Utilising Supervised Paramatric Classification to Assess the Quality of the UK Rural Road Network using Aerial LiDAR Data

201374125

Abstract

An automated method for the identification of rural road geometric features is presented for use in a national rural road assessment in England. This method includes the use of LiDAR for height and road surface intensity information, as well as aerial imagery, and road centrelines enabling automated classification of road surfaces prior to road width extraction. The method presented ensures scalability, allowing for an extension beyond the 1km^2 area proposed, given existing data and suitable computational power. A unique classification method is proposed, utilising a linear probability model given points determined to be road surface, derived from the road centreline information, while implementing a sampling methodology which primarily reduces the computational overhead. Results were broadly assessed qualitatively and quantitatively giving around a 70% accuracy in road width measurements. Additional geometric features extracted from roads include the change in road elevation and road surface quality, derived through LiDAR data, and road bend sharpness, derived through the OS Road Centrelines data.

Keywords: LiDAR; Aerial Imagery; Linear Probability Classification; Rural Road Quality

In Partial Fulfillment of the Requirements for the Degree of
Geographic Data Science MSc

UNIVERSITY OF
LIVERPOOL

Acknowledgements

I would like to thank my supervisor Paul Williamson for his continued support through the writing of this dissertation. Thank you to John Murray for his help with the various mathematical problems and solutions presented through the methodology, particularly for providing some of the core functions essential for the road width and angle extraction. I would also like to thank Jean Romain (and package contributors) for his excellent work producing the lidR R package, without which the manipulation of LiDAR data in this paper would have been far more limited. Finally, thank you to Jenny Gibson, who always provides love and support throughout my studies.

Contents

List of Figures

List of Tables

Local Definitions

- **Road:** a section of a full named road that is extracted from Ordinance Survey road centrelines. Often (but not always), a section of road that is in-between two junctions. Due to the nature of roads, it is essential to define a start and end point to ensure no ambiguity. Figure 1.1 gives and overview of what is considered to be a road in this analysis.
- **Node:** each road may be split into nodes, defined by the point at which a road changes direction, as indicated on the Ordinance Survey road centrelines, a node may also be the start and end of a road. In computational geometries, nodes are the start and end point of a `LINESTRING`, which may be part of a `MULTILINESTRING`.

Broad Definitions

- **Light Detection and Ranging (LiDAR):** similar to radar, a laser pulse is sent out of a transmitter on an aircraft or ground vehicle and the light is reflected back to a receiver.
- **Billion Vehicle Miles (BVM):** the total number of miles travelled by all vehicles divided by 1 billion. Typically all traffic is measured in vehicles miles.
- **Digital Terrain Model (DTM):** digital representation of the land surface topography without surface objects.
- **Digital Surface Model (DSM):** digital representation of the land surface capturing all natural and built features.
- **Transport Accessibility:** blah...
- **Social Exclusion:** exclusion from the prevailing social system and its rights and privileges, typically as a result of poverty or the fact of belonging to a minority social group.
- **Read-eval-print loop (REPL):** a simple interactive computer programming environment that takes single user inputs, evaluates them, and returns the results to the user.
- **Root Mean Square Error (RMSE):** the square root of the second sample moment of the differences between predicted values and observed values. Commonly used to measure the difference between sample or population values.
- **Scalability:** the property of a system to handle an increasing amount of work. In this context, a methodology that may handle an increasing area of interest.

1. Introduction

ROAD usage in the United Kingdom has been steadily increasing by year with the highest ever billion vehicle miles travelled in 2018 (318.1 BVM; Department for Transport 2019). Characterised by tall hedgerows and winding turns, rural roads in the UK are often unsuitable for higher traffic flow due to the obstruction of view from protected hedgerows, narrow lanes and often poor condition (Department for Transport 2018*a*). Due to the abundance of these roads, with "Unclassified" local network roads making up 60% of all roads in the UK (Department for Transport 2012), and their varying nature, the national assessment of these roads into appropriate speed limits on an individual basis has been considered impractical (Taylor et al. 2002). Due to this, there have been no individual assessments for the majority of rural roads, and instead, given their nature, they are classified as unlit, single carriageway roads and thus assigned a default speed of 60mph (UK Government 2019*b*). Highways England manages the motorways and trunk roads within the UK, but local road networks are maintained by Local Authorities, and as such there is at present no comprehensive information regarding these smaller road networks (Highways England 2019). Rural roads in the UK are often cited as by far the most dangerous road type with studies suggesting that up to two thirds of vehicle accidents occur on rural roads (Corben et al. 2005).

The Rural Urban Classification defines a rural area as one outside of a settlement with more than a 10,000 resident population (UK Government 2011), therefore a road could be considered rural, if either connecting or present within small settlements in the UK. This study will focus particularly on rural connecting roads, outside of rural towns. The purpose of this is to select roads that are unlit, and are unlikely to have been individually assessed, as opposed to roads that are present within more built up rural areas. These roads are considered to likely have the designated national speed limit of 60 mph, and feature hedgerows, narrow road surfaces, and bends, key features to be considered in this dissertation.

A Governmental review of speed policy considered the need for the role of speed and accidents on rural roads to be further addressed (Road Safety and Environment 2000), suggesting a framework for individual classification of roads, taking into account local considerations of the road to implement more suitable speed limits. In 2012, draft guidance for rural roads was presented by the Department for Transport suggesting a blanket reduction in rural single carriageway road speed limits from 60mph to 40mph including a reduction to 50mph for lower quality A and B roads (BBC 2012). However, this draft guidance has yet to be implemented, likely due to the costs involved in a blanket change to speed limits. For example, the cost for a complete change in national speed limits from imperial to metric in Ireland cost an estimated €30 million in speed limit signs alone (Noctor 2004). These costs suggest that an alternative to blanket implementation may be required.

National speed limits have seen little variation for a number of years, with the majority of roads following the broad criteria for the three main roads types. The three national speed limits are:

- the 30 mph speed limit on roads with street lighting (sometimes referred to as Restricted Roads)
- the national speed limit of 60 mph on single carriageway roads
- the national speed limit of 70 mph on dual carriageways and motorways.

(UK Government 2019*b*)

The Department for Transport (2013*a*) outline, in *Setting Local Speed Limits*, that national speed limits are not appropriate for all roads, where local road conditions present the requirement for alternative speed limits. The majority of the rural road network in the UK follows the national speed limit of 60mph for single carriageway roads, and 70mph for dual carriageway roads, despite driver speed often being far below the speed limit. The Department for Transport (2013*a*) note that this is especially common on C and Unclassified roads due to the narrow roads, frequent bends, junctions and access roads. In 2011, an estimated 66% of total road deaths in Britain occurred on rural roads, with 51% on single carriageway rural roads with the national speed limit of 60mph Department for Transport (2012).

The Department for Transport (2013*a*) suggest that selecting alternative speed limits for single carriageway rural roads should consider:

- History of collisions;
- The road's function;
- Existing mean traffic speed;
- Use by vulnerable road users;
- The road's geometry and function;
- and the road environment, including road-side development.

The Road Safety Management Capacity Review (Department for Transport 2018*b*) outlines the current limitations with road safety management, with the lack of defined and measurable safety performance framework, noting that such a framework should set out the long term goal of total prevention of road deaths and injuries, achieving this through a reduction in average speeds on different road types, and an improvement in emergency response times. This review states that at present there is a distinct lack of both urban and rural road hierarchies, which could be used to better match appropriate speed limits, with function, layout and design. Again, this review notes that posted speed limits often allow for speed far in excess of the design limits of single carriageway rural roads, with inappropriate but allowable speed often a contributing factor in rural accidents. Finally the report calls for a review of national speed limits as soon as possible.

A recent development for guidance in setting local speed limits is the production of the *Speed Limit Appraisal Tool* (Department for Transport 2013*b*). This tool provides an automated method for the introduction of new speed limits for local councils. This tool takes observed traffic flow, accidents, speeds, descriptive information regarding the network and current costs, outputting projections in these data to advise speed limit changes. While this tool introduces a quantitative method for individual road speed limit assessments, it misses some key features outlined in past government framework proposals (e.g. Department for Transport 2018*b*), particularly in relation to road geometry.

The innovative methodology presented in this dissertation will focus particularly on the call for an improved understanding of rural road geometry to support the production of appropriate and justified speed limits for rural single carriageway roads. Road geometry is defined here as the parameters of roads relating to geometric design, particularly relating to the appropriate road speed, stopping sight distance, road width, road bends and surface quality (Jaakkola et al. 2008).

Some road geometric information may be extracted through the readily available OS Road centreline geometries. However, the extraction of road widths poses a complication as this information is not readily available, and automated extraction requires techniques that enable road classification through the data available, either aerial imagery, or LiDAR. Road classification techniques have more recently been aided through

the introduction of LiDAR data, as an alternative to aerial imagery classification, allowing for more reliable results due to the presence of additional information that LiDAR provides. The following section outlines LiDAR in this context, and presents how LiDAR may be used to extract these features of roads.

1.1. Introduction to LiDAR

LiDAR data is collected by emitting rapid laser pulses from an aircraft towards the ground which are reflected back, measuring the distance between the aircraft and surface objects at up to 500,000 measurements per second (Environment Agency 2019). This method produces a set of highly accurate three dimensional points which collectively are known as a LiDAR point *cloud*. As LiDAR data detects all surface objects, the resultant point cloud produced will include all natural and man made structures, including buildings, roads and trees in addition to the natural variation in the terrain height, known as a digital surface model (Hatger 2005).

The main features unique to LiDAR, as opposed to similar aerial data collection techniques such as true colour imagery are outlined below:

- **Pulses**: LiDAR systems record the data by emitting a laser pulse which is reflected back at the aircraft by ground objects. If the laser hits a solid object such as ground or a building roof, this laser pulse is entirely reflected back towards the aircraft, giving a single point. However, if the laser pulse hits a soft object such as a tree canopy, the pulse may be partially returned, giving multiple return pulses (Rottensteiner et al. 2003). Therefore, these multiple pulse returns give information regarding objects at an exact xy location but with varying heights.
- **Intensity**: LiDAR systems also give intensity values for return pulses, which gives information regarding the reflectance of the surface of objects that are hit by the laser pulses. If intensity is given I then reflectance R may be represented as $R = \frac{I}{E_T}$ where E_T refers to the first pulse signal intensity (Charaniya et al. 2004).
- **Elevation:** In addition to x and y coordinates, the distance between the plane and the reflected ground or object is recorded and assigned a z value.

1.1.1. Benefits over Aerial Imagery

Rural roads in the UK are often characterised by dense hedgerows either side, with large trees that extend over the road surface. In addition to the reduction in corner visibility on these roads, standard aerial imagery suffers from the road surface being obscured by shadows from these trees and hedgerows, and the tree canopy itself. Additionally, aerial imagery often suffers from obstruction due to clouds (Li et al. 2016). Due to the inclusion of pulses with modern LiDAR data, the road surface can often be detected through the canopy by selecting the final pulse returns, the infrared laser pulses also have smaller shadows, due to the narrow scanning angle of LiDAR (Wang & Shan 2009). Non LiDAR imagery often suffers from scene complexity, where road patterns, vehicles and lane markings reduce road heterogeneity (Li et al. 2016).

The 3D z value information provided by LiDAR data allows for the separation of ground and objects on the surface, meaning roads and buildings are often easily separated, despite having similar reflectance (Sampath & Shan 2008). Additionally, the reflectance of roads is often homogeneous, and distinctly separate from vegetation (Clode et al. 2004).

1.1.2. Limitations

LiDAR lacks any texture or spectral information, and often studies in road classification have combined LiDAR with aerial imagery to alleviate this issue (Hu et al. 2004, Zhang et al. 2003), with the inclusion of lumines-

Fig. 1.1: *Study area highlighting road centrelines, each colour represents a seperate 'road' as defined by the OS Data provided. Thicker roads are B roads, while thinner are Unclassified. The inset map gives shows an example road, with features typical of roads within this study area.*

cence information to aid with road classification (e.g. Charaniya et al. 2004). LiDAR points are distributed irregularly and with varying density, with point density often higher where flight strips overlap, and tall objects can occlude points, leaving more limited data surrounding trees or buildings (Li et al. 2016).

Often road classification methodologies use LiDAR height data to identify kerbs to separate streets from pavement (Kumar et al. 2013, Vosselman & Zhou 2009), however rural roads often have no kerb, and are often at the same level as the surrounding vegetation (Yadav et al. 2018).

LiDAR data often requires a large amount of processing due to the irregular distribution of points, presence of noise and the number of variables that have to be considered, Yadav et al. (2018) note that often papers do not include information regarding the computational time for processing this data which may cause practical limitations at larger scales.

1.2. Objective of this paper

This dissertation will present a method for rural road classification and width extraction for a 1km^2 region in the North West of England. The methodology is produced in order to ensure scalability and automation, allowing for replication for any area where data is available. Data used will include road centreline geometries, LiDAR point cloud, and aerial imagery to extract road widths through linear probability models. Additionally, this dissertation aims to extract other features of roads such as elevation changes, surface quality, and the sharpness of bends. The extraction of such features aim to build upon past road classification studies, combined

with a more refined methodology that aims to ensure a higher accuracy for rural British roads.

Key Aims:

- Produce and assess an automated method using LiDAR, aerial imagery and OS road geometry to determine the true width of roads within the chosen study area, outlining the particular limitations and solutions when considering the rural British road network.
- Using OS Road and LiDAR Data produce an automated method for determining the characteristics of rural roads that relate to overall road quality. These are;

  Bend sharpness

  Road steepness

  Surface quality
- Using extracted road features, outline the overall quality of the road network, and allow for direct comparison between each road.

Unlike previous road classification methodologies, this paper aims to focus primarily on road feature extraction, and not the accurate extraction of road locations, as road centerline locations provided by Ordnance Survey already exists. This paper is organised into chapters, first a literature review, outlining the broad implications of speed limits, rural road networks, and object extraction particularly in relation to LiDAR aerial point clouds. Second, a detailed description of the methodology involved in this paper will outline the techniques used to classify road widths, in addition to the other road geometric information. A results section will primarily assess the method for road extraction, through sensitivity analysis and some qualitative observations, a section will then explore the findings. Finally a discussion will detail the implications of the findings, and suggest areas for methodological improvement.

2. Literature Review

TYPICAL road classification techniques have focused purely on urban road networks and involved methods which can be both computationally intensive and time consuming. Given the pressure for a full quantitative assessment of the current speed limits for the rural road network in the United Kingdom, there is a demand to produce comprehensive methods for rural road feature extraction. This paper primarily focuses on techniques for assessing the road geometry for roads considered to be rural connecting roads in the United Kingdom. This literature review will first outline the current understanding of the rural road network, considering the role of speed and speed limits in accident likelihood, rural accessibility, and a detailed look at current road extraction techniques involving aerial imagery and LiDAR, presenting the key differences and limitations of these studies when considering the rural road network in the UK.

2.1. British Rural Road Network

Taylor et al. (2002) conducted a study outlining the key features of British rural roads, in an effort to improve the understanding of the characteristics associated with accident rates, beyond the past *Speed-accident relationship on European Roads* (MASTER) study which primarily consisted of European road data, with limited data for England (Baruya 1998). Taylor et al. (2002) identify key features of 174 selected rural British roads across England which they use to classify roads into certain categories. This data was obtained through drive-through video recordings.

Notably, the features unique to rural roads were observed to be the land use either side of a road, consisting mainly of residential, farming, wooded, open or industrial. Verges varied from grass verges, pavement, low banks, ditches or none. The most dominant vertical feature closest to roads was recorded, which included trees (some overhanging), hedges, banks, fences and so on. This study also manually measured the road width for each site, and to determine the "hilliness" of roads, the number of 10m contour lines crossed were counted to give the total change in height. Additional key features of British rural roads are given on Table 2.1.

Taylor et al. (2002) categorised these roads into four key groups:

- **Group 1:** Roads which are very hilly, with a high bend density and low traffic speed. *These are low quality roads.*
- **Group 2:** Roads with a high access density, above average bend density and below average traffic speed. *These are lower than average quality roads.*
- **Group 3:** Roads with a high junction density, but below average bend density and hilliness, and above average traffic speed. *These are higher than average quality roads.*
- **Group 4:** Roads with a low density of bends, junctions and accesses and a high traffic speed. *These are high quality roads.*

Table 2.1: *Features Identified as Important to Rural Road Analysis (Taylor et al. 2002)*

Type of data	Examples
Discrete data	Type of junction Minor junctions Accesses Number of bends, classified into: Sharp (warning signposts) Medium Slight
Semi-continuous data	Lighting Reflecting road studs Kerbs Number of lanes Road markings Land use
Continuous data	Visibility Verge width and type Roadside type

This study therefore attempted to outline a rural road hierarchy in relation to road function, and certain road geometries, which addresses issues outlined in the Government's review of speed policy (Road Safety and Environment 2000). However, due to the nature of the data collection for this study, time constraints mean that producing a full road hierarchy for all rural roads within England using this methodology is impractical. Additionally the speed policy review notes that road hierarchies should inform appropriate speeds for each road, rather than concentrating on current speed limits and average vehicle speed.

2.1.1. Rural Speed Limits

An observation of 270 single carriageway rural roads in England found that the distribution of mean speeds was wide, and often significantly below the 60mph limit (Department for Transport 2006). Accidents on rural roads often occur within the 60 mph speed limit meaning a distinction between what is an appropriate speed should be made that does not relate to a given speed limit. Baruya (1998) suggest a distinction between both *excess* and *inappropriate* speed. *Excess* when driving above the speed limit, and therefore directly breaking the law; *inappropriate* speed, when driving too fast for the conditions of the road, not necessarily above the speed limit, often considered dangerous driving. A study by the Department for Transport (2013*a*) assessed the impact of inappropriate speed on rural roads, which contributed to 20% of all crashes on minor rural roads with a 60mph limit, whereas excess speed accounted for around 16% of collisions. The Department for Transport (2013*a*) note that this high proportion of inappropriate speed on rural roads reflects the inappropriate speed limits that are given on the majority of rural roads

The Department for Transport found that rural roads account for around 66% of all road deaths, despite accounting for around 42% of the total distance travelled by all vehicles. Notably 51% of all deaths in Britain in 2011 occurred on rural single carriageway roads, with the national speed limit of 60mph (Department for Transport 2011).

2.1.2. Speed, Road Geometry and Accidents

Lowering the speed limit on roads has been shown to result in an overall reduction in the average speed of vehicles. Finch et al. (1994) found that a reduction in the speed limit of a road resulted in a mean speed reduction of around one quarter of the difference, noting that drivers will often obey speed limits that they determine to be reasonable. A reduction in average speed subsequently leads to a reduction in road traffic

accidents (Finch et al. 1994, Taylor et al. 2002), Taylor et al. (2000) produced a model to predict accident frequencies given the proportion of drivers exceeding the speed limit and the average speed, finding that excess speed and a higher speed limit were both associated with a higher accident frequency. Particularly, the risk of death at various speeds has been assessed in various studies, Richards & Cuerden (2009) found that at 60mph the risk of a driver dying in a head on collision between two cars is around 90%, but with a reduction in speed, this drops to around 50% at 48mph.

Taylor et al. (2000) demonstrated that traffic flow, link length, and the number of minor junctions all directly increased the number of accidents, while wider roads were associated with a reduction in the number of accidents. The *MASTER: Speed-accident relationship on European roads* (Baruya 1998), assessed road geometry and other features of rural roads in Europe, however road data for the United Kingdom was limited to a small area in the South East, suggesting that a comprehensive methodology for the extraction of UK rural road geometry is required for a more comprehensive study.

Newer developments like the Speed Limit Appraisal Tool mean that automated and quantitatively informed speed limits may be imposed on rural roads. However, this tool does not take into account key features such as road geometry, and simply builds on existing speed data and accidents (Department for Transport 2013*b*).

2.1.3. Accessibility

Rural accessibility can be defined in terms of economic and social opportunity as the proximity or ability for spatial interaction (Gutiérrez 2009). In relation to transport specifically, accessibility may be defined as the ability or opportunities by which basic services can be reached by either public or private transportation (Gutiérrez 2009).

Journey time on rural roads is often a primary concern when considering a reduction in speed limits, often higher speed is perceived to bring with it shorter travel times, and lower accessibility for people and goods (Department for Transport 2013*a*). Despite this, there is evidence to support that traffic travelling at constant, and lower speeds may result in overall more reliable journey times, and the time saved by travelling at faster speed is often overestimated (Stradling et al. 2008).

Transport accessibility for rural communities is far more limited than for urban communities, where often rural areas have limited or no public transport, meaning there is a heavy reliance on personal transport (Gray et al. 2001). Accessibility in this context therefore in turn limits social mobility, influencing social exclusion as there is a reliance on owning expensive personal transport to improve accessibility (Gutiérrez 2009).

Despite this, there has been little focus on the improvement of transport technologies in rural areas, with the potential for new technologies already implemented into urban areas to improve rural accessibility (Velaga et al. 2012). A key area to address is the level of accessibility towards hospital services for rural communities, where recent centralisation of these services has negatively impacted the level of access for rural communities (Mungall 2005). This also impacts the level of access for hospital services to reach rural areas, where distance and time taken to a hospital directly correlates with a patients mortality (Nicholl et al. 2007). Emergency vehicles are often larger than personal vehicles and as such it can be assumed that accessibility for these types of services is often more limited depending on the quality of rural roads.

Velaga et al. (2012) suggest that technological innovations targeting rural areas may alleviate this problem through the production of user generated data. They note that certain challenges will come with this technology, notably their first point;

> "*Understanding basic technological infrastructure requirements in rural areas.*"

This suggests a greater understanding of the existing rural road infrastructure is required before the introduction of existing urban technologies may improve access in rural areas.

2.2. LiDAR Data Classification

Aerial LiDAR classification typically follows two objectives, the classification of ground and non-ground points, and the classification of surface objects, including buildings, trees or roads (Charaniya et al. 2004). Classification takes two forms, *supervised* and *unsupervised*, supervised classification taking a *training* dataset, using it to estimate the parameters associated with the outcome hoping to be classified. These parameters are then used on unknown data, with a similar distribution to the training set, and used to classify features (Charaniya et al. 2004).

2.2.1. Digital Terrain Models

Early LiDAR classification primarily focused on the production of digital terrain models (DTM). Kraus & Pfeifer (1998) used an iterative linear prediction algorithm which used residuals to compute weights for each LiDAR point. Ground points produced residuals with negative weighting, and vegetation was more likely to produce residuals with higher weighting. From this 47.8% of points were classified as vegetation. Kraus & Pfeifer (1998) note that advances in the technology, such as the inclusion of multiple laser pulses would enable for a higher quality DTM.

The observation of height textures are another method for DEM extraction enabled through the use of LiDAR data. Maas (1999) used height variation between neighbouring LiDAR points to classify buildings, trees, and flat terrain, with enough detail to determine whether a building had a flat or sloped roof. Similarly Elberink & Maas (2000) noted that man made structures often have smooth, regular height textures, with small variations in height, while trees and other vegetation give an irregular height pattern, and used this to separate man made structures from vegetation. The accuracy of their technique was assessed through a comparison with a set of known labelled features, and resulted in a 98% point accuracy for buildings, and 97% for trees, however ground based objects such as roads could only be detected with an accuracy of below 70%. Elberink & Maas (2000) suggest that multi-spectral data should be included to achieve better results.

This dissertation will utilise a recent method for DEM production in order to classify ground and non ground points for subsequent road classification. The method chosen was proposed by Zhang et al. (2016) using *cloth simulation* to generate a DTM from LiDAR data. This algorithm, unlike other filtering algorithms, allows for a simplistic input, without the need for numerous parameters to ensure an accurate DTM. This method consists of four main steps:

- *Initial State.* A simulated *cloth*[1] is placed above the inverted LiDAR measurements. A series of points that lie flat to the surface and are allowed to move based on the influence of gravity.
- The displacement of each LiDAR point is calculated under the influence of gravity, meaning some points appear below ground measurements.
- *Intersection check.* For any points detected as being under the ground, they are moved to ground level and set to be unmovable.
- *Considering internal forces.* Movable points are moved according to neighbouring points.

Quantitative accuracy assessment of this methodology gave results similar to top existing DTM production algorithms, but with a far more simplistic implementation.

[1]Used in 3D modelling, a simulation of particles with a mass, connected by a mesh, following Newton's Second Law: $\vec{F} = m\vec{a}$ (Karthikeyan & Rnaganatan 2001)

2.2.2. Feature Classification

Developments in LiDAR enabled the possibility of classification by using laser intensity information and multiple returns, features of more advanced LiDAR systems. The TopEye system used by Axelsson (1999) allowed for classification of buildings and electrical power-lines using reflectance to obtain radiometric information about the area and note that this can be used to separate paved area from grassland. Power lines in particular benefited from the multiple returns produced by the LiDAR system used as they often gave one return from the power line, and one from ground.

Vegetation in particular exhibits multiple returns, whereas most man made surface objects do not. By considering the number of returns and homogeneous height variation Hui et al. (2008) were able to categorise surface vegetation into both high vegetation, low vegetation as well as smooth man made surfaces.

2.3. Road Classification

In comparison to the extraction of vegetation and buildings from LiDAR, the extraction of roads poses far more of a challenge, due to there being less prominent height differences (Vosselman & Zhou 2009). Road classification is essentially a data clustering method to categorise data into road and non-road points, enabled through discovering patterns and relationship between variables and validation of findings (Saeedi et al. 2009). Clustering may be achieved through various algorithms, categorised generally into partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods (Saeedi et al. 2009). Yadav et al. (2018) note that the periodic assessment of roads is important due to the changing traffic load, which is generally increasing over time. New automated techniques will enable this in areas where in the past it had not been feasible. Due to the heterogeneous nature of certain roads types, the road environment is often complex, meaning collection and accurate processing of road features is challenging (Yadav et al. 2018).

2.3.1. Aerial Imagery

Road classification methodologies have historically used purely aerial imagery, providing only road pixels and 2D location information (Yadav et al. 2018). Ferchichi & Shengrui Wang (2005) used high resolution satellite imagery for road centreline extraction from a suburban area based on cluster coverage. They used image segmentation through maximum likelihood to assign pixels to road and non-road classes, making the assumption that both classes gave a Gaussian distribution. Features for classification include the angular second-movement, contrast, and entropy of the image, based on image texture analysis by Dubes & Ohanian (1992). Results of the classification gave significant noise, noted as a possibility due to non conformance with a Gaussian distribution. However, centreline extraction from the resultant classification removed any pixels with low density and a K-means clustering algorithm was ran iteratively to determine road clusters. Wan et al. (2007) produced an automated method for mapping urban and suburban roads using high resolution satellite imagery, using spectral, context, shape and the structural features or roads. Using fuzzy segmentation, classification of buildings, parking areas, and road clusters were obtained, but with large levels of noise. Angular texture analysis was able to reduce this noise, however large areas of incorrectly classified roads appear present in results. Additionally, the method proposed fails to address the differentiation between road and pavement, a key issue with the use of 2D imagery.

2.3.2. LiDAR Road Classification

The more accurate intensity data, and inclusion of multiple returns in LiDAR data enabled methods for categorising roads, and differentiating them from the surrounding ground, despite the similar elevation.

The majority of current road classification techniques using LiDAR have focused on unsupervised classification, often with the goal in vehicle automation using mobile LiDAR data (e.g. Yadav et al. 2018, Kumar et al. 2013, Smadja et al. 2010, Jaakkola et al. 2008). Applications using aerial LiDAR have also followed this trend for unsupervised classification. Clode et al. (2004) used a sequential Hough filtering transformation to classify roads using aerial LiDAR data, achieved by first taking only the last pulse LiDAR coordinates, considered a digital surface model (DSM)[2]. To obtain a digital terrain mode (DTM) from the DSM, Clode et al. (2004) used a method proposed by Rottensteiner et al. (2003), using a square structural element and grey scale opening to filter non-terrain objects. Clode et al. (2004) note that, at the time of production, intensity data with LiDAR had been often subject to large amounts of noise, and their use of intensity values for road detection was only a recent possibility. Road classification was subsequently achieved through both intensity and height filtering. Minimum and maximum intensity values for the road surface type (bitumen) were used to remove points that fell outside these values, and points that were outside a selected tolerance from the DTM were also removed. This methodology achieved a completeness of 86%, correctness of 69% and quality of 65%, quantified by the use of a ground truth model. This methodology provides a simplistic method for unsupervised road classification that reached similar levels of accuracy as more complex methods for the time, at the expense of certain assumptions, namely the road surface material.

Many subsequent methods for road classification from LiDAR follow similar patterns, first identification of ground points through height data, then DTM extraction through interpolation, followed by classification based on the attributes of the point cloud (Vosselman 2009). Due to the 3D information provided by LiDAR data, kerbs can be segmented from roads, achieving a more accurate road edge location when kerbs are present. However, due to the often small jump between kerb and road, this methodology is limited to new terrestrial mobile LiDAR data collection. Jaakkola et al. (2008) produced a methodology using mobile LiDAR data to classify road edges by segmenting kerbstones based on the height gradient along the scanned profile, with a final completeness of 73.9%, and correctness of 85.6%. Additionally, road markings were segmented from the road surface using intensity information. The mobile data collection for this study obtained a point resolution of between $10cm^2$ and $100cm^2$, with driving speeds of 20km/h to 50km/h. Jaakkola et al. (2008) note that due to the size of the data produced through this method, computation times were lengthy. Yoon & Crane (2009) similarly developed an approach for urban drivable terrain detection for automated vehicles through mobile LiDAR data, road edges were obtained through the slope and standard deviation of the height of points, finding the standard deviation to be far more efficient for edge detection. They note that while results were promising, it would take a very long time to collect data for a large number of roads, and at higher vehicle speeds LiDAR points density was reduced below a usable threshold.

Other road classification techniques rely on the detection of planar or smooth surfaces at ground level, indicative of man made objects which, unlike vegetation, do not display sharp variation in height values (Vosselman et al. 2004, Darmawati 2008).

2.3.3. Supervised Methods

Few supervised methodologies exist for LiDAR road extraction, Guan et al. (2013) used both aerial imagery and LiDAR data in combination with a training dataset to classify urban roads. The training dataset gave human labelled features, and used to train a maximum likelihood classification model. Matkan et al. (2014) extracted roads from LiDAR using a Support-vector Machine (SVM) classification. Training samples enabled classification into roads, trees, buildings, grassland and cement. Accuracy was determined through testing on three known road datasets, the number of LiDAR points correctly categorised ranged from 63% to 66% depending on the classification. Ferraz et al. (2016) used supervised random forest classification to detect

[2]Containing all surface objects, including ground points, vegetation, and man made structures such as buildings and roads.

large-scale forest roads using LiDAR. They note particularly that given the scale of these roads, the efficiency of road extraction is most important, managing to achieve 80% accuracy with individual roads processed at 2 minutes per kilometer. Despite being forest roads however, the canopy was often not obscuring roads due to their reasonably wide surface, and as such this method produced inaccuracies in areas where the canopy was dense and covering the detected road. Charaniya et al. (2004) trained a mixture of Gaussian models using key features of both LiDAR and aerial imagery data. They found that for classification of buildings and roads, the key features of LiDAR that enable extraction were the height, intensity, and the number of returns, in addition to luminescence information obtained from the aerial imagery. Results correctly categorised from 66% to 84% of LiDAR points when compared with a labelled dataset.

Hatger (2002) used 0.5m resolution LiDAR data to extract the properties of roads given a database of known road centreline locations using a simplification algorithm. Properties included height variation, curvature, and width. This paper used only LiDAR height information for road extraction and so produced a large amount of noise. To combat this, a RANSAC algorithm was used to determine the most likely position of the straight line road edges.

Elberink & Maas (2000) produced an automated method for modelling highway infrastructure using airborne LiDAR, 2D topographic information, and a database of known road locations. True road polygons were extracted using a seed-growing algorithm, and hough transformation.

These supervised techniques give insight into the feasibility for rural road extraction, given a dataset of known road centrelines. With improvements to the quality of LiDAR data more recently, a methodology for road feature extraction using road centrelines and LiDAR may more comprehensively include features that distinguish LiDAR from surrounding objects, including intensity, the number of returns, and aerial luminescence.

2.3.4. Rural Road Extraction

Many recent road extraction techniques have relied on kerb extraction or the detection of buildings, features uncommon to rural roads in the UK, where non road surfaces often include managed grass verges, with little height difference in comparison to roads (Kumar et al. 2013). Rural classification therefore must rely on alternative features of roads, notably the differences in intensity produced by vegetation, explored in various studies (e.g. Vosselman 2000). Additionally, overhanging tree canopies are uncommon features of roads that have been previously classified, and as such, a methodology for classifying rural British roads must take this limitation into into particular consideration.

2.4. Overview of this Paper

This dissertation aims to extract key features of a selection of rural roads in England through a combination of LiDAR point clouds, OS open road geometries, and aerial imagery, selecting road features considered to be important in past literature and government reviews. The key features considered when determining road quality are;

- **Width:** Narrower roads are associated with an increased number of accidents in many studies of rural roads (Taylor et al. 2002, Aarts & van Schagen 2006, Taylor et al. 2000)
- **Surface quality:** Poor quality road surfaces have been shown to increase the number of road traffic accidents (Fleming et al. 2009)
- **Blind corners/winding roads:** Blind corners increase the risk of accidents, and higher speeds mean stopping distances are often above the distance visible around rural road corners in the UK. Blind corners

are particularly an issue due to the tall hedgerows that often bank rural roads (Aarts & van Schagen 2006, Wu et al. 2013)

- **Road Steepness:** Steeper roads have an increased skid risk, and their quality is more likely to deteriorate (Moore et al. 2006, Viner et al. 2004)

While the focus of past road classification methods typically aim to classify the entire road surface, this isn't necessary for the feature extraction of roads as proposed in this dissertation. To extract roads widths, only LiDAR points at each edge of the road surface are required, and points along the road surface may be sampled at regular intervals, removing the majority of unnecessary LiDAR points and improving computation time, often a key limitation when working with LiDAR data (Zhang et al. 2018). Additionally, this paper aims to concentrate on a supervised classification of roads, by taking known centreline locations, many LiDAR points may be excluded, and remaining points concentrated towards the centre of the centrelines may be used as a training data set. Unlike road widths, other road features do not require classification with the inclusion of known road geometries. Road bends may be determined through the existing road linestrings, while surface quality and road steepness may be extracted at known road locations in the LiDAR point cloud. The method as presented aims to allow for the potential expansion beyond the dataset considered in this analysis, providing the requisite data is available.

3. Methodology

THIS dissertation primarily makes use of the free open source statistical language **R** (R Core Team 2019). Managing the large LiDAR datasets from my personal computer was mode possible through the `lidR` **R** package (Roussel & Auty 2019). Further details regarding the **R** environment and computer setup used for this dissertation is given in **Appendix A**. All content was written using LaTeX combined with the `rnoweb` file type (Ihaka 2011), for *Literate Programming*[1]. The template is built from scratch but takes much inspiration (and code) from the R-LaTeX-Template.

All code is hosted on my personal GitHub account, along with my my complete dotfiles, used in conjunction with the Linux distribution Manjaro, and the i3 window manager. All writing and code was produced using Neovim with my personal configuration to implement integrated development environment (IDE) style features for writing R code, while also providing essential features for writing in LaTeX. Neovim has the benefit of being both highly customisable, and lightweight, which allows for much lower system utilisation compared with R Studio when working with large datasets. One essential Vim plugin to mention is Nvim-R, providing an **R** REPL connection to vim, and other useful functions.

Given in **Appendix A** are the code snippets utilised in this methodology, for many equations, the relevant code is given as a reference to the appendix location, in the form **A.x.x**. Due to the nature of the functions used in this analysis, a single function often contains multiple equations, and so a reference to a particular appendix number may be repeated. Additionally, not all functions are referenced in text, so there are gaps in references to functions, and they do not necessarily appear sequentially.

3.1. Data

LiDAR point cloud data was downloaded through the UK Government's open data repository which uses the Open Government Licence, allowing for:

- Copying, publishing, distributing and transmission of the data
- Adaptation of the data
- Commercial and Non-commercial use of the information

LiDAR data used in this paper is available HERE under this licence (UK Government 2019*a*). This data was given as a compressed LAS file format (`.laz`), the **R** package `lidR` provided the function `lidr::catalog()` which enabled each separate `.laz` to be combined into one object of class `LAScatalog`. Analysis on this object could then be split into chunks (selected as 500m^2), allowing for multi-core threading to speed up analysis, and a reduction in the memory overhead when reading in data, often a limitation of the **R** language as objects are stored entirely into memory when read (Wickham 2014). The `LAScatalog` object did not

[1] See Knuth (1984); *"Let us change our traditional attitude to the construction of programs: Instead of imagining that our main task is to instruct a computer what to do, let us concentrate rather on explaining to human beings what we want a computer to do."*

require the compressed `.laz` files to be read into memory as `.las` files, meaning memory limitations were far less of a problem.

Aerial imagery was downloaded through Digimap® which uses the *Aerial Digimap Educational User Licence*, allowing for free use of the data for educational purposes (The University of Edinborough 2019).

Road centreline geometries were accessed through the Ordnance Survey Open Data repository which shares the Open Government licence. These were downloaded in the GeoPackage format (`.gpkg`) nationally and cropped to the extent of the LiDAR point cloud data.

3.2. LiDAR Preprocessing

The total number of LiDAR points used in this study is 9,419,272. All LIDAR data has a vertical accuracy of +/-15cm Root mean square error (RMSE). An overview of the LiDAR data selected for this study is given on Table 3.1. The variables of primary interest are:

- **z:** The distance a laser pulse is reflected back to to scanner, calculated by the time taken for a return pulse to be detected.
- **Intensity:** The amplitude of the return pulse, reflected back by the surface terrain or objects.
- **ReturnNumber:** A number of range 1-5, indicating for a point, the corresponding order of a reflected laser pulse. A ReturnNumber of 1 indicates the first return for a pulse (and highest z value), a return number of 5 indicates the last return (and lowest z value).
- **NumberOfReturns:** The number of return pulses for a single laser pulse (maximum of 5).
- **Classification:** A number given to a point indicating a specific numeric classification. Of interest in this study is a classification of 2, indicating a ground point. More information is given by ESRI (2019), which outlines numerical classifications for various vegetation types and man made structures.

3.2.1. Last Pulse

The LiDAR point cloud data used in this paper gives the values for 5 pulse returns. The canopy above roads may be excluded through ignoring early pulses (higher Z values), therefore only the last pulse values for any point are selected, this can be expressed as;

$$\mathbf{p}_i = (lpx, lpy, lpz, lpi), \tag{A.2.1}$$

where $\mathbf{p_i}$ is a single instance of a LiDAR point within the chosen point cloud, lpx is the last pulse x coordinate, lpy the last pulse y coordinate, lpz the last pulse z coordinate, and lpi the last pulse intensity value.

3.2.2. Normalisation

Ground points were classified using the Cloth Simulation Filtering (CSF) algorithm, as described in Zhang et al. (2016). Points were already classified in the data provided, however, as the classification technique was unknown, reclassification was considered necessary. The general implementation simulates the movements of a piece of cloth lying over the inverse of a point cloud, as the point cloud is flipped, the cloth settles beneath ground points, while covering points that lie separate to the ground, essentially forming a digital terrain model (DTM), cloth simulations are described in more detail in Bridson et al. (2005) and subsection 2.2.1. The CSF algorithm is given;

Table 3.1: *LiDAR Point Cloud Summary Data*

	Mean	SD	Min	Max
Z	80.58	5.97	64.85	115.79
Intensity	177.10	124.85	1.00	4064.00
ReturnNumber	1.47	0.95	1.00	5.00
NumberOfReturns	1.94	1.42	1.00	5.00
ScanDirectionFlag	0.50	0.50	0.00	1.00
EdgeOfFlightline	0.00	0.03	0.00	1.00
Classification	3.04	1.70	1.00	8.00
ScanAngleRank	-2.01	13.16	-22.00	22.00

$$X(t+\Delta t) = 2X(t) - X(t-\Delta t) + \frac{G}{m}\Delta t^2,$$

A.2.1

where m is the mass of a single LiDAR point (set to 1), Δt is the time step between points and G represents the gravity constant. The implementation of this algorithm was given as part of the `lidR` package. Reclassification resulted in an increase in the number of classified ground points by 50.77%. Reflecting primarily the simplification of classifications into ground and non-ground.

With the classification of ground points, (given $Classification = 2$), a full DTM may be produced through spatial interpolation of the classified points. Interpolation uses the inverse distance weighting and k nearest neighbours algorithms to produce the DTM. Nearest neighbours were selected as $k = 10$, with $q = 2$ for the inverse weighting, and used to produce a DTM with a resolution of 1m. This particular technique was selected over more comprehensive methods such as kriging as the number of points is very high, and the small benefit of kriging was considered minimal compared with the increase in computational load. The z values from the DTM were then subtracted from the LiDAR point cloud, leaving a normalised point cloud. This ensures that when extracting height information, any observed values are due to objects on the surface of the terrain, and not due to the lie of the terrain itself.

3.2.3. Points Extent

With the normalised last pulse point cloud, the point cloud was clipped to within a 30m extent of each known road location, using the OS road shapefiles;

$$\mathbf{p}_i \in A(r_i) \times 30,$$

A.2.5

where $A(r_i)$ are the geometric areas of each road in the study area. Selecting a 30m extent ensured that even with slight inaccuracy in road location, the road LiDAR points would likely not be excluded. A large number of unimportant points were therefore removed, saving on computational resources. Additionally this extent ensured that both road and non road points were included, but reduced the chance of false positives from occurring as fewer non road points were now included in the analysis.

3.2.4. Noise Filtering

Intensity noise was filtered through area based outlier detection, measuring the 95th percentile values within a $10m^2$ area, removing all points above the 95% percentile. This can be expressed as;

$$\mathbf{c}_k = \left(\mathbf{p}_i \in \left[\frac{95}{100} \times \mathbf{p}_{i_{lpi}}\right]\right)$$
$$A(\mathbf{c_k}) = 10m^2$$

A.2.4

where $\mathbf{c}_k$ represents a 10m^2 selection of LiDAR points, where each point ($\mathbf{p_i}$) has an intensity value within the 95% percentile intensity for all points ($\mathbf{p}_{i_{lpi}}$) within $\mathbf{c}_k$.

3.2.5. LiDAR Catalog

As mentioned in Section 3.1, objects of class LiDAR catalog enabled more efficient processing through allowing the LiDAR point cloud to be processed in predefined batch sizes. Considering a collection of processed LiDAR points, with last pulse, normalised, clipped to 30m road extents, and intensity noise filtered, points were grouped into 500m^2 areas;

$$\mathbf{C}_N = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k\}$$
$$A(\mathbf{C}_N) = 500m^2,$$

A.2.4

and each $500m^2$ area collectively represents the overall processed point cloud;

$$\mathbf{S} = \{\mathbf{C}_1, \mathbf{C}_2 \ldots, \mathbf{C}_N\}.$$

A.2.9

3.3. Road Analysis

Next, the preprocessed LiDAR data was combined with the OS road shapefiles and aerial imagery to obtain a set of criteria to assess the chosen road network. A summary of the information provided alongside OS road shapefiles is given on Table 3.2. While both `roadNumberTOID` and `roadNameTOID` do provide true road identification for many roads, this was not true for each road in the area chosen. Due to this, it was impossible to identify what could be considered an individual road, meaning a "road" will now be defined as indicated on Figure 1.1, selected based on the shapefile geometry provided.

The roads in this paper consist of these functions;

- B Road
- Minor Road
- Restricted Local Access Road
- Secondary Access Road
- Local Road

B roads are classified roads, while other functions are unclassified. All roads are single carriageway, and so for the purpose of this analysis it is assumed they likely have the default national speed limit of 60mph. All private roads were removed, as were roads with a length of less than 50m, often those clipped by the extent of the LiDAR data.

Table 3.2: *OS Roads Data Summary*

	Example
id	idE381337E-E88D-4232-8CAD-F543F178EBE4
endNode	id42B6F387-D838-445C-AA7A-6558362B7B9F
startNode	idC8EE8B4C-D965-436A-BA02-A0925A6EA1B8
roadNumberTOID	osgb4000000013398492
roadNameTOID	
fictitious	FALSE
roadClassification	B Road
roadFunction	B Road
formOfWay	Single Carriageway
length	241
loop	FALSE
primaryRoute	FALSE
trunkRoad	FALSE
roadClassificationNumber	B5392
name1	
name2	
roadStructure	

3.4. Road Angles

The angle of bends in roads were identified through the nodes produced in the creation of the road shapefiles. First the road linestrings were split into points, with coordinates representing each node within a road, a point at which the orientation of the linestring is altered.

The bend in a road was considered to be the *bearing angle* θ, from a node $\mathbf{N}_i = \left(i_x, i_y\right)$ to a node $\mathbf{N}_{i+1} = \left(i+1_x, i+1_y\right)$, with the angle measured in a clockwise direction from north. This is represented on Figure 3.1.

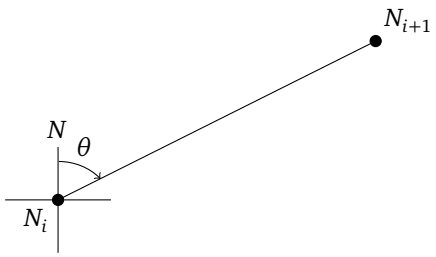

Fig. 3.1: *Bearing Angle Between Road Nodes*

To find the angle θ, the node N_{i+1} can be represented into relation to point N_i as;

$$\left(i+1_x, i+1_y\right) = \left(i_x + r\sin\theta, i_y + r\cos\theta\right)$$

where r is the length of the line segment N_iN_{i+1}. Rearranging the equation for θ gives;

$$\tan\theta = \frac{b_x - a_x}{b_y - a_y}$$

this equation can be rewritten to calculate the value of θ using the *atan2* function;

$$\hat{\theta} = \text{atan2}\left(b_1 - a_1, b_2 - a_2\right) \in [-\pi, \pi]$$

finally the bearing angle $\theta \in [0, 2\pi]$ is given as;

$$\theta = \begin{cases} \hat{\theta}, & \hat{\theta} \geq 0 \\ 2\pi + \hat{\theta}, & \hat{\theta} < 0 \end{cases}$$

A.2.21

with the bearing angle of the first line segment ($\hat{\theta}_{AB}$) for a particular road, the change in orientation of the second road segment BC may be given as $\theta_{BC} = \theta_{BC} - \theta_{AB}$, with additional nodes following the pattern $\theta_{n+1} = \theta_{n+1} - \theta_n$.

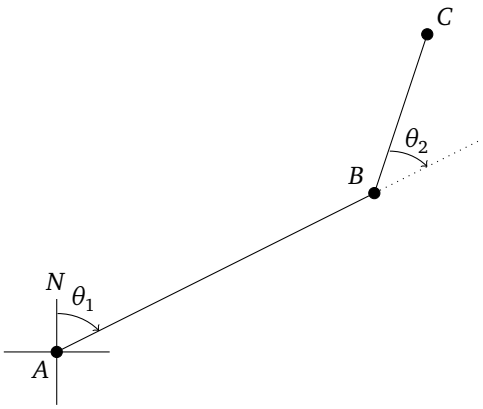

Fig. 3.2: *Bearing Angle Between two Road Segments*

For each road the maximum bearing angle between two nodes was selected, as well as the average bearing angle between two nodes. Given the bearing angle between the first two nodes gives only the initial direction of the road, this was set to zero; $\hat{\theta}_{AB} = 0$.

3.5. Road Node Elevation Change

The elevation change between two road node points was calculated by first selecting non-normalised LiDAR points at a geometric node within a $1m^2$ area. LiDAR points were then filtered by those only classified as ground, and with only a single return, to reduce the likelihood of inaccurate z values from canopy or other vegetation and vehicles. The mean z value of points were found for each node, and elevation change between each node was calculated. For each road, the total elevation change per kilometer was calculated.

/// add an equation + explanation here.

$$\mathbf{W}_n = \frac{\mathbf{W}}{W_e \times 100}$$

3.6. Surface Quality

Surface quality was assessed roughly through the range in intensity values found in each known road point, and the average number of returns for a road for sample lines not obstructed by canopy. To ensure there was no inaccuracy in intensity values caused by later returns passing through a canopy, only points that had a single return pulse were used in this analysis.

/// add an equation + explanation here.

$$\mathbf{W}_n = \frac{\mathbf{W}}{W_e \times 100}$$

3.6.1. Road Sampling

The LiDAR point cloud data was sampled at regular 10 meter intervals for each road, perpendicular to the road direction, ensuring that when road direction changed, the sampling locations remained perpendicular. To achieve this, each road was first split into nodes at which road direction changed, with a single road consisting of multiple nodes with xy coordinates, indicating a point along a road where the road direction changed. From this, points with xy coordinates were created at 10 meter intervals beginning at the start of a road, (considered Node 1; $\mathbf{N}_1$), until the next node along the road ($\mathbf{N}_2$). To calculate these points along each line between two nodes ($\mathbf{N}_i = (x_i, y_i)$ and $\mathbf{N}_{i+1} = (x_{i+1}, y_{i+1})$), first the individual change in x and y values was calculated, expressed as;

$$\begin{aligned} |x| &= x_{i+1} - x_i \\ |y| &= y_{i+1} - y_i, \end{aligned}$$

A.2.10

along with the euclidean distance between these nodes;

$$d(\mathbf{N}_i, \mathbf{N}_{i+1}) = \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2}.$$

A.2.7

A point (p_i) along these two nodes, at an interval distance I_k, was achieved through these equations;

$$\begin{aligned} px &= \frac{x_i + |x|}{d(\mathbf{N}_i, \mathbf{N}_{i+1})} \times I_k \\ py &= \frac{y_i + |y|}{d(\mathbf{N}_i, \mathbf{N}_{i+1})} \times I_k, \end{aligned}$$

A.2.10

giving a point $p_k = (px, py)$ at a distance I_k from N_i in the direction of N_{i+1}. Where I_k is the increment, which increases by 10m until the length of the node is covered, given $I_1 = 10, I_2 = 20, \ldots, I_k < d(\mathbf{N}_i, \mathbf{N}_{i+1})$ and $k \geq 2$. To create a perpendicular sample from a point p_k at position I_k between two nodes N_i and N_{i+1}, first two points at a perpendicular distance δ from the bearing angle between the two nodes, at a point p_k were created, with δ selected as 30m. First the euclidean distance from N_{i+1} to the point p_i was calculated;

$$d(N_{i+1}, p_i) = \sqrt{(x_{i+1} - px)^2 + (y_{i+1} - py)^2},$$

A.2.7

with this perpendicular distance, the value required for each xy coordinate to achieve a distance of δ from a point p_i may be calculated by;

$$\begin{aligned} \delta_x &= \frac{\delta}{d(N_{i+1}, p_i)} \times (x_{i+1} - px_i) \\ \delta_y &= \frac{\delta}{d(N_{i+1}, p_i)} \times (y_{i+1} - py_i), \end{aligned}$$

A.2.10

then to create perpendicular points at length δ from the point p_i, the value δ_y was added to the x value of the point p_i, while the value δ_y was subtracted from the y value of the point p_i. This was then inverted to produce a second point. This may be expressed as;

$$P_{perp} = \begin{cases} px_i + \delta_y, & py_i - \delta_x \\ px_i - \delta_y, & py_i + \delta_x \end{cases},$$

A.2.8

where P_{perp} is a collection of two points at distance δ from the point p_i. From these two points, a linestring was created between them, which was then buffered to 2m. This gave sample lines, with an area of 2m by 60m, at 10m intervals along each road. The total point cloud was then clipped to only include points with fell inside these sample lines:

$$\mathbf{S} = \mathbf{p}_i \in A(s_i)$$

A.2.3

3.6.2. Aerial Imagery

With the perpendicular sample lines extracted for the length of every road, to assist with the prediction of correct road locations, true colour aerial imagery was included. This imagery was first converted from three band RGB raster images, to a single-band grey-scale raster brick with values ranging 0 to 255. Combining the three bands into a single band produces a grey scale image, that more accurately portrays luminescence information from the aerial image, which has been included in past road classification methodologies.

$$lum = \frac{Band_1 + Band_2 + Band_3}{3}$$

A.2.12

3.7. Linear Probability Models and Road Width

For a supervised classification of roads, first the outcome variable "road" was estimated by classifying all points within a 2m buffer of the known road centrelines as road, and all points outside this as non-road. To classify road and non-road, linear models were constructed in relation to this outcome variable, and compared to assess effectiveness. A maximal approach was chosen, selecting all appropriate predictor variables, iterating through models by removing variables that did not significantly influence the model outcome, or created noise.

In addition to the variables provided by the LiDAR and aerial data, the variable $Dist$ was created and included, representing the shortest distance from a point to the centreline of the road it is associated with, considering that road points should be weighted more towards points that are closer to the centre-point of the road.

Linear probability models essentially follow the same formula as a linear regression model:

$$Y_i = \beta_0 + \beta_1 + X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$$

but given a binary outcome variable Y_i, this is considered to be a linear probability model, taking the form;

$$E(Y|X_1, X_2, \ldots, X_k) = P(Y = 1|X_1, X_2, \ldots, X_3)$$

where;

$$P(Y = 1|X_1, X_2, \ldots, X_k) = \beta_0 + \beta_1 + X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}$$

β_j therefore may be interpreted as the change in the probability that $Y_i = 1$, with all other predictor variable constant. β_j may be estimated using Ordinary Least Squares regression (Hanck et al. 2019).

Likelihood values from the predictions gave a range of numerical values. Points that fell below a certain threshold were removed, leaving only points that were most likely correctly identified as road points. This threshold was assessed qualitatively through both observation of the distribution of probability ranges for each model, and results gained through different thresholds. Considering a threshold x, this may be expressed as;

$$\mathbf{S} = \left(\mathbf{p}_i \in \left[\frac{x}{100} \times \mathbf{p}_{i_{lm}}\right]\right),$$

A.2.12

where $\mathbf{S}$ is the total point cloud.

Further qualitative assessment of the results revealed that some points considered to be noise were still present, but often isolated. To ensure no isolated points were present, the minimum distance between each point, and the nearest neighbouring point was checked, if a single point was considered isolated, with over 1m between it and any other point, it was removed. This may be expressed as;

$$D = \sqrt{\delta x i^2 + \delta y^2}$$
$$\mathbf{S} = (\mathbf{p}_i \in [D \leq 1]),$$

A.2.14

given D is the minimum distance between a point and any other point.

$\mathbf{S}$ now gave of a collection of predicted road points for each sample line along a road segment, with noise removed. To obtain road widths from these points, the maximum distance between two points in a particular sample was determined, these points were kept and all others removed. A linear section of road with two samples resembles Figure 3.3.

A.2.15

Given the two selected points at every sample with a maximum distance between them, trigonometry could be used to determine the width of the road at that particular location. Road width is considered to be the adjacent line length ($b_{\mathbf{K}}$; where $\mathbf{K}$ is a single sample location), perpendicular to the road segment, considering the distance between the two points to be the hypotenuse ($|\mathbf{K}_1\mathbf{K}_2| \equiv h_{\mathbf{K}}$) of a right angled triangle (Figure 3.3). The average width for each road identifier was then found, for a sample location $\mathbf{K}$, this may be expressed using trigonometry as;

$$b_{\mathbf{K}} = |\mathbf{K}_1\mathbf{K}_2| \times cos(\theta_{\mathbf{K}})$$

A.2.19

where $b_{\mathbf{K}}$ gives the true predicted of a road for the sample $\mathbf{K}$. With a complete set of calculated road widths for each sample, any width above 8m was removed, in addition to any width below 2m, under the assumption that a width calculated outside these limits would be caused due to noise or inaccuracy.

There is the possibility that the maximum distance between two points does not provide the maximum perpendicular distance across a road section. Such a situation would arise given a triangle formed that is has an opposite length ($a_{\mathbf{K}}$) above the adjacent length ($b_{\mathbf{K}}$), giving a hypotenuse ($h_{\mathbf{K}}$) with a longer vertical length. However, given the maximum opposite line length between two points in a sample line is 2m, (each sample is 2m by 60m), for any line where the opposite length is greater than the adjacent length, the adjacent length must therefore be below 2m and thus, the road width calculated from this sample is removed during the filtering process.

3.7.1. Improved Road Centrelines

During the analysis of the linear models, it was noted that road centrelines were often inaccurate, giving road outcome values that were not representative of the road surface. In an attempt to adjust for this, road centrelines were adjusted based on the centre location of road points in each sample, classified through an initial linear probability model. The mid point between two points (x_1, y_1) and (x_2, y_2) can be expressed as;

$$\left(\frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2}\right)$$

A.2.17

With the mid point of each classified road sample, these were than joined for form new centrelines, and a further linear probability model was ran and compared.

Given the improvement in road centreline locations, it was considered feasible to construct linear probability models individually for each sample location on each road. This technique would allow for per road variation in material type or quality, and potentially reduce the amount of noise brought in from inaccurate centrelines. Preliminary testing of this method revealed that it was essential to remove any sample containing tree canopy, as given the small number of points, the predictions were based off points that misrepresented true road points. While points are generally able to penetrate the canopy, the intensity values produces by these points was reduced, and removed the distinction between road intensity values, and vegetation. This analysis of the LiDAR data is described in more detail in chapter 4. Additionally, due to the reduced number of points, it was considered feasible to filter out points that gave model p values below 0.05. For each individual model, if the p value of any predictor or the outcome variable road was above 0.05, the sample location it was associated with was removed. This may be expressed as;

$$\mathbf{S} = \left(s_i \in \left[p_s < 0.05\right]\right).$$

A.2.12

3.7.2. Final Model Analysis

To aid with model interpretability, the direct comparison between each variable in the analysis was enabled through centering and scaling through the use of beta coefficients (Peterson & Brown 2005);

$$\beta_p = \frac{\mathrm{Cov}\left(r_p, r_b\right)}{\mathrm{Var}\left(r_b\right)}$$

A.2.24

3.7.3. Estimate of True Widths

QGIS (QGIS Development Team 2019) was used to measure the width at various points along each road using 25cm resolution aerial imagery, avoiding stretches of road with canopy cover that obscured the true road width. With the widths, the results of each model was compared to assess model accuracy. Each width was normalised to allow comparison between each road, and to give a final average accuracy value. Normalisation was achieved through finding the relative difference in width as a percentage;

$$\mathbf{W}_n = \frac{\mathbf{W}}{\mathbf{W}_e \times 100},$$

where $\mathbf{W}_n$ is the normalised width, $\mathbf{W}$ is the average width per road derived from the linear model, and $\mathbf{W}_e$ is the qualitatively estimated width. Given some widths occasionally were overestimated, to ensure the outcome of this calculation gave a relative value, any normalised width given a value above 100 was reassigned;

$$\mathbf{W}_n = 100 - \mathbf{W}_n,$$

given $\mathbf{W}_n > 100$.

3.8. Road Quality Assessment

To provide a method for direct comparison between each road, the extracted features are normalised and combined as one to produce the Road Quality Index (RQI).

Normalisation of each road feature was achieved through a simple range normalisation;

$$m \mapsto \frac{m - r_{\min}}{r_{\max} - r_{\min}},$$

where $r_{\min}$ denotes the minimum of the range of a variable, $r_{\max}$ denotes the maximum of the range of a variable, and $m \in [\, r_{\min}\, , r_{\max}\,]$ denotes the variable to be scaled. As an increase in road width is associated with a higher quality road, as opposed to larger values of each other variable indicating a poorer quality road, the width values were first inversed before normalisation. An additional variable, relitability was presented in addition to the RQI, which gives a value for the number of points per road length, (P_n/L), allowing for some information regarding the density of sample points to be considered in analysis.

Following normalisation, the sum of all normalised variables for a particular road were taken, and subtracted from 1 to give positive values indicating better quality roads, and lower values indicating lower quality roads. These variables involved in the creation of this index are;

- **Road Width:** Extracted through linear probability classification of the road surface using LiDAR data (Section 3.7)
- **Road Angles:** The bearing angle change in road direction, considering the initial road direction as a bearing angle of 0° (Section 3.4)
- **Road Elevation Change:** The change in elevation for each node in a road, extracted from LiDAR data, giving a total elevation change for a full road (Section 3.5)
- **Road Surface Quality:** The total range of intensity values for each node in a road, extracted from the LiDAR point cloud for points that did not return more than a single pulse (Section 3.6)

The value obtained from this is referred to as the Road Quality Index, presented in full on Table 4.3.

4. Results

RESULTS for the overall methodology are given in this chapter, covering the initial preprocessing of LiDAR and other data, and following onto the width extraction of roads, in addition to other geometric features. The primary goal is to critique the effectiveness of the proposed methodology, and provide a baseline for future improvements particularly in road classification and width extraction, while presenting the quantifiable results in a way that relates to the overall quality of each road. Outlined in detail therefore is sensitivity analysis of the road classification models, presenting both qualitative and quantitative assessments of accuracy. Assessment of the improvement to road centreline locations is also covered, before a detailed look at the final results of the analysis, demonstrating how road feature extraction may inform the overall quality of a road, comparing the extracted data to aerial imagery for a visual assessment of the results.

As noted in Section 2.4, computation time is considered an important aspect of this analysis. The total time taken, including all data preprocessing, perpendicular sample line extraction, LiDAR sample extraction, construction of linear models, reprocessing of road centrelines, road feature extraction, and further analysis is 15.98 minutes.

4.1. Data Preprocessing

Table 3.1 indicates that there are likely some points with noise, particularly reflected by the highest intensity value (4064) relative to its standard deviation (125), with 99% of observations within the range of 0 to 417. As noted in previous LiDAR classification methods, intensity is often subject to noise, therefore a simplistic noise exclusion technique (Roussel & Auty 2019) was considered, as described in Section 3.2.4.

```
Error in as.list.environment(x, all.names = TRUE): cannot open file '/home/cjber/drive/uni/envs492/main/data/cache/c
No such file or directory

Error in grid_canopy(las_lp, res = 2, p2r()):  cannot open file '/home/cjber/drive/uni/envs492/main/data/cache/unnam
No such file or directory
```

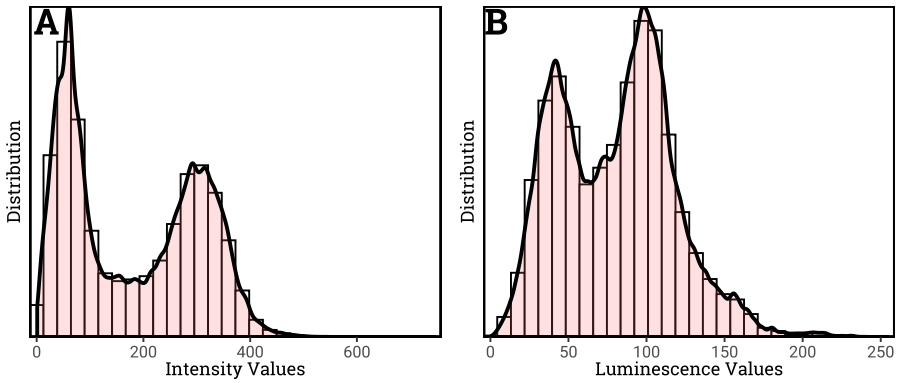

Fig. 4.1: *Post noise filtering LiDAR point cloud distribution of;* ***(A)*** *Intensity, and* ***(B)*** *Luminescence*

```
Error in .class1(object):  cannot open file '/home/cjber/drive/uni/envs492/main/data/cache/unnamed-chunk-6_ff6c72158
No such file or directory

Error in as.data.frame(chm_lp):  cannot open file '/home/cjber/drive/uni/envs492/main/data/cache/unnamed-chunk-6_ff6
No such file or directory

Error in colnames(chm_lp) <- c("value", "x", "y"):  cannot open file '/home/cjber/drive/uni/envs492/main/data/cache/
No such file or directory

Error in lasground(las_lp, csf()):  cannot open file '/home/cjber/drive/uni/envs492/main/data/cache/unnamed-chunk-6_
No such file or directory

Error in grid_terrain(las_lp, 2, knnidw(k = 10, p = 2)):  cannot open file '/home/cjber/drive/uni/envs492/main/data/
No such file or directory

Error in lasnormalize(las_lp, dsm_lp):  cannot open file '/home/cjber/drive/uni/envs492/main/data/cache/unnamed-chun
No such file or directory

Error in grid_canopy(norm_lp, res = 2, p2r()):  cannot open file '/home/cjber/drive/uni/envs492/main/data/cache/unna
No such file or directory

Error in .class1(object):  cannot open file '/home/cjber/drive/uni/envs492/main/data/cache/unnamed-chunk-6_ff6c72158
No such file or directory

Error in as.data.frame(dsm_lp):  cannot open file '/home/cjber/drive/uni/envs492/main/data/cache/unnamed-chunk-6_ff6
No such file or directory

Error in colnames(dsm_lp) <- c("value", "x", "y"):  cannot open file '/home/cjber/drive/uni/envs492/main/data/cache/
No such file or directory

Error in grid_canopy(norm_lp, res = 2, p2r()):  cannot open file '/home/cjber/drive/uni/envs492/main/data/cache/unna
No such file or directory

Error in .class1(object):  cannot open file '/home/cjber/drive/uni/envs492/main/data/cache/unnamed-chunk-6_ff6c72158
No such file or directory

Error in as.data.frame(las_lp):  cannot open file '/home/cjber/drive/uni/envs492/main/data/cache/unnamed-chunk-6_ff6
No such file or directory

Error in colnames(las_lp) <- c("value", "x", "y"):  cannot open file '/home/cjber/drive/uni/envs492/main/data/cache/
No such file or directory

Error in grid_metrics(norm_lp, ~myMetrics(Z, Intensity), res = 2):  cannot open file '/home/cjber/drive/uni/envs492/
No such file or directory

Error in .class1(object):  cannot open file '/home/cjber/drive/uni/envs492/main/data/cache/unnamed-chunk-6_ff6c72158
No such file or directory

Error in as.data.frame(metrics):  cannot open file '/home/cjber/drive/uni/envs492/main/data/cache/unnamed-chunk-6_ff
No such file or directory

Error in colnames(metrics) <- c("zmean", "imean", "x", "y"):  cannot open file '/home/cjber/drive/uni/envs492/main/d
No such file or directory

Error in fortify(data):  cannot open file '/home/cjber/drive/uni/envs492/main/data/cache/unnamed-chunk-6_ff6c72158dc
No such file or directory

Error in fortify(data):  cannot open file '/home/cjber/drive/uni/envs492/main/data/cache/unnamed-chunk-6_ff6c72158dc
No such file or directory
```

```
Error in plot_grid(chm_gg, dsm_lp_gg, las_i_gg, aerial_gg, labels = "AUTO", :  lazy-load database '/home/cjber/drive
is corrupt
```

Following intensity noise filtering, the highest intensity value was now 746, with a mean of 182. Figure 4.1 **(A)** gives the distribution of Intensity values for all points within the study area, showing two clear spikes in intensity, at at value of around 50, with another around 350. This is reflected similarity in the Luminescence values, with two peaks at around 50 and 120 (Figure 4.1 **(B)**).

Figure **??** gives the results of further LiDAR preprocessing, comparing Figure **?? (A)** and Figure **?? (B)**, shows how last pulse LiDAR filtering allows for the removal of the majority of tree canopies, leaving only ground points that are considered hard surfaces, and as such are the lowest point the laser pulse has penetrated. Additionally, Figure **?? (B)** shows how a digital terrain model, created through interpolation techniques, using only the base point cloud may be used to normalise the points, giving a digital surface model which only shows the true height of surface objects, without having to consider the variation in lie of the land. However, Figure **?? (C)** indicates that while filtering for last pulse returns may appear to remove much of the canopy, reflected in the z values, the intensity values for points that have penetrated the canopy are lower than those that did not (See the tree just below the centre of the road section). This particularly creates issues in the distinction between road and non road in neighbouring areas where the intensity "shadow" created removes the distinct difference in intensity. This suggests that for ground points with multiple returns, the intensity values are likely far less reliable for road classification. Quantitative analysis of this limitation reveals that for ground classified points with a single return the average intensity value is 214.9, while for ground classified points with multiple returns, the average intensity value is 88.15.

4.2. Perpendicular Sampling

Using the 30m buffer from known road locations, and sample line extraction, the number of points from the original LiDAR point cloud for the 1km^2 area was reduced from $9,419,272$ to $616,015$ giving a reduction in number of points by 93.46%. Additionally, including sample lines allowed for filtering based on features of each sample, allowing for samples fully obscured by canopy to be identified through the number of returns, and excluded easily if required. See **Appendix B**, Figure B.1 for an overview of all the sample lines produced in this analysis.

4.3. Linear Probability Model Sensitivity Analysis

Selected based on literature, and correlation analysis of the variables (Table B.1), the first model was constructed to include all variables of importance from the LiDAR point cloud, z, Intensity, and Number of Returns. Additionally, luminescence from aerial imagery was included, and the minimum distance of a point from the known road centreline location.

This first (maximal) model was constructed as;

$$\begin{aligned} \text{Road}_t = \alpha &+ \beta_1 \text{Intensity}_t \\ &+ \beta_2 \text{Luminescence}_t \\ &+ \beta_3 \text{Z}_t \\ &+ \beta_4 \text{NumberOfReturns}_t \\ &+ \beta_5 \text{Dist}_t + \epsilon \end{aligned} \tag{4.1}$$

As proposed in Chapter 3, Section 3.7, the road outcome variable was given as points that fell within a 2m buffer of the known road centreline locations. As such, this meant that a fair number of false negative points are expected to have occurred, where points outside 2m of a road centreline location would be incorrectly classified as non-road. Due to this, the classification of non-road and road was not a simple selection of points that were above a 50% threshold prediction as being road. To determine an appropriate cutoff for road predictions a histogram was produced which gave insight into the distribution of the linear prediction values (Figure 4.2).

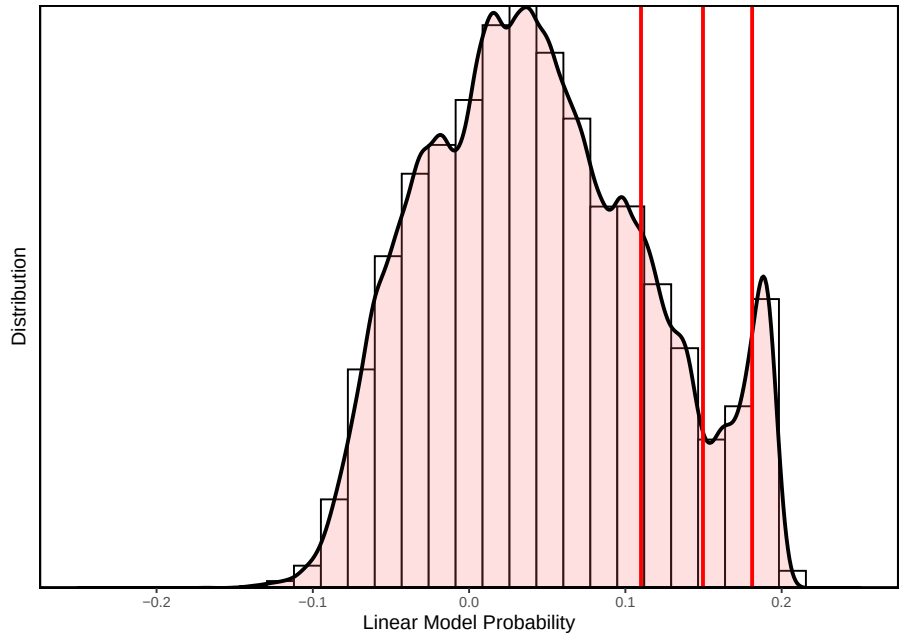

Fig. 4.2: *Linear Model Probability Distributions for the Maximal Model, showing vertical lines at the 95th, 90th, and 80th quantile of the distribution*

Figure 4.2 shows that there is a clear separation between the majority of points, and higher probability values. This therefore gives insight into the true divide between true road and non-road points, allowing for a qualitative analysis to select the most appropriate quantile of probability values. Three quantiles were chosen, the 95th, 90th and 80th, as indicated on Figure 4.2.

Figure 4.3 reveals that qualitatively, the optimal choice for a quantile filtering of the linear probability distribution is likely the 95th quantile (Figure 4.3 **(A)**). However, observation of the southern section of Figure 4.3 **(A)** reveals that inaccurate centreline locations have led to an incomplete linear model analysis. To compensate for this, a further method proposed aims to improve the accuracy of the given road centreline locations. Additionally, Figure 4.2 reveals that for the 95th quantile probability values, shadow from road hedgerows appears to reduce the model accuracy, as noticeable towards the top end of the road. For this reason, a second model was constructed for later comparison, which removes the *luminescence* information provided by the aerial imagery.

$$\begin{aligned} \text{Road}_t = \alpha &+ \beta_1 \text{Intensity}_t \\ &+ \beta_2 \text{Z}_t \\ &+ \beta_3 \text{NumberOfReturns}_t \\ &+ \beta_4 \text{Dist}_t + \epsilon \end{aligned} \tag{4.2}$$

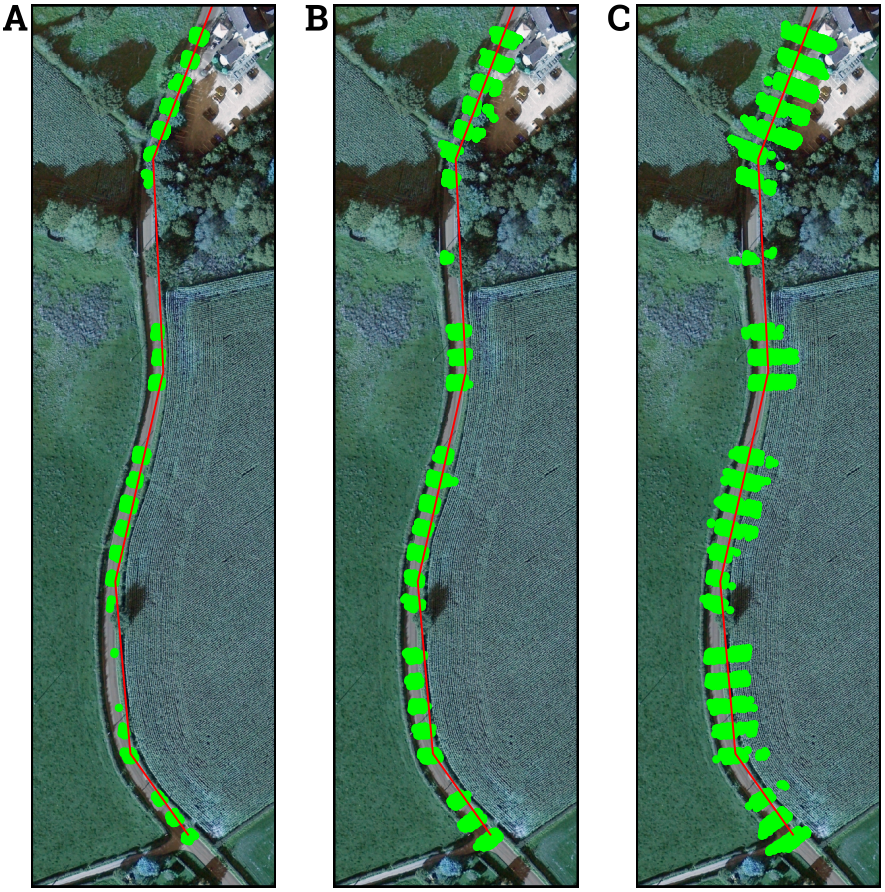

Fig. 4.3: *Comparison between Linear Prediction Quantiles;* ***(A)*** *95th Quantile,* ***(B)*** *90th Quantile,* ***(C)*** *80th Qunatile*

4.4. Corrected Centreline Extraction

To improve road centreline location accuracy, the 90th quantile results from the first linear probability analysis were used, due to there being a more complete selection of points, but without compromising the true location of roads by including too many outside points.

New road centrelines are given on Figure 4.4 **(A)**. Particular improvements are given where the road curves between two open fields, but the original centreline was given as a straight line, covering the hedgerow, and no road surface.

Qualitative comparison between the Linear Probability Model based off the original centreline locations reveals an improvement in overall road detection, particularly towards the edge of roads, while additional samples are achieved in areas which previously had no coverage due to the incorrect centreline placement (Figure 4.4 **(A)**). However, it appears that in areas where there are higher levels of linear predictive inaccuracy, the new centrelines are less accurate. Thankfully, noise exclusion techniques employed have removed samples that fall within these areas, particularly noticeable at the northern end of Figure 4.4 **(C)**. Figure 4.4 also gives information regarding the distance based noise exclusion technique, which has allowed for the exclusion of isolated points accurately on Figure 4.4 **(B)**. Improved centreline locations allowed for individual linear models (Figure 4.4 **(D)**). While it was assumed that individual linear models would potentially produce more accurate width estimations, it is hard to differentiate between the global and individual linear models (Figure 4.4 **(C)** and **(D)**).

4.5. Final Model Analysis

For direct comparison between the two selected global linear probability models, centering and scaling of the predictor variables allowed for an easier interpretation of results, without affecting any statistical inferences. This was considered necessary as both P values and standard errors produced by global models offered little in terms of interpretability due to the very large number of points involved in this study. Centering and scaling was obtained through the production of beta coefficients with results given on Table 4.1. While the removal of $luminescence$ has had little effect on the other predictor coefficients, due to the very small influence of this coefficient, reflected by the normalised value (0.01), and the qualitative analysis of the issues due to shadows, is was considered an unnecessary addition. The other coefficients all give insight into their influence of the road outcome, for example for every 1 increase in the standard deviation in $dist$, the likelihood a point is to be a road point decreases by a standard deviation of 0.33. This therefore suggests that the inclusion of the $dist$ coefficient is important, despite not considered in other supervised road detection techniques.

Table 4.2 gives a normalised comparison between each linear model, and its associated estimated road width. This gives insight into the effectiveness of various linear probability models for each road, and road type. While average values all give relative accuracy in the region of 70%, it appears likely that without adjusting the road centrelines, model probabilities are reduced given **LM 0** gives the lowest average accuracy in relation to the other three models using improved centrelines for outcome variable creation. Unexpectedly, the roads with the highest accuracy are Local and Minor Roads, rather than B roads. For certain roads, the accuracy of the derived centrelines appear to vastly reduce model accuracy, (For example Road 14; Table 4.2

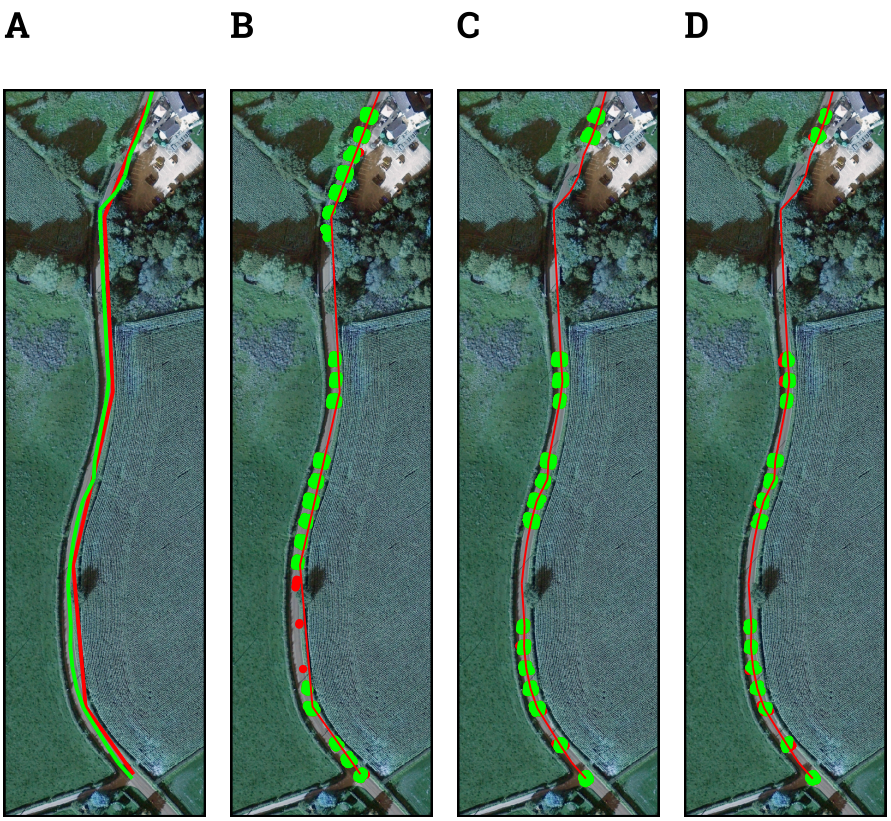

Fig. 4.4: *Comparison between Linear Probability models applied to; (A) OS Road Centrelines, and (B) Derived Centrelines. Green points give classified road locations, red show classified road locations removed through isolation based filtering*

and Figure 4.5).

4.6. Road Assessment

Given Linear Model 2 with corrected road centrelines appears to produce the most accurate estimates for road widths, this was used in the final road assessment. Table 4.3 gives the full results of the road geometric extraction, along with an estimate of overall road quality given by the Road Quality Index (RQI). Qualitative assessment of the RQI may be achieved through observation of the highest and lowest values (Figure 4.5). It appears to produce reliable results, as the road with the highest RQI is straight, wider than the road with the lowest RQI, and is likely flat given it is neighboured by two agricultural fields. Additionally, this figure gives both the original, and derived centrelines for both roads. Notably the reduction in accuracy in the centreline for road 14 has greatly reduced the accuracy of the road width estimate (Table 4.2)

Table 4.1: *Model Coefficients, Comparison between Linear Probability Model 1 and 2*

Variable	LM 1	LM 2
Intensity	-0.30	-0.30
dists	-0.33	-0.33
Z	-0.07	-0.07
NumberOfReturns	-0.22	-0.22
lum	0.01	

Table 4.2: *Comparison between different linear probability models for each individual road, in relation to estimated true road width*

ID	Class	LM *i*	LM 1	LM 2	LM 0
1	B Road	67.86	66.55	70.47	64.10
4	Minor Road	73.77	87.62	84.58	61.90
6	Minor Road	64.93	81.88	75.97	80.59
7	B Road	76.50	59.95	61.13	76.61
8	Minor Road	75.71	73.61	74.95	85.43
9	Minor Road	82.76	70.92	81.73	80.13
10	Minor Road	83.55	96.37	95.77	61.43
12	Minor Road	82.74	88.24	89.02	67.28
13	B Road	53.62	53.66	52.34	51.97
14	Local Road	53.60	56.81	64.28	98.88
15	B Road	61.63	91.10	86.31	76.50
16	B Road	61.23	40.34	59.83	42.91
18	Minor Road	59.53	73.73	58.16	71.71
20	Local Road	91.04	84.95	72.05	58.86
21	B Road	48.68	51.08	50.51	38.17
30	Minor Road	87.48	78.25	90.22	50.53
31	Minor Road	60.75	61.11	63.96	70.55
33	Minor Road	71.40	67.01	69.11	61.29
34	Minor Road	67.31	81.14	80.26	75.91
36	Minor Road	87.43	79.20	85.97	90.87
Means:		**70.58**	**72.18**	**73.33**	**68.28**

LM *i*: Individual Linear Models
LM 1: Linear Model inc. lum
LM 2: Linear Model ex. lum
LM 0: Linear Model 1 with original road centrelines

Table 4.3: *Overall Features Extracted from Roads in the Study Area, in descending order by RQI value*

Road ID	Road Function	Max Angle	Total Z	Intensity	Width (LM 2)	Reliability (P_n/L)	RQI
34	Minor Road	0.00	2.82	40	4.87	14.19	0.89
15	B Road	0.01	7.49	62	4.99	0.93	0.60
20	Local Road	0.00	1.26	319	4.95	3.90	0.18
16	B Road	0.00	7.55	167	3.88	1.04	0.00
1	B Road	0.00	3.02	281	3.91	1.26	-0.08
21	B Road	0.01	2.41	155	2.89	2.02	-0.14
9	Minor Road	17.15	7.31	64	4.07	0.69	-0.19
18	Minor Road	9.51	3.84	101	2.95	0.46	-0.34
36	Minor Road	14.58	6.95	125	3.79	0.84	-0.36
13	B Road	0.00	3.62	313	3.14	0.74	-0.52
7	B Road	0.06	7.10	299	3.39	3.83	-0.55
4	Minor Road	12.81	7.29	382	4.73	4.33	-0.79
10	Minor Road	17.95	12.08	277	4.85	4.97	-0.89
30	Minor Road	24.53	4.93	252	3.74	3.63	-0.96
8	Minor Road	7.08	10.63	346	3.73	0.38	-0.97
31	Minor Road	10.33	11.04	251	3.33	0.53	-0.99
33	Minor Road	18.16	11.33	309	3.52	4.08	-1.35
12	Minor Road	18.51	15.85	370	4.17	2.03	-1.55
6	Minor Road	30.38	14.03	384	4.22	2.55	-1.87
14	Local Road	16.66	19.91	233	2.52	0.37	-2.11

Fig. 4.5: *Visual comparison between road with the highest RQI **(A)**, and lowest RQI **(B)**, original centrelines are given in red, with derived centrelines given in green*

5. Discussion

THE proposed methodology provides a road classification method which considers the need for an optimised and efficient extraction of road widths to combine with other road features for use in an automated national rural road assessment. This method builds upon past classification methodologies with the inclusion of Ordinance Survey road centreline geometries allowing for a supervised classification, without the need for the manual and time consuming creation of a training dataset.

While considered to be a supervised method, the *training* data used in this method is derived from the preexisting OS road centrelines, and as such may be recreated automatically for any road in England, given the comprehensive coverage of this data (Ordnance Survey 2019). This differs from existing supervised road classification methods, where a training dataset is created and manually labelled (e.g. Charaniya et al. 2004). Additionally, the majority of LiDAR road classification techniques focus on unsupervised methods (e.g. Clode et al. 2004, Vosselman 2009, Jaakkola et al. 2008, Darmawati 2008), and as such do not consider the use of any training data, with the primary goal of obtaining accurate road locations. Such techniques often limit their scope to road centreline extraction which has a limited use case with a preexisting national road centreline database. Notably, Zhang et al. (2018) assess the accuracy of their technique by comparing derived road centrelines to an existing road centreline database which renders little in terms of practical application.

Therefore, the method presented in this dissertation considers a more practical approach, providing an intermediate between unsupervised and supervised methods, which integrates the known road centrelines into the road classification. Rather than attempting to classify road centrelines, or the full road surface, this dissertation concentrates on the requirement for road feature extraction as presented in various UK government rural road studies (Department for Transport 2018*a*, 2013*b*, Road Safety and Environment 2000). With the primary focus of road classification on the automated road width extraction. Focusing on solely width extraction enables road surface sampling, primarily enabling a reduction in computational overhead, while adding the benefit of more simplistic noise filtering techniques.

5.1. Effectiveness of the Method

5.1.1. Computational Efficiency

Direct comparison between the computational efficiency of this method and past methods is not possible due to the unique data used in this study. However, Zhang et al. (2016) outline some specifications of the dataset used in their supervised road centreline extraction using LiDAR and aerial imagery. With aerial imagery at a resolution of 15cm, and a total 5200 by 5000 pixels, the study area therefore covers 780m by 750m. Similar to the 1km^2 for this study. The total time taken for object extraction for this area in their method was 37.87 minutes, compared with the 15.98 minutes proposed in this paper. It is expected that the method proposed by Zhang et al. (2016) should take far longer to compute due to being the inclusion of complex methodologies such as image segmentation through random forest classification.

5.1.2. Comparison with Similar Studies

Despite several road classification methodologies proposing the inclusion of aerial imagery to assist with accuracy (e.g. Charaniya et al. 2004, Hui et al. 2008, Guan et al. 2013), this study reveals that for roads overlooked by tall features such as hedgerows, the shadow created reduces overall classification accuracy. Additionally, the inclusion of aerial data provided little benefit in areas without shadows, likely due to the more distinct separation in intensity values from road surface and surrounding vegetation, unlike that found in a more urban setting.

Due to the irregularity of LiDAR and the large number of points, many past road classification techniques have relied on regulating the data into a grid (Hatger 2005). For example, Clode et al. (2004) used LiDAR with a resolution of 0.8m, and regularised this into a grid to produce a DTM and subsequently extract road centrelines. Due to this aggregation, they were able to filter points through a density threshold, and produce road centrelines. However, using this method to find road widths proved more difficult, and as Hatger (2005) note, the function that derived road widths in this paper resulted in some ambiguity.

The focus of many road classification techniques are primarily directed towards either centreline extraction (Clode et al. 2004, Zhang et al. 2018, Matkan et al. 2014), or the use of ground based LiDAR for use in automated vehicles (Jaakkola et al. 2008, Yoon & Crane 2009), and almost all studies appear to focus on urban road classification (Li et al. 2016, Vosselman 2009, Zhao & You 2012), while even studies considering "rural" areas, do not represent roads that would be found in the context of rural England (Azizi et al. 2014, Mena & Malpica 2005), and exclude key features such as hedgerows, overhanging vegetation, and appear fully distinct from surrounding areas. Additionally these studies do not focus primarily on road width extraction, and as such are limited by the requirement for the inclusion of all points of data, to obtain a full road extraction. The methodology presented has a clear focus on the road feature extraction, without the requirement for road centreline extraction which is already accounted for nationally. Due to this, the novel technique for extraction allows for both sampling of the point cloud, leading to far lower computational requirements, while enabling simplistic filtering for noise and width extraction.

The study of the speed accident relationship on rural British roads by Taylor et al. (2002) outlined some techniques for the extraction of rural British road features, and as such was able to begin an assessment for the classification of rural road hierarchies. However, the techniques employed did not allow for a scalable approach, using drive through video recordings, and often lacked in accuracy, taking the road height variation from OS 10m contour lines. The method proposed here aims to alleviate these problems by ensuring a higher level of accuracy in height variation, through the use of LiDAR data with a +/-25cm RMSE. As well as allowing for a computational technique that does not rely on the manual collection of road level data, and instead uses aerial LiDAR which is more practically feasible to obtain for a comprehensive study, given the multiple use cases LiDAR has.

5.2. Applications of this Methodology

5.2.1. Stopping Sight Distance

Stopping Sight Distances are an important consideration for rural British roads. From qualitative observation of aerial imagery, and personal knowledge, hedgerows that bank the verges either side of many rural British roads often fully obscure the sight line around sharp bends, meaning it is often impossible to see oncoming traffic or obstacles, which, given the nature of these roads can often be large farm vehicles which spill into multiple lanes, or hazards such as farm animals, or unsafe road conditions. It is worth mentioning that the majority of hedgerows have automatic protection under the Hedgerow Regulations 1997 (UK Government

Table 5.1: *Recommended minimum Stopping Sight Distances at certain speeds (Layton & Dixon 2012)*

Speed	**Stopping Sight Distance**		**Typical Emergency Stopping Distance (m)**	
Design Speed (km/h)	Calculated (2.5^s, a=3.4m/s^2)	Design (2.5^s, a)	Wet Pavement (1^s, f_{wet})	Dry Pavement (1^s, f_{dry})
30	31.2	35	17.1	14.2
40	46.2	50	27.7	21.6
50	63.5	65	42.0	30.3
60	83.0	85	59.6	40.3
70	104.9	105	81.7	51.6
80	129.0	130	106.1	64.2
90	155.5	160	131.2	78.1
100	184.2	185	163.4	93.4
110	215.3	220	200.6	110.0
120	248.6	250	235.7	127.9

1997) for numerous historic and environmental reasons (e.g. protected species; UK Government 1981), as such, their removal for road safety is rarely granted.

Stopping Sight Distance is defined as the ability to see an object in the roadway with enough distance to stop, 5.1 outlines the calculated stopping sight distances at certain speeds, giving a rough indication of the distance required between a car and bend in a road. For example, 5.1 indicates that at 100kph (60mph), stopping sight distance is recommended to be 185m. Broadly, for a road to be considered appropriate for a 60mph limit, it could be said that it should not have a bend which impairs the line of sight within 185m such a speed limit. For a rough idea of the number of bends per road (ignoring bend sharpness), there is an additional table given in **Appendix B**, Table B.2.

Assumptions for certain road regulations are made that drivers will slow to appropriate speeds to adapt to road conditions, either in poor weather, or to approach a sharp bend, however Layton & Dixon (2012) note that often this is not that case, and drivers often do not slow appreciably to account for these conditions. Therefore, the consideration that speed limits should more accurately reflect the conditions of the road. Additionally, stopping sight distances observed by Layton & Dixon (2012) are significantly longer for larger vehicles such as trucks, given the large farm vehicles often present on country roads, speed policy should take this into account.

Road features extracted in this dissertation may be used to inform the current likely stopping sight distances, combining key features such as the width of roads, which influences the sight line, the max bend angle within a road segment, and the elevation change. For information regarding the calculated number of bends per kilometer in this study, see Table B.2.

5.2.2. Improving Rural Transport Accessibility

Transport disadvantage is a key limitation of transport accessibility, that may be due to lack of public transport, a poor road network, or a persons physical inability to reach a destination due to disability (Smith et al. 2012). Often transport disadvantage may be alleviated through access to public transport, as this removes the requirement for private transport ownership, limited by both income and ability to drive. However, public transport in rural areas is often limited or absent, meaning rural transport predominantly relies on private road vehicles, limiting access for those two are unable to drive, such as children, the elderly, and people with disabilities (Manthorpe et al. 2008). Additionally, this reliance of private transport increases the minimum cost of living in rural communities, given car ownership is often considered mandatory and often isn't taken into account when assessing the minimum cost of living in rural areas (Smith et al. 2012). For those who are unable to access private transport, this limited accessibility is considered limited through capability, rather than pure accessibility through journey times and other factors (Currie 2010). Rural areas in particular often

have larger elderly populations, meaning capability is often a key issue in areas with poor public transport, and can lead to social exclusion for those without cars (Solomon & Titheridge 2009).

Transport accessibility is defined by the UK Government through journey time estimates for populations to particular key services. In particular, the UK Government uses official accessibility indicators to set minimum thresholds for journey time access to education, health services, employment and retail hubs (Department for Transport 2016), also taking into account the availability of public transport services. Accessibility in rural areas is found to be far poorer than urban areas based off minimum travel times to various services, and while travel by car generally reduces travel times, the rate is still far below urban areas. This journey time data is simplified, giving the start point of journeys as a single point within Output Area census units, and aggregated road speeds. The output of this data is given at the LSOA level which is then used for accessibility analysis. Journey times are obtained through mass collection of GPS data by INRIX (INRIX 2019) which is then used in TRACC software (TRACC 2019).

Due to the limited use of rural roads, the GPS data obtained for speed estimates are likely far less reliable than for urban areas. Additionally, calculating journey time in rural areas should consider the road geometry, which the above method ignores. By considering features such as road width, quality, bends, and elevation change, the suitability of certain roads for particular vehicles types may be informed. For example, rural healthcare accessibility is becoming more of an issue given fewer healthcare professionals now live in rural communities, (Farmer et al. 2003), and the urban centralisation of hospital services (Mungall 2005) means that understanding the level of access that each rural community has to these services is more important than ever. Emergency vehicles are often far larger than personal transportation, meaning GPS times derived from smaller personal transportation likely does not provide an accurate journey time estimate for these vehicles types. Aggregation and simplification of journey time estimates also do not provide a comprehensive estimate of the true journey times for each rural road, which would be achievable through the road feature extraction presented in this dissertation. The individual rural road features may provide further insight into specific roads which require targeted improvements to alleviate accessibility problems, without suffering from inherent geographical limitations such as the MAUP (Openshaw 1984) imposed by journey time aggregation.

5.2.3. New Forms of Public Transport

There is a strong urban bias for the development of new transport technologies (Malecki 2003), explained through key issues particular to rural transport systems;

- **Service area:** Rural transport agencies often serve large areas with long trips. As a results, assisting passengers needs is not easy and attending immediately to a problem that arises on the road is difficult (e.g. rescheduling trips when an incident occurs.)
- **Service Coordination:** There are different basic public services e.g. healthcare and education with overlapping areas of services. It is challenging to coordinate services and resources among the agencies and other providers.
- **Infrastructure:** Rural areas suffer a lack of communication infrastructure e.g. wireless communications services, real-time communication from and to rural passengers.
- **Fleet size:** Although tech can solve several transportation problems in remote rural areas, it might be difficult to fund and develop at a small scale.

(Riva et al. 2011)

While it may appear that many of these issues are inherent to rural areas, and unsolvable, the optimisation of transport technologies for rural areas may be made more achievable through access to the comprehensive

road data provided through the methods proposed in this dissertation. Palmer et al. (2004) state that flexible integrated transport services are a likely public transport implementation that would benefit rural areas without the limitations outlined above, such a technology would rely extensively on a full understanding of the road network on which it would be dispatched. The Department for Transport (2016) call for "Unconventional modes" of public transport in such areas, building mainly on a bottom up approach to meet direct demand. Additionally, vehicles supplied through such an implementation would account for suitability to both road conditions and consumer demand, allowing for vehicles smaller than a typical bus for example on narrower roads (Mulley & Nelson 2009). There is also a significant call for the inclusion of a more comprehensive understanding of the road network through advanced computational techniques to improve the efficiency and quality of existing transport systems (Deeter 2009), including a more flexible transport management system (Robinson 2008).

5.2.4. Other Applications

Supply chains rely on a well maintained road infrastructure, and as such, many rural areas are considered to be (economically) "Lagging Rural Regions", due to their geographical remoteness, poor infrastructures, low population density and limited employment opportunities, often supported economically by an agricultural backbone (Ilbery et al. 2004). Improving the economy of such areas therefore relies primarily on the effectiveness of the supply chains, often limited due to the poor infrastructures (Marsden et al. 2002), and recent demand for large scale supply chains is limited in these areas due to the overall quality of the road network. To further understand the limitations of rural supply chains inherently relies on a full understanding of the road network, whether to acknowledge where limitations exist, or to develop opportunities for optimisation of the supply chains. Similarly, Bosona & Gebresenbet (2011) call for location analysis of supply chains through quantifiable data, to better optimise supply routes.

5.3. Current Limitations of this Method

As results have revealed, it is relatively hard to quantify the results of this particular linear probability analysis, and often the accuracy of results has been assessed qualitatively. The ultimate goal with this method would be to produce a model which may be assessed quantitatively, allowing for a more conclusive and full automation. It should also be noted that the quantile selection at present is based on a qualitative observation of the distribution of linear probability values for the current $1km^2$ area, and as such it would be essential to find a method to quantitatively assess the cutoff for road and non-road points in order for this method to be used with other roads.

At present the removal of noise at the final stage of the road classification comes from both identifying isolated points (See Function A.2.16), and the removal of calculated widths that are above 8m and below 2m C. While logically it makes sense to include limitations for widths, given a road below 2m would not support even single way traffic, and a 8m road is unexpected for any rural single carriageway, these limitations are still arbitrary, and for all unclassified roads in England, there is no minimum width required (Highways England 2016).

Alternatives to Linear Probability models do exist when considering binary outcome variables, one being probabilistic regression, which takes the cumulative standard normal distribution function (Φ) to model the regression. Interpretability of results may be aided through this method as it includes consideration of the quantiles associated with a unit change in outcome variables. Additionally logistic (and probabilistic) regression, unlike simple linear regression do not take the assumption that there is a linear distribution in the outcome, weighting values more towards 1 or 0, conforming more with the distribution of a binary outcome variable (See Figure 5.1; Hanck et al. 2019). However, preliminary analysis of the methodology performed

in this study did consider a logistic regression, but found that interpretability of the quantiles and results was difficult, and qualitative observation of the results did not appear to provide much benefit over linear regression.

Alternative methods to reduce the noise produced by this method may include the identifications of straight line road edges, a distinguishing feature of man made structures Guan et al. (2013). This could enable point exclusion if outside of a particular threshold in relation to other points. LiDAR classification techniques often make use of segmentation algorithms to identify objects from geometric features such as planes, and straight edges (Wang & Shan 2009). These include Hough transforms (Vosselman et al. 2004, Clode et al. 2004), RANSAC (Smadja et al. 2010, Sampath & Shan 2008), and least square fitting (Matkan et al. 2014).

Driving behaviour is affected by particular features that are not covered in this methodology, namely the delineation of road centrelines and edges (Steyvers & De Waard 2000, Charlton et al. 2018). The detection of these features is influenced by both the quality of the road, if a road edge is easily detectable, or if a road has painted markings to indicate a centreline, both of which are not observable through this methodology, and would likely rely on mobile LiDAR data collection.

An assumption is made that roads are a constant width, however, the roads observed in this study were relatively short, with an average road length of 236.86, as such, it is unlikely that width changes significantly over this generally. The current inaccuracy of this method however relies on the aggregation of road sampled widths to provide an overall mean for a road, and as such is unable to provide accurate information regarding a road at any given point.

Additionally Hatger (2005) note that given a road network is an interconnected grid, there should be more emphasis on bridging gaps between road segments that are unable to be identified. While this would introduce some assumptions, it would be worth considering the possibility of producing a result that connects all samples to form a true representation of road geometries.

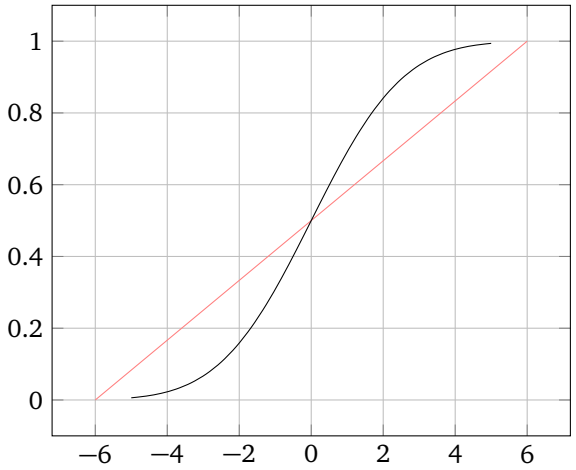

Fig. 5.1: *Comparison between a Linear Probability Model Distribution (Red) and Probit/Logit Cumulative Standard Normal Distribution (Black) (Approximation credit Bowling et al. 2009))*

6. Conclusion

This dissertation presents a methodology for the extraction of road features to enable a fuller understanding of the rural British road network. The method presented considers the requirement for supervised classification of roads to determine road width, that utilises existing OS open road data which is freely available. Concentrating on the practical applications of determining road width enables the use of a sampling classification, which reduces computational load, and enables sample based filtering.

Bibliography

Aarts, L. & van Schagen, I. (2006), 'Driving speed and the risk of road crashes: A review', *Accident Analysis & Prevention* **38**(2), 215–224. ZSCC: 0000994.

Aquino, J. (2019), *nvimcom: Intermediate the Communication Between R and Either Neovim or Vim*. R package version 0.9-83.
URL: *https://github.com/jalvesaq/nvimcom*

Arnold, J. B. (2019), *ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'*. R package version 4.2.0.
URL: *https://CRAN.R-project.org/package=ggthemes*

Axelsson, P. (1999), 'Processing of laser scanner data—algorithms and applications', *ISPRS Journal of Photogrammetry and Remote Sensing* **54**(2-3), 138–147. ZSCC: 0000940.

Azizi, Z., Najafi, A. & Sadeghian, S. (2014), 'Forest Road Detection Using LiDAR Data', *Journal of Forestry Research* **25**(4), 975–980. ZSCC: 0000030.

Bache, S. M. & Wickham, H. (2014), *magrittr: A Forward-Pipe Operator for R*. R package version 1.5.
URL: *https://CRAN.R-project.org/package=magrittr*

Baruya, A. (1998), 'MASTER: Speed-accident relationship on European roads'. ZSCC: 0000014.

BBC (2012), 'Plan for 40mph country road limit', *BBC News* . ZSCC: 0000000[s1].

Bengtsson, H. (2019), *future: Unified Parallel and Distributed Processing in R for Everyone*. R package version 1.14.0.
URL: *https://CRAN.R-project.org/package=future*

Bivand, R., Keitt, T. & Rowlingson, B. (2019), *rgdal: Bindings for the 'Geospatial' Data Abstraction Library*. R package version 1.4-4.
URL: *https://CRAN.R-project.org/package=rgdal*

Bosona, T. & Gebresenbet, G. (2011), 'Cluster building and logistics network integration of local food supply chain', *Biosystems Engineering* **108**(4), 293–302. ZSCC: 0000139.

Bowling, S. R., Khasawneh, M. T., Kaewkuekool, S. & Cho, B. R. (2009), 'A logistic approximation to the cumulative normal distribution', *Journal of Industrial Engineering and Management* **2**(1), 114–127. ZSCC: 0000087.

Bridson, R., Marino, S. & Fedkiw, R. (2005), Simulation of clothing with folds and wrinkles, *in* 'ACM SIGGRAPH 2005 Courses on - SIGGRAPH '05', ACM Press, Los Angeles, California, p. 3. ZSCC: 0000550.

Charaniya, A., Manduchi, R. & Lodha, S. (2004), Supervised Parametric Classification of Aerial LiDAR Data, *in* '2004 Conference on Computer Vision and Pattern Recognition Workshop', IEEE, Washington, DC, USA, pp. 30–30. ZSCC: 0000166.

Charlton, S. G., Starkey, N. J. & Malhotra, N. (2018), 'Using road markings as a continuous cue for speed choice', *Accident Analysis & Prevention* **117**, 288–297. ZSCC: 0000005.

Clode, S., Kootsookos, P. & Rottensteiner, F. (2004), 'The Automatic Extractin of Roads from LiDAR Data', p. 7. ZSCC: NoCitationData[s1].

Corben, B., Oxley, J., Koppel, S. & Johnston, I. (2005), 'Cost-effective measures to improve crash and injury risk at rural intersections.', p. 10. ZSCC: 0000004.

Currie, G. (2010), 'Quantifying spatial gaps in public transport supply based on social needs', *Journal of Transport Geography* **18**(1), 31–41.

Darmawati, A. (2008), Utilization of multiple echo information for classification of airborne laser scanning data, ITC. ZSCC: 0000013.

Deeter, D. (2009), 'Real-time traveler information systems. NCHRP report 399', *Transport Research Board, USA* . ZSCC: 0000004.

Department for Transport (2006), 'Speed Assessment Framework', http://www2.dft.gov.uk/. ZSCC: NoCitationData[s0].

Department for Transport (2011), 'Strategic framework for road safety'. ZSCC: 0000006.

Department for Transport (2012), 'Guidance on road classification and the primary route network', p. 26. ZSCC: 0000005.

Department for Transport (2013*a*), 'Setting local speed limits', p. 42. ZSCC: 0000004[s1].

Department for Transport (2013*b*), 'The Speed Limit Appraisal Tool: User Guidance', p. 93. ZSCC: 0000000[s1].

Department for Transport (2016), 'Overall measure of accessibility of services'. ZSCC: NoCitationData[s0].

Department for Transport (2018*a*), 'Road Conditions in England 2017', p. 8. ZSCC: NoCitationData[s0].

Department for Transport (2018*b*), 'Road Safety Management Capacity Review'. ZSCC: NoCitationData[s1].

Department for Transport (2019), 'Road traffic statistics - Summary statistics', https://roadtraffic.dft.gov.uk/summary. ZSCC: 0000000[s1].

Dowle, M. & Srinivasan, A. (2019), *data.table: Extension of 'data.frame'*. R package version 1.12.2.
URL: *https://CRAN.R-project.org/package=data.table*

Dubes, R. & Ohanian, P. (1992), 'Performance evaluation for four classes of textural features', *Pattern Recognition* **25**, 819–833. ZSCC: 0000550.

Elberink, S. O. & Maas, H.-G. (2000), 'The Use of Anisotropic Height Texture Measures for the Segmentation of Airborne Laser Scanner Data', p. 8. ZSCC: 0000116.

Environment Agency (2019), 'LiDAR', https://data.gov.uk/dataset/977a4ca4-1759-4f26-baa7-b566bd7ca7bf/lidar-point-cloud. ZSCC: NoCitationData[s1].

ESRI (2019), 'Lidar point classification—Help | ArcGIS Desktop', http://desktop.arcgis.com/en/arcmap/10.3/manage-data/las-dataset/lidar-point-classification.htm.

Farmer, J., Lauder, W., Richards, H. & Sharkey, S. (2003), 'Dr. John has gone: Assessing health professionals' contribution to remote rural community sustainability in the UK', *Social Science & Medicine* **57**(4), 673–686. ZSCC: 0000092.

Ferchichi, S. & Shengrui Wang (2005), Optimization of cluster coverage for road centre-line extraction in high resolution satellite images, *in* 'IEEE International Conference on Image Processing 2005', Vol. 2, pp. II–201. ZSCC: 0000010.

Ferraz, A., Mallet, C. & Chehata, N. (2016), 'Large-scale road detection in forested mountainous areas using airborne topographic lidar data', *ISPRS Journal of Photogrammetry and Remote Sensing* **112**, 23–36. ZSCC: 0000038.

Finch, D., Kompfner, P., Lockwood, C. & Maycock, G. (1994), 'Speed, Speed Limits and Accidents', https://trl.co.uk/sites/default/files/PR058.pdf. ZSCC: 0000234.

Fleming, P., Frost, M. & Lambert, J. (2009), Lightweight deflectometers for quality assurance in road construction, *in* 'IN: Tutumluer, E. and Al-Qadi, IL (Eds). Bearing Capacity of Roads, Railways and Airfields: Proceedings of the 8th International Conference (BCR2A'09)', Taylor & Francis Group, pp. 809–818. ZSCC: 0000017.

Francois, R. (2017), *bibtex: Bibtex Parser*. R package version 0.4.2.
URL: *https://CRAN.R-project.org/package=bibtex*

Garnier, S. (2018*a*), *viridis: Default Color Maps from 'matplotlib'*. R package version 0.5.1.
URL: *https://CRAN.R-project.org/package=viridis*

Garnier, S. (2018*b*), *viridisLite: Default Color Maps from 'matplotlib' (Lite Version)*. R package version 0.3.0.
URL: *https://CRAN.R-project.org/package=viridisLite*

Gillespie, C. (2019), *benchmarkme: Crowd Sourced System Benchmarks*. R package version 1.0.2.
URL: *https://CRAN.R-project.org/package=benchmarkme*

Gray, D., Farrington, J., Shaw, J., Martin, S. & Roberts, D. (2001), 'Car dependence in rural Scotland: Transport policy, devolution and the impact of the fuel duty escalator', *Journal of Rural Studies* **17**(1), 113–125. 00075.

Guan, H., Ji, Z., Zhong, L., Li, J. & Ren, Q. (2013), 'Partially supervised hierarchical classification for urban features from lidar data with aerial imagery', *International Journal of Remote Sensing* **34**(1), 190–210. ZSCC: 0000020.

Gutiérrez, J. (2009), 'Transport and accessibility'. ZSCC: 0000038.

Hanck, C., Arnold, M., Gerber, A. & Schmelzer, M. (2019), 'Introduction to Econometrics with R', p. 392. ZSCC: NoCitationData[s0].

Harrell Jr, F. E., with contributions from Charles Dupont & many others. (2019), *Hmisc: Harrell Miscellaneous*. R package version 4.2-0.
URL: *https://CRAN.R-project.org/package=Hmisc*

Hatger, C. (2002), 'On the use of Airborne Laser Scanning Data to Verify and Enrich Road Netwirk Features', p. 6. ZSCC: 0000013.

Hatger, C. (2005), 'Road extraction by use of airborne laser scanner data', p. 19. ZSCC: 0000001.

Highways England (2016), 'Letter in response to road width restrictions. FOI: 734,857'.

Highways England (2019), 'Network management'. ZSCC: 0000000.

Hijmans, R. J. (2019), *raster: Geographic Data Analysis and Modeling*. R package version 3.0-2.
URL: *https://CRAN.R-project.org/package=raster*

Hu, X., Tao, C. V. & Hu, Y. (2004), 'Automatic Road Extraction from Dense Urban Area by Integrated Processing of Height Resolution Imagery and LiDAR Data', p. 5. ZSCC: 0000149.

Hui, L., Di, L., Xianfeng, H. & Deren, L. (2008), Laser Intensity Used in Classification of Lidar Point Cloud Data, *in* 'IGARSS 2008 - 2008 IEEE International Geoscience and Remote Sensing Symposium', IEEE, Boston, MA, USA, pp. II–1140–II–1143. ZSCC: 0000016.

Ihaka, R. (2011), 'Rnoweb: Literate Programming with and for R', p. 5. ZSCC: 0000000.

Ilbery, B., Maye, D., Kneafsey, M., Jenkins, T. & Walkley, C. (2004), 'Forecasting food supply chain developments in lagging rural regions: Evidence from the UK', *Journal of Rural Studies* **20**(3), 331–344. ZSCC: 0000148.

INRIX (2019), 'INRIX', http://inrix.com/. ZSCC: NoCitationData[s0].

Jaakkola, A., Hyyppä, J., Hyyppä, H. & Kukko, A. (2008), 'Retrieval Algorithms for Road Surface Modelling Using Laser-Based Mobile Mapping', *Sensors* **8**(9), 5238–5249. ZSCC: 0000162.

Karthikeyan, P. & Rnaganatan, P. (2001), 'Tutorial on cloth modelling', *ACM Student Tutorial Contest, India* . ZSCC: 0000004.

Kassambara, A. (2019), *ggpubr: 'ggplot2' Based Publication Ready Plots*. R package version 0.2.3.
URL: *https://CRAN.R-project.org/package=ggpubr*

Knuth, D. E. (1984), 'Literate programming', *The Computer Journal* **27**(2), 97–111. ZSCC: 0002155.

Kraus, K. & Pfeifer, N. (1998), 'Determination of terrain models in wooded areas with airborne laser scanner data', *ISPRS Journal of Photogrammetry and Remote Sensing* **53**(4), 193–203. ZSCC: 0001405.

Kumar, P., McElhinney, C. P., Lewis, P. & McCarthy, T. (2013), 'An automated algorithm for extracting road edges from terrestrial mobile LiDAR data', *ISPRS Journal of Photogrammetry and Remote Sensing* **85**, 44–55. ZSCC: 0000088.

Layton, R. & Dixon, K. (2012), 'Stopping sight distance', *Kiewit Center for Infrastructure and Transportation, Oregon Department of Transportation* . ZSCC: 0000037.

Leutner, B., Horning, N. & Schwalb-Willmann, J. (2019), *RStoolbox: Tools for Remote Sensing Data Analysis*. R package version 0.2.6.
URL: *https://CRAN.R-project.org/package=RStoolbox*

Li, Y., Hu, X., Guan, H. & Liu, P. (2016), 'An Efficient Method for Automatic Road Extraction Based on Multiple Features from LiDAR Data', *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **XLI-B3**, 289–293.

Maas, H.-G. (1999), 'The Potential of Height Texture Measures for the Segmentation of Airborne Laserscanner Data', p. 8. ZSCC: 0000134.

Mahmoudian, M. (2018), *varhandle: Functions for Robust Variable Handling*. R package version 2.0.3.
URL: *https://CRAN.R-project.org/package=varhandle*

Malecki, E. (2003), 'Digital Development in Rural Areas: Potentials and Pitfalls', *Journal of Rural Studies* **19**, 201–214.

Manthorpe, J., Iliffe, S., Clough, R., Cornes, M., Bright, L., Moriarty, J. & Older People Researching Social Issues (2008), 'Elderly people's perspectives on health and well-being in rural communities in England: Findings from the evaluation of the National Service Framework for Older People', *Health & social care in the community* **16**(5), 460–468. ZSCC: NoCitationData[s0].

Marsden, T., Banks, J. & Bristow, G. (2002), 'The social management of rural nature: Understanding agrarian-based rural development', *Environment and planning A* **34**(5), 809–825. ZSCC: 0000228.

Matkan, A. A., Hajeb, M. & Sadeghian, S. (2014), 'Road Extraction from Lidar Data Using Support Vector Machine Classification', *Photogrammetric Engineering & Remote Sensing* **80**(5), 409–422. ZSCC: 0000022.

Mena, J. & Malpica, J. (2005), 'An automatic method for road extraction in rural and semi-urban areas starting from high resolution satellite imagery', *Pattern Recognition Letters* **26**(9), 1201–1220. 00195.

Moore, R., Carey, J., Mills, A., Martin, S., Irinder, S., Kerry, L., Leask, G., Simmons, A. & Ashaari, M. (2006), Recent landslide impacts on the UK Scottish road network: Investigation into the mechanisms, causes and management of landslide risk, *in* 'Proceedings of the International Conference on Slopes, Kuala Lumpur, Malaysia', Public Works Department, Kuala Lumpur, Malaysia, pp. 223–237. ZSCC: NoCitationData[s0].

Mulley, C. & Nelson, J. D. (2009), 'Flexible transport services: A new market opportunity for public transport', *Research in Transportation Economics* **25**(1), 39–45. ZSCC: 0000133.

Mungall, I. (2005), 'Trend towards centralisation of hospital services,'. 00000.

Nicholl, J., West, J., Goodacre, S. & Turner, J. (2007), 'The relationship between distance to hospital and patient mortality in emergencies: An observational study', *Emergency Medicine Journal* **24**(9), 665–668. 00162.

Noctor, I. (2004), 'Change to kph limit to cost €30m', https://www.irishtimes.com/life-and-style/motors/change-to-kph-limit-to-cost-30m-1.1133469. ZSCC: NoCitationData[s0].

Openshaw, S. (1984), 'The modifiable areal unit problem', *Concepts and techniques in modern geography* .

Ordnance Survey (2019), 'OS Open Roads', p. 28.

Palmer, K., Dessouky, M. & Abdelmaguid, T. (2004), 'Impacts of management practices and advanced technologies on demand responsive transit systems', *Transportation Research Part A: Policy and Practice* **38**(7), 495–509.

Pebesma, E. (2018), 'Simple Features for R: Standardized Support for Spatial Vector Data', *The R Journal* **10**(1), 439–446.
URL: *https://doi.org/10.32614/RJ-2018-009*

Peterson, B. G. & Carl, P. (2019), *PerformanceAnalytics: Econometric Tools for Performance and Risk Analysis*. R package version 1.5.3.
URL: *https://CRAN.R-project.org/package=PerformanceAnalytics*

Peterson, R. A. & Brown, S. P. (2005), 'On the Use of Beta Coefficients in Meta-Analysis.', *Journal of Applied Psychology* **90**(1), 175–181.

QGIS Development Team (2019), *QGIS Geographic Information System*, Open Source Geospatial Foundation.
URL: *http://qgis.org*

Qiu, Y. & authors/contributors of the included software. See file AUTHORS for details. (2019), *showtext: Using Fonts More Easily in R Graphs*. R package version 0.7.
URL: *https://CRAN.R-project.org/package=showtext*

R Core Team (2019), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Ram, K. & Wickham, H. (2018), *wesanderson: A Wes Anderson Palette Generator*. R package version 0.3.6.
URL: *https://CRAN.R-project.org/package=wesanderson*

Richards, D. & Cuerden, R. (2009), 'The Relationship between Speed and Car Driver Injury Severity', p. 16. ZSCC: 0000045.

Riva, M., Curtis, S. & Norman, P. (2011), 'Residential mobility within England and urban–rural inequalities in mortality', *Social Science & Medicine* **73**(12), 1698–1706. ZSCC: 0000050.

Road Safety and Environment (2000), 'New directions in speed management: A review of policy'. ZSCC: 0000002.

Robinson, D. & Hayes, A. (2019), *broom: Convert Statistical Analysis Objects into Tidy Tibbles*. R package version 0.5.2.
URL: *https://CRAN.R-project.org/package=broom*

Robinson, S. P. (2008), 'Determining London bus stop locations by means of an automatic vehicle location system', *Transportation Research Record* **2064**(1), 24–32. ZSCC: 0000007.

Rottensteiner, F., Trinder, J., Clode, S. & Kubik, K. (2003), 'Building Detection Using LIDAR Data and Multi-spectral Images', p. 10. ZSCC: NoCitationData[s0].

Roussel, J.-R. & Auty, D. (2019), *lidR: Airborne LiDAR Data Manipulation and Visualization for Forestry Applications*. R package version 2.1.2.
URL: *https://CRAN.R-project.org/package=lidR*

Saeedi, S., Samadzadegan, F. & El-Sheimy, N. (2009), 'Object Extraction from LiDAR Data using an Artificial Swarm Bee Colony Clustering Algorithm', p. 6. ZSCC: 0000018.

Sampath, A. & Shan, J. (2008), 'Building Roof Segmetnation and Reconstruction from LiDAR Point Clouds using Clustering Techniques', p. 6. ZSCC: 0000022.

Smadja, L., Ninot, J. & Gavrilovic, T. (2010), 'Road Extraction and Environment Interpretation from LiDAR Sensors', p. 6. ZSCC: 0000040.

Smith, N., Hirsch, D. & Davis, A. (2012), 'Accessibility and capability: The minimum transport needs and costs of rural households', *Journal of Transport Geography* **21**, 93–101. ZSCC: 0000052.

Solomon, J. & Titheridge, H. (2009), 'Setting accessibility standards for social inclusion: Some obstacles', p. 11.

Solymos, P. & Zawadzki, Z. (2019), *pbapply: Adding Progress Bar to '*apply' Functions*. R package version 1.4-2.
URL: *https://CRAN.R-project.org/package=pbapply*

Steyvers, F. J. J. M. & De Waard, D. (2000), 'Road-edge delineation in rural areas: Effects on driving behaviour', *Ergonomics* **43**(2), 223–238. ZSCC: 0000078.

Stradling, S. G., Great Britain & Department for Transport (2008), *Understanding Inappropriate High Speed: A Quantitative Analysis*, Dept. for Transport, London. ZSCC: NoCitationData[s0] OCLC: 277067852.

Taylor, M. C., Baruya, A. & Kennedy, J. V. (2002), 'The relationship between speed and accidents on rural single-carriageway roads', p. 32. ZSCC: 0000125.

Taylor, M. C., Lynam, D. A. & Baruya, A. (2000), 'The effects of drivers' speed on the frequency of road accidents', p. 56. ZSCC: 0000002.

The University of Edinborough (2019), 'Aerial Digimap Educational User Licence', https://digimap.edina.ac.uk/. ZSCC: 0000000[s0].

TRACC (2019), 'TRACC', https://www.basemap.co.uk/tracc/.

UK Government (1981), 'Wildlife and Countryside Act 1981'.

UK Government (1997), 'The Hedgerow Regulations 1997'.

UK Government (2011), 'Rural Urban Classification', https://www.gov.uk/government/collections/rural-urban-classification. ZSCC: 0000347[s0].

UK Government (2019*a*), 'Find open data - data.gov.uk', https://data.gov.uk/.

UK Government (2019*b*), 'Speed limits', https://www.gov.uk/speed-limits. ZSCC: 0000773[s0].

Velaga, N. R., Beecroft, M., Nelson, J. D., Corsar, D. & Edwards, P. (2012), 'Transport poverty meets the digital divide: Accessibility and connectivity in rural communities', *Journal of Transport Geography* **21**, 102–112. ZSCC: 0000141.

Viner, H., Sinhal, R. & Parry, T. (2004), 'Review of UK skid resistance policy', *Preprint SURF* . ZSCC: 0000023.

Vosselman, G. (2000), 'Slope Based Filtering of Laser Altimetry Data', p. 9. ZSCC: 0000833.

Vosselman, G. (2009), 'Advanced Point Cloud Processing', p. 10. ZSCC: 0000074.

Vosselman, G., Gorte, B. G., Sithole, G. & Rabbani, T. (2004), 'Recognising structure in laser scanner point clouds', *International archives of photogrammetry, remote sensing and spatial information sciences* **46**(8), 33–38. ZSCC: 0000541.

Vosselman, G. & Zhou, L. (2009), 'Detection of curbstones in airborne laser scanning data', *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* **38**(Part 3/W8), 111–116. ZSCC: 0000030.

Wan, Y., Shen, S., Song, Y. & Liu, S. (2007), A Road Extraction Approach Based on Fuzzy Logic for High-Resolution Multispectral Data, *in* 'Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)', Vol. 2, pp. 203–207. ZSCC: 0000014.

Wang, J. & Shan, J. (2009), 'Segmentation of LiDAR Point Clouds for Building Extraction', p. 11. ZSCC: 0000076.

Wickham, H. (2014), *Advanced r*, Chapman and Hall/CRC. ZSCC: 0000183.

Wickham, H. (2017), *tidyverse: Easily Install and Load the 'Tidyverse'*. R package version 1.2.1.
URL: *https://CRAN.R-project.org/package=tidyverse*

Wickham, H. (2018), *scales: Scale Functions for Visualization*. R package version 1.0.0.
URL: *https://CRAN.R-project.org/package=scales*

Wickham, H., Hester, J. & Chang, W. (2019), *devtools: Tools to Make Developing R Packages Easier*. R package version 2.2.0.
URL: *https://CRAN.R-project.org/package=devtools*

Wilke, C. O. (2019), *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*. R package version 1.0.0.
URL: *https://CRAN.R-project.org/package=cowplot*

Williamson, P. (n.d.), *ENVS450: Helper functions for ENVS450*. R package version 0.1.0.

Wu, K.-F., Donnell, E. T., Himes, S. C. & Sasidharan, L. (2013), 'Exploring the Association between Traffic Safety and Geometric Design Consistency Based on Vehicle Speed Metrics', *Journal of Transportation Engineering* **139**(7), 738–748. ZSCC: 0000027.

Yadav, M., Lohani, B. & Singh, A. K. (2018), 'Road Surface Detection from Mobile LiDAR Data', *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* **IV-5**, 95–101. ZSCC: 0000000.

Yoon, J. & Crane, C. D. (2009), Evaluation of terrain using LADAR data in urban environment for autonomous vehicles and its application in the DARPA urban challenge, *in* '2009 ICCAS-SICE', pp. 641–646. ZSCC: 0000011.

Zhang, Shu-Ching Chen, Whitman, D., Mei-Ling Shyu, Jianhua Yan & Chengcui Zhang (2003), 'A progressive morphological filter for removing nonground measurements from airborne LIDAR data', *IEEE Transactions on Geoscience and Remote Sensing* **41**(4), 872–882. ZSCC: NoCitationData[s0].

Zhang, W., Qi, J., Wan, P., Wang, H., Xie, D., Wang, X. & Yan, G. (2016), 'An Easy-to-Use Airborne LiDAR Data Filtering Method Based on Cloth Simulation', *Remote Sensing* **8**(6), 501. ZSCC: 0000112.

Zhang, Z., Zhang, X., Sun, Y. & Zhang, P. (2018), 'Road Centerline Extraction from Very-High-Resolution Aerial Image and LiDAR Data Based on Road Connectivity', *Remote Sensing* **10**(8), 1284. ZSCC: 0000000.

Zhao, J. & You, S. (2012), Road network extraction from airborne LiDAR data using scene context, *in* '2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops', IEEE, Providence, RI, USA, pp. 9–16. ZSCC: 0000042.

Zhu, H. (2019), *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. R package version 1.1.0.
URL: *https://CRAN.R-project.org/package=kableExtra*

Word Count: 13007

A. Environment and Functions

A.1. Session Information

```
Machine:
[1] "AMD Ryzen 5 2600 Six-Core Processor"
Num cores:
[1] 12
Num threads:
[1] 12
RAM:
33.7 GB
```

```
R version 3.6.1 (2019-07-05)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Manjaro Linux

Matrix products: default
BLAS:   /usr/lib/libopenblasp-r0.3.6.so
LAPACK: /usr/lib/liblapack.so.3.8.0

attached base packages:
[1] parallel  stats     graphics  grDevices utils     datasets  methods
[8] base

other attached packages:
 [1] wesanderson_0.3.6          data.table_1.12.2
 [3] showtext_0.7               showtextdb_2.0
 [5] sysfonts_0.8               benchmarkme_1.0.2
 [7] bibtex_0.4.2               cowplot_1.0.0
 [9] pbapply_1.4-2              rgdal_1.4-4
[11] future_1.14.0              varhandle_2.0.3
[13] forcats_0.4.0              stringr_1.4.0
[15] dplyr_0.8.3                purrr_0.3.2
[17] readr_1.3.1                tidyr_0.8.3
[19] tibble_2.1.3               tidyverse_1.2.1
[21] lidR_2.1.2                 raster_3.0-2
[23] sp_1.3-1                   scales_1.0.0
[25] kableExtra_1.1.0           sf_0.7-7
[27] ggpubr_0.2.3               magrittr_1.5
[29] viridis_0.5.1              viridisLite_0.3.0
[31] broom_0.5.2                RStoolbox_0.2.6
[33] ggthemes_4.2.0             PerformanceAnalytics_1.5.3
[35] xts_0.11-2                 zoo_1.8-6
[37] Hmisc_4.2-0                Formula_1.2-3
[39] survival_2.44-1.1          lattice_0.20-38
[41] devtools_2.2.0             usethis_1.5.1
[43] ENVS450_0.1.0              ggplot2_3.2.1
[45] pacman_0.5.1               knitr_1.24
[47] nvimcom_0.9-83             colorout_1.2-1
```

A.2. Functions

A.2.1. LiDAR Clean

```
lidr_clean <- function(cluster) {
  las <- readLAS(cluster)
  if (is.empty(las)) {
    return(NULL)
  }
  # remove all but last return
  las <- lasfilter(las, NumberOfReturns == ReturnNumber)

  # find ground points
  las <- lasground(las, csf())

  ## Create Point DEM
  # interpolate ground points to create raster dtm. Uses Classification = 2
  # very large number of points, therefore idw used as opposed to kriging
  dtm <- grid_terrain(las, 1, knnidw(k = 10, p = 2))
  # normalise heights using dtm
  las <- lasnormalize(las, dtm)
  return(las)
}
```

A.2.2. Catalog to Dataframe

```
ctg_to_df <- function(cluster, aerial = NULL) {
  # read cluster as LAS
  las <- readLAS(cluster)
  # dont read empty clusters
  # all subsequent ctg funcs need these
  if (is.empty(las)) {
    return(NULL)
  }
  # to sp then tibble
  las <- las %>%
    as.spatial()

if (is.null(aerial) == FALSE){
  las@data$lum <- as.numeric(raster::extract(aerial, las))
}
  # sp to df
  las <- as.data.frame(las)
  return(las)
}
```

A.2.3. Clip Samples

```
clip_samples <- function(cluster, x) {
  las <- readLAS(cluster)
  if (is.empty(las)) {
    return(NULL)
  }
  # las to sp, sf then spatial join
  las <- las %>%
    as.spatial() %>%
    st_as_sf(las) %>%
    st_set_crs(27700) %>%
    st_join(x)

  # clip points by removing NA values
  las <- las[is.na(las$sample_id) == FALSE, ]
  return(las)
}
```

A.2.4. Filter LAS Noise

```
las_filter_noise <- function(cluster, sensitivity = 1) {
  las <- readLAS(cluster)
  if (is.empty(las)) {
```

```
    return(NULL)
  }
  # find 95th quantile intensity values per 10m^2
  p95i <- grid_metrics(las, ~ quantile(Intensity, probs = 0.95), 10)
  p95z <- grid_metrics(las, ~ quantile(Z, probs = 0.95), 10)
  # join by merging
  las <- lasmergespatial(las, p95i, "p95i")
  # remove above 95th quantile
  las <- lasfilter(las, Intensity < p95i * sensitivity)

  las <- lasmergespatial(las, p95z, "p95z")
  # remove above 95th quantile
  las <- lasfilter(las, Z < p95z * sensitivity)
  # remove unneeded var
  las$p95i <- NULL
  las$p95z <- NULL
  return(las)
}
```

A.2.5. Extract Buffer

```
extract_buff <- function(cluster, clip_input) {
  las <- readLAS(cluster)

  if (is.empty(las)) {
    return(NULL)
  }

  # ensure no null input
  if (!is.null(clip_input)) {
    las <- lasclip(las, clip_input)

    # bind clipped inputs together
    # as gives list depending on number of
    # sp objects
    if (length(las) > 1) {
      for (i in 1:length(las)) {
        if (!is.empty(las[[i]])) {
          las <- do.call(rbind, las)
          return(las)
        }
      }
    }
  }
}
```

A.2.6. Find Distances

```
find_dists <- function(x, y) {
  # euclidean distance with sf
  d <- st_distance(x, y)
  return(d)
}
```

A.2.7. Euclidean Distance

```
# Function to calculate Euclidean distance between 2 points
# using coordinate data
euclidean_distance <- function(p1, p2) {
  return(sqrt((p2[1] - p1[1])**2 + (p2[2] - p1[2])**2))
}
```

A.2.8. Perpendicular Sampling

```
# Function to calculate 2 points on a line perpendicular to another defined by 2 points p1,p2
# For point at interval, which can be a proportion of the segment length, or a constant
# At distance n from the source line
calc_perp <- function(p1, p2, n, interval = 0.5, proportion = TRUE) {
  # Calculate x and y distances
```

```
  x_len <- p2[1] - p1[1]
  y_len <- p2[2] - p1[2]

  # If proportion calculate reference point from tot_length
  if (proportion) {
    point <- c(p1[1] + x_len * interval, p1[2] + y_len * interval)
  }
  # Else use the constant value
  else {
    tot_len <- euclidean_distance(p1, p2)
    point <- c(
      p1[1] + x_len / tot_len * interval,
      p1[2] + y_len / tot_len * interval
    )
  }

  # Calculate the x and y distances from reference point
  # to point on line n distance away
  ref_len <- euclidean_distance(point, p2)
  xn_len <- (n / ref_len) * (p2[1] - point[1])
  yn_len <- (n / ref_len) * (p2[2] - point[2])

  # Invert the x and y lengths and add/subtract from the refrence point
  ref_points <- rbind(
    point,
    c(point[1] + yn_len, point[2] - xn_len),
    c(point[1] - yn_len, point[2] + xn_len)
  )

  # Return the reference points
  return(ref_points)
}
```

A.2.9. Combine Catalog

```
comb_ctg <- function(x) {
  las <- readLAS(x)
  if (is.empty(las)) {
    return(NULL)
  }
  return(las)
}
```

A.2.10. Compute Samples

```
# default of 10m increments and 30m width either side of a line
compute_samples <- function(x, increment = 10, width = 30) {
  sample_lines <- c()
  if (nrow(x) > 1) {
    # split linestring into coordinates
    road_node <- st_coordinates(x)
    tot_len <- 0
    len_inc <- increment
    len_ofs <- len_inc

    # for each linestring "node"
    # find dist between them
    for (i in 2:nrow(road_node) - 1) {
      n1 <- road_node[i, ]
      n2 <- road_node[i + 1, ]

      len_seg <- euclidean_distance(n1, n2)
      len_ofs <- len_ofs + len_inc

      # max length of linestring
      while (len_ofs <= tot_len + len_seg) {
        len_ofs <- len_ofs + len_inc

        # Add results to output vector
        # for each node of a linestring
        perp_segments <- calc_perp(
```

```
          n1, n2, width,
          len_ofs - tot_len,
          proportion = FALSE
        )

        # combine to multipts
        multipoints <- st_multipoint(matrix(perp_segments, ncol = 2))
        pts <- st_cast(st_geometry(multipoints), "POINT")
        n <- length(pts)

        # points to perp lines
        pair <- st_combine(c(pts[1], pts[2], pts[3]))
        # then to linestring + buffer to polygon
        linestring <- st_cast(pair, "LINESTRING") %>%
          st_buffer(2) %>%
          st_sf() %>%
          mutate(road_id = as.character(unique(x$road_id)))
        sample_lines <- rbind(sample_lines, linestring)
      }
      tot_len <- tot_len + len_seg
    }
  }
  return(sample_lines)
}
```

A.2.11. Greyscale

```
# combine three band rgb
greyscale <- function(x) {
  x <- (x[[1]] + x[[2]] + x[[3]]) / 3
}
```

A.2.12. Compute Individual Linear Model

```
# function to compute individual linear models per
# sample
lm_compute <- function(x, f) tryCatch({
    m <- lm(formula = f, data = x)

    # find p vals
    p <- m %>%
      tidy() %>%
      dplyr::select(p = p.value)

    pred_m <- predict(m, x, type = "response")

    # remove average p val above 0.05
    if (sum(p) / nrow(p) < 0.05) {
      x$lm <- pred_m
    } else {
      x$lm <- NA
    }

    # find 95th quantiles
    x$I_dum <- ifelse(x$lm > quantile(x$lm, .95), 1, 0)

    return(x)
  }, error = function(e) NULL)
```

A.2.13. Filter Returns

```
# remove samples with any road points with a return above 1
filter_returns <- function(x) {
  road <- x[x$road == 1, ]
  if (max(road$NumberOfReturns) == 1) {
    return(x)
  }
}
```

A.2.14. Filter Samples

```
filter_samples <- function(s) {
  # find rows with fewer than 8 samples
  # 8 chosen as ~2m^2 given 25cm res
  if (nrow(s) > 8) {
    # remove outlier points
    # distance based isolation filtering
    distances <- s %>%
      st_distance() %>%
      apply(1, FUN = function(y) {
        min(y[y > 0])
      }) %>%
      as.data.frame() %>%
      mutate(rowid = row_number()) %>%
      select(min_dist = ".", rowid)

    # given min dist between two points
    # remove any above 1m from any other point
    distances <- distances[distances$min_dist < 1, ]

    s <- s %>% mutate(rowid = row_number())

    # remove excluded index values
    s <- s[s$rowid %in% distances$rowid, ]
    return(s)
  }
}
```

A.2.15. Max Dist

```
# two furthest points in a sample
# convert to a linestring to assume max detected road points
max_dist <- function(x) {
  tot_dists <- c()
  # gives largest distances for a collection of pts
  distances <- x %>%
    st_distance(by_element = FALSE) %>%
    unclass() %>%
    "[<-"(lower.tri(., diag = TRUE), NA) %>%
    as_tibble() %>%
    rowid_to_column() %>%
    gather(colid, distance, starts_with("V"),
      na.rm = TRUE
    ) %>%
    arrange(desc(distance))

  # use colid to find index of pts with largest distances
  if (nrow(distances) > 0) {
    distances$colid <- gsub("[^0-9.-]", "", distances$colid)
    tot_dists <- rbind(tot_dists, max(distances$distance))

    distances <- as.list(distances[1, 1:2]) %>%
      unlist() %>%
      as.numeric()

    # convert two pts to linestring
    x <- x[distances, ] %>%
      st_combine() %>%
      st_sf() %>%
      st_cast("LINESTRING")
    return(x)
  }
}
```

A.2.16. Max Lines

```
# combines points filtering and max dist linestrings
# adds linestring length for later
max_lines <- function(x, cents) {
  road_lm <- split(x, f = x$sample_id)
```

```
  road_lm <- road_lm %>% compact()

  # filter samples with few points and isolated points  >1m
  road_lm <- lapply(road_lm, filter_samples)
  road_lm <- road_lm %>% compact()
  # create linestrings
  road_lm <- lapply(road_lm, max_dist)
  road_lm <- do.call(rbind, road_lm)
  road_lm$length <- as.numeric(st_length(road_lm))
  # find intersecting buffers, ensure intersects centreline
  # prevents lines taller than wide
  road_lm <- st_join(road_lm, cents)

  return(road_lm)
}
```

A.2.17. Mid Points

```
# find mid point between linestring
mid_pts <- function(x) {
  fixed_cents <- st_coordinates(x)[, 1:2]
  x_mid <- mean(fixed_cents[, 1])
  y_mid <- mean(fixed_cents[, 2])
  mid_point <- cbind(x_mid, y_mid)
  mid_point <- as.data.frame(mid_point)
  mid_point <- mid_point %>%
    st_as_sf(coords = c("x_mid", "y_mid"), crs = 27700)
  return(mid_point)
}
```

A.2.18. True Centrelines

```
# using mid points convert a list of mid points into
# linestring, i.e. new road centreline
true_cents <- function(x) {
  rd <- unique(x$road_id)
  y <- x %>%
    distinct()
  n <- nrow(y) - 1
  if (nrow(y) > 2) {
    y <- lapply(X = 1:n, FUN = function(i) {
      pair <- y[c(i, i + 1), ] %>%
        st_combine()
      line <- st_cast(pair, "LINESTRING")
      return(line)
    })
    y <- do.call(c, y)
    # remove some noise through filtering out v large lines
    # optimal was qualitatively assessed
    y <- y[as.numeric(st_length(y)) <
      sum(as.numeric(st_length(y))) / (length(y) / 4)]

    y <- y %>%
      st_combine() %>%
      st_cast("MULTILINESTRING")
    y <- y %>% st_sf()

    y <- y[is.na(rd)]

    y$road_id <- as.character(rd)
    return(y)
  }
}
```

A.2.19. Opposite Length

A.2.20. Model Comparison

```
# find estimated mean widths per road
# remove noise given no road above 8m and below 2m
```

```
model_comparison <- function(model) {
  road_lm <- model[!is.na(model$road_id), ]
  rds <- unique(model$road_id)
  road_lm <- split(road_lm, f = road_lm$road_id)

  samp <- Filter(function(x) dim(x)[1] > 0, road_lm)
  cent <- centrelines[centrelines$road_id %in% rds, ]
  cent <- split(cent, f = cent$road_id)
  cent <- Filter(function(x) dim(x)[1] > 0, cent)

  widths <- mapply(adjacent_length, samp, cent)
  widths <- do.call(rbind, widths)
  widths <- as.data.frame(widths)

  widths$adjacent <- as.numeric(unfactor(widths$adjacent))

  widths <- widths[widths$adjacent > 2 & widths$adjacent < 8, ]

  widths <- widths %>%
    group_by(V2) %>%
    select(road_id = V2, adjacent) %>%
    summarise(
      mean_width = mean(adjacent)
    )

  return(widths)
}
```

A.2.21. Road Angles

```
# atan2 to find angle between two centreline segments
# relative to previous centreline orientation
road_angles <- function(rd) {
  coords <- rd %>% st_coordinates()
  angle <- c()
  if (nrow(coords) > 1) {
    for (i in 1:(nrow(rd) - 1)) {
      n1 <- coords[i, ]
      n2 <- coords[i + 1, ]
      x <- n1[1] - n2[1]
      y <- n1[2] - n2[2]
      ang_rad <- atan2(y, x)
      ang_deg <- ang_rad / pi * 180

      angle <- append(angle, ang_deg)
      # left of N same as right of N
      angle <- abs(angle)
    }
  }

  # normalise angle, i.e. use prev orientation to find true difference in angle
  normal_ang <- c()
  for (i in 2:length(angle)) {
    # here i - 1 is theta 1, i is theta 2
    normal <- abs(angle[i] - (angle[i - 1]))
    normal_ang <- rbind(normal_ang, normal)
  }
  normal_ang <- cbind(
    normal_ang,
    as.character(rep(unique(rd$road_id), nrow(normal_ang)))
  )
  return(normal_ang)
}
```

A.2.22. Height Change

```
# find difference in average height between two samples
height_change <- function(x) {
  elev <- c()
  samples <- split(x, x$sample_id)
  if (length(samples) > 2) {
```

```
    for (s in 2:length(samples) - 1) {
      pair <- samples[c(s, s + 1)]
      n1 <- mean(pair[[1]]$Z)
      n2 <- mean(pair[[2]]$Z)
      e <- abs(n1 - n2)
      e <- cbind(
        as.character(unique(samples[[s]]$road_id)), e
      )
      elev <- rbind(elev, e)
    }
  }
  return(elev)
}
```

A.2.23. Formatting

```
make_table <- function(df, cap = "", dig = 2, col_names = NA, table_env = "table", ...) {
  require(kableExtra)
  require(tidyverse)

  options(knitr.kable.NA = "")
  kable(df,
    digits = dig, caption = cap,
    linesep = "", # remove 5 row spacing
    longtable = FALSE, booktabs = TRUE, # latex opts
    format = "latex",
    escape = F, # allow maths chars
    col.names = col_names,
    table.env = table_env # change to figure*
  ) %>%
    kable_styling(font_size = 9, position = "center") %>%
    row_spec(0, bold = TRUE)
}
```

A.2.24. Beta Coefficients

```
lm_beta <- function(model) {
  b <- summary(model)$coef[-1, 1]
  sx <- apply(model$model[-1], 2, sd)
  sy <- apply(model$model[1], 2, sd)
  beta <- b * sx / sy
  return(beta)
}
```

B. Additional Tables and Figures

Table B.1: *Spearman's rank correlation coefficients for all variables in relation to the road outcome variable*

Variable	Rho	Lower CI †	Upper CI †
dists	**-0.28590 ****	**-0.279**	**-0.274**
Intensity	**-0.23199 ****	**-0.211**	**-0.207**
gpstime	-0.02286 **	-0.021	-0.016
Z	**-0.01694 ****	**-0.030**	**-0.025**
lum	0.01617 **	0.022	0.027
ReturnNumber	**-0.01271 ****	**-0.020**	**-0.015**
NumberOfReturns	**-0.01271 ****	**-0.020**	**-0.015**
ScanDirectionFlag	-0.00402 **	-0.007	-0.002
ScanAngleRank	-0.00081	0.000	0.005
EdgeOfFlightline	0.00057	-0.002	0.003

* Significant at the 0.05 level;
** Significant at the 0.01 level;
*** Significant at the 0.001 level;
† 95% Confidence Interval

Table B.2: *Estimated number of bends per road*

Road ID	Number of Bends	Road Length (km)	Bends per Kilometer
6	8	0.36	22.29
8	4	0.19	20.79
5	2	0.11	18.77
9	4	0.21	18.68
7	4	0.26	15.34
36	4	0.28	14.27
38	1	0.07	13.34
14	3	0.24	12.59
32	3	0.24	12.42
2	4	0.38	10.57
21	1	0.11	9.05
4	3	0.38	7.94
12	2	0.25	7.93
39	1	0.13	7.82
31	1	0.13	7.74
15	1	0.13	7.58
10	2	0.26	7.58
33	4	0.54	7.37
13	1	0.17	5.96
17	1	0.18	5.62
30	2	0.41	4.85
34	1	0.21	4.85
25	1	0.23	4.36
24	1	0.24	4.16
1	1	0.24	4.15
18	2	0.74	2.70
16	0	0.18	0.00
20	0	0.07	0.00
40	0	0.08	0.00
41	0	0.08	0.00

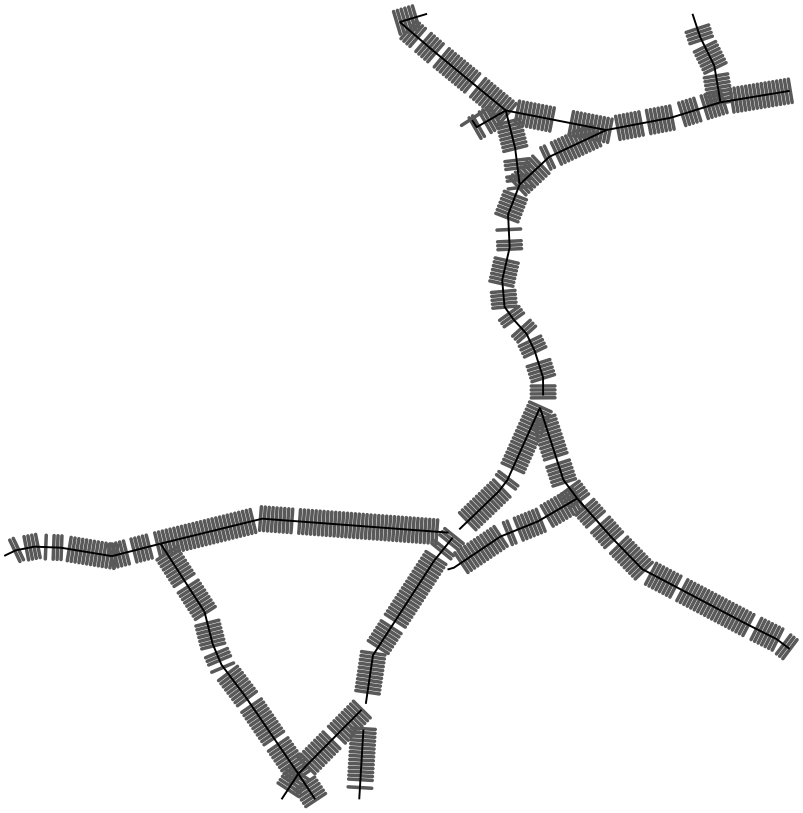

Fig. B.1: *Sample lines extracted based on known road locations*

C.1. Clean Data

```
# Source Scripts
source("./functions.r")

# Create las catalog with all .laz files
ctg <- catalog("../data/point/")
opt_chunk_size(ctg) <- 500
opt_chunk_buffer(ctg) <- 20

# create lax file to index + speed up process
plan(multisession, workers = 6L)
set_lidr_threads(12L)
# speed up lax computation time
lidR:::catalog_laxindex(ctg)

# ctg to points csv
las <- catalog_apply(ctg, ctg_to_df)
las <- do.call(rbind, las)
las <- las %>%
    select(-c(
        Synthetic_flag,
        Keypoint_flag,
        Withheld_flag
    ))

fwrite(las, "../data/point/points.csv")

# filter using sql expressions why not
# very very slow to read in full gpkg, don't run unless new data added
#roads <- st_read("../data/osroads/oproad_gpkg_gb/data/oproad_gb.gpkg",
#    layer = "RoadLink", query =
#        "SELECT * FROM RoadLink WHERE
#         formOfWay = \"Single Carriageway\" AND
#         roadFunction <> \"Restricted Local Access Road\" "
#) %>%
#    st_zm() # remove z axis
#
#roads <- as_Spatial(roads)
#roads <- raster::crop(roads, as.matrix(extent(ctg))) %>%
#    st_as_sf()

#st_write(roads, "../data/osroads/oproad_crop.gpkg")

roads <- st_read("../data/osroads/oproad_crop.gpkg") %>%
    mutate(
        len = as.numeric(st_length(geom)),
        road_id = paste0("road_", row_number())
    ) %>%
    select(c(road_id, roadFunction, len, geom)) %>%
    subset(len > 50)

# keep line polys
roads_line <- roads

# one buffer to include non road points, 1m buffer to show only road points
roads_buff <- st_buffer(roads, 30)
roads <- st_buffer(roads, 1)
roads_buff_union <- st_union(roads_buff)
```

```
# write all outputs to files
st_write(roads, "../data/derived/roads/roads.gpkg",
    delete_layer = TRUE
)
st_write(roads_line, "../data/derived/roads/roads_line.gpkg",
    delete_layer = TRUE
)
st_write(roads_buff, "../data/derived/roads/roads_buff.gpkg",
    delete_layer = TRUE
)

st_write(roads_buff_union, "../data/derived/roads/roads_buff_diss.gpkg",
    delete_layer = TRUE
)

roads_buff <- st_read("../data/derived/roads/roads_buff.gpkg") %>%
    as_Spatial()

ctg <- catalog("../data/point/")
opt_output_files(ctg) <- "../data/derived/ctg_clean/{ID}_clean"
opt_chunk_size(ctg) <- 500
opt_chunk_buffer(ctg) <- 20
catalog_apply(ctg, lidr_clean)

ctg <- catalog("../data/derived/ctg_clean/")
opt_output_files(ctg) <- "../data/derived/ctg_buff/{ID}_tile"
opt_chunk_size(ctg) <- 500
opt_chunk_buffer(ctg) <- 20
catalog_apply(ctg, extract_buff, roads_buff)

ctg <- catalog("../data/derived/ctg_buff/")
opt_output_files(ctg) <- "../data/derived/ctg/{ID}_tile"
opt_chunk_size(ctg) <- 500
opt_chunk_buffer(ctg) <- 20
catalog_apply(ctg, las_filter_noise, sensitivity = 1.2)

# non normalised ctg
ctg_notnorm <- catalog("../data/point/")
opt_output_files(ctg_notnorm) <- "../data/derived/ctg_notnorm/{ID}_tile"
opt_chunk_size(ctg_notnorm) <- 500
opt_chunk_buffer(ctg_notnorm) <- 20
catalog_apply(ctg_notnorm, extract_buff, roads_buff)

# read in written roads file
roads <- read_sf("../data/derived/roads/roads.gpkg")

# find roads extent shows study area + used for aerial imagery from digimaps
extent <- st_as_sfc(st_bbox(roads))

# Write extent shapefile
st_write(extent, "../data/derived/extent/extent.gpkg", delete_layer = TRUE)
```

C.2. Create Sample lines

```
source("./functions.r")

centrelines <- read_sf("../data/derived/roads/roads_line.gpkg") %>%
    st_set_crs(27700)

roads_split <- centrelines %>% st_cast("POINT")

roads_split <- split(roads_split, f = roads_split$road_id)

sample_lines <- lapply(roads_split, compute_samples)
sample_lines <- do.call(rbind, sample_lines)

sample_lines <- sample_lines %>%
    st_set_crs(27700)
# label each sample
sample_lines$sample_id <- seq.int(nrow(sample_lines))
```

```
write_sf(sample_lines, "../data/derived/roads/sample_lines.gpkg")

ctg <- catalog("../data/derived/ctg_buff/")

opt_chunk_size(ctg) <- 500
plan(multisession, workers = 6L)
set_lidr_threads(12L)

# remove points outside samples
comb <- catalog_apply(ctg, clip_samples, sample_lines)
comb <- comb <- do.call(rbind, comb)

roads <- st_read("../data/derived/roads/roads.gpkg") %>%
    st_transform(27700)
roads_df <- roads %>% st_drop_geometry()

comb <- comb %>%
    st_transform(27700)

joined_output <- merge(comb, roads_df, by = "road_id")

int <- st_contains(roads, joined_output, sparse = FALSE) %>%
    colSums()

joined_output$road <- int

# turn to binary, some road buffers overlap
joined_output$road <- as.numeric(joined_output$road > 0)

# aerial data
jpgs <- Sys.glob("../data/aerial/*.jpg")
jpgs <- lapply(jpgs, brick)
grey_rasters <- lapply(jpgs, greyscale)
grey_rasters <- lapply(grey_rasters, brick)
aerial <- do.call(merge, grey_rasters)
aerial <- crop(aerial, roads)

writeRaster(aerial, "../data/derived/aerial/aerial_crop.tif",
    format = "GTiff", overwrite = TRUE
)

# crop aerial data
lum <- raster::extract(aerial, joined_output)
joined_output$lum <- as.numeric(lum)


# find dists from centrelines
joined_output <- split(joined_output, f = joined_output$road_id)
centrelines <- split(centrelines, centrelines$road_id)

centrelines <- centrelines[names(joined_output)]

dists <- mapply(
    find_dists,
    joined_output,
    centrelines
)

joined_output <- do.call(rbind, joined_output)
dists <- do.call(rbind, dists)
joined_output$dists <- dists

coords <- joined_output %>%
    st_coordinates()

# change to data.frame
joined_output <- joined_output %>%
    st_drop_geometry() %>%
    mutate(
        X = coords[, 1],
        Y = coords[, 2]
    )
```

```
fwrite(joined_output, "../data/derived/model_data/sampled_las.csv")
```

C.3. Linear Models and Improve Centrelines

```
source("./functions.r")
sampled_las <- fread("../data/derived/model_data/sampled_las.csv")

# ground pts only
sampled_las <- sampled_las[sampled_las$Classification == 2, ]

# global linear model: unfiltered
# for this section see social survey + ss assessment 2
f1 <- as.formula("road ~ Intensity + lum + dists + Z + NumberOfReturns")
lm1 <- lm(data = sampled_las, formula = f1)
lm1_pred <- predict(lm1, sampled_las, type = "response")

sampled_las$lm1_pred <- lm1_pred
sampled_las$lm1_dum <- ifelse(sampled_las$lm1_pred >
    quantile(sampled_las$lm1_pred, .95), 1, 0)

sampled_las$lm1_pred <- lm1_pred
sampled_las$lm1_dum90 <- ifelse(sampled_las$lm1_pred >
    quantile(sampled_las$lm1_pred, .90), 1, 0)

sampled_las$lm1_pred <- lm1_pred
sampled_las$lm1_dum80 <- ifelse(sampled_las$lm1_pred >
    quantile(sampled_las$lm1_pred, .80), 1, 0)

fwrite(sampled_las, "../data/derived/model_data/linearmodels.csv")
```

C.4. Road Widths

```
source("./functions.r")
## ---- widths
road_lm <- fread("../data/derived/model_data/linearmodels.csv") %>%
    as.data.frame() %>%
    st_as_sf(coords = c("X", "Y"), crs = 27700)

roads <- st_read("../data/derived/roads/roads_line.gpkg")
roads_5m <- st_read("../data/derived/roads/roads_line.gpkg") %>%
    st_buffer(5)

road_lm90 <- road_lm[road_lm$lm1_dum90 == 1, ]
# find improved centrelines
fixed_cents <- list(
    road_lm90
)

# includes all filtering, max dist points
fixed_cents <- lapply(fixed_cents, max_lines, cents = roads)

fixed_cents <- do.call(rbind, fixed_cents)
fixed_cents <- fixed_cents %>%
    mutate(rowid = row_number())

mid_point <- split(fixed_cents, fixed_cents$rowid)
mid_points <- lapply(mid_point, mid_pts)

mid_points <- do.call(rbind, mid_points)
mid_points <- mid_points %>%
    st_join(roads_5m)
mid_rds <- split(mid_points, mid_points$road_id)

# remove empty geoms
mid_rds <- Filter(function(x) dim(x)[1] > 0, mid_rds)
cents <- lapply(mid_rds, true_cents)
cents <- compact(cents)
cents <- do.call(rbind, cents)
```

```
st_write(cents, "../data/derived/roads/cent_iteration1.gpkg",
    layer_options = "OVERWRITE=yes"
)

## ---- angles
roads_split <- st_read("../data/derived/roads/roads_line.gpkg") %>%
    st_cast("POINT") %>%
    st_set_crs(27700)

roads_split <- split(roads_split, roads_split$road_id)

angles <- lapply(roads_split, road_angles)
angles <- do.call(rbind, angles)
row.names(angles) <- NULL
angles <- angles %>%
    as.data.frame()
names(angles) <- c("angle", "road_id")
angles$angle <- as.numeric(unfactor(angles$angle))

angles <- angles %>%
    group_by(road_id) %>%
    summarise(
        mean_angle = mean(angle),
        max_angle = max(angle)
    )

roads <- merge(roads, angles, by = "road_id")

## ---- heights
# Non-normalised las files
sample_lines <- st_read("../data/derived/roads/sample_lines.gpkg") %>%
    st_set_crs(27700)
roads_1m <- st_read("../data/derived/roads/roads.gpkg")
ctg <- catalog("../data/derived/ctg_notnorm/")
opt_chunk_size(ctg) <- 500
plan(multisession, workers = 6L)
set_lidr_threads(12L)

# remove points outside samples
las_rds <- catalog_apply(ctg, clip_samples, sample_lines)
las_rds <- do.call(rbind, las_rds)

las_rds <- las_rds[las_rds$NumberOfReturns == 1 &
    las_rds$Classification == 2, ]

rds <- st_read("../data/derived/roads/roads.gpkg") %>%
    st_transform(27700)

rd_line <- st_read("../data/derived/roads/roads_line.gpkg", quiet = TRUE) %>%
    mutate(len = as.numeric(st_length(geom))) %>%
    select(c(road_id, len)) %>%
    st_drop_geometry()

roads_df <- rds %>% st_drop_geometry()

las_rds <- las_rds %>%
    st_transform(27700)

las_rds <- merge(las_rds, roads_df, by = "road_id")

int <- st_contains(roads_1m, las_rds, sparse = FALSE) %>%
    colSums()

las_rds$road <- int

# remove overlapping road points
las_rds <- las_rds[las_rds$road < 2, ]
# turn to binary (might not be needed)
las_rds$road <- as.numeric(las_rds$road > 0)

las_rds <- las_rds[las_rds$road == 1, ]
```

```
las_rds <- split(las_rds, las_rds$sample_id)

las_rds <- lapply(las_rds, filter_returns)

las_rds <- las_rds %>%
    compact()

las_rds <- do.call(rbind, las_rds)

las_height <- split(las_rds, las_rds$road_id)

las_height <- lapply(las_height, height_change)

las_height <- do.call(rbind, las_height)
las_height <- as.data.frame(las_height)

names(las_height) <- c("road_id", "Z")
las_height <- las_height %>%
    merge(rd_line, by = "road_id")

las_height <- las_height %>%
    group_by(road_id) %>%
    summarise(
        tot_z = sum(as.numeric(unfactor(Z))) / (mean(len) / 1000),
    ) %>%
    drop_na()

roads <- merge(roads, las_height, by = "road_id")

## ---- surface_qual
las_qual <- las_rds %>%
    group_by(road_id) %>%
    summarise(
        mean_int = mean(Intensity),
        range_int = max(Intensity) - min(Intensity)
    ) %>%
    drop_na() %>%
    select(c(road_id, mean_int, range_int))

roads <- merge(roads, las_qual, by = "road_id") %>%
    st_drop_geometry()

write.csv(roads, "../data/final_data/final.csv")
```

C.5. Improved Centreline Models

```
source("./functions.r")
cent1 <- st_read("../data/derived/roads/cent_iteration1.gpkg") %>%
    st_transform(27700)
sampled_las <- fread("../data/derived/model_data/sampled_las.csv") %>%
    as.data.frame() %>%
    st_as_sf(coords = c("X", "Y"), crs = 27700)
aerial <- raster("../data/derived/aerial/aerial_crop.tif")


# improved roads centrelines
roads <- cent1 %>%
    st_buffer(2)

roads_df <- roads %>% st_drop_geometry()

joined_output <- merge(sampled_las, roads_df, by = "road_id")

int <- st_contains(roads, joined_output, sparse = FALSE) %>%
    colSums()

joined_output$road <- int

# turn to binary, some road buffers overlap
joined_output$road <- as.numeric(joined_output$road > 0)
```

```
# crop aerial data
lum <- raster::extract(aerial, joined_output)
joined_output$lum <- as.numeric(lum)

# find dists from centrelines
joined_output <- split(joined_output, f = joined_output$road_id)
centrelines <- split(cent1, cent1$road_id)

centrelines <- centrelines[names(joined_output)]

dists <- mapply(
    find_dists,
    joined_output,
    centrelines
)

joined_output <- do.call(rbind, joined_output)
dists <- do.call(rbind, dists)
joined_output$dists <- dists

coords <- joined_output %>%
    st_coordinates()
cent1_las <- joined_output %>%
    st_drop_geometry() %>%
    mutate(
        X = coords[, 1],
        Y = coords[, 2]
    )

fwrite(cent1_las, "../data/derived/model_data/cent1_lm.csv")

# linear models with improved centrelines
# for this section see social survey + ss assessment 2
f1 <- as.formula("road ~ Intensity + lum + dists + Z + NumberOfReturns")
lm1 <- lm(data = cent1_las, formula = f1)
lm1_pred <- predict(lm1, cent1_las, type = "response")

f2 <- as.formula("road ~ Intensity + dists + Z + NumberOfReturns")
lm2 <- lm(data = cent1_las, formula = f2)
lm2_pred <- predict(lm2, cent1_las, type = "response")

cent1_las$lm1_pred <- lm1_pred
cent1_las$lm1_dum <- ifelse(cent1_las$lm1_pred >
    quantile(cent1_las$lm1_pred, .95), 1, 0)

cent1_las$lm2_pred <- lm2_pred
cent1_las$lm2_dum <- ifelse(cent1_las$lm2_pred >
    quantile(cent1_las$lm2_pred, .95), 1, 0)


# individual linear probability model: has to filter out canopy: proof of concept
cent1_las <- split(cent1_las, cent1_las$sample_id)
cent1_las <- lapply(cent1_las, filter_returns)
f1 <- as.formula("road ~ Intensity + dists + Z + NumberOfReturns")
cent1_las <- lapply(cent1_las, lm_compute, f = f1)
cent1_las <- do.call(rbind, cent1_las)

fwrite(cent1_las, "../data/final_data/cent_lm.csv")

lmi <- cent1_las[cent1_las$I_dum == 1, ] %>%
    as.data.frame() %>%
    st_as_sf(coords = c("X", "Y"), crs = 27700)
lm1 <- cent1_las[cent1_las$lm1_dum == 1, ] %>%
    as.data.frame() %>%
    st_as_sf(coords = c("X", "Y"), crs = 27700)
lm2 <- cent1_las[cent1_las$lm2_dum == 1, ] %>%
    as.data.frame() %>%
    st_as_sf(coords = c("X", "Y"), crs = 27700)

lm2 <- split(lm2, lm2$road_id)
tot_pts <- lapply(lm2, function(x) {
    tot_pts <- nrow(x)
    return(tot_pts)
```

```
})
lm2 <- do.call(rbind, lm2)

lm_max_widths <- list(lmi, lm1, lm2)

road_buff <- st_read("../data/derived/roads/roads_buff.gpkg")

centrelines <- do.call(rbind, centrelines)
# includes all filtering, max dist points
lm_max_widths <- lapply(lm_max_widths, max_lines, cents = centrelines)

lm_max_widths <- lapply(lm_max_widths, function(x) {
    x <- x[x$length < 8 & x$length > 2, ]
    x <- x[!is.na(x$road_id), ]
})

# save lines for comparison
for (i in 1:length(lm_max_widths)) {
    st_write(lm_max_widths[[i]], paste0("../data/final_data/widths_", i, ".gpkg"),
        layer_options = "OVERWRITE=YES"
    )
}

####

centrelines <- st_read("../data/derived/roads/cent_iteration1.gpkg")
linear_widths <- lapply(lm_max_widths, model_comparison)
linear_widths <- linear_widths %>%
    reduce(left_join, by = "road_id")

names(linear_widths) <- c(
    "road_id",
    "lmi_mean",
    "lm1_mean",
    "lm2_mean"
)
tot_pts <- do.call(rbind, tot_pts) %>%
    as.data.frame() %>%
    rownames_to_column()
names(tot_pts) <- c("road_id", "tot_pts")

linear_widths <- merge(linear_widths, tot_pts, by = "road_id")

roads <- fread("../data/final_data/final.csv")

roads <- merge(roads, linear_widths, by = "road_id")

######

# old centrelines
sampled_las <- fread("../data/derived/model_data/sampled_las.csv") %>%
    as.data.frame() %>%
    st_as_sf(coords = c("X", "Y"), crs = 27700)

f1 <- as.formula("road ~ Intensity + lum + dists + Z + NumberOfReturns")
lm0 <- lm(data = sampled_las, formula = f1)
lm0_pred <- predict(lm0, sampled_las, type = "response")
sampled_las$lm0_pred <- lm0_pred
sampled_las$lm0_dum <- ifelse(sampled_las$lm0_pred >
    quantile(sampled_las$lm0_pred, .95), 1, 0)

fwrite(sampled_las, "../data/final_data/lm0.csv")

lm0 <- sampled_las[sampled_las$lm0_dum == 1, ]

lm_max_widths <- list(lm0)

road_buff <- st_read("../data/derived/roads/roads_buff.gpkg")
centrelines <- st_read("../data/derived/roads/roads_line.gpkg")
# includes all filtering, max dist points
lm_max_widths <- lapply(lm_max_widths, max_lines, cents = centrelines)

lm_max_widths <- lapply(lm_max_widths, function(x) {
```

```
    x <- x[x$length < 8 & x$length > 2, ]
    x <- x[!is.na(x$road_id), ]
})

# save lines for comparison
st_write(lm_max_widths[[1]], paste0("../data/final_data/widths_0.gpkg"),
    layer_options = "OVERWRITE=YES"
)

####
linear_widths <- lapply(lm_max_widths, model_comparison)
linear_widths <- linear_widths %>%
    reduce(left_join, by = "road_id")

names(linear_widths) <- c(
    "road_id",
    "lm0_mean"
)

roads <- merge(roads, linear_widths, by = "road_id")
fwrite(roads, "../data/final_data/final.csv")

# aerial data

# ctg to points csv
ctg <- catalog("../data/derived/ctg/")
las <- catalog_apply(ctg, ctg_to_df, aerial)
las <- do.call(rbind, las)
las <- las %>%
    select(-c(
        Synthetic_flag,
        Keypoint_flag,
        Withheld_flag
    ))

fwrite(las, "../data/point/points_clean.csv")
```