THE UNIVERSITY *of* EDINBURGH

Exploration of walking speed prediction: a data-driven approach

Andrew Wood

Doctor of Philosophy

THE UNIVERSITY OF EDINBURGH

2022

Abstract

Hikers and hillwalkers typically use the gradient in the direction of travel (walking slope) as the main variable in established methods for predicting walking speeds along a route. Research into fell-running has suggested further variables which impact speed in this context. Recent improvements in data availability, as well as widespread use of GPS tracking now make it possible to test these variables on a large scale. Here we tested various models used to predict walking speed against public GPS data from almost 93,000 km of UK walking / hiking tracks. Tracks were filtered to remove breaks and non-walking sections. A generalised linear model (GLM) was found to be most accurate at determining walking speeds. Key differences between the GLM and commonly used rules were that the GLM considered the gradient of the terrain (hill slope) irrespective of walking slope, as well as the terrain type and level of terrain obstruction in off-road travel. All of these factors were shown to be highly significant, and this is supported by a lower root-mean-square-error compared to existing functions, particularly in the areas where the majority of travel occurs. We also noted an increase in RMSE between the GLM and established methods as hill slope increases, further exemplifying the importance of this variable. As well as providing a new walking speed formula, the underlying dataset can be used in future work to test alternate models.

Lay Summary

Predicting the time taken to travel a walking route on a map is needed in a wide range of scenarios, from planning a day's journey through to vital emergency response. While there are many variables which affect walking speed, a large number of these depend on who is participating in the hike, and what the weather conditions are at the time. As such, existing methods use an average walking speed based on fixed variables, and users then determine whether they will walk faster or slower based on their personal circumstances.

In fell-running, three fixed variables have been suggested to affect speeds. These are the the slope in the direction of travel (walking slope), the gradient of the terrain (hill slope) and the level of terrain obstruction. However, these variables have not all been investigated to ascertain their effect on walking speeds, and the most commonly used methods rely on the walking slope as the main variable to to estimate walking speed.

In this work we investigate the impact that the remaining two variables (hill slope and terrain obstruction) have on walking speeds, and build a new model for walking speeds which takes all three variables into account.

To build a robust model, we need to use a large volume of data which covers the full scope of possible terrain types, and is from a wide enough range of participants that individual factors which affect the walking speed are averaged out. In order to do this, we make use of publicly available GPS data which has been collected by users of Hikr.org and OpenStreetMap.org, and which we filter to remove non-walking sections.

We begin by exploring the inclusion of hill slope and walking slope in a walking speed model, using GPS data from across Scotland. We find that both factors are highly significant when predicting the walking speeds. In order to validate this model, we extend it to data from across the rest of the UK. When doing this we find that the results seen for Scotland are different to those in the rest of the UK. To look into this further, we explore whether the differences in models can be explained by differences in terrain types in the data of the two regions.

We find that while walking along paved roads, there is no significant difference between Scotland and the rest of the UK. However, while on unpaved roads, or in off-road conditions, the data for Scotland is at the extreme end of what one may expect to see throughout the rest of the UK. We are not sure what causes this difference, but believe that it may be due to environmental factors such as the weather, which make hiking more difficult (and therefore slower) in Scotland.

Following this, we continue to explore the effect of terrain on walking speeds, looking at the level of obstruction in off-path regions. We find that this has a large effect on walking speeds, with just 10 cm of terrain obstruction (i.e. walking through low-level vegetation) reducing walking speeds by over 0.5 km/h.

We ultimately end up with a model for predicting walking speeds which takes into account all three of the variables which have been proposed to affect fell running speeds; the walking slope, the hill slope, and the terrain type (including the road type for on-road travel, or the level of obstruction for off-road travel).

Finally, we conduct a fieldwork study in which we can test our model under more controlled conditions. We find that our model for walking speed performs better than the current most commonly used functions for predicting hiking speeds, particularly in low-slope regions where most walking occurs.

While the final model produced is computationally complex (and thus not suited to calculating walking speeds by hand), the impact that each of the three variables has on walking speeds can be used in addition to existing methods for calculating walking speed when selecting a route while on a hike. For example, a route traversing a steep hill is likely to take longer than would be predicted by existing methods (as they do not account for hill slope), and alternative routes may be preferred. Thus, our new model for walking speed can be beneficial to hikers both while initially planning a route (when the full model can be used), and while on the hill.

Acknowledgements

I would like to thank my supervisors Douglas and William, for their help and advice throughout my project.

I would also like to thank my friends and family for their constant support and encouragement over the last few years.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Andrew Wood

Contents

Figures and Tables

Figures

Tables

Chapter 1

Introduction

1.1 Background

Knowing how fast people are able to walk between locations is key information in many situations. This knowledge is used everyday when using apps to provide directions; they take an estimate for walking speed to calculate the optimal route to present to the user. In hiking and hillwalking scenarios, this information is even more vital for safety reasons: if you are leaving in the morning for a hike then it is standard practice to provide an estimated return time such that emergency services can be contacted if you get into difficulty and do not return. An inaccurate estimate for how long a route will take could lead to unnecessary callouts, or delay a callout in a situation where every minute is important. Furthermore, in circumstances where a hiker has gone missing, an accurate measure of walking speed can help to restrict a potential search area around a last known location. Finally, when out on a hike there are situations where hikers may be deciding whether to follow a footpath, or take a more direct cross-country route. Accurate estimates of the walking speed and time for both scenarios are required to be able to select the optimal route.

There are a multitude of factors which can impact the walking speed and time predictions for a route, although these can generally be split into two categories. The first is the individual effects which depend on who precisely is undertaking the walk and when they are doing it. These include the group size (larger groups often walk slower), age or fitness of the participants, and weather conditions, as well as the aim of the walk (afternoon stroll vs. specific hike). The second category covers the fixed effects which will be consistent across all individuals who attempt such a given route. These include factors such as how steep the terrain is, or whether the route is following a road.

As most of the individual effects cannot be modelled without knowing information about the person who is planning a route, most existing hiking route planners calculate the walking speed based solely on the terrain, and this is presented as the average time (or time range) it takes to complete a hike. It is then left up to the individual to allow more or less time for a hike given their knowledge about personal ability and circumstances.

Formulae of varying complexity have been proposed to estimate human walking speed or time along a projected path. A popular early method that is still widely used was put forward by Naismith (1892) which calculates walking time under normal conditions as:

> "*an hour for every three miles on the map, with an additional hour for every 2,000 feet of ascent.*"

This approximates to a walking speed of 5 km/h with 10 minutes added on for every 100 m of ascent. Naismith's rule is still widely used today by Scout groups and other casual hikers due to the ease of calculating the walking time by hand using a paper map.

Although Naismith's rule is still very widely used it does have a well-known downside; namely the fact that is does not predict a reduced walking speed regardless of how steep a descent the user is on. As such, a number of updates to Naismith's rule have been proposed over the years, with the aim of improving the accuracy of walking speed predictions. Aitken (1977) introduced a reduced base movement speed of 4 km/h on surfaces which are not paths or roads, and Langmuir (1984) included extra terms to account for descents. Langmuir put forward that walking time should be calculated as per Naismith's rule (with Aitken's reduced off-path speed), and:

- 10 minutes should be added per 300 m of descent at an angle greater than 12 degrees.
- 10 minutes should be subtracted per 300 m of descent at an angle between 5 and 12 degrees.

Langmuir's rule does predict lower speeds on steep descents but also suggests a top speed of 12 km/h on shallow descents, which is much faster than can be achieved. Furthermore, it also implies a sudden jump from 12 km/h down to 3 km/h on slightly steeper slopes which is unrealistic.

An alternative hiking function proposed by Tobler (1993), has become more popular in recent research and situations where speeds do not need to be calculated by hand:

$$W = 6 * exp(-3.5|S + 0.05|),$$

where

W = velocity (km/h)
S = gradient of slope.

Like Naismith's rule, this gives a speed of 5 km/h on flat ground, though the maximum speed is 6 km/h on a mild descent (around 3 degrees). In a similar manner to Aitken's correction, a factor of 0.6 is applied to the calculated speed for all off-road travel. Tobler's function does avoid the issues seen in Naismith's and Langmuir's functions, but does predict a sharp peak in walking speed, which may be unrealistic. The formulae discussed here are all shown in Figure 1.1.

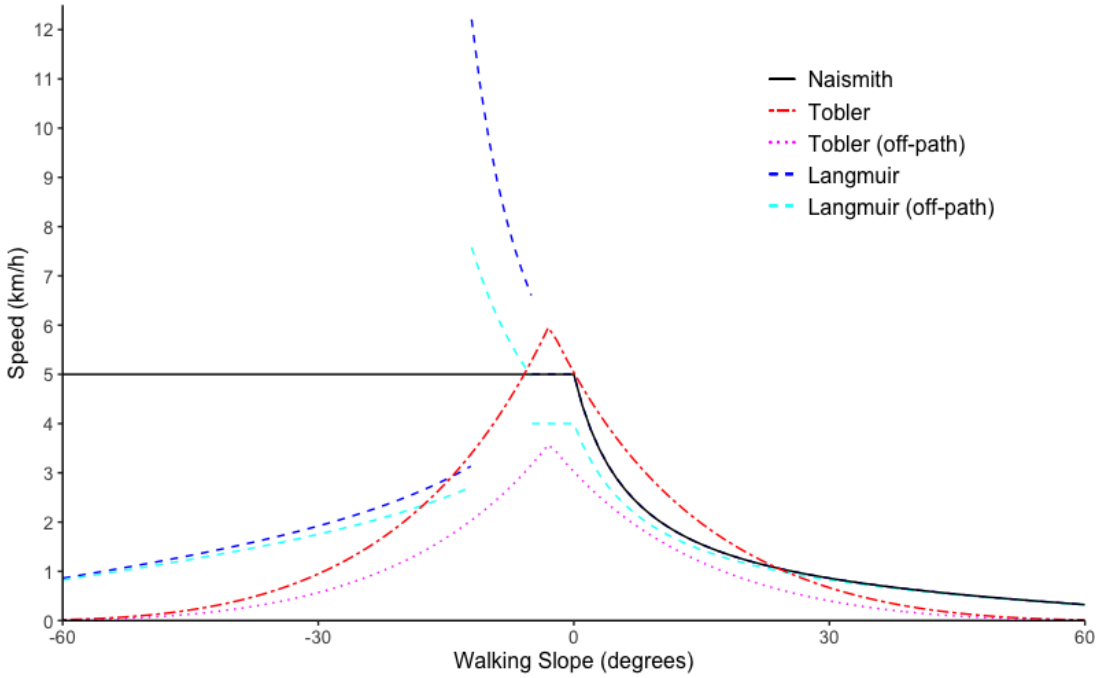


Figure 1.1: The most commonly used functions to calculate walking speed as a function of walking slope angle: Naismith's rule (black), Langmuir's rule (blue, cyan) and Tobler's hiking function (red, magenta).

A number of other works have also looked at providing alternative methods to calculate walking speeds (Campbell, Dennison, Butler, & Page, 2019; Davey, Hayes, & Norman, 1994; Irmischer & Clarke, 2018; Rees, 2004). The majority of these are small scale studies using either very few of individuals, or taking data from a limited area (where a wide range of terrain is not encountered), so their applicability to a wider population is not clear. The work by Campbell et al. (2019) uses a crowdsourced dataset from Strava, Inc. (2022), however there are a number of problems which come with this. The data used was anonymised and aggregated for each segment of the trail, and there was no separation between data recorded while walking, jogging or running. This lack of separation shows up in the resulting model. When simulating predicted travel time for a given hiking trail, the existing hiking functions (Naismith and Tobler) match most closely to just the 5th percentile of speeds predicted by Campbell et al.. This clearly suggests that the majority of the data which was used in this work comes from faster travel methods (i.e jogging or running), and so is not relevant to calculating walking speeds. A further issue resulting from the use of the anonymised data is that it is not possible to follow a single individual's path from start to finish. This has the knock-on effect that it is not possible to identify points along the route where a user may have stopped for a break. This will have resulted in underestimated movement speeds, as break points will be included in the data for a given trail segment. Furthermore, breaks will not be evenly distributed throughout the data. They are likely to occur at specific locations such as

viewpoints, and are more likely to be taken when hiking than when running or jogging, due to the nature of each activity. It is therefore difficult to use this dataset specifically to predict hiking speeds, as the breaks and other non-walking data will have confounding effects on the results.

A further drawback of all of these proposed alternate methods to Naismith's rule is that they continue to use walking slope as the main variable to determine walking speed (with simple multiplicative factors applied for off-road travel), and other variables which may affect the walking speed have not been considered. When exploring speeds of fell-runners, Arnet (2009) suggests that the movement velocity is dependent on three factors: obstruction (with different factors applied depending on the kind of obstruction), ascent in the run direction (walking slope) and slope of the terrain (hill slope). The variable values used in Arnet's calculations cannot be directly applied to walking speeds as they were based on orienteering championships which are very different situations.

Experience tells us that traversing on a steep hill (while maintaining constant elevation) is more difficult than traversing flat ground. However, the existing methods estimate the same walking speed for both situations. Similarly, high levels of terrain obstruction in off-road areas (such as a thick gorse bush) are much more difficult to walk through than empty fields. The simple multipliers for off-road travel in Aitken's correction and Tobler's function do not provide any distinction between two such regions.

N. J. Wood and Schmidtlein (2012), in which all three of Arnet's factors were taken into account, looked at evacuating citizens in the event of a hurricane. They applied Tobler's function to both the hill slopes and walking slopes, and calculated the terrain obstruction coefficients based on energy usage rather than walking speed (using Soule and Goldman (1972)). They accepted that these were likely not the correct values, but they were unable to find any better alternatives. Campbell, Dennison, and Butler (2017) conducted a study to explore the effects of ground roughness and vegetation density on firefighter evacuation speeds, but they did not consider the hill slope separately.

As demonstrated by Campbell et al. (2019), the rise in use of GPS tracking in recent years means that a data-driven approach to modelling the walking speed is now possible. This makes it possible to access GPS tracks from a wide variety of regions and terrains, without the difficulty of organising large-scale route recording. A second benefit of utilising GPS data (which was not available for Campbell et al. (2019) due to their use of Strava data) is that tracks can easily be broken down into individual sections, enabling specific route features to be investigated.

There are a number of factors which have also been previously observed, and should be taken into consideration when exploring walking speeds. While existing formulae suggest that walking on very steep slopes is achievable, in practice the majority of slopes encountered day-to-day rarely exceed 10 degrees (Proffitt, Bhalla, Gossweiler, & Midgett, 1995). It is therefore most important to provide accurate walking speed predictions in this region, as it will be of greatest practical use.

Pitman, Zanker, Gamper, and Andritsos (2012) looked at producing personalised walking speed predictions based on current progress along a hike. Instead of looking at the value of the walking speed, they investigate how an individual's speed changes over the course of a route. Most of these features require knowledge about current progress along a route to determine the walking speed, however they do suggest that, in general, the walking speed is faster on longer routes, with the explanation that these are only performed by more experienced, faster, hikers. Interestingly, they also found a decrease in walking speed on medium length walks (7 - 17 km), which we can explore.

A further feature which has been noted in fell-running and hillwalking is the existence of a 'critical gradient'; the angle at which it is faster to zig-zag up a hill, rather than ascend directly. This was first identified as occurring at approximately 17 - 20 degrees based on treadmill experiments for fell-runners (Davey et al., 1994). More recent work also looking at fell-running data has found the critical gradient at a similar point (gradients between '0.276' and '0.382', or 15 – 21 degrees) (Kay, 2012). There also evidence to suggest that a critical gradient occurs when walking, and that it occurs at approximately the same point: Balstrøm (2002) says that 40% (approximately 22 degree) slopes are manageable, but hairpins are generally found in paths starting at 30% (approximately 17 degrees). Furthermore, Llobera and Sluckin (2007) found a critical gradient of '0.287' (approximately 16 degrees), when exploring energy usage rather than walking speeds. Based on this evidence, a new model for walking speed should predict the critical gradient to occur in this region.

In this work we aim to use a data-driven approach to explore the impact of all three factors discussed by Arnet on walking speeds. These are the walking slope, the hill slope and the terrain obstruction. By including all three of these variables, we will produce a model for walking speed which is more accurate than existing methods in a wider range of circumstances (such as when traversing a hill). Further, the majority of the existing methods were created from small sample sizes. A crowdsourced model built on data from a wide variety of people and terrains will improve the reliability of walking speed estimates compared to these existing methods. Unlike Campbell et al. (2019), we aim to include only walking or hiking data in our dataset, to improve the accuracy of the resulting model. The downside of using a crowdsourced dataset is that we will have very little control over what data is collected. This means that the data will be biased towards more popular walking areas, so very large quantities of data are required to ensure that less-travelled terrain conditions are taken into

account by the model. On top of the data bias, we will not have information regarding the conditions the walks took place in, as GPS tracks do not typically include any metadata regarding the individual participants. This means that a model based on crowdsourced data is unable to take into account any individual effects which may impact the walking speed. Instead, we will aim to use a sufficient volume of data so as to determine the walking speed for an average individual solely based on the non-temporal environmental conditions. It will then be up to users to determine their own adjustments to this speed according to personal circumstance. This is in line with how many existing route planners already function. Providing more accurate and reliable estimates for the average walking speed will allow users to make more informed choices about their route selection, and thus improve the safety and enjoyment of their hike.

Chapter 2

Materials and Methods

2.1 Methodology

In order to build an improved model for walking speeds we needed data. This had to provide us with the walking speeds people achieved over a wide variety of terrains (to give us sufficient information to build our model), and also record the speeds achieved by a mixed range of individuals across multiple weather conditions to account for the individual variances in speed discussed in Section 1.1. Although it would have been possible to manually collect this data, it would be very challenging to collect a large and diverse enough sample from which to build a robust model. For this reason, we used existing data tracks which have been uploaded to the internet by hikers and hillwalkers, and are publicly available to access.

The tracks used had to contain enough datapoints along their length to enable us to calculate walking speeds for small, specific sections of the route. Therefore tracks which only provided the total distance and time taken were not useful to us. By having data along an entire route we could isolate different sections and determine what factors were affecting the walking speed at that point. Furthermore, in order for tracks to be useful to us, we needed to know the location of each point (to crosscheck against terrain information), and information about the walking speed (or which could be used to calculate the speed) at that point, therefore tracks which provided segment information, but not the location of where or when these points occurred (for user privacy reasons) could not be utilised.

While using an existing dataset allowed us to explore a wider range of terrains and individuals than would be possible if we were collecting the data manually, it had downsides as well. By not having control over the data, we could not be sure that all of the tracks available in a dataset were from valid walks or hikes, so a check to ensure that non-walking data was not present had to be included. A further complication of using crowdsourced data was that we didn't know when the participants took breaks during their walk. As we were attempting to calculate walking speed, not total route time, our walking speed model should only be based on the active components of a route. Performing analysis on the data without first removing

breaks would have likely resulted in inaccurate estimates for the speed. We therefore needed to find a method to detect and filter any breaks which were found in the data. The methods which were developed to read and filter the data were dependent on features of the data itself and are described in Section 3.1.

Due to these data processing requirements, we chose to initially explore a simpler model while also building the data extraction methods. For this reason we focused only on the hill and walking slope components at first, leaving the terrain factors to be added in once a proof of concept was established. In order to calculate the two slope components, another data source was required which could be crosschecked to provide the slope at each location. This data needed to be of high enough resolution that details in the terrain slopes could be distinguished. For low resolution data, hill and walking slopes associated with each data point would not be precise enough to produce an accurate model.

We aimed to conduct this work over two stages, first using data from Scotland to build a model (Chapter 3), before expanding the area of interest in order to validate the model on new data. All types of walk are available in Scotland, from flat coastal walks to steep mountain ridges. Starting in Scotland therefore allowed us to build and refine our model on a relatively smaller amount of data, while still encountering the full range of slopes.

The model itself had a small number of prior assumptions based on the work discussed in Section 1.1. Firstly that the speed predictions when walking directly up or down a hill on neutral terrain, would be close to those of existing functions seen in Figure 1.1. These models have been widely used and accepted as being accurate, on average, so there are not likely to be any large deviations from them. The second assumption which we had was that any new model for walking speed should reach its 'critical gradient' at slopes of around 15 – 21 degrees. Thirdly, based on intuition, walking speeds would be slower when on a steeper hill slope compared to flat ground.

While we were aware of the constraints, there were a large number of models which met these criteria. For example, in the hill slope direction, models which predicted a linear decrease in walking speed and those which suggested an exponential decay in walking speeds were equally possible based on our initial assumptions. To reduce the quantity of models to be considered, we allowed the data to determine which were most appropriate. Plots of the walking speed against each variable showed us the relationship and helped to determine which models should be explored. Cross-fold validation was then used to determine the model formulation which best fit the data. In order to test the performance of our model, we compared it to the two most widely used existing functions (Naismith's rule and Tobler's function). Two different metrics were looked at: how well each model predicted the walking speed over individual sections of the route, and how well each model predicted the total time taken to complete a route. For both of these metrics, a number of comparison methods

were explored, including the average error and the root-mean-square error (RMSE). These provided a measure of the different performance of each of the models, which we could use to determine which models were most successful at predicting walking speeds in different scenarios.

Once our two-variable model had been successfully applied to the data for Scotland, we validated it against data from the rest of the UK (Chapter 4). When doing this, differences in walking speed predictions were found between the models for each region. We were able to bring in the terrain variable, and explore whether this accounted for the differences seen between the two regions. When looking at terrains, we started off on a broad scale, looking at the difference between on- and off-road walking speeds as these are the factors used in some of the existing methods. We then extended this further to explore the types of roads (paved or unpaved) which were being walked on.

As we were not in control of the data collection, we had to develop a method of determining which of our data tracks were providing us with on-road data, and which represented off-road data. In order to do this, we required a dataset which provided the true locations of roads (and the road type) to compare against our location data. Similarly, we also required data which could be used to calculate the obstruction level for the off-road sections of our routes. Due to data constraints (described in Section 4.3.2), we were not able to explore the terrain obstruction effects in Scotland. We could, however, for other areas of the UK.

Using these additional datasets, we were able to build a model for the walking speed which takes into account all three of the factors which have been suggested to affect the speed. This model was tested in the same manner as the initial 2-variable one, namely by comparing its performance in predicting walking speeds against the existing methods.

As previously mentioned, the downside of using crowdsourced data is that we had no control over the range of walking types or terrains which were encountered, and we did not have access to any ground truth data regarding whether points we had identified as break points were actually breaks. We therefore included a fieldwork component to this work (Chapter 5) which had a number of goals. Firstly, it allowed to us validate our break-finding and data filtering methods to provide confidence that our walking speed model was built using valid walking data. It also allowed us to test our model under controlled circumstances, and specifically measure its performance in regions where we only had small volumes of data.

2.2 Materials Used

A variety of different data sources were used throughout this work. Figure 2.1 indicates how each dataset contributed to the creation of the walking speed model, and the details of each dataset are provided below.

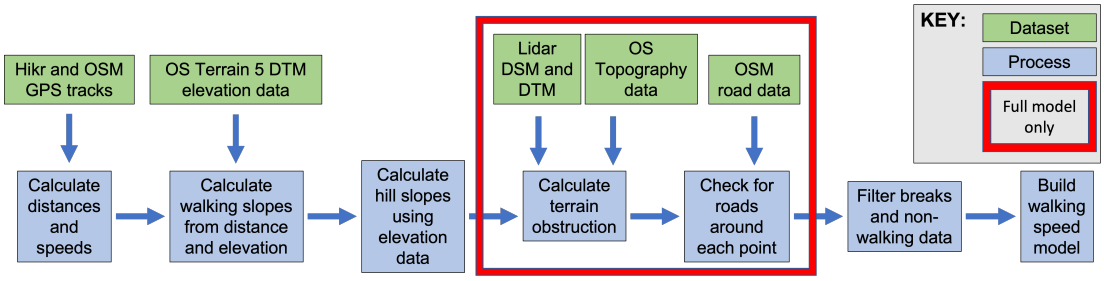


Figure 2.1: Data processing schema indicating how different datasets were combined and processed to build a walking speed model. Items in the red box were not used for the initial stage of this work (Chapter 3), only for the subsequent extension (Chapter 4).

2.2.1 GPX Data

The Global Positioning System (GPS) is a method for determining the location of a receiver using a network of satellites. GPS satellites transmit a signal which can be processed by a receiver to determine the distance between the receiver and the satellite. By combining the measurements from multiple (usually at least 4) satellites, the exact location of the receiver can be found (El-Rabbany, 2002).

The majority of the data used throughout this work came from OpenStreetMap (**OSM**) (OpenStreetMap contributors, 2021a), while some came from Hikr.org (Hikr.org, 2021b), both of which provide the data in GPX format. This is the most common format for GPS data. In general, GPX files consist of routes, tracks and waypoints. Both routes and tracks are ordered lists of waypoints; routes are used when planning a trip, whereas tracks provide a record of where the user has actually been, so are what this work is interested in. A **track** is a record of a trip, and is made up of **track segments**, where each track segment is a continuous run of points. The exact nature of tracks and track segments depends on the individual device settings being used. Some devices will record points at fixed time intervals, others may only record a new point after travelling a sufficiently large distance away from the previous point. In general, a new track or a new track segment is created whenever the device is switched off or loses signal for a period, although this is not always the case.

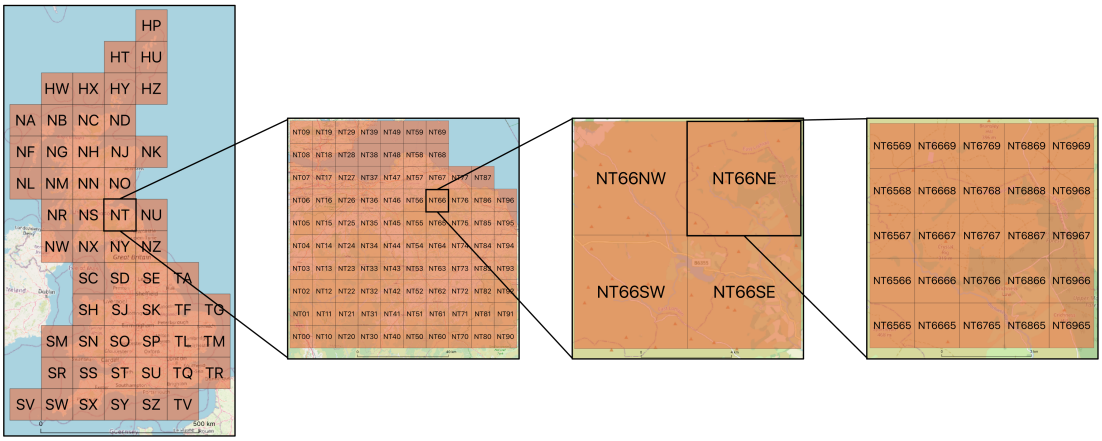


Figure 2.2: Maps showing different subdivisions of the National Grid. The grid squares shown have side lengths of 100 km, 10 km, 5 km and 1 km respectively. Grid data from Ordnance Survey (GB) (2020a), background map from OpenStreetMap, visualised using QGIS (see 2.2.5).

The downloadable OSM data consisted of a GPS data dump of all tracks on OpenStreetMap which contain points within the UK, as of April 2013. It is split into three different groups; Identifiable, Trackable and Public, depending on the user-specified privacy setting when the file was created. The Public tracks do not contain any timestamps so were ignored, as movement speeds could not be calculated. This dataset did not contain any information about the mode of transport being used when the track was recorded, or the device used for the recording.

The **Hikr** data was downloaded using a custom web scraper. This looked at each of the Hikr reports for a given region (Scotland, or the UK as a whole), checked if the associated GPX file contained timestamps for the route, and downloaded it if so. Although only relatively small quantities of Hikr data are available, all of the tracks used were explicitly tagged as hiking reports, so could be used as the basis of a filter to determine which of the OSM tracks contained walking data (this method is discussed in Section 3.1.3). The Hikr dataset contained all of the Hikr reports which were uploaded prior to July 2021 (the earliest tracks were recorded in July 2009).

2.2.2 National Grid

The slope and terrain data used in this work was either presented in, or converted to, the Ordnance Survey (**OS**) National Grid reference system. This system divides Great Britain into 100 km x 100 km squares. These 100 km squares have 2 letter designations and are divided into smaller squares by using numerical grid references (Ordnance Survey (GB), 2020a). The different subdivisions of the National Grid are shown in Figure 2.2.

2.2.3 Elevation Data

The Ordnance Survey Terrain 5 Digital Terrain Map (DTM), accessed through Digimap (Ordnance Survey (GB), 2020c) provides the elevation over the whole of Great Britain at 5 m intervals, with an accuracy of greater than 2.5 m RMSE. It is presented in ASCII format, with each file representing a 5 km x 5 km National Grid Square (1,000,000 datapoints). An example of the data is shown in Figure 2.3, where lighter regions indicate higher elevation values.

2.2.4 Terrain Data

Road and Path Data

Road and path data used was from OpenStreetMap, downloaded from GeoFabrik.de (OpenStreetMap contributors, 2021b) and contains the data as of August 2021. Although similar data is available from Ordnance Survey, the OSM data was preferred due to it providing a more detailed classification than Ordnance Survey, in terms of the road or path type. A full list of the different classifications of road and path types is given in Section 4.3.1. Each region was downloaded individually (England, Scotland and Wales) and within each download, only the `gis_osm_roads_free_1` shapefile was retained (other files detailed features such as watercourses or buildings which were not required for this work). Each of the shapefiles was broken up into 10 km x 10 km tiles, matching National Grid squares. For locations which occurred in multiple data sets (e.g. on the border of England and Wales), the data were combined. This led to duplicates of some data in these regions, but this was not an issue as we were only interested in whether a road or path was present, not how many there were. The NT26 tile is shown in Figure 2.4. This tile contains a wide range of the different road and path types contained within the dataset, from main and local roads in the urban areas to footpaths and tracks through the hills.

Terrain Type

The Ordnance Survey MasterMap Topography dataset was obtained via Digimap (Ordnance Survey (GB), 2020b). The data was provided in the form of 100 km x 100 km National Grid tiles, each as an individual file geodatabase. Within each file, only the 'Land' layer, and within this the 'topographicarea', was used. This provides descriptions of the terrain type for the whole of Great Britain. Each feature can have multiple terrain types associated to it, with the full list of possible terrain types given in Section 4.3.2. Figure 2.5 shows a small area of the NT tile where single terrain feature has been highlighted, and the corresponding terrain description shown (Boulders, Heath, Rock, Rough Grassland). Terrain features which crossed National Grid tile borders led to duplicate data, but, similar to the road data above, this was not an issue as we only needed to know if a type of terrain was present.

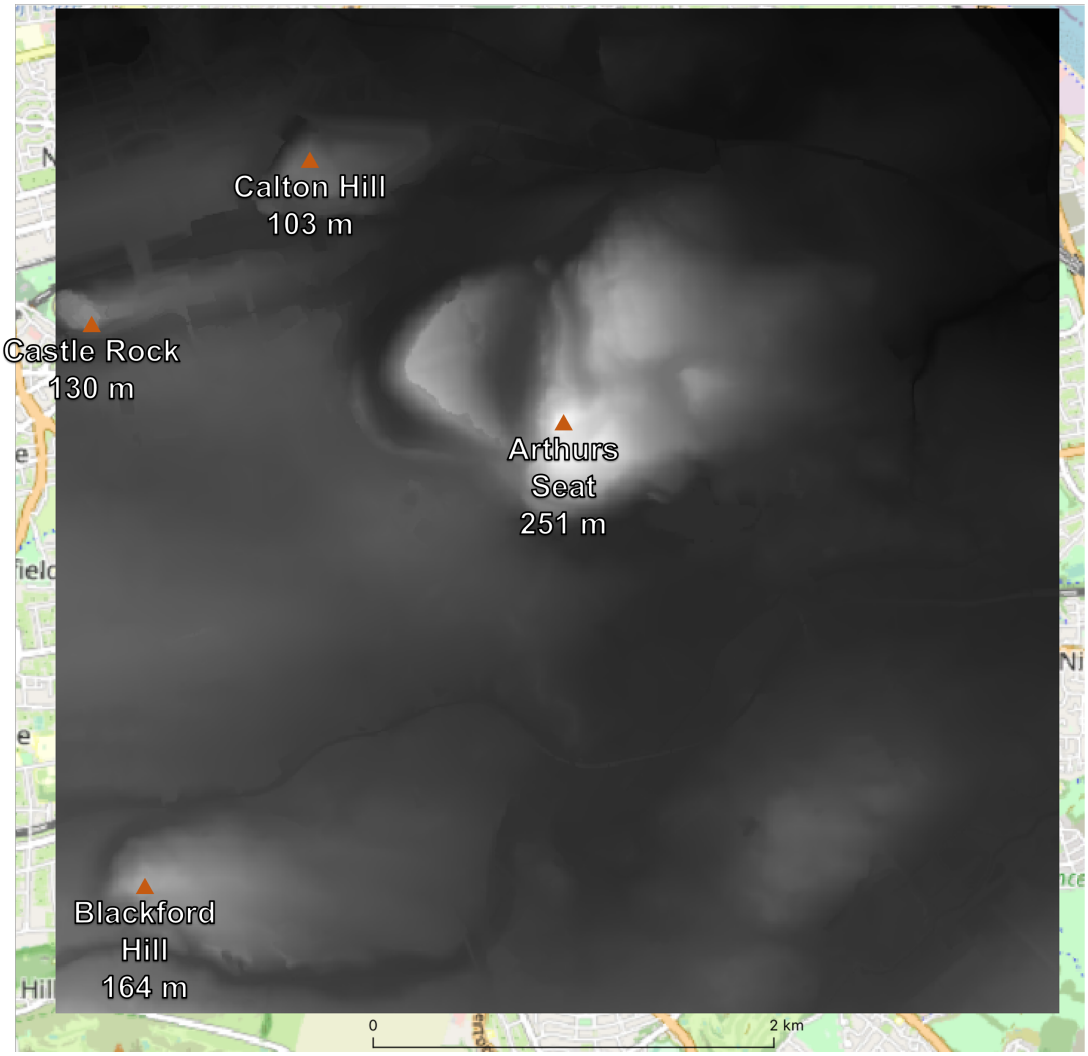


Figure 2.3: Example of OS Terrain 5 DTM data. The NT27SE tile is shown, shaded to indicate elevation from black (0m) through white (250m). The specific locations and heights of four hills are also indicated. Terrain data from Ordnance Survey (GB) (2020c), background map from OpenStreetMap, visualised using QGIS (see 2.2.5).

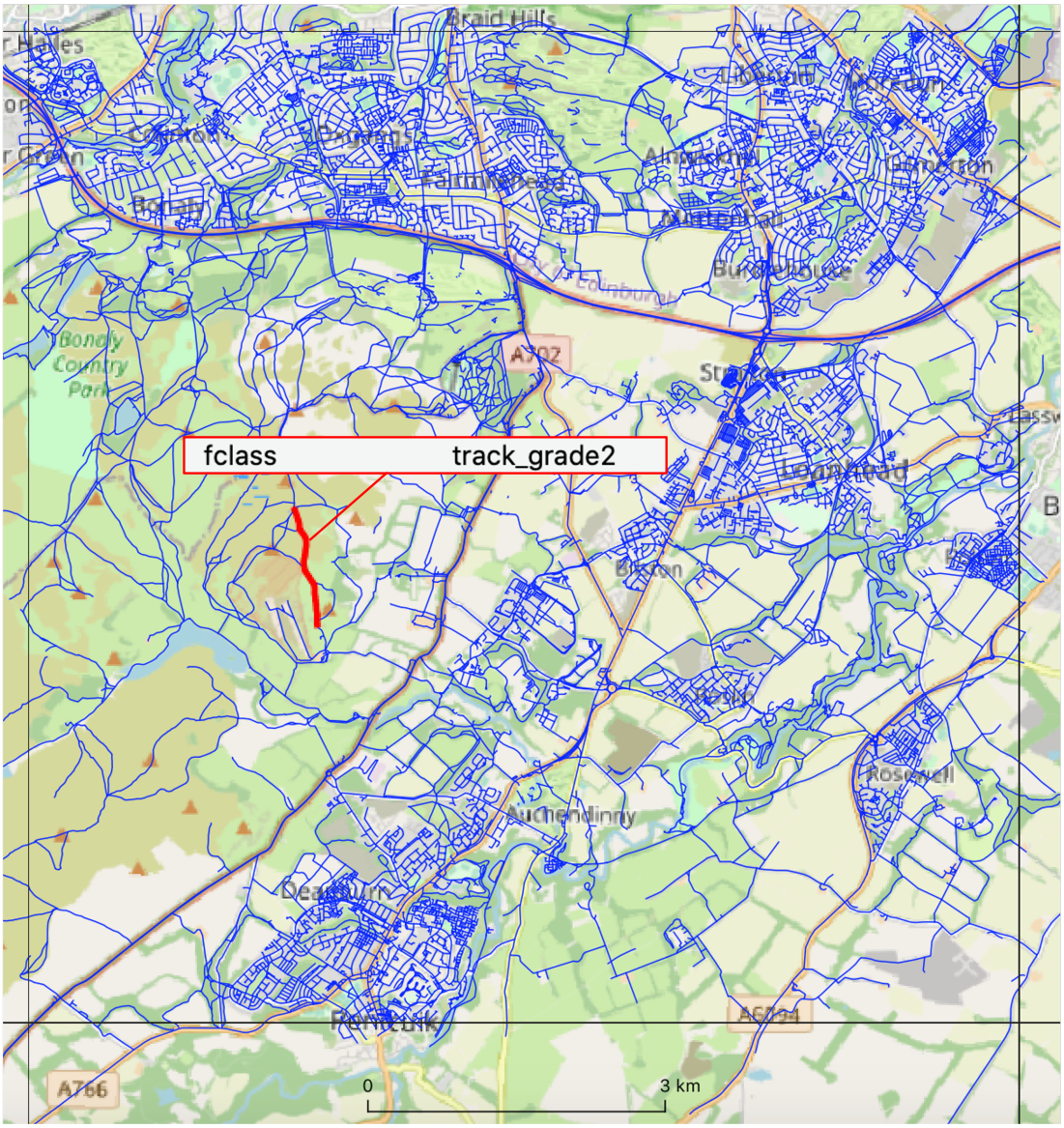


Figure 2.4: Example of OSM Road data. The NT26 tile road data is shown (blue), overlaid onto a map of the area. A single feature is highlighted (red) with the road type shown. Road and path data from OpenStreetMap contributors (2021b), background map from OpenStreetMap, visualised using QGIS (see 2.2.5).

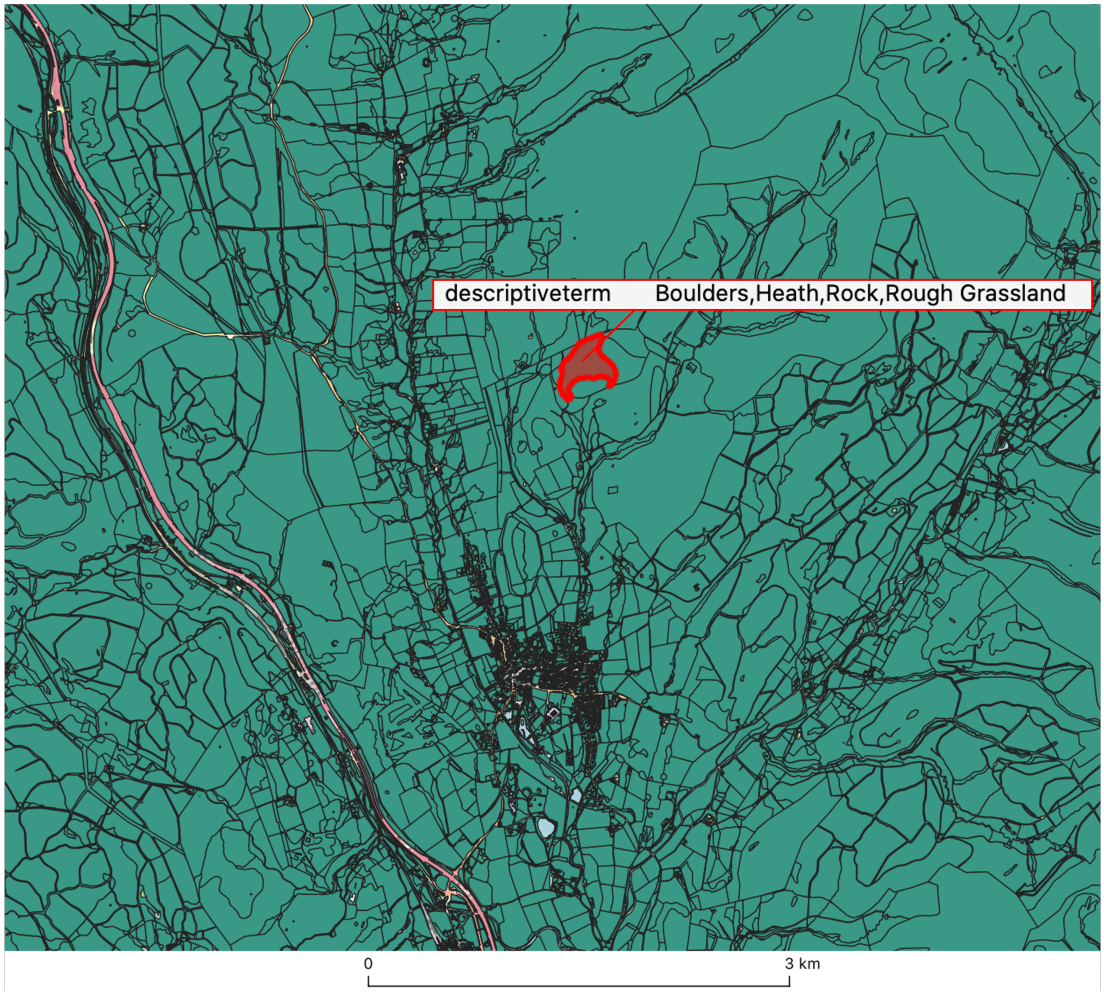
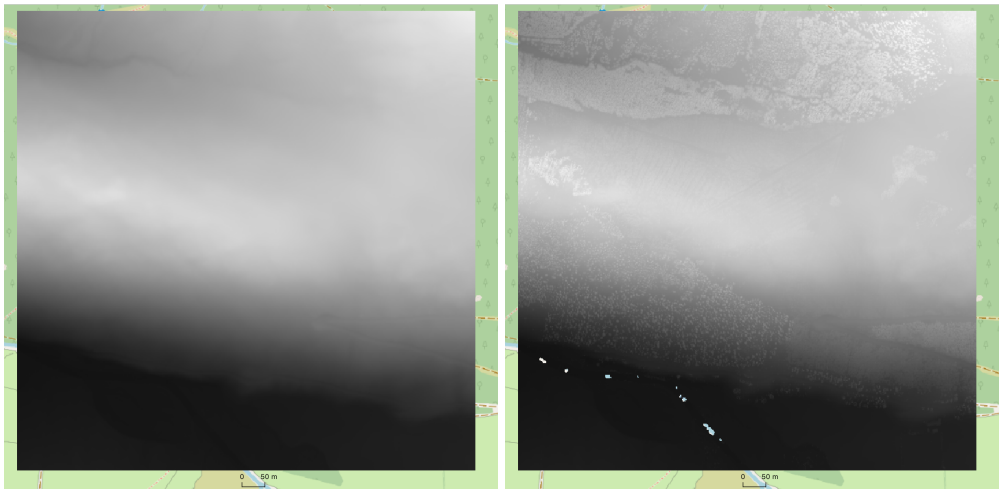


Figure 2.5: Example of OS MasterMap Topography data. A region of the NT tile is shown, with a single feature highlighted (red), and the terrain description for that feature shown. Topography data from Ordnance Survey (GB) (2020b), background map from OpenStreetMap, visualised using QGIS (see 2.2.5).



(a) DTM **(b)** DSM

Figure 2.6: Example of lidar DTM and DSM data. Both images show the same tile (NY1314) shaded to indicate elevations from 100 m (black) to 300 m (white). The missing regions in the DSM tile are areas where the lidar data is not available. Lidar data from Environment Agency (2017, 2020), background map from OpenStreetMap, visualised using QGIS (see 2.2.5).

Lidar Data

Due to some problems with the OS Topography dataset (discussed in Section 4.3.2), lidar data was also used when looking at terrain obstruction. Lidar data is available for large areas of the UK, at various resolution levels. However, for Scotland, there is limited coverage in more rural (i.e. off-road) areas, and most of the available data is at very high resolution (25-50 cm). This is generally greater than the accuracy of the GPS devices used, so it was decided that including lidar data for Scotland was not worth the increased computational storage and processing time. For England and Wales, a lidar Digital Terrain Map (DTM) and Digital Surface Map (DSM) were downloaded, both at 2 m resolution, with an accuracy of 15 cm RMSE (Environment Agency, 2017, 2020; Natural Resources Wales, 2016). The DTM gives the ground height above sea level every 2 m, while the DSM provides the surface height (i.e. taking into account buildings or trees etc). Coverage of the two datasets is not complete, with some areas in England and Wales unavailable, but all available data was requested. This was provided by Digimap in ASCII format, with each file representing a 1 km x 1 km National Grid square. The data was provided separately for England and Wales, but was manually put into a single folder structure for ease of processing. In cases where both datasets contained data from a given grid square (regions on the England-Wales border where data was captured twice), the file from the England dataset was preserved. Figure 2.6 shows examples of both the DTM and DSM data. Individual trees can be seen in the DSM data (most clearly in the top right corner of Figure 2.6b), and their heights calculated as the difference in elevation between the two datasets.

2.2.5 Computational Materials

Computers

Preprocessing of the GPX files made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF) (U of Edinburgh, 2022).

Data modelling and analysis was performed on a Macbook Air 2018, 1.6 GHz Dual-Core Intel Core i5 processor, 8 GB 2133 MHz LPDDR3 RAM.

QGIS

GPS and terrain data was visualised using QGIS version 3.14.0-Pi (Open Source Geospatial Foundation Project, 2020).

OpenStreetMap

Images throughout make use of map data from OpenStreetMap (OpenStreetMap contributors, 2022).

R

Data modelling and analysis was completed using R-Studio Version 1.2.5019 (R Core Team, 2019).

The following R packages were used throughout this work:

- `caret` (Kuhn, 2020)
- `Hmisc` (Harrell Jr, 2021)
- `reshape2` (Wickham, 2007)
- `dplyr` (Wickham, François, Henry, & Müller, 2021)
- `mgcv` (S. N. Wood, 2017)
- `sandwich` (Zeileis, 2006)
- `lmtest` (Zeileis & Hothorn, 2002)
- `ggplot2` (Wickham, 2016)
- `rgl` (Murdoch & Adler, 2021)
- `visreg` (Breheny & Burchett, 2017)
- `DescTools`(Andri et mult. al., 2022)
- `plotrix`(J, 2006)
- `cowplot` (Wilke, 2020)
- `stringr` (Wickham, 2019)

2.2.6 Github Repository

The code written for this project is stored in the Github repository AndrewWood94/PhDThesis, and is licensed under the terms of the GNU General Public License v3.0. The structure of the files is as follows:

- Remapping: Contains the code used to convert the OSM data files into National Grid file structure.

- Preprocessing: Contains the code used to read in the GPX files, run the preprocessing methods, lookup the elevation and terrain data and prepare the data prior to analysis.
  - Scotland contains the snapshot of the code as it was when the data was processed as described in Section 3.1.
  - ROUK contains the up-to-date codebase, incorporating the changes and additional terrain information which were used for the processes discussed in Chapter 4.

    Both of these folders contain a python package which can import and process the data using a single script (`run_gpx_importer`). Links to the code for specific elements of the processing pipeline, and the sections of this document to which they relate, are detailed below. (Links lead to the ROUK file versions, but the file structure is the same in both cases)

    * Web Scraper - Downloads the GPX tracks from Hikr.org (Section 3.1).
    * File Reader - This is an adapted version of the GPX Segment Importer plugin for QGIS (Simon Gröchenig, 2019), which can import GPX files and calculate basic statistics about each point (duration, speed etc.). It has been updated to ensure the track falls within the desired boundaries and allows for terrain and slope data to be incorporated (Section 3.1.1).
    * Terrain Classifier - Used to read the slope and terrain information for each datapoint (Sections 3.1.1 and 4.3).
    * Break Finder - Used to identify breaks found within each GPS track (Section 3.1.2).
    * Filtering and Merging - Used to filter the data to remove breaks or non-walking track segments, and merge the data into 50 m sections for analysis (Sections 3.1.3, 4.1).

- Analysis: Contains the R files which were used for data modelling and analysis.
  - Scotland contains the code used to determine the initial model type (Sections 3.2 and 3.3), and to compare with existing datasets (Section 3.4).

- `ROUK` contains the code used to analyse the extended dataset, and extend the model further to explore additional variables (Sections 4.2, 4.4 - 4.7).
- `Fieldwork` contains the code written to analyse the model performance across different terrain environments encountered during the fieldwork (Section 5.3).

Chapter 3

Modelling walking slope and hill slope: Scotland as a case study

For initial processing of the data, we limited our dataset to only include tracks in Scotland while we explored the methods required to read and filter the data. Scotland has widely varying terrain, including steep regions in the Cairngorms and Highlands, as well as large flat areas, so is representative of the full range of possible slopes. For this first exploration of the data, we were also only investigating the impact of the walking slope and hill slope on walking speeds.

3.1 Data Preparation

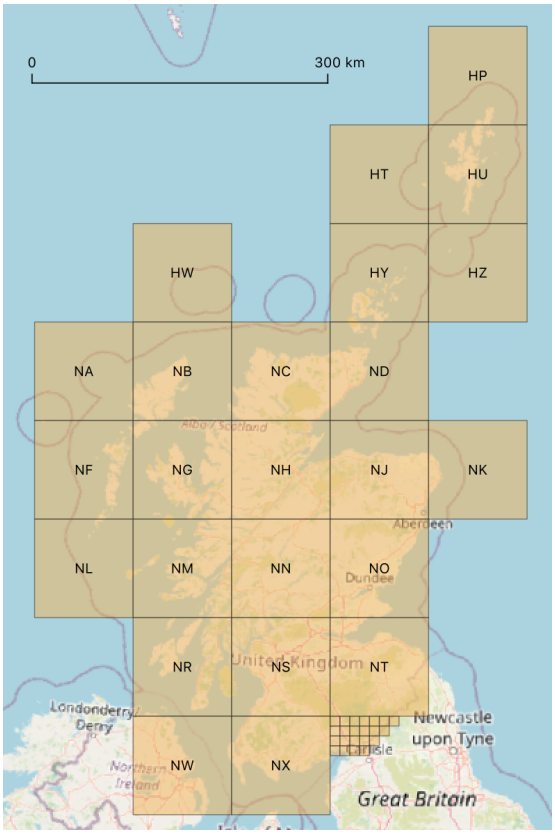
Both the Hikr and OSM GPS track repositories were able to be filtered such that only those tracks which were tagged as Scotland were included in the dataset (Hikr.org, 2021a; OpenStreetMap contributors, 2021c). For the OSM data, this consisted of all tracks which contained points in Scotland, but was not limited to those which were fully contained within Scotland. Therefore only tracks segments where every point was within the region covered by the OS DTM elevation data for the following Ordnance Survey tiles were considered as for this work:

HP HT HU HW HX HY HZ NA NB NC ND NF NG NH NJ NK NL NM NN NO NR NS NT NU NW NX NY

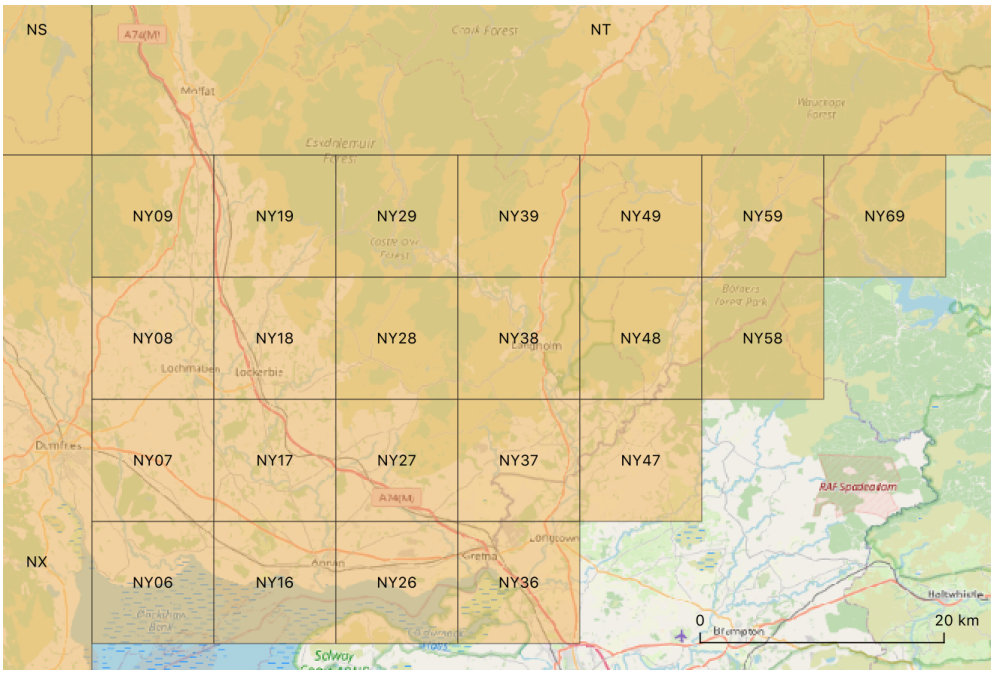
Within NY, only the following tiles were used:

09 19 29 39 49 59 69 08 18 28 38 48 58 07 17 27 37 47 06 16 26 36

A map of the region used can be seen in Figure 3.1. Note that NX tile covers part of northern England which does not border Scotland. There were no tracks included in the download which extended into this region, so a smaller breakdown of the NX tile was not used. Similarly the NT tile was not broken up as only 5% of the tiles do not overlap the Scottish border, so it was felt to be unnecessary to remove them (compared to 78% non-overlap for NY tile). The NW tile covers a large portion of Northern Ireland, but in practice the tile data received from Ordnance Survey is limited to only the areas in Scotland, as the Ordnance Survey DTM elevation data used does not cover Northern Ireland.



(a) Full view of National Grid tiles covering Scotland



(b) Closeup of the tiles used within the NY National Grid tile

Figure 3.1: Maps showing the National Grid tiles within which tracks were considered to take place in Scotland. Every point in the track had to be within the tiles shown. Grid data from Ordnance Survey (GB) (2020a), background map from OpenStreetMap, visualised using QGIS (see 2.2.5).

3.1.1 Importing GPS tracks

The GPX files were read using a heavily modified version of the `GPX Segment Importer` plugin for QGIS (Simon Gröchenig, 2019) (see 2.2.6). A GPX file can contain multiple separate tracks, potentially recorded across different days. However, when processing the data, we combined all data in a single file into one track, but kept the individual track segments separate. The reason for this is that tracks within the same file were highly likely to be undertaken by the same individual, and thus have correlated walking speeds. By grouping them together, we were able to account for this in later processing. It was important, however, that we distinguished between different track segments and processed them individually. When looking for points in the data where people took breaks along their route (Section 3.1.2), we used the median speed and distance of a track segment to help identify breakpoints. If instead we used the median speed and distance for the track as a whole, then this could cause problems in situations where a track contained multiple segments with different speed profiles.

For each track segment, the list of points was converted to a series of connected linestrings, with the following properties calculated and attached to each one:

- Start coordinate
- End coordinate
- Start time
- End time
- Duration
- Distance
- Speed

Any segments less than 250 m in length or 2.5 minutes in duration were ignored, as it was felt they could not represent a real walk.

Elevation and Slope

After reading in each track, the Ordnance Survey Terrain 5 DTM was used to calculate the following:

- Elevation
- Walking slope
- Hill slope

Although the GPX files contained elevation data which could be used to calculate the walking slope, there were a number of reasons to prefer the Ordnance Survey data:

- Multiple tracks were found where the elevation was measured to the nearest half metre or metre on the GPS device. Using this value would lead to rounding errors in slope calculations over short distances. The Ordnance Survey values for elevation have a higher level of precision (0.01 m) which could alleviate these errors.
- Using OS values avoids any concern about potential discrepancies arising from differences in calibration across GPS devices, so we could ensure that the same location on different tracks would have the same elevation.
- We were also investigating the impact of hill slope, which could not be calculated from the GPS track data, and using the same data source for both calculations should allow for better evaluation of interactions between the two variables.

When first reading each data track, the elevation DTM was sampled to provide the spot height of the start co-ordinate, and the hill slope at this point was calculated using the quadratic surface method (Dunn & Hickey, 1998; Zevenbergen & Thorne, 1987). As explored in previous work which compared a variety of hill slope algorithms (Jones, 1998), the quadratic surface method produces the most accurate slope estimates, especially given the high resolution of the DTM being used for calculations. The walking slope was calculated using the spot heights of the start and end coordinates of each linestring, and the distance between those points.

3.1.2 Break Finding

For this work, we were interested in calculating the active components of a route (as opposed to breaks or rests), as that is the data on which a walking speed model should be based. Performing analysis on the data without first removing breaks would likely result in inaccurate estimates for the movement speed. Similarly, when developing the filter to find walking routes in the OpenStreetMap dataset we wanted to ensure that we were only considering the active route components.

Initial inspection of the walking tracks showed that a large number contained obvious breaks, while the device was still recording. The simplest method to find breaks, and the first implemented, was to tag all points where there was no movement (i.e. walking speed = 0 km/h). The next step was to tag any individual points which represented over 1 km or 10 minutes of travel as breaks. These points could occur in areas where the device lost signal for a period. Although this should result in the creation of a new track segment, there were a number of tracks where this was clearly not the case.

Unfortunately, this did not capture all of the break points due to GPS drift; the distance between the measured position of the GPS device and the true location. GPS signal is subject to satellite visibility; tall buildings and mountains can affect this as the satellite signal reflects off the walls, leading to errors in the location (Merry & Bettinger, 2019). On top of this, tree cover, or similar things (e.g. overhanging rocks) can cause the signal to be weakened, reducing the location accuracy. Particularly enclosed spaces, such as steep-sided valleys can

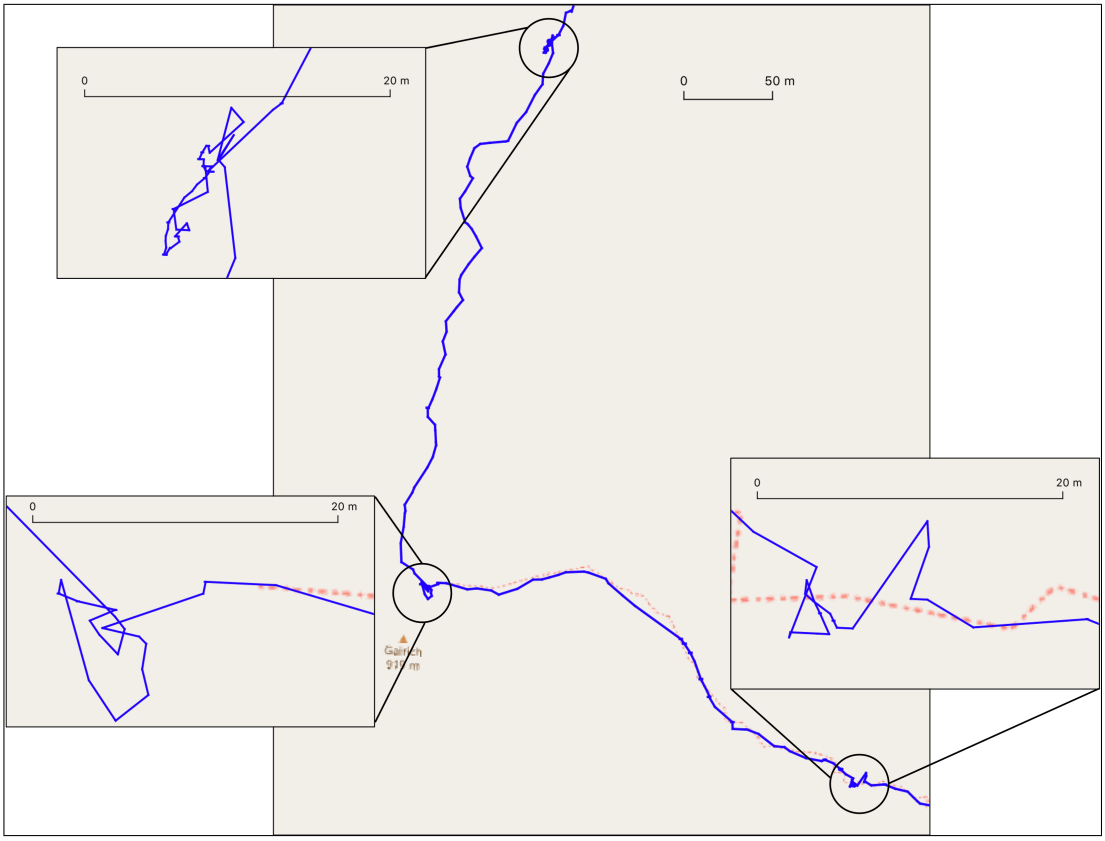


Figure 3.2: A GPS track where three breaks can identified by finding GPS drift clusters. Background map from OpenStreetMap, visualised using QGIS (see 2.2.5).

also suffer from reduced GPS accuracy, as there is less sky (and therefore fewer satellites) visible (D'eon, Serrouya, Smith, & Kochanny, 2002). Furthermore, the OSM GPS tracks were all from a data dump produced in 2013. We do not know what devices were used to record the OSM tracks, but it is likely that some were recorded on smartphones. Smartphones at the time had more basic GPS chips which were less accurate than dedicated GPS devices (Zandbergen, 2009), and their GPS accuracy depended on the particular device and app being used (Bauer, 2013; Hess, Farahani, Tschirschnitz, & von Reischach, 2012), with an average error of up to 9 m being identified in some circumstances. This error can be magnified when stationary, as a large number of points are recorded in the same general area, forming clusters. Examples of these clusters can be seen in Figure 3.2.

Although these locations are easy to identify when the route is visualised, we needed an automatic filter in order to remove them from our analysis. The movement speeds calculated from these drift points can vary greatly, depending on the sampling rate of the device and the amount of GPS drift. A drift measurement of 6 m in a very short time (1-2 seconds) would mean a very high speed (10-20 km/h) is found, but a series of very low speeds in a row could be due to a break with a small amount of drift, or it could indicate a particularly difficult (and therefore slow) section of the route.

A number of previous studies explored methods to classify GPS tracks into activity types (Alvares et al., 2007; Biljecki, 2010; De Vries & Sterkenburg, 2012; Palma, Bogorny, Kuijpers, & Alvares, 2008; Schuessler & Axhausen, 2009; Tsui & Shalaby, 2006; Wan & Lin, 2016; Zhou, Jia, Juan, Fu, & Xiao, 2017), however none could be directly applied to this problem. This is because they were usually trying to identify different modes of transport, which can be clearly distinguished by different travel speeds. However, as mentioned above, the speed measurements caused by GPS drift can easily be in the same range as an expected walking speed. Methods which look for clusters similar to those being investigated here were also not useable as they generally require a known sampling rate for the GPS device (De Vries & Sterkenburg, 2012), or involve checking clusters against a pre-existing database of features (Alvares et al., 2007; Palma et al., 2008). Our data consisted of tracks created using a wide range of devices and settings, and over a very large area, so it was not possible to either assume a fixed sampling rate, or to pre-select features where breaks were likely. Instead, various ideas from a number of works were combined and adapted to find clusters of points, which were then checked to see if they should be identified as a break.

Firstly, based on Palma et al. (2008), a modified version of DBSCAN (Ester, Kriegel, Sander, & Xu, 1996) was used to identify clusters. Unlike in the standard algorithm, each point only looked for neighbours whose timestamp was within 10 minutes of the point being checked. This prevented clusters being found on routes which doubled back on themselves or contained loops. Secondly, as sampling rates were not consistent across devices and the tracks covered a variety of terrains, we could not assume a fixed radius to find neighbouring points. Instead the median travel distance for the particular track segment being investigated was used (r_{median}).

> *DEFINITION 1. Neighbourhood of a point: Let* $\{p_0, p_1, ..., p_n\}$ *be points on a GPS track segment, with timestamps* $\{t_0, t_1, ..., t_n\}$*, a median distance* r_{median} *between consecutive points and a median point speed* s_{median}*. The neighbourhood* N_k *of a point,* p_k*, is the set of points* p_i *such that:*
>
> $$dist(p_i, p_k) < r_{median} \text{ and } |t_i - t_k| < 600s$$

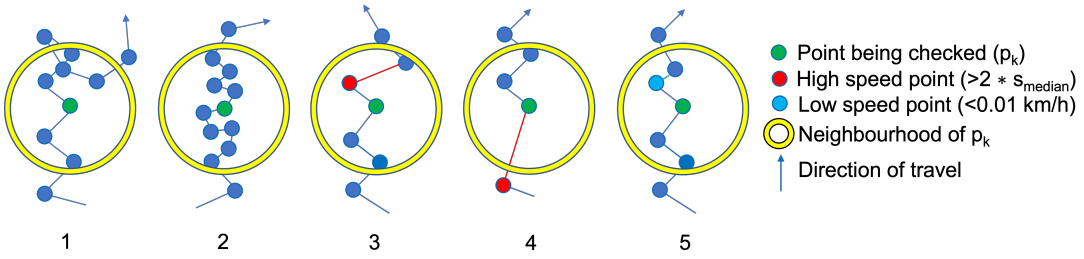


Figure 3.3: The five scenarios in which a point cluster is identified around a point.

Using these conditions, all points along each track segment were tested to find point clusters.

DEFINITION 2. Point cluster: A point cluster, C, is formed from a point neighbourhood, N_k, if any of the following hold:

1. *At least 5 non-consecutive points are found in N_k*
2. *At least 10 consecutive points are found in N_k*
3. *A point within N_k has a 'high speed'; a speed greater than $2 * s_{median}$*
4. *A point p_k is immediately preceded by a 'high speed' point*
5. *A point within N_k has a 'very low speed'; a speed less than 0.01 km/h*

Examples of each of these conditions are shown in Figure 3.3. The first two conditions worked together to prevent finding clusters in unusually slow sections of a route, such as a steep hill climb, but allowed for areas where GPS drift was small and an entire cluster was contained within N_k. The third and fourth conditions found 'high speed' points which occurred when a large amount of GPS drift was measured, and could spread a cluster out over a wider area. Without accounting for these points separately, the algorithm would often end up registering a single break as multiple short breaks separated by high speed movements. The fifth condition was necessary for situations where the device was set to only record new points once a minimum distance has been travelled from the previous location. Note that in the third and fifth cases above, the point immediately following the high- or low-speed point was added to the cluster, even if it was not included in N_k. Similarly, in the fourth case, the cluster included the preceding high-speed point regardless of whether it was in N_k.

Once a point cluster was identified, each point within it was subsequently tested and any further clusters found were added to the original. This continued recursively until every point within the cluster had been checked and no new clusters were found. This allowed us to identify clusters (and breaks) which lasted longer than the 10 minute Neighbourhood threshold, as each point added to a cluster subsequently identified later points.

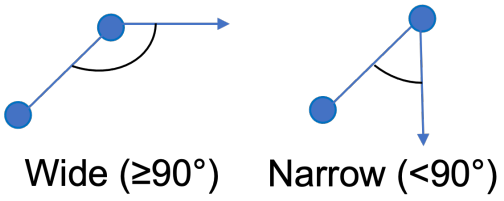


Figure 3.4: Examples of wide and narrow point angles.

		Point Speed		
		Low	Medium	High
Point Angle	Narrow	High	Medium	High
	Wide	Medium	Low	Medium

Table 3.1: Break likelihood classifications based on point speeds and angles.

Following this, further steps were taken to try and ensure that only legitimate breaks were classified, rather than slow sections of movement. Firstly, each point within the cluster was tested for 'break likelihood' using a simple classification methodology based on the approach used by Wan and Lin (2016), which uses two variables; the point speed, s_i, and the point angle, α_i.

DEFINITION 3. Point speed: The speed, s_i, at a point, p_i is categorised as follows:
- *Low:* $s_i < s_{median}/2$
- *Medium:* $s_{median}/2 < s_i < 10m/s$
- *High:* $s_i > 10m/s$

DEFINITION 4. Point angle: The point angle, α_i, for a point, p_i, is the angle created at p_i by the lines connecting it to p_{i-1} and p_{i+1}. It is then categorised as follows:
- *Narrow:* $\alpha_i < 90^\circ$
- *Wide:* $\alpha_i \geq 90^\circ$

Example of narrow and wide point angles are given in Figure 3.4. As discussed in Wan and Lin (2016), normal movement is unlikely to result in point angles below 90 degrees, unlike the random motion caused by the GPS drift which often produces narrower point angles.

The break likelihood for each point in a cluster was found using the classifications in Table 3.1. Once the break likelihoods were identified, the cluster was checked to find a potential break. Performing this step helped to limit the sizes of breaks, as walking points immediately preceding or following a break were often caught in the cluster.

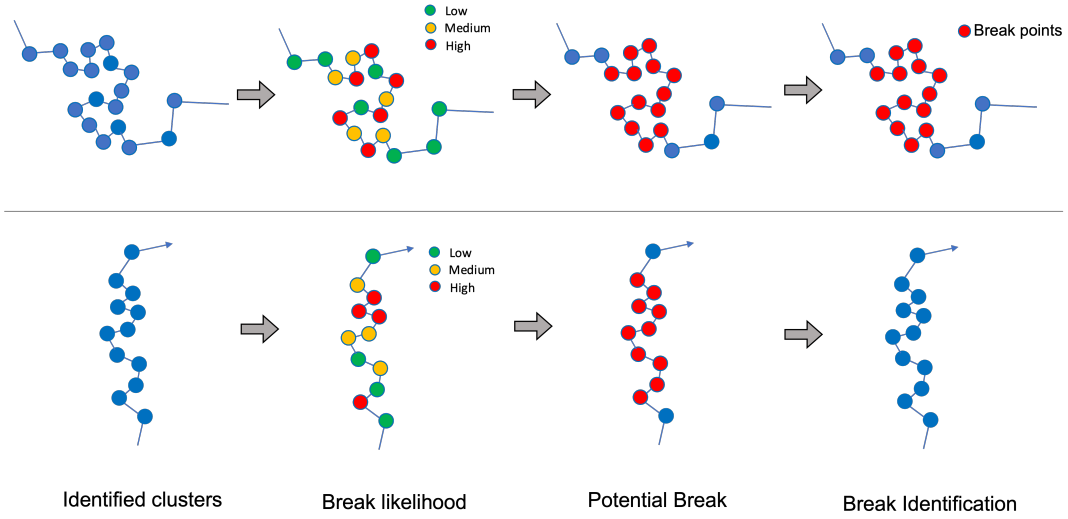


Figure 3.5: Demonstration of how breaks are identified from point clusters. Points in a cluster are checked to find their break likelihood, and see if a potential break can be identified. Potential breaks are then checked to see if a break can be formed.

> *DEFINITION 5. Potential Break: A potential break, B^*, is created by ordering the points in a cluster and identifying the first and last points with a break likelihood of medium or high. All points in the segment between these points (including the points themselves) form the potential break.*

As a break by definition implies no movement, any motion in a given direction should be cancelled out by motion in the opposite direction. For this reason, the bearing of all points within the potential break was found and assigned a quadrant, and a break was only formed if there was travel in opposite quadrants.

$$
\begin{aligned}
Q_i = True \; iff \\
&\exists \; p \in B^* \mid (90 * (i-1))^\circ \leq p_{bearing} < (90 * i)^\circ, \\
&i = (1,2,3,4)
\end{aligned}
$$

> *DEFINITION 6. Break: A break, B, is created from a potential break, B^*, if both of the following hold:*
> - $Q_1 = True \; and \; Q_3 = True, \; or \; Q_3 = True \; and \; Q_4 = True.$
> - *Less than half the points in B^* have a low break likelihood.*

Figure 3.5 provides examples of two point clusters and the process to identify breaks. Note that not every point in each cluster becomes part of the potential break; low break likelihood points at the ends of a cluster are excluded. After processing the second cluster, a break is not formed because the cluster does meet the requirement of travel in opposite quadrant directions. While there is a substantial amount of east-to-west deviation in the track, it consistently heads north. Regions such as this identified in GPS tracks are more likely to indicate a period of challenging (and therefore slow) movement.

Pseudocode outlining the breakfinding method is shown in Algorithm 1. A ground truth dataset of routes where the breaks are tagged does not exist, so we were unable to measure the accuracy of this break-finding algorithm (although this is explored in Section 5.2.1). However, a visual inspection of the tracks suggested that it classified breaks well, with over-classification in some circumstances. This was preferable to under-classification for two reasons; firstly, as discussed in Section 3.1.3, only the longer breaks were actually removed from calculations so small breaks found as a result of over-classification have no impact. Secondly, we were performing analysis on very large datasets, so we expect the analysis to be robust to the incorrect removal of a small number of data points.

Algorithm 1 Breakfinding process for Scotland data

Breakpoint_list $= \emptyset$
Find the median distance (r_{median}) and speed (s_{median}) of the segment
for point (p_i) in segment **do**
    Calculate travel direction quadrant and point angle
    Calculate break likelihood using the point speed and angle ▷ see Table 3.1
    if speed == 0 or distance > 1 km or duration > 10 minutes **then**
        Breakpoint_list += p_i
    end if
end for
for point (p) in segment **do**
    if Neighbourhood of p is a cluster (C) **then** ▷ See Definitions 1 & 2
        for point (p_c) in C **do**
            if Neighbourhood of p_c is a new cluster (C_n) **then**
                $C = C \cap C_n$
            end if
        end for
        Remove points at the ends of the cluster with low break likelihood
        Add 'missing' point ids to the cluster (to make a continuous run of points) to form a Potential Break (B*)
        if less than half the points in B* have low break likelihood and there is travel in opposite quadrants (Q1 & 3 or Q2 & 4) **then**
            Breakpoint_list += B*
        end if
    end if
end for

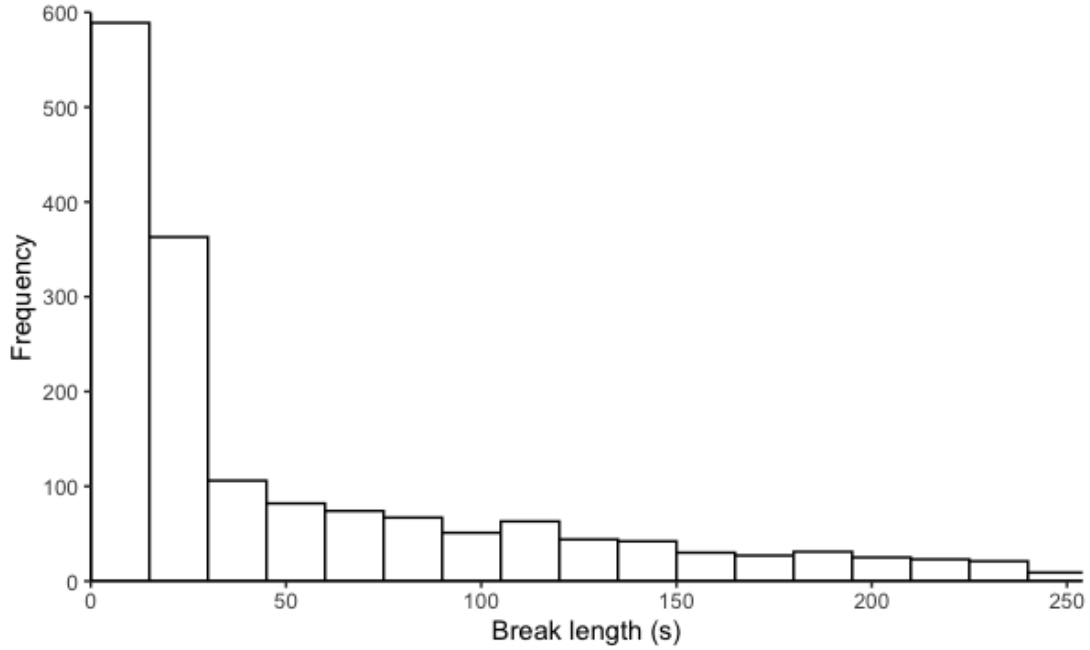


Figure 3.6: Histogram of break lengths found in the Scottish Hikr dataset, with bin widths of 15 seconds. (Note: for readability this graph only shows the subset of breaks whose length is under 250 s.)

3.1.3 Data Filtering

Before processing our data, we wanted to use information about walking speeds found in the Hikr data to identify and remove non-walking tracks from the OSM dataset. When doing this, we only wanted to compare the active component of each of the tracks. However, not all break points should necessarily be considered 'inactive'. Figure 3.6 shows the frequencies of break lengths found in the Scottish Hikr data, and it is clear that the majority of the breaks were under 30 seconds in length; we defined these sub-30 second breaks as '**micro-breaks**'. Micro-breaks were felt to be a constituent part of the walk which most people would have to do, such as pausing to catch your breath or check a map, and were not excluded from further analysis. The exception to this was if the micro-break contained a single point with over 1 km of movement, or points with speeds of over 10 km/h. Speeds over 10 km/h in a region which had been identified as a break were very likely to be errors caused by GPS drift, and as such these points would not be useful in representing a typical walking speed. It was felt that 30 seconds is a reasonable length such that breaks of this length or longer would be a conscious choice by the walker to stop, rather than being necessary for the route. After excluding long breaks, it was also decided to remove any breaks which occurred at the start or end of a track segment, as these were considered likely to be legitimate breaks, regardless of their duration.

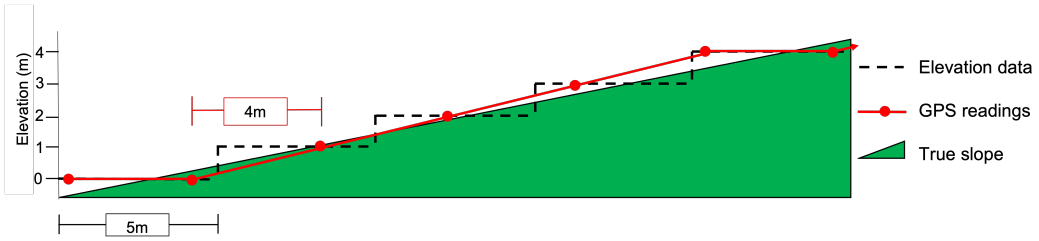


Figure 3.7: Impact of elevation data resolution on slope measurements. A simulated route section on a slope measured with 5 m resolution elevation tiles and 4 m between datapoints. Measured datapoints record either no walking slope, or a value steeper than the true walking slope.

The remaining data were merged together into continuous sections at least 50 metres in length, to remove some of the variability caused by the GPS devices and elevation data resolution. Devices generally measure the time to the nearest second, so rounding errors over short distances could have a large impact on the estimated speeds, especially in combination with small inaccuracies in the location data. For example, a travel distance of 4 m recorded in a time rounded to 3 seconds could indicate a speed of anywhere from 4.11 km/h to 5.76 km/h. By merging the data together into longer sections the impact of these errors was greatly reduced, as any error made up a smaller percentage of the merged section. A similar inaccuracy existed with walking slope values. As our elevation DTM had a resolution of 5 m, a distance of under 5 m between two consecutive points could result in a walking slope of 0 degrees, regardless of how steep the terrain is in reality. This is demonstrated in Figure 3.7, which shows a simulated section of a route, with the device recording a new point every 4 m. In this example, the true slope is approximately 11.3 degrees, however without merging the data, we would record 4 points on a 14 degree slope and 2 points on a 0 degree slope. Without merging the datapoints, we would have inaccurate data on which to base a walking speed model. By merging the points together, we smoothed out the steps in the data and produced a single datapoint with a slope value closer to the true value, which will help produce a model with more accurate walking speed predictions.

When merging data, the distance was taken as the cumulative distance across all points, and not the direct start point to end point distance, to avoid cutting corners and artificially decreasing the movement speeds. After merging the data, a small number of points less than 50 m long remained immediately preceding a break or the end of the segment. These were ignored as they were under the 50 m threshold used. When merging the points together into 50m intervals, both the hill slope and walking slope were calculated as the weighted average of the slope at each datapoint, weighted by the duration of each point. Note that the hill slope value used was that at the start coordinate of each datapoint. The importance of weighting by time and not distance is demonstrated in Figure 3.8, which gives an example of a steep (30 degree) hill being descended for a short distance, before travelling on flat terrain, with an

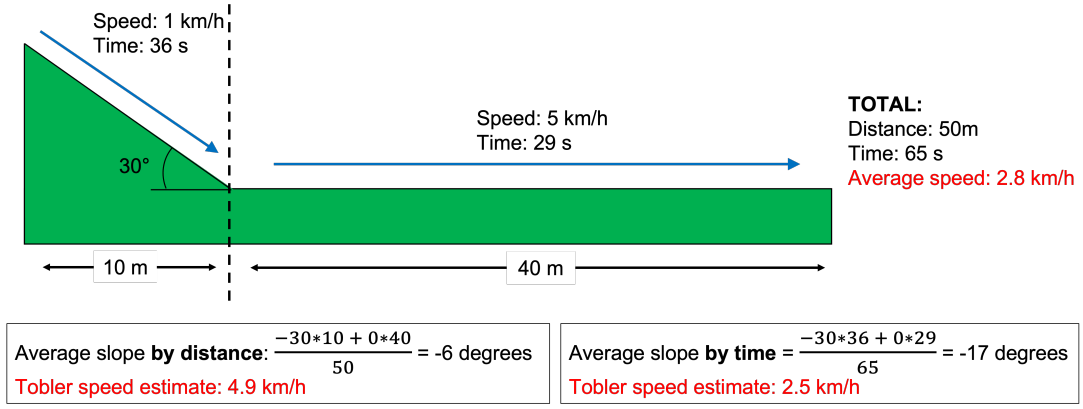


Figure 3.8: Comparison of merging slope angles by distance or by time. Demonstrating how calculation of the average slope when merging datapoints together produces different results if weighting the average by the distance travelled, or by the time taken in each section.

average speed of 2.8 km/h. The average slope if weighted by the distance travelled on each slope is -6 degrees, whereas weighting by the time spent on each slope produces an average slope of -17 degrees. The Tobler hiking speed predictions for these slope angles are 4.9 km/h and 2.4 km/h respectively. Weighting the slopes angles by time therefore produces an average slope which can explain the relatively slow average speed observed.

Although all of the Hikr data was tagged as a walk or hike, there were a small number of individual track segments with high average speeds. Upon further inspection it was clear that these were segments, within a larger walking track, where other modes of transport were used and should be filtered out. Furthermore, a minimum distance between breaks was added to ensure that we were only investigating sustained periods of movement.

1. Segments with an average speed greater than 10km/h were removed.
2. All points where there was less than 250m of usable data between breaks were ignored. These points could occur through people moving during their break, therefore we assumed they were unlikely to be part of a sustained walk.

For each remaining track segment the median, upper quartile and maximum speed were calculated, and statistics from these were found to use in filtering the OpenStreetMap data to remove non-walking tracks:

The upper quartile of the maximum Hikr speeds (approx. 5.9 km/h)	(3.1)
The median of the median Hikr speeds (approx. 3.0 km/h)	(3.2)
The upper whisker of the maximum Hikr speeds (approx. 7.5 km/h)	(3.3)
The minimum of the upper quartile Hikr speed (approx. 2.4 km/h)	(3.4)

The OpenStreetMap data were also merged into intervals of at least 50 m, while ignoring short micro-breaks. Similarly to the Hikr tracks, this dataset contained a number of tracks where there were clearly multiple movement methods, often when the user was driving or cycling to a hike location. Unlike in the Hikr data however, a change in transport method was often not accompanied by the start of a new track segment. Instead, individual route segments appeared to contain a variety of transport methods, often separated by a number of very extreme points (likely where the device lost signal for a period). As we did not want to remove these segments entirely if they contained valid walking data, we used the extreme points as markers to break the segment down into smaller sections. The 'key points' were defined as the start and end points of the segment, as well as any point with a distance greater than 500 m, a duration greater than 10 minutes, or a speed over 100 km/h. An example of a track which is clearly made up of two distinct regions separated by a key point is shown in Figure 3.9. After identifying the key points, the following conditions were applied:

- If only a single data point existed between a pair of key points it was ignored
- If the median speed between a pair of key points was greater than (3.1), then all points in the range were ignored

Following this, all segments were checked and the steps outlined below were carried out to remove unwanted data. These were repeated until no further data was removed.

1. If the segment contained less than 2.5 minutes of useable data it was removed
2. Segments were removed if any of the following were true:
    - The median speed was greater than (3.1)
    - The minimum speed was greater than (3.2)
    - The upper quartile speed was greater than (3.3)
    - The upper whisker speed was less than (3.4)
3. All points with a speed above 10 km/h which were at the start or end of a segment, or next to a break point were considered part of the break.

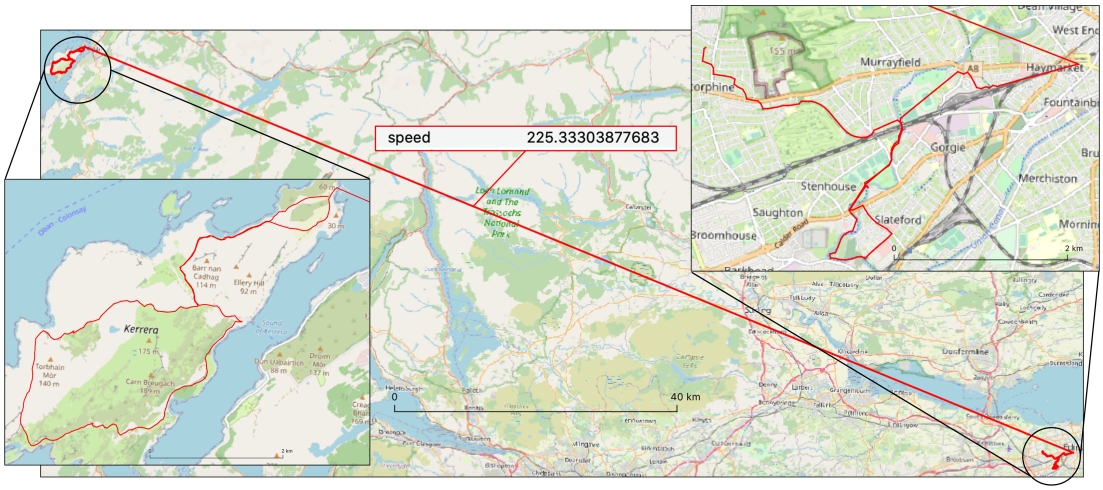


Figure 3.9: Example of a 'key point' separating two distinct track sections. Background map from OpenStreetMap, visualised using QGIS (see 2.2.5).

4. All points where there was less than 250 m of usable data between breaks were ignored.

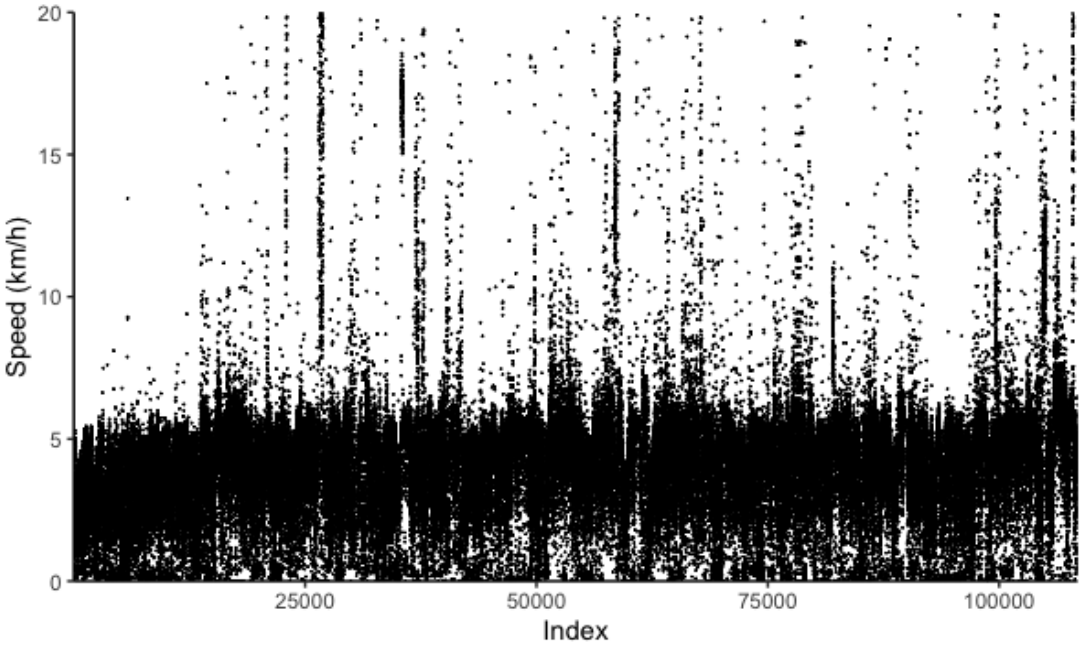
Algorithm 2 summarises the filtering process which was applied to each GPS track segment. For the remaining segments, the Hikr and OpenStreetMap data were then combined together into a single dataset. The impact of removing breaks and non-walking sections can be seen in Figure 3.10. The mean walking speed in 3.10a is 12.33 km/h, while the mean walking speed in 3.10b is 4.14 km/h. A small number of high-speed points can be seen in 3.10b, likely due to noise in the GPS data which was not picked up by the break filter. These were left in as they made up a very small proportion of the overall dataset (<0.25% of the points had a speed greater than 10 km/h), and there was no objective way to remove such points without potentially removing valid data as well.

All of the data remaining was assumed to consist solely of hiking or walking tracks, although there are likely to be a number of areas where this was not the case. There is not a large difference in speed profile when walking or cycling up a steep incline and without in-depth analysis of each individual track we could not be certain that only hiking data remained in the dataset. However, the volume of data (approximately 80,000 points or 4,500 km of travel) should alleviate errors arising from this issue. A map showing all of the data used can be seen in Figure 3.11.

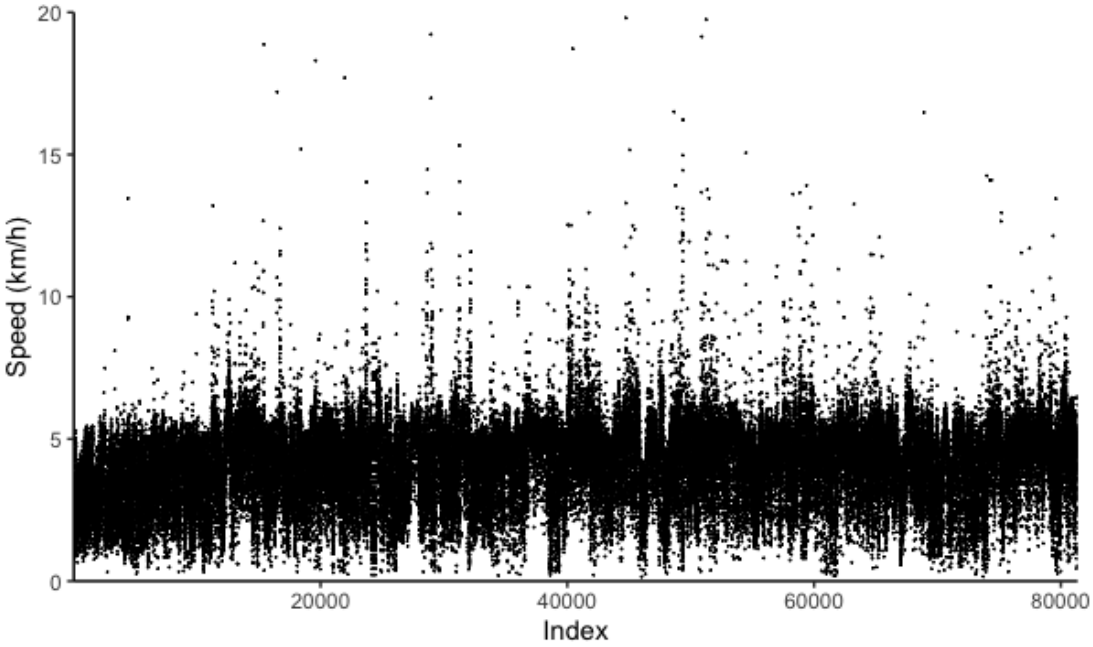
Algorithm 2 Filtering process for Scotland data

```
Remove all breaks with duration > 30 seconds
Remove all breaks containing points with speed > 10 km/h or distance > 1km
Merge remaining points into 50 m sections. (Remove sections under 50 m in length which precede a break or the end of the file)

if Hikr data then
    if segment mean speed > 10 km/h then
        Remove segment
    end if
    Remove points where there is less than 250 m of movement between breaks
    Calculate filtering bounds (Equations 3.1 - 3.4)
else
    Identify Key Points (start, end, distance > 500m, duration > 10 min or speed > 100 km/h)
    Remove single datapoints between Key Points
    Remove points where median speed between consecutive key points > (3.1)
    while segment length is not consistent do
        Remove points with speed > 10 km/h adjacent to a break, or the end of the track
        Remove points where there is less than 250 m of movement between breaks
        if segment median speed > (3.1) or segment minimum speed > (3.2) or segment upper quartile speed > (3.3) or segment upper whisker speed < (3.4) or segment contains less than 2.5 minutes of data then
            Remove segment
        end if
    end while
end if
```



(a) Walking speeds before breaks and non-walking sections are removed.



(b) Walking speeds after breaks and non-walking sections are removed.

Figure 3.10: Walking speeds of the dataset before and after points identified as breaks or non-walking sections are removed. The x-axis in each case represents an ordered list of all datapoints. Note that both graphs have their y-axes limited to 20 km/h for clarity.

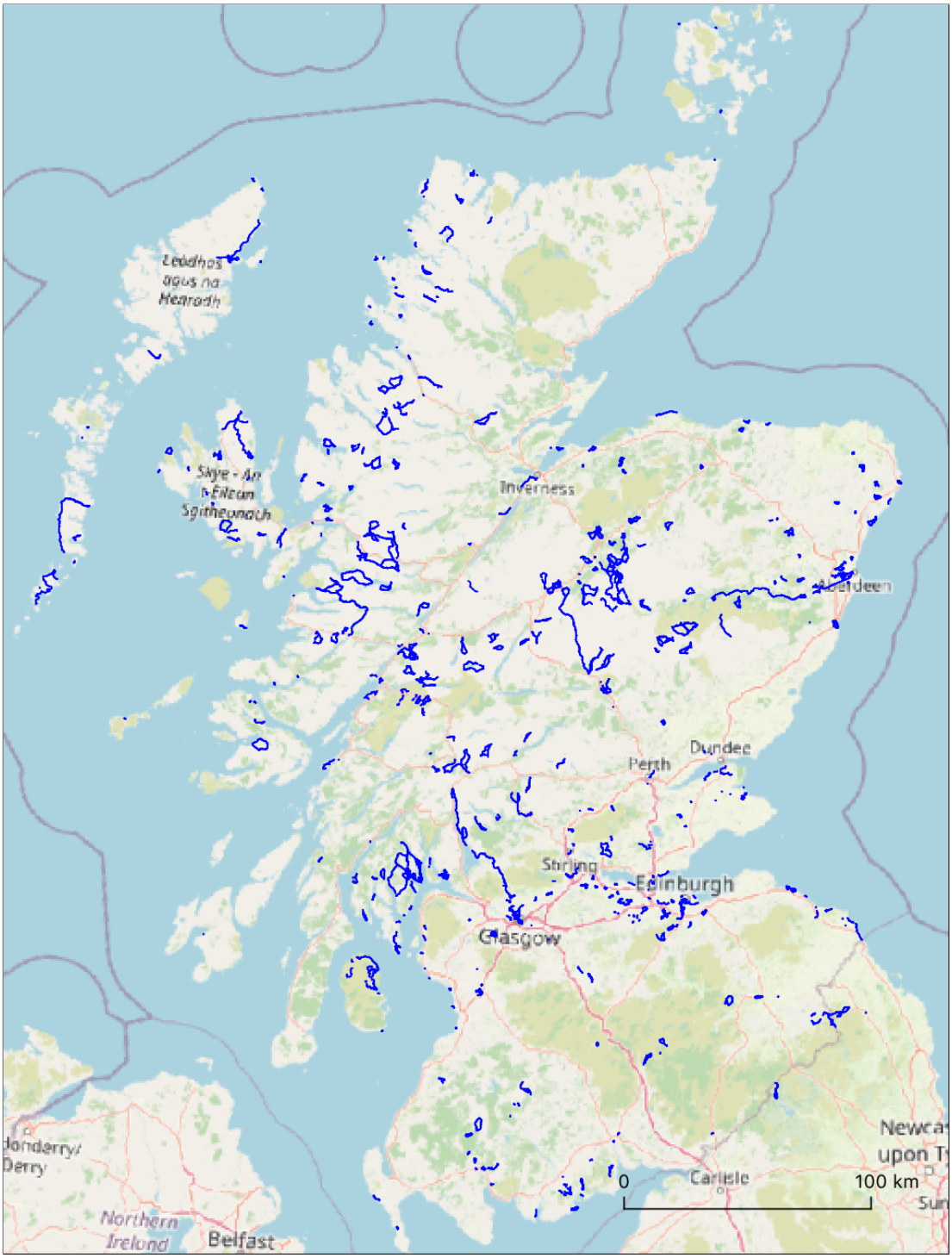


Figure 3.11: Map showing all of the GPS tracks used in data analysis for Scotland. Background map from OpenStreetMap, visualised using QGIS (see 2.2.5).

3.2 Modelling

Once the data were filtered and processed, we were able to use them to build a model to predict the walking speed. Two different approaches were explored in order to model the data: a generalised linear model (GLM) and a generalised additive model (GAM). The GLM method allows us to produce an easily interpretable model which can be applied to real world scenarios. A GAM approach allows us to explore more complex models, by fitting a flexible function (spline) to each coefficient (James, Witten, Hastie, & Tibshirani, 2013).

In general a more complex model will fit the data better. For GLMs, increasing the complexity involves increasing the number of terms in the model, while GAM complexity increases as the number of knots in each spline is increased. Although more complex models fit the data better, this comes with the trade off of being more likely to overfit the data; that is to also fit the random variance found within the modelling dataset. In this case the resulting model would not generalise well to new data.

We know that predictions for walking speeds must be non-negative, and two different setups were explored to achieve this: a Gaussian distribution with log link function and a Gamma distribution with inverse link function. The GAM approach was also deployed with both thin plate spline or cubic regression basis functions. Initial investigations into different models showed that there was no improvement to model fit beyond cubic terms in a GLM, or 7 knots in each GAM smoothing term, so more complex models than this were not considered for selection.

Both model types were created in R:

$$\texttt{glm}(v \sim a\phi + b\phi^2 + c\phi^3 + d\theta + e\theta^2 + f\theta^3, \texttt{distribution})$$
$$\texttt{gam}(v \sim s(\phi, k, \beta) + s(\theta, k, \beta), \texttt{distribution})$$

where

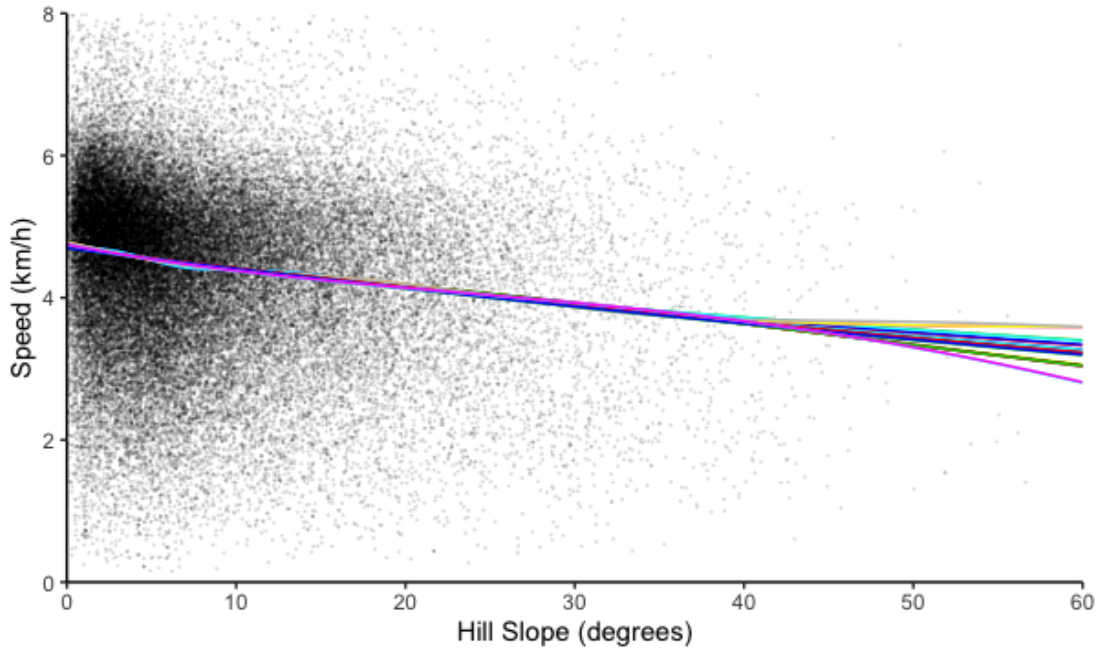
$v =$ walking speed
$\phi =$ hill slope angle (degrees)
$\theta =$ walking slope angle (degrees)
k = knots used in spline (up to 7)
$\beta =$ basis function

3.3 Model Selection

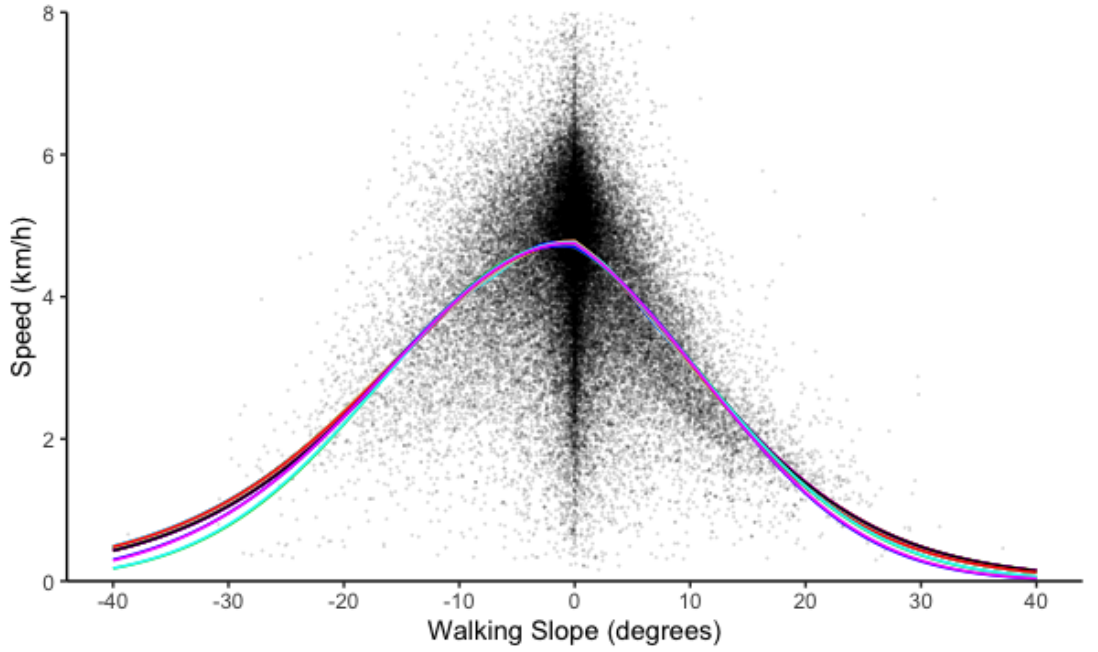
Initially, 10-fold cross-validation was used to compare the model parameters, looking at R-squared values, root-mean-squared error (RMSE) and mean absolute error. All models produced very similar results, with no change in the RMSE to 2 decimal places, although there was a general trend of marginal improvements as the model complexity increased. As no best model could be chosen based on the cross-validation, each was checked in more detail. Firstly, the hill slope component was isolated by investigating the speed when the walking slope was zero (i.e when traversing across a slope). Intuitively, and from experience, this should be a decreasing function; as the slope gets steeper it is harder to traverse, so the walking speed will decrease. Models which failed to predict this were removed under the assumption that the data were overfitted. Following this, the walking slope component was investigated, specifically looking at the walking speed when travelling directly up- or down-hill. By inspection of the data, existing functions, and intuition, this should be modelled as a roughly bell-shaped function with the peak at, or close to, 0 degrees. Any models which predicted an increase in speed as walking slope steepness increased (from a minimum magnitude of 10 degrees) were removed. Secondly, we know from existing work that there exists a critical gradient at a walking slope of around 15 – 21 degrees, at which it becomes more efficient to zig-zag up a steep hill rather than going directly uphill. Models which failed to predict the critical gradient occurring below 21 degrees when travelling uphill were also removed.

This resulted in 21 model configurations remaining, although it is clear from Figure 3.12 that the speed predictions are very similar in most circumstances. Figure 3.12a shows that all of the remaining models predict very similar speeds when traversing a slope of up to 40 degrees, after which there is more deviation in predictions. Similarly, when travelling in the slope direction (Figure 3.12b), all of the models are broadly similar on slopes up to approximately ± 15 degrees. More than 96% of the data is contained within this area, and the relative lack of data outside this region explains the divergent speed predictions. As all of the models provided both very similar R-squared values and very similar predictions over the vast majority of the dataset, we used the following points to make our final selection:

- It is easier to apply GLMs than GAMs to future work, as a simple formula to predict the walking speed can be produced for application elsewhere, without needing to recreate the model from the original data.
- In general, simpler models are easier to interpret, and we had no clear evidence that a more complex model would perform better.



(a) Walking speed predictions for traversing across hills of varying slope, overlaid on GPS data where walking slope is below 5 degrees



(b) Walking speed predictions for travelling directly up or down hills of varying slope, overlaid on GPS data where walking slope is within 5 degrees of hill slope

Figure 3.12: Walking speed predictions from 21 possible models (coloured individually) generated from the Scotland GPS dataset.

3.4 Results

GPS tracks were obtained for hikes in Scotland from Hikr and OpenStreetMap. To get the best accuracy for predicting walking speeds, the data were filtered to remove significant breaks and tracks or segments which were not hike or walking based (Sections 3.1.2, 3.1.3). This gave us a dataset of 596 GPS tracks; 54 from Hikr.org and the rest from OpenStreetMap, consisting of over 80,000 individual data points and over 4,500 km of travel. Each datapoint contained:

- Start coordinate
- End coordinate
- Start time
- End time
- Duration
- Distance
- Speed
- Elevation
- Walking slope
- Hill slope

The elevation and slope values were calculated using data from the Ordnance Survey Terrain 5 dataset. Hill slope values were found using the quadratic surface method, and measure the hill slope at the starting co-ordinate. Where the datapoints in the original GPS track were under 50 m in length, they were merged together to minimise the effects of errors in the GPS location values. While doing this, the resulting distance was the sum of all distances in the constituent GPS points, so may be longer than the straight line distance between co-ordinates. Similarly, both hill and walking slope values were calculated as the weighted average of constituent points, weighted by point duration.

This dataset was then used to create a GLM for walking speed taking into account both walking slope and hill slope (Sections 3.2, 3.3):

$$v = exp(1.547 - 0.006\phi - 0.013\theta - 0.002\theta^2) \tag{3.5}$$

where

$v =$ walking speed (km/h)
$\phi =$ hill slope angle (degrees)
$\theta =$ walking slope angle (degrees)

In this model, all coefficients are highly significant (p-value < 2×10^{-16}). This shows the importance of including the hill slope alongside the walking slope when calculating speeds.

The GLM predicts the critical gradient when travelling uphill to be at a walking slope of 14 degrees, which is slightly lower than has been previously suggested. It also finds a critical gradient of 17 degrees for downhill travel.

When comparing the total walking time for routes as a whole, the GLM improves on existing models at predicting the total walking time, but they are all very similar (15.9% average error in the GLM vs 16.2% for Naismith and 19.7% for Tobler). We suspected the similarity to be caused by the summation of individual errors throughout the route. Instead of looking at routes as a whole we investigated the accuracy of predicted speeds in 50 m segments. As shown in Figure 3.13A, the proposed new GLM has a lower RMSE than existing models when traversing a slope (walking slope = 0 degrees), and the difference in RMSE between the models increases as the hill slope increases, providing further evidence that hill slope is an important variable when predicting walking speeds. Figure 3.13B shows the predicted walking speeds when traversing a hill of a given slope , overlaid onto all points where the magnitude of the walking slope is less than 5 degrees. Also shown is the mean walking speed from this data, and a 95% confidence interval for the population mean. There is a clear downward trend in walking speed which is correctly predicted by the model, but there are relatively fewer points at high slope angles (above 40 degrees), evidenced by the increasing width of the confidence interval for the mean value. Figures 3.13C and D show the same information as A and B, when looking at the effect of walking slope on speeds. These clearly show that Naismith's rule overestimates walking speeds when descending a slope. Furthermore, the new GLM and Tobler's function predict similar walking speeds at most slope angles, with the exception of mild descents, where Tobler's function reaches a maximum of 6 km/h. Comparing the predicted walking speeds to the mean speed, and looking at the RMSE values, we have some evidence to suggest that Tobler's function overestimates walking speeds on mild descents, and our new model improves predictions in this area.

We also see that the uncertainty in our estimate for the mean walking speed increases once the walking slope is steeper than approximately 20 degrees. This is due to a lack of data in this region, which is also likely to be the reason for the larger RMSE values which are observed. It is unclear whether walking is not possible on these slopes (in which case a model should predict a speed of 0 km/h), or whether people are choosing not to (the critical gradient at which walking is most efficient has previously been found on walking slopes of less than 20 degrees, see Section 1.1), so this is a region which should be investigated more explicitly in future.

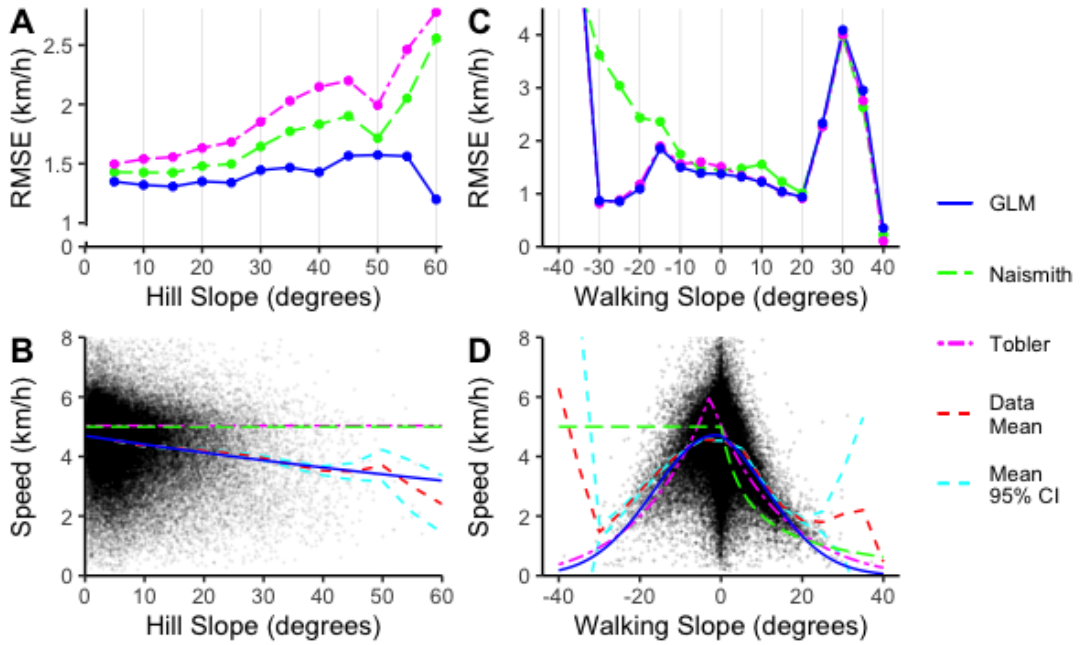


Figure 3.13: A: Root-mean-square error of walking speed when traversing a slope, calculated Naismith's rule (green), Tobler's hiking function (magenta), and the new GLM proposed here (blue). The RMSE is calculated from all data where the walking slope is below 5 degrees, separated into (overlapping) bins of width 10 degrees by hill slope. **B**: Walking speed predictions using Naismith's rule (green), Tobler's hiking function (magenta), and the new GLM proposed here (blue) when traversing a hill of varying slope angle, overlaid on all data points where the magnitude of the walking slope angle is less than 5 degrees. Also shown is the data mean (red) and a 95% confidence interval for the mean value (cyan), separated into (overlapping) bins of width 10 degrees by hill slope. **C**: Root-mean-square error of walking speed when directly climbing or descending a slope, calculated Naismith's rule (green), Tobler's hiking function (magenta), and the new GLM proposed here (blue). The RMSE is calculated from all data where the difference between hill slope and walking slope is below 5 degrees, separated into (overlapping) bins of width 10 degrees by walking slope. **D**: Walking speed predictions using Naismith's rule (green), Tobler's hiking function (magenta), and the new GLM proposed here (blue) when directly climbing or descending hill of varying slope angle, overlaid on all data points where the magnitude of the walking slope angle is within 5 degrees of the hill slope angle. Also shown is the data mean (red) and a 95% confidence interval for the mean value (cyan), separated into (overlapping) bins of width 10 degrees by walking slope.

3.5 Discussion

We have taken a crowdsourced dataset of GPS data, and filtered it to remove breaks and non-walking sections. We then explored multiple model formulations to predict the walking speed based on the hill slope and the walking slope, and found a GLM model was able to accurately explain the data, whilst also being easy to interpret and apply to practical implementations.

Our new model represents a new walking speed function generated using over 4,500 km of real-world walking data. This model provides two clear improvements over the most widely used existing hiking functions. Firstly that it was generated from such a large dataset increases the reliability of the results, and secondly that it takes into account the impact that hill slope has on walking speeds (a factor we show to be highly significant) which has not been included in previous methods.

One reason for the lack of inclusion of hill slope in previously described walking speed formulae is the difficulty of calculating the slope value. Previously, this would have required either measuring the slope value along routes as part of the data collection process, or only collecting data in a relatively small area, as wide-scale slope data did not exist. However, the release of the OS Terrain 5 DTM (Ordnance Survey (GB), 2020c), which provides elevation data at 5 m intervals across the whole of the UK, allowed us to include hill slopes in the model.

While building this model, we only used a small subset of the total data available to us (Scotland), as there were multiple processing methods which had to be designed and implemented. If we expand our dataset to explore the full region, then we will be able to both test the choices made during the break finding and data filtering process, as well as whether our model still fits a larger dataset. Furthermore, to this point we have only considered one of the two additional variables which we are investigating. The next evolution of this walking speed model should explore the impact of different terrain types on walking speeds.

Chapter 4

Expanding the Dataset: Model Validation and Extension

4.1 Introduction

In the previous chapter we described an improved model for walking speed that includes hill slope. Here we attempt to validate this model on a larger dataset. Next we assess the impact of terrain on the model.

The dataset used for validation was made up of all UK GPS tracks available on Hikr.org (prior to July 2021) and through the OpenStreetMap data dump. This dataset was processed in generally the same manner as described in Chapter 3, with a number of changes. These changes were mainly required due to features of the additional data which were not removed by the previous filtering method, but visual inspection showed that they were not part of a valid walk. This was likely due to the additional tracks being recorded on previously unseen devices, which had different settings to those previously encountered. The changes to the processing methods were as follows:

- All GPS track segments were now fully contained within the area covered by the full OS Terrain 5 DTM, rather than the reduced grid used in Section 3.1.
- Before identifying breaks, any track segments where the median speed was greater than 10 km/h were automatically removed.
- Individual points representing 3 minutes of travel were now tagged as breaks (down from 10 minutes as used in Section 3.1.2).
- High speed (>10 km/h) points occurring immediately following a long (>3 minute) point were now tagged as breaks. Situations like this occurred on a number of devices, likely when a device automatically paused recording for a break. Once significant movement was detected, two points were then added in quick succession; the first at the original location with the time the movement started again, and the second at the new position of the device. This resulted in one high duration point with very little movement, followed by single-second duration movement with high speed, where the device ‘caught up’ with the correct location.

- When merging data into 50 m sections, sections under 50 m in length immediately preceding a break or the end of the segment were now combined with the previous section, rather than ignored as was done in Section 3.1.3.
- After merging the data into 50 m sections, any section with a speed above 10 km/h which was at the start or end of a segment, or next to a break point was considered to be part of the break. This was repeated recursively until no more high speed points were found next to breaks. (This filter was already used for the OSM data - see Section 3.1.3, and was added to the Hikr data process).
- The requirement for 250 m of movement between break points was removed. In difficult to navigate regions, such as climbing a steep hill, we found that valid data was being incorrectly removed.
- When looking for 'key points' to filter out non-walking track sections, the duration required to be tagged as a 'key point' was reduced to 3 minutes (down from 10 minutes used in Section 3.1.3).
- Duplicate track segments (those which contained a section with the same start and end coordinates, the same starting date and time, and the same duration) were tagged, and only the first instance of each track was retained. It is believed that these duplicates were a result of either users uploading the same track multiple times, or GPS devices recording multiple versions of the same track with automatic filtering, or different levels of accuracy.
- The fastest and slowest 0.5% of the merged datapoints were removed as outliers (classified as breaks for further processing) and were not included in any modelling.

The updated breakfinding and data filtering processes are shown in Algorithms 3 and 4 with changes highlighted in blue. On top of these changes to the filtering process, a small number of outlier Hikr tracks were checked manually and removed; looking at the metadata or track description showed that the routes in question were labelled as trailruns, and so should not be considered for this work. This left us with a final dataset of almost 93,000 km and over 8,200 tracks.

The dataset was then split in two: Scotland and the rest of the UK (ROUK). The Scotland dataset consisted of all of the track segments which were used to create the model described in Chapter 3, while all remaining tracks were part of the ROUK dataset (to be used validate the model described in Chapter 3). In practice this meant that the ROUK dataset was generally made up of GPS traces covering England and Wales (as Northern Ireland is not included in the OS Terrain DTM). There were a few exceptions to this, such as one Hikr track in the Shetland Islands (a result of the Shetland Islands being a distinct region from Scotland on Hikr.org, and thus not included in the Scotland download), as well as segments which had points in both England and Scotland. The Scotland dataset contained 4,950 km of walking data, while the ROUK dataset contained approximately 88,000 km.

Algorithm 3 Breakfinding process for full dataset

Breakpoint_list $= \emptyset$
Find the median distance (r_{median}) and speed (s_{median}) of the segment
for point (p_i) in segment **do**
    Calculate travel direction quadrant and point angle
    Calculate break likelihood using the point speed and angle ▷ see Table 3.1
    if speed == 0 or distance > 1 km or duration > 3 minutes **then**
        Breakpoint_list += p_i
    end if
    if speed > 10 km/h and duration(p_{i-1}) > 3 minutes **then**
        Breakpoint_list += p_i
    end if
end for
for point (p) in segment **do**
    if Neighbourhood of p is a cluster (C) **then** ▷ See Definitions 1 & 2
        for point (p_c) in C **do**
            if Neighbourhood of p_c is a new cluster (C_n) **then**
                $C = C \cap C_n$
            end if
        end for
        Remove points at the ends of the cluster with low break likelihood
        Add 'missing' point ids to the cluster (to make a continuous run of points) to form a Potential Break (B*)
        if less than half the points in B* have low break likelihood and there is travel in opposite quadrants (Q1 & 3 or Q2 & 4) **then**
            Breakpoint_list += B*
        end if
    end if
end for

Algorithm 4 Filtering process for full dataset

Remove duplicate segments (containing sections with identical start location, end location, start time and duration)
Remove all segments with median speed > 10 km/h
Remove all breaks with duration 30 seconds
Remove all breaks containing points with speed > 10 km/h or distance > 1km
Merge remaining points into 50 m sections. (Sections under 50 m in length which precede a break or the end of the file are combined with previous section)
Recursively remove points with speed > 10 km/h adjacent to a break, or the end of the track

if Hikr data **then**
    if segment mean speed > 10 km/h **then**
        Remove segment
    end if
    ~~Remove points where there is less than 250 m of movement between breaks~~
    Calculate filtering bounds (Equations 3.1 - 3.4)
else
    Identify Key Points (start, end, distance > 500m, duration > 3 min or speed > 100 km/h)
    Remove single datapoints between Key Points
    Remove points where median speed between consecutive key points > (3.1)
    while segment length is not consistent **do**
        Remove points with speed > 10 km/h adjacent to a break, or the end of the track
        ~~Remove points where there is less than 250 m of movement between breaks~~
        if segment median speed > (3.1) **or** segment minimum speed > (3.2) **or** segment upper quartile speed > (3.3) **or** segment upper whisker speed < (3.4) **or** segment contains less than 2.5 minutes of data **then**
            Remove segment
        end if
    end while
end if

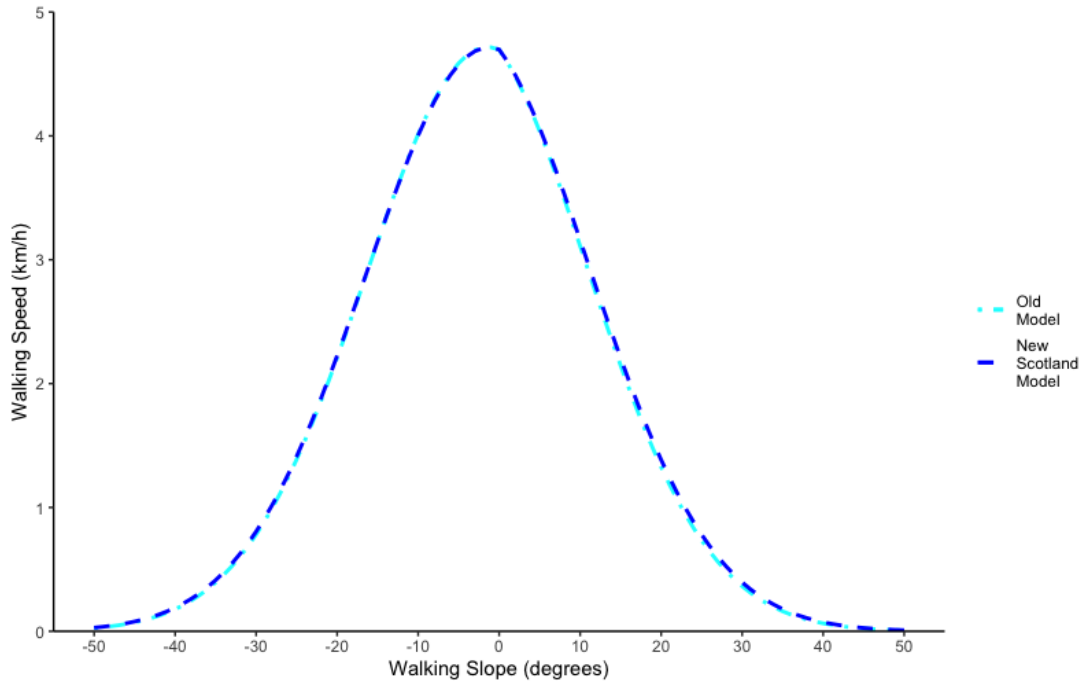
Combine all segments into single dataset
Remove the fastest and slowest 0.5% of the data

As a first step, we wanted to ensure that our adjusted filtering methods did not affect the model which we found previously. The same GLM model form used in Chapter 3 was run on the new Scotland dataset and compared to the previous result. As shown in Figure 4.1, the two models are almost identical. This showed that our change in filtering methods did not have a large impact on the model, and allowed us to directly compare and contrast results found using the Scotland and ROUK datasets, as we could compare results where both datasets were subject to the same processes. It is not surprising that changing the filtering method had very little impact on the Scotland model; the changes were generally required due to features of the additional ROUK data which were not present in the Scotland data (due to different devices and settings being used), and thus the majority of the Scotland data was unaffected by the change. The relatively small number of points which were removed under the updated filtering method were enough to significantly affect the results of the model.

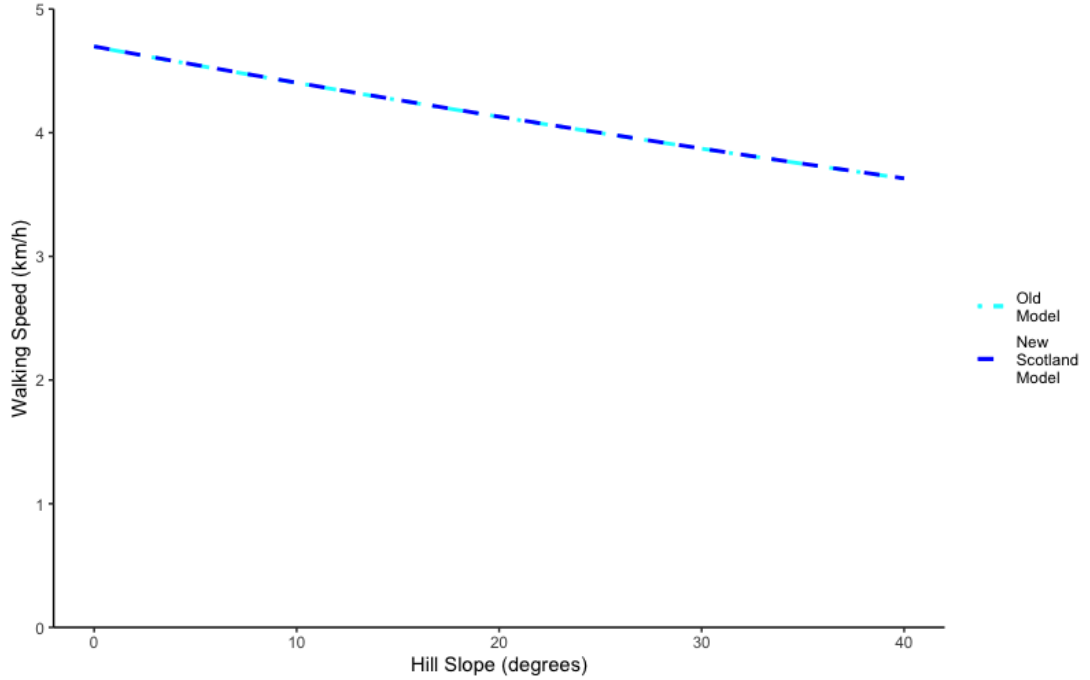
4.2 Initial Model

The GLM model formulation found in Chapter 3 was applied to both the Scotland and ROUK datasets, and the results can be seen in Figure 4.2. Two things are visible from this model: firstly, once again all of the variables were found to be highly significant and secondly, we see faster predicted walking speeds in the ROUK model than in the Scotland model when traversing a slope, or when walking uphill.

Before looking for causes of this speed difference, we wanted to check the significance of the result. There were a couple of factors which had to be taken into account when checking the confidence intervals for the two models. Firstly, the datapoints were not independent; each track contained data from a single GPS device, (and likely a single individual), so all datapoints in a given track would be correlated. This meant that the standard deviations produced by the GLM formula in R could not be used (as this assumes independence). We therefore needed to calculate adjusted standard errors for each variable, to take into account the clustering by track. After applying this adjustment, we found that the confidence intervals did not overlap for 3 of the 4 variables, suggesting that the two datasets were different. However, we also needed to take into account the fact that the ROUK datset was much larger than the Scotland dataset (7636 tracks vs 648). We wanted to test the likelihood of finding the Scotland data result within the ROUK dataset (i.e. was the Scotland dataset a possible sample of the ROUK data?). We took 100 samples of 650 tracks from the ROUK dataset (to form sample sets of comparable size to the Scotland data) and modelled the walking speed for each one. The results are visualised in Figure 4.3.

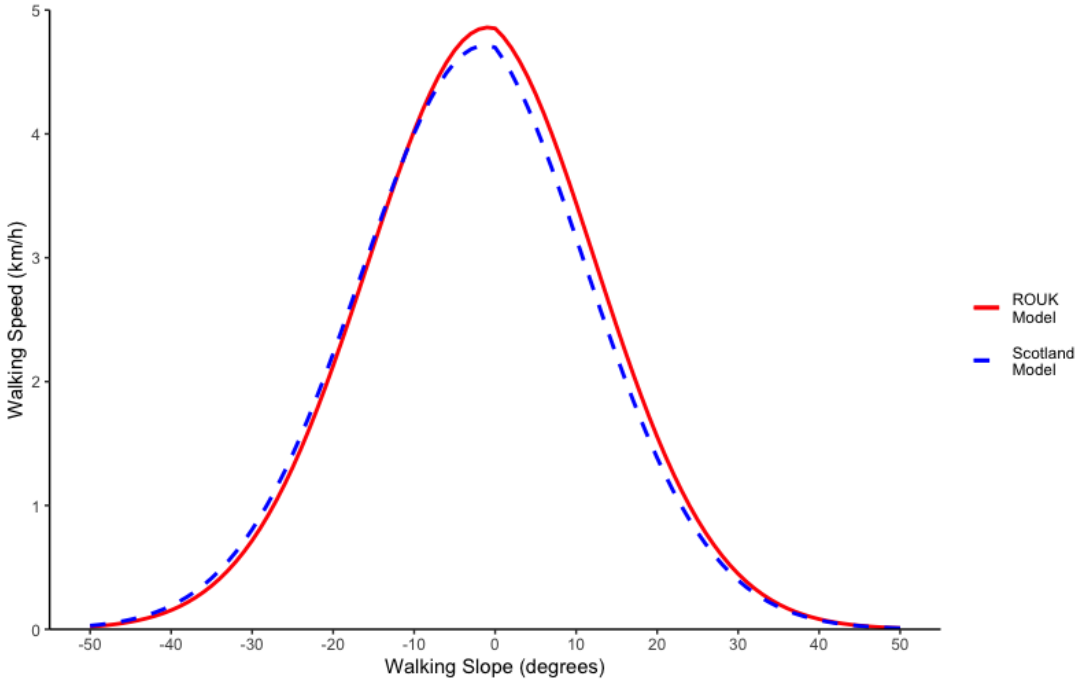


(a) Walking speed predictions for travelling directly up or down hills of varying slope.

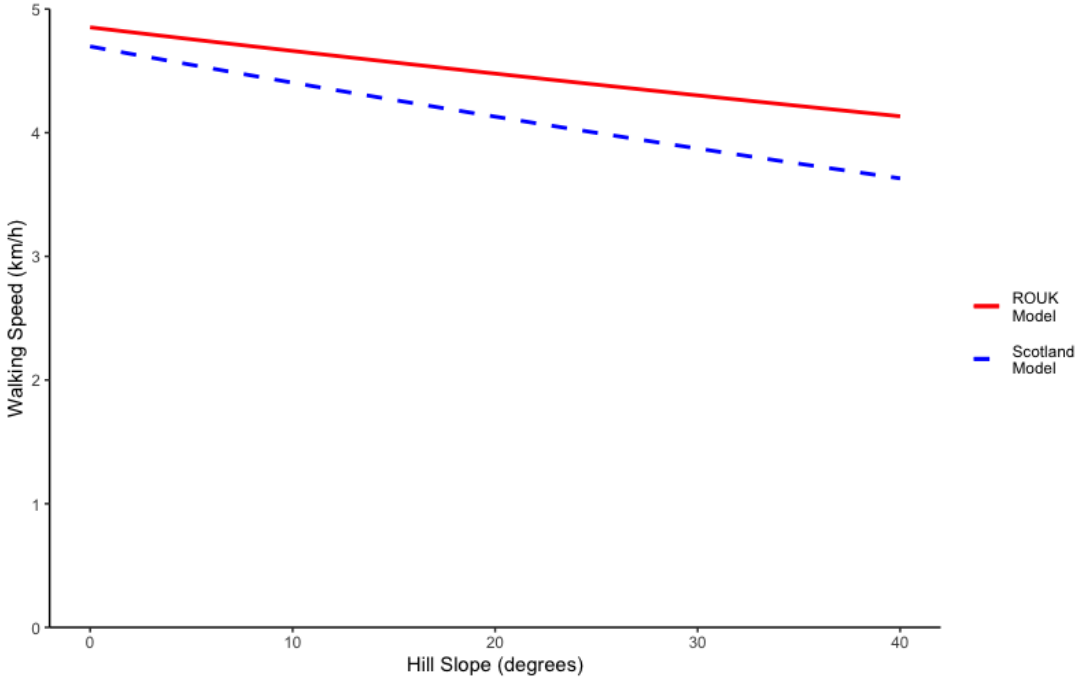


(b) Walking speed predictions for traversing across hills of varying slope.

Figure 4.1: Comparison of walking speed models produced when Scotland data is processed using different filtering methods.

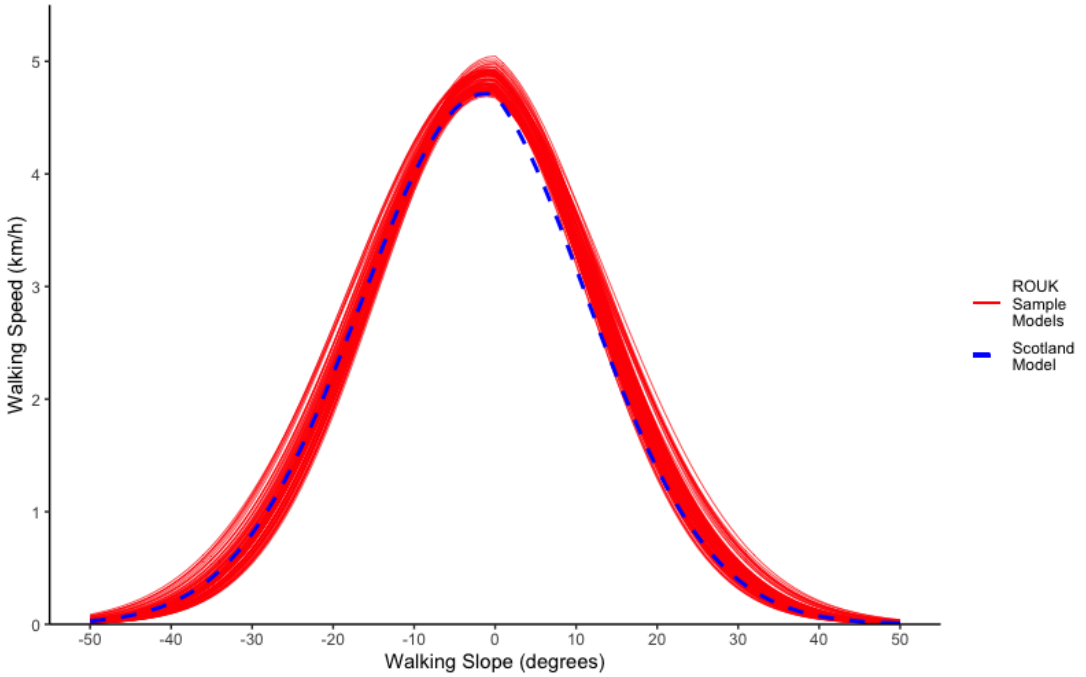


(a) Walking speed predictions for travelling directly up or down hills of varying slope.

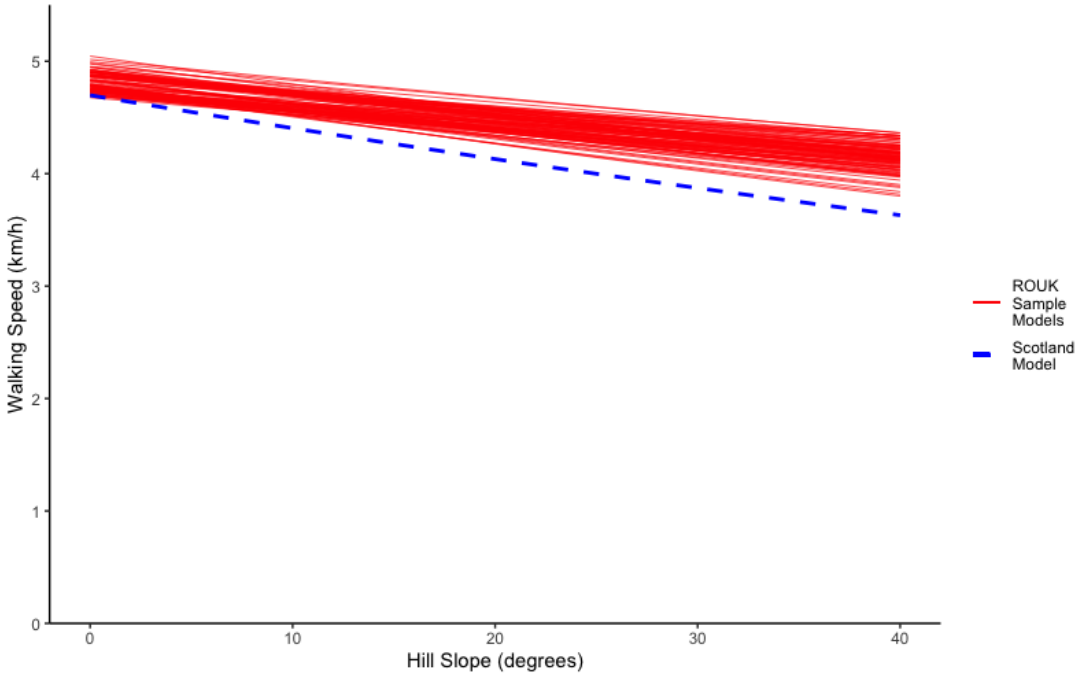


(b) Walking speed predictions for traversing across hills of varying slope.

Figure 4.2: Comparison of walking speed models produced using data from Scotland and the rest of the UK.



(a) Walking speed predictions for travelling directly up or down hills of varying slope.



(b) Walking speed predictions for traversing across hills of varying slope.

Figure 4.3: Comparison of walking speed models produced using data from Scotland against 100 sampled datasets from the rest of the UK.

When ascending or descending the slope, the Scotland data falls within the sample range (although it is at the extreme end of the range in some cases). However, when we look at traversing a hill, it is clear that the two datasets are distinct, as the model for Scotland is outside the range of results seen in the ROUK sample models. Up to this point, we had not considered the terrain type within our models. We wanted to extend our models to take the terrain information into account, and see if this could explain the difference between the two datasets.

4.3 Terrain Classification

The OSM road and path network and OS Topography dataset described in Section 2.2.4 were used to identify the terrain types for each point.

4.3.1 Roads and Paths

The OSM road and path data classifies roads into the following types:

- Bridleway
- Cycleway
- Footway
- Living_street
- Motorway
- Motorway_link
- Path
- Pedestrian
- Primary
- Primary_link
- Residential
- Secondary
- Secondary_link
- Service
- Steps
- Tertiary
- Tertiary_link
- Track
- Track_grade1
- Track_grade2
- Track_grade3
- Track_grade4
- Track_grade5

- Trunk
- Trunk_link
- Unclassified
- Unknown

The road and path data is made up of linestrings (single vector lines, as opposed to area features which match the width of the road) and it would be very unlikely for all of the points on a GPS route which follows a road to fall exactly along that line, making it difficult to determine which points should be classified as following the road or path. This problem was then exacerbated by two further factors; inaccuracy within the GPS readings and GPS drift (previously discussed in Section 3.1.2), and inaccuracies within the map data itself.

The OSM road data uses a combination of crowdsourced GPS data, alongside aerial photography and older, out of copyright, maps to build their road network. Each of these methods is susceptible to some error, so we could not be certain that the mapped locations of roads and paths indicated their true location. Note also that OpenStreetMap suggests that users not update road locations unless their GPS data is greater than 20 m from the mapped location (OpenStreetMap Wiki, 2022a). An example of these errors is shown in Figure 4.4, in which a GPS track (red) is following a footpath (purple). The track follows the shape of the footpath, but has a consistent deviation of up to 20 m from the position of the path in the OSM data, likely as a result of both mapping and GPS error. The combination of potential GPS and map error meant that we needed to decide upon a radius around each point to search for roads or paths.

During the first exploration into the impact of roads and paths we did not distinguish between the different road or path types. Instead, all GPS points were classified as either on-road or off-road. A point was classified as being on-road if a single feature of the OSM road dataset was found in a 50 m radius around the point. This radius was determined after a number of iterations. It is likely that this was too far in many cases, and would lead to over classification of on-road points, but this was accepted as a necessary limitation of the work. Due to the relative numbers of on-road vs off-road points, it was preferential to mis-classify points as being on a road rather than the other way around, as this would have a much smaller impact on any resulting models. (Note that when we merged our data points into 50 m intervals, the merged point was classified as on-road if at least one of the constituent points was classified as being on-road.)

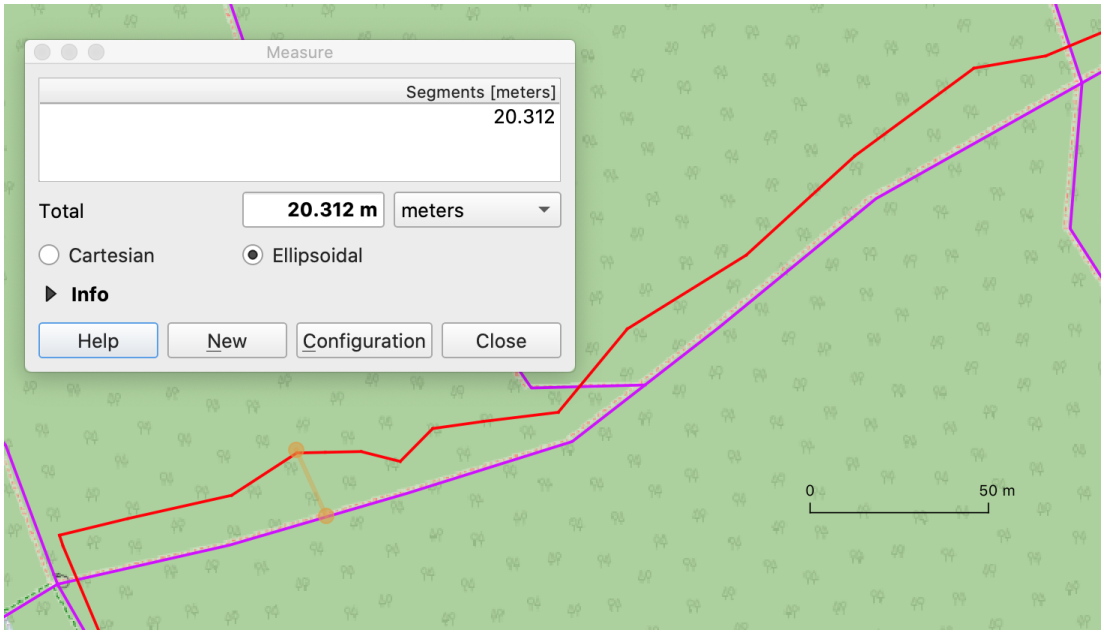


Figure 4.4: Deviation between a GPS track (red) following a footpath and the OSM data footpath position (purple). A point on the track approximately 20 m away from the footpath is marked. Path data from OpenStreetMap contributors (2021b), background map from OpenStreetMap, visualised using QGIS (see 2.2.5).

4.3.2 Off-Road Terrain

Terrain Description

The OS MasterMap Topography dataset provides standardised descriptions of the terrain type. Each feature in the dataset can have any number of the following associated with it:

- Agricultural Land
- Cliff
- Slope
- Boulders
- Boulders (Scattered)
- Coniferous Trees
- Coniferous Trees (Scattered)
- Coppice
- Heath
- General Surface
- Marsh
- Saltmarsh
- Nonconiferous Trees
- Nonconiferous Trees (Scattered)
- Orchard

- Osiers
- Reeds
- Rock
- Rock (Scattered)
- Rough Grassland
- Scree
- Scrub

Unlike for the road data, a radius was not used to search around points for terrain features, instead only the terrain feature at the point location was sampled. The terrain data is presented as a series of polygons which are usually relatively large, so error in the GPS device was unlikely to cause us to miss the feature. Furthermore, neighbouring features are generally closely related, with identical, or almost identical, lists of terrain types. (Note that within towns and cities the features are much smaller, denoting individual buildings, but these are all classified as 'General Surface'.) For these reasons, it was felt that it was unnecessary to expand the search radius for terrain types.

Initial investigations into the terrain descriptions revealed a number of problems. Firstly, because each feature could have multiple terrain types associated with it, there were a very large number of different terrain combinations, and it would have been very difficult to separate these to determine individual terrain effects. Furthermore, the low resolution of the data meant that it was often imprecise in its classification. A large area may be classified as a single feature with multiple terrain types, but a GPS point in that feature would not be affected by all terrain types simultaneously. An example of this is shown in Figure 4.5, which shows the feature highlighted in Figure 2.5 in more detail, alongside an aerial view of the same area (aerial imagery via Bing Maps (2022)). The feature has an area of almost 200,000 m^2 and 4 terrain types associated with it, including heath. However, looking at the aerial imagery, we can see that the terrain is not uniform throughout the entire region, and the areas of heath are restricted to the right hand side of the feature (the patches of darker green). If a GPS point were recorded within the area of this feature, the terrain type dataset would not allow us to distinguish which individual terrain type was affecting the walking speed. For this reason, we chose to use an alternate measure of terrain obstruction, namely the obstruction height, which we would subsequently attempted to link back to the original terrain descriptions (Section 4.7.3).

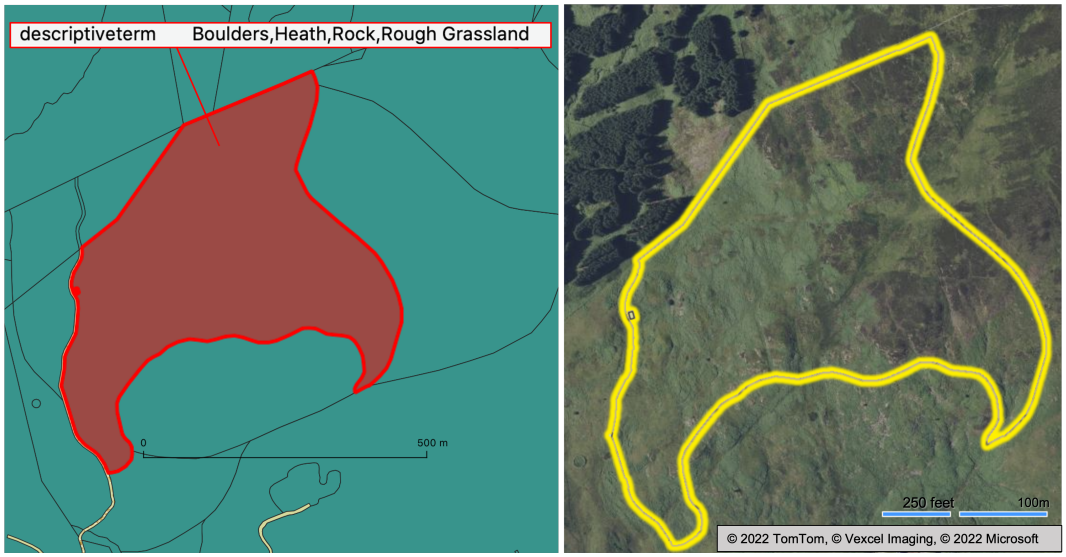


Figure 4.5: Comparison of OS MasterMap Topography data (left), and an aerial view of the same region (right). Topography data from Ordnance Survey (GB) (2020b), visualised using QGIS (see 2.2.5). Aerial imagery via Bing Maps (2022)

Obstruction Height

We calculated the obstruction height for each point in the dataset using the lidar DTM and DSM described in Section 2.2.4. These provide us with the ground height and the surface height respectively. The starting coordinate for each point in a GPS track was sampled in both the DTM and DSM, and the difference between the two values taken as the level of terrain obstruction for the point. For example, a point in a woodland may have a DTM height of 80 m above sea level, and a DSM height of 85 m (the height at the top of the tree canopy), giving us a terrain obstruction value of 5 m. Figure 2.6 provides an example of this, where heights of individual trees and bushes can be determined by taking the difference between the two images. When merging points into 50 m sections, we calculated the average obstruction height as the weighted average (by time) of all constituent points. As described in Section 2.2.4, the lidar data was only available for areas in England and Wales (at 2 m resolution), so we were unable to use it on our Scotland dataset.

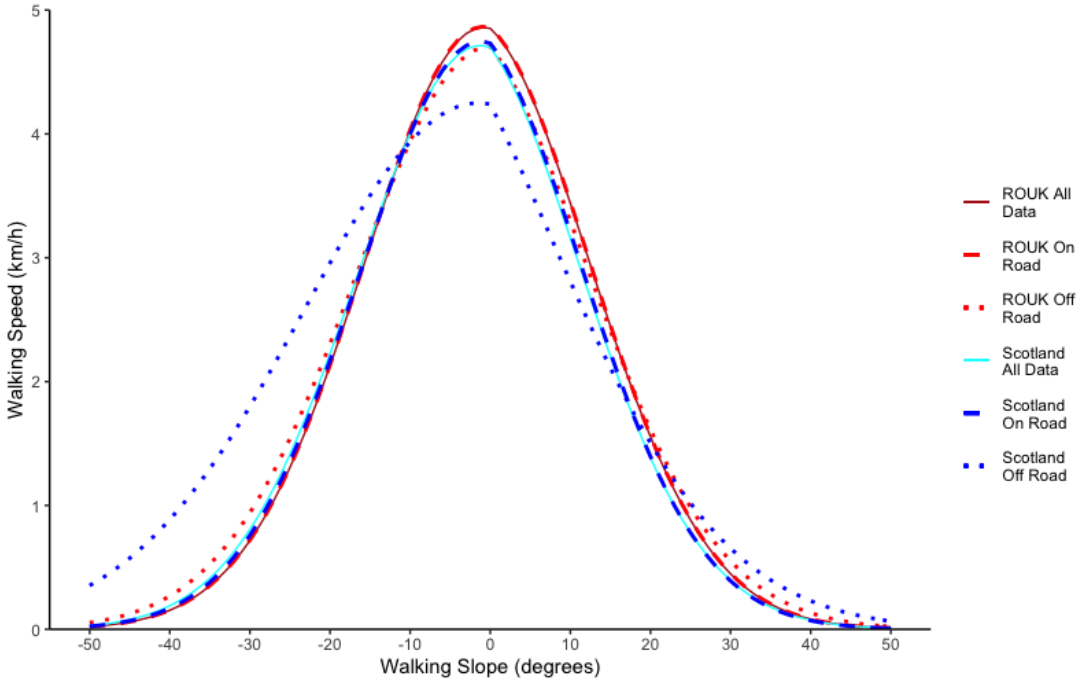
4.4 Roads and Paths

When exploring for potential differences in terrain type between Scotland and the ROUK datasets, the first factor we considered was whether the walk was on- or off-road. Previous walking speed methods have included off-road factors to lower the predicted walking speed, so if the Scotland dataset had a greater proportion of off-road data than the ROUK dataset, this could explain the observed slower walking speeds. For our initial inspection we were interested solely in the presence or absence of a road, not what type of road it was. We found that 96.5% of points in the ROUK dataset were identified as being on-road, while this number was 90.7% for the Scotland dataset. Although this difference is not large, and probably could not explain the full difference in speeds between the different models, it could still be significant. On top of this, we were now able to test the modifiers which have been previously used in walking speed predictions for off-road travel.

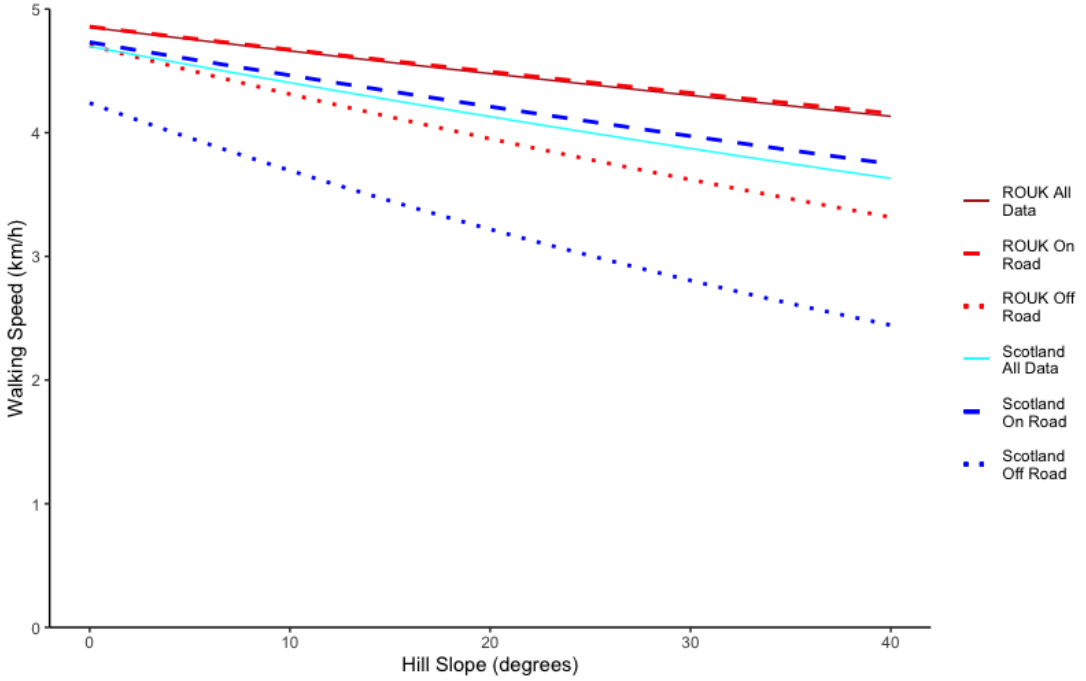
To start, 'OnRoad' was added as a variable into the GLM model found in Chapter 3, both as a factor variable in its own right, and as an interaction term with each of the existing variables. In all cases, this variable was found to be highly significant (at the 95% level), so we can be confident that there is a relationship between the presence of roads, and the attainable walking speed. Each dataset was split into two subsets to investigate this more closely; on-road and off-road. Once again, we fit the GLM model to each of the datasets, and the results are shown in Figure 4.6.

In the most common walking slope ranges (up to ± 10 degrees), we see the expected drop off in speed between being on-road and off-road. However, the decline is not as great as described in previous models. For walking slope angles up to ± 10 degrees, the off-road walking speed is, on average, 96% value of the on-road speed in the ROUK model, and 89.9% in the Scotland models. This is much greater than the factors applied by Tobler (60%), or Aitken (80%). Furthermore, on steeper slopes, the data suggests that it its faster to walk off-road rather than on-road. This occurs at roughly the same slope angle for both datasets (approximately 10-12 degrees when walking downhill and 17-18 degrees when walking uphill), although the difference is much more noticeable in the Scotland data. One potential reason for this is that off-road travel is more likely to have vegetation which could be used to 'cushion' each step when walking downhill, thus making walking easier. When walking uphill, this same vegetation may be crushed underfoot and be used to create artificial 'steps', thus making slipping less likely on the steep slope and enabling faster walking speeds.

As well as identifying the impact of off-road travel, we have also identified a new difference between the Scotland and ROUK datasets. Previously, both models predicted very similar walking speeds when descending a steep slope, however, we can see from Figure 4.6a that the Scotland off-road model is faster than the ROUK off-road model in these regions. Similarly, the impact of leaving the road in the Scotland data is much greater than in the ROUK data.



(a) Walking speed predictions for travelling directly up or down hills of varying slope.



(b) Walking speed predictions for traversing across hills of varying slope.

Figure 4.6: Comparison of on- and off-road walking speed models produced using data from Scotland and the rest of the UK. Also shown are the models with all data included for reference.

Although we have identified the road status (on-road or off-road) as a significant factor which should be included when determining the walking speed, this has not explained the differences found between the Scotland and ROUK models. As previously, the confidence intervals for each model show that we still have a significant difference between the model coefficients. Once again the ROUK dataset is much larger than the Scotland dataset in both the on- and off-road cases. Like we did before, 100 samples of tracks were taken from the ROUK dataset, to form a sample set of comparable size to the Scotland dataset (650 tracks for on-road and 200 tracks for off-road), and the resulting models can be seen in Figures 4.7 & 4.8.

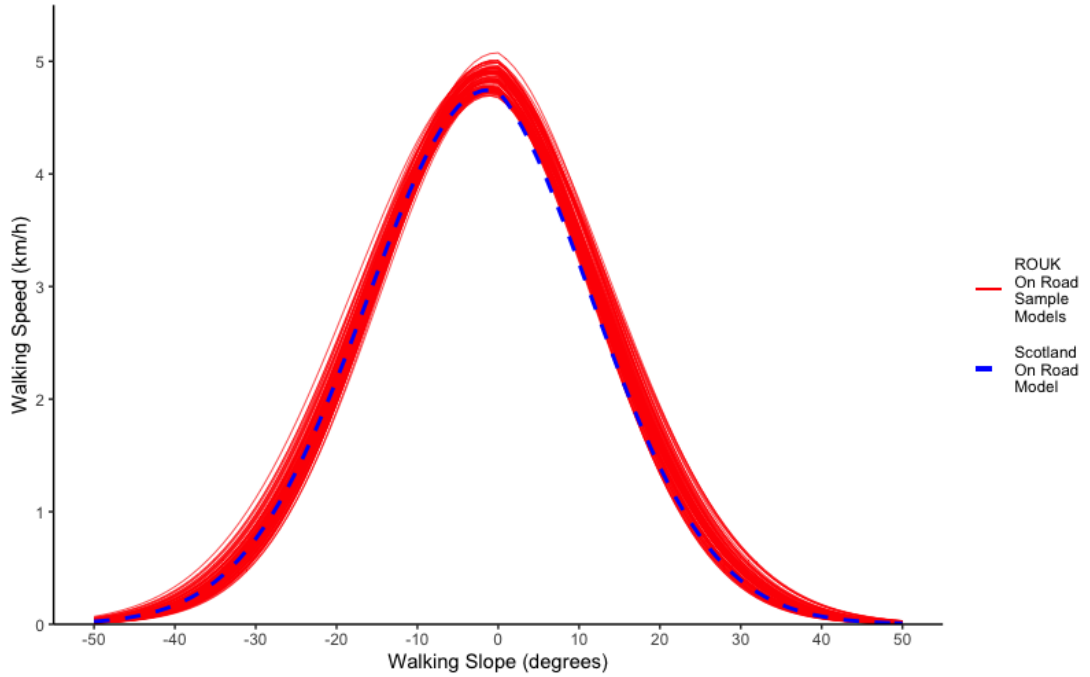
In both on- and off-road cases, the Scotland model is within the range of ROUK samples when ascending or descending the slope (Figures 4.7a, 4.8a), and is much closer to the sample range when traversing the slope than we saw previously (Figures 4.7b, 4.8b vs Figure 4.3b). This suggests that we are closer to explaining the differences between the two datasets, and it is reasonable now to conclude that the Scotland data is an extreme sample of the ROUK data.

As we continued to explain the differences between the datasets, we first focused on the on-road travel, as this made up more than 90% of our data, and came back to look at off-road travel later.

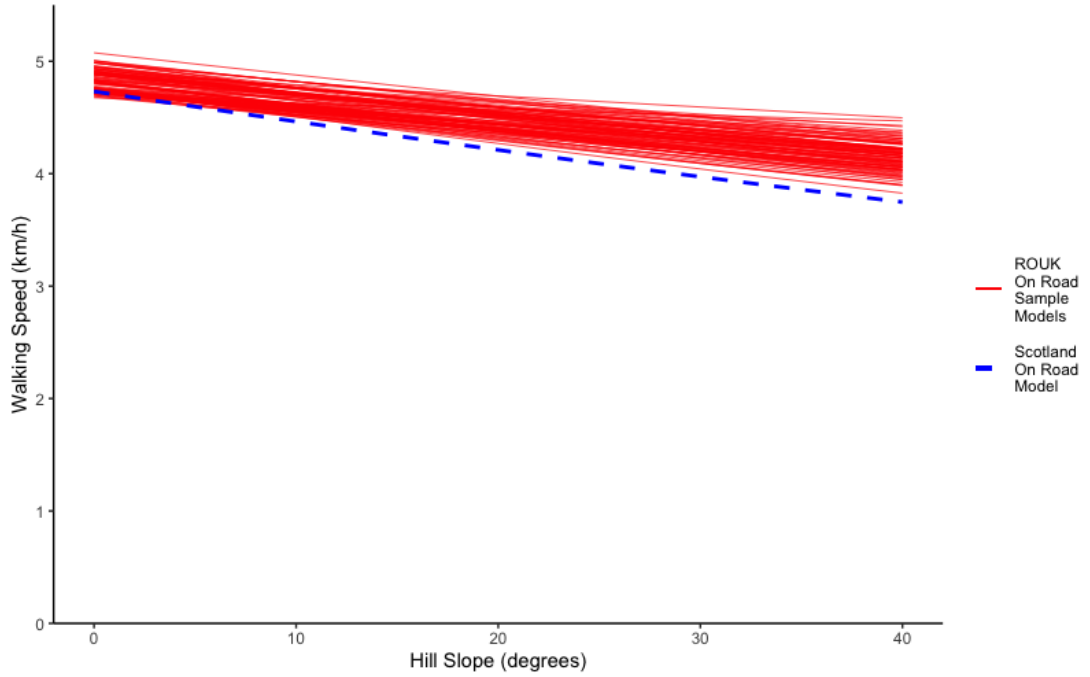
4.4.1 On-Road

Previously, we did not differentiate between the types of road or path which were being used, so this was now looked into in more detail. By consulting the descriptions of the OSM road-type definitions (OpenStreetMap Wiki, 2022b), we could separate the roads into two categories, paved and unpaved. When doing this, we assumed that the standard road type was paved. Therefore, if a route section contained multiple road types, we only considered it to be unpaved if none of the road types detected were paved. The paved road types are the following:

- Cycleway
- Footway
- Living_street
- Motorway
- Motorway_link
- Pedestrian
- Primary
- Primary_link
- Residential
- Secondary
- Secondary_link
- Service

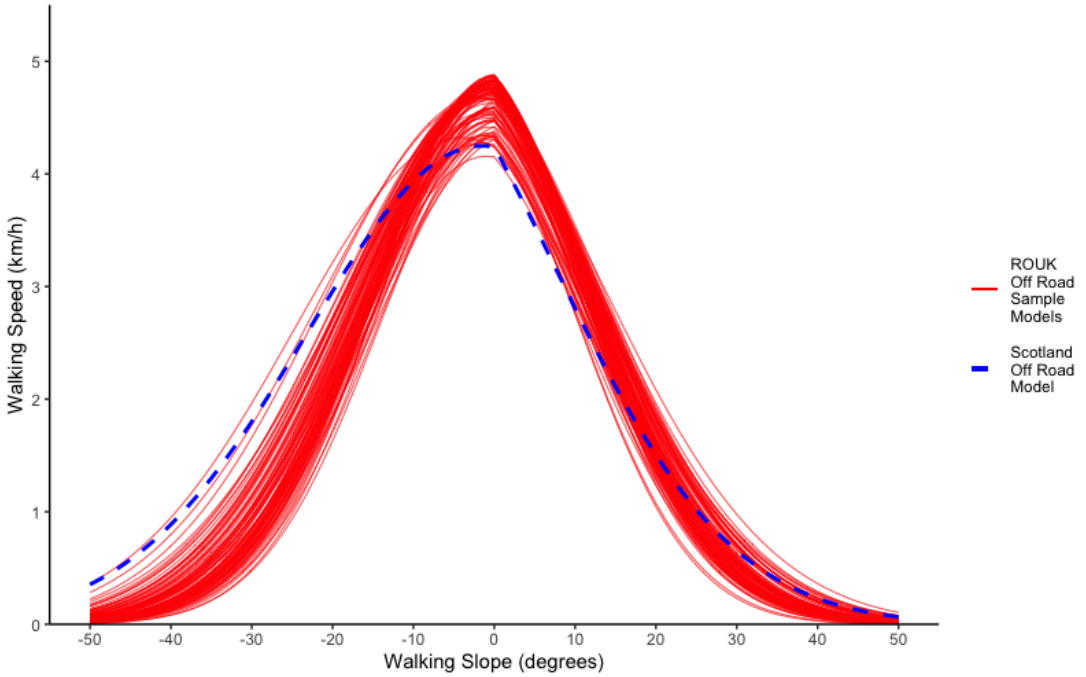


(a) Walking speed predictions for travelling directly up or down hills of varying slope.

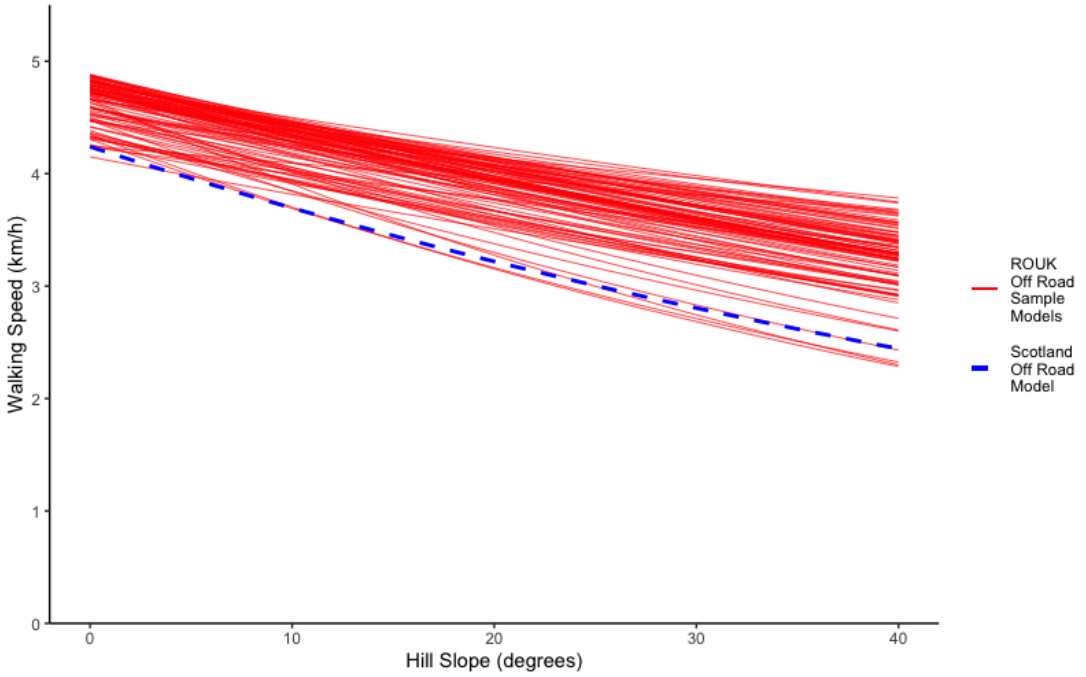


(b) Walking speed predictions for traversing across hills of varying slope.

Figure 4.7: Comparison of on-road walking speed models produced using data from Scotland against 100 sampled datasets from the rest of the UK.



(a) Walking speed predictions for travelling directly up or down hills of varying slope.



(b) Walking speed predictions for traversing across hills of varying slope.

Figure 4.8: Comparison of off-road walking speed models produced using data from Scotland against 100 sampled datasets from the rest of the UK.

- Steps
- Tertiary
- Tertiary_link
- Trunk
- Trunk_link
- Unclassified
- Unknown

While these are unpaved:

- Bridleway
- Path
- Track
- Track_grade1
- Track_grade2
- Track_grade3
- Track_grade4
- Track_grade5

Using this metric, we found that 84.9% of the on-road ROUK data was on a paved road, while only 46.5% of the on-road Scotland was on a paved road. Due to our method of considering a road to be paved by default, we were likely overestimating the amount of paved roads we considered. However, in a similar argument to our decision to use a 50 m radius to identify roads and paths, we preferred to overclassify paved roads rather than underclassify due to the relative volumes of each within our data.

In the same manner as was done with the 'OnRoad' variable, we included 'Paved' as a factor in the model, both on its own, and as an interaction term with the slope terms. Not every term was significant in both the ROUK and Scotland datasets, but in both cases we had some terms which were significant at the 95% level (adjusted for track dependence), suggesting that the road surface was important to the walking speed. Once again we chose to split the dataset up for ease of processing, so we had a paved and unpaved dataset for each region.

As expected from the significance values, the separation of paved and unpaved had much less impact in the ROUK dataset than in Scotland . However, we can see in Figure 4.9 that the paved road models are are much more similar than the models looking at all road types (Figure 4.6), and the majority of the difference in walking speeds between the two models comes from the unpaved road data. We formalised this by once again sampling the ROUK dataset such that it was of comparable size to the Scotland dataset (sample sizes of 600 tracks for paved roads and 450 for unpaved). The results of this are shown in Figures 4.10 and 4.11. We can clearly see now that our model for paved roads in the Scotland data is comfortably

within the range of samples of the ROUK data. It is reasonable to suggest, therefore, that there is no difference in walking on a paved road in Scotland compared to the rest of the UK. However, our unpaved road model for Scotland lies at the extreme edge, or outside of the range of sample models taken from the ROUK unpaved data.

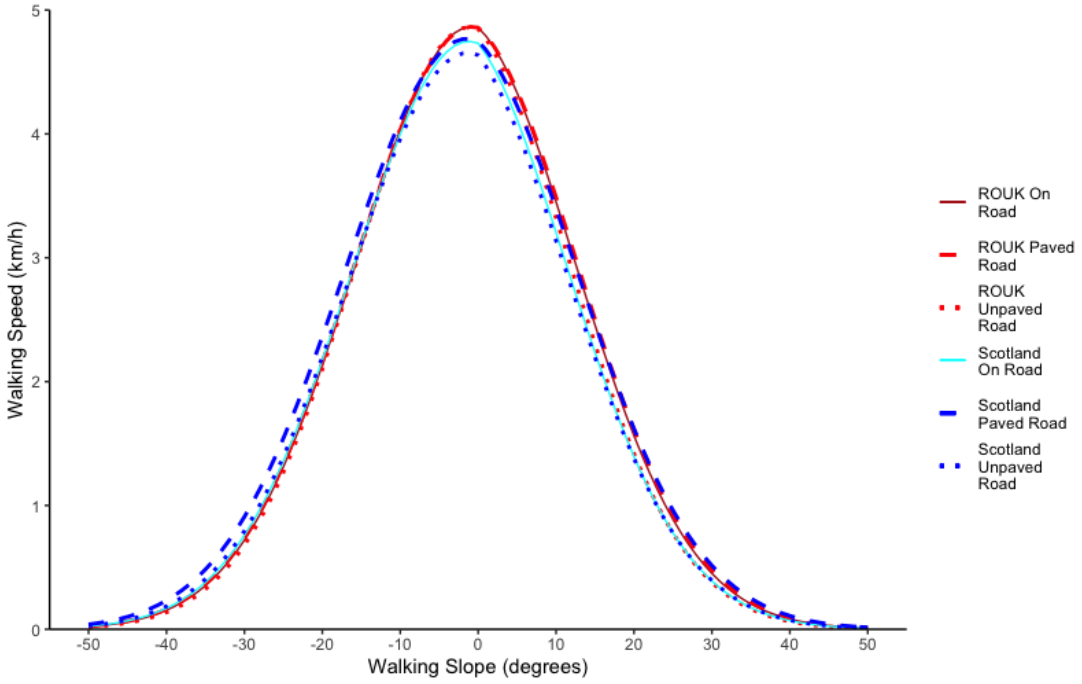
4.4.2 Unpaved and Off-Road Differences

For our unpaved road datasets (Figure 4.11), the greatest difference in walking speed predictions between the Scotland and ROUK models occurs in the -10 – 10 degree walking slope range. Similarly, in the off-road datasets (Figure 4.8), we can see that our Scotland model is at the extreme end of the ROUK sample models on walking slopes of 0 – 10 degrees (as well as slopes steeper than -20 degrees). As suggested by Proffitt et al. (1995), the -10 – 10 degree region is where most walking takes place, so it is important to have accurate speed predictions here.

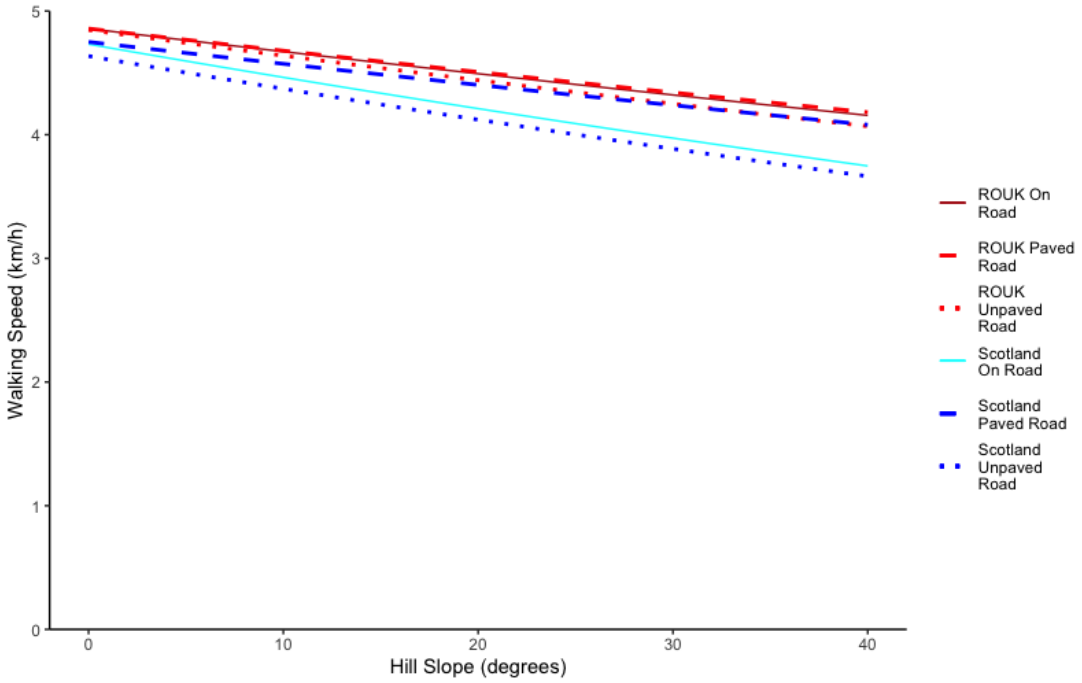
While we know that terrain obstruction may also affect walking speeds, this would only affect walking speeds in the off-road areas of routes. Prior to exploring this, we wanted to see if we could find another variable which would account for the lower walking speeds seen in Scotland on both unpaved roads and when off-road.

We have previously discussed the fact there are many variables which will affect walking speeds such as age, group size and composition, or walk motivation. However, we did not have any reason to suggest that there would be a large difference in most of these between the two regions. The main factor which we believed may be different between Scotland and the rest of the UK (and predominantly affect travel on unpaved roads or when off-road) would be the weather, with more extreme weather often found in Scotland. Instead of modelling the weather directly (i.e. by checking historical weather data for each track), we instead attempted to look at exposure. A high exposure area may be on the edge of a cliff, or a narrow ridge. These regions can have higher wind speeds, and hikers may walk more slowly, or exhibit more caution because of this. Higher exposure can also cause more anxiety or fear, leading to slower speeds. Furthermore, high exposure regions are less likely to occur on a paved road, and are more likely in unpaved mountain paths, or off-path regions; where we had the greatest difference between modelled speeds. If the exposure impacts walking speed, then we would expect this to be worse in Scotland, due to the more inclement weather.

In order to capture the exposure fully, we would have to model the terrain globally around each point, and come up with some measure to determine an exposure value. However, a simpler method which we could use for an initial investigation was to look at the elevation value of a point. Although not all high-altitude places will have high exposure, and vice-versa, it was felt that in general, higher elevations will have higher exposure.

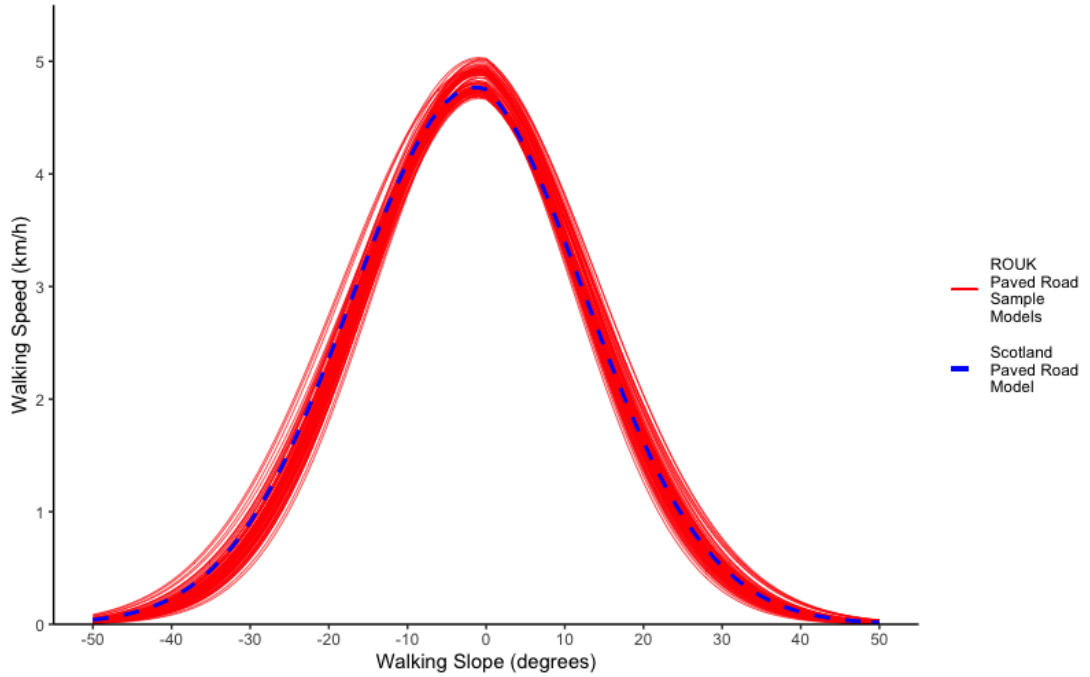


(a) Walking speed predictions for travelling directly up or down hills of varying slope.

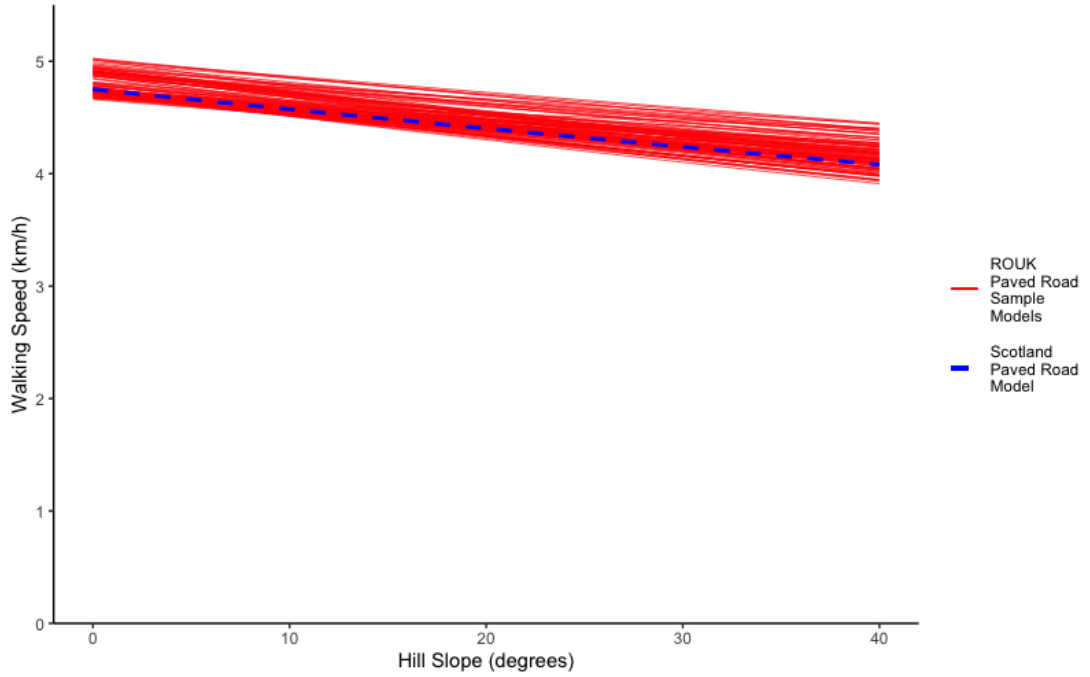


(b) Walking speed predictions for traversing across hills of varying slope.

Figure 4.9: Comparison of walking speed models on paved roads and unpaved roads produced using data from Scotland and the rest of the UK. Also shown are the models with all on-road data for reference.

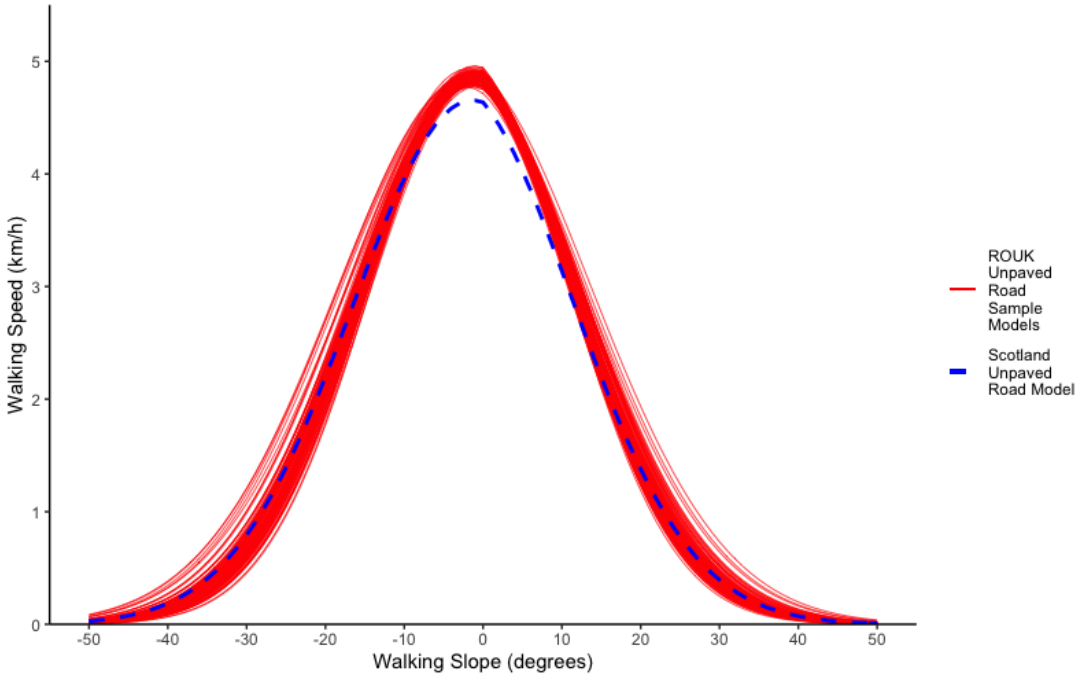


(a) Walking speed predictions for travelling directly up or down hills of varying slope.

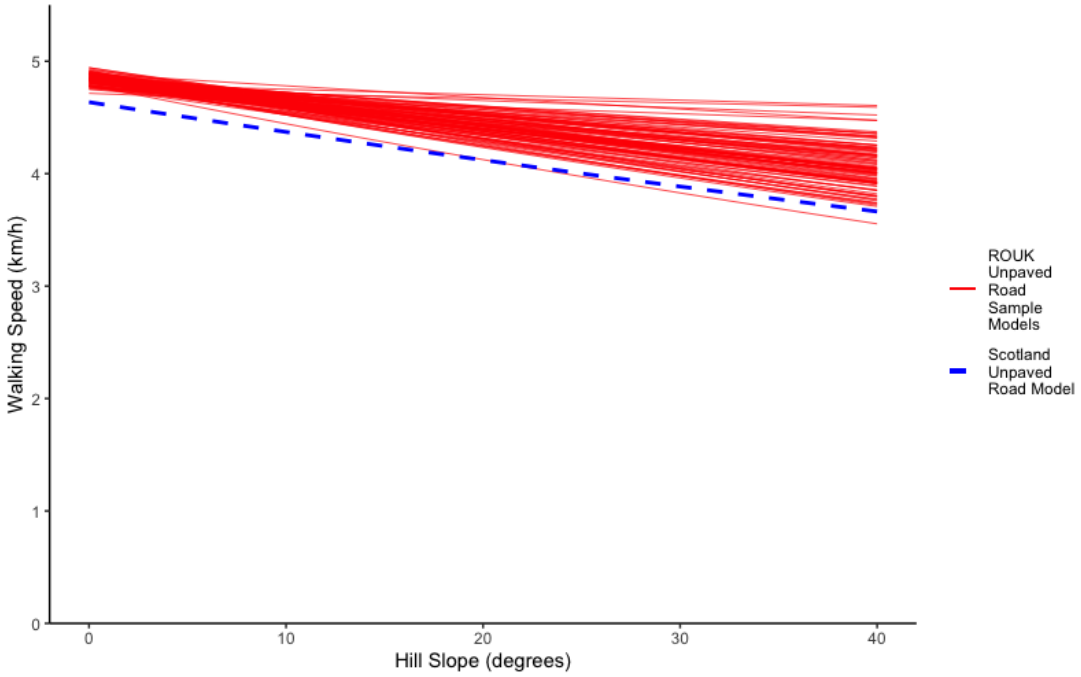


(b) Walking speed predictions for traversing across hills of varying slope.

Figure 4.10: Comparison of paved road walking speed models produced using data from Scotland against 100 sampled datasets from the rest of the UK.



(a) Walking speed predictions for travelling directly up or down hills of varying slope.



(b) Walking speed predictions for traversing across hills of varying slope.

Figure 4.11: Comparison of unpaved road walking speed models produced using data from Scotland against 100 sampled datasets from the rest of the UK.

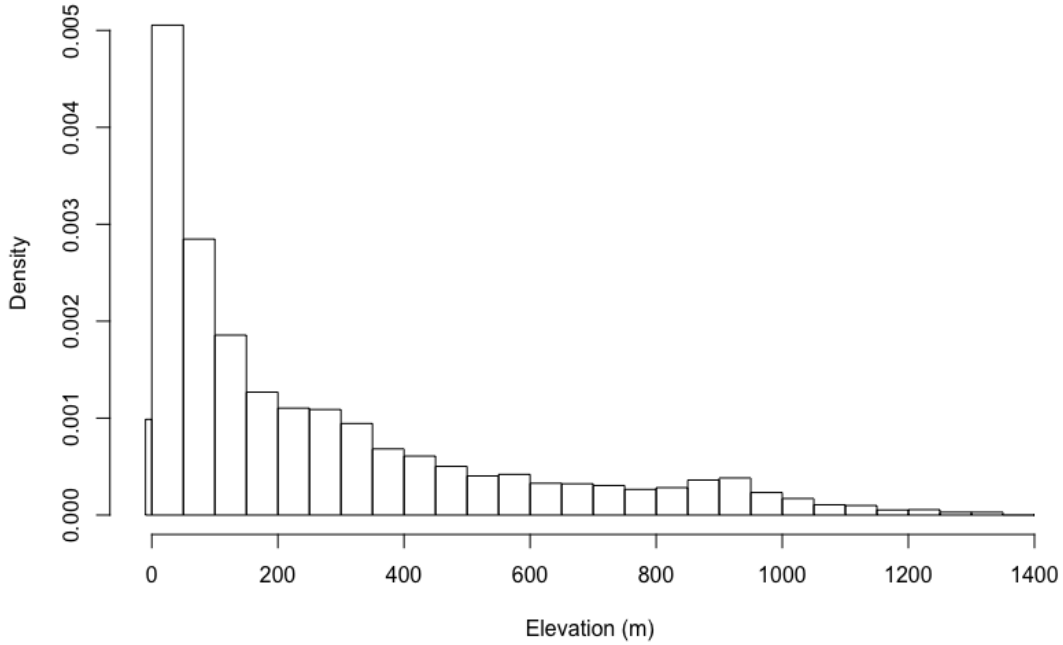
When we compared the data structures of each region, we could see that there was a much greater proportion of data at high elevation (>500 m) in the Scotland dataset than the ROUK dataset (Figure 4.12). Furthermore, this difference was heavily skewed towards the regions where we have the most difference between the walking speed models. In both of the paved road datasets, the proportion of data above 500 m was very low (1.4% and 4.1% for ROUK and Scotland respectively, Figure 4.13). However in the unpaved datasets, a much greater proportion of the data occurred at high elevation in Scotland (4.9% and 27.8% for ROUK and Scotland respectively, Figure 4.14). Similarly, our Scotland off-road dataset had a much greater proportion of points above 500 m than seen in the ROUK dataset (43.2% compared to 2.8%, Figure 4.15).

Figures 4.14 and 4.15 show us that we had a large difference in the characteristics of the two datasets in the areas where we saw the greatest difference between the models. For this reason we included elevation as a model variable, both as a continuous variable or as a factor variable classifying all points as either high elevation or low elevation (where high elevation consisted of all data >500 m). However, in both cases we found this to not be a significant factor in the model.

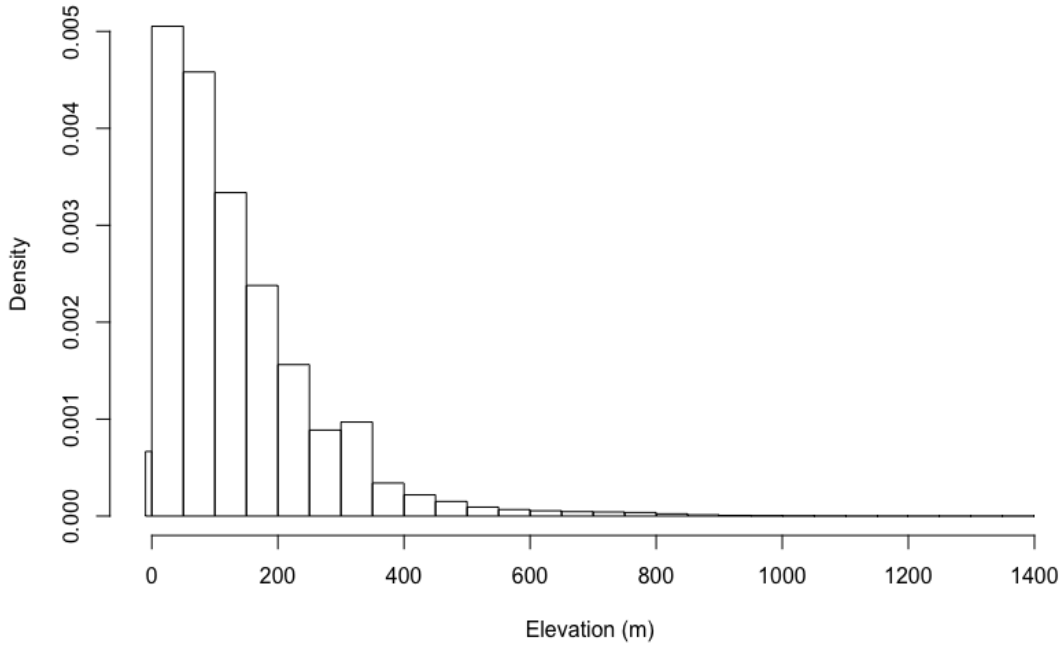
We conclude that the three separate road distinctions (paved, unpaved, off-road) we made are all significant. Based on the sample data taken, we suggest that the model formulated using the Scotland data is an extreme sample of the ROUK data, where a greater-than-average portion of the data has been sampled from high elevation regions. However, the high elevations themselves are not the cause of the difference between the model coefficients. Further investigation is required into the difference, but we could speculate that it is likely something which is connected to, but not caused by, the elevation differences. For example, snow or other bad weather will affect high-elevation areas more than low-elevation ones, and non-paved or off-road areas more than paved roads.

4.5 Terrain Obstruction

Despite not having suitable lidar coverage in Scotland (see Section 2.2.4), we could explore the impact of terrain obstruction for the data from the rest of the UK. As introduced in Section 1.1, we suggested that terrain was a missing factor that could be used to improve walking speed precision for off-road travel, and used lidar data to test this. We had access to lidar data at 2 m resolution covering large areas of England and Wales, but the coverage was not complete. Of our off-road data ($\sim$2,900 km, spread across over 1,200 tracks), over 2,000 km had lidar data available. Before checking the impact of the terrain obstruction on walking speed, we wanted to check that there was not a systematic difference between the walking speeds in regions where we had lidar data, and regions where we did not. If the two regions were not found to be different, then any findings about the effects of terrain obstruction in

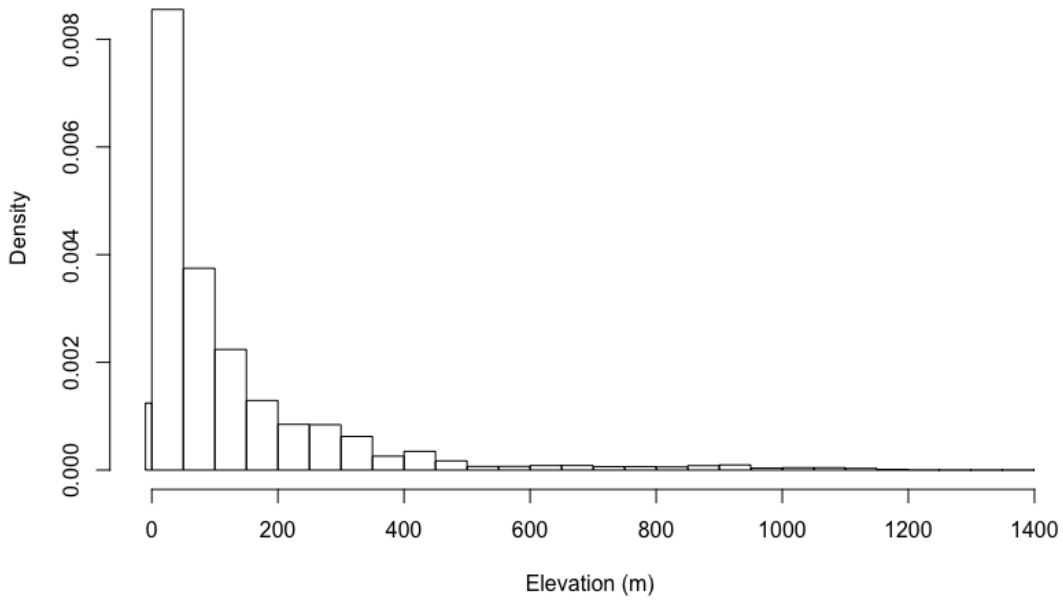


(a) Scotland dataset elevations

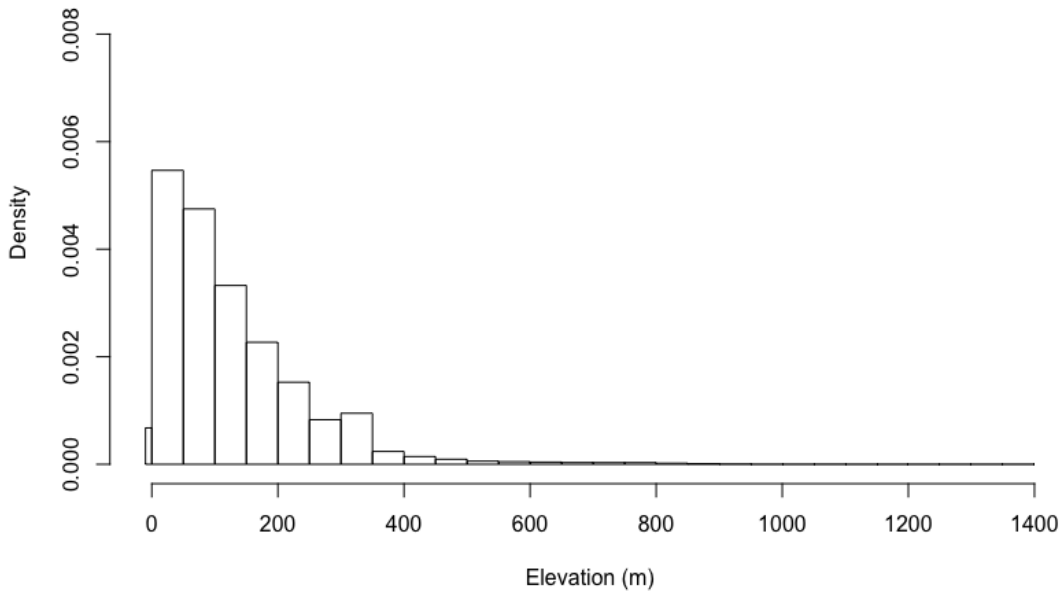


(b) ROUK dataset elevations

Figure 4.12: Comparing elevations of tracks between Scotland and the rest of the UK.

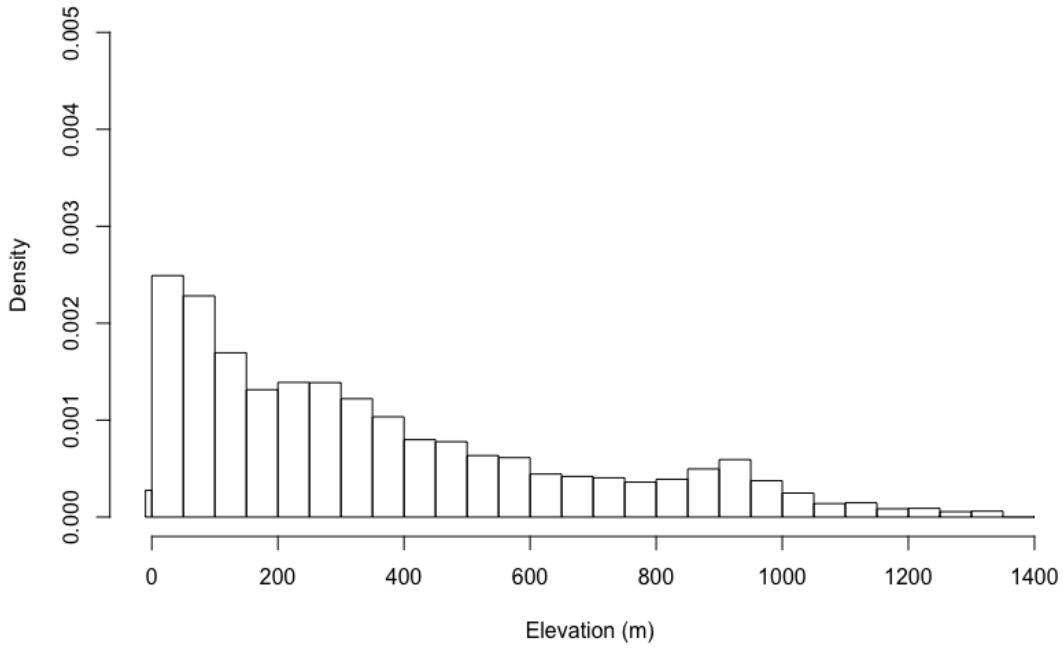


(a) Scotland paved road dataset elevations

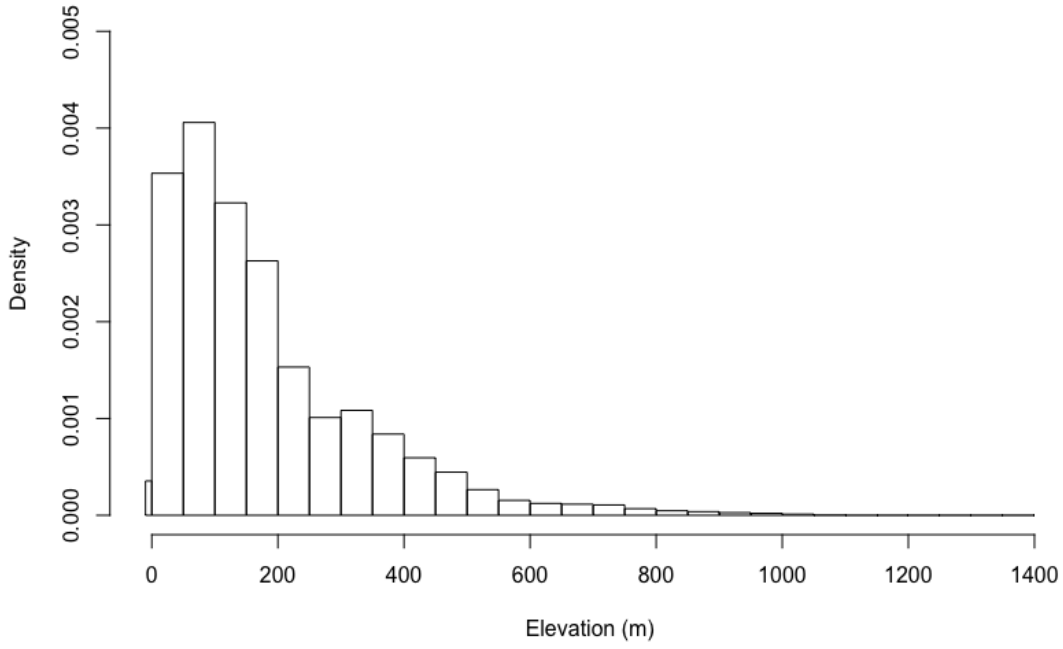


(b) ROUK paved road dataset elevations

Figure 4.13: Comparing elevations of paved road track sections between Scotland and the rest of the UK.

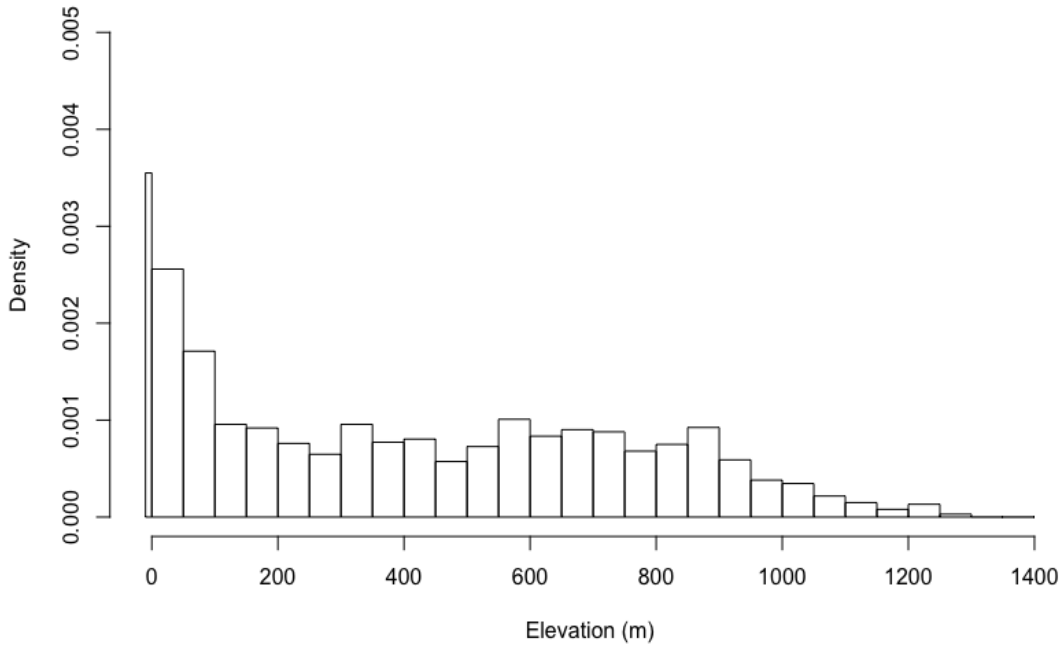


(a) Scotland unpaved road dataset elevations

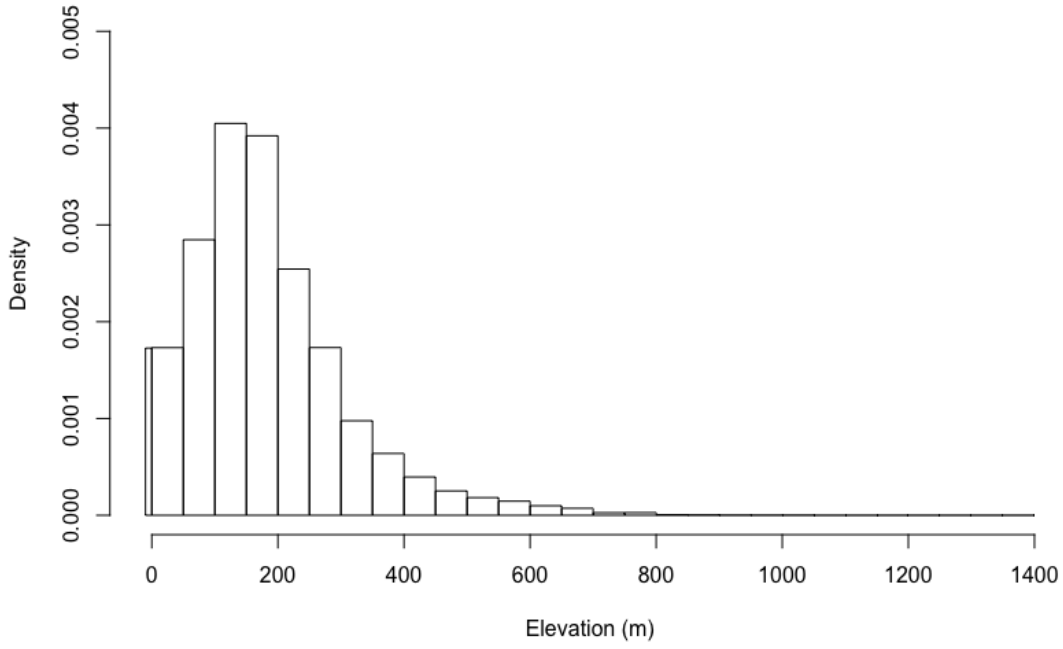


(b) ROUK unpaved road dataset elevations

Figure 4.14: Comparing elevations of unpaved road track sections between Scotland and the rest of the UK.



(a) Scotland off-road dataset elevations



(b) ROUK off-road dataset elevations

Figure 4.15: Comparing elevations of off-road track sections between Scotland and the rest of the UK.

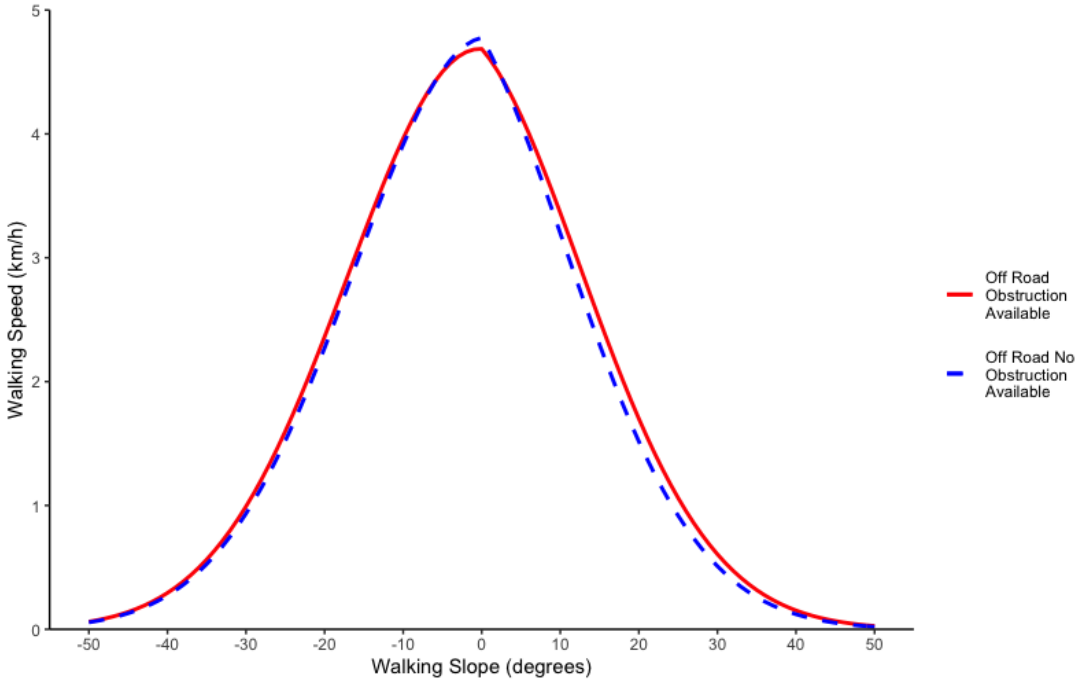
regions where we had lidar data could also be applied to areas where we didn't have the data. As we only had obstruction data available for areas within the ROUK dataset, when comparing the two cases we did not include the Scotland data. We concluded in the previous section that the Scotland dataset is a plausible, but extreme, subset of the ROUK data so it would not be valid to include it in a comparison without also including a similarly extreme dataset where lidar data is available.

When modelling the data for the separate datasets ('obstruction available' vs 'no obstruction available'), we see that the models are very similar when ascending or descending a slope (Figure 4.16a). This was not the case when traversing the slope however, as the 'no obstruction available' model predicts that hill slope has a greater impact on reducing walking speed than the 'obstruction available' data model (Figure 4.16b).

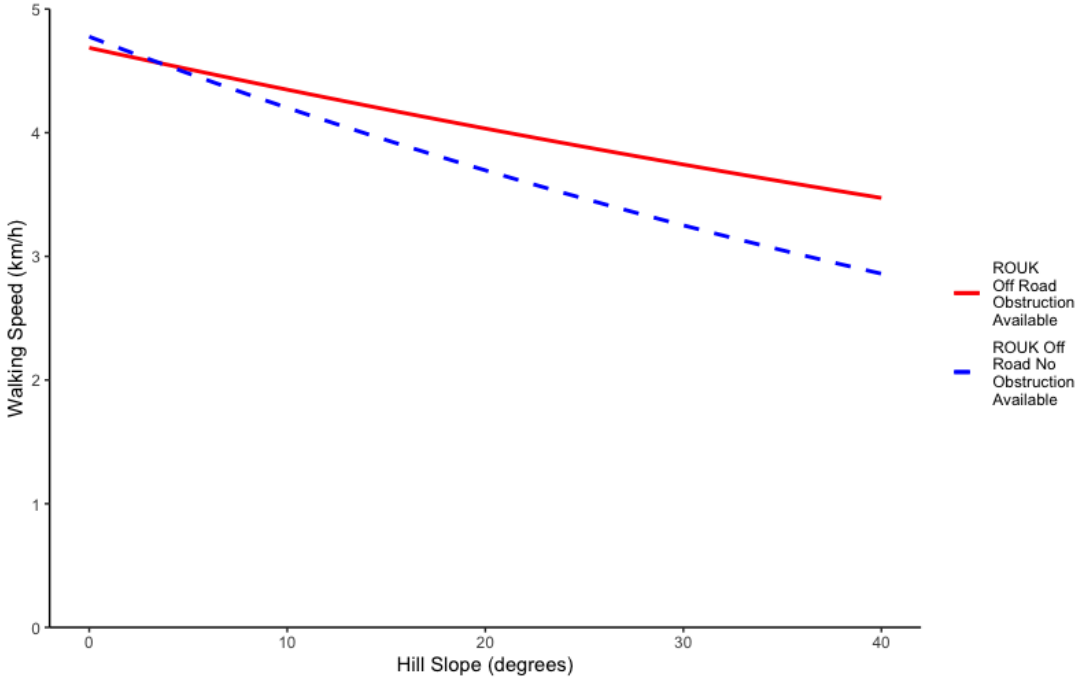
As we did when exploring the impact of roads and paths, we sampled our larger ('obstruction available') dataset, so that we had a similar number of tracks as in our smaller dataset, and compared models made from more equal volumes of data. When doing this (Figure 4.17), we found that the 'no obstruction available' model is within the range of sample models for traversing the slope, albeit at an extreme end. This is likely due to the low volume of data which we had at high hill-slopes. (Only 50 km of data had a hill slope greater than 15 degrees with no lidar data available, and only 130 km with lidar data available). Going forward, we assumed that the regions where we had lidar data were representative of all off-road regions, and so any findings could be applied to both areas.

To begin exploring the effects of terrain obstruction, we first looked at the range of speeds across the different obstruction values. The data was split into 25 quantiles, and the average walking speed for each was calculated. The results are shown in Figure 4.18a. This shows us two things; firstly the vast majority of our data had very little, or no obstruction (as most of the quantile points occur below 0.5 m of obstruction). Secondly we can see that there is a very steep drop off in walking speed initially, and it then remains relatively constant across obstruction levels. Our initial assumption was that walking would be relatively easy with no, or very little obstruction, and then much slower at obstruction values of approximately 0.5 m - 4 m when it would involve walking through thick vegetation, before getting slightly faster again at higher obstruction values (as you would be walking through a forest and could walk between the trees below the canopy). The data shows this not to be the case, although this may be a result of our data only showing us regions where walking was possible. Due to the crowdsourced nature of our GPS tracks, we had no data showing us the walking speed when in 2 m of thick vegetation, as it is very unlikely that people would have chosen to walk there.

Figure 4.18b shows a close-up of the steep speed drop off, and we can see that the average speed dropped from approximately 4.8 km/h when there was no obstruction down to about 4 km/h once there was more than 10 cm of obstruction. We used this information to classify all points into heavy obstruction (>10 cm) or light obstruction (<=10 cm). Although the figure

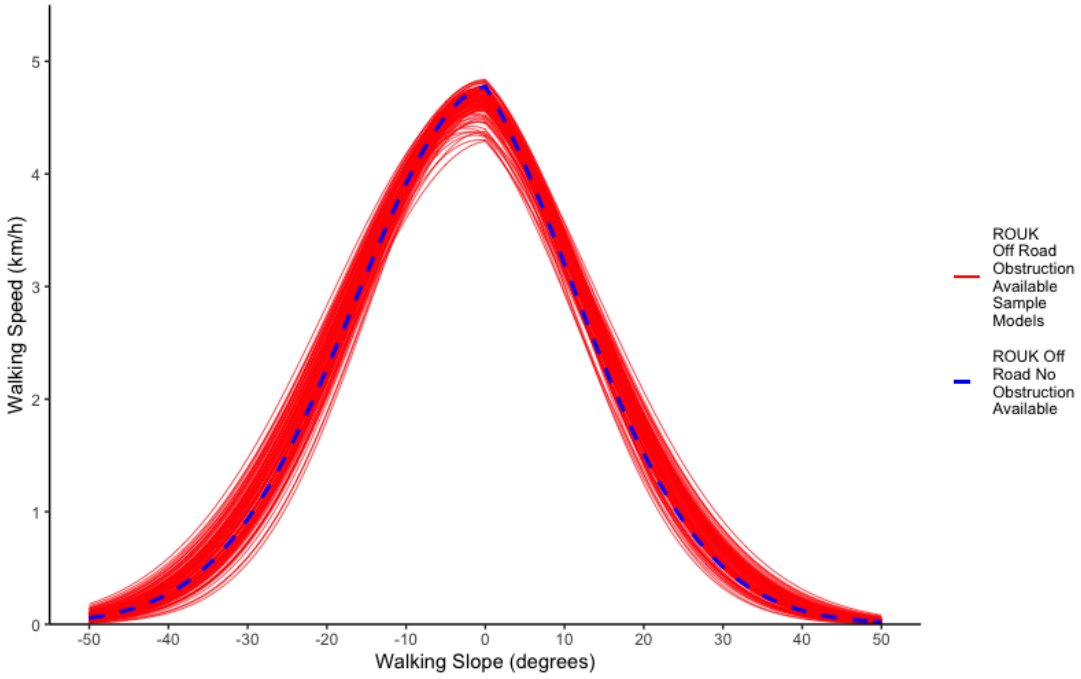


(a) Walking speed predictions for travelling directly up or down hills of varying slope.

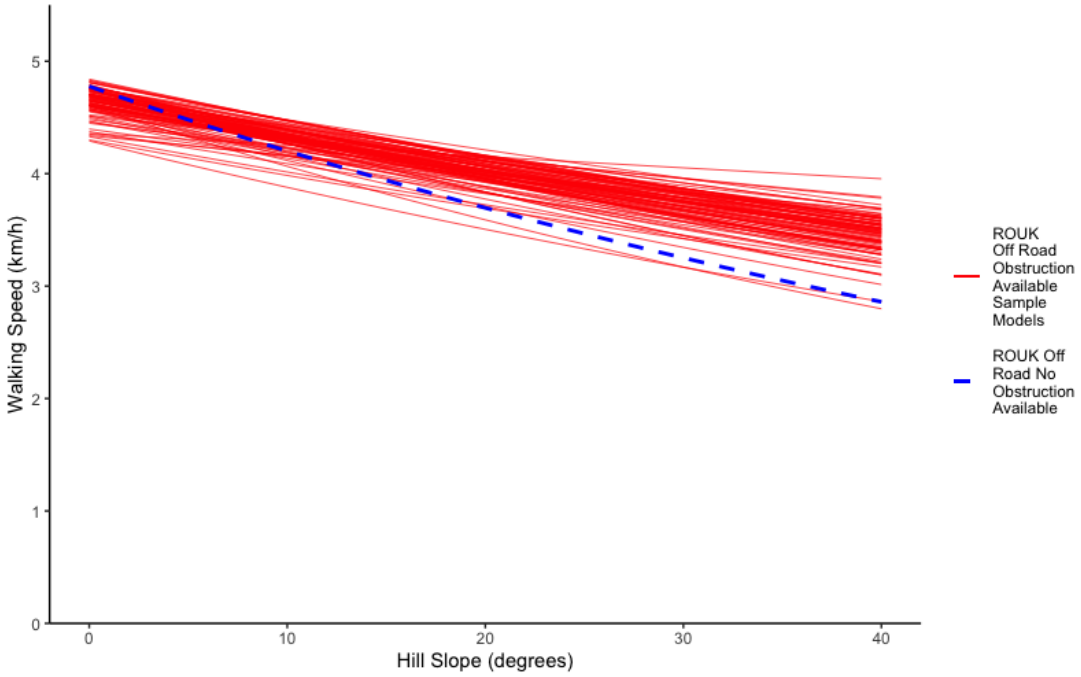


(b) Walking speed predictions for traversing across hills of varying slope.

Figure 4.16: Comparison of off-road walking speed models where obstruction data is, or is not, available.



(a) Walking speed predictions for travelling directly up or down hills of varying slope.



(b) Walking speed predictions for traversing across hills of varying slope.

Figure 4.17: Comparison of off-road walking speed models produced using a dataset where obstruction data isn't available against 100 sampled datasets where obstruction data is available.

	a	b	c	d
Paved Road	1.580	-0.00389	-0.00726	-0.00218
Unpaved Road	1.580	-0.00389	-0.00965	-0.00248
Off Road (Obstruction Unknown)	1.536	-0.00731	-0.00965	-0.00187
Off Road (Light Obstruction)	1.580	-0.00731	-0.00965	-0.00187
Off Road (Heavy Obstruction)	1.443	-0.00731	-0.00965	-0.00187

Table 4.1: Final model variable coefficients using the ROUK dataset.

suggests a gradual decrease in walking speed between 0 and 10 cm of obstruction, we chose not to model this. Vegetation length is highly variable throughout the year, and it is more practical to classify regions as light or heavy obstruction when discussing walking speeds. The obstruction classification was added to the walking speed model as a factor variable, and the results can be seen in Figure 4.19. The obstruction level was found to be highly significant in the model, and adding a small amount (10 cm) of obstruction to off-road terrain can reduce the walking speeds by up to 0.6 km/h when walking on flat ground.

4.6 Final Model

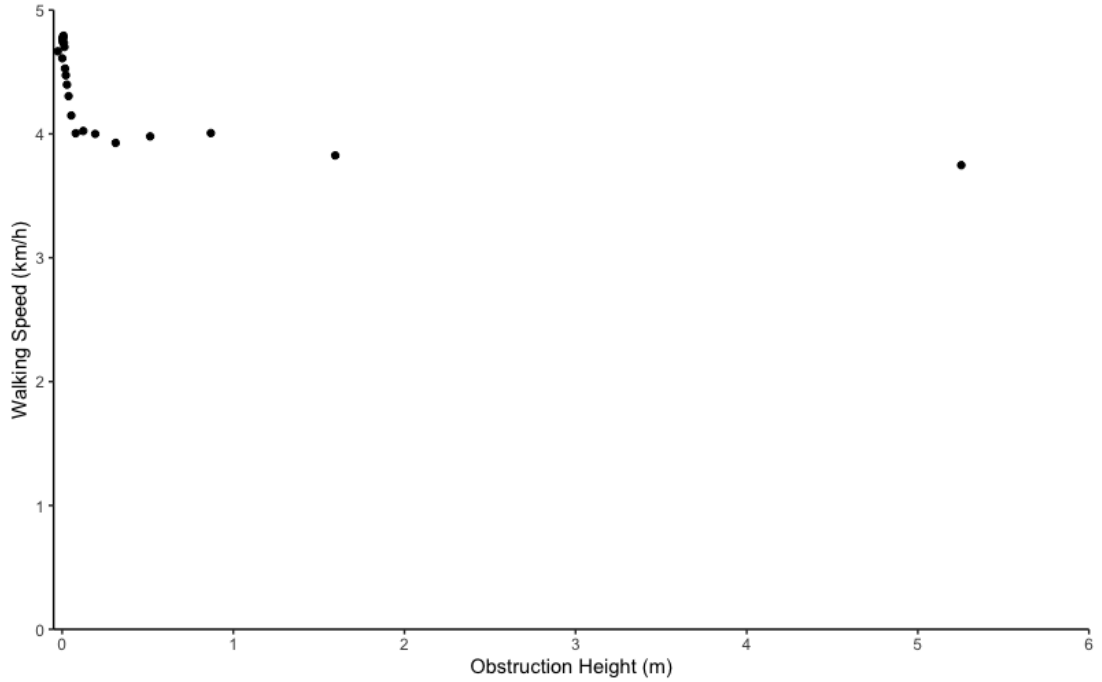
We were now able to build a full model for walking speeds which took into account all three of the variables suggested by Arnet (2009). The basic model formulation in general was the same as that used in Chapter 3:

$$v = exp(a + b\phi + c\theta + d\theta^2) \tag{4.1}$$

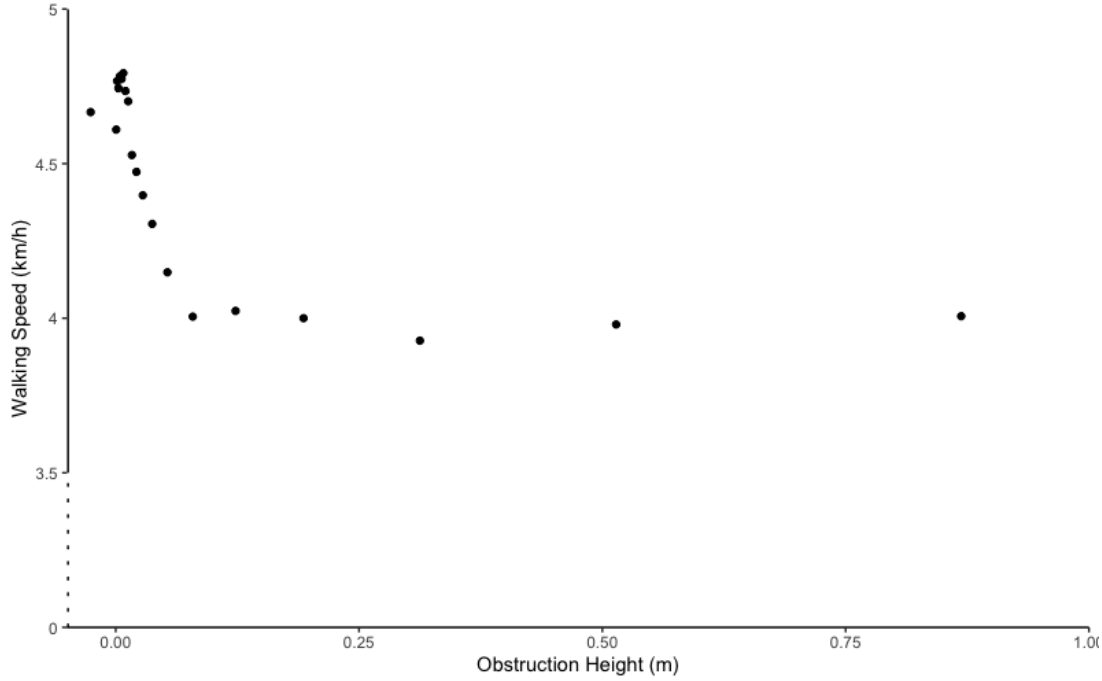
where

$v =$ walking speed (km/h)
$\phi =$ hill slope angle (degrees)
$\theta =$ walking slope angle (degrees)

Unlike the model found in Chapter 3, we now considered the terrain types as factor variables and interaction terms. As mentioned in Section 4.4.1 when looking at the road conditions, not all additional terrain factors had a significant effect on all variables. We therefore created a model with all possible terms, and removed terms one at a time (in order of least significance) until all remaining terms were significant at the 95% level. The final values for a,b,c and d (for the ROUK dataset) are given in Table 4.1. (Scotland model values and combined dataset model values can be seen in Appendix A). The critical gradient for this model is between 14 – 16 degrees when walking uphill and -16 – -18 degrees when walking downhill (depending on road and obstruction conditions).

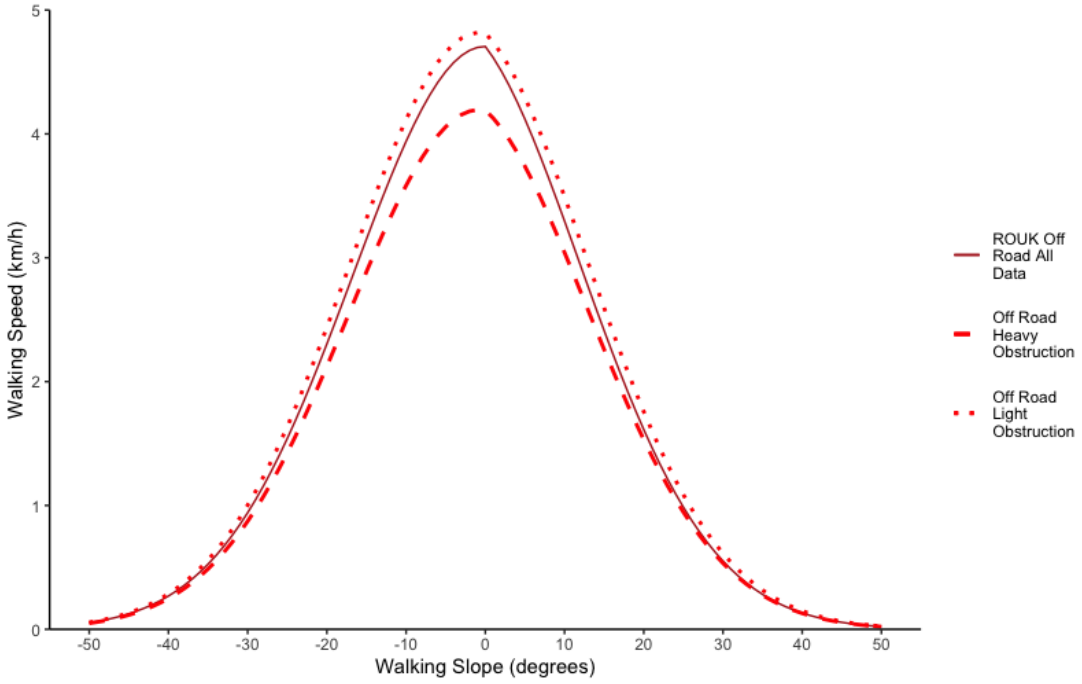


(a) Full range.

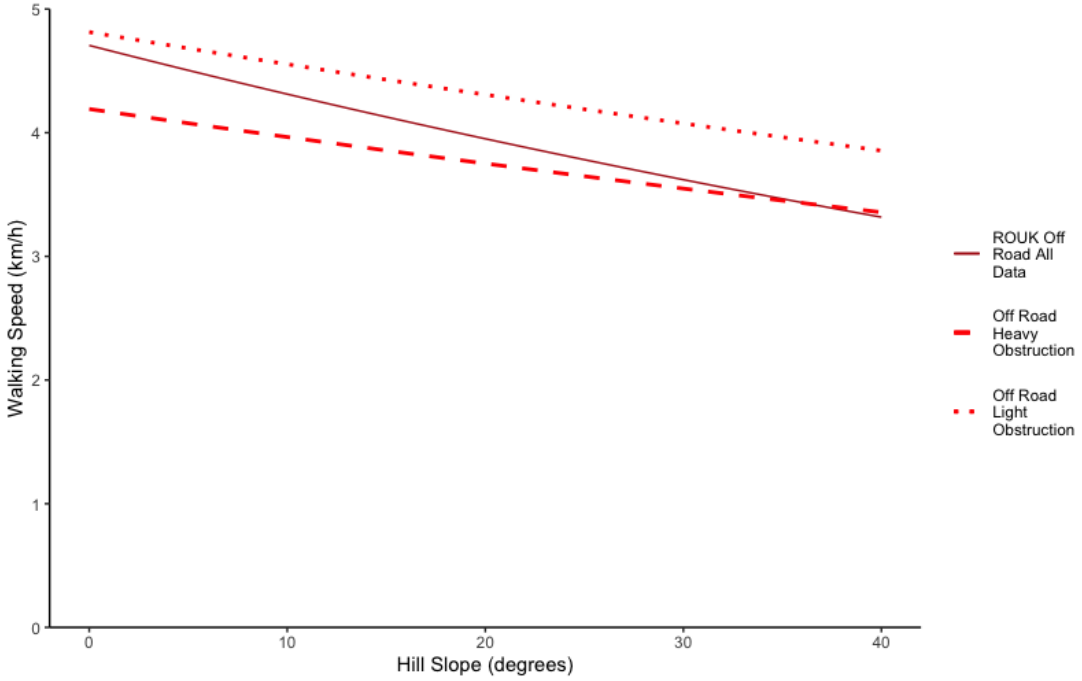


(b) Zoomed range.

Figure 4.18: Binned average walking speeds across different levels of obstruction. Each bin contains 1/25th of the datapoints



(a) Walking speed predictions for travelling directly up or down hills of varying slope.



(b) Walking speed predictions for traversing across hills of varying slope.

Figure 4.19: Comparison of walking speed models under heavy or light terrain obstruction. Also shown is the model using all off-road data for reference.

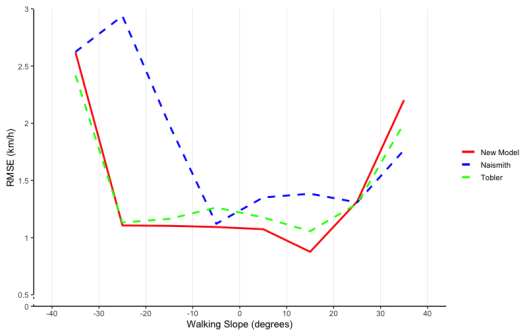
	New Model	Naismith	Tobler
Average % error	23.68	26.36	26.17
MSE	1.20	1.61	1.53
RMSE	1.10	1.27	1.24
R^2	0.09	-0.22	-0.16

Table 4.2: Comparison of new model against existing methods to calculate walking speeds.

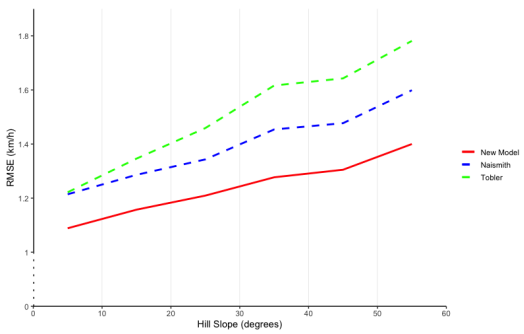
As we did with the model found for Scotland in Chapter 3, we compared the predictions of this model against those of Naismith's and Tobler's models, as shown in Table 4.2. Firstly, the predicted speeds for individual 50 m sections had a lower RMSE and percentage error, and a higher R squared value in this model than in other models. However, this did not translate to similar results when looking at predicted walking times (for individual 50 m sections). While the average percentage error for predicted time was lower in the new model than existing ones, the RMSE value was substantially higher (103 s vs 22 s). Investigation into the most extreme error values showed us that this was caused by errors in the data, rather than problems with the model. The most extreme difference between predicted walking time and actual walking time for a single 50 m section was over 33.5 hours. Upon inspection we found that this was caused by an error with the OS elevation DTM. A single 5 km x 5 km tile received from Ordnance Survey contained elevation data which did not match up with neighbouring tiles, leading to apparent steep slopes on the tile borders. We confirmed this data was incorrect by comparing elevation values against a paper OS map of the region. No other instances of this were found in the data, and less than 0.5% of the GPS track segments intersected the affected tile, so we do not believe that the overall walking speed model will have been detrimentally affected. In order to account for these (or other similar) data errors when calculating the RMSE values for predicted walk times individual 50 m sections, we adjusted our calculations to only look at the middle 99.9% of the values for each model. This reduced the RMSE value of the new model to 19.5 s, lower than for both Naismith's rule and Tobler's function.

As we found in Chapter 3 there was very little difference between the new and existing models, when looking at the time estimates for routes as a whole. Our model has an average percentage error of 14.1% compared to 13.6% for Naismith's rule and 15.5% for Tobler's hiking function. These equate to approximately 8-9 minutes of error per hour of walking. Naismith's rule is still a good method to estimate walking time for a hike as a whole (as speed over-estimations on descents and under-estimations on ascents 'cancel each other out'), but particularly steep or difficult sections will likely not be estimated correctly, and time estimates for individual sections of a route will be less accurate than using the new model found here.

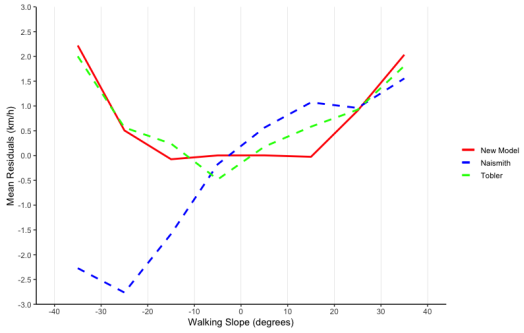
Figure 4.20 shows the RMSE and mean residuals for each of the models, looking only at data which was within 5 degrees of directly climbing (a,c) or traversing (b,d) hills of varying slope. There are some interesting points to note here. Firstly, Naismith's rule consistently overestimates walking speeds when descending a slope, and underestimates speeds when



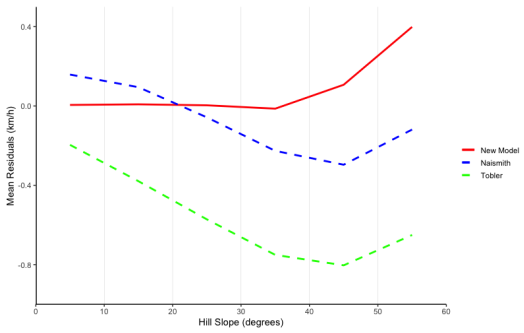
(a) RMSE values when travelling directly up or down hills of varying slope.



(b) RMSE values when traversing across hills of varying slope.



(c) Mean residual values when travelling directly up or down hills of varying slope.



(d) Mean residual values when traversing across hills of varying slope.

Figure 4.20: Comparing RMSE and mean residual values for the new model (red), Naismith's rule (blue) and Tobler's function (green).

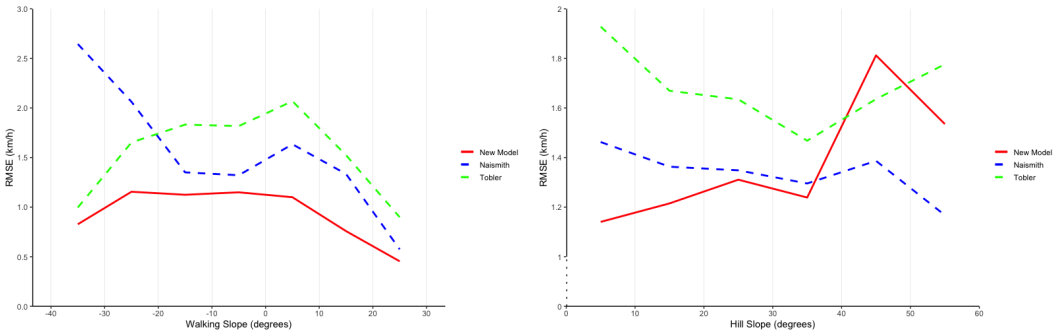
climbing a slope. This backs up our previous assertion that the errors in predicting route times as a whole were reduced as a result of the errors throughout the route 'cancelling each other out'. When ascending or descending a slope, the RMSE of our GLM is similar to that of Tobler's hiking function. However, one of the main areas where we see an improvement using our model is on slight declines. Tobler's hiking function suggests that walking speed increases on mild descents up to a maximum of 6 km/h (seen in Figure 1.1). It is clear from Figure 4.20c, that Tobler's function overestimates the walking speed in this region. We know from existing research that most walking takes place on low walking slopes, and this is evidenced by our data ($\sim$98% of our data was from walking slopes of under 10 degrees). The improved walking speed predictions of our model in this region therefore have the greatest impact in real-world situations. We also see an improvement in RMSE when using our model to predict speeds for hill traversals (Figure 4.20b). We can note from Figure 4.20d that both Naismith's rule and Tobler's hiking function consistently overestimate the walking speed when traversing a slope, as they do not take into account the impact that the hill slope has on reducing walking speeds. We do see that the average error in our model increases as the hill slope increases, but we believe that this is due to limited volumes of data at high hill slopes ($\sim$0.5% of our data occurs on hill slopes steeper than 40 degrees).

As well as looking at the overall performance of our new model, we looked to explore how well our model performed in off-road conditions, compared the existing functions. Note that when doing this we compared our predicted speeds to those from Naismiths's rule with Aitken's correction applied (a reduced base speed of 4 km/h), and Tobler's function with the off-road multiplicative factor of 0.6 (both discussed in Section 1.1). Figure 4.21 shows the RMSE and mean residuals for each of the models in off-road conditions when climbing (a,c) or traversing (b,d) hills of varying slope. From Figures 4.21a and 4.21c it is clear that Tobler's hiking function consistently underestimates the walking speed when off-road. The factor of 0.6 is a larger reduction in walking speed than is observed in practice. As we found when looking at our data as a whole, Naismith's rule underestimates the walking speed when climbing a slope and overestimates when descending a slope. Our new model does not suffer from these problems, with both a lower RMSE and lower absolute mean residual value across all walking slopes. Both of the existing models also consistently underestimate walking speeds when traversing a slope, unlike our new model which has a mean residual of less than 0.4 km/h on slopes of up to 35 degrees. The error in predictions of our new model does increase as the hill slope increases, though the RMSE is generally lower than seen in the existing models. On the steepest hill slopes our model appears to perform less well than the existing ones, though only 0.2% of our off-road data occurred on a hill slope steeper than 40 degrees.

Although the off-road data made up only a small portion of our total dataset, improvements to off-road walking speed predictions are very important. In an emergency scenario people want to navigate the fastest route to safety, whether that is following a path or cutting cross-country. Improvements to the off-road walking speeds enable more accurate calculation of what the fastest escape route may be.

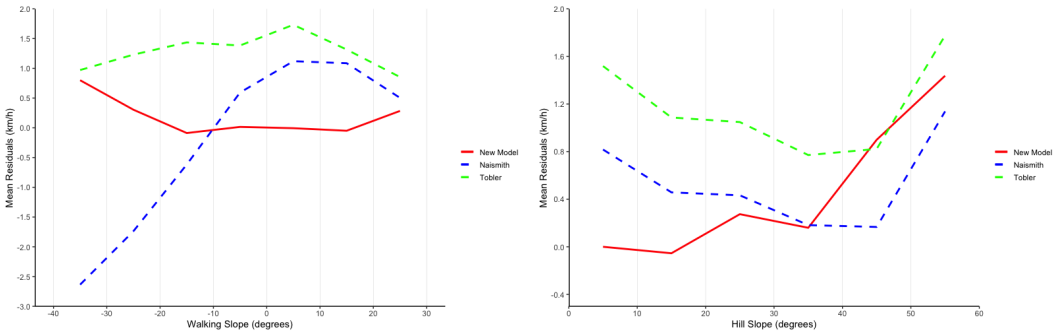
4.7 Further Investigations

We were able to use the dataset which we used to build our walking speed model to explore other characteristics of walking routes. These included how the walking speed and amount of time spent on breaks changes depending on the length of the route. We also looked to find a relationship between the OS Terrain Type data, and the lidar obstruction data (both described in Section 4.3.2), which may allow us to estimate the terrain obstruction level in regions where the lidar data is unavailable.



(a) RMSE values when travelling directly up or down hills of varying slope.

(b) RMSE values when traversing across hills of varying slope.



(c) Mean residual values when travelling directly up or down hills of varying slope.

(d) Mean residual values when traversing across hills of varying slope.

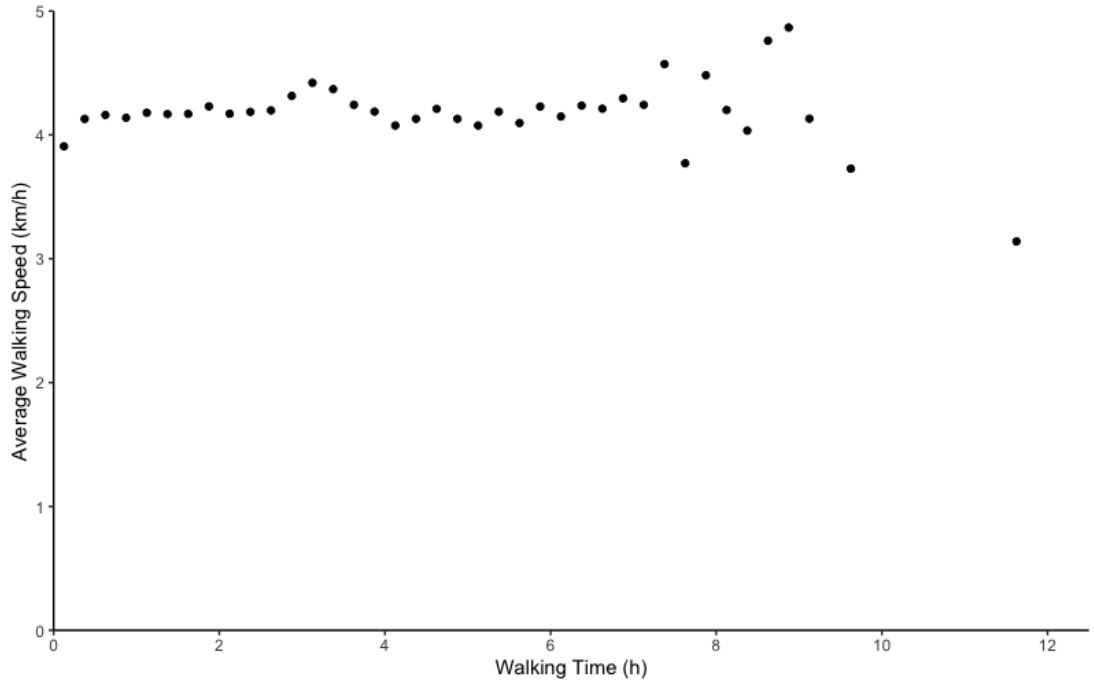
Figure 4.21: Comparing RMSE and mean residual values for the new model (red), Naismith's model (blue) and Tobler's function (green) in off-road conditions.

4.7.1 Walking Speed Variance by Hike Length

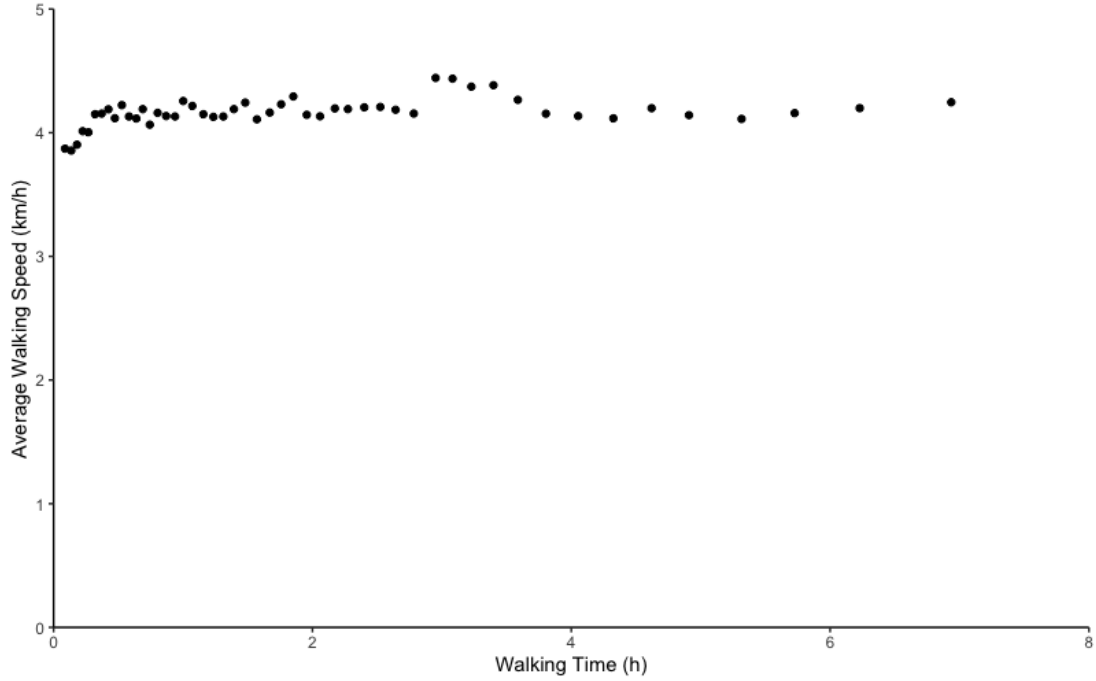
Pitman et al. (2012) proposed that the walking speed changes depending on the length of a hike. While exploring this, we were concerned with the total hike length in a given day, and so we considered all tracks or segments within a file as a single group. This allowed us to account for tracks where a new track or track segment was created following loss of connection, or if multiple recordings were made throughout the day (e.g. if a recording was stopped at lunch time and a new track created after lunch). In cases where a hike was spread over multiple days, or a single file contained many tracks from different days, we separated them such that each data point considered data from a single day.

Figure 4.22a shows the average walking speed of the tracks, where the tracks have been grouped by total walking time (into bins of 15 minutes). There is no large change in walking speed as the length of walking time increases, although the variance increases for longer walks. There is a slight increase in walking speed as walking time increase from 0-15 minutes to 15-30 minutes, and this is shown more clearly in Figure 4.22b, where we split the data into 50 equal bins by walk length. We can also see that most of the walks were short, under two hours in length (this is backed up by the histogram of walk lengths, Figure 4.23). We chose not to model the speed increase in the first half-hour of walking for a number of reasons. Firstly, in an implementation of the walking speed which is to be used practically, the walking speed should be easily understood. Having the predicted time for a section of a route change as the route length increases would not accomplish this. Secondly, over such a short timespan, the difference in distance walked would be very small ($\sim$50 m over 15 minutes). The wider variance of points in Figure 4.22a for walks over 8 hours is also explained by the histogram in Figure 4.23, as we can see we had much less data for these longer walks, leading to less precise average speeds.

We have confirmed that average walking speed does not increase as the walking time increases, however if we look at Figure 4.24a (which shows the average walking speed of the tracks, where the tracks have been grouped by total walking distance instead of time), we can see that the average walking speed of our tracks did increase as walks got longer. Figure 4.24b shows a zoomed in version of this, where the relationship is clearer. Average walking speeds increase on walks of up to 5 km in length, then level off at approximately 4.5 km/h until the walk reaches 20 km in length. Beyond this, the walking speed again increases. This relationship is very similar to that noted by Pitman et al. (2012), although they suggested that the walking speed decreases on hikes of length 7 to 17 km (in the region where we see a relatively constant speed). Our overall conclusion however is the same as that put forward by Pitman et al., namely that the increased speed is a result of higher fitness. In other words, people who have the ability to walk long distances in a single day are fit enough to walk at a



(a) Average walking speed of tracks, where tracks are grouped by total walking time into 15-minute bins.



(b) Average walking speed of tracks, where tracks are grouped by total walking time into 50 bins containing equal numbers of tracks.

Figure 4.22: Plots showing how average walking speed changes depending on the total walk duration.

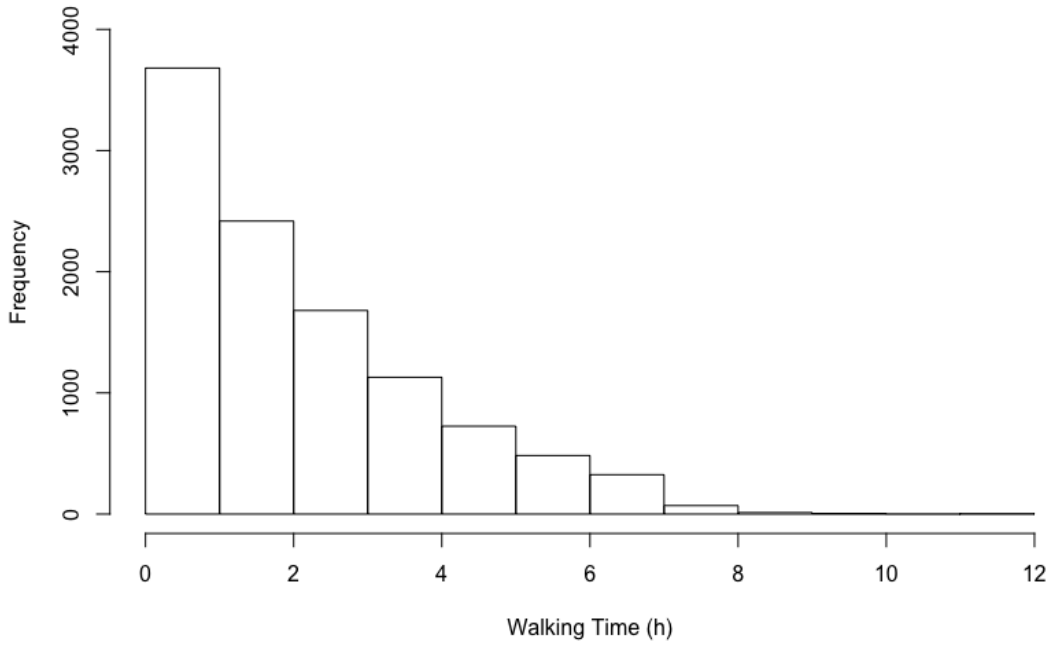
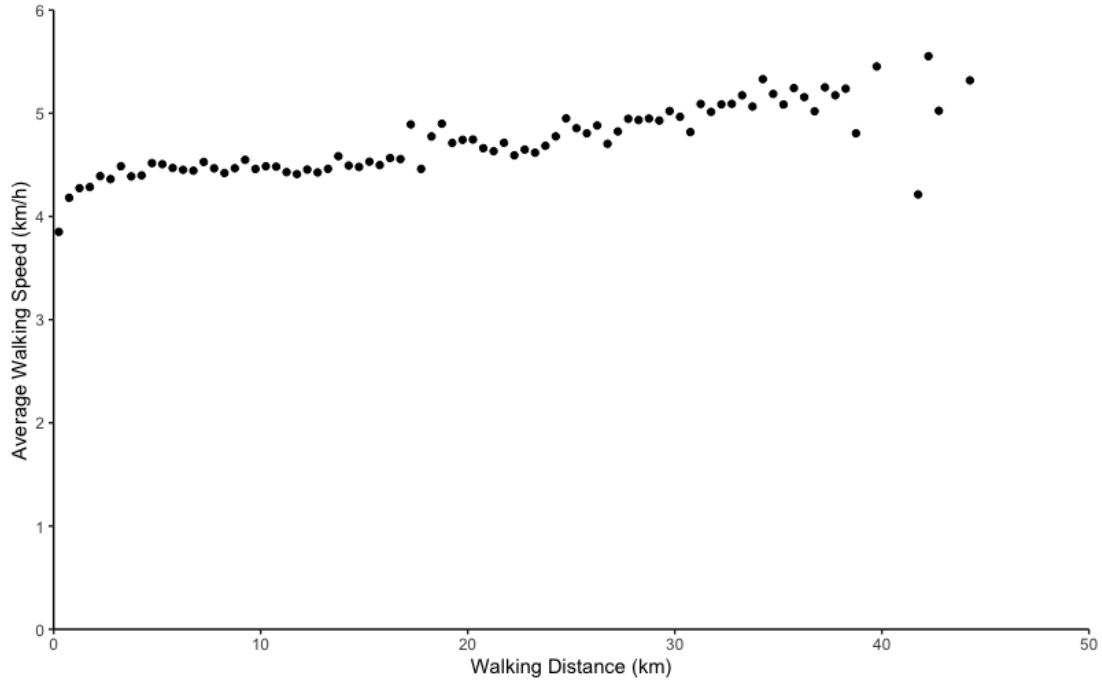


Figure 4.23: Histogram of the total walking times of tracks.

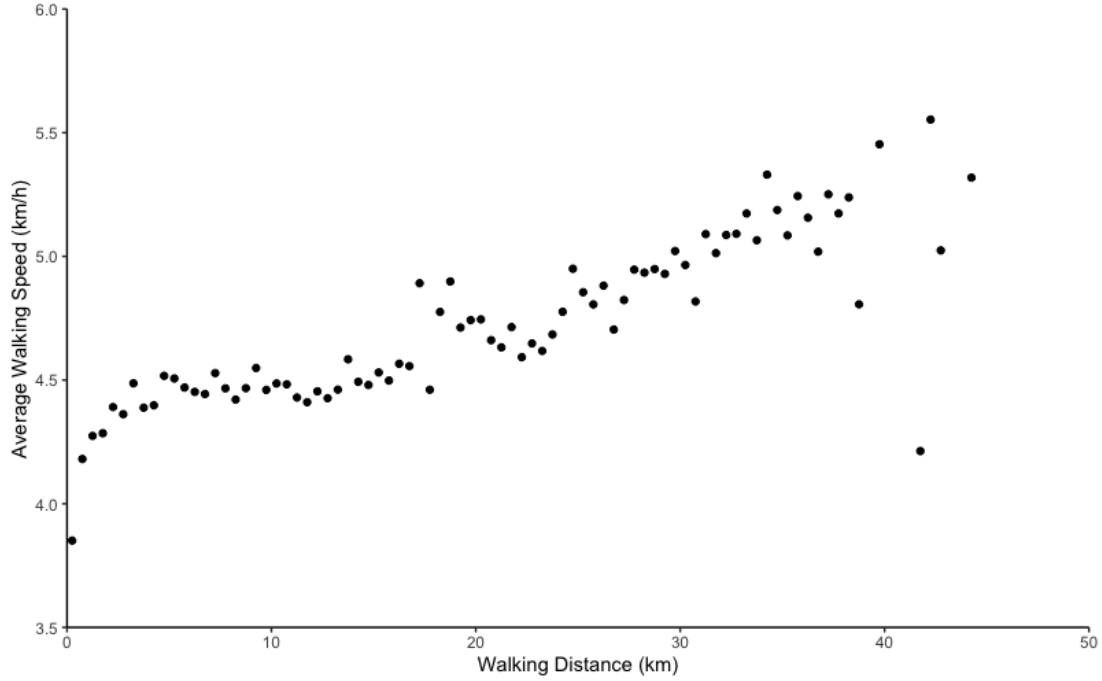
higher speed than average. This is backed up by our previous findings from Figure 4.22 where we did not see a similar increase in speed simply as a result of walk duration. We can infer that walks of 8+ hours in duration are undertaken by all types of people, and (as expected) fitter people can walk further in that time.

4.7.2 Total Break Duration Variance by Hike Length

Although the movement speed does not change as the walking time increases, it is still possible that people's overall speed over the course of a long walk decreases, as a result of taking more breaks. To check this, we looked at the total break time and the total walking time for each day. The tracks were grouped by walking time (as in Section 4.7.1), and the median total break time calculated for each group. The median break length was used rather than the mean to account for tracks where a device may have been kept running overnight. The overnight time would be recorded as a very long break (split over 2 days), and would artificially increase the mean. Figure 4.25a shows the break length plotted against walking time. We can see that there is a linear relationship between total walking time and total break time for walks up to approximately 5 hours in length (with around 1.5 hours of break taken). Beyond this however, the total duration of breaks taken begins to decrease. This pattern seems to end once the walking time is above 8 hours, although this is due to less data existing at these points. This is evidenced by Figure 4.25b, where the data was split into 500 bins, each containing an equal number of tracks. The same pattern to the data can be clearly seen, and no bins had an average walking time greater than 8 hours.

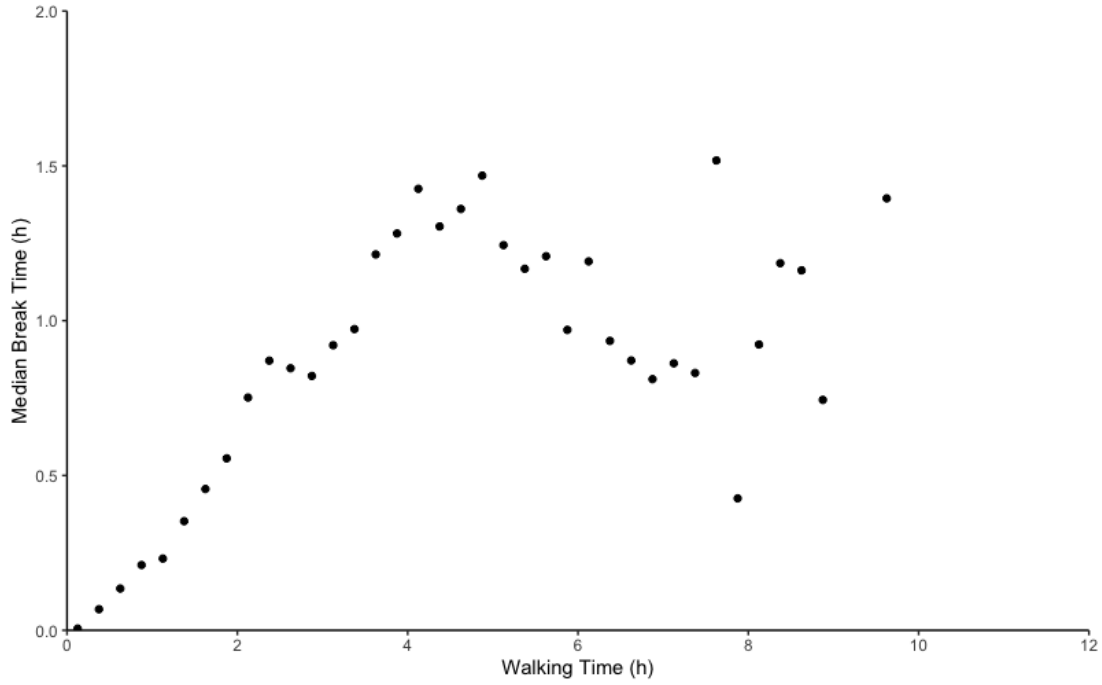


(a) Average walking speed of tracks, where tracks are grouped by total distance into 500 m bins.

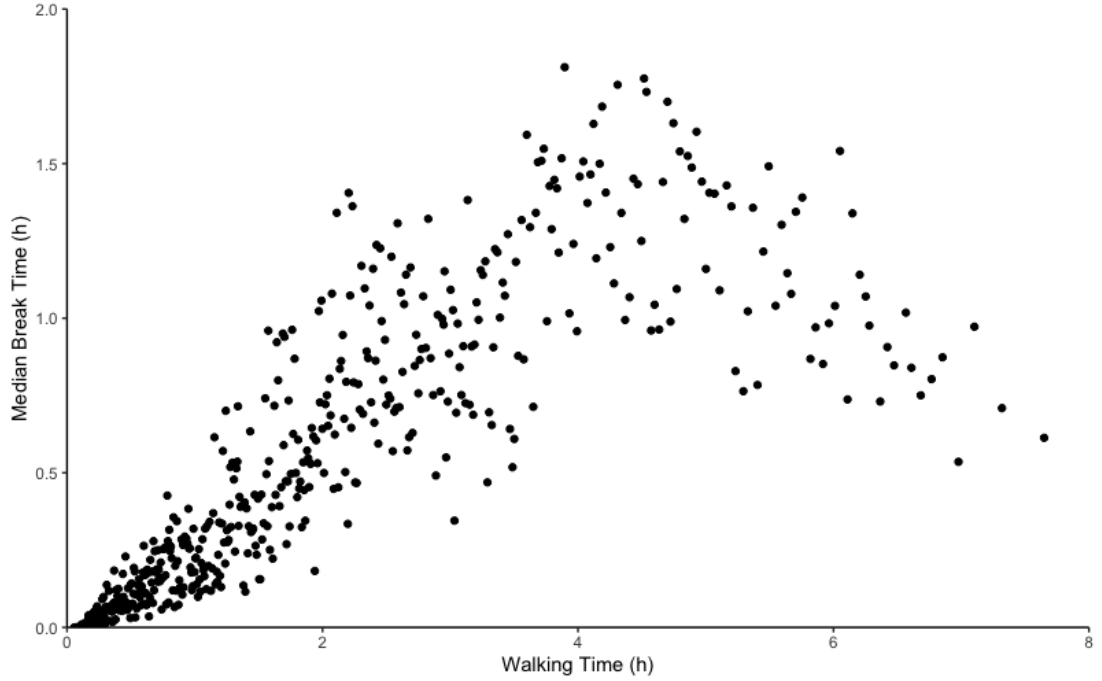


(b) The figure in 4.24a with the y-axis reduced to better show the relationship between walking speed and walk distance.

Figure 4.24: Plots showing how average walking speed changes depending on the total walk distance



(a) Average total break time of tracks, where tracks are grouped by total walking time into 15-minute bins.



(b) Average total break time of tracks, where tracks are grouped by total walking time into 500 bins containing equal numbers of tracks.

Figure 4.25: Plots showing how total break time changes depending on the total walk duration.

The reason for the decrease in break lengths on longer walks is unknown, but we suggest a number of possible explanations. People undertaking hikes with over 5 hours of movement will generally be physically fit, and may not need to take as many breaks as those who are unable to walk for such lengths of time. Secondly, walks over 5 hours are more likely to be specific 'hikes', where the aim is to complete the route, and so the participants may not want to stop for as long. On top of this, the longer walks may also have daylight hours as a determining factor. If the aim of the day is to complete an 8 hour hike, then the number and duration of breaks may be deliberately limited, so that the hikers can complete the walk in daylight. Shorter walks are less likely to have this time pressure, so more leisurely breaks can be taken. These considerations do not need to be taken into account in the walking speed model, but the findings here could be applied to a hiking route planner, as estimates for both walking time and total hike duration (including breaks) could be provided.

4.7.3 Obstruction Values of Different Terrain Types

We have classified the off-road data into two types in the walking speed model; light obstruction and heavy obstruction. However, for large portions of the ROUK dataset, and the entirety of the Scotland dataset, we did not have the lidar data required to classsify the terrain (see Section 4.5). The solution to this was to use a third model setup (off-road, obstruction unknown). Here we investigate the possibility of an alternate solution, namely using the Ordnance Survey Terrain Type data which is available for the whole of the UK (and described in Section 4.3.2). We looked at the obstruction levels from the lidar data in each of the OS terrain types, to see if it is possible to determine whether each terrain type should be classified as light or heavy obstruction.

Figures 4.26 - 4.28 show the obstruction values for each terrain type while off-road. It is important to note the different count ranges on each graph, as we had very little data for some of these terrain types. We have ignored terrain types with less than 20 occurrences in the data. From this we can see that there are a number of terrains which are classified as heavy obstruction more than half the time. As expected, the wooded areas (Coniferous and Non-Coniferous trees, and Orchard) are usually classified as heavy obstruction, and we can see the clear difference between these regions and their 'Scattered' counterparts. We also see that Scrub regions are usually heavy obstruction (the same is true of Cliffs, although the Cliff data is relatively limited in number).

Further work is needed to incorporate this data into the walking speed model. We have very few terrain types which are overwhelmingly measured as a single obstruction class, and of those which are, most have relatively low rates of occurrence. Furthermore, as we discussed in Section 4.3.2, each feature of the OS data can be associated with multiple different terrain types, and this is exacerbated when we merge the data into 50m segments, combining the terrains for all constituent points. We have over 100 unique combinations of the different

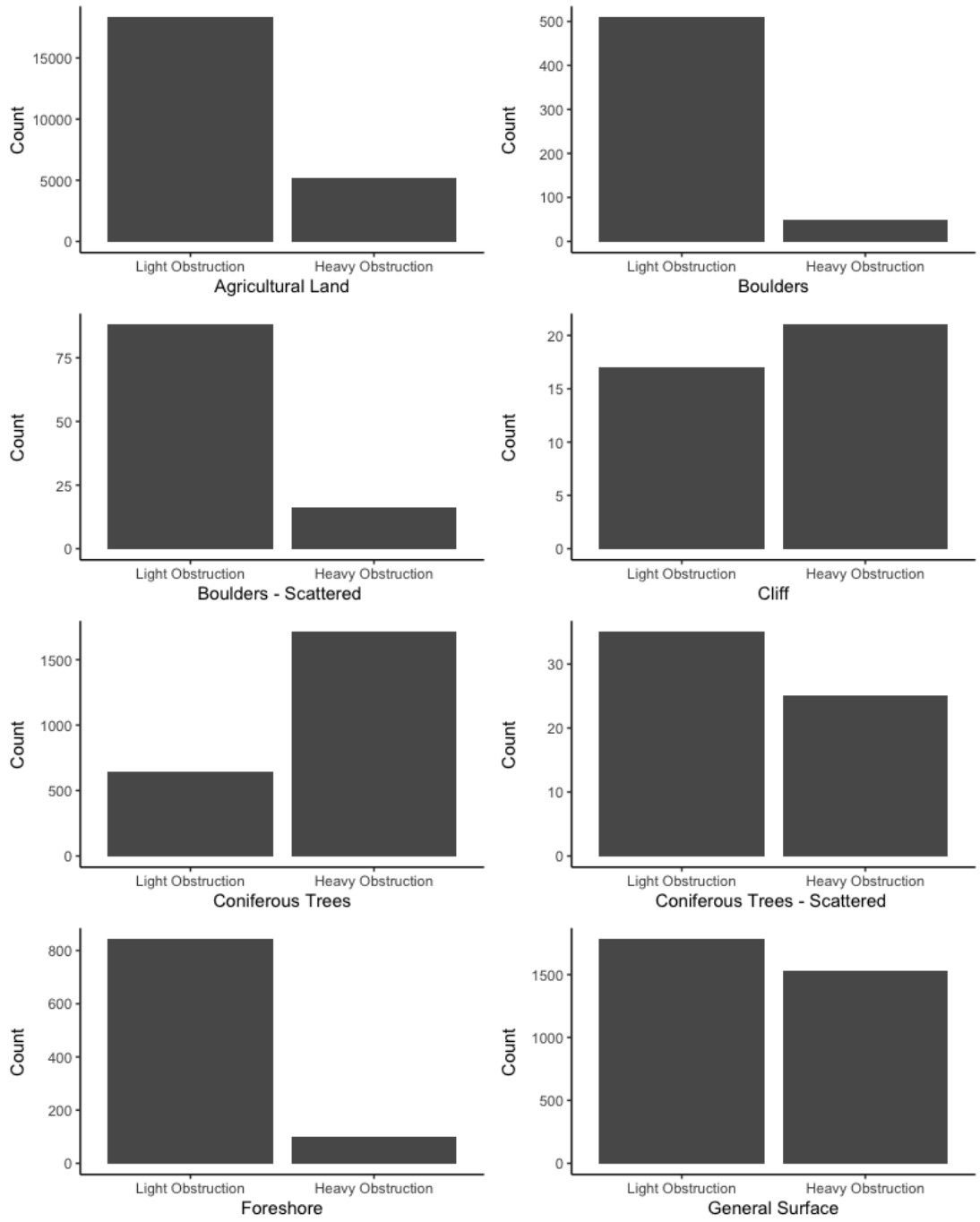


Figure 4.26: Numbers of light and heavy obstruction points observed in different terrain types (1/3).

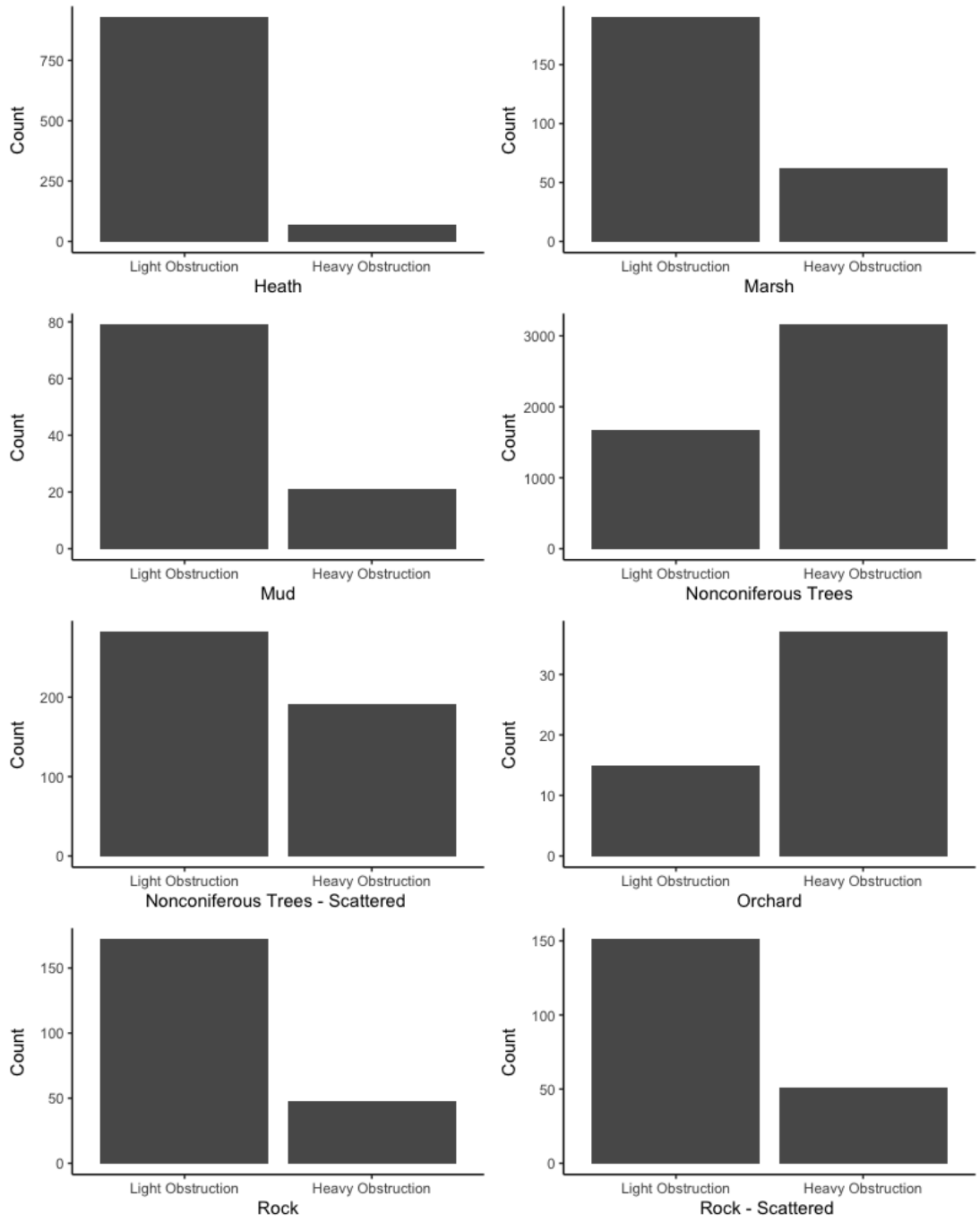


Figure 4.27: Numbers of light and heavy obstruction points observed in different terrain types (2/3).

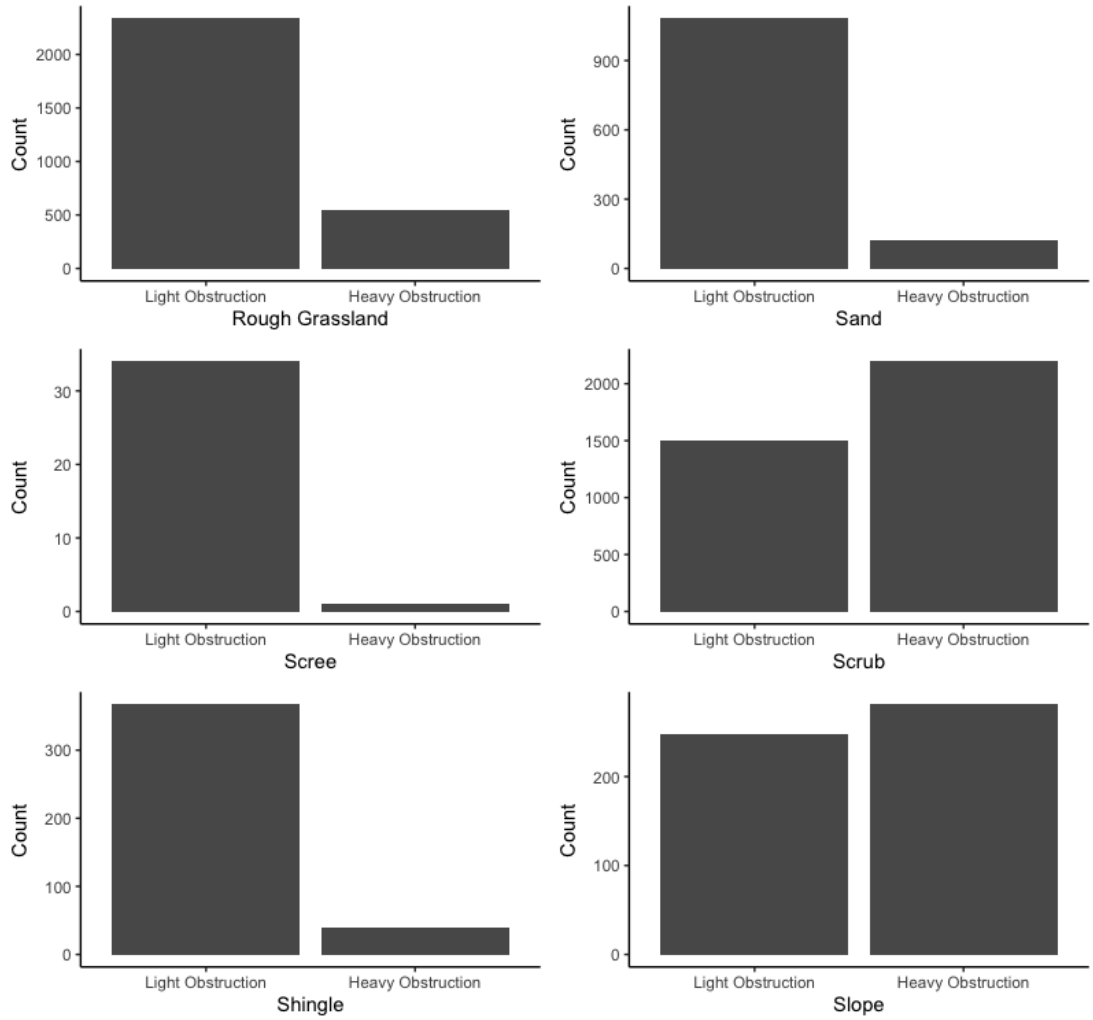


Figure 4.28: Numbers of light and heavy obstruction points observed in different terrain types (3/3).

terrain types which contain at least one of Coniferous, Non-Coniferous trees, Orchard or Scrub. More than half of these also contain terrain types which tend heavily towards light obstruction (Rough Grassland, Heath, Agricultural Land, Boulders etc). It is therefore very difficult to currently draw conclusions about how to interpret the data and include it in a model. One example of this is the Heath terrain type, which we may expect to be considered heavy obstruction, due to the thick shrubs which can be present, but is much more likely to be light obstruction. This is most likely due to the data resolution issues previously mentioned in Section 4.3.2. Many of the features with the Heath tag also include the Rough Grassland tag, indicating a field with some Heath patches. It is unlikely that, given the choice, hikers would walk through a patch of heath when they could simply go around it. However, all tracks which pass through fields with both Heath and Rough Grassland tags will have both terrain types associated with them.

In future, further investigation could be done into this, firstly to explore if there are less terrain combinations prior to the data merge stage, or if there is a correlation between specific combinations and terrain obstruction. A more complex solution may be to use satellite imagery and computer vision methods to determine which terrain types associated with a given terrain feature apply only to specific areas. For now, a solution for use in practical settings may be to warn users that off-road travel through woodland or scrub may be slower than the speeds predicted, if no lidar data is available.

4.8 Discussion

We have taken the model which we found for Scotland and attempted to validate it against data from the rest of the UK, as well as expand the model to include terrain factors. When expanding our dataset to validate our previous model, we found differences in the models produced using data from Scotland compared to the ROUK data. These were in part explained by investigating terrain differences between the two regions (with a much greater proportion of our Scottish data taking place on unpaved roads), and doing so reduced the difference such that it only occurred on unpaved roads or off-road (Section 4.4.1). Due to data limitations, we were unable to explore whether differences in the terrain obstruction levels were responsible for the differences observed in off-road walking speeds. However, we did note that the Scotland dataset has clearly distinct characteristics from the ROUK dataset in terms of the elevation profile (Section 4.4.2). We concluded from this, and the differences between the models, that the data seen in Scotland would be a valid, though extreme, sample of that seen in the rest of the UK.

When exploring terrain obstruction, we were unable to use the terrain description as a variable due to the low resolution of the data. However, by using lidar data to estimate the obstruction height, we have found an objective and easily measurable method to determine obstruction value (Sections 4.3.2, 4.5). While we have found a coefficient to determine the walking speed when lidar data is not available, we believe that as the scope of the lidar data coverage expands, our model accuracy will increase further.

Overall, we have found a model for walking speed which we believe is very robust, due the large volume of data (93,000 km) which was used to build it, and which correlates with the data over a wider range of conditions than commonly used formulae. We have successfully confirmed that each of the walking slope, the hill slope and the terrain type or obstruction are highly significant in determining walking speeds, and shown that this method improves on existing methods to predict walking speeds. We have also shown the specific improvement that our new model has on walking speeds in off-road conditions, compared to the simple off-road speed reductions used by existing models.

The benefit of using crowdsourced GPS data to build our model is also a limitation of the approach, as we did not have control over data collection. This meant that models were unable to account for group size, ability and composition, or other potential variables such as weather conditions, as factors in determining walking speed (although the volume of data should cause these effects to average out). The datasets were also filtered to remove as many non-walking or hiking tracks as possible and to identify and remove substantial breaks. While the methods used to filter the datasets were blinded to the outcome of the model generation, the choice of filtering methods will have had an impact on the dataset and subsequent model. Furthermore, the use of crowdsourced data meant that all of our data came from ‘walkable’ regions by definition. When exploring the terrain obstruction, we were unable to determine if there is a level of terrain obstruction which makes walking impossible. Similarly, the vast majority of the data was collected on shallow hill- and walking slopes, leading to a lack of data in steep areas. While this does mean that we can be very confident about our walking speed predictions in less steep regions (where most walking occurs), it is unclear whether the lack of data on steeper regions is a result of steep slopes being relatively rare, or that they cannot be easily navigated, so hikers chose an alternate path.

Chapter 5

Fieldwork

In the previous chapter, we found a new model to predict walking speeds which took into account the walking slope, the hill slope and the terrain type or obstruction level. We showed that this model provides improvements over the most widely used existing models, although we noted that the crowdsourced nature of our underlying dataset limited our ability to test methods used in data filtering, and made it difficult to establish the limits of regions where the model could be applied. Fieldwork was therefore carried out to validate and test our model, while exploring the limits of where it is applicable. Originally it was intended for the fieldwork to be a much larger component of this research, which could provide us with quantitative data to explore the reliability of the model in areas where we had limited quantities of data (as described in Section 6.3). However, due to the impact of the Covid-19 pandemic it was not possible to arrange for groups to meet up, so the scale of the fieldwork was reduced.

5.1 Setup

The main aspects of the fieldwork were to answer the following questions:

- Were our methods to remove breakpoints and filter non-walking tracks valid?
- Is the new model accurate under normal walking conditions? If so, does it provide an improvement over the existing hiking functions?
- Is the model accurate at more extreme slope values, and what are the limits of applicability?

As explained in Section 3.1.2, we did not have access to any ground truth data when creating the algorithm to define breakpoints in a GPS track. A visual inspection of tracks showed that the algorithm was successfully identifying clusters of points believed to be breaks, while not identifying walking points as breaks, although no formal testing was conducted. As the original dataset came from a large number of sources using a variety of GPS devices and settings, we were not able to ensure that the algorithm is fully accurate in all scenarios. By conducting

fieldwork we were able to control the device settings, and manually record where breaks were taken, to compare with the algorithm results. This enabled us to check our assumption that we were able to identify significant breaks in a GPS track, or if this was not the case, we would gain information about what methods could be employed to improve the algorithm.

We already had some evidence that our model is accurate under normal walking conditions, through our comparison of walking speed predictions of our crowdsourced data against those of existing hiking functions (Section 4.6). The fieldwork gave us a more controlled dataset to further our conclusions. However, our existing datasets did not contain many points with steep (>20 degrees) walking or hill slopes. We aimed to test the model accuracy in these regions. One potential explanation for the low volume of data may be due to it not being possible to walk in such areas. If this is the case, we would like to investigate at what slope angle this happens, and whether the walking speed function gradually decays to zero at this point or whether it should be be discontinuous.

The experiment took place in the Pentland Hills, just south of Edinburgh, using a group of Scouts as the test participants. Each member of the group was carrying a device to record their speed throughout the day. The majority of the day consisted of a standard Scout training day; we tracked the Scout group to test whether the walking speed model is accurate under normal conditions, in places where the Scouts chose to walk. However, at two points in the day, specific experiments were conducted on high-slope surfaces in order to test the model validity at extreme values (see 5.3.3). The start- and end-times of these experiments were noted, so we were able to separate the specific experimental data from the data for the rest of the day. Ethical approval was obtained for this and all data was anonymous. The Scout group provided qualified leaders to supervise the day, and briefed the participants and their parents about the use of GPS devices. Following the fieldwork, the participants were debriefed regarding the findings, and how their data would be used.

The following recording devices were used to track the participants:

- 5 iPhone devices using the myTracks app (Dirk Stichling, 2021)
- 1 Garmin watch

Due to battery issues, one of the iPhone tracks only recorded part of the route.

The myTracks apps were set up with a 1 second recording interval, and the criteria to continue recording at breaks (in order to validate the break-finding methods used) and also to not automatically smooth out the track. These features had to be manually configured, as the default settings turn them off. The settings may have led to some problems as detailed in the following sections. The Garmin watch track was recorded using the default device settings. We were aware that the number of participants in the fieldwork was not large enough to provide quantitative results, however we were able to produce qualitative feedback on the

model and the methods used in its formulation, which could provide guidance on where future work should focus. As all of the tracks were known to be walks, the filtering methods used on the OSM data were not necessary. The tracks were therefore processed in the same manner as the Hikr tracks used in Chapter 4.

5.2 Model Creation Validation

Before exploring how well our model predicted the walking speeds in our test data, we tested the methods which were used in filtering the data in order to create the model

5.2.1 Breakfinding

The first method tested was the break finding algorithm (described in Section 3.1.2). The majority of the fieldwork consisted of a standard Scout training day, where the participants were hiking normally. This included both long and short breaks, and is representative of a typical hike. We also monitored the participants during the high-slope experiments which were undertaken, allowing us to test our break finding algorithm under less typical circumstances. Notes were taken about where and when breaks were taken during the fieldwork, and these were compared to the breaks identified by the algorithm. In general, the break finding algorithm was successful, with breaks being correctly identified in almost all situations. For all but one of the tracks, all of major the breaks taken throughout the day were found, although there was also a small amount of over classification. The remaining track had a large over-classification issue as almost the entire length of the track was identified as a break, and this is investigated in the next section.

The algorithm, which identified breaks by looking for clusters of points, was found to over-classify regions as breaks. The majority of this over-classification was subsequently removed by implementation of the 30-second minimum break length, designed to ensure that only valid breaks were identified. As discussed in Section 3.1.3, only breaks of over 30 seconds were felt to be a constituent part of the walk. Figure 5.1 shows one of the tracks in full, where the breaks identified by the algorithm are highlighted. Those which are under 30 seconds in length are coloured green, while the ones which were considered to be valid breaks (and thus the points were removed from processing) are pink.

Almost all of the other tracks follow the same pattern, with sub-30 second breaks present throughout the whole track, but occurring most commonly on the steeper, slower sections. Without the 30 second threshold as part of the break identification process, over-classification would be a much greater problem, however in most cases we do still have some issues with over-classification of breaks. We explore the reasons behind this in the following sections.

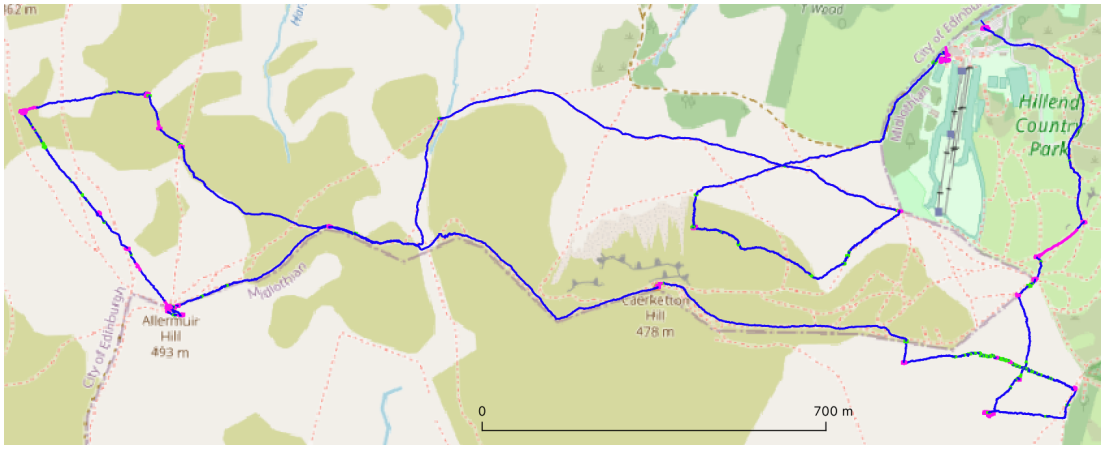
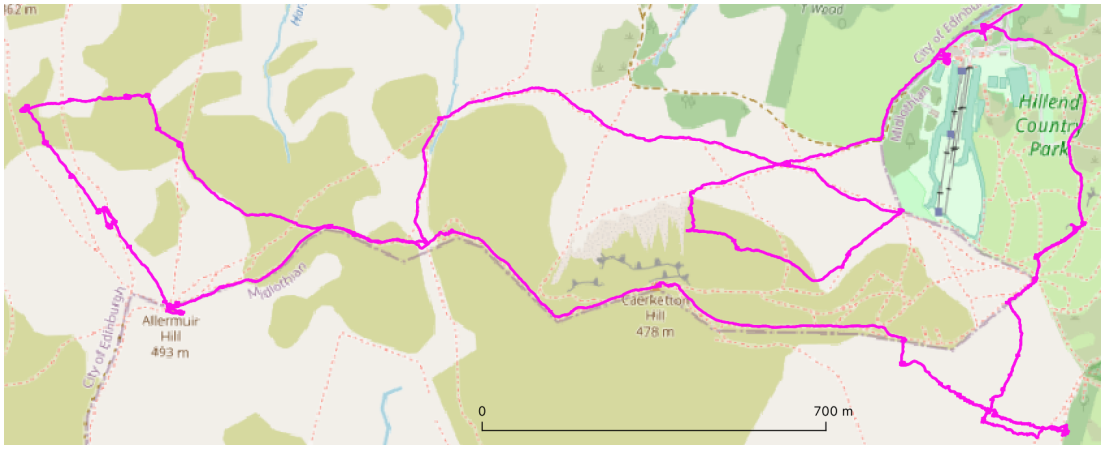


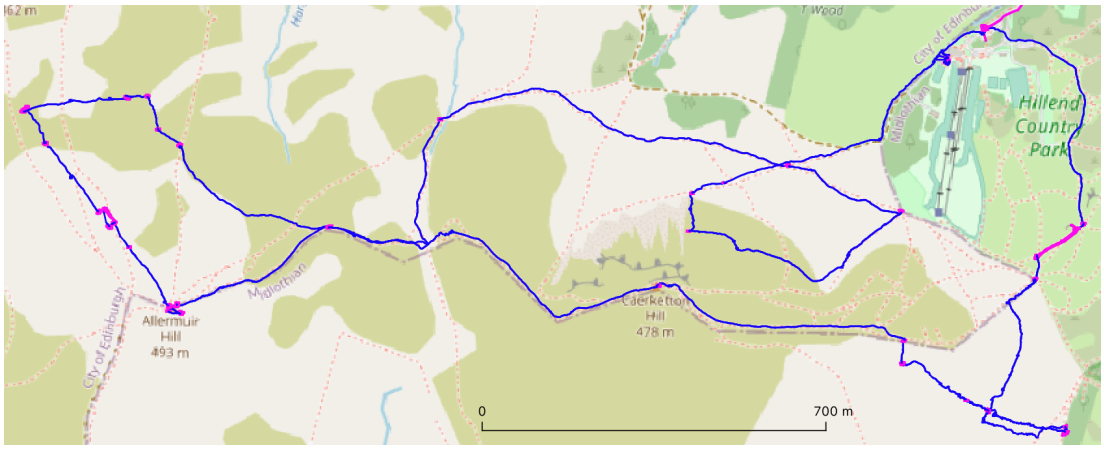
Figure 5.1: One of the GPS traces from the fieldwork, with breaks highlighted. Breaks under 30 seconds are coloured green, and breaks over 30 seconds are coloured pink. Background map from OpenStreetMap, visualised using QGIS (see 2.2.5).

Zero-speed Medians

As mentioned above, one of the tracks from the fieldwork was classified almost entirely as a break. Upon investigation, we found that this was a result of the median speed for the track being 0 km/h. The reasons for this are twofold: we set up the iPhones to not automatically pause recording when the app detected a break, and the recording interval was fixed at 1 second. Each time a break was taken we would expect a cluster of points to form as a result of GPS drift. However, in this particular case there were a large number of points during breaks which experienced no GPS drift and therefore had a speed of 0 km/h. Multiple long breaks were taken throughout the day, and there were a large enough number of points with no GPS drift to cause the median speed for the whole track to be 0 km/h. Subsequently, all of the points recorded during movement had a speed greater than twice the median speed and were thus classified as ‘high-speed points’. As described in Section 3.1.2, this caused clusters to propagate along the whole length of the track, until every point was included. The break was then defined as being between the first and last points with medium or high ‘break likelihood’, which encompassed almost the entire track. It is unknown why only one of the tracks was affected in this manner, as the same settings were used on all of the iPhones. The device in question was not the latest iPhone model used during the fieldwork, so it is unlikely to be a result of newer hardware and software eliminating GPS drift. The device settings to bring about this issue were configured deliberately, and most of the tracks in our original datasets would not be affected. This is evidenced by the data from the Garmin watch (which was using the default settings), which did not contain any points with a speed of 0 km/h. Although unlikely



(a) Breaks (pink) identified by the algorithm when all points are included. The entire route is tagged as a break.

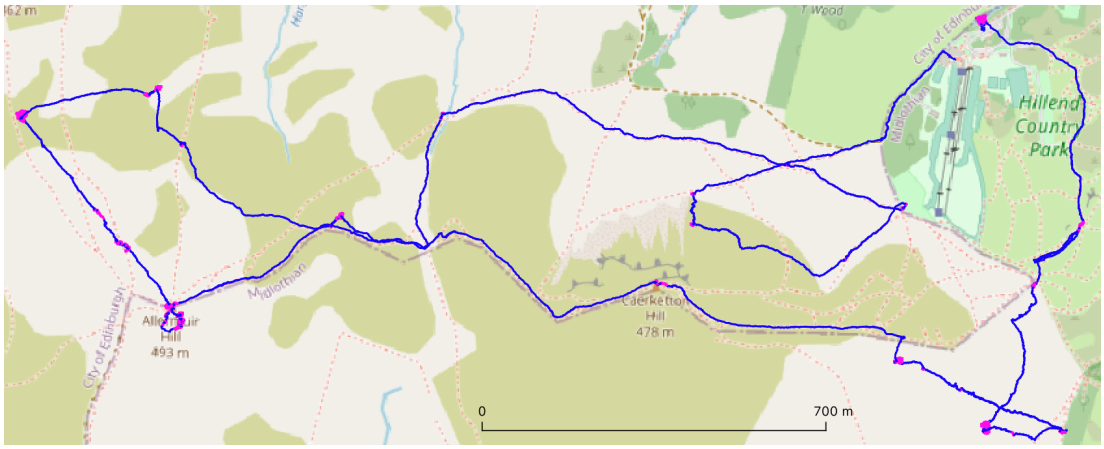


(b) Breaks (pink) identified by the algorithm when zero-speed points are ignored.

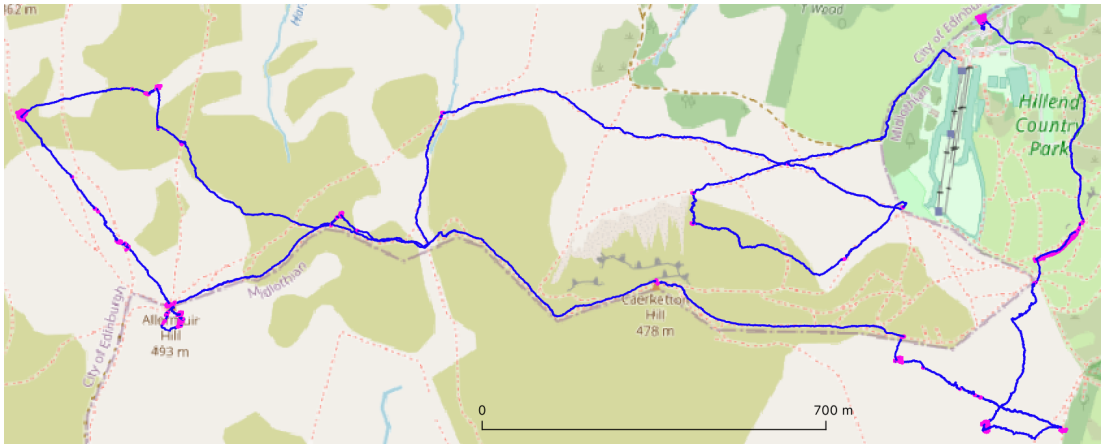
Figure 5.2: Comparing results of the break finding algorithm when we a) include zero-speed points when calculating median speeds and distances, and b) remove zero-speed points when calculating median speeds and distances. Background maps from OpenStreetMap, visualised using QGIS (see 2.2.5).

to be a widespread issue, this can be alleviated by updating the break-finding algorithm, such that the median distance and speed values are changed to be the median values, when zero-distance points have been removed. We tested this method on the track in question, and the majority of the over-classification was removed (Figure 5.2).

We also tested this updated method on each of the tracks which were not affected, and there was very little difference in the breaks identified (Figure 5.3). While we believe it is unlikely that this issue will have affected our data, it is possible that a small number of tracks were inadvertently ignored by classifying them as breaks. In future, the break finding algorithm



(a) Breaks (pink) identified by the algorithm when all points are included.



(b) Breaks (pink) identified by the algorithm when zero-speed points are ignored.

Figure 5.3: Comparing results of the break finding algorithm when we a) include zero-speed points when calculating median speeds and distances, and b) remove zero-speed points when calculating median speeds and distances. Background maps from OpenStreetMap, visualised using QGIS (see 2.2.5).

Figure 5.4: A steep section of one of the GPS traces from the fieldwork, with breaks highlighted. Breaks under 30 seconds are coloured green, and breaks over 30 seconds are coloured pink. 10 m contour lines are also shown to indicate hill slope. Background map from OpenStreetMap, visualised using QGIS (see 2.2.5).

should be updated as we have described, as this will prevent these cases from occurring, while having a minimal impact on other tracks. Although a small number of tracks may have been incorrectly ignored during processing, we do not believe that it will have affected our model due to the number of tracks we were still able to work with.

Overclassification through GPS Drift

As previously mentioned, the majority of the overclassification in the break clustering algorithm occurred on the steepest sections of the route. Figure 5.4 shows a portion of a track, where the walking slope was very steep. Two breaks were identified (pink) as well as multiple sub-30 second 'micro-breaks' (green). No breaks were taken during this section, therefore all of these are points where the algorithm has overclassified breakpoints, and it is easy to see why this has occurred. The track doubles back on itself a lot and there are often a large number of points grouped over a small area, causing multiple point clusters to form.

The excess clustering was likely caused by the slow walking speeds on steep terrain combined with GPS drift. At very low speeds the distance between points is also low, so small amounts of GPS drift cause larger deviations than in other, faster regions. We can see this by looking at Figure 5.5, which shows more of the same track (blue). The right-to-left portion is recording the participant climbing the steep slope, while the top-to-bottom line shows the participant traversing the slope. Higher speeds are seen when traversing the slope, along with less deviation and GPS drift, so there are no mis-classified breaks, and fewer places where short 'micro-breaks' (green) were found and ignored for not meeting the 30 second threshold. (The single identified (pink) break in the traverse portion of the route is correctly classified.)

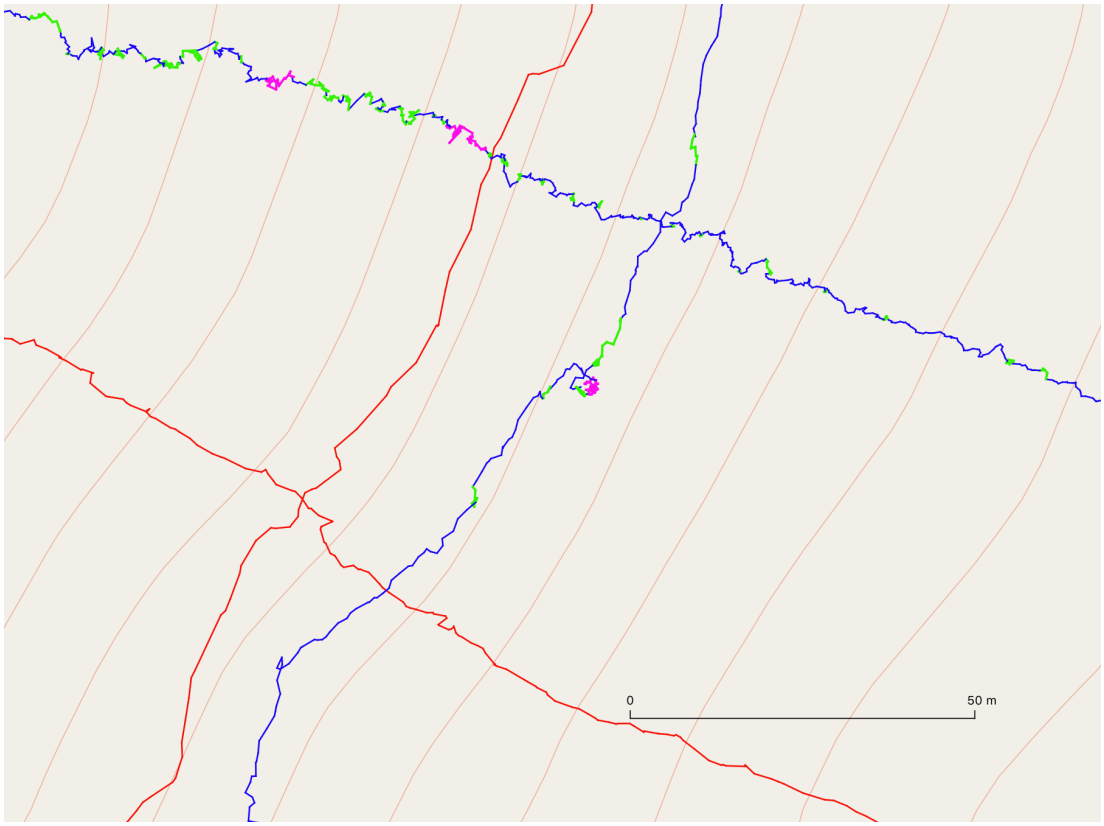


Figure 5.5: Sections of two GPS tracks on a steep section of the route. Each track contained both a hill traverse (top-to-bottom lines) and a hill climb (right-to-left lines). One track was recorded using an iPhone and the MyTracks app (blue), while the other was recorded on a Garmin watch with default settings (red). Breaks under 30 seconds are coloured green, and breaks over 30 seconds are coloured pink. 10 m contour lines are also shown to indicate hill slope. Background map from OpenStreetMap, visualised using QGIS (see 2.2.5).

The slow uphill speeds themselves are not the direct cause of the over classification. Figure 5.5 also shows the track recorded by the Garmin device (red) on the same slope, and visually we can notice that it is much smoother. This is backed up by the data, where we did not detect any breaks on the steep uphill section. The main reason for this is that the iPhone tracks were set up to record a new point each second, whereas the Garmin device only recorded points once sufficient movement was detected. The average duration for points in the Garmin track is 5.53 s. This is enough time such that, even when travelling at slow speeds, the GPS drift did not cause clustering. In order to test this theory, the iPhone tracks were processed a second time, but a minimum duration of 5 seconds between consecutive points was set. The result of this change can be seen in Figure 5.6. The overclassification as a result of excessive GPS drift in this track has been removed, while the Garmin track is largely unchanged. This does not entirely fix the problem, as some of the tracks do still contain excess breaks, although the number of instances of this are reduced. While implementing this new method, none of the correctly identified breaks in other areas of the track were affected.

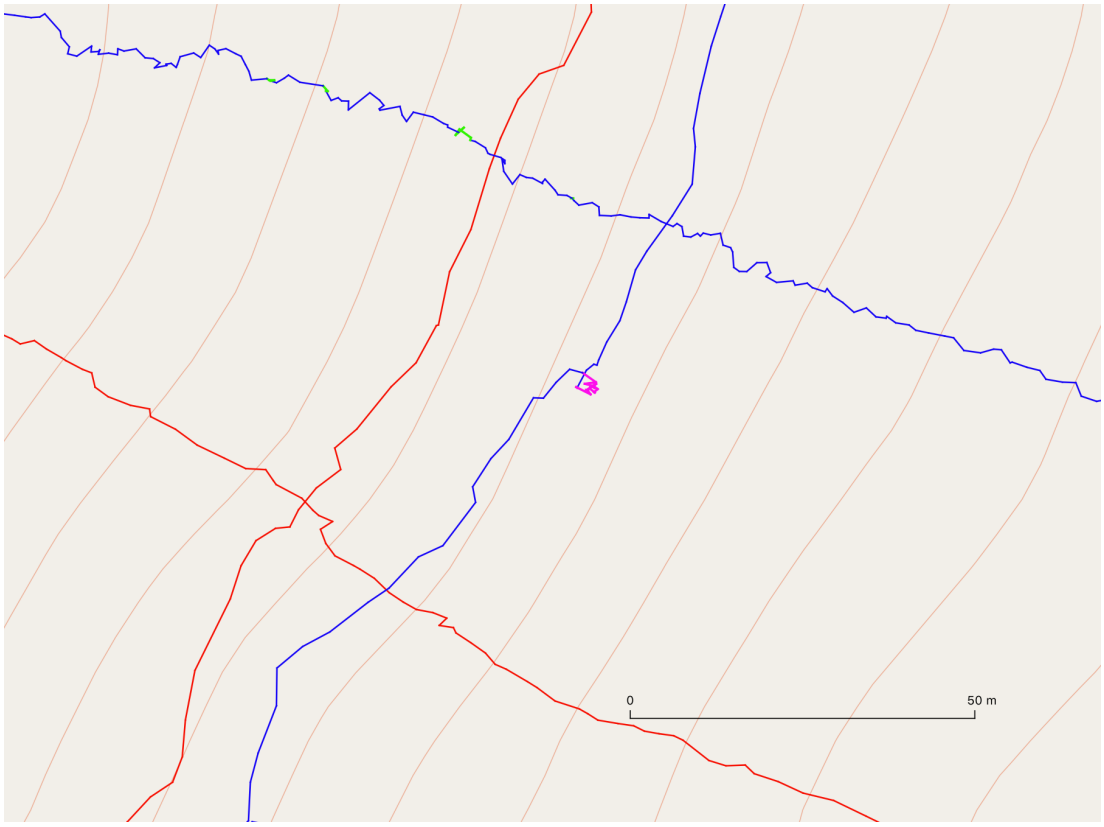


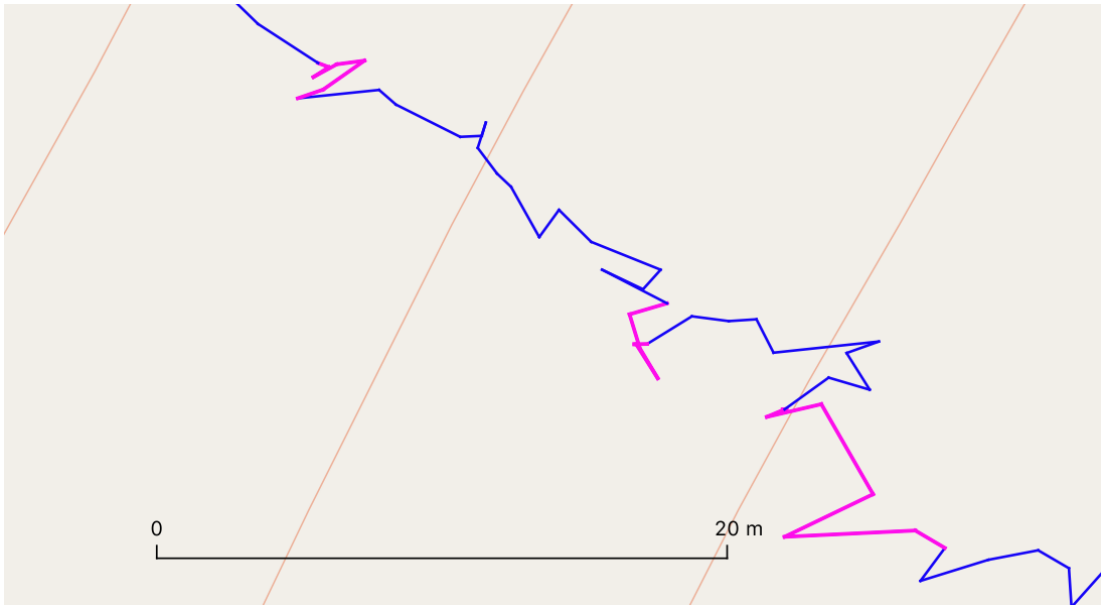
Figure 5.6: The sections of GPS tracks from Figure 5.5, where a 5 second minimum time has been applied between track points. Each track contained both a hill traverse (top-to-bottom lines) and a hill climb (right-to-left lines). One track was recorded using an iPhone and the MyTracks app (blue), while the other was recorded on a Garmin watch with default settings (red). Breaks under 30 seconds are coloured green, and breaks over 30 seconds are coloured pink. 10 m contour lines are also shown to indicate hill slope. Background map from OpenStreetMap, visualised using QGIS (see 2.2.5).

We have seen that slow regions of a route can result in overclassification from a combination of high device recording rate and GPS drift. In future iterations of this work a 5 second filter, or other filtering method, could be applied to reduce this. However we do not believe this overclassification will have had a major impact on the model as a whole for a number of reasons. Firstly, during this experiment, the devices were deliberately set up to record a point every second. In practice, very few of the devices which were used to record the tracks used in the model would have been set up like this, as it is a non-standard device setting. Secondly, we have seen that overclassification occurs only in the slowest regions (generally very steep slopes), so the majority of our data will be unaffected. Finally, despite multiple clusters being erroneously tagged as breaks initially, very few of these clusters were long enough to reach the 30 second threshold and be tagged as a break. Therefore, only a very small portion of our data will have been misclassified as a result of excessive GPS drift in very slow regions.

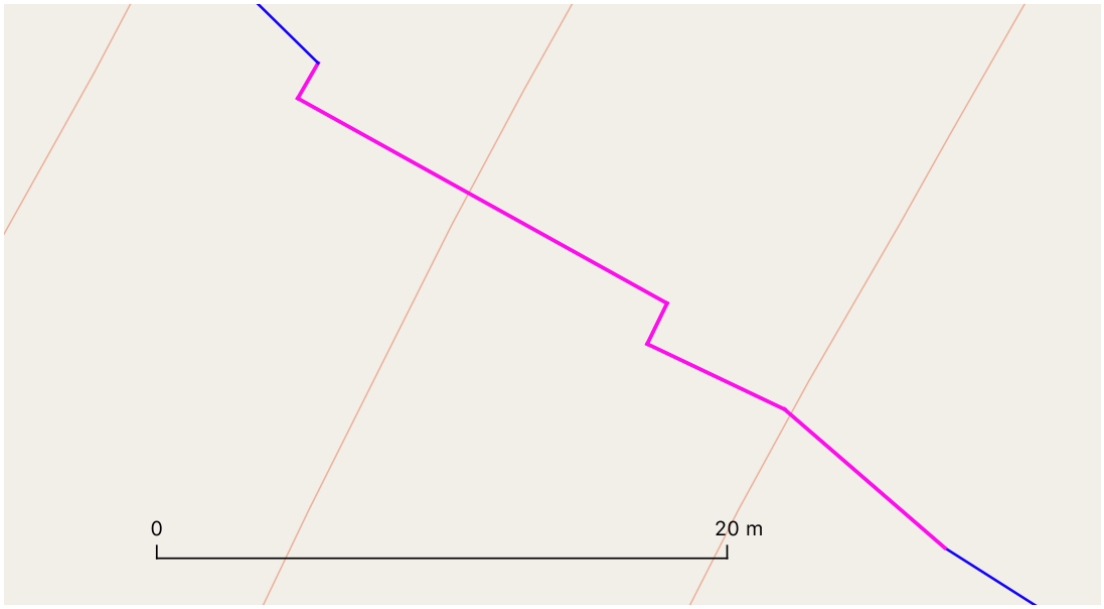
Extending Overclassification

As mentioned in the previous section, even allowing a 5 second gap between points did not remove all instances overclassification in slow regions, and this was be exacerbated when we merged the data into 50 m segments. If the algorithm identified two breaks with less than 50 m of walking between them, then this walking section would also be identified as a break. An example of this can be seen in Figure 5.7, where the pink regions are incorrectly identified as breaks due to of GPS drift. The walking distance between these points is less than 50 m (the maximum travelled distance between consecutive breaks shown in Figure 5.7a is 22 m), so when we merge the data into 50 m sections, then entire region is tagged as a break, as seen in Figure 5.7b.

This does not appear to be a widespread problem, both due to the unusual device settings required, and the relative rarity of such slow sections. However, the rarity of the regions also means that while uncommon, this issue should be looked at in future. We did not have a large amount of data on very steep walking slopes, and by incorrectly removing valid data we reduced our already limited dataset even further. A number of potential solutions to this problem exist and could be looked into. Firstly, the 50 m merge distance could be reduced. This would prevent such areas from occurring, although situations where we have a ‘moving break’ may be affected. This would occur for example at the summit of a hill, where it is common to stop for a break and take photos. If someone is moving around a hill summit taking pictures, this can appear in a GPS trace as multiple breaks, separated by very short sections of movement. An alternative option would be to implement a minimum duration, rather than a minimum distance, between breaks. If calibrated correctly, this should allow for ‘moving breaks’ to continue to be classified as a single continuous break, while also not removing slow travel over short distances between break points. Both of these methods could also be explored in future work to find a compromise which maximises correct data classification.



(a) A section of one of the GPS traces from the fieldwork, with breaks highlighted in pink. 10 m contour lines are also shown to indicate hill slope.



(b) The section of GPS route seen in 5.7a, once a minimum distance of 50 m has been applied to all walking sections. 10 m contour lines are also shown to indicate hill slope.

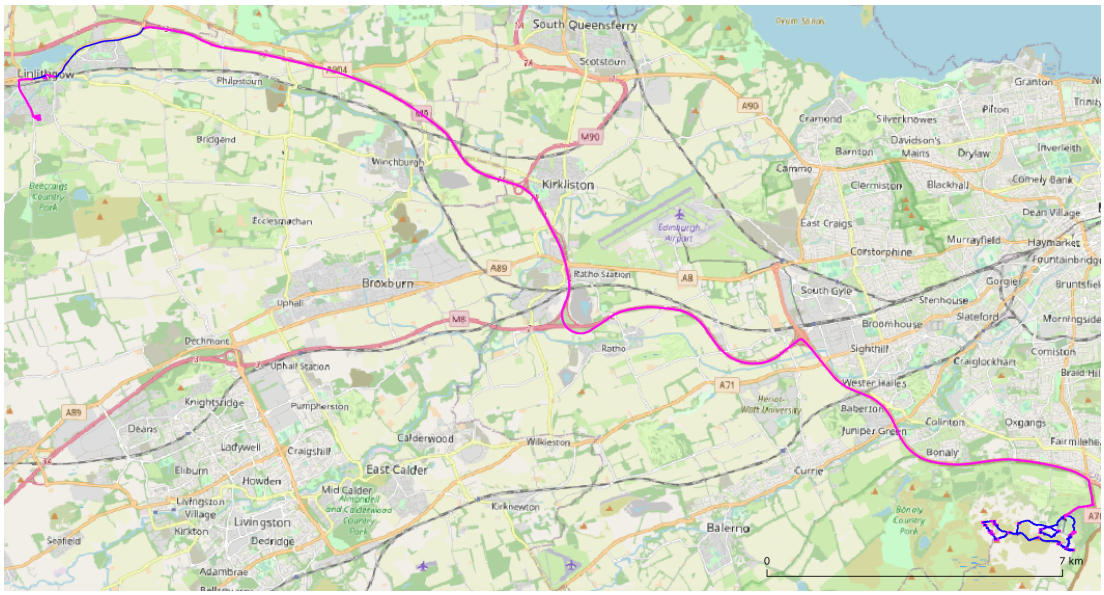
Figure 5.7: Demonstrating how overclassification of breaks can lead to further removal of valid data when data points are merged into 50 m sections. Background maps from OpenStreetMap, visualised using QGIS (see 2.2.5).

Non-walking Regions

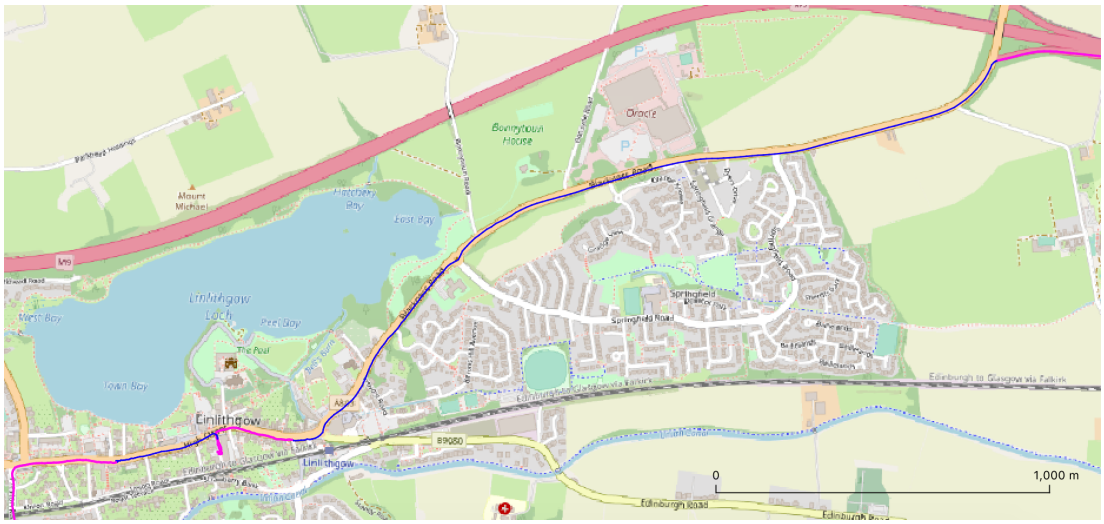
It was important that our Hikr data contained only tracks which were valid walks or hikes, as this dataset was used to determine the filter values of our OSM data. When processing our Hikr data, we manually removed segments which had been incorrectly tagged as hiking reports (when the metadata or description labelled them as trailruns), and any track segment with an average speed of over 10 km/h was removed as a non-walking segment. (The filtering process and methods are described in Sections 3.1.3 and 4.1.) However, this did not account for any track segments which contained contained a small amount of driving, usually in order to get to or from the walk location. In these situations, we relied on the break-finding algorithm to detect and remove these points. One of the fieldwork GPS tracks also contained a period of driving after the walk concluded, so we were able to test how accurate our method was at detecting and eliminating non-walking regions in a track segment. Note, this track is also the one which was identified as being almost entirely a break previously. Therefore, in order to see how well driving sections are filtered out, we must use the updated break-finding algorithm which ignores zero-speed points (discussed in Section 5.2.1). As mentioned, this change to the algorithm had very little impact on the classification of breaks for other tracks, so conclusions found here will still apply to the rest of our data.

Figure 5.8 shows the GPS track in question, with areas identified as breaks highlighted in pink. The majority of the driving portion has been highlighted as a break but there are two sections where this is not the case (Figure 5.8b). Both of these regions consist of a run of high-speed points, bounded at either end by a single slower point (i.e. when stopped in traffic). The consecutive high-speed points caused these regions to be identified as a cluster, however they were not subsequently flagged as a break. This is because the constraint requiring travel in opposite directions within a break was not met (see Section 3.1.2 for details).

Although not flagged as a break by the clustering algorithm, the longer of these two regions was subsequently removed from the dataset. When we merged the track into 50 m sections, the short periods of slow travel at either end of the region were combined with their neighbours to produce sections with high average speeds (>10 km/h). As described in Section 4.1, high speed sections next to existing breaks were considered to be part of the break, and were therefore removed. As all of the merged sections in the longer region had average speeds of over 10 km/h, each was added to the break in turn, until the whole region had been classified as a break. This did not occur for the second, smaller region because the average speeds of the 50 m sections adjacent to the identified breaks were not as high, so high-speed, break-adjacent points were not found. This shows us that while our current method is able to capture the majority of driving regions, there are conditions under which it may not work; specifically a period of low speed driving, such as when stuck in traffic or along a high-street with many pedestrians.



(a) A GPS track which included a period of driving. Points identified as breaks are highlighted in pink.



(b) Closeup on the a portion of the track in 5.8a, where the algorithm has failed to identify two separate periods of driving as a break.

Figure 5.8: Demonstrating the efficiency of the break finding algorithm at also identifying driving sections of a GPS route. Background maps from OpenStreetMap, visualised using QGIS (see 2.2.5).

Overall however, we can have confidence that the vast majority of any driving sections contained in Hikr tracks were automatically flagged and removed, and thus that our dataset which was used as a basis from which to classify OSM track sections as either walking, or non-walking, was valid. Furthermore, prior to building our walking speed model, we filtered the fastest 0.5% of speeds in our dataset and removed them as outliers. This speed cutoff was 8.52 km/h. Had this track been a part of our full dataset, only a single section (with an average speed of 7.05 km/h) would have remained and been incorporated into the model.

In order to test the filtering process used for the OSM tracks, we also processed the fieldwork tracks according to the same rules that were used for that dataset (described in Sections 3.1 and 4.1). In this instance, the driving points were all successfully removed from the dataset. There were also 3 valid walking points which were removed (across all 6 tracks), which occurred when a single 50 m track section was encountered between two relatively long breaks. From this, we can be confident that periods of driving within a GPS track will have been successfully removed. This is evidenced by the lack of high-speed points in our final dataset; the 99.5% upper quantile of walking speeds was 8.52 km/h.

Through this test, we have shown that the majority of driving sections within a route will automatically be removed as a break by our filtering methods, however there are circumstances where this will not be the case. If a longer period of slow driving happened (for example along a high street with lots of traffic and pedestrian crossings), then this could cause data to pass through both filtering methods erroneously. In order to distinguish these sections, more work needs to be done. This could include categorising breaks into separate groups: legitimate walking breaks, or high speed breaks (i.e. driving). The high speed breaks should not have the requirement of travel in opposite directions which we used when identifying point clusters. Furthermore, periods of travel between high-speed breaks should be examined in more detail to determine whether they are walks, or slow sections of driving.

Underclassification

We do have some evidence to suggest that while we were capturing break locations correctly, they were not being fully identified by the break finding algorithm. In our final (50 m merged) data, we found 29 points with speed predictions greater than 2 km/h different to the observed values. 6 of these occurred in driving conditions (and so would be removed by our data filter as described in Section 5.2.1), but of the remaining 23, 13 occurred immediately preceding or following a tagged break. This implies that we may be correctly identifying breaks, but not finding all of the points associated with them. Note that the majority of these sections contained a ‘micro-break’ identified by the algorithm (a break shorter than the 30 second threshold). It is generally clear from the GPS trace that these identified ‘micro-breaks’ were

distinct from the main break cluster, suggesting that the participant either slowed down prior to reaching the eventual break point, or paused briefly after leaving it. One solution to this may be to extend breaks, such that micro-breaks within a certain distance or time of the main identified break are added to it.

Of the remaining 10 points with high residuals, 8 featured a break which was identified by the algorithm, but under 30 seconds in length. As discussed in Section 3.1.3, 30 seconds was chosen as the threshold using the Scotland data, in conjunction with the idea that some breaks are 'necessary' for the route. More work should be conducted to look at this threshold in order to determine if it should be lowered. Furthermore, although currently necessary to prevent a large amount of overclassification as a result of GPS drift, future improvements to break identification may also enable us to remove the threshold without this overclassification occurring.

We also found a very small number of places where the initial break finding algorithm failed to find a break. In all cases, this is because the total break itself was short (under 1 minute), and made up of only a very small number of points. None of these points had a speed below 0.01 km/h (the 'very low speed' cutoff which caused a cluster to be formed - see Section 3.1.2), so a break cluster was not created. 0.01 km/h was chosen as the minimum speed because it successfully captured the breaks in the tracks which were used in testing. However, given the evidence from this fieldwork, this value should be increased to account for other device types and settings. Using the results seen here, it could be raised to 0.5 km/h, or possibly higher, although care should be taken to avoid identifying breaks in legitimate, slow sections of routes.

In our current implementation, the most extreme cases of break underestimation will have been removed when we filtered our dataset to remove the slowest 0.5% of walking speed. Furthermore, within any given walk the percentage of points which were affected by this will be very low, so we are not considering it to be a problem which affected our work. However, in future, updates should be made to the break identification algorithm.

Overall, we can be confident that the break finding algorithm successfully identifies breaks in GPS tracks. We do have some areas where over-identification occurs, specifically slower sections, and a small number of very short breaks may be missed by the algorithm in its current form.

The changes to the break algorithm which could be explored to improve it are the following:

- Adjust the median values to ignore zero-speed points
- Adjust the minimum speed required to trigger a break cluster
- Adjust the minimum time threshold to classify a break
- Implement a minimum distance or duration between consecutive points in the track

- Adjust the merge distance of points, to reduce removal of otherwise valid data
- Investigate a more targeted method to remove driving and other non-hiking sections

All of these except the first would require testing in order to determine the new values of each threshold. For now however, we do not believe that the current break implementation has had a detrimental effect on the walking speed model. The majority of the overclassification seen in this fieldwork required specific device settings, and very slow walking speeds in order to occur. Furthermore, due to the volume of data we were working with in the model, we do not think that misclassification of a small number of breaks will have affected our final results.

5.2.2 Road Classification

A further area of data classification we were able to explore in the fieldwork was the classification of on-road and off-road sections. All of the points on roads were correctly identified as being so. Similarly, on the roads which were paved, all of the points were correctly identified. However, this did not extend in the opposite direction. There were a number of points where the route ran along an unpaved track adjacent to a road, and was classified as a paved road. Similarly, a number of regions where participants crossed open, off-road terrain were classified as being on a road. This was not a problem caused by the data itself, but by our choice of a 50 m search radius around each point to detect a road. As discussed in Section 4.3.1, we knew that this would likely result in overclassification of roads, but accepted it as a necessary limitation of our approach. A decision was made that it was better to misclassify points as being on-road when they were not, due to the relative frequencies of each type. Off-road points incorrectly labelled as being on-road would have less impact on the model because we had such a large quantity of on-road data. Even in this small study, there were areas where the GPS tracks deviated consistently from the road by up to 10 m, and when selecting the search radius on the original data, a number of areas with greater deviation were found. If a smaller radius were used, and these points were incorrectly classified as off-road, this would have a greater impact on the model as we had less off-road data to start with. However, it is possible that we erred too far on the side of caution. By using a 50 m radius to find roads we accepted a certain level of misclassification, and this misclassification was compounded by the decision to require only a single on-road point to classify a 50 m section as on-road when merging the data. Future work should look into adjusting this classification, either finding a smaller distance which may minimise total misclassification, or using a more nuanced approach, determining whether a run of points is following the route of a road or path, and if so, classifying them as such. We do not believe that this has greatly affected the model, as we still had large quantities of data to work with, but improving this will help ensure the accuracy and reliability of the model, particularly in off-path regions.

5.3 Model Validation

One of the main focuses of the fieldwork was to gather data in order to validate the model found in Chapter 4. In order to maximise the data available for this, we processed the data using the break formula updated to ignore zero-speed points when calculating the median speed (described in Section 5.2.1). This allowed us to use data from the GPS track which would otherwise be entirely tagged as a break, while having minimal impact on the breaks found throughout the other tracks. During model validation we did not initially remove outlier speeds (unlike in the modelling data where the fastest and slowest 0.5% of data points were removed). Instead, we can investigate these individually should any issues arise. We first explored our walking speeds over the range of 'normal walking', and subsequently explored high slope regions.

5.3.1 General Walking Speeds

Throughout most of the day, the Scout group was not under our instruction about where to go, so they were able to choose a route as they normally would. We separated this data from the specific high-slope experiments which were conducted, and considered it to be demonstrative of normal walking conditions and terrain.

After processing the data, there were seven outlier points (points with speeds greater than 8.52 km/h; the cutoff point used in the model data to detect outliers). Five of these occurred under driving conditions while the remaining two occurred immediately preceding or following a break. As discussed in Section 5.2.1, outlier points adjacent to breaks may indicate areas where our method to identify a break has not captured all of the points which should be associated with it. There was also one point remaining where the GPS device was recording a driving section and was not flagged as a break, or an outlier (speed of 7.05 km/h). This highlights that more work should be done to remove such non-walking points if the model is to be recreated.

On the low end of walking speeds, our minimum walking speed was 1.23 km/h, higher than the 1.14 km/h outlier cutoff used in modelling. This gives us some evidence that our cutoff did not incorrectly identify and remove valid walking data (although we were only using a small sample size here). Of the 12 slowest points (those with speeds below 2 km/h), 10 occurred either immediately next to a break, or prior to one of the experimental regions (where a brief pause was taken to ensure all participants began simultaneously). This may suggest that our break identification methods need further adjustment.

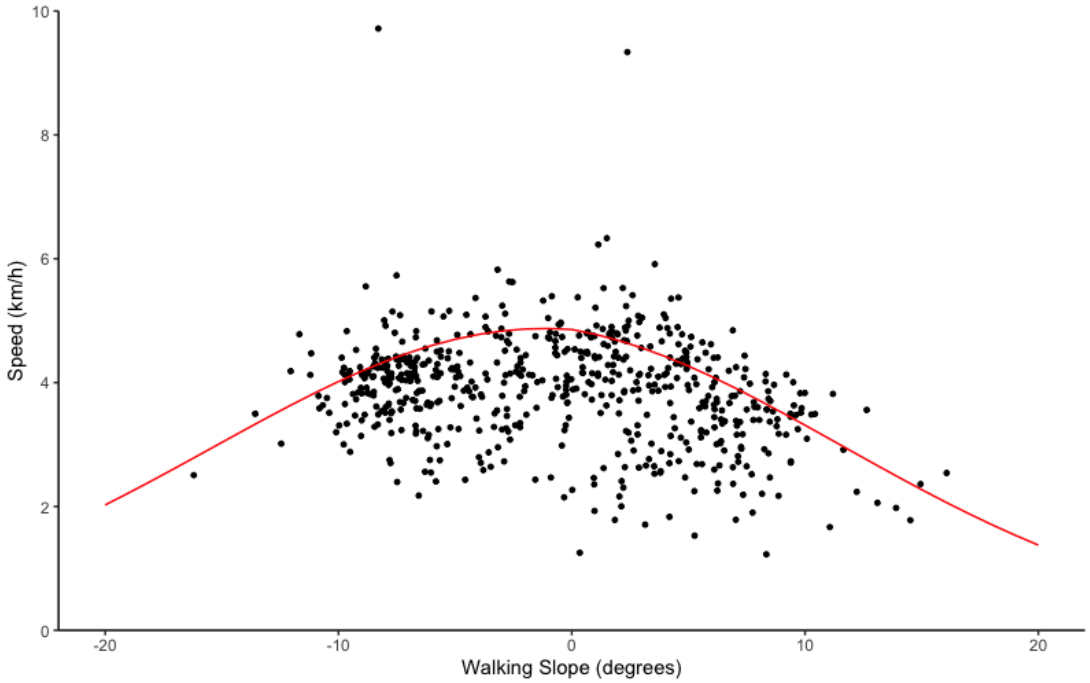
Although we have noted that a number of points may have been misclassified, this only affected a very small proportion of the data collected. As such, we were confident that the fieldwork data is an accurate representation of walking under standard conditions, and we could use it to test our model's prediction ability.

A small number of manual adjustments were made to the data to test speed prediction. First, we manually set the road and path sections to have the correct attributes. Similarly, we ignored the outlier points identified above, which were a result of driving. If the model were being used to predict the walking speeds for an individual planning to walk this route then only the walking sections would be modelled, and the road or path indicators would be taken directly from the map data, not the GPS data. Note that we did not remove the high or low speed points which were adjacent to breaks. We could not be certain whether these speeds were valid or not, so they were kept in. Finally, we did not have lidar data available for the experiment region to measure terrain obstruction. However, during the off-path sections the heavy obstruction locations were noted, and later confirmed against satellite imagery, so we were able to use the more accurate model with heavy/light obstruction for model testing.

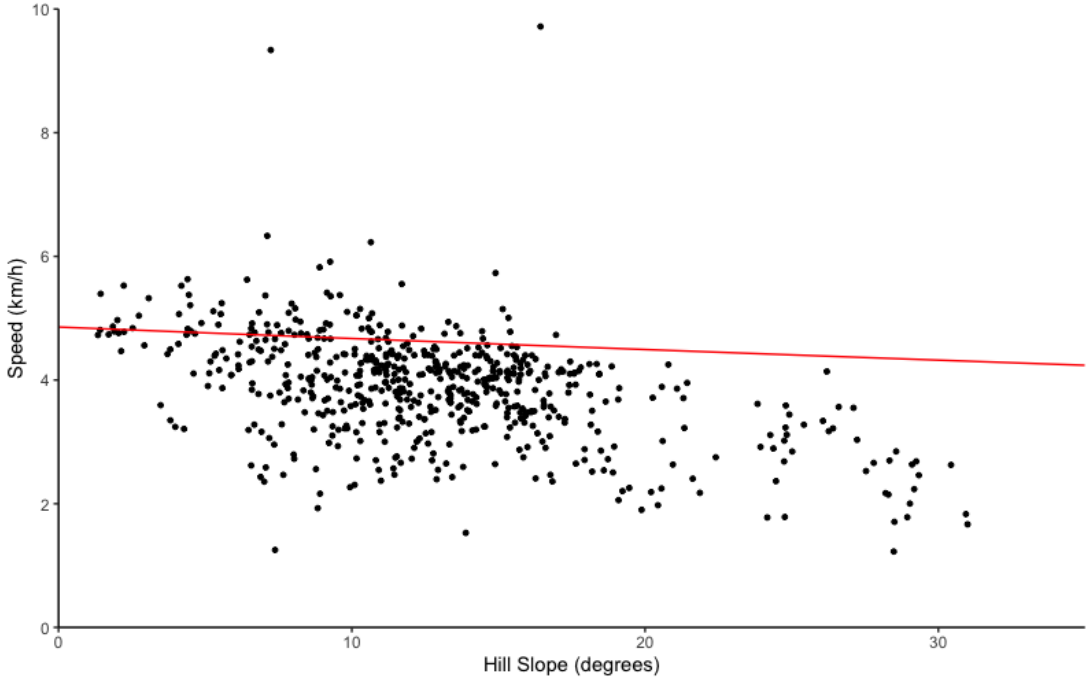
Figure 5.9 shows the walking speeds plotted against both hill slope and walking slope. Also drawn is the predicted walking speed for an unpaved road (the most common terrain type encountered), when either directly climbing (Figure 5.9a) or traversing (Figure 5.9b) the slope. Note that the model lines indicate the predicted maximum speeds, and most of the time we would expect slower speeds as the walk was not perpendicular or parallel to the slope. As expected, we can see the walking speeds decrease with increasing slope steepness. The two previously mentioned high-speed points are clearly visible as outliers.

We calculated the residuals of the walking speed predictions for this data and they can be seen in Figure 5.10. Our residuals appear to be vertically centred around zero, although there is a general trend of slightly overestimating the walking speed. Rather than suggesting the model predictions are inaccurate, we suggest that the participants were walking at a slower than average speed. This is likely because of the instructional nature of the Scouts' day - there were a large number periods where the Scouts were walking while following a compass bearing for example. Slower speeds are not surprising here as the group was focused on navigation while also walking. We can also see that on steeper hill slopes, most of our residuals are negative, where the model predicts higher speeds than were achieved. This suggests that the model may not be treating hill slope as a large enough factor, and perhaps a greater penalty term must be used to reduce the predicted walking speed. It is difficult to tell from these residuals whether the current linear hill-slope term is correct (though the slope is not steep enough), or whether the hill slope effect may be non-linear, with the walking speed decreasing more rapidly at steeper slopes. This should be investigated in the future by repeating the fieldwork with a greater number of participants.

We also calculated the residuals if we predicted the walking speed using either Naismith's rule (Figures 5.11a, 5.11b) or Tobler's hiking function (Figures 5.11c, 5.11d). It is easy to see from these figures that Naismith's function overestimates speed when walking downhill. This is not unexpected, as Naismith's rule predicts a fixed walking speed of 5 km/h in downhill areas. However, we also see that Naismith's rule underestimates speeds when walking uphill.

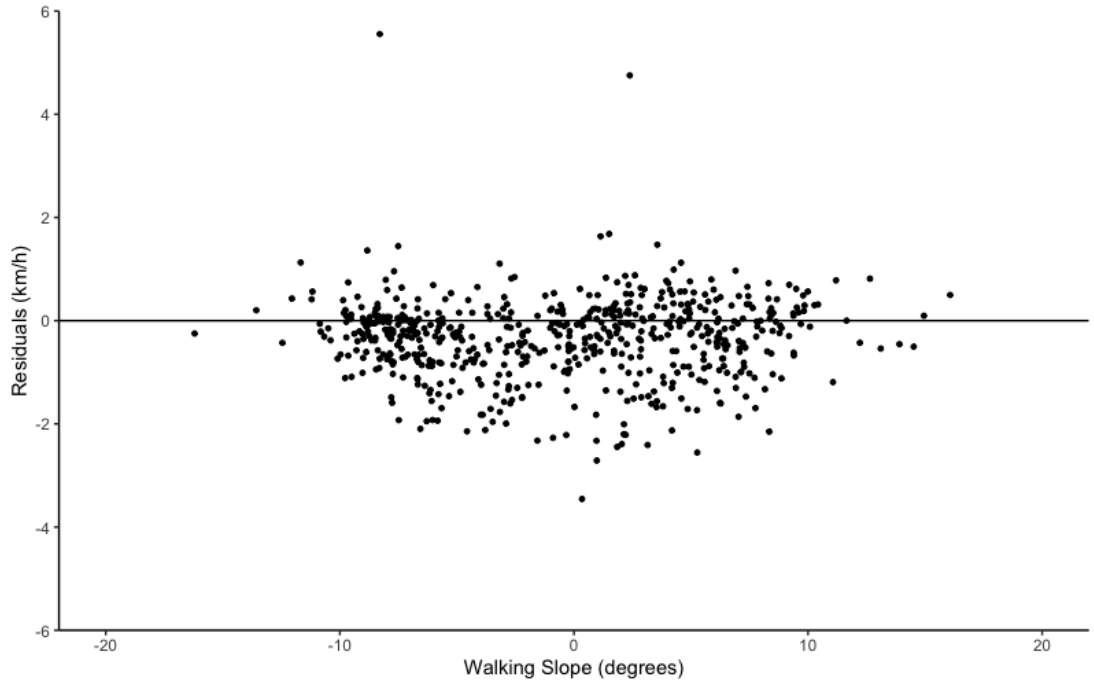


(a) Walking speed plotted against walking slope for our Scout fieldwork data under normal walking conditions. Also shown is the predicted walking speed when directly ascending or descending the slope (red).

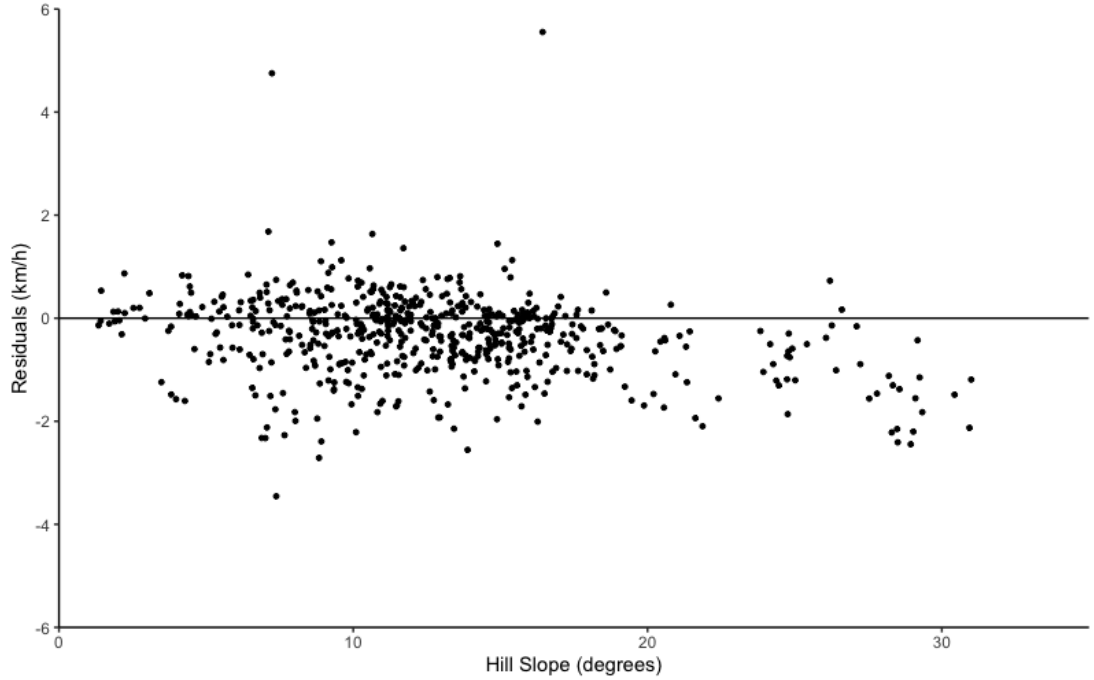


(b) Walking speed plotted against hill slope for our Scout fieldwork data under normal walking conditions. Also shown is the predicted walking speed when traversing the slope (red).

Figure 5.9: Plots showing the walking speed against slope values for our Scout data, alongside the speeds predicted by our model to traverse or climb a slope on an unpaved road.

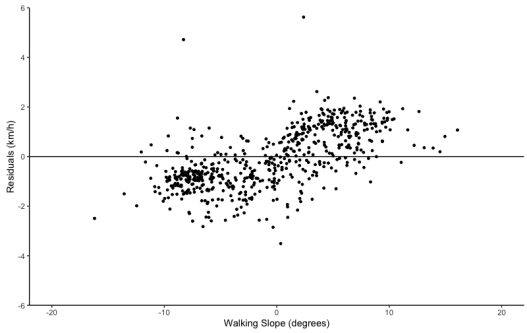


(a) Model speed residuals plotted against walking slope for our Scout fieldwork data under normal walking conditions.

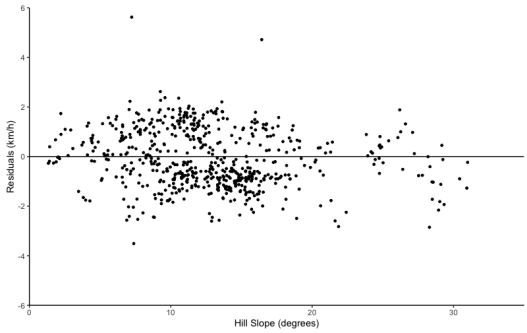


(b) Model speed residuals plotted against hill slope for our Scout fieldwork data under normal walking conditions.

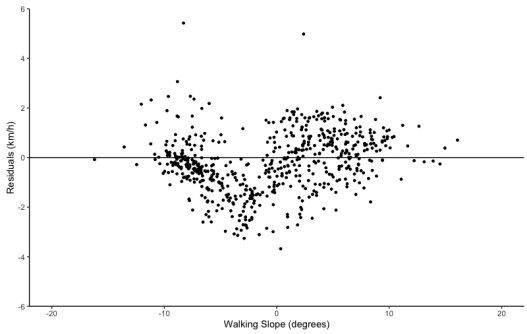
Figure 5.10: Plots showing the residuals of walking speeds predicted by our model for the Scout data.



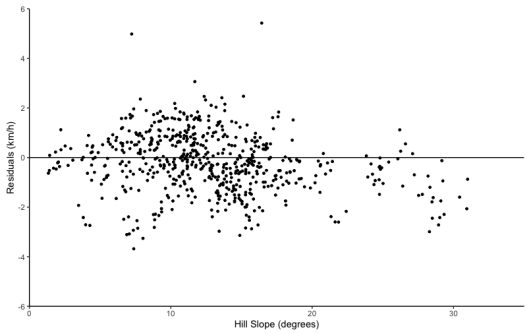
(a) Naismith speed residuals plotted against walking slope for our Scout fieldwork data under normal walking conditions.



(b) Naismith speed residuals plotted against hill slope for our Scout fieldwork data under normal walking conditions.

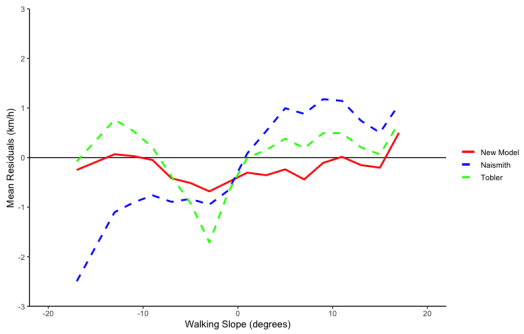


(c) Tobler speed residuals plotted against walking slope for our Scout fieldwork data under normal walking conditions.

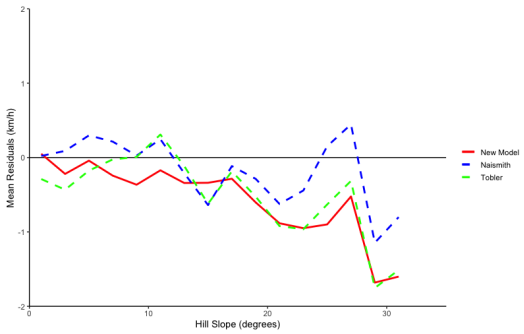


(d) Tobler speed residuals plotted against hill slope for our Scout fieldwork data under normal walking conditions.

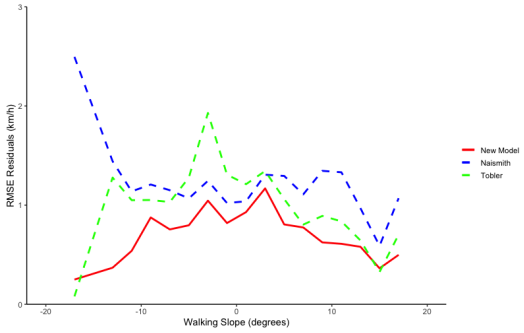
Figure 5.11: Plots showing the residuals of walking speeds predicted by Naismith's rule and Tobler's function for our Scout data.



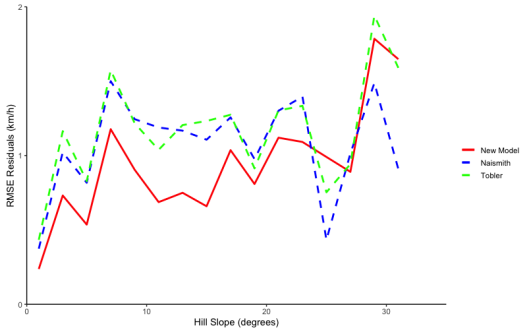
(a) Comparison of mean values for the residuals of walking speed plotted against walking slope.



(b) Comparison of mean values for the residuals of walking speed plotted against hill slope.



(c) Comparison of RMSE values for the residuals of walking speed plotted against walking slope.



(d) Comparison of RMSE values for the residuals of walking speed plotted against hill slope.

Figure 5.12: Comparing mean and RMSE residual values for the new model (red), Naismith's rule (blue) and Tobler's function (green).

Similarly, the walking speeds calculated according to Tobler's function are overestimated when descending at a shallow slope angle. The difference between each set of residuals is less clear when looking at the hill slope, although the residuals from our new model are more tightly clustered around zero. When we look at the average residuals we can see this tighter clustering more clearly. Figures 5.12a and 5.12b plot the binned mean residuals for each model (bin widths of 2 degrees). Both of the existing models have walking slopes at which they predict speeds on average at least 1 km/h different to the observed speeds, whereas our new model does not. We also look at the RMSE values of the residuals, shown in Figures 5.12c and 5.12d. Although Figure 5.12b, appears to suggest that the existing functions may be better for predicting hiking speeds when traversing a hill at most slope values (as they have a lower absolute mean residual), this is refuted by 5.12d. The average residual may be close to 0, but the RMSE value is generally higher than those from our new model. We also note that over 90% of our data occurs on hill slopes of less than 20 degrees, and over 95% on walking slopes within ± 10 degrees. Within these regions, our model has a lower RMSE than the existing models.

These residual plots help to explain why predictions for total route times were not found to be greatly improved when using the our model compared to the existing hiking functions. The existing models' overestimations of speed walking downhill are cancelled out by underestimations of speed when walking uphill. The results from all of these comparisons give us confidence in our previous conclusion that the new model for predicting walking speeds is more accurate than the most commonly used existing walking speed functions.

We also looked further at the extreme residuals found in our data. Of the 20 residuals where the difference between predicted and observed speed was greater than 2 km/h, 18 were where the speed was being overestimated by the model. Furthermore, 17 of these occurred either next to a break point, or prior to an experimental region, where the group paused briefly. As noted in Section 5.2.1, this suggests that rather than the model being inaccurate, we were not fully identifying breaks, and a small amount of break was included in the next walk section, reducing the observed walking speed. The break-finding algorithm should be updated to counter this.

The fieldwork has confirmed that our new model is both successful at predicting walking speeds, and provides an improvement over existing models over the range of common walking slopes suggested by Proffitt et al. (1995).

5.3.2 Feasible Slopes

The next area we wished to explore was the extreme ends of the model, i.e. is the model accurate when the walking slope or hill slope are very high? In order to test this, two specific regions within the fieldwork area were selected, both with steep hill slopes. Before analysing the walking speeds attained in these regions, we noticed a disparity between the expected walking slopes and the measured walking slopes. During planning of the fieldwork, slope values of up to 37 degrees were identified. However, when participants were instructed to ascend or descend the slope directly, almost all of the calculated walking slope values were below 20 degrees, and none were above 30 degrees. This can be seen in Figure 5.13. Furthermore, the points highlighted in red in the figure (which include most of the steepest points) all came from a single track. This track is the Garmin watch data, where the time interval between consecutive points was greater than in the iPhone tracks. As discussed when investigating the break finding algorithm (Section 5.2.1), the high sampling rate caused more clustering to occur in slower sections of the route, namely on higher walking slopes. Not only did this clustering present a problem for identifying breaks, it also lead to artificially low walking slopes. When we merged datapoints into 50 m sections, the walking slope was calculated as the weighted average of the walking slope for each constituent point, weighted by point duration. The heavy clustering found in the steep sections of the route meant that we had a large number of points, all pointing in different directions, thus with very different walking slopes, leading to very low averages.

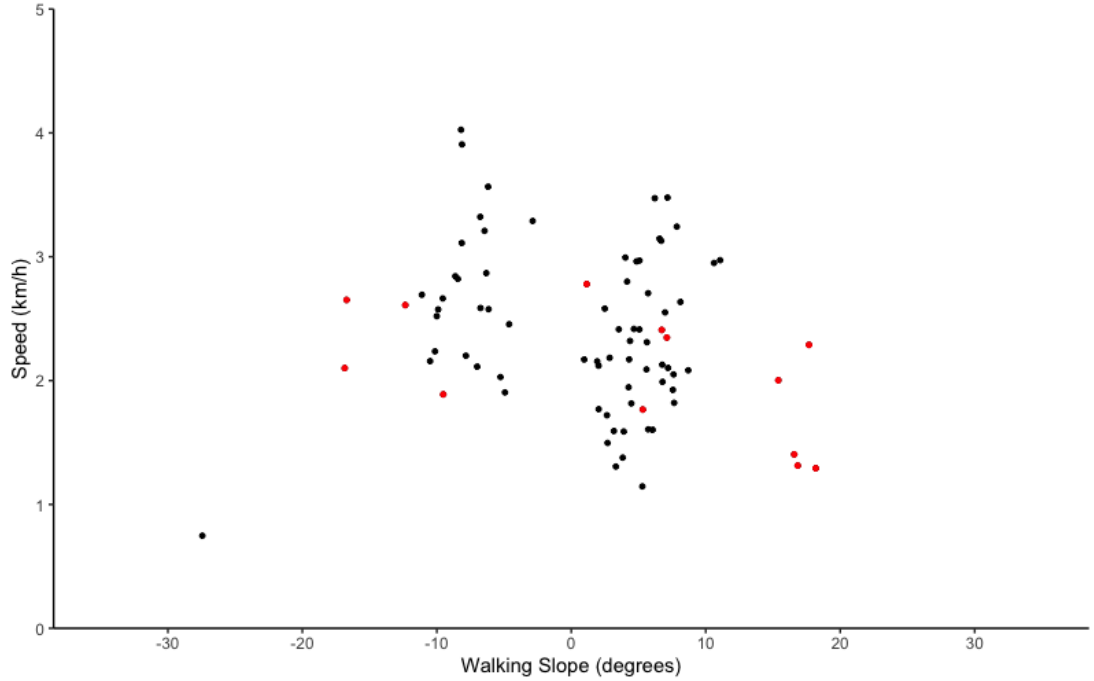


Figure 5.13: Calculated walking speeds and slopes when participants were instructed to directly ascend, or descend a steep slope. Points in red are those measured by a Garmin watch, the remainder were recorded on iPhones.

To counter this, we employed the same method as used to reduce the break overclassification found in Section 5.2.1; instead of taking every point, we implemented a 5 second delay between consecutive points in order to remove the excess clustering. The reduction this has to the clustering is shown in Figure 5.6, and the updated walking slopes can be seen in Figure 5.14. The walking slope values are clearly greater than before, and the red (Garmin) points are less isolated at the extremes, particularly when walking downhill.

While we have identified a scenario whereby walking slope values can be underestimated, we do not believe it will have affected much of our model data, due to the specific device settings required to generate such high levels of point clustering. However, even if over clustering did lead to underestimated walking slopes, we believe it was unlikely to affect the model predictions for two reasons:

1. Our model predictions for steep walking slopes will still be based on the unaffected data.
2. Our model predictions for shallow walking slopes will not be greatly affected by any slope underestimation, due to the quantity of valid data we had. Less than 4% of the model data had a hill slope of over 20 degrees, so the quantity of valid shallow-slope data will overcome any misclassification.

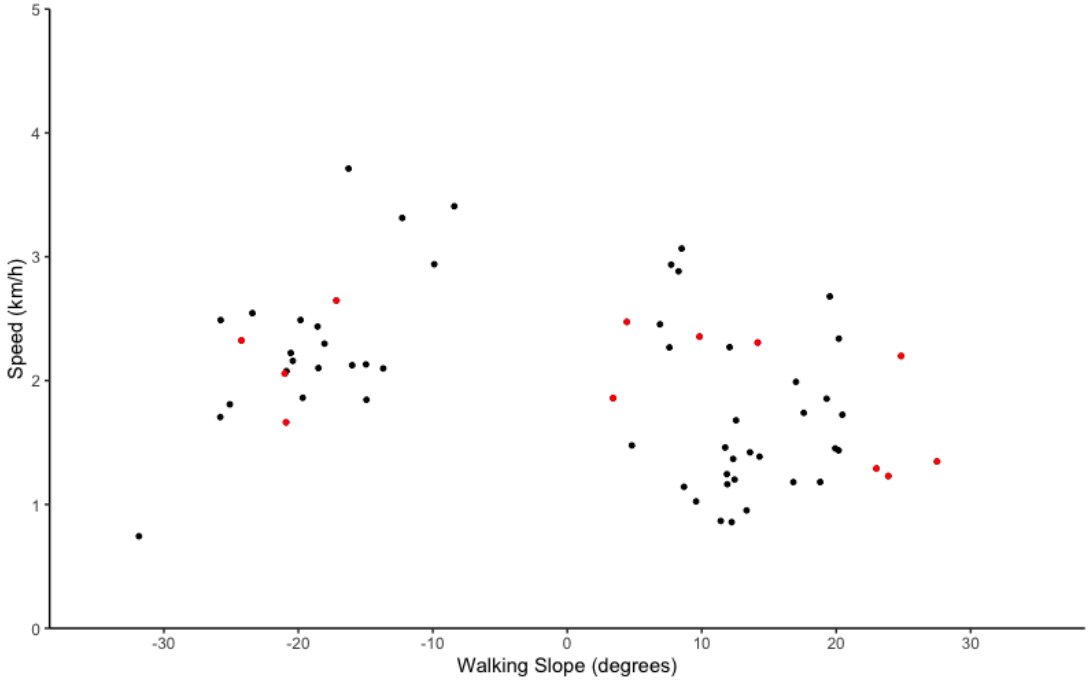


Figure 5.14: Calculated walking speeds and slopes when participants were instructed to directly ascend or descend a steep slope, when individual points have a minimum interval of 5 seconds before being merged into 50 m sections. Points in red are those measured by a Garmin watch, the remainder were recorded on iPhones.

Even after adding a 5 second interval to the data, the measured walking slopes were lower than we expected. Less than 10% of our data had a walking slope greater than 25 degrees, while over 70% of the hill slopes were over 25 degrees. Looking at a map of the region (Figure 5.15), we see that the areas of steepest slope are relatively narrow. It is therefore possible that the walking slope values in the steepest areas were being lowered, as they were averaged with data from surrounding, less steep areas. To ensure that this was not the case and that we could evaluate the model predictions for all available slope values, we reduced the data merge distance from 50m down to 25m. 25m was chosen as a compromise between lowering the distance, and the limitations of both the GPS timestamp precision and the 5m resolution of the elevation data (discussed in Section 3.1.3). The resulting walking slopes can be seen in Figure 5.16. We still do not see walking slopes as high as would be expected, and the proportion of walking slopes greater than 25 degrees remains very low (11%), so more exploration was required.

Figure 5.17 shows the hill slope, plotted against the walking slope, in the areas where participants were instructed to directly ascend or descend a hill. Added to the plot are lines indicating a direct ascent or descent (black, solid), as well as horizontal lines indicating a 28 degree (blue, dashed) walking slope. From this we can see that the walking slopes do not

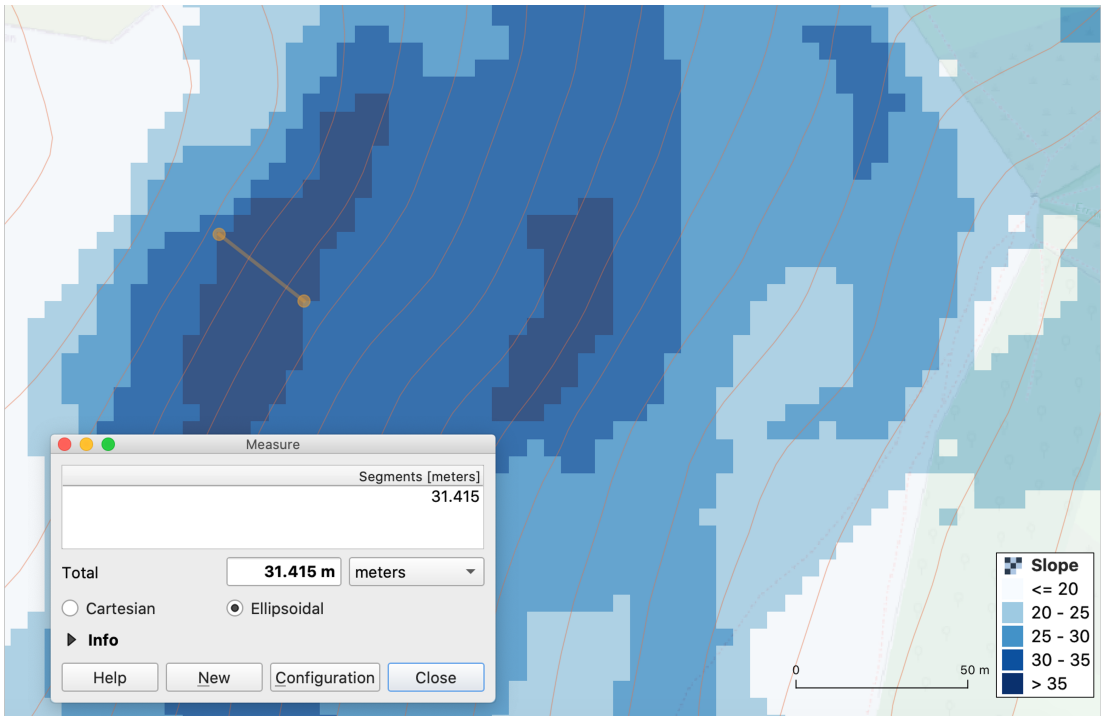


Figure 5.15: Map of the first high-slope experimental region, overlaid with 10 m contour lines and coloured based on hill slope values. Colours are in 5 degree bands, with the darkest blue indicating regions with a slope greater than 35 degrees. Background map from OpenStreetMap, visualised using QGIS (see 2.2.5).

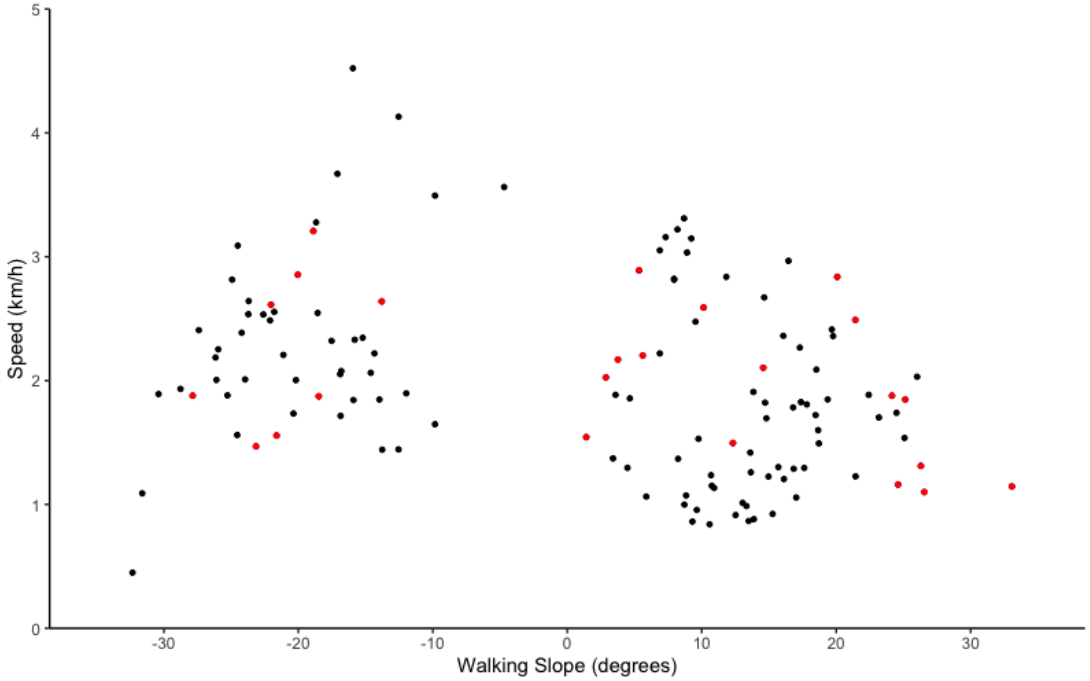


Figure 5.16: Calculated walking speeds and slopes when participants were instructed to directly ascend or descend a steep slope, where individual points have a minimum interval of 5 seconds before being merged into 25 m sections. Points in red are those measured by a Garmin watch, the remainder were recorded on iPhones.

match the hill slopes, and thus that the participants were not directly climbing the slope. This appears to get worse as the hill gets steeper, as the distance between the points and the black 'direct climb' lines increases; over 75% of our data with a hill slope over 30 degrees has a walking slope of under 25 degrees.

Figure 5.18 plots the average walking slope within hill slope bins (bin width 2 degrees). We can see how the walking slope 'levels off' and does not increase as the hill slope increases, after hill slopes of approximately 24 degrees. We can visualise the reason for the discrepancy between hill slope and walking slope by looking at the GPS traces (Figure 5.19). In this image the deviation between the GPS track and the track when merged into 25 m sections is clearly visible. This example shows how the participant appears to zig-zag somewhat while ascending the slope. While we believe that some of the deviation in the GPS track is caused by GPS error (and thus the walking slope is still slightly underestimated), the majority is legitimate, caused by the participants being unable to directly walk up the hill, and instead having to zig-zag in order to make it up or down the slope. While conducting the experiment, participants stated they had difficulty in maintaining a straight line, and there were instances where they had to deviate around a particularly steep region.

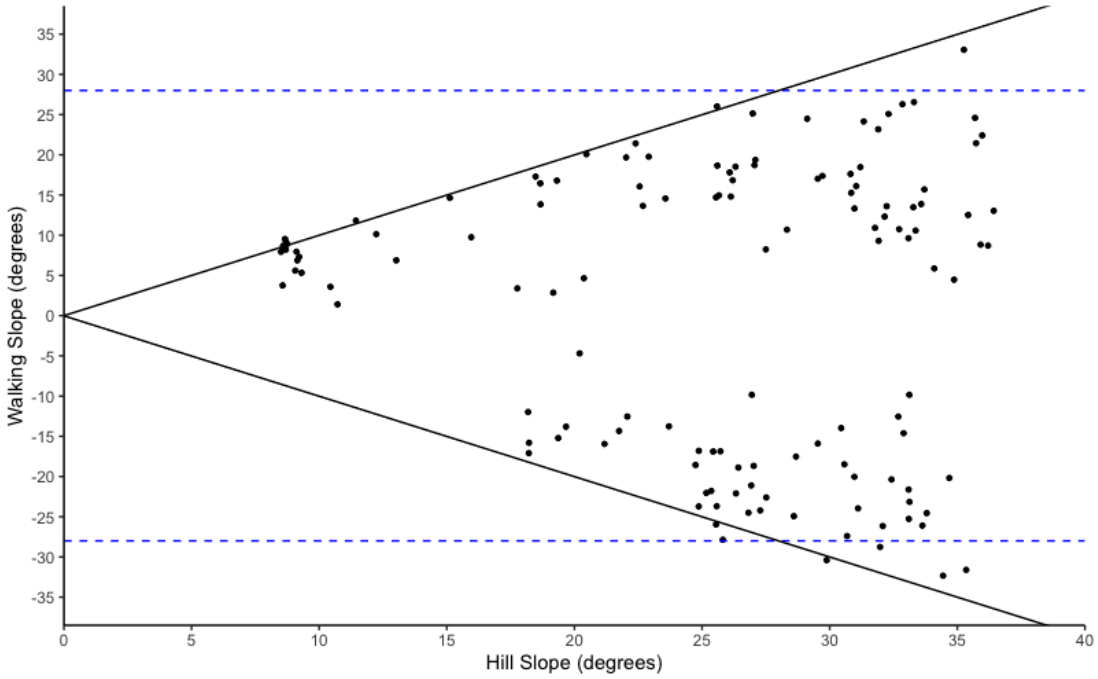


Figure 5.17: Calculated walking slopes plotted against hill slopes when participants were instructed to directly ascend, or descend a steep slope. The black solid lines indicate direct ascent or decent. Also shown is are blue dotted lines indicating a 28 degree walking slope.

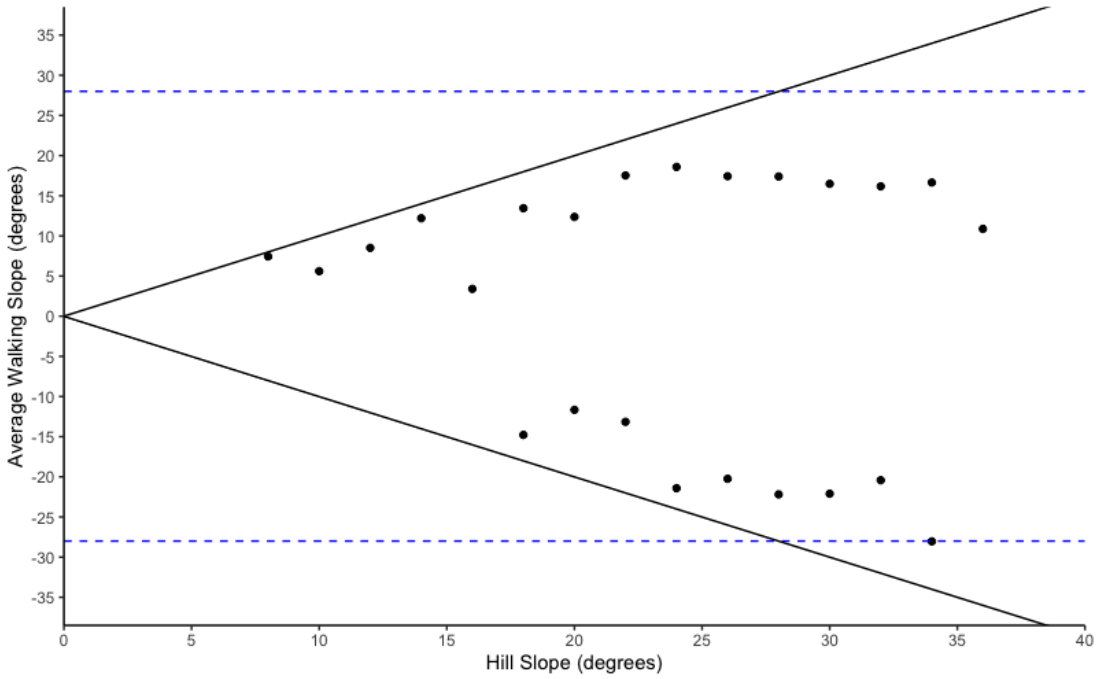


Figure 5.18: Average walking slopes plotted against hill slopes when participants were instructed to directly ascend, or descend a steep slope. Walking slopes are the average value found in hill slope bins of width 2 degrees. The black solid lines indicate direct ascent or decent. Also shown is are blue dotted lines indicating a 28 degree walking slope.

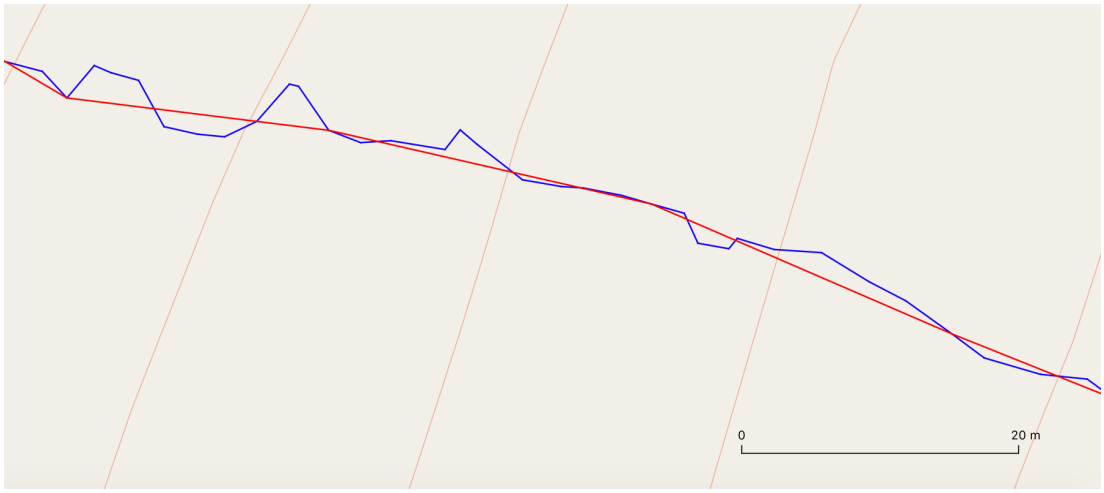


Figure 5.19: The difference between a GPS track with points at 5 second intervals (blue), and the same track when merged into 25m sections (red), during a section where the participant was asked to directly ascend the slope. 10 m contour lines are also shown to indicate hill slope. Background map from OpenStreetMap, visualised using QGIS (see 2.2.5).

Previously, we have discussed the critical gradient; the point at which it is faster to zig-zag rather than walk straight up or down a slope. Our model puts this figure between 14 – 16 degrees uphill and -16 – -18 degrees downhill (depending on path/obstruction conditions). This follows the existing work, which suggests the critical gradient occurs around 15 – 21 degrees. From this experiment, we suggest that not only is it faster to zig-zag up or down a steep hill, but there is also a point at which zig-zagging will occur regardless of intent. It is difficult to pinpoint the exact slope angle at which this occurs from our data here, but we believe it is around a 28 degree hill slope (indicated on Figures 5.17 and 5.18). There is some evidence to suggest that this angle may be slightly greater when walking downhill, as we do have a small number of data points with steeper walking slopes. Further work is needed to classify this in more detail. An experiment should be conducted on a larger scale, and with fixed gradients, rather than the changing hill slope which was encountered in this fieldwork.

We have now established a limit for the maximum feasible walking slope, but we also know that there must be a maximum hill slope on which it is possible to walk. When participants were climbing up the slope, there were a number of areas where they reported having to use their hands briefly, in order to aid travel up the hill. The maximum hill slope encountered during this experiment was approximately 37 degrees, and we suggest that these slopes are walkable, although only with relative difficulty. Furthermore, due to the difficulty the particpants encountered on these slopes, we suggest that the slopes encountered here are very close to the limit of what is feasible to ascend or descend safely.

Care should be taken when attempting to implement this limit into a practical application of our model, as the combination of terrain type and elevation data resolution is very important at steep slopes. A 37 degree hill slope implies 3.77 m of ascent over a 5m horizontal distance, which was the resolution of our elevation data. Our experiment took place on relatively smooth terrain (see Figure 5.20a), so we are confident that our 37 degree value was correct. However, at our resolution level a small, sheer cliff containing the entire elevation change would be recorded as having the same slope angle despite being much steeper in practice. To apply the slope restriction to a practical implementation of the model, a much finer elevation map would be required.

We have shown that it is possible to walk on slopes of up to 37 degrees, or possibly higher (though this is likely approaching the limit). However, walking directly up or down a slope becomes very challenging after values of approximately 28 degrees, and people will begin to zig-zag up the slope. Both of these factors have implications on the model in order for it to have practical use. For example, if a user requested the time taken to walk directly up a steep hill, should a distance and time be calculated assuming the user will zig-zag up the hill, because it would take significant effort to do otherwise? Taking the speed values from our model, a participant attempting to walk directly up a 35 degree slope for 1 km would take almost 4 hours. However, based on this experiment, the participant would naturally zig-zag regardless of intention at around 28 degrees. Our model estimates that doing this would take approximately 2 hours, with a distance of 1.37 km travelled. Furthermore, if the participant instead walked at the critical gradient of 16 degrees they could reach the top of the slope in approximately 1.5 hours, with a walking distance of 2.44 km. Further work must be undertaken to understand more about what the most important information is for hikers. A route going up or down steep slopes, calculated exactly as drawn on a map with straight lines, is likely to both underestimate the total distance travelled and overestimate the total walking time. An alternative to this could be to provide the walking time and distance, assuming that the user will zig-zag, although this distance may not match the route drawn on a map by the user.

5.3.3 Model Validation at High Hill Slopes

Once we had established the range of feasible walking slopes, we could look at the model predictions at high slopes. Lidar data was not available in the region, however, as we did when looking at standard walking conditions (Section 5.3.1), we were still able to use the model which specifies obstruction level, as we have observed the physical characteristics of the areas. Images of the two experiment regions can be seen in Figure 5.20, and it is clear that the first (Figure 5.20a) has minimal terrain obstruction, while the other (Figure 5.20b) is heavily obstructed by plantlife. 5 distinct experiments were taken: in the first region, the slope was traversed, ascended and descended; in the second, the slope was traversed and

descended. Participants began each experiment simultaneously, with starting positions 5-10 m apart in order to maximise the area that was covered and increase the range of slopes encountered. They were instructed to walk at a comfortable pace, and only deviate from a direct ascent/traversal if they were unable to continue, or felt it was unsafe to do so.

In Section 5.2.1, we noted that breaks were overclassified in the steepest regions due to slow walking speeds and excessive GPS drift. In order to analyse the data from these regions, all of the incorrectly identified breaks were manually removed. Furthermore, to both maximise the data available to us, and have the most accurate estimates for the walking slope values, a 5 second delay between points and a 25 m merge distance was applied to the data (as used in Section 5.3.2).

Traversing the Slope

Our experiment had 2 specific sections where steep slopes were traversed. Both of these were off-road, one in the light obstruction region, and the other in the heavy obstruction region. Figure 5.21 shows the walking slopes plotted against the hill slope. We can see that the majority of the data (96%) had a walking slope of under 10 degrees, with most (76%) under 5 degrees. We are therefore confident in stating that the participants were traversing the hill successfully. We compared the routes taken to a contour map of one of the regions to confirm this (Figure 5.22). The tracks clearly follow the shape of the hill, and there is very little travel perpendicular to the contour lines.

Figure 5.23 shows the walking speed plotted against the hill slope, and the points are coloured by obstruction amount. We can see from this that the obstructed slope generally had a lower gradient than unobstructed one. This was unavoidable due to the nature of the region where the experiments took place. To explore the difference between the two regions we grouped the data into bins of width 2 degrees, and found the average walking speed within each bin for each obstruction level (Figure 5.24). Generally our speeds are lower in the heavy obstruction areas, although the difference in speed between the two regions appears to diminish at the higher slopes. More work is needed to draw a fuller comparison, as we do not have large amounts of data in the region where the hill slope values overlap. Also plotted on both figures are the lines indicating the model predictions for traversing each of the slopes. Note that these show the predicted value for a walking slope of 0 degrees. In practice our walking slopes were rarely exactly zero, so we would expect the data to be below this speed, due to the negative effect walking slopes have on speed. The convergence which is seen in the observations is contrary to what our model suggests (as the model lines in Figure 5.24 do not converge), suggesting that more data is required in the high slope regions.

(a) The high slope experimental region with light terrain obstruction.

(b) The high slope experimental region with heavy terrain obstruction.

Figure 5.20: Images showing the two high slope regions used for experiments.

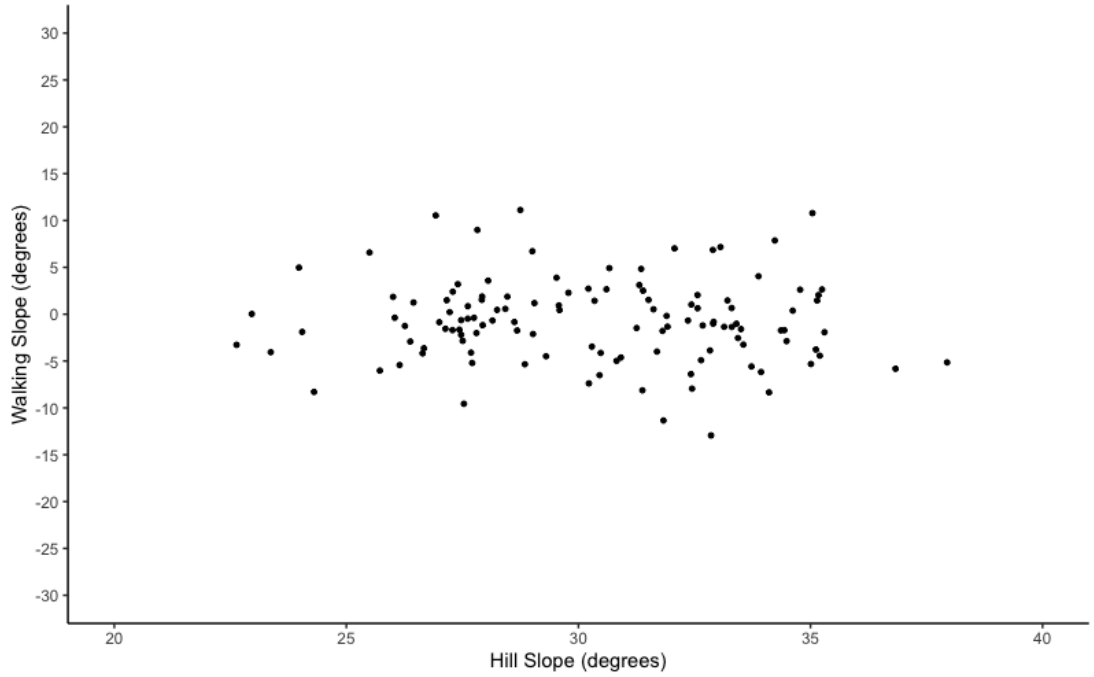


Figure 5.21: Calculated walking slopes plotted against hill slopes when participants were instructed to traverse a steep slope.

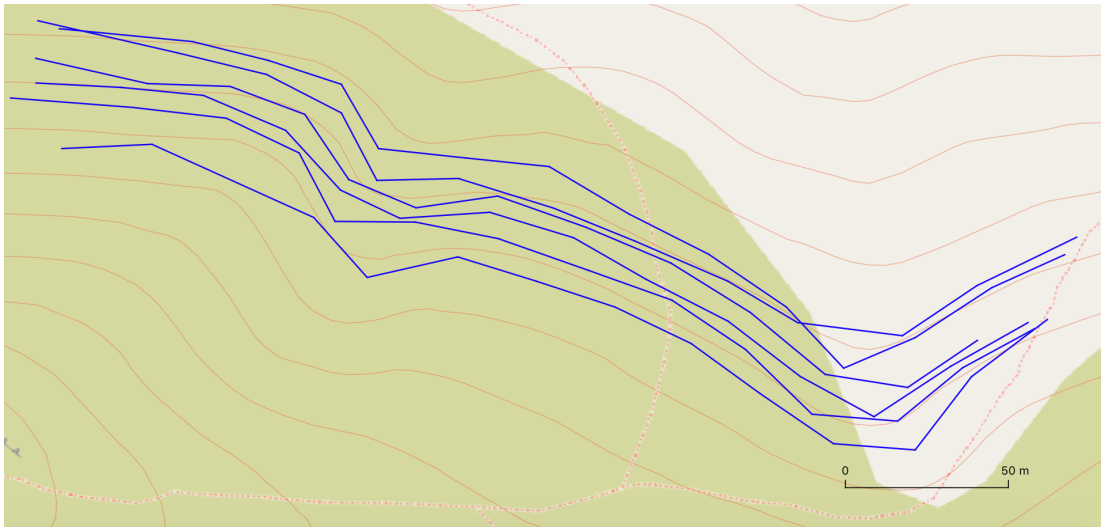


Figure 5.22: Map showing the GPS tracks (merged into 25 m sections) when participants were asked to traverse a steep slope. Contour lines with an interval of 10 m are also shown to indicate the hill slope and direction. Background map from OpenStreetMap, visualised using QGIS (see 2.2.5).

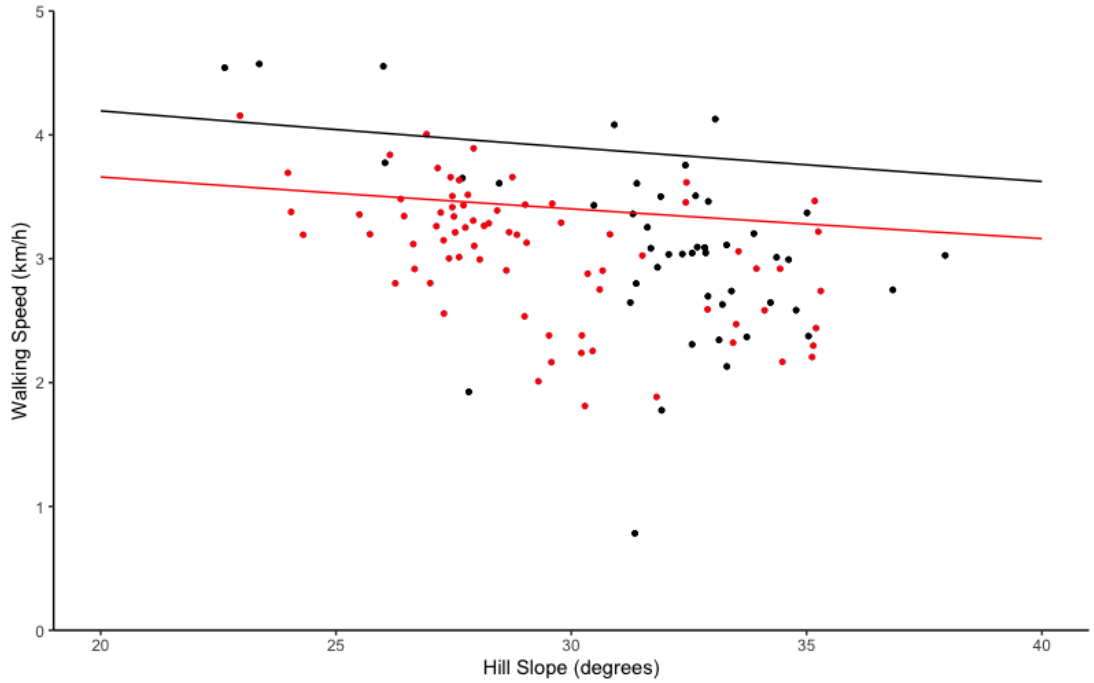


Figure 5.23: Walking speed plotted against hill slope for our Scout fieldwork data when traversing a steep hill. Points are coloured based on the obstruction level: heavy (red), or light (black). Also drawn are lines showing the walking speeds predicted by our model when traversing a heavy obstruction (red) or light obstruction (black) slope.

To investigate the ability of our model to predict the walking speeds, we looked the residuals (Figure 5.25). We can see that we appear, in general, to be overestimating the walking speed, in a similar manner to what we saw on steeper slopes in the Section 5.3.1. We also suggest that this underestimation gets worse as the hill slope increases, and this appears to be relatively consistent across both terrain types. Once again, we compared our model to the existing hiking functions. The experiments all took place in off-road conditions, so we applied the off-road corrections to both Naismith's rule and Tobler's function, as we did in Section 4.6 when investigating off-road regions. Figure 5.26 shows the residual plots under Naismith's rule (5.26a) and Tobler's hiking function (5.26b). Naismith's rule consistently underestimates the walking speed, and has a greater residual spread than we see in our new model. We confirm this by looking at the average and RMSE values for the residuals, seen in Figure 5.27. Contrary to what we saw in Figure 4.21d, our new model appears to overestimate the walking speed when traversing steep slopes, and this overestimation gets worse as the hill slope increases. This suggests that more data is required in this region to determine whether the participants were walking slower than average, or whether the model needs to be adjusted such that the hill slope has a greater negative effect on predicted walking speeds. Tobler's function initially underestimates the walking speed, but on steeper slopes it provides more accurate speed predictions than we see from our new model. However, it is important to note

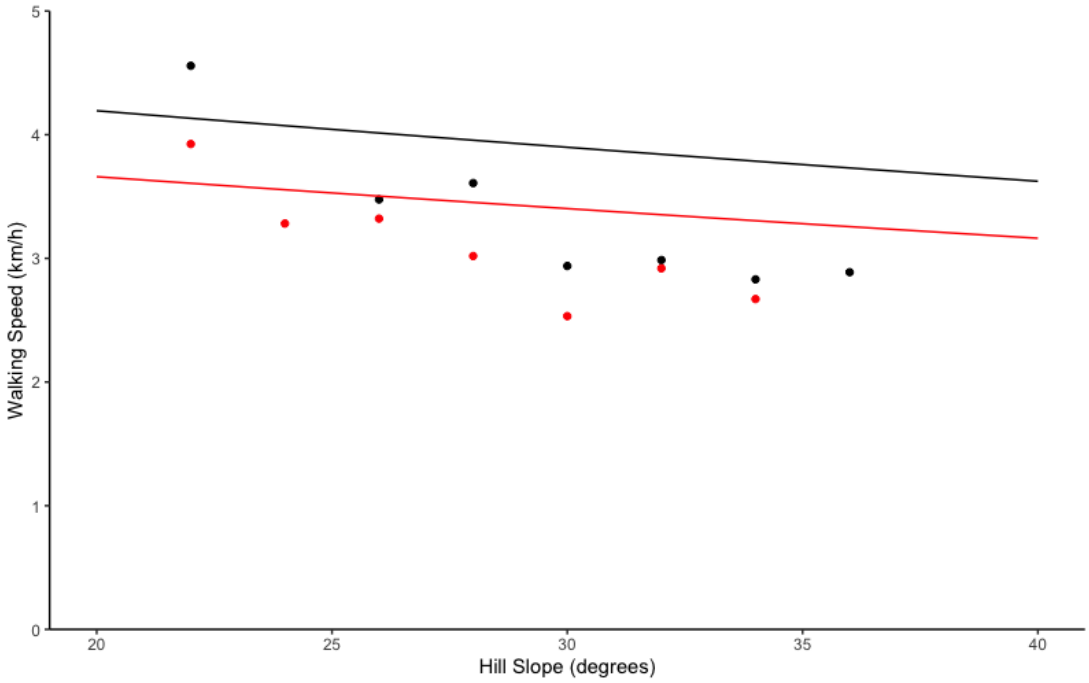


Figure 5.24: Average walking speed plotted against hill slope for our Scout fieldwork data when traversing a steep hill. Each point represents the average walking speed found in a hill slope bin of width 2 degrees. Points are coloured based on the obstruction level: heavy (red), or light (black). Also drawn are lines showing the walking speeds predicted by our model when traversing a heavy obstruction (red) or light obstruction (black) slope.

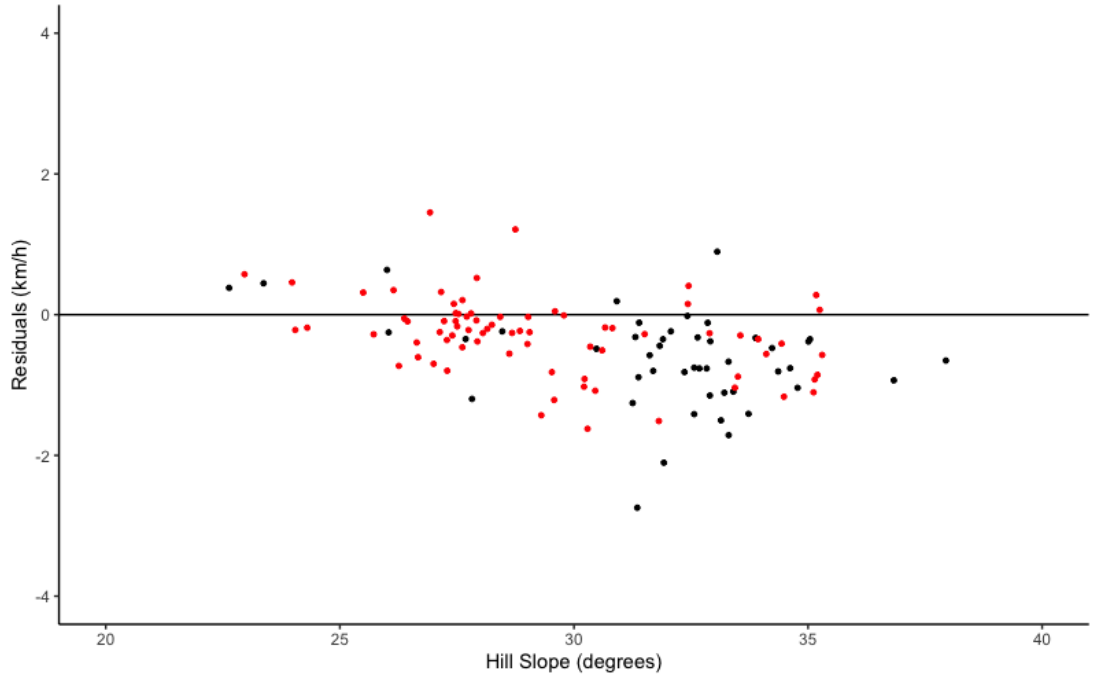
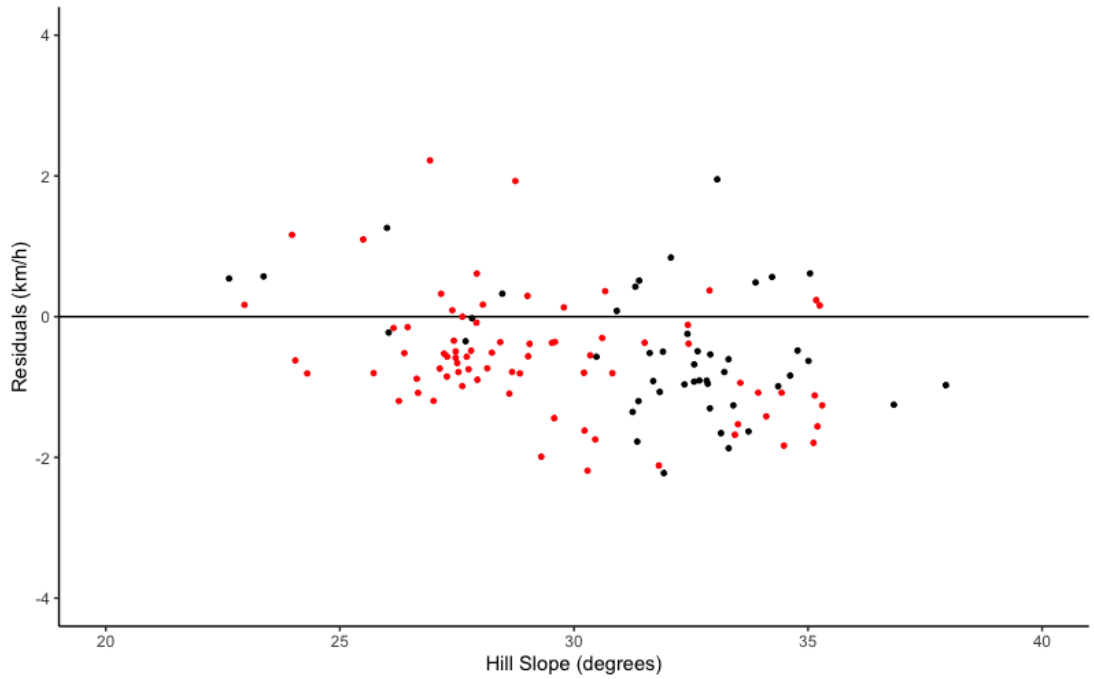


Figure 5.25: Model speed residuals plotted against hill slope for our Scout fieldwork data when traversing a steep slope. Points are coloured based on the obstruction level: heavy (red), or light (black).

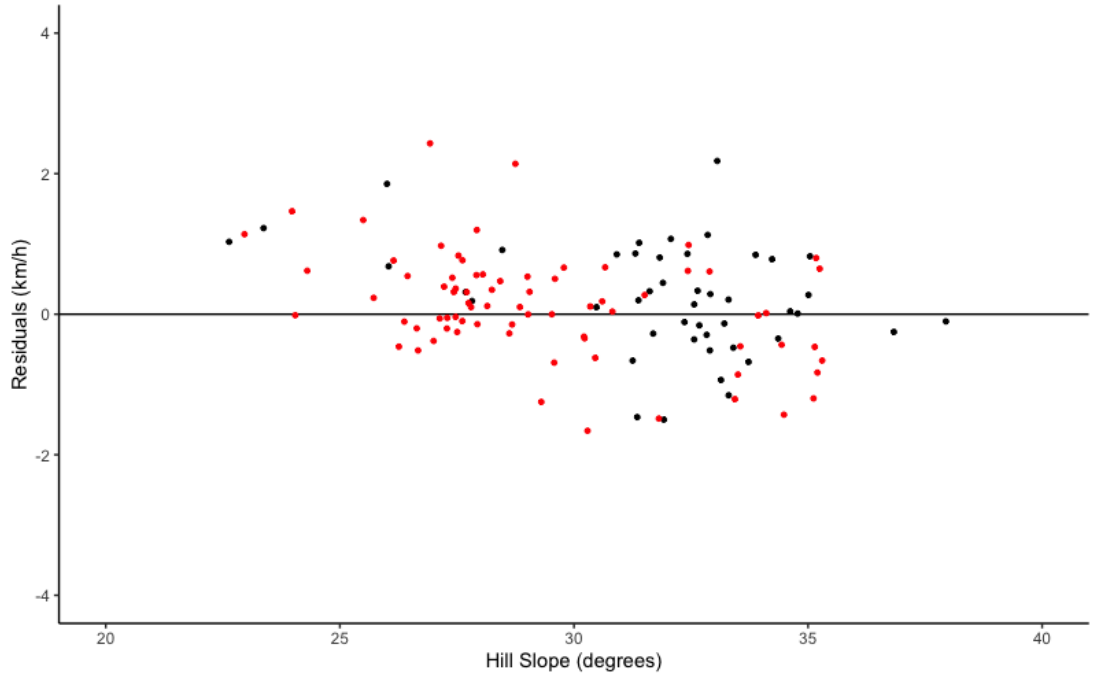
here that the predictions of Tobler's function are dependent only on the walking slope value. Therefore when traversing the hill in off-road conditions, Tobler's function always predicts a walking speed of approximately 3 km/h. From this data we suggest that when traversing an off-road slope, Tobler's hiking function is therefore only accurate once the hill slope is approximately 30 degrees.

Climbing the Slope

As discussed in Section 5.3.2, participants were unable to directly climb steep slopes, however, we could still compare the speeds attained when attempting to climb the slope, with those predicted by the observed hill slope/walking slope combination. The walking speeds attained are shown in Figure 5.28, coloured by obstruction level. Once more, we also show the model predictions for directly ascending or descending the slope. To explore the difference in walking speed across the two obstruction levels, we plotted the average speed seen at each walking slope value (bin widths of 2 degrees). This is shown in Figure 5.29. Heavy obstruction walking speeds are generally lower than those seen in light obstruction, although our limited volume of data makes this comparison difficult.

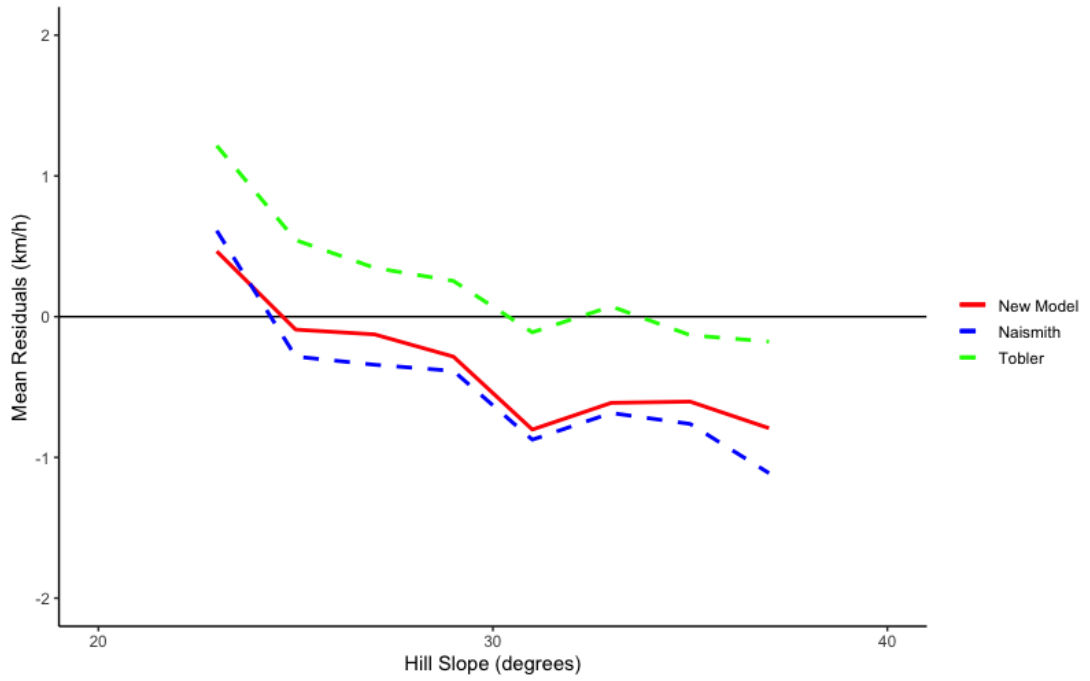


(a) Naismith speed residuals plotted against hill slope for our Scout fieldwork data when traversing a steep slope. Points are coloured based on the obstruction level: heavy (red), or light (black).

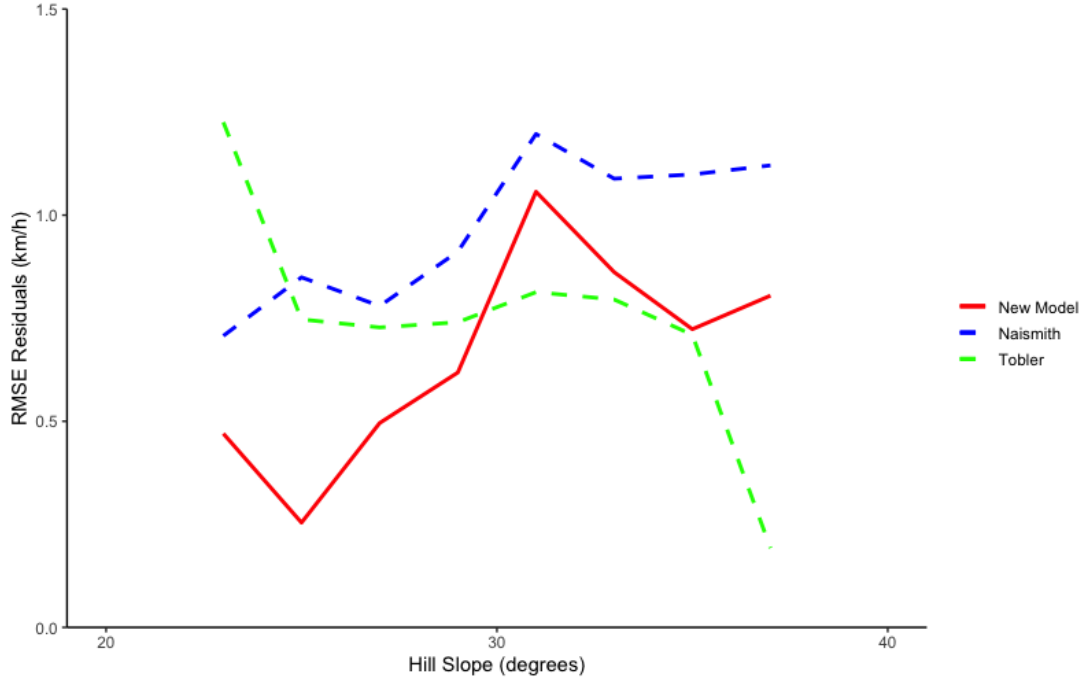


(b) Tobler speed residuals plotted against hill slope for our Scout fieldwork data when traversing a steep slope. Points are coloured based on the obstruction level: heavy (red), or light (black).

Figure 5.26: Plots showing the residuals of walking speeds predicted by Naismith's and Tobler's speed functions when traversing a steep slope.



(a) Comparison of mean values for the residuals of walking speed plotted against hill slope, when traversing a steep slope.



(b) Comparison of RMSE values for the residuals of walking speed plotted against hill slope when traversing a steep slope.

Figure 5.27: Comparison of mean and RMSE residual values for the new model (red), Naismith's model (blue) and Tobler's function (green), when traversing a steep slope.

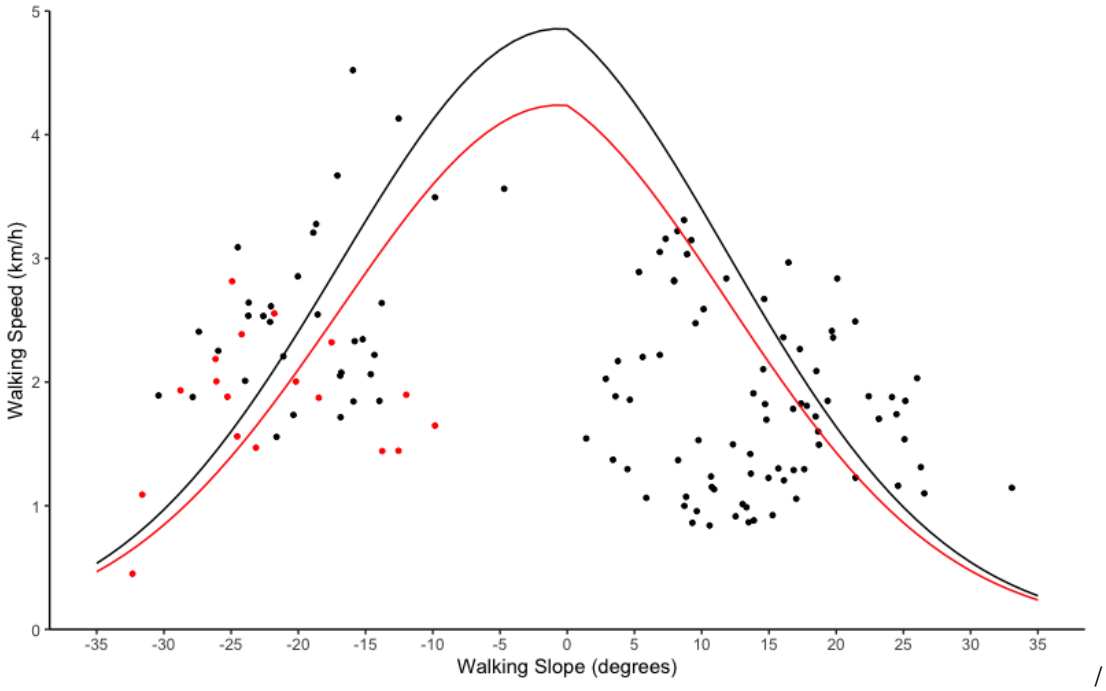


Figure 5.28: Walking speed plotted against walking slope for our Scout fieldwork data when ascending or descending a steep hill. Points are coloured based on the obstruction level: heavy (red), or light (black). Also drawn are lines showing are the walking speeds predicted by our model when directly ascending or descending a slope in heavy obstruction (red) or light obstruction (black).

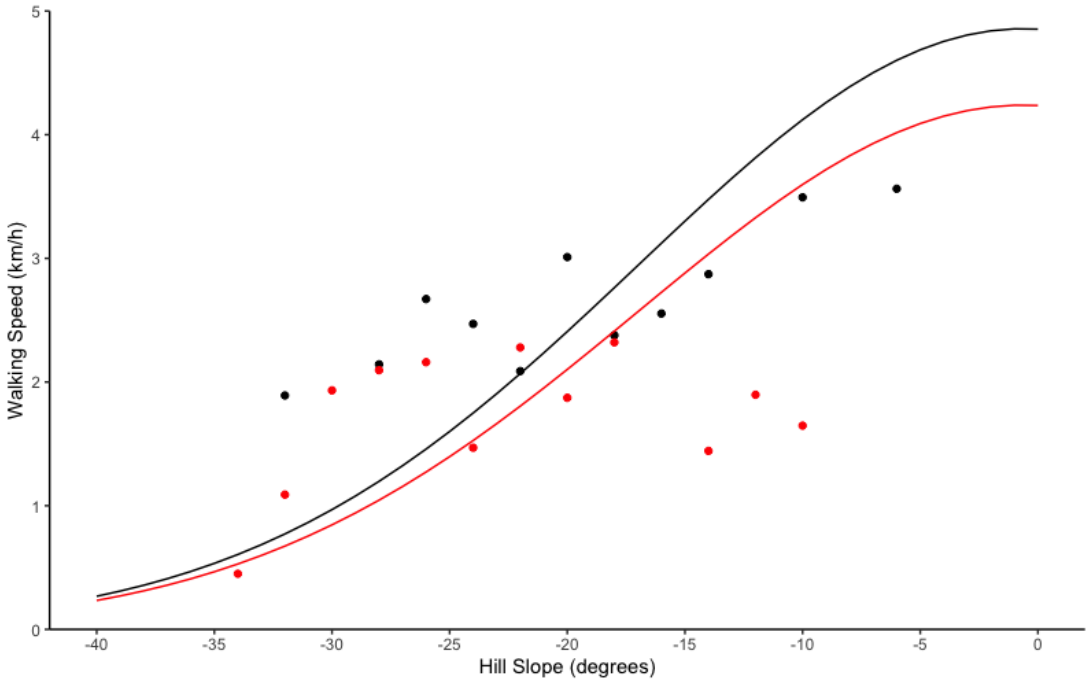


Figure 5.29: Average walking speed plotted against walking slope for our Scout fieldwork data when descending a steep hill. Each point represents the average walking speed found in a walking slope bin of width 2 degrees. Points are coloured based on the obstruction level: heavy (red), or light (black). Also drawn are lines showing are the walking speeds predicted by our model when directly descending a heavy obstruction (red) or light obstruction (black) slope.

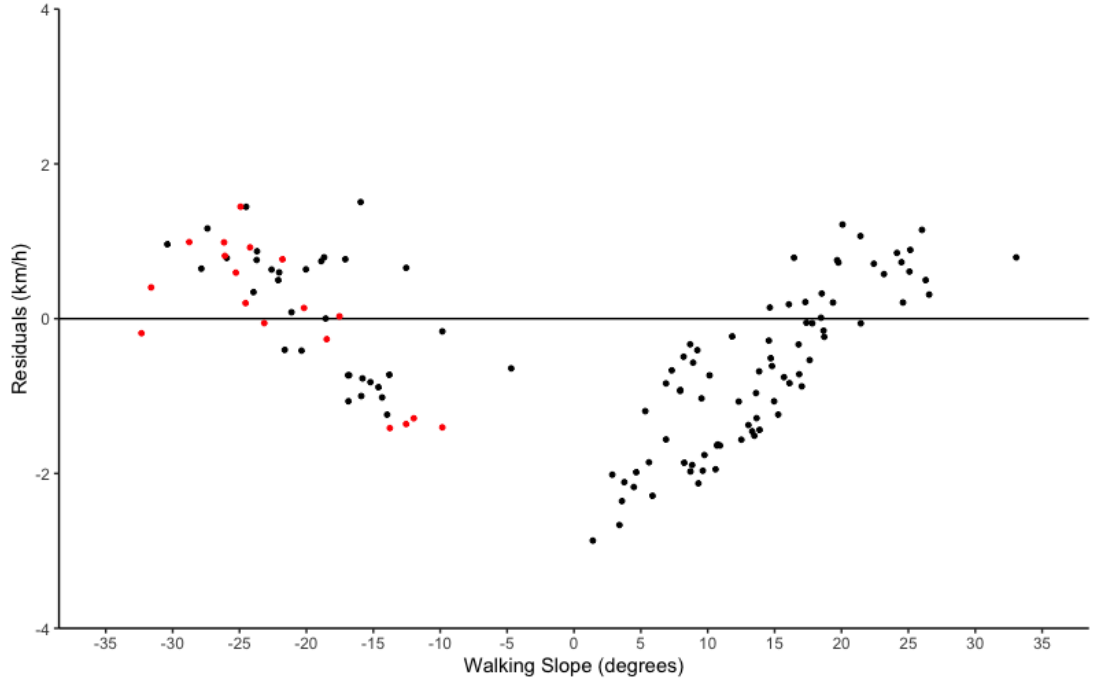


Figure 5.30: Model speed residuals plotted against walking slope for our Scout fieldwork data when ascending or descending a steep slope. Points are coloured based on the obstruction level: hevay (red), or light (black).

One of the main conclusions we can draw from Figure 5.28 is that we appear to be largely overestimating walking speeds at low walking slopes, and underestimating the walking speed at higher slopes. This is confirmed by the residual plot of the data (Figure 5.30). We found previously that our model was overestimating the walking speeds of the participants at higher hill slopes. Although we do not know whether this was a result of our participants walking slower than average, or our model understating the effect of hill-slope, it can explain a large amount of the overestimation seen at lower walking slopes in Figure 5.30. Points where the difference between hill slope angle and walking slope angle was over 15 degrees made up over 60% of all points where we were over-estimating the walking speed by at least 1 km/h. We suggest that the overestimation of walking speeds at these points was largely a result of the overestimation already identified on steep hill-slopes.

A second explanation for some of the overestimation at lower walking slopes can be found if we look further into the data. As the participants were nearing the summit of the hill climb, the hill slope reduced, thus reducing the walking slope. However, the participants were tired from the exertion of ascending the hill, and so walked slower than the model predicted (these points are those with hill slopes below approximately 15 degrees, seen in Figure 5.17). The model does not factor in how exertion will impact walking speeds. A break was taken after the climb, and subsequently the participants were able to once again walk at the predicted

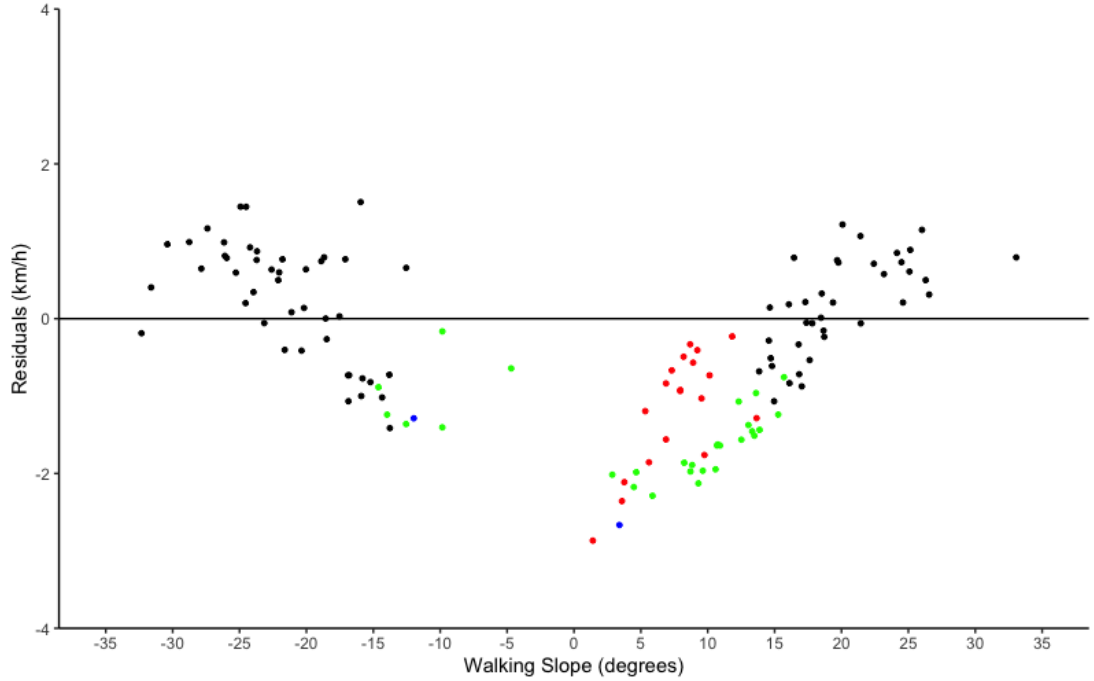


Figure 5.31: Model speed residuals plotted against walking slope for our Scout fieldwork data when ascending or descending a steep slope. Red points indicate those towards the end of the ascent, where the terrain flattened out. Green points are those where there is a greater than 15 degree difference between the hill slope and the walking slope. Blue points indicate those adjacent to a break.

speeds. Therefore, if exertion is a factor in a model, then it must only affect a short period of time after a particularly steep section. We also found a small number of points which we believe have slower walking speeds than expected due to them occurring either immediately preceding or following a break.

Figure 5.31 shows the residual plot recoloured such that all points in the categories described above are highlighted. The majority of the overestimation in walking speeds when climbing a slope has been explained by failings in the model (either not taking hill slope into account enough, or not allowing for a slowdown in speed after a period of exertion). However, the residuals also show that we underestimate the walking speeds on steeper slopes. This suggests one of two things: either that our model is overstating the drop in speed as the walking slopes reach their steepest values, or the group were faster than average in this section. As the participants were told specifically to climb/descend the steep hill, they may have approached it as a ‘challenge’, and thus pushed to walk faster in order to complete it. It may also be that this overestimation is balanced out by the underestimation we saw when the participants were near the summit of the hill. The fact that the participants walked faster than expected at the start of the steep ascent, may have forced them to walk slower than expected at the top of the slope as they tired out. A different group of participants may walk slower up the slope,

but subsequently be less tired and able to walk faster once the slope flattens out. Over the course of the whole hill, the estimated time to travel from the bottom to the top of the hill would be correct. To confirm this, we looked at the total time taken for the experiment for each participant, and found that on average we underestimated the time taken for the hill ascent by approximately 16.3%. However, if we acknowledge that our model may underestimate the impact of hill slope, and only consider points where the walking slope is within 15 degrees of the hill slope, then our walking time prediction now overestimates the observed time by an average of 0.25%. (Note here that the range of estimates was still between an underestimation of 29.9% and an overestimation of 36.3%).

While we have been able to come up with potential explanations as to why our model may be inaccurate on the steepest ascents, we have not explained the similar residual pattern seen when descending the hill. More data is required here, as our data and residuals suggest that when descending a slope, the walking speed in fact drops rapidly on walking slopes of between 10 and 15 degrees (we showed previously that our model is accurate at walking slopes of up to 10 degrees), before levelling off and remaining at approximately 2 km/h even at steeper slopes. Further experiments with more participants would allow us to explore the following ideas:

- A different group of people may respond differently to the same hill, initially walking slower in the steepest sections, but enabling them to walk faster once the summit was reached.
- The fact that the hill was relatively short (approximately 350 m horizontal distance) may have meant that our participants did not get tired until they had overcome the steepest parts of the route. A longer hill may produce some speed underestimation on high slopes once the participants tire.
- The model may be underestimating walking speeds on steep walking slopes. We did not have access to large amounts of very steep data when building the model, so repeating the experiment may show that the model must be updated for steep walking slopes.
- If more participants display the same speed decrease due to exertion as was seen here, then a tiredness factor should be added into the model.

Figure 5.32 shows the mean and RMSE residuals when comparing our model to the existing functions when ascending or descending a steep slope. We can see that, as expected, Naismith's rule overestimates walking speeds when walking downhill (due to the fixed 4 km/h speed). Comparing our model with Tobler's function, we see that our model has a lower RMSE value and lower absolute mean residual value than Tobler's function on the steepest slopes (walking slopes steeper than $\pm$20 degrees). This suggests that it is an improvement over the existing functions when walking on steep walking slopes. On the shallower ascents (0 – 15 degrees), both Naismth's rule and Tobler's hiking function appear produce more accurate walking speed predictions than our model. However, we have already shown in Section 5.3.1

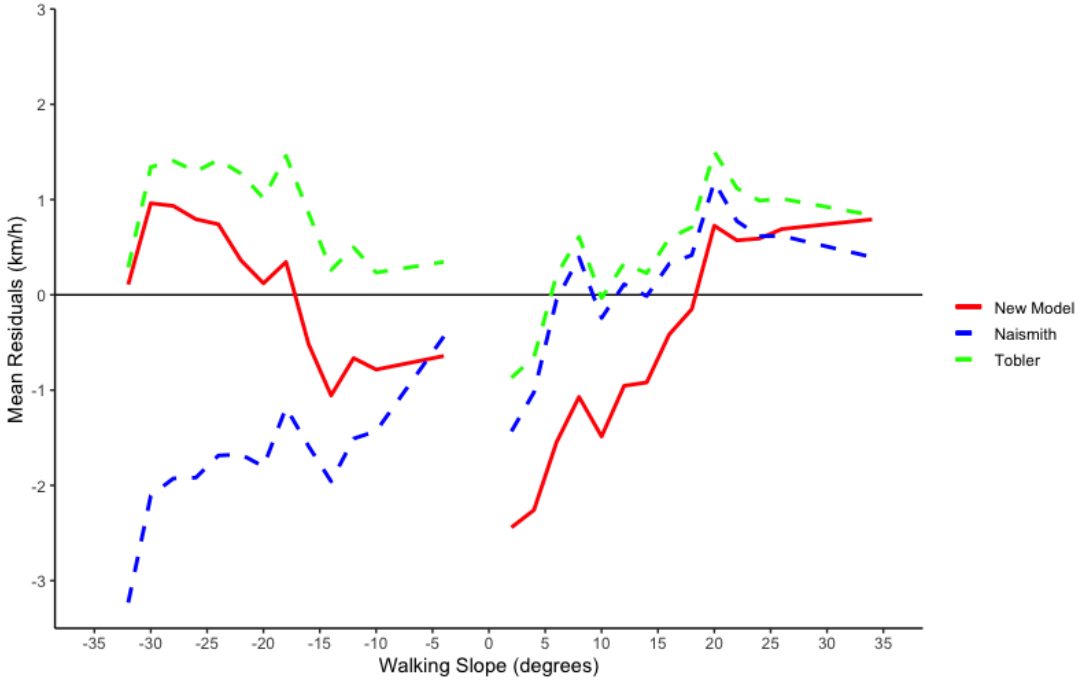
that our model is more accurate in this region under normal conditions. The discrepancy here shows that we need to repeat the fieldwork to determine whether an exhaustion factor should be applied to our model after climbing a steep slope, or whether this is not necessary as, on average, our speed predictions are accurate.

5.4 Discussion

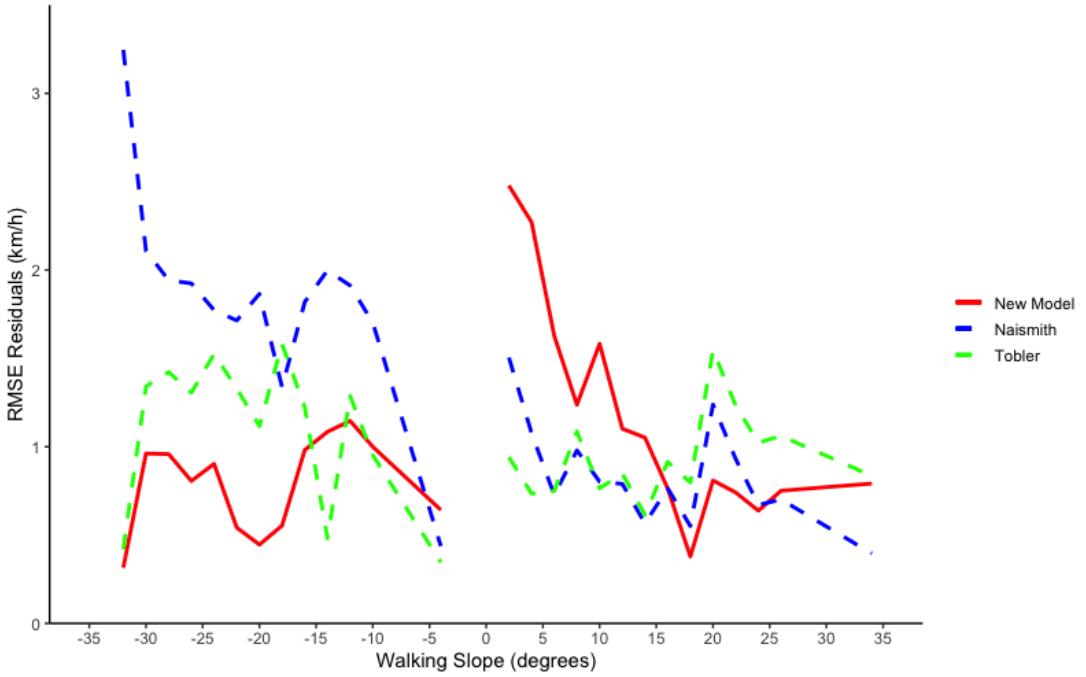
Overall, all of the aims of the fieldwork were met, and we were able to test the methods used in creating the model, as well as the model's prediction ability in both normal and extreme walking conditions. We have confirmed that our breakfinding algorithm is very accurate at finding the main breaks in hiking tracks, although a number of areas for improvement were identified. We have also seen how our current road or path identification is overclassifying points on a road. Similarly, we have noted that our filters to remove non-walking tracks, or track segments, are likely to be broadly successful, but further work should be conducted to ensure that slower driving sections are fully identified. We do not believe that classification and filtering errors which were noted will have affected the final walking speed model, due to the volume of data which was used to create the model.

One of the aims of the fieldwork was to establish the limits of our model; the position of the slope angles at which walking is no longer possible. While we were unsuccessful in this, as walking was possible on the steepest slopes measured here (37 degrees), we believe that this is very close to the upper limit. Furthermore, we did establish that there is a point, at around 28 degrees, at which walking directly up a hill is no longer possible, and zig-zagging will start to occur. This builds upon the idea of the critical gradient established by previous work, where it is faster to zig-zag up a slope once it reaches an incline of approximately 16 degrees.

In exploring the walking speed predictions of our model, compared to the existing methods, we found (as we did in Chapter 4) that our model provided greater accuracy than the existing speed functions, specifically under normal walking conditions (walking slopes of under 10 degrees). However, we also consistently overestimated the walking speeds on higher hill slopes, and this overestimation increases as the slope angle increases. More work needs to be conducted with a larger number of participants to determine whether the model needs refinement in this region, or if our group of participants walked at a slower than an average speed. When directly ascending or descending on (off-road) steep slopes, our model shows improvement over both Tobler's function and Naismith's rule. This is the area where most work is needed to understand the reasons for our error in estimation. Due to the low number of participants, our results here may be overfitting to a limited dataset. However, we have some evidence to suggest that walking speeds on the steepest slopes are not as low as the our model predicts, and also of the potential need to include an 'exhaustion' factor into the model, which reduces walking speeds following a difficult climb.



(a) Comparison of mean values for the residuals of walking speed plotted against walking slope, when ascending or descending a steep slope.



(b) Comparison of RMSE values for the residuals of walking speed plotted against walking slope, when ascending or descending a steep slope.

Figure 5.32: Comparison of mean and RMSE residual values for the new model (red), Naismith's rule (blue) and Tobler's function (green), when ascending or descending a steep slope.

One further point of interest to note is that during the fieldwork, measurements of the hill slope were taken using a clinometer. During the steepest sections, slope angles of up to 45 degrees were measured. This is steeper than we saw in the data (37 degrees), and may be a result of the inaccuracy in our elevation data; either through error in the data values, or a result of the data resolution (5 m) not being able to describe the detail of the slope. This would affect the accuracy of our model, but it would not necessarily affect an implementation if the same data sources used to create the model were also used when planning a route.

Chapter 6

Discussion

6.1 Project Conception

This project began as an exploration into finding the fastest route to use in an emergency situation when hiking. This required being able to accurately predict walking speeds in all scenarios. While exploring which method to predict walking speeds should be used, it quickly became clear that existing methods had all been built using relatively limited sample sizes, and that there were unexplored variables which could significantly affect the walking speed. Furthermore, the impact of the Covid-19 pandemic meant that the quantity of fieldwork required to sufficiently test a route planning method would not be possible. A decision was made to then focus entirely on the creation of a new model to better predict walking speeds, built upon large quantities of data, which could be used in the future in many different settings, including in calculation of the originally proposed escape routes.

6.2 Project Summary

In Chapter 3, we began our explorations into how crowdsourced GPS data could be harnessed to use in a new formula to predict walking speeds, using Scotland as a test area. Before any modelling could be done, we had to remove breaks and non-walking tracks from our dataset. As discussed in Section 3.1, we were not able to use any existing methods to identify breaks in each track. The majority of the existing methods are used to identify breaks in different modes of transport, where the differences in travel speed between moving and stationary are much more pronounced. As such, we developed new methods to identify and filter breaks and non-walking sections in hiking and walking tracks, and applied it to our data. We then used this filtered data to produce a walking speed model which expanded on existing models by including both walking slope and hill slope as input parameters. The resulting model for predicting walking speed was shown to be an improvement over the most commonly used existing methods (Section 3.4).

In Chapter 4, we looked to validate this model on an extended dataset by expanding our area of interest to the whole of the UK. When doing this we found that different models were required for Scotland and the rest of the UK (Section 4.2). We explored whether terrain type could explain this variation, and found that on paved roads, the two regions are comparable. However for unpaved roads and when off-road, our Scottish data would be a very extreme sample of the data seen from the rest of the UK (Section 4.4). Investigating this further, we found that the elevation profile for Scotland is very different to the ROUK, with much more data seen at higher elevations. The elevation difference itself is not a factor which explains the differences between the models, but it suggests that there could be other environmental factors which have not been taken into account.

We extended our ROUK model further by looking at the obstruction levels in off-road conditions using lidar data (Section 4.5). We showed a clear drop in walking speeds once 10 cm of terrain obstruction is present, and found a significant difference in walking speeds between the heavy obstruction and light obstruction regions. We used this to produce a final model for walking speed which takes into account all three variables which we originally wanted to focus on, namely the walking slope, the hill slope and the terrain obstruction level. This model showed a clear improvement over existing hiking functions, with a lower RMSE when predicting walking speeds (Section 4.6). We also briefly looked into how the walk length impacts the amount of breaks taken, and observed that breaks increase as walk length increases, up to walking time of approximately 5 hours, beyond which the total break length on a walk decreases (Section 4.7.2). This information could be very useful when planning hikes in future, as people would be able to estimate the required break time as well as the required walking time.

Finally, in Chapter 5 we were able to test our methods used to identify breaks and filter our tracks, and test our model predictions under controlled conditions. Our filtering methods were broadly successful, and we are confident that our model accuracy has not been impacted by the inclusion of invalid data, although several areas of future improvement were noted. Under standard walking conditions we found results similar to those seen in Chapter 4, namely that our new model performed better than existing hiking functions at predicting walking speeds (Section 5.3.1). We found a similar result at steeper, more extreme, slopes although our model may be underestimating the impact that very steep hill slopes have on reducing the walking speed (Section 5.3.3). We also looked to identify the limits of the regions where our model is valid. We suggest that walking slopes greater than 28 degrees are not possible, and the maximum possible hill slope is close to 37 degrees (Section 5.3.2). Both of these results were found using a very small sample size, so the fieldwork should be repeated again in future to further explore these limits.

Overall, the methods used to create our dataset were successful and could easily be generalised to filter and model other walking data. This would be useful for future studies which may wish to explore other aspects of hiking speeds. For example, participants could be asked to upload their GPS tracks to a data portal, along with additional metadata such as the weather conditions, or their previous hiking experience. Each track could be processed using the methods we have established here to remove breaks, allowing data to be collected without needing to monitor every participant's walk to check for breaks, or requiring them to manually note the timestamps of each break taken. This data could then be used to improve the model found here, by providing speed estimations which can account for more individual factors.

We believe that the methods could also be adapted to identify and filter cycling tracks, although some parameter tuning would be required. When cycling, you can get a much wider range of speeds than are seen when walking. The values we have used in our breakfinding methods were found through exploration of walking data, and therefore could not be directly applied to cycling data. For faster modes of transport, such as driving, existing work has shown that classifying the transport type based on movement speed works well, so while it may be possible to adapt our approach to these, it is not necessary.

Throughout this work the use of crowdsourced GPS data has been one of the biggest advantages. It allowed us to include much more data than would have been possible to collect manually, and from a wider variety of terrains and conditions. This gives us confidence that our model provides a reliable representation of an average hiker's walking speed, unlike previous methods (see Section 1.1) which were often based on very small sample sizes. However, the crowdsourced approach did also come with some significant limitations. While we had access to data from a wide variety of regions and individuals, we did not have access to any track metadata. Much of the original downloaded GPS data was from tracks which were not recording a walk or a hike. Websites such as Strava (Strava, Inc., 2022), which tag tracks by their activity, do not allow for bulk downloads of GPS data over a large area (likely due to GDPR laws). The OSM data was the only large scale and wide ranging dataset of GPS data available. Although we implemented a number of filtering methods to identify non-walking track sections (described in Sections 3.1.3 and 4.1), we still had points in our final dataset with relatively high speeds. We believe that our final model is accurate, due to the volume of data used which we are confident came from hikes or walks, but it is likely that a small amount on non-walking data was included in the modelling dataset.

As well as not knowing what activity our original data was recording, we also didn't have an objective way to determine when breaks were taken. We developed our breakfinding algorithm using a selection of tracks, and checked whether clusters which we identified visually were identified by the algorithm (Section 3.1.2). We did not perform a quantitative analysis on the breakfinding performance as we did not have access to any ground truth data regarding

when breaks were actually taken. Any analysis would first require us to manually identify the locations of breaks in the GPS tracks, which would still be a subjective identification method. In Section 5.2.1 we analysed our breakfinding method on a dataset where the breaks were known, and found a number of areas where it could be updated to improve classification.

A further limitation of our data came when we looked to separately model on-road and off-road scenarios. We did not have knowledge of whether the routes were following a road or path, and a combination of GPS drift and map error meant that we had to use a search radius around each data point to identify roads (Section 4.3.1). While doing this, we were aware that we were likely to be overclassifying the roads, and this was confirmed in our fieldwork data. While we believe that our model was robust to this overclassification (due to the volumes of valid on-road data used), the overclassification left us with a reduced number of off-road datapoints from which to predict off-road travel speeds.

The final limitation of our crowdsourced dataset was that it limited our available data to only regions where walking was feasible. While this means that we can be very confident in our model predictions at low walking slopes, the relative lack of data at high slopes may mean that our model is less accurate in these regions. Similarly, our off-road terrain data was heavily skewed towards regions of low terrain obstruction. While this itself gave us an indication that heavy obstruction regions are less navigable, it did not enable us to empirically estimate walking speeds for heavy obstruction regions. We attempted to explore the steepest slopes in our fieldwork, although that was only a very small study to gain an indication of the current model accuracy, and the limits of where walking is possible. In future work, a power analysis calculation should be done to identify the number of participants required to produce statistically significant results to both identify the limits of the range of slopes and obstruction values which are walkable, and also ensure that the walking speed model is accurate over this full range.

6.3 Future Work

There are a number of areas which we have identified for improvement in future work, in both the data processing and modelling areas.

- The breakfinding algorithm is currently successful at identifying all of the major breaks in a hike, but still has some areas where classification can be improved.
  - The algorithm should be updated to ignore zero-speed points when calculating the median speed and distance, in order to prevent tracks with relatively high volumes of breaks from being discarded.

  - Filtering methods to reduce GPS drift in the steepest, slowest sections should be explored. We originally chose not to filter the tracks, so as to ensure that break clusters were not smoothed out of the tracks. However, we showed in Section 5.2.1 that this can lead to overclassification of breaks. Methods already exist which could be tested to filter the tracks (Lee & Krumm, 2011), although care must be taken when applying these for multiple reasons. Our breakfinding method relies on identifying point clusters, and smoothing could prevent identification of some smaller clusters. Furthermore, some of the tracks were recorded on devices which automatically smooth out the tracks. We would need to take care not to oversmooth and remove detail from such track sections.
  - Explore reducing the minimum distance between breaks. This will prevent us from removing otherwise valid data, and would likely lead to the presence of more data in the most extreme areas, where more breaks are likely to be necessary.
  - Ensure that the full duration of breaks are captured by the algorithm. A large proportion of our most inaccurate predictions in the fieldwork occurred immediately preceding or following a break, indicating that breaks may not be fully captured.
  - Investigate reduction of the 30 second threshold for breaks which are 'necessary' for the hike.
  - Increase the minimum speed which is considered to be 'extreme' (currently 0.01 km/h). This would ensure that all short breaks where the user is still moving are identified, but calibration would be needed to ensure genuine slow sections are not misclassified.
  - Add a separate filtering method which can be used to specifically identify driving, or other non-walking sections of a route. Currently it would be possible for a very slow period of driving to pass through the filters. Methods to do this could involve a wider check of the surrounding points; if a slow run of points occurs amongst a series of faster points, then it is likely not to be a valid walk.
- Tagging of roads and paths currently overestimates the volume of data which is on a road or path. This could be improved by reducing the distance a GPS track must be from a path in order to tag it. Furthermore, work could look at using the overall shape of a route to determine whether paths are being followed. This would also make it easier to identify individual location errors where the device records a small number of points off-line.
- The fieldwork should be repeated, following a power analysis calculation to determine a statistically significant sample size. Furthermore, efforts should be taken to identify regions with greater quantities of steep terrain, which is relatively consistent in gradient. Our work was limited by the fact that the steepest slopes were only present in narrow bands 5.15. An ideal hill would be long enough such that if a participant initially walks up it 'too fast', then they tire and must slow down before reaching the summit, so a true

average speed for the gradient can be recorded. Furthermore, during our fieldwork, the participants were spaced out to increase the range of slope values which were encountered. However, this meant that each 'line' was only covered by a single participant. With a greater quantity of participants, each 'line' could be repeated multiple times by different individuals, to generate more accurate average speeds. These changes would enable us to confirm some of the ideas presented here:

- Are we correct in our belief that zig-zagging will occur on slopes of over 28 degrees, even if the particpant is attempting to ascend directly?
- We found that hill slopes of up to 37 degrees were walkable, and believe that this is close to the limit (due to participants having to use their hands occasionally). Steeper slopes should be tested until the limit is found.
- More levels of terrain obstruction should be tested to determine the limits of walkability. We have classified terrain into light or heavy obstruction, but we know from experience that there exists an obstruction level at which walking is no longer possible.
- Our model dataset had limited quantities of data on the steepest hill slopes, and we have some evidence from our fieldwork (Figures 5.12b, 5.27a) that our model is underestimating the reduction of walking speed as hill slope increases. If more data can be gathered on these steep hill slopes, then we can determine if the model should be improved, or if our model is correct on average when looking at a larger dataset.
- Similarly, we found that the model may underestimate the walking speed when participants were ascending the steepest slopes. By repeating the experiment we would be able to determine if this is a failure of the model, or if this overestimation is balanced out by the underestimation seen later in the experiment when the participants were tired.
- Descending slopes of between 10-20 degrees should be explored further. We found some evidence from our fieldwork to suggest that our model underestimates the rate of speed reduction in this region, while simultaneously underestimating walking speeds on the steepest descents.

6.4 Conclusion

Overall the goals of this project have been met. We were able to take a large dataset of GPS tracks and produce new methods to detect and filter breaks, or other non-walking sections. We used almost 93,000 km of walking and hiking data to produce a model to predict walking speeds which takes into account the walking slope, the hill slope and the level of terrain obstruction, all of which were shown to be highly significant factors. Currently, we suggest that this model is valid 'assuming walking is feasible in that region', however we have also explored

what the limits of this feasibility are, with a maximal walking slope identified and a maximal hill slope value proposed. We have shown that this new model is more accurate at predicting walking speed times than the most common methods, and believe its implementation would improve the performance of hiking route planners.

Naismith's rule is still a good rule-of-thumb to use when estimating the total walking time for a route, especially in situations where the calculation must be done by hand. However, the findings here can be used as an addition to Naismith's rule; it is likely that (under Naismith's rule) the predicted ascent time will be overestimated and the predicted descent time will be underestimated. It is not uncommon for hikers to contact one another when they reach the summit of a hill, and provide an estimated arrival time back at the campsite. Knowing that the descent will likely take longer than estimated by Naismith's rule will result in more accurate arrival estimations being given. Similarly, the knowledge of how the hill slope reduces walking speeds, or that just 10 cm of vegetation can reduce walking speeds by up to 0.6 km/h may well affect route choices made when out on a walk. For example, if a hiker is following a footpath, but can see from their map that the path forms a large curve then they can use our findings to decide whether it will be faster to travel off-road and cut the corner. On flat terrain with heavy levels of obstruction, our model suggests that such a short cut will be faster if the distance covered on the path is more than 15% longer than the off-road distance. Speed is not the only factor which would affect this decision, as safety and navigability are also important variables, but these results can help people make more informed choices when on a hike.

Finally, we have demonstrated that our model provides more accurate walking speeds than the existing methods in all scenarios, and particularly in off-road regions. Both Naismith's and Tobler's adjustments for off-road travel estimate much lower walking speeds than we have seen in our data. By providing improved walking speed predictions in these off-road regions, we have enabled more accurate calculations of the fastest route to or from any given location, which could save minutes in an emergency situation where every second is important.

Appendix A

Model coefficients using alternate datasets

Final model coefficients using the Scotland dataset

	a	b	c	d
Paved Road	1.558	-0.00365	-0.00952	-0.00212
Unpaved Road	1.534	-0.00595	-0.0116	-0.00212
Off Road	1.445	-0.0138	-0.0169	-0.00106

Final model coefficients using all available data (combined Scotland and ROUK datasets)

	a	b	c	d
Paved Road	1.580	-0.00377	-0.00734	-0.00220
Unpaved Road	1.580	-0.00575	-0.0105	-0.00232
Off Road (Obstruction Unknown)	1.518	-0.0104	-0.0105	-0.00157
Off Road (Light Obstruction)	1.591	-0.0104	-0.0105	-0.00157
Off Road (Heavy Obstruction)	1.458	-0.0104	-0.0105	-0.00157

Bibliography

Aitken, R. (1977). *Wilderness areas in Scotland* (Unpublished doctoral dissertation). University of Aberdeen.

Alvares, L. O., Bogorny, V., Kuijpers, B., de Macedo, J. A. F., Moelans, B., & Vaisman, A. (2007). A model for enriching trajectories with semantic geographical information. In *Proceedings of the 15th annual acm international symposium on advances in geographic information systems - gis '07* (p. 1). New York, New York, USA: ACM Press. Retrieved from http://portal.acm.org/citation.cfm?doid=1341012.1341041 doi: 10.1145/1341012.1341041

Andri et mult. al., S. (2022). DescTools: Tools for descriptive statistics [Computer software manual]. Retrieved from https://cran.r-project.org/package=DescTools (R package version 0.99.47)

Arnet, F. (2009). Arithmetical Route Analysis with examples of the long final courses of the World Orienteering Championships 2003 in Switzerland and 2005 in Japan. *Scientific Journal of Orienteering*, *17*, 3–21.

Balstrøm, T. (2002, 1). On identifying the most time-saving walking route in a trackless mountainous terrain. *Geografisk Tidsskrift-Danish Journal of Geography*, *102*(1), 51–58. Retrieved from https://www.tandfonline.com/doi/full/10.1080/00167223.2002.10649465 doi: 10.1080/00167223.2002.10649465

Bauer, C. (2013). On the (In-)Accuracy of GPS Measures of Smartphones. In *Proceedings of international conference on advances in mobile computing & multimedia - momm '13* (pp. 335–341). New York, New York, USA: ACM Press. Retrieved from http://dl.acm.org/citation.cfm?doid=2536853.2536893 doi: 10.1145/2536853.2536893

Biljecki, F. (2010). *Automatic segmentation and classification of movement trajectories for transportation modes* (Doctoral dissertation). doi: 10.4233/uuid:654587d2-6e93-4619-ab9a-29d95f843f35

Bing Maps. (2022). *Bing Maps Aerial View.* https://www.bing.com/maps/aerial.

Breheny, P., & Burchett, W. (2017). Visualization of regression models using visreg. *The R Journal*, *9*(2), 56–71.

Campbell, M. J., Dennison, P. E., & Butler, B. W. (2017). A LiDAR-based analysis of the effects of slope, vegetation density, and ground surface roughness on travel rates for wildland firefighter escape route mapping. *International Journal of Wildland Fire*, *26*(10), 884. Retrieved from http://www.publish.csiro.au/?paper=WF17031 doi: 10.1071/WF17031

Campbell, M. J., Dennison, P. E., Butler, B. W., & Page, W. G. (2019, 5). Using crowdsourced fitness tracker data to model the relationship between slope and travel rates. *Applied Geography*, *106*, 93–107. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0143622818307859 doi: 10.1016/j.apgeog.2019.03.008

Davey, R. C., Hayes, M., & Norman, J. M. (1994, 1). Running Uphill: An Experimental Result and Its Applications. *The Journal of the Operational Research Society*, *45*(1), 25. Retrieved from https://www.jstor.org/stable/2583947?origin=crossref doi: 10.2307/2583947

De Vries, S. I., & Sterkenburg, R. P. (2012). *Filtering GPS tracks: cluster detection, cluster classification and transportation mode classification* (Tech. Rep.). Retrieved from https://www.researchgate.net/publication/328131279 doi: 10.13140/RG.2.2.24503.98726

Dirk Stichling. (2021). *mytracks - the gps-logger.* Retrieved from https://apps.apple.com/us/app/mytracks-the-gps-logger/id358697908 (Version 7.3.1)

Dunn, M., & Hickey, R. (1998, 6). The effect of slope algorithms on slope estimates within a GIS. *Cartography*, *27*(1), 9–15. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/00690805.1998.9714086 doi: 10.1080/00690805.1998.9714086

D'eon, R., Serrouya, R., Smith, G. C., & Kochanny, C. O. (2002). GPS radiotelemetry error and bias in mountainous terrain. *Wildlife Society Bulletin*, *30*(2), 430–439. Retrieved from http://www.jstor.org/stable/3784501 doi: 10.2307/3784501

El-Rabbany, A. (2002). *Introduction to GPS: The Global Positioning System*. Norwood, MA: Artech House.

Environment Agency. (2017). *LIDAR Composite DSM 2017 - 2m [ASC geospatial data], Scale 1:8000, Tiles: England.* Using: EDINA LIDAR Digimap Service, (https://digimap.edina.ac.uk). Retrieved from https://www.data.gov.uk/dataset/fba12e80-519f-4be2-806f-41be9e26ab96/lidar-composite-dsm-2017-2m (Downloaded: 16-09-2021)

Environment Agency. (2020). *LIDAR Composite DTM 2020 - 2m [ASC geospatial data], Scale 1:8000, Tiles: England.* Using: EDINA LIDAR Digimap Service, (https://digimap.edina.ac.uk). Retrieved from https://www.data.gov.uk/dataset/a58f4e0d-27ba-440a-9a9c-274bc76500f5/lidar-composite-dtm-2020-2m (Downloaded: 16-09-2021)

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In E. Simoudis, J. Han, & U. M. Fayyad (Eds.), *Second international conference on knowledge discovery and data mining* (p. 226–231). AAAI Press. Retrieved from https://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=9E6FE78EE86D98E605DEDBAF0582CF09?doi=10.1.1.121.9220&rep=rep1&type=pdf doi: 10.5555/3001460.3001507

Harrell Jr, F. E. (2021). Hmisc: Harrell miscellaneous [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=Hmisc (R package version 4.6-0)

Hess, B., Farahani, A. Z., Tschirschnitz, F., & von Reischach, F. (2012). Evaluation of fine-granular GPS tracking on smartphones. In *Proceedings of the first acm sigspatial international workshop on mobile geographic information systems - mobigis '12* (p. 33). New York, New York, USA: ACM Press. Retrieved from http://dl.acm.org/citation.cfm?doid=2442810.2442817 doi: 10.1145/2442810.2442817

Hikr.org. (2021a). *Scotland Hiking Reports.* https://www.hikr.org/region518/ped/?gps=1. (Downloaded: 12-04-2021)

Hikr.org. (2021b). *United Kingdom Hiking Reports.* https://www.hikr.org/region516/ped/?gps=1. (Downloaded: 01-07-2021)

Irmischer, I. J., & Clarke, K. C. (2018, 3). Measuring and modeling the speed of human navigation. *Cartography and Geographic Information Science*, *45*(2), 177–186. Retrieved from https://www.tandfonline.com/doi/full/10.1080/15230406.2017.1292150 doi: 10.1080/15230406.2017.1292150

J, L. (2006). Plotrix: a package in the red light district of r. *R-News*, *6*(4), 8-12.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103). New York, NY: Springer New York. doi: 10.1007/978-1-4614-7138-7

Jones, K. H. (1998, 5). A comparison of algorithms used to compute hill slope as a property of the DEM. *Computers & Geosciences*, *24*(4), 315–323. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0098300498000326 doi: 10.1016/S0098-3004(98)00032-6

Kay, A. (2012, 4). Route Choice in Hilly Terrain. *Geographical Analysis*, *44*(2), 87–108. Retrieved from https://www.researchgate.net/publication/228719038_Route_Choice_in_Hilly_Terrain doi: 10.1111/J.1538-4632.2012.00838.X

Kuhn, M. (2020). caret: Classification and regression training [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=caret (R package version 6.0-86)

Langmuir, E. (1984). *Mountaincraft and leadership : a handbook for mountaineers and hillwalking leaders in the British Isles*. Edinburgh. Retrieved from http://www.worldcat.org/title/mountaincraft-and-leadership-a-handbook-for-mountaineers-and-hillwalking-leaders-in-the-british-isles/oclc/34730649

Lee, W.-C., & Krumm, J. (2011). Trajectory Preprocessing. In *Computing with spatial trajectories* (pp. 3–33). New York, NY: Springer New York. Retrieved from http://link.springer.com/10.1007/978-1-4614-1629-6_1 doi: 10.1007/978-1-4614-1629-6{_}1

Llobera, M., & Sluckin, T. J. (2007, 11). Zigzagging: Theoretical insights on climbing strategies. *Journal of Theoretical Biology*, *249*(2), 206–217. Retrieved from https://www.sciencedirect.com/science/article/pii/S0022519307003542?via%3Dihub doi: 10.1016/j.jtbi.2007.07.020

Merry, K., & Bettinger, P. (2019, 7). Smartphone GPS accuracy study in an urban environment. *PLOS ONE*, *14*(7), e0219890. Retrieved from https://dx.plos.org/10.1371/journal.pone.0219890 doi: 10.1371/journal.pone.0219890

Murdoch, D., & Adler, D. (2021). rgl: 3d visualization using opengl [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=rgl (R package version 0.106.8)

Naismith, W. W. (1892). Cruach Ardran, Stobinian, and Ben More. *Scottish Mountaineering Club Journal*, *2*. Retrieved from http://gdl.cdlr.strath.ac.uk/smcj/smcj009/smcj0090603.htm

Natural Resources Wales. (2016). *LiDAR Composite Dataset [ASC geospatial data], Scale 1:8000, Tiles: Wales.* Using: EDINA Digimap Ordnance Survey Service, (https://digimap.edina.ac.uk). Retrieved from http://lle.gov.wales/Catalogue/Item/LidarCompositeDataset?lang=en (Downloaded: 16-09-2021)

Open Source Geospatial Foundation Project. (2020). *QGIS.org.* Retrieved from https://qgis.org/en/site/

OpenStreetMap contributors. (2021a). *OpenStreetMap.org.* Using: Planet OSM regional extracts http://zverik.openstreetmap.ru/gps/files/extracts/europe/great_britain.tar.xz. (Downloaded: 01-07-2021)

OpenStreetMap contributors. (2021b). *OpenStreetMap.org.* Using: Planet OSM regional extracts: http://download.geofabrik.de/europe/great-britain.html. (Downloaded: 04-08-2021)

OpenStreetMap contributors. (2021c). *OpenStreetMap.org.* Using: Planet OSM regional extracts: http://zverik.openstreetmap.ru/gps/files/extracts/europe/great_britain/scotland.tar.xz. (Downloaded: 16-04-2021)

OpenStreetMap contributors. (2022). *OpenStreetMap.org.* https://www.openstreetmap.org.

OpenStreetMap Wiki. (2022a). *Accuracy — openstreetmap wiki,.* Retrieved from https://wiki.openstreetmap.org/w/index.php?title=Accuracy&oldid=2079309 (Online; accessed 31-October-2022)

OpenStreetMap Wiki. (2022b). *Key:highway — openstreetmap wiki,.* Retrieved from https://wiki.openstreetmap.org/w/index.php?title=Key:highway&oldid=2426496 (Online; accessed 9-November-2022)

Ordnance Survey (GB). (2020a). *GB National Grid Squares [SHAPE geospatial data], Scale 1:250000, Tiles: GB.* Using: EDINA Digimap Ordnance Survey Service, (https://digimap.edina.ac.uk). (Downloaded: 27-03-2021)

Ordnance Survey (GB). (2020b). *OS MasterMap Topography Layer [GeoDatabase data], Tiles: GB.* Using: EDINA Digimap Ordnance Survey Service, (https://digimap.edina.ac.uk). Retrieved from https://www.ordnancesurvey.co.uk/business-government/products/mastermap-topography (Downloaded: 06-10-2021)

Ordnance Survey (GB). (2020c). *OS Terrain 5 [ASC geospatial data], Scale 1:10000, Tiles: GB.* Using: EDINA Digimap Ordnance Survey Service, (https://digimap.edina.ac.uk). Retrieved from https://www.ordnancesurvey.co.uk/business-government/products/terrain-5 (Downloaded: 05-08-2021)

Palma, A. T., Bogorny, V., Kuijpers, B., & Alvares, L. O. (2008). A clustering-based approach for discovering interesting places in trajectories. In *Proceedings of the 2008 acm symposium on applied computing - sac '08* (p. 863). New York, New York, USA: ACM Press. Retrieved from http://portal.acm.org/citation.cfm?doid=1363686.1363886 doi: 10.1145/1363686.1363886

Pitman, A., Zanker, M., Gamper, J., & Andritsos, P. (2012, 9). Individualized Hiking Time Estimation. In *2012 23rd international workshop on database and expert systems applications* (pp. 101–105). IEEE. Retrieved from http://ieeexplore.ieee.org/document/6327410/ doi: 10.1109/DEXA.2012.51

Proffitt, D. R., Bhalla, M., Gossweiler, R., & Midgett, J. (1995, 12). Perceiving geographical slant. *Psychonomic Bulletin & Review*, *2*(4), 409–428. Retrieved from http://link.springer.com/10.3758/BF03210980 doi: 10.3758/BF03210980

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Rees, W. (2004, 4). Least-cost paths in mountainous terrain. *Computers & Geosciences*, *30*(3), 203–209. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0098300404000226 doi: 10.1016/j.cageo.2003.11.001

Schuessler, N., & Axhausen, K. W. (2009). Processing Raw Data from Global Positioning Systems Without Additional Information. *Transportation Research Record: Journal of the Transportation Research*, *2105*, 28–36. Retrieved from https://journals.sagepub.com/doi/pdf/10.3141/2105-04 doi: 10.3141/2105-04

Simon Gröchenig. (2019). *QGIS GPX Segment Importer.* GitHub Repository. Retrieved from https://github.com/SGroe/gpx-segment-importer

Soule, R. G., & Goldman, R. F. (1972, 5). Terrain coefficients for energy cost prediction. *Journal of Applied Physiology*, *32*(5), 706–708. Retrieved from https://www.physiology.org/doi/10.1152/jappl.1972.32.5.706 doi: 10.1152/jappl.1972.32.5.706

Strava, Inc. (2022). *Strava.* https://www.strava.com/.

Tobler, W. (1993). Three Presentations on Geographical Analysis and Modelling. *National Center for Geographic Information and Analysis Technical Report*, *93*(1). Retrieved from http://www.geodyssey.com/papers/tobler93.html

Tsui, S., & Shalaby, A. (2006, 1). Enhanced System for Link and Mode Identification for Personal Travel Surveys Based on Global Positioning Systems. *Transportation Research Record: Journal of the Transportation Research Board*, *1972*, 38–45. Retrieved from http://trrjournalonline.trb.org/doi/10.3141/1972-07 doi: 10.3141/1972-07

U of Edinburgh. (2022). *Edinburgh Compute and Data Facility web site.* www.ecdf.ed.ac.uk.

Wan, N., & Lin, G. (2016, 12). Classifying Human Activity Patterns from Smartphone Collected GPS data: A Fuzzy Classification and Aggregation Approach. *Transactions in GIS*, *20*(6), 869–886. Retrieved from https://onlinelibrary.wiley.com/doi/10.1111/tgis.12181 doi: 10.1111/tgis.12181

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, *21*(12), 1–20. Retrieved from http://www.jstatsoft.org/v21/i12/

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H. (2019). stringr: Simple, consistent wrappers for common string operations [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=stringr (R package version 1.4.0)

Wickham, H., François, R., Henry, L., & Müller, K. (2021). dplyr: A grammar of data manipulation [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=dplyr (R package version 1.0.5)

Wilke, C. O. (2020). cowplot: Streamlined plot theme and plot annotations for 'ggplot2' [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=cowplot (R package version 1.1.1)

Wood, N. J., & Schmidtlein, M. C. (2012, 6). Anisotropic path modeling to assess pedestrian-evacuation potential from Cascadia-related tsunamis in the US Pacific Northwest. *Natural Hazards*, *62*(2), 275–300. Retrieved from http://link.springer.com/10.1007/s11069-011-9994-2 doi: 10.1007/s11069-011-9994-2

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R* (2nd ed.). Chapman and Hall/CRC.

Zandbergen, P. A. (2009, 6). Accuracy of iPhone Locations: A Comparison of Assisted GPS, WiFi and Cellular Positioning. *Transactions in GIS*, *13*(SUPPL. 1), 5–25. Retrieved from https://onlinelibrary.wiley.com/doi/10.1111/j.1467-9671.2009.01152.x doi: 10.1111/j.1467-9671.2009.01152.x

Zeileis, A. (2006). Object-oriented computation of sandwich estimators. *Journal of Statistical Software*, *16*(9), 1–16. doi: 10.18637/jss.v016.i09

Zeileis, A., & Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News*, *2*(3), 7–10. Retrieved from https://CRAN.R-project.org/doc/Rnews/

Zevenbergen, L. W., & Thorne, C. R. (1987, 1). Quantitative analysis of land surface topography. *Earth Surface Processes and Landforms*, *12*(1), 47–56. Retrieved from https://onlinelibrary.wiley.com/doi/10.1002/esp.3290120107 doi: 10.1002/esp.3290120107

Zhou, C., Jia, H., Juan, Z., Fu, X., & Xiao, G. (2017, 8). A Data-Driven Method for Trip Ends Identification Using Large-Scale Smartphone-Based GPS Tracking Data. *IEEE Transactions on Intelligent Transportation Systems*, *18*(8), 2096–2110. Retrieved from http://ieeexplore.ieee.org/document/7778251/ doi: 10.1109/TITS.2016.2630733