Utilising Supervised Parametric 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 uses Government produced LiDAR point cloud data to obtain road surface height, and intensity information. Road centrelines provided by Ordnance Survey are used to enable an automated classification of road surfaces, allowing for road width extraction. Aerial imagery is additionally used to provide luminescence information to aid with road classification. Additional geometric features extracted from the chosen rural 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. The method presented ensures scalability, allowing for an extension beyond the $1km^2$ area proposed, given access to the appropriate data and suitable computational power. For road classification to be used in width extraction, a unique classification method is proposed, utilising a linear probability model, given points determined to be road surface, derived from the road centreline information. A sampling methodology is implemented, which primarily reduces the computational overhead through a reduction in LiDAR points required for classification. Results were broadly assessed qualitatively and quantitatively giving around a 70% accuracy in road width measurements. A direct comparison of individual road quality is presented as the Road Quality Index (RQI), which combines the extracted geometric features of roads.

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

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

Acknowledgements

I would like to thank my supervisor Paul Williamson for his continued support throughout 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 and insights that were essential for this dissertation. 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 vehicle 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:** a measure of the ease of reaching (and interacting with) destinations or activities distributed in space.
- **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, 2018b). 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, 2019a). 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 and 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, 2019a)

The Department for Transport (2013b) 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 (2013b) note that this is especially common on C and Unclassified roads due to the narrow width, 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 (2013b) 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, 2018a) 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, 2013a). 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, 2018a), 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, and to inform future requirements for a full rural road hierarchy. 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

Aerial 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, unlike 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 aircraft 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 and 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 and 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 luminescence 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 and 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 Dissertation

This dissertation will present a method for rural road classification and width extraction for a 1km^2 region in the North West of England (Figure 1.1). 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

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

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 aims to build upon past road classification studies, combined with a more refined methodology that aims to ensure a higher accuracy for rural British roads. Unlike previous road classification methodologies, this dissertation 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.

Key Aims:

- 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

  Road width

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

This dissertation 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 dissertation 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 classification, 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 that may be applied nationally. 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, 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.

Features that Taylor *et al.* (2002) suggest to consider in an analysis of British rural roads are given on Table 2.1. This study 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.

Table 2.1: *Features identified as important to British rural road analysis.*

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

(Taylor *et al.*, 2002)

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.*

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.

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 (2013b) 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 (2013b) 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 and 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, 2013a).

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, and 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) (e.g. Kraus and Pfeifer, 1998; Maas, 1999; Elberink and Maas, 2000), by segmenting vegetation, and man made structures from ground. This dissertation will utilise a recent method for DTM 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 by Zhang *et al.* (2016) gave results similar to top existing DTM production algorithms, but with a far more simplistic implementation, and reduced computation times.

2.2.2. Feature Classification

Developments in LiDAR enabled the possibility of classification beyond ground and non-ground, 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

[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 and Rnaganatan, 2001)

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 and Zhou, 2009). Road classification is essentially a data clustering method to categorise data into road and non-road points, enabled through discovering patterns and relationships 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, and 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 and Shengrui Wang, 2005; Wan *et al.*, 2007). Such techniques rely solely on image texture analysis of the coloured pixels (Dubes and Ohanian, 1992), and appear to be focused entirely in urban settings. Additionally, the methods proposed in aerial image classification fail to address the differentiation between road and pavement, simply classifying building and non-building, rather than roads.

2.3.2. LiDAR Road Classification

Intensity, elevation and the inclusion of multiple returns in LiDAR data enabled methods for categorising roads, and differentiating them from the surrounding ground, despite the similar elevation (Clode *et al.*, 2004).

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), and applications using aerial LiDAR have also followed this trend for unsupervised classification (Clode *et al.*, 2004; Vosselman, 2009).

Many 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). 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. However, this mobile LiDAR classification required driving speeds of 20km/h to 50km/h, and lengthy computation times. Additionally noted by Yoon and Crane (2009) in a similar study. Ground based LiDAR collection techniques often provide a resolution far higher than aerial LiDAR data, however the time taken for a comprehensive coverage of all British rural roads would not allow for its use in a national road assessment.

Other unsupervised 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

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 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.

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 roads 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 the segmentation between roads and buildings, relying on a clear height difference between road and non-road surfaces (Kumar *et al.*, 2013), a feature uncommon to rural roads. Rural road classification therefore must rely on alternative features of roads, notably the difference in intensity produced by vegetation compared with the surface of man made objects, 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 particular consideration.

2.4. Overview of this Dissertation

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 and 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 and 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 made possible through the `lidR` **R** package (Roussel and Auty, 2019). Further details regarding the **R** environment and computer setup used for this dissertation are given in **Appendix A**. Also 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.

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, 2019b). 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 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 Edinburgh, 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 return number 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.

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

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, taking only points where the **ReturnNumber** equals the **NumberOfReturns**. Last pulse points may be expressed as;

$$\mathbf{p}_i = (lpx, lpy, lpz, lpi),$$

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;

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

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 existing 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^2. 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;

$$\mathbf{p}_i \in \left[A(r_i) \times 30m^2\right],$$

A.2.2

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, and removing all points above the 95% percentile. This can be expressed as;

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

A.2.3

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 original points in $\mathbf{c}_k$.

3.2.5. LiDAR Catalog

As mentioned in Section 3.1, objects of class `LASCatalog` enabled more efficient processing by allowing the LiDAR point cloud to be processed in predefined batch sizes. Considering a collection of processed LiDAR points; last pulse, normalised, clipped to 30m road extents, and intensity noise filtered. Points were then 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.2.3

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.2.4

3.3. Road Analysis

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.

```
Reading layer `oproad_crop' from data source `/home/cjber/drive/gds/envs492/data/osroads/oproad_crop.gpkg' using dri
Simple feature collection with 41 features and 20 fields
geometry type:  LINESTRING
dimension:      XY
bbox:           xmin: 380000 ymin: 368000 xmax: 382000 ymax: 370000
epsg (SRID):    NA
proj4string:    +proj=tmerc +lat_0=49 +lon_0=-2 +k=0.9996012717 +x_0=400000 +y_0=-100000 +ellps=airy +units=m +no_de
```

Table 3.2: *OS Roads Data Summary*

Variable	Example
id	id3DF463D5-C191-4F41-B941-CB69E13E53DA
roadNameTOID	osgb4000000013342919
roadNumberTOID	
roadClassification	Unknown
roadFunction	Local Road
formOfWay	Single Carriageway
length	466
name1	Pitt Lane

The roads in this paper consist of these functions;

- Local Road
- Minor Road
- B Road
- Secondary Access 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 *Restricted Local Access Roads* were removed, as were roads with a length of less than 50m, often those clipped by the extent of the LiDAR data.

3.4. Road Angles

The angle of each bend in a road was 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, (See Figure 3.1 and 3.2 for illustrations of road nodes).

The direction of a road was considered to be the *bearing angle* $\hat{\theta}$, from a node $\mathbf{N}_i = (x_i, y_i)$ to a node $\mathbf{N}_{i+1} = (x_{i+1}, y_{i+1})$, with the angle measured in a clockwise direction from north. This is represented on Figure 3.1.

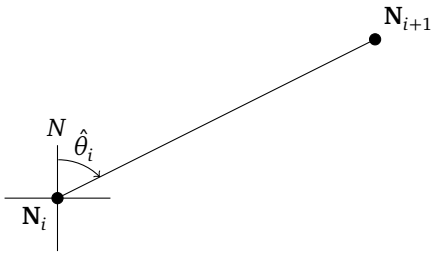

Fig. 3.1: *Bearing Angle (θ) between two sequential road Nodes; $\mathbf{N}_i$ and $\mathbf{N}_{i+1}$. North is given by N.*

To find the angle $\hat{\theta}$, the node $\mathbf{N}_{i+1}$ can be represented into relation to node $\mathbf{N}_i$ as;

$$(x_{i+1}, y_{i+1}) = (x_i + r\sin\theta, y_i + r\cos\theta)$$

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

$$\tan\theta = \frac{x_{i+1} - x_i}{y_{i+1} - y_i}$$

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

$$\theta = \text{atan2}\,(x_{i+1} - x_i, y_{i+1} - y_i) \in [-\pi, \pi]$$

finally the bearing angle $\hat{\theta} \in [0, 2\pi]$ may be obtained by the addition of 2π to any value below 0;

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

A.2.5

With the bearing angle of the first line segment ($\hat{\theta}_1$) for a particular road, the change in orientation of the second line segment between nodes $\mathbf{N}_2$ and $\mathbf{N}_3$ may be given;

$$\begin{aligned}\hat{\theta}_2 =& \text{atan2}\,(x_3 - x_2, y_3 - y_2) - \\ & \text{atan2}\,(x_2 - x_1, y_2 - y_1),\end{aligned}$$

A.2.5

or simply written as $\hat{\theta}_2 = \hat{\theta}_2 - \hat{\theta}_1$, with additional nodes following the pattern $\hat{\theta}_i = \hat{\theta}_i - \hat{\theta}_{i-1}$. Figure 3.2 illustrates this for nodes 1 to 3, indicating the bearing angle $\hat{\theta}_2$ in relation to the bearing angle $\hat{\theta}_1$, rather than in relation to the north (N).

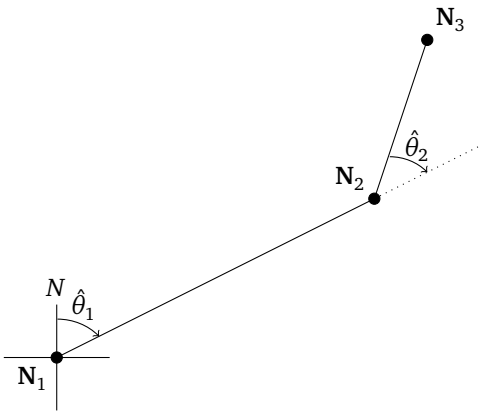

Fig. 3.2: *Bearing Angle between the first ($\mathbf{N}_1$), second ($\mathbf{N}_2$) and third node ($\mathbf{N}_3$) of a road; giving $\hat{\theta}_1$ in relation to the north (N), and $\hat{\theta}_2$ in relation to the bearing angle $\hat{\theta}_1$.*

As the bearing angle between the first two nodes gives only the initial direction of the road, this was set to zero; $\hat{\theta}_1 = 0$. In the final analysis, for each road the maximum bearing angle between any two nodes was selected, as well as the average bearing angle between all nodes.

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;

$$\Delta e(\mathbf{N}_i, \mathbf{N}_{i+1}) = \Delta \begin{cases} \overline{spz} \in \left[A(\mathbf{N}_i) \times 1m^2\right], \\ \overline{spz} \in \left[A(\mathbf{N}_{i+1}) \times 1m^2\right] \end{cases},$$

A.2.6

where spz are the z values for ground classified points with a single return, and $\Delta e(\mathbf{N}_i, \mathbf{N}_{i+1})$ is the change in elevation between sequential nodes $\mathbf{N}_i$ and $\mathbf{N}_{i+1}$, taking the change in mean single return pulse point z values ($\overline{spz}$) within a 1m^2 buffer of each node. For each road, the total elevation change per kilometer was calculated by dividing the sum of all elevation changes between two neighbouring road nodes by the length of a road in kilometers;

$$\Delta e = \sum \frac{\Delta e(\mathbf{N}_i, \mathbf{N}_{i+1})}{L_k \times 1000},$$

A.2.6

where L_k is the length of a road k in meters.

3.6. Surface Quality

Surface quality was assessed through the range in intensity values found at a 1m^2 area around each road node. Again, 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 and classified as ground were used in this analysis;

$$\begin{aligned}\Delta q(\mathbf{N}_1, \mathbf{N}_2, \ldots, \mathbf{N}_i) =& \Delta(\overline{spi}) \in A(\mathbf{N}_i) \times 1m^2 \\ q_r =& \Delta q(\mathbf{N}_1, \mathbf{N}_2, \ldots, \mathbf{N}_i)_{max} - \\ & \Delta q(\mathbf{N}_1, \mathbf{N}_2, \ldots, \mathbf{N}_i)_{min},\end{aligned}$$

where spi represents LiDAR point intensities with a single pulse return and ground classified, and $\Delta q(\mathbf{N}_1, \mathbf{N}_2, \ldots, \mathbf{N}_i)$ represents each individual intensity value for each node in a road, giving the range q_r.

3.7. Road Width

3.7.1. Road Sampling

The LiDAR point cloud was sampled at 60m by 2m bounding regions, 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 neighbouring 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;

$$\begin{aligned}|x| =& x_{i+1} - x_i \\ |y| =& y_{i+1} - y_i,\end{aligned} \tag{3.1}$$

A.2.7

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.2.8

A point (p_k) along these two nodes, at an interval distance I_k, was determined 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.2.7

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 $\mathbf{N}_{i+1}$ to the point p_k was calculated;

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

A.2.8

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

$$\delta_x = \frac{\delta}{d(N_{i+1}, p_k)} \times (x_{i+1} - px_k)$$
$$\delta_y = \frac{\delta}{d(N_{i+1}, p_k)} \times (y_{i+1} - py_k),$$

A.2.7

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

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

A.2.9

where P_{perp} is a collection of two points at distance δ from the point p_k. 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 (s_i);

$$\mathbf{S} = \mathbf{p}_i \in \left[A(s_i) \times 2m^2\right].$$

A.2.10

3.7.2. 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.2.11

3.7.3. 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 further 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 (-1.1 to 0.22). 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;

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

A.2.12

where $\mathbf{S}$ is the total point cloud and lm is the value assigned to a point $\mathbf{p}_i$, indicating the likelihood that the point is part of the road surface.

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;

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

A.2.13

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

The full point cloud $\mathbf{S}$ 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.

To find the angle $\theta_{\mathbf{K}}$, the difference in x and y coordinates between two nodes $\mathbf{N}_i$ and $\mathbf{N}_{i+1}$ was calculated as in Equation 3.1 to obtain the bearing angle between these nodes, represented by the grey line on Figure 3.3. Similarly, the difference in x and y coordinates were found for the perpendicular points associated with the sample. With this, $\theta_{\mathbf{K}}$ is given;

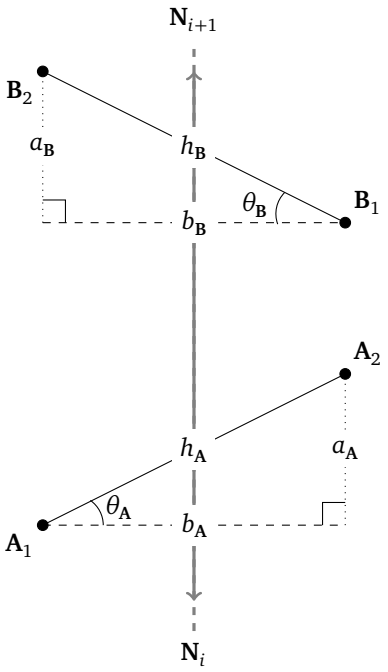

Fig. 3.3: *Road LiDAR points at maximum distance apart, showing two example sample locations (**A** and **B**). Road centreline represented by the thick grey line, as a line joining between two nodes. True road width is indicated by the dashed lines $b_\mathbf{A}$ and $b_\mathbf{B}$, considered to be the adjacent side of a triangle in relation to be bearing angles between the first and second point of each sample, $\theta_\mathbf{A}$ and $\theta_\mathbf{B}$. The distance between the two points per sample are considered to give the hypotenuse length of a triangle ($h_\mathbf{A}$ and $h_\mathbf{B}$).*

A.2.14

$$\theta_\mathbf{K} = \text{atan2}\,(|x|_l, |y|_l) - \text{atan2}\,(|x|_s, |y|_s)$$

A.2.15

where the difference in node coordinates gives $L = (|x|_l, |y|_l)$, and the difference in sample line coordinates gives $s = (|x|_s, |y|_s)$.

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 **K** is a single sample location), perpendicular to the road segment, considering the distance between the two points to be the hypotenuse of a right angled triangle (Figure 3.3; $|\mathbf{K}_1\mathbf{K}_2| \equiv h_\mathbf{K}$). The width $b_\mathbf{K}$ for each sample location was found, this may be expressed using trigonometry as;

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

A.2.16

where $b_\mathbf{K}$ gives the predicted width of a road at a sample **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.

3.7.4. 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, new road centrelines were derived 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;

$$\left(\frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2}\right). \tag{A.2.17}$$

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 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 in an attempt to improve the reliability of results. This may be expressed as;

$$\mathbf{S} = \left(s_i \in \left[p_s < 0.05\right]\right). \tag{A.2.12}$$

3.7.5. Final Model Analysis

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

$$\beta_p = \frac{\mathrm{Cov}\left(r_p, r_b\right)}{\mathrm{Var}\left(r_b\right)}. \tag{A.2.18}$$

3.7.6. Estimate of True Widths

QGIS (QGIS Development Team, 2019) was used to manually 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, reliability 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 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).
- **Road Width:** Extracted through linear probability classification of the road surface using LiDAR data (Section 3.7).

The value obtained from this is referred to as the Road Quality Index, presented in full on Table 4.1.

4. Results

RESULTS for the overall methodology are presented in this chapter, covering the initial preprocessing of LiDAR and other data, 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 improvements made 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 21.17 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 the standard deviation (125), with 99% of observations within the range of 1 to 417. As noted in previous LiDAR classification methods, intensity is often subject to noise, therefore a simplistic noise exclusion technique (Roussel and Auty, 2019) was implemented, as described in Section 3.2.4.

Following intensity noise filtering, the highest intensity value was now 746, with a standard deviation of 124. 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 similarly in the Luminescence values, with two peaks at around 50 and 120 (Figure 4.1 **(B)**).

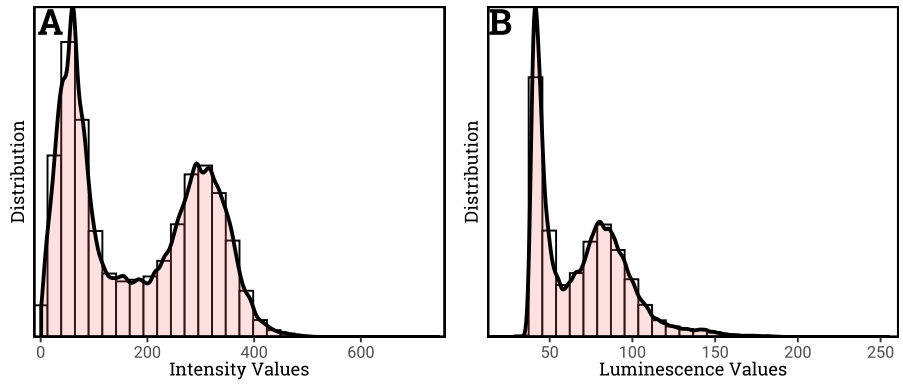

Fig. 4.1: *Post noise filtering LiDAR point cloud distribution; of **(A)** Intensity, and **(B)** Luminescence*

Figure 4.2 gives the results of further LiDAR preprocessing, comparing Figure 4.2 **(A)** and Figure 4.2 **(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 4.2 **(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 4.2 **(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). 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

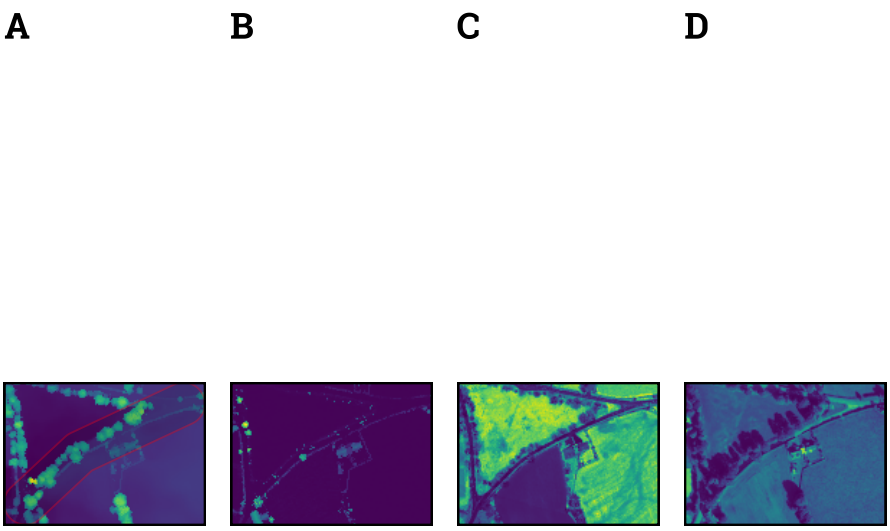

Fig. 4.2: *LiDAR point clouds for one selected road aggregated into 2m^2 grids;* ***(A)*** *Base point cloud z values, road location indicated with a 30m buffer* ***(B)*** *Normalised point cloud z values for only last returns (lpz)* ***(C)*** *Normalised point cloud $Intensity$ values for last returns,* ***(D)*** *Aerial data combined to 1 band*

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 (See **Appendix B**, 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 3.7.3, 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 4.3).

Figure 4.3 shows that there is a clear separation between the majority of points, and higher probability

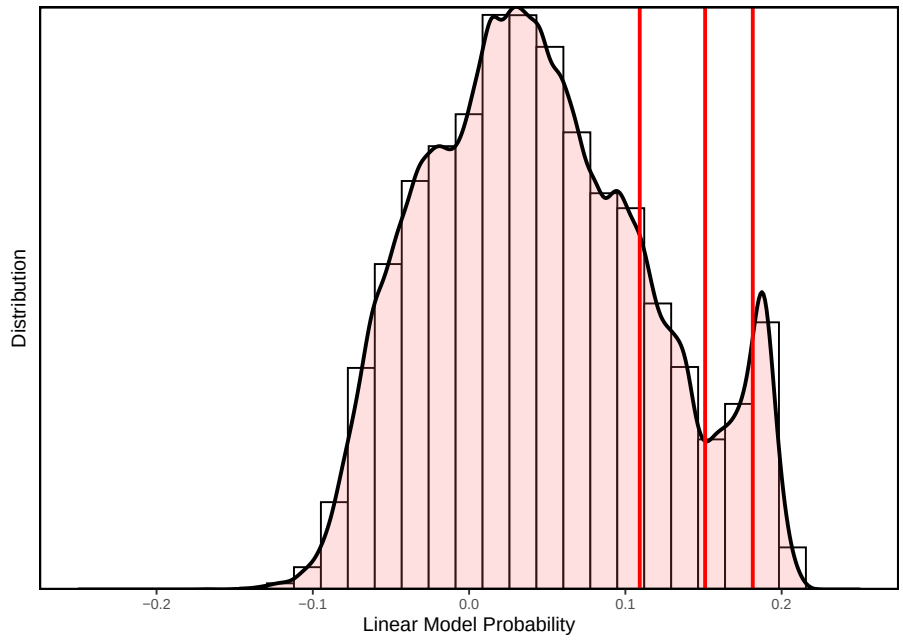

Fig. 4.3: *Linear model probability distributions for the maximal model (LM 1); showing vertical lines at the 95th, 90th, and 80th quantile of the distribution*

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 4.3.

Figure 4.4 reveals that qualitatively, the optimal choice for a quantile filtering of the linear probability distribution is likely the 95th quantile (Figure 4.4 **(A)**). However, observation of the southern section of Figure 4.4 **(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 4.4 **(A)** reveals that for the 95th quantile probability values, shadow from road hedgerows appears to reduce the model accuracy, as noticeable towards the centre of the road. For this reason, a second model was constructed for later comparison, which removes the *luminescence* information provided by the aerial imagery;

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

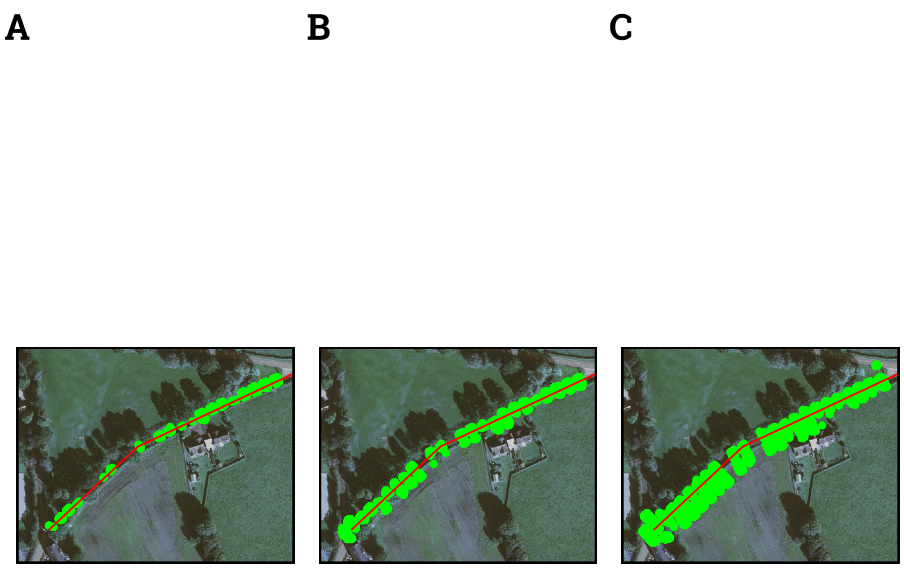

Fig. 4.4: *Comparison between linear prediction quantiles;* ***(A)*** *95th quantile,* ***(B)*** *90th quantile,* ***(C)*** *80th quantile.*

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 4.5 **(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 4.5 **(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 4.5 **(C)**. Figure 4.5 also gives information regarding the distance based noise exclusion technique, which has allowed for the exclusion of isolated points accurately on Figure 4.5 **(B)**. Improved centreline locations allowed for individual linear models (Figure 4.5 **(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 4.5 **(C)**

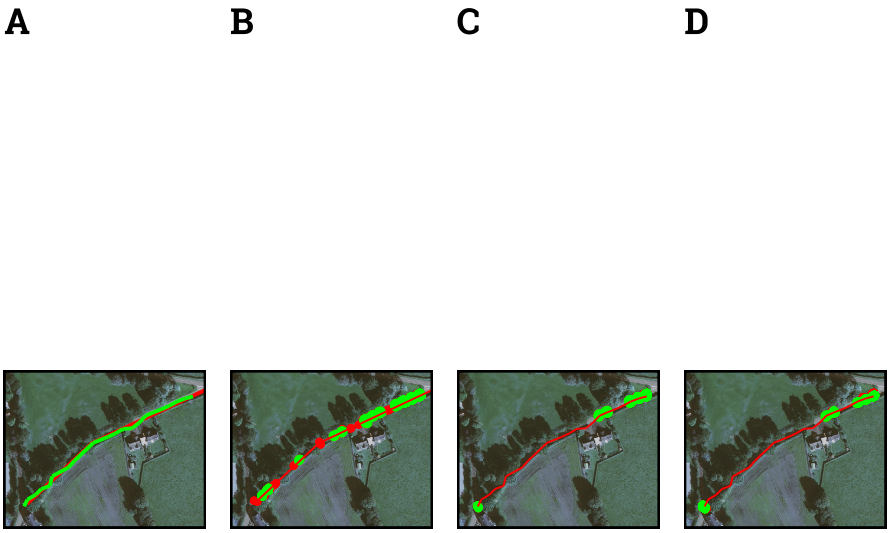

Fig. 4.5: *Comparison between original and derived centrelines showing differences in linear models; showing **(A)** Comparison between original (red) and derived road centrelines (green). **(B)** Second linear model (LM 2) applied to original centrelines. **(C)** Second linear model (LM 2) applied to derived centrelines. **(D)** Individual linear model (LM i) applied to derived centrelines. Each linear model probability quantile is selected as 95%, green points indicate road classified points, red indicate road classified points, removed through noise isolation filtering.*

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 **??**. The removal of luminescence 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, as such, it 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 being considered in other supervised road detection techniques.

Table **??** 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 that at present, the method for centreline improvements does not appear to improve road width estimates by much.

```
Error:  Cannot open "../data/osroads/roads_estwidth.gpkg"; The file doesn't seem to exist.

Error in merge(., road_est, by = c("road_id")):  object 'road_est' not found

Error in eval(lhs, parent, parent):  object 'widths' not found

Error in eval(lhs, parent, parent):  object 'widths' not found

Error in norm_means$road_id <- "Means:":  object 'norm_means' not found

Error in gsub("road_", "", norm_widths$road_id):  object 'norm_widths' not found

Error in eval(lhs, parent, parent):  object 'norm_widths' not found
```

```
Error in kable(norm_widths, format = "latex", booktabs = TRUE, digits = 2, :  object
'norm_widths' not found

Error in cat(c("\\begin{table}[!b]\n    \\begin{minipage}{.45\\linewidth}\n    \\caption{Model
Coefficients, Comparison between Linear Probability Models 1 and 2}\\label{tab:coeftest}\n
\\centering", :  object 't2' not found
```

4.6. Road Assessment

As LM i gave the highest mean accuracy for roads (Table **??**), it was selected for the final road width predictions. From this, Table 4.1 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 4.6). It appears to produce reliable results, as the road with the highest RQI is straighter and wider than the road with the lowest RQI, and is likely flat given it is neighboured by houses.

Table 4.1: *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 i)	Reliability (P_n/L)	RQI
14	B Road	0.01	2.41	155	3.84	1.94	0.34
11	B Road	0.01	7.49	62	4.03	0.96	0.34
13	B Road	0.00	7.55	167	3.63	0.83	-0.18
29	Minor Road	14.00	3.75	85	3.46	0.16	-0.21
25	Minor Road	14.58	6.95	125	3.77	0.78	-0.42
4	Minor Road	17.15	7.31	64	3.60	0.72	-0.43
28	Minor Road	9.51	3.84	101	2.94	0.39	-0.47
12	B Road	0.00	3.02	281	3.10	0.95	-0.54
9	B Road	0.06	7.10	299	3.66	3.30	-0.55
39	Minor Road	24.53	4.93	252	4.81	3.58	-0.58
8	B Road	0.00	3.62	313	2.86	0.62	-0.87
24	Minor Road	12.81	7.29	382	3.75	4.53	-1.19
15	Minor Road	10.33	11.04	251	3.28	0.53	-1.25
6	Minor Road	17.95	12.08	277	3.32	5.14	-1.64
16	Minor Road	18.16	11.33	309	3.40	4.05	-1.64
7	Minor Road	7.08	10.63	346	2.94	0.27	-1.66
2	Minor Road	18.51	15.85	370	3.87	2.03	-1.92
5	Minor Road	30.38	14.03	384	3.37	2.58	-2.49

Fig. 4.6: *Visual comparison between the RQI of roads.* ***(A)*** *highest RQI* ***(B)*** *lowest RQI*

5. Discussion

THE proposed methodology presents a road classification technique 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 road classification techniques 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, 2018b, 2013a; 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, and a per road width classification.

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 21.17 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, and without any preliminary removal of LiDAR points such as the 30m buffer from known road centrelines used in this dissertation.

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 a rural 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 and Crane, 2009), and almost all studies appear to focus on urban road classification (Li *et al.*, 2016; Vosselman, 2009; Zhao and 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 and Malpica, 2005), and exclude key features such as hedgerows, overhanging vegetation, with the road surface appearing fully distinct from neighbouring verges. 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 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 data collection technique employed included drive-through video recordings, not allowing for a scalable approach. Results also often lacked in accuracy, taking the road height variation from OS 10m contour lines. The method proposed in this dissertation 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 ground-based data, and instead uses aerial LiDAR which is more practically feasible to obtain for a comprehensive study. LiDAR has multiple use cases, meaning the national production of this data is well funded.

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, 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, Table 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, Table 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 and Dixon (2012) note that often this is not that case, and drivers often do not slow appreciably to account for these conditions. Therefore suggesting that speed limits should more accurately reflect the conditions of the road. Additionally, stopping sight distances observed by Layton and Dixon (2012) are significantly longer for larger vehicles such as trucks, and 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.

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 who are unable to drive, such as children, the elderly, and people with disabilities (Manthorpe *et al.*, 2008). Additionally, this reliance on private transport increases the minimum cost of living in rural communities, given car ownership is often considered mandatory but 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, 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 and Titheridge, 2009).

Transport accessibility is defined by the UK Government through journey time estimates for populations to

Table 5.1: *Recommended minimum Stopping Sight Distances at certain speeds (Layton and 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

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 vehicle 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 vehicle types, as the geometric features of certain roads likely do not allow for these vehicles. 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 result, 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 and 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 and 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 accuracy of this particular linear probability analysis, and often the accuracy 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.3.4), and the removal of calculated widths that are above 8m and below 2m (See **Appendix 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 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.

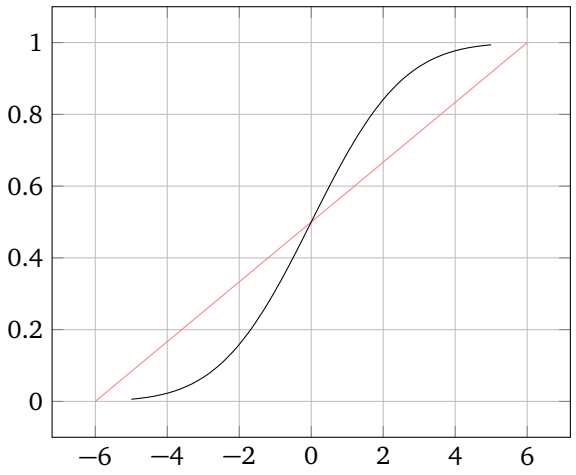

Fig. 5.1: *Comparison between a Linear Probability Model Distribution (Red) and Probit/Logit Cumulative Standard Normal Distribution (Black) (Approximation credit Bowling* et al.*, 2009))*

Alternative methods to reduce the noise produced in road classification 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 and Shan, 2009). These include Hough transforms (Vosselman *et al.*, 2004; Clode *et al.*, 2004), RANSAC (Smadja *et al.*, 2010; Sampath and Shan, 2008), and least square fitting (Matkan *et al.*, 2014).

Additionally, unsupervised LiDAR classification techniques have considered the texture of road and non-road, where man made structures often have smooth, regular height textures, and small height variations, while trees and other vegetation give an irregular height pattern, using this to separate man made structures, such as roads, from vegetation (Vosselman *et al.*, 2004; Darmawati, 2008).

Driving behaviour is affected by particular features that are not covered in this methodology, namely the delineation of road centrelines and edges (Steyvers and De Waard, 2000; Charlton *et al.*, 2018). The detection of these features is influenced by either 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.

5.4. 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.

Results generally provide insight into the overall road quality of individual roads in relation to each other, but work is required to more accurately extract key features in road quality assessment, particularly road widths. Further development of this method would provide road features for use in a national assessment of rural roads in Britain, to aid with improvements with the understanding of rural accessibility, the response time of emergency services, more appropriate speed limits, and rural road hierarchies.

Bibliography

Aarts L, van Schagen I (2006). "Driving Speed and the Risk of Road Crashes: A Review." *Accident Analysis & Prevention*, **38**(2), 215–224. ISSN 00014575. doi:10/djq87b.

Aquino J (2019). *nvimcom: Intermediate the Communication Between R and Either Neovim or Vim*.

Arnold JB (2019). *ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'*.

Axelsson P (1999). "Processing of Laser Scanner Data—Algorithms and Applications." *ISPRS Journal of Photogrammetry and Remote Sensing*, **54**(2-3), 138–147. ISSN 09242716. doi:10/c4g47v.

Azizi Z, Najafi A, Sadeghian S (2014). "Forest Road Detection Using LiDAR Data." *Journal of Forestry Research*, **25**(4), 975–980. ISSN 1007-662X, 1993-0607. doi:10/f6nx23.

Bache SM, Wickham H (2014). *magrittr: A Forward-Pipe Operator for R*.

Baruya A (1998). "MASTER: Speed-Accident Relationship on European Roads."

BBC (2012). "Plan for 40mph Country Road Limit." *BBC News*.

Bengtsson H (2019). *future: Unified Parallel and Distributed Processing in R for Everyone*.

Bivand R, Keitt T, Rowlingson B (2019). *rgdal: Bindings for the 'Geospatial' Data Abstraction Library*.

Bosona T, Gebresenbet G (2011). "Cluster Building and Logistics Network Integration of Local Food Supply Chain." *Biosystems Engineering*, **108**(4), 293–302. ISSN 15375110. doi:10/dg2mcd.

Bowling SR, Khasawneh MT, Kaewkuekool S, Cho BR (2009). "A Logistic Approximation to the Cumulative Normal Distribution." *Journal of Industrial Engineering and Management*, **2**(1), 114–127. ISSN 2013-0953. doi:10/b89r64.

Bridson R, Marino S, Fedkiw R (2005). "Simulation of Clothing with Folds and Wrinkles." In *ACM SIGGRAPH 2005 Courses on - SIGGRAPH '05*, p. 3. ACM Press, Los Angeles, California. doi:10/c8qsq9.

Charaniya A, Manduchi R, Lodha S (2004). "Supervised Parametric Classification of Aerial LiDAR Data." In *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pp. 30–30. IEEE, Washington, DC, USA. doi:10/d2qv8m.

Charlton SG, Starkey NJ, Malhotra N (2018). "Using Road Markings as a Continuous Cue for Speed Choice." *Accident Analysis & Prevention*, **117**, 288–297. ISSN 00014575. doi:10/gdvrxj.

Clode S, Kootsookos P, Rottensteiner F (2004). "The Automatic Extractin of Roads from LiDAR Data." p. 7.

Corben B, Oxley J, Koppel S, Johnston I (2005). "Cost-Effective Measures to Improve Crash and Injury Risk at Rural Intersections." p. 10.

Currie G (2010). "Quantifying Spatial Gaps in Public Transport Supply Based on Social Needs." *Journal of Transport Geography*, **18**(1), 31–41. ISSN 09666923. doi:10/dmv2nd.

Darmawati A (2008). "Utilization of Multiple Echo Information for Classification of Airborne Laser Scanning Data." ITC.

Deeter D (2009). "Real-Time Traveler Information Systems. NCHRP Report 399." *Transport Research Board, USA*.

Department for Transport (2006). "Speed Assessment Framework." http://www2.dft.gov.uk/.

Department for Transport (2011). "Strategic Framework for Road Safety."

Department for Transport (2012). "Guidance on Road Classification and the Primary Route Network." p. 26.

Department for Transport (2013a). "Setting Local Speed Limits." p. 42.

Department for Transport (2013b). "The Speed Limit Appraisal Tool: User Guidance." p. 93.

Department for Transport (2016). "Overall Measure of Accessibility of Services."

Department for Transport (2018a). "Journey Time Statistics."

Department for Transport (2018b). "Road Safety Management Capacity Review."

Department for Transport (2019). "Road Traffic Statistics - Summary Statistics." https://roadtraffic.dft.gov.uk/summary.

Dowle M, Srinivasan A (2019). *data.table: Extension of 'data.frame'*.

Dubes R, Ohanian P (1992). "Performance Evaluation for Four Classes of Textural Features." *Pattern Recognition*, **25**, 819–833. doi:10/cx7ktb.

Elberink SO, Maas HG (2000). "The Use of Anisotropic Height Texture Measures for the Segmentation of Airborne Laser Scanner Data." p. 8.

Environment Agency (2019). "LiDAR." https://data.gov.uk/dataset/977a4ca4-1759-4f26-baa7-b566bd7ca7bf/lidar-point-cloud.

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. ISSN 02779536. doi:10/dq6rh3.

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*, volume 2, pp. II–201. doi:10/fjcjsn.

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. ISSN 09242716. doi:10/gddkjk.

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

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)*, pp. 809–818. Taylor & Francis Group.

Francois R (2017). *bibtex: Bibtex Parser*.

Garnier S (2018a). *viridis: Default Color Maps from 'matplotlib'*.

Garnier S (2018b). *viridisLite: Default Color Maps from 'matplotlib' (Lite Version)*.

Gillespie C (2019). *benchmarkme: Crowd Sourced System Benchmarks*.

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. ISSN 0143-1161, 1366-5901. doi:10/gf6d37.

Hanck C, Arnold M, Gerber A, Schmelzer M (2019). "Introduction to Econometrics with R." p. 392.

Harrell Jr FE, with contributions from Charles Dupont, many others (2019). *Hmisc: Harrell Miscellaneous*.

Hatger C (2005). "Road Extraction by Use of Airborne Laser Scanner Data." p. 19.

Highways England (2016). "Letter in Response to Road Width Restrictions. FOI: 734,857."

Highways England (2019). "Network Management."

Hijmans RJ (2019). *raster: Geographic Data Analysis and Modeling*.

Hu X, Tao CV, Hu Y (2004). "Automatic Road Extraction from Dense Urban Area by Integrated Processing of Height Resolution Imagery and LiDAR Data." p. 5.

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*, pp. II–1140–II–1143. IEEE, Boston, MA, USA. ISBN 978-1-4244-2807-6. doi:10/cmb4tq.

Ihaka R (2011). "Rnoweb: Literate Programming with and for R." p. 5.

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. ISSN 07430167. doi:10/d5r24k.

INRIX (2019). "INRIX." http://inrix.com/.

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. ISSN 1424-8220. doi:10/d2p8gr.

Karthikeyan P, Rnaganatan P (2001). "Tutorial on Cloth Modelling." *ACM Student Tutorial Contest, India*.

Kassambara A (2019). *ggpubr: 'ggplot2' Based Publication Ready Plots*.

Knuth DE (1984). "Literate Programming." *The Computer Journal*, **27**(2), 97–111. ISSN 1460-2067. doi:10/c73jjw.

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. ISSN 09242716. doi:10/c37dsh.

Kumar P, McElhinney CP, 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. ISSN 09242716. doi:10/f5gjhk.

Layton R, Dixon K (2012). "Stopping Sight Distance." *Kiewit Center for Infrastructure and Transportation, Oregon Department of Transportation*.

Leutner B, Horning N, Schwalb-Willmann J (2019). *RStoolbox: Tools for Remote Sensing Data Analysis*.

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. ISSN 2194-9034. doi:10/gf3tdf.

Maas HG (1999). "The Potential of Height Texture Measures for the Segmentation of Airborne Laserscanner Data." p. 8.

Mahmoudian M (2019). *varhandle: Functions for Robust Variable Handling*.

Malecki E (2003). "Digital Development in Rural Areas: Potentials and Pitfalls." *Journal of Rural Studies*, **19**, 201–214. doi:10/fhms4b.

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. ISSN 0966-0410. doi:10/ct29jh.

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. ISSN 0308-518X. doi:10/d8b774.

Matkan AA, Hajeb M, Sadeghian S (2014). "Road Extraction from Lidar Data Using Support Vector Machine Classification." *Photogrammetric Engineering & Remote Sensing*, **80**(5), 409–422. ISSN 00991112. doi:10/f52cdx.

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. ISSN 01678655. doi:10/dhn6ck.

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*, pp. 223–237. Public Works Department, Kuala Lumpur, Malaysia.

Mulley C, Nelson JD (2009). "Flexible Transport Services: A New Market Opportunity for Public Transport." *Research in Transportation Economics*, **25**(1), 39–45. ISSN 07398859. doi: 10/chspr5.

Mungall I (2005). "Trend towards Centralisation of Hospital Services,."

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.

Openshaw S (1984). "Ecological Fallacies and the Analysis of Areal Census Data." *Environment and Planning A: Economy and Space*, **16**(1), 17–31. ISSN 0308-518X, 1472-3409. doi:10/bdn84m.

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. ISSN 09658564. doi:10/bkqhnk.

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

Peterson BG, Carl P (2019). *PerformanceAnalytics: Econometric Tools for Performance and Risk Analysis*.

Peterson RA, Brown SP (2005). "On the Use of Beta Coefficients in Meta-Analysis." *Journal of Applied Psychology*, **90**(1), 175–181. ISSN 1939-1854, 0021-9010. doi:10/fksgxd.

QGIS Development Team (2019). *QGIS Geographic Information System*. Open Source Geospatial Foundation.

Qiu Y, authors/contributors of the included software See file AUTHORS for details (2019). *showtext: Using Fonts More Easily in R Graphs*.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ram K, Wickham H (2018). *wesanderson: A Wes Anderson Palette Generator*.

Richards D, Cuerden R (2009). "The Relationship between Speed and Car Driver Injury Severity." p. 16.

Riva M, Curtis S, Norman P (2011). "Residential Mobility within England and Urban–Rural Inequalities in Mortality." *Social Science & Medicine*, **73**(12), 1698–1706. ISSN 02779536. doi: 10/c5wkhn.

Road Safety and Environment (2000). "New Directions in Speed Management: A Review of Policy."

Robinson D, Hayes A (2019). *broom: Convert Statistical Analysis Objects into Tidy Tibbles*.

Robinson SP (2008). "Determining London Bus Stop Locations by Means of an Automatic Vehicle Location System." *Transportation Research Record*, **2064**(1), 24–32. ISSN 0361-1981. doi: 10/bj5m3d.

Rottensteiner F, Trinder J, Clode S, Kubik K (2003). "Building Detection Using LIDAR Data and Multi-Spectral Images." p. 10.

Roussel JR, Auty D (2019). *lidR: Airborne LiDAR Data Manipulation and Visualization for Forestry Applications*.

Saeedi S, Samadzadegan F, El-Sheimy N (2009). "Object Extraction from LiDAR Data Using an Artificial Swarm Bee Colony Clustering Algorithm." p. 6.

Sampath A, Shan J (2008). "Building Roof Segmetnation and Reconstruction from LiDAR Point Clouds Using Clustering Techniques." p. 6.

Smadja L, Ninot J, Gavrilovic T (2010). "Road Extraction and Environment Interpretation from LiDAR Sensors." p. 6.

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. ISSN 09666923. doi:10/f3w23n.

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*.

Steyvers FJJM, De Waard D (2000). "Road-Edge Delineation in Rural Areas: Effects on Driving Behaviour." *Ergonomics*, **43**(2), 223–238. ISSN 0014-0139. doi:10/ddvg7m.

Taylor MC, Baruya A, Kennedy JV (2002). "The Relationship between Speed and Accidents on Rural Single-Carriageway Roads." p. 32.

Taylor MC, Lynam DA, Baruya A (2000). "The Effects of Drivers' Speed on the Frequency of Road Accidents." p. 56.

The University of Edinburgh (2019). "Aerial Digimap Educational User Licence." https://digimap.edina.ac.uk/.

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.

UK Government (2019a). "Find Open Data - Data.Gov.Uk." https://data.gov.uk/.

UK Government (2019b). "Speed Limits." https://www.gov.uk/speed-limits.

Viner H, Sinhal R, Parry T (2004). "Review of UK Skid Resistance Policy." *Preprint SURF*.

Vosselman G (2000). "Slope Based Filtering of Laser Altimetry Data." p. 9.

Vosselman G (2009). "Advanced Point Cloud Processing." p. 10.

Vosselman G, Gorte BG, 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.

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.

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)*, volume 2, pp. 203–207. doi:10/cpf8x8.

Wang J, Shan J (2009). "Segmentation of LiDAR Point Clouds for Building Extraction." p. 11.

Wickham H (2014). *Advanced r*. Chapman and Hall/CRC. ISBN 1-4665-8697-4.

Wickham H (2017). *tidyverse: Easily Install and Load the 'Tidyverse'*.

Wickham H (2018). *scales: Scale Functions for Visualization*.

Wickham H, Hester J, Chang W (2019). *devtools: Tools to Make Developing R Packages Easier*.

Wilke CO (2019). *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*.

Williamson P (????). *ENVS450: Helper functions for ENVS450*.

Wu KF, Donnell ET, Himes SC, 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. ISSN 0733-947X, 1943-5436. doi:10/f4575w.

Yadav M, Lohani B, Singh AK (2018). "Road Surface Detection from Mobile LiDAR Data." *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, **IV-5**, 95–101. ISSN 2194-9050. doi:10/gf3tdd.

Yoon J, Crane CD (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.

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. ISSN 0196-2892. doi:10/dv3889.

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. ISSN 2072-4292. doi:10/gftcvs.

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. ISSN 2072-4292. doi:10/gd9j5f.

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*, pp. 9–16. IEEE, Providence, RI, USA. ISBN 978-1-4673-1612-5 978-1-4673-1611-8 978-1-4673-1610-1. doi:10/gf3tc9.

Zhu H (2019). *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*.

Word Count: 12070

A. Environment and Functions

All code is hosted on my personal GitHub account, along with 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.

This dissertation 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.

A.1. Packages and Machine Environment

```
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-r0.3.7.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_1.12.6
 [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_1.4-7
[11] future_1.14.0              varhandle_2.0.4
[13] forcats_0.4.0              stringr_1.4.0
[15] dplyr_0.8.3                purrr_0.3.3
[17] readr_1.3.1                tidyr_1.0.0
[19] tibble_2.1.3               tidyverse_1.2.1
[21] lidR_2.1.4                 raster_3.0-7
[23] sp_1.3-1                   scales_1.0.0
[25] kableExtra_1.1.0           sf_0.8-0
[27] ggpubr_0.2.3               magrittr_1.5
```

[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."*

```
[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_2.2.1             usethis_1.5.1
[43] ENVS450_0.1.0              ggplot2_3.2.1
[45] pacman_0.5.1               knitr_1.25
[47] nvimcom_0.9-83
```

A.2. Referenced 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. 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.2.3. 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.2.4. Combine Catalog

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

A.2.5. 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
      # same as + 2pi
      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.2.6. 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.2.7. 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.2.8. 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.2.9. 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.2.10. 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.2.11. 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 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.2.14. 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.2.15. Adjacent Length

```
# use atan2 to find true width of roads given
# a non perpendicular line, convert to perpendicular to find width
adjacent_length <- function(samp, cent) {
  tot_width <- c()
  cent <- cent %>% st_cast("POINT")
  n <- nrow(cent) - 1
  nodelines <- lapply(X = 1:n, FUN = function(i) {
    pair <- cent[c(i, i + 1), ] %>%
      st_combine()
    line <- st_cast(pair, "LINESTRING")
    return(line)
  })

  samp <- samp %>%
    mutate(row_id = row_number())
  samp <- split(samp, samp$row_id)

  for (n in nodelines) {
    for (s in samp) {
      # find which centreline it is associated with
      # as road consist of multiple
      int <- as.numeric(st_crosses(n, s))
      int[is.na(int)] <- 0
      # with correct line, find perpendicular angle
      # and length
      if (int == 1) {
        n1 <- st_coordinates(n)[1, ]
        n2 <- st_coordinates(n)[2, ]
        x <- n1[1] - n2[1]
        y <- n1[2] - n2[2]
        ang_rad <- atan2(y, x)
        ang_deg <- ang_rad * 180 / pi

        n1 <- st_coordinates(s)[1, ]
        n2 <- st_coordinates(s)[2, ]
        x <- n1[1] - n2[1]
        y <- n1[2] - n2[2]

        ang_rad <- atan2(y, x)
        ang_deg_c <- ang_rad * 180 / pi

        theta <- abs(ang_deg) - abs(ang_deg_c)

        c1_len <- st_length(s)
        # pythagoras to find adjacent line length
      # left of N same as right of N
      # same as + 2pi
        adjacent <- abs(as.numeric(c1_len) * cos(as.numeric(theta)))
        adjacent <- cbind(
          adjacent, as.character(unique(cent$road_id)),
          as.character(unique(cent$sample_id))
        )
        tot_width <- rbind(tot_width, adjacent)
      }
    }
  }
  return(tot_width)
}
```

A.2.16. Find Distances

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

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. 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)
}
```

A.3. Additional Functions

A.3.1. 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.3.2. 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.3.3. 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.3.4. 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.3.5. 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.3.6. 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)
}
```

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**
ReturnNumber	-0.01271 **	-0.020	-0.015
NumberOfReturns	**-0.01271 ****	**-0.020**	**-0.015**
ScanDirectionFlag	**-0.00402 ****	**-0.007**	**-0.002**
lum	-0.00142 **	0.013	0.018
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
5	8	0.36	22.29
7	4	0.19	20.79
10	2	0.11	18.77
4	4	0.21	18.68
9	4	0.26	15.34
25	4	0.28	14.27
20	1	0.07	13.34
1	3	0.24	12.59
29	3	0.24	12.42
17	4	0.38	10.57
14	1	0.11	9.05
24	3	0.38	7.94
2	2	0.25	7.93
26	1	0.13	7.82
15	1	0.13	7.74
11	1	0.13	7.58
6	2	0.26	7.58
16	4	0.54	7.37
8	1	0.17	5.96
34	1	0.18	5.62
39	2	0.41	4.85
27	1	0.21	4.85
32	1	0.23	4.36
30	1	0.24	4.16
12	1	0.24	4.15
28	2	0.74	2.70
13	0	0.18	0.00
31	0	0.08	0.00
33	0	0.08	0.00
41	0	0.07	0.00

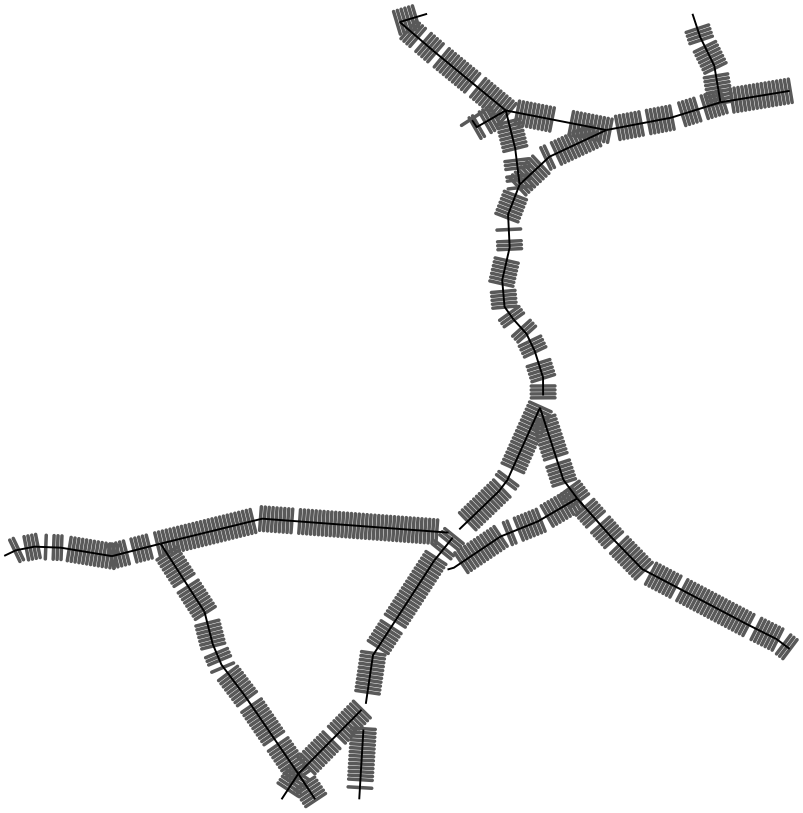

Fig. B.1: *Sample lines extracted based on known road locations*

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_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")
plot(roads)

# 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.shp", 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 Improved 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()

## ---- asdf
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")
```