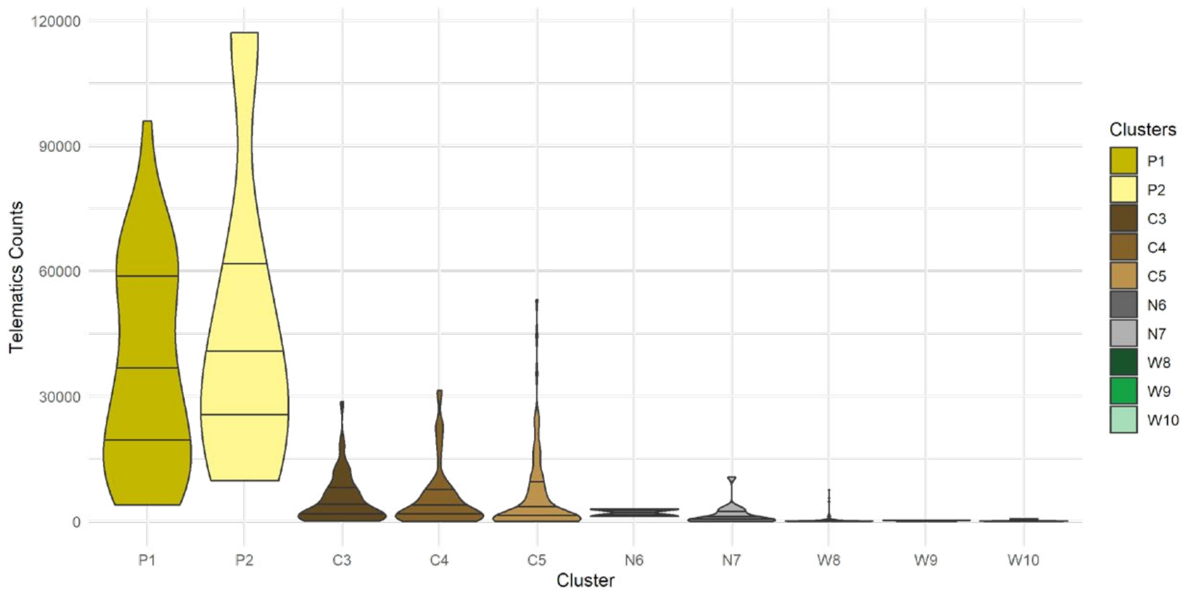


Figure 7.13: Travel time to towns distributions for rural road clusters

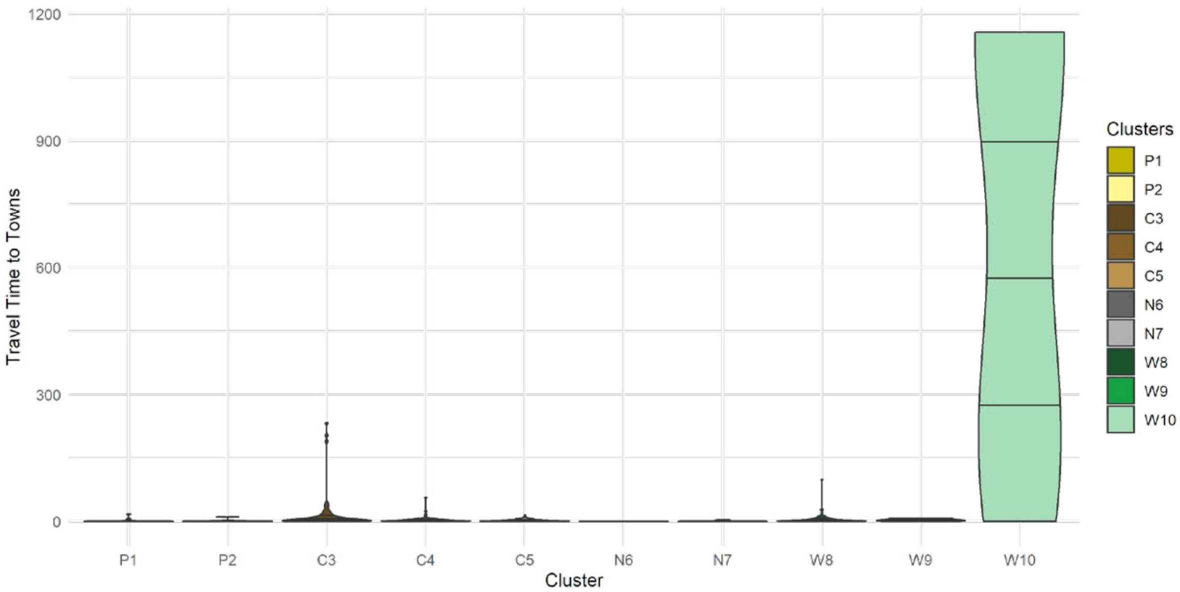


Figure 7.14: Travel time to urban areas distributions for rural road clusters

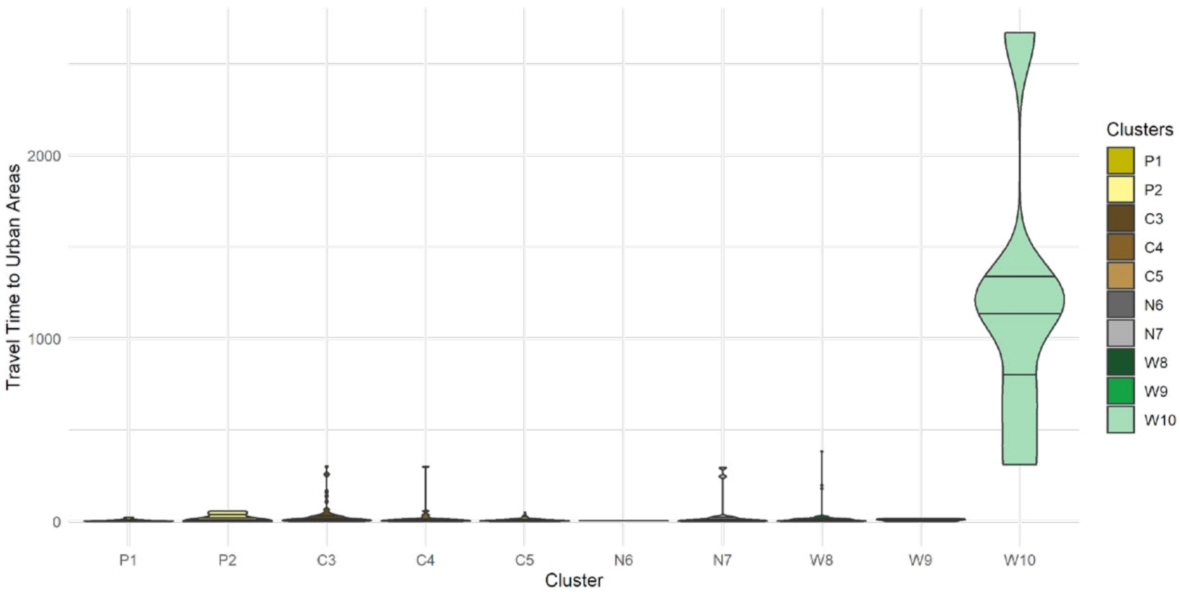


Figure 7.15: Network density distributions for rural road clusters

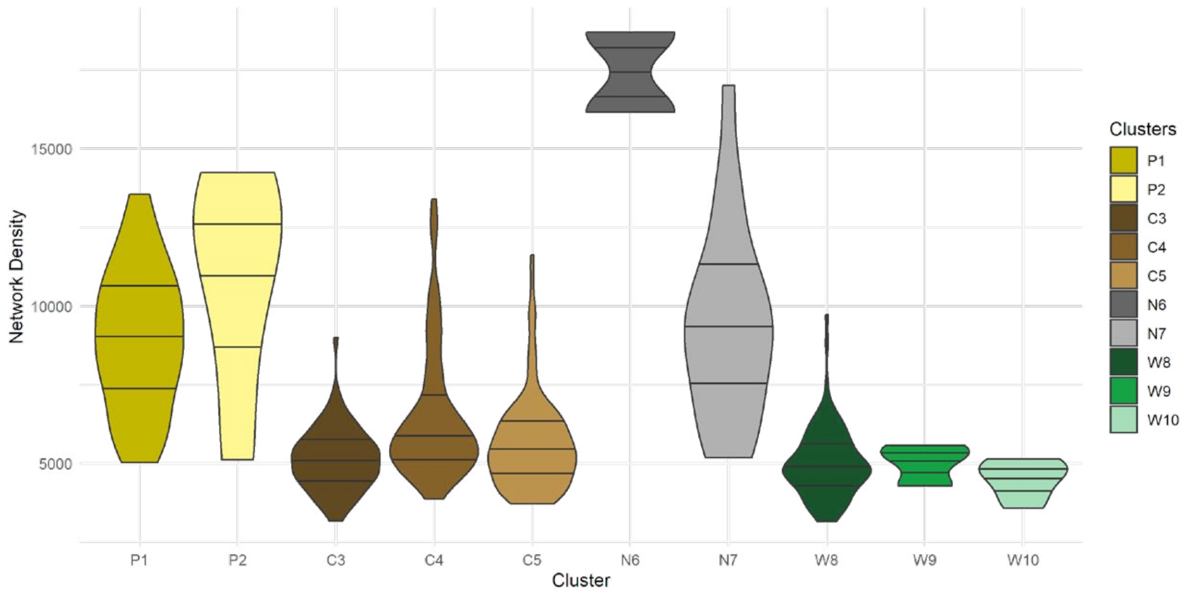


Figure 7.16: Population density distributions for rural road clusters

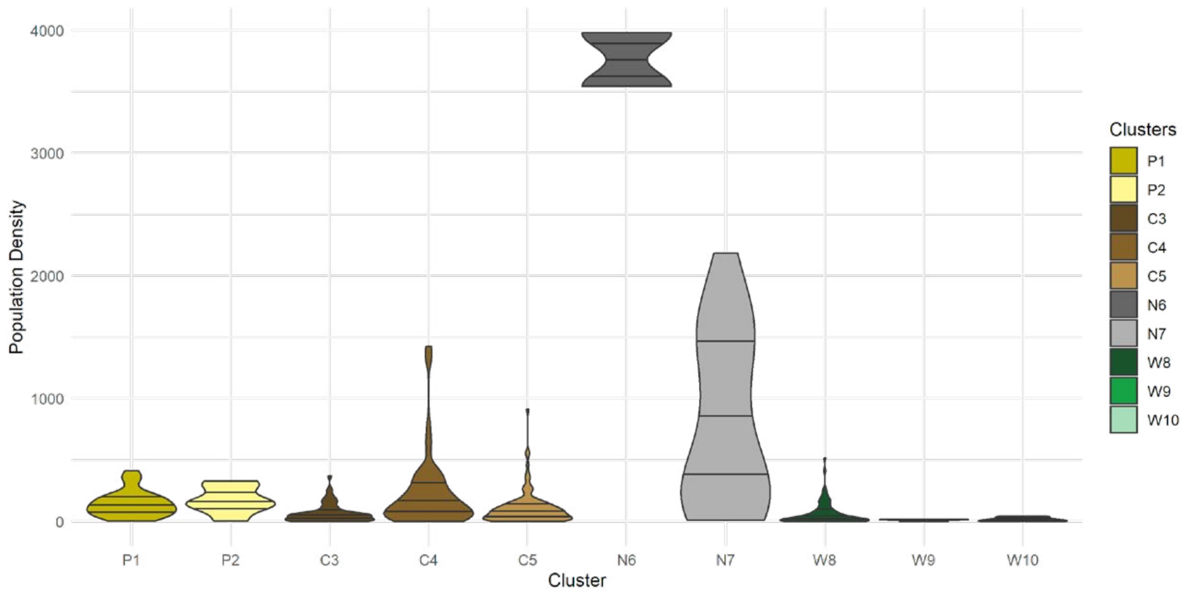


Figure 7.17: Arable land distributions for rural road clusters

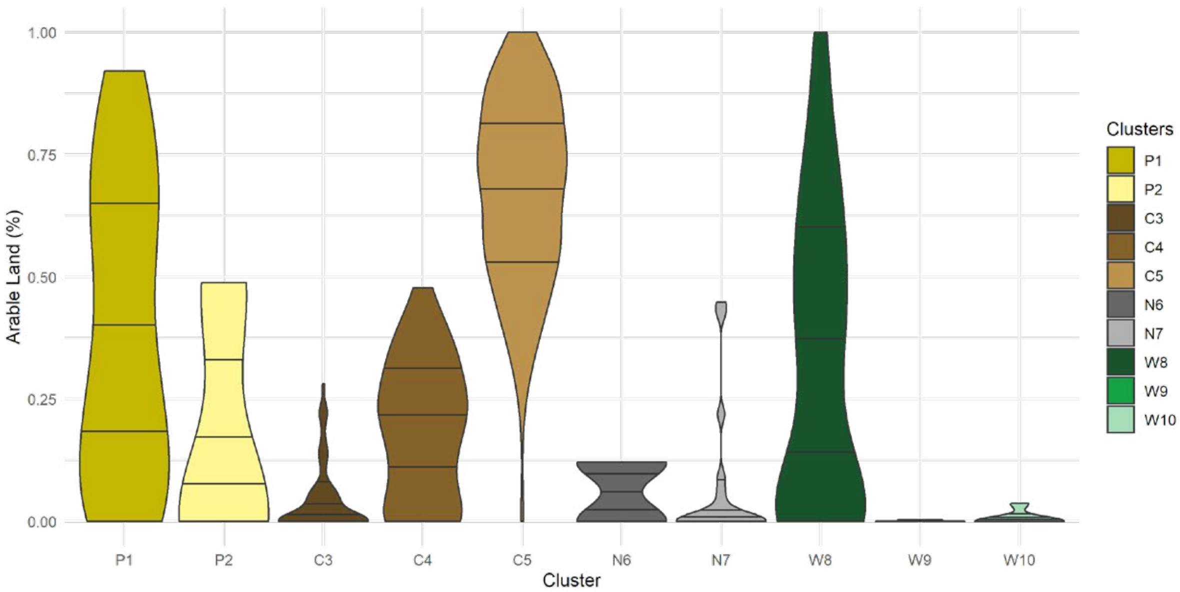


Figure 7.18: Broadleaf woodland distributions for rural road clusters

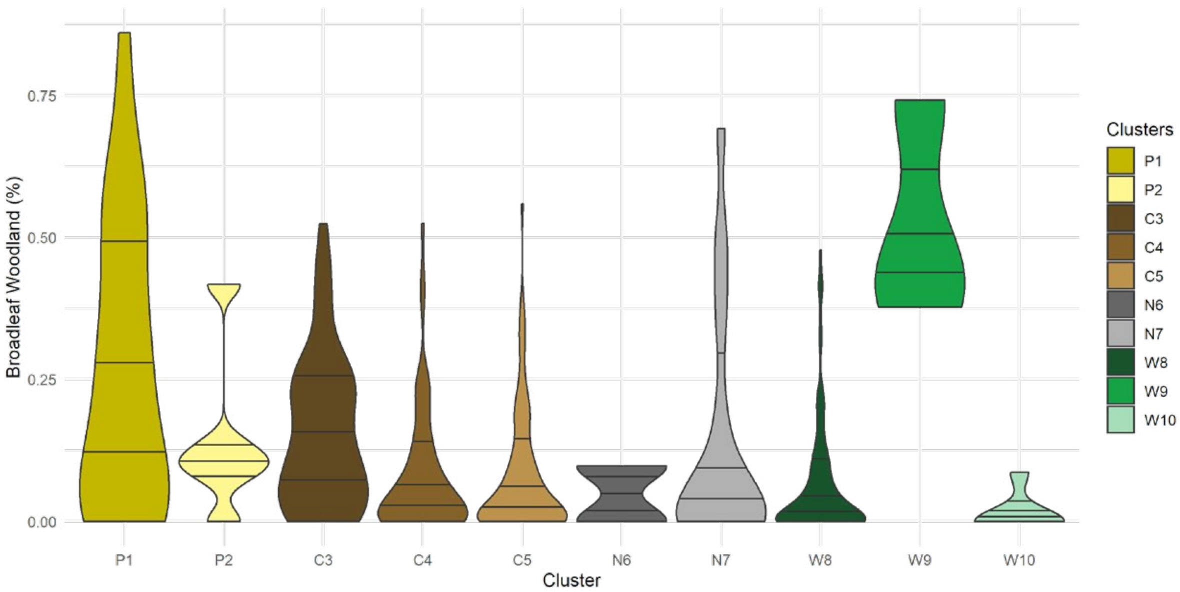


Figure 7.19: Coniferous woodland distributions for rural road clusters

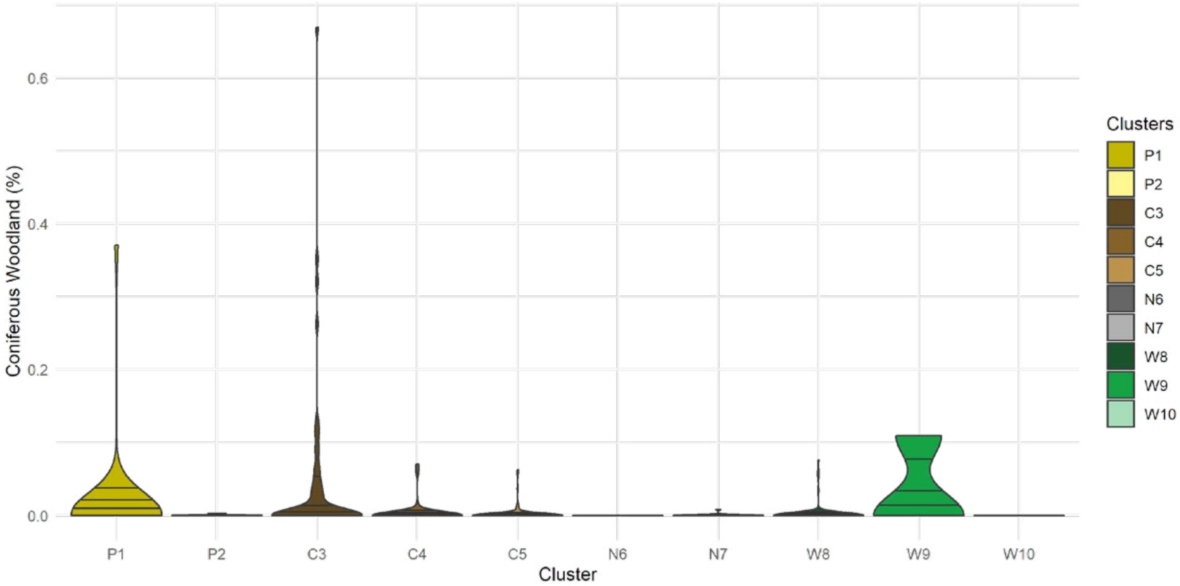


Figure 7.20: Improved grassland distributions for rural road clusters

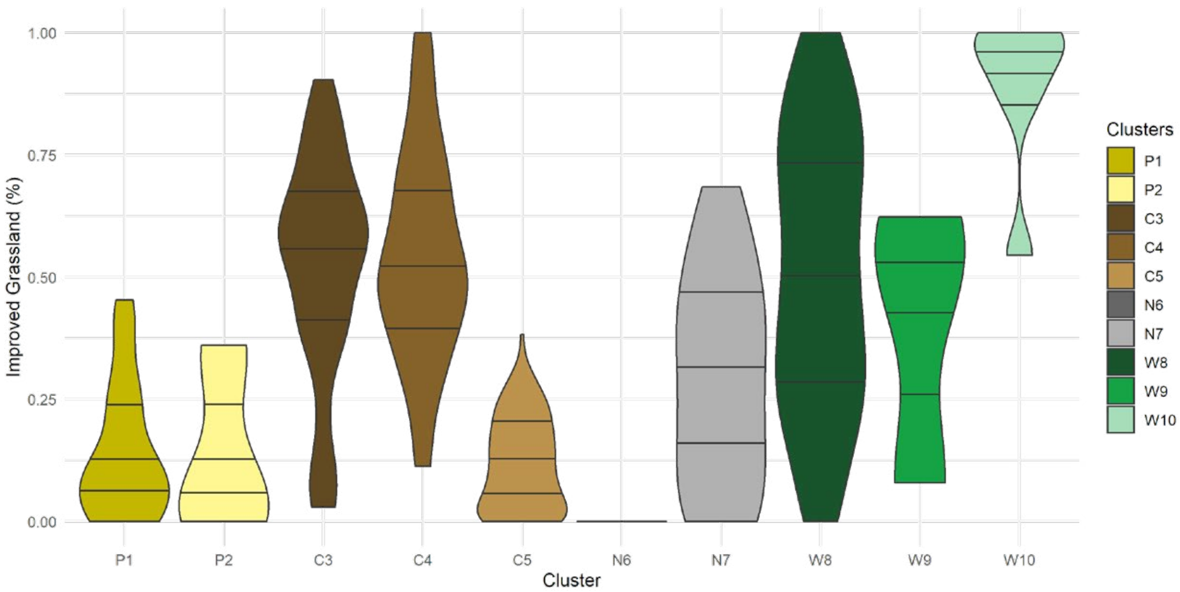


Figure 7.21: Semi-natural grassland distributions for rural road clusters

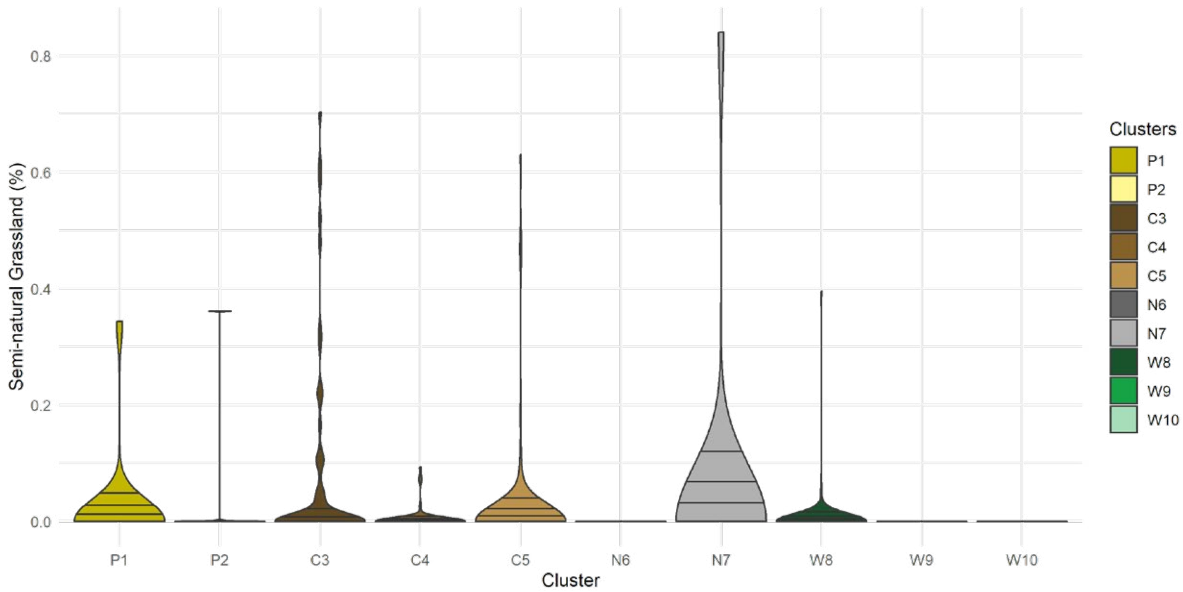


Figure 7.22: Built-up areas and garden distributions for rural road clusters

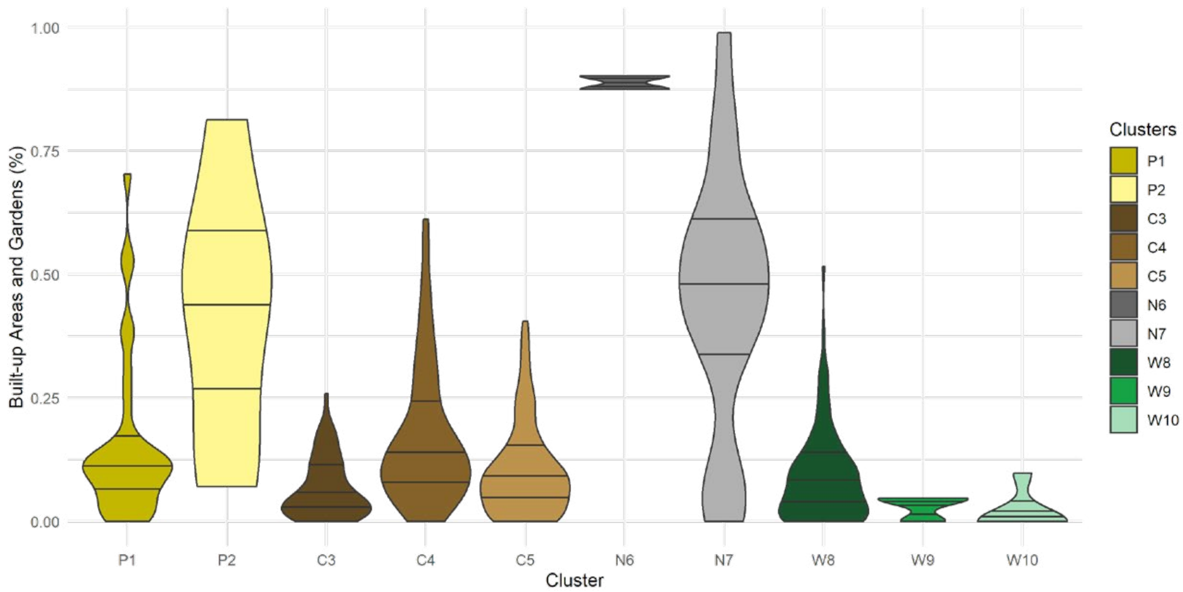


Figure 7.23: Mountain, heath and bog distributions for rural road clusters

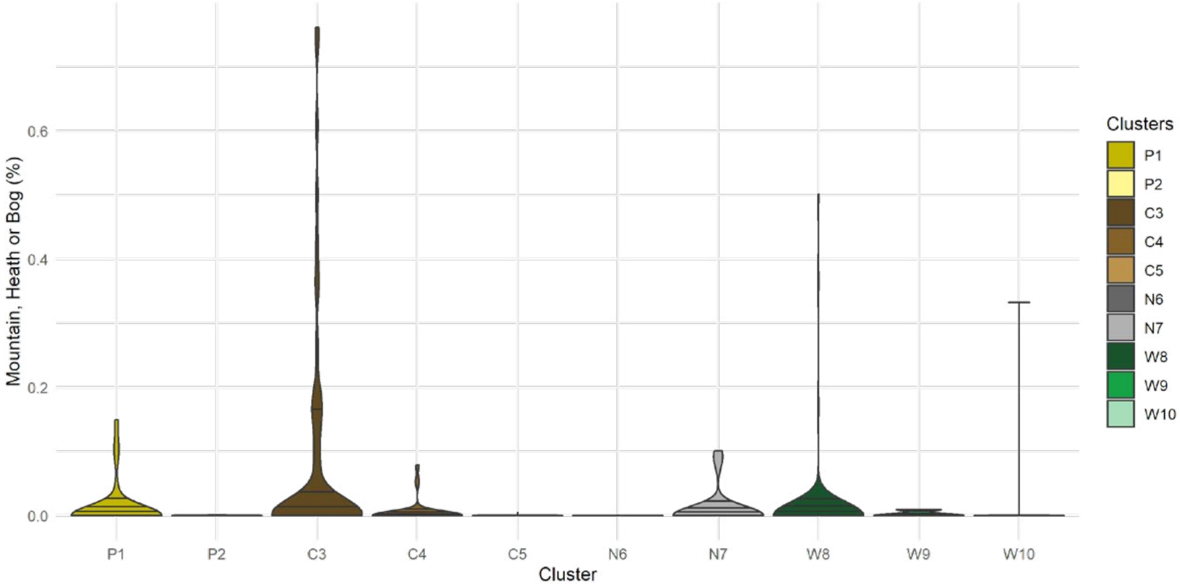
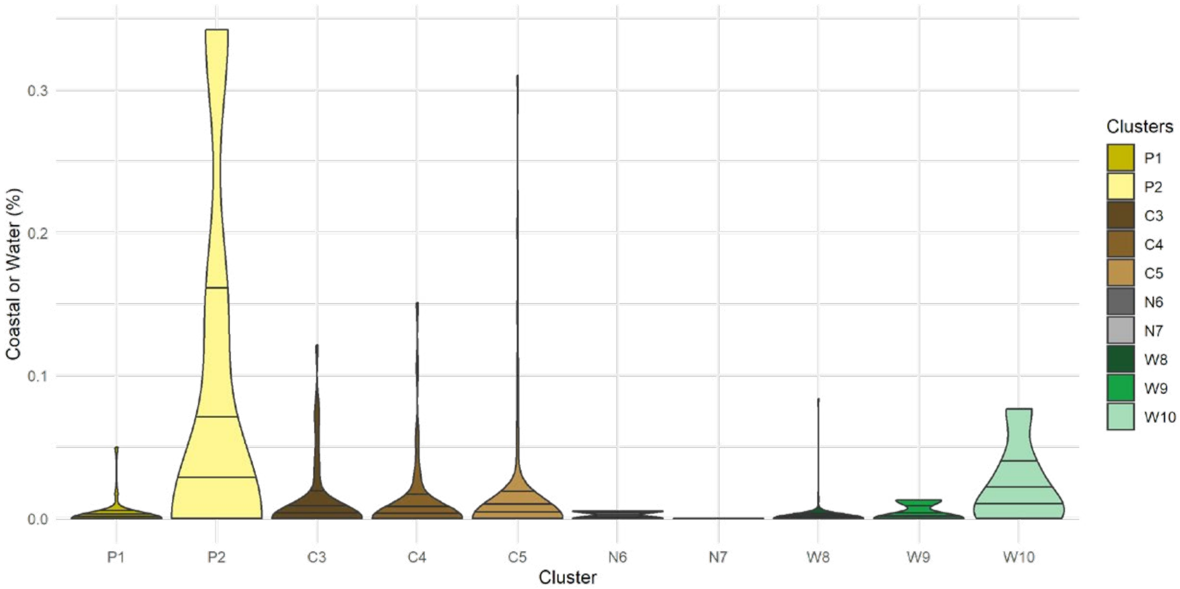


Figure 7.24: Coastal and water distributions for rural road clusters



7.2 Additional route examples

Figure 7.25: Regional Distributor example

Figure 7.26: Main Road example

Figure 7.27: Main Road example

Figure 7.28: Primary Route example

Figure 7.29: Primary Route example

Figure 7.30: Secondary Route example

Figure 7.31: Secondary Route example

Figure 7.32: Tertiary Route example

Figure 7.33: Country Lane example

Figure 7.34: Country Lane example

Figure 7.35: Hill Pass example

Figure 7.36: Remote Road example

7.3 Technical appendix on road network construction

The underlying spatial layer is derived from the latest Ordnance Survey Open Roads network. Data from both the link and node layers is used to determine the shape of the route network. The route network only covers named roads, as a road name is essential for determining the extent of a route. This covers all numbered roads and most unclassified roads.

Initially, major junctions are extracted from the network and dissolved into individual features, based on their road number. Major junctions consist of slip road and roundabout links that collectively meet more than one numbered road. On major junctions, slip roads are split where they meet another numbered road, but roundabouts do not split.

Once major junctions have been dissolved, the algorithm moves on to the rest of the road network.

Numbered roads are dissolved based on their road number, excluding major junctions. These routes are then split wherever they meet a numbered road section with a different road number, or where they meet any major junction. Note that the algorithm does not treat bridges over, or tunnels under, roads as meeting points, so routes do not split at bridges or tunnels.

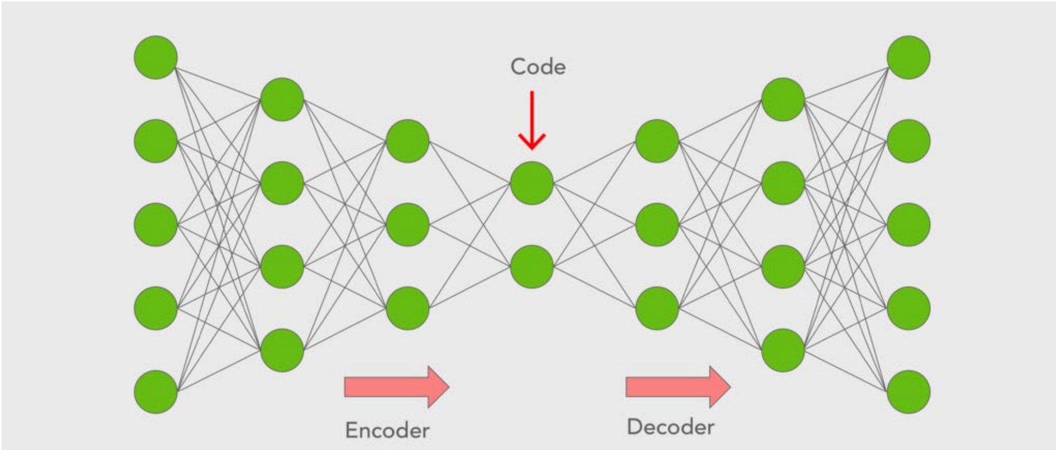
Unnumbered roads are dissolved based on their road name. These routes are then split wherever they meet a numbered road section or a roundabout, but do not split where they meet other unnumbered roads at a normal junction.

It should be noted that routes do not split where they meet other classified roads if the junction is designed to allow traffic to flow seamlessly past the intersection; for example, at hamburger junctions. This is because Ordnance Survey does not class these as roundabouts in Open Roads, and so the algorithm cannot distinguish them from the main carriageway.

7.4 Technical appendix on deep autoencoders

As outlined in section 3.1, deep autoencoders are a neural network architecture, formed from a series of dense layers that gradually decrease in size to a central bottleneck and then increase in size back to the original input dimensions (see Figure 7.37). The first series of layers form the *encoder*, which compresses the input data into the bottleneck. The second series of layers form the *decoder*, which decompresses the data following the bottleneck. Deep autoencoders are trained to ensure that the output data matches the input data as closely as possible, after having passed through both the encoder and decoder, despite the data compression occurring at the bottleneck of the model.

Figure 7.37: Structure of a deep autoencoder

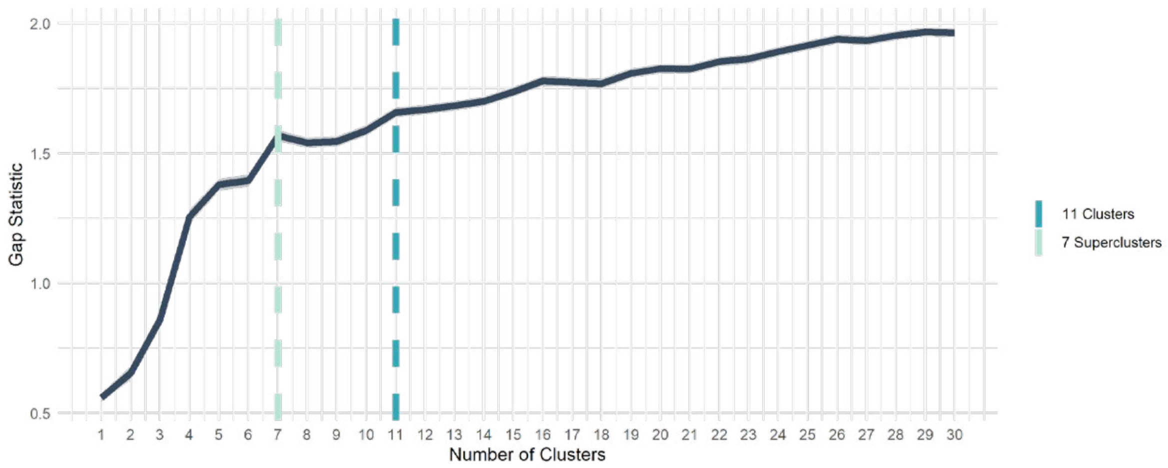


In this project, the deep autoencoder takes the high-dimensional input data containing information on the physical characteristics of the road, its usage and the local environment, encodes this using the first half of the neural networks into a low-dimensional latent space and then decodes the data again into the original high-dimensional state. In theory, to have the most efficient autoencoders, similar roads are placed closer together in the latent space, while dissimilar roads are placed further apart.

Once roads have been encoded into the latent space and arranged by their levels of similarity, an agglomerative hierarchical clustering algorithm is applied. This process starts with all roads as separate clusters and iteratively joins the most similar clusters together until all roads are in a single cluster. Towards the end of this process, the gap statistic is measured at each stage; this measures the within-cluster dispersion (how similar the roads are within the same cluster) relative to the expected within-cluster dispersion of randomly distributed data. This value is used to determine the optimal stage at which clusters are taken from the hierarchical clustering algorithm and hence influences the number of clusters to be analysed.

If the change in this gap statistic from n clusters to $n+1$ clusters is smaller than previous increases (or even reduces), then is considered a 'good' number of clusters. This occurs because dividing a cluster in two at this stage has less effect on improving cluster cohesion. As shown in Figure 7.38, the sharp peak at around seven clusters suggests an optimal number of seven superclusters. The second peak at 11 clusters suggests that these could then be subdivided into a further 11 clusters.

Figure 7.38: Gap statistic of agglomerative clustering



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 metric. However, determining the ideal metric is an intractable problem, and so the Euclidean metric is the best compromise. But this means that some caution should be exercised when considering roads on the fringes of clusters in the latent space, as cluster assignment may not be perfect at borders where multiple clusters meet. To address this issue, manual quality assurance was carried out on the clusters, resulting in the reassignment of some roads to different clusters.

7,5 Collision trends

Table 7.1: Machine-learning collision analysis results

Collision characteristic	Supercluster P	Supercluster C	Supercluster N	Supercluster W
% KSI	34%	37%	20%	32%
% Junction collisions	18%	37%	57%	22%
% Single vehicle collisions	35%	36%	43%	51%
% Multiple vehicle collisions	65%	64%	57%	49%
Predominant point of impact	57% (Rear)	30% (Head-on)	-	38% (Other impact)[16]
% Run-off	45%	43%	29%	37%
% VRU involvement	19%	19%	29%	28%

16 In addition to these Other Impact collisions, 31% of supercluster W collisions were Head-on

Table 7.2: Qualitative analysis collision analysis results

Collision characteristic	Supercluster P	Supercluster C	Supercluster N	Supercluster W
% KSI	32%	38%	20%	30%
% Junction collisions	28%	34%	57%	17%
% Single vehicle collisions	29%	40%	43%	46%
% Multiple vehicle collisions	71%	60%	57%	54%
Predominant point of impact	53% (Rear)	31% (Head-on)	-	39% (Head-on)
% Run-off	38%	47%	29%	29%
% VRU involvement	16%	21%	29%	29%

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

Designed and printed by
The Javelin Partnership Ltd
Tel: 0118 907 3494

Produced on paper from a managed
sustainable source which is FSC certified
as containing 50% recycled waste.