A LARGE SCALE METHOD FOR EXTRACTING GEOGRAPHICAL FEATURES ON BUS ROUTES FROM OPENSTREETMAP AND ASSESSMENT OF THEIR IMPACT ON BUS SPEED AND RELIABILITY

L. Dunne[1*], G. McArdle[1]

[1]School of Computer Science, University College Dublin, Belfield, Dublin 4, Ireland -
laura.dunne2@ucdconnect.ie, gavin.mcardle@ucd.ie

Commission IV, WG IV/9

KEY WORDS: GIS, OSM, Spatial Features, Geographical Features, Public Transport, Bus Speed, Bus Reliability, Bus Route

ABSTRACT:

Geographical features on bus routes impact a bus's performance, and as a consequence affect human mobility through cities. Analysis of these geographical features is non-trivial because they often must be manually recorded, limiting the ability to extract these features on a large scale. This paper proposes a novel method of extracting features from crowd-sourced OpenStreetMap (OSM) data and compares this method to the ground truth data for 539 stop pair segments in Dublin, Ireland. This paper also proposes algorithms to detect turns and the direction taken by buses at roundabouts, using the angle between points on the segment lines. Statistical analysis was performed, and elastic net linear regression models were developed with a subset of the route features to show their effect. The results show over 97% accurate identification of most individual features using the novel technique, with most errors resulting from OSM quality issues. The features that most negatively affected the average speed and reliability of the bus with statistical significance ($p < 0.025$) were: retail land use, turns, traffic lights, and roundabouts. The average speed limit and the length of the segment had a positive impact on the average speed but not on the reliability. This method can be used with any bus performance metric to obtain a deeper understanding of the dynamics of bus travel, provide detailed information for bus travel time simulations and more accurately predict bus journey times to improve scheduling on the overall bus network.

1. INTRODUCTION

Making bus transport an attractive option to commuters is an important component of the global drive towards modern, smart and sustainable cities. Accurately predicting bus journey times is essential for scheduling bus services, and bus reliability is an important indicator of a bus route's performance and passenger satisfaction (Dastjerdi et al., 2019). Various metrics are used to measure bus performance depending on the application. Average speed and reliability are commonly used metrics (Hu and Shalaby, 2017). There are several variations of the reliability metric, and can either refer to variability in travel time or to degree of deviation from a timetable (El-Geneidy et al., 2011). A related metric is headway variability. Headway is the time interval between two consecutive buses on the same route (Soza-Parra et al., 2021). When headway variability on a route is poor, the extreme case is bus bunching, when buses arrive at a stop in very close succession (Chioni et al., 2020). Geographical features such as traffic lights or bus lanes are known to impact upon these metrics of bus performance and optimising these geographical features can even increase the use of public transport (Arasan and Vedagiri, 2010). However, analysis of these features is non-trivial because they are often manually recorded, limiting the ability to extract these features on a large scale. Many of the analyses are limited to a small number of bus routes. Small sample sizes make it difficult to detect patterns due to the highly correlated nature of geographical data and the complex environment buses operate in.

Some of the earliest work looking at factors impacting bus performance dates from the 1970s (Sterman and Schofer, 1976) and 1980s (Abkowitz and Engelstein, 1983). These studies were conducted before Automatic Vehicle Location (AVL) data was available, and both the factors being studied and the bus travel times had to be collected manually. This data collection was prohibitively expensive to do at a large scale, and many of these studies were very small. AVL data has now been available for many years, and historical or real-time bus journey times are commonly available, but collecting the geographical data remains challenging. There are some recent studies in this area that highlight this continuing limitation. Feng et al. (2015) quantified the joint impacts of stop locations, signalised intersections and traffic conditions on bus travel time on a 41-mile urban stretch of road with 21 stop pair segments. Hu and Shalaby (2017) evaluated multiple features on two bus routes in Toronto, Canada, with the geographical data provided by the City of Toronto. Cui et al. (2019) analysed the impact of driveway density, bus volume, the number of bus routes, bus stop density and traffic signals on 180 road segments. The geographical data in this study was obtained from field surveys and Baidu Street Maps. Almeida et al. (2022) state that the spatial characteristics of the 36 segments in their study were collected from Google Maps. They evaluated the impact of the number of traffic lanes, bus lanes, land use zones, bus stops and traffic signals on bus' speed. Similarly, Kaewunruen et al. (2021) also examined 36 segments in Birmingham, UK, while evaluating the impact of segment length, pedestrian crossings, and intersections with traffic lights on bus reliability. However, it is unclear where the geographical data is coming from in this study. Soza-Parra et al. (2021) looked at the entire bus network in Santiago, Chilli in their analysis of factors that affect headway variability but the geographical features examined were limited to segregated bus lanes and traffic lights. The location of these

* Corresponding author

features was provided by the local public transport metropolitan area and the local metropolitan traffic control centre, respectively. Chioni et al. (2020) look at 360 segments in Athens, Greece, in their study to evaluate factors that cause bus bunching. The geographical features evaluated were: traffic signals, bus lanes, the position of the bus stops, number of lanes, number of routes and land use, but it is unclear how the data on the factors was obtained. Lyan (2021) extracted factors such as traffic lights, stops, and roundabouts for 1400 segments from OSM data but does not validate the accuracy of the results compared to a manual survey or use any proxy measures of quality. The geographical data in this study also comes from OSM, the most successful crowd-sourced geographic database (Bertolotto et al., 2020). While the aim of OSM is to produce a free world map, it has been used as a datasource in academic research (Grinberger et al., 2022). The quality of OSM data and possible ways of evaluating the quality has also been the focus of much research and the results of these studies vary by location and by application (Alghanim et al., 2021; Rabiei-Dastjerdi et al., 2020). To our knowledge, there are currently no studies evaluating the quality of OSM data as it pertains to the factors that influence bus performance.

To summarise, the problems that currently exist in this research area are that geographical features often have to be manually collected, either by field survey or using online maps, and the studies in the are tend to be quite small, and/or use a very limited set of geographical feature. Volunteered geographic information like OSM shows promise but there is no research on the quality of OSM data for the geographical features that affect buses. This paper attempts to solve these problems with the following four main contributions:

- Presents a framework for extracting OSM data relevant to the factors that impact bus routes that is transferable between geographical regions and works with any bus metric.
- Presents novel methods for detecting turns on bus routes and the direction taken by a bus at roundabouts.
- Compares the result of these approaches to the ground truth data for a large dataset of 15 bus routes consisting of 539 unique stop pair segments.
- Performs in-depth analysis of the resulting data for average speed and reliability.

2. DATA

Several types of data are used in this study, including historical bus journey data, static route data and geographical data. Ireland's National Transport Authority (NTA) provided the historical bus journey data. It contained details of the journeys on the Dublin Bus network of 253 routes from 1st January 2018 to 31st December 2018. Fifteen bus routes from the 253 routes in the city met the following inclusion criteria: they must be outbound routes originating in the city centre of Dublin and terminating in a suburban area; they must have good data quality with at least 70% of the original dataset being usable and the route must also have a large test size with at least 4000 unique bus journeys on that route each year. No other selection criteria were considered when selecting the bus routes. The selected bus routes were 15A, 15B, 25A, 26, 27A, 41, 42, 49, 54A, 56A, 65B, 66, 69, 79A, and 130. The static route data included the shapefiles for the routes, and the location of the bus stops on the routes is available from the NTA in General Transit Feed Specification (GTFS) format. OSM has a structure of nodes (points), ways (lines connecting points), and polygons (areas defined by lines) that associate the spatial data component with contextual tags. For example, roads and bus routes are represented by ways, bus stops and traffic lights are represented by nodes, and industrial land use is represented as a polygon. The contextual tags for nodes and ways include details like the road's name or type or the bus stop's identifying name or number. This study extracted the OSM data using the QuickOSM extension in QGIS, an open-source geographic information system. The categories of OSM data used were nodes (traffic lights, pedestrian crossings and mini-roundabouts), ways representing the roads and polygons for land use.

3. METHODOLOGY

After obtaining OSM data, it was visualised in QGIS to validate the data was extracted correctly. The OSM data is then exported to a PostgreSQL database with a PostGIS extension. SQL queries was used with python 3.6 to manipulate the data in Jupyter Notebooks. Statistical analysis was performed with statmodels.

3.1 Defining Time Groups

The two biggest contributing factors to bus travel time variability are passenger load and traffic conditions (Mazloumi et al., 2011). However, these are complex to measure directly, but as they tend to be cyclical, the day of the week and time of day can be used as proxy measures (Mazloumi et al., 2011). These impactful factors account for much travel time variability, so it is important to control for these factors to see the impact of geographical features. The data was analysed to find the average whole route travel time for each of the routes at each of the time periods that occurred in the data, and based on that, the data was split into eight groups for analysis. This resulted in weekdays being split into five time groups, and weekends were split into three time groups and analysed separately, as shown in Table 1. The long morning peak period may seem counterintuitive but makes sense considering that these outbound routes originate in the city centre. The direction of travel also impacts journey time in how it relates to passenger load and traffic conditions at different times of the day. This is why only outbound bus routes that originate in the city centre and terminate in the suburbs were selected.

3.2 Creating Segments

Bus routes exist within OSM as a relation (a collection of ways and nodes), but very few bus routes had been added in Dublin, and most were either out of date, incomplete or had errors. For this reason, it was decided to use shapefiles of the bus route from the bus operator in GTFS format. These route shapefiles and bus stop locations are uploaded to the PostgreSQL database. Functions built into PostGIS were used to split the route into segments. The bus route shapefiles did not perfectly align with the roads in OSM as they were collected using Global Positioning System (GPS) data, so the route had to be snapped to the road. It was then possible to split the route into line segments representing the bus' journey between two consecutive bus stops. Firstly, the ST_LineLocatePoint function returns the fractional location of the closest node on the bus route (way) for each bus stop along the route. These fractional locations are used with the ST_LineSubstring function to return the route

	Weekday Early Morning	Weekday Morning Peak	Weekday Afternoon	Weekday Evening Peak	Weekday Late Evening	Weekend Morning	Weekend Afternoon	Weekend Evening
Start Time	04:00	07:30	11:00	16:30	19:00	04:00	12:00	21:00
End Time	07:30	11:00	16:30	19:00	End Service	12:00	21:00	End Service
Average Speed/km	27.42	23.21	21.19	18.61	24.16	27.13	22.95	26.6
Reliability	0.09	0.057	0.05	0.049	0.062	0.07	0.06	0.08

Table 1. Time groups used in this study and the average speed and reliability of the buses during these times.

segment between each consecutive pair of bus stops. Many segments were part of two or more bus routes, especially in the city centre, so a data frame of unique segments was created. The ratio of the total length that each segment represented was calculated from the fractional locations, and the total length of the route was calculated using ST_Length function. The ratio was maintained as a feature in the data frame and multiplied by the length of the route to get the length of each segment in metres.

The average journey time, the average speed and the reliability index for each segment were calculated using the historical bus journey data and the length of the segment. In this study, the reliability index is taken to be one divided by the standard deviation of the average journey time for that segment. This is aligned with the reliability measure described by Sterman and Schofer (1976). As shown in Table 1 the weekday evening peak is the time group with the lowest average speed and reliability, while weekday early mornings have the highest. Reliability shows greater relative variation than average speed. Bus frequency on each segment was also calculated using the historical bus journey data. The analysis of frequency was limited to the 15 routes in the study and did not consider other buses on the network. Due to data quality issues inherent with GPS technology (Lyan, 2021), there was, on average, 15% missing data, so the true frequency cannot be determined with absolute accuracy, but the analysis indicates relative frequency between segments.

3.3 Extracting Node Features

Some of the features were readily available in OSM as nodes: traffic lights, pedestrian crossings and mini-roundabouts. Datasets containing each of these features separately in the relevant geographical area were created using QGIS and stored in the PostgreSQL database. The features associated with a bus route could be easily extracted using the ST_DWithin function. In this case, the function was used to return the node features (e.g. traffic lights) within a certain distance of the way representing the bus route segment. The node features used coincided with the bus route segments, so the radian value was set low at 0.00002 (which corresponds to approximately 2m) to minimise false positives on adjacent roads. The number of node features per km for all segments was calculated and added as features to the segment data frame.

3.4 Extracting Way Features

To extract the desired tags of the roads, the bus routes segments had to be correlated with a road. The tag for roads in OSM is "highway", and it has many different subtypes, including roads that buses can not drive on: driveways, footpaths and cycle lanes. The bus route was matched to the closest OSM highway that was one of the following types: 'primary', 'secondary', 'tertiary', 'residential', 'motorway', 'unclassified', 'primary_link', 'secondary_link', 'motorway_link', 'trunk_link',

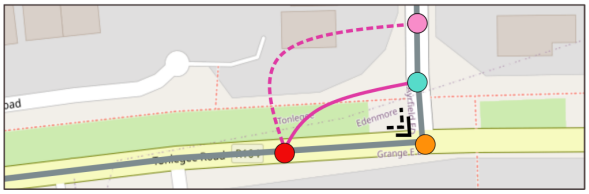

Fig 1(a): Left Turn

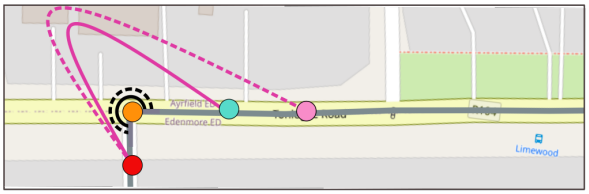

Fig 1(b): Right Turn

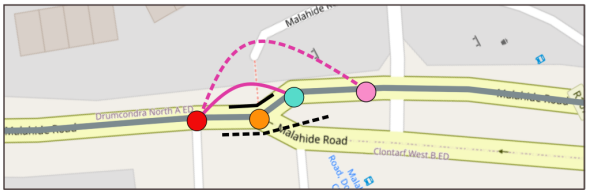

Fig 1(c): Avoiding false positives

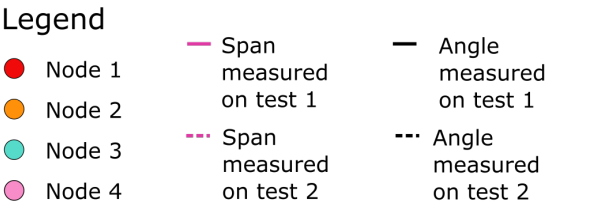

Figure 1. Visualisation of the turn identification process. (a) Two tests are performed. In test 1, the angle between nodes 1, 2 and 3 is recorded. In this case, the angle is less than 150°. For test 2, the angle between nodes 1, 2 and 4 is recorded. In this case, the angle is less than 150° also, so a left turn is recorded for that segment. (b) Similar to the left turn, except in this case, both tests return angles greater than 220°, so a right turn is recorded. (c) The second test exists to avoid splits in the road, like the one shown, being recorded as a turn. It does not always occur, but if the nodes are positioned unfavourably, the threshold for a turn can be surpassed. In this case, test 1 will return an angle of less than 150°, but test 2 will not, so no turn will be recorded. If node 3 is the last node on the segment, then the decision about turns is made solely on test 1.

'trunk', 'service'. This process results in a list of roads on each segment. Some segments will have only one road, and others will have multiple. The tags of the OSM road were then easily extracted, such as the average speed limit on that segment and the type of road (in the analysis, only if it was a secondary, primary, tertiary or residential road was considered), if a bus lane or bridge was present and the number of lanes. The average speed limit was treated as an average speed limit on all the roads on the segment; the proportion of the segment that each road occupied was not considered. The road types were recorded as binary - either a segment contained a residential road, or it did not. A segment can contain more than one type of road.

3.5 Extracting Polygon Features

There are land-use tags already present in OSM. These tags were used for residential land use, commercial land use, and industrial land use. The recreation land use tag was combined with the leisure tag to get a complete picture of land used for recreation. Retail land use was combined with the shop tag to get a full picture of this type of land use. Railway and bus stations were combined to produce a transport land use feature. In this case, it was decided to remove the node features and keep only the polygon features to avoid false positives. Education land use was combined with school and university tags. Finally, multiple services were combined into a services land use tag; this included fire stations, hospitals, nursing homes, churches, and police stations. The land-use features were recorded as a binary - either a segment had that type of land use, or it did not. A segment can contain more than one type of land use.

3.6 Change of Direction Features

3.6.1 Turns Turns are considered significant factors in a bus's speed and reliability but are less well studied than other geographical factors (Alfa et al., 1988). Turns can either be against traffic where it is necessary to wait for oncoming traffic to stop before crossing their lane, or with traffic. In the geographical area studied, left turns are with traffic and right turns are against traffic. It was therefore expected, that right turns would have a greater impact on bus performance than left turns. To detect direction change on a bus segment, the ST_Angle query was used to measure the angle between each three consecutive nodes on the way as depicted in Figure 1. An angle of less than 150° was a left turn, and an angle greater than 220° was a right turn. To minimise false positives when a road slightly changes direction when it divides, an additional check was included, where the angle between the first, second and the node after the third node was also checked. Direction changes coincident with roundabouts were excluded, as they are accounted for separately. The number of left and right turns per km was added to the data frame.

3.6.2 Roundabouts In OSM, roundabouts are implemented as ways, like other roads, but with a specific tag (junction=roundabout) that identifies them as roundabouts. Similar to the node features, roundabouts on bus routes could be easily identified using a ST_DWithin proximity query. We are not solely interested in the presence of the roundabout but also in the direction the bus took at the roundabout. Each segment was provisionally assigned a roundabout score of zero. For each node on the bus route segment, a ST_DWithin query was performed with a dataset of roundabouts produced using QGIS. As shown in Figure 2 the first node on the segment that coincided with the roundabout, the last node that coincided with

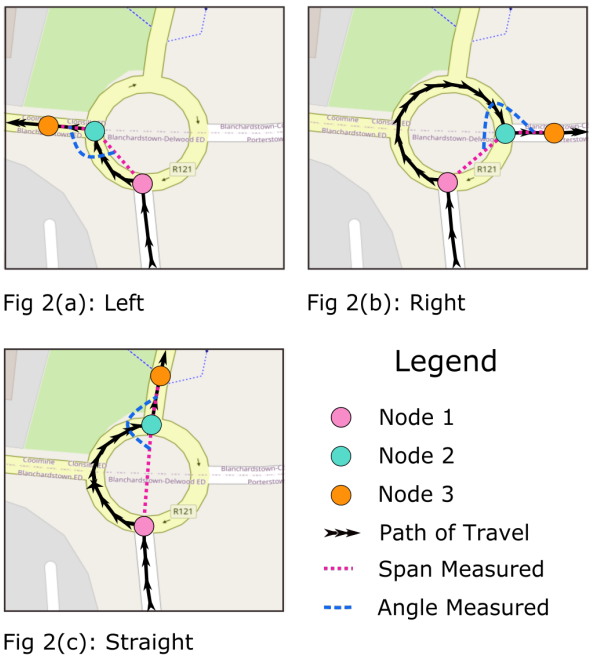

Figure 2. Visualisation of process for detecting the direction a bus takes at a roundabout.

the roundabout and the node after that (i.e. the first node after the route leaves the roundabout) were identified. The S T_Angle function was used with these three nodes to determine the angular change in direction after the roundabout. While roundabouts can vary in reality, they were conceptually simplified as a left, straight, right or complete turn in this experiment. If the returned angle is greater than 315° degrees, that is considered a full circle of the roundabout and four is added to the roundabout score for that segment. Similarly, an angle of greater than 215° and less than 315° is considered a right turn and a score of three is added to the roundabout score. Angles between 162° and 215° are deemed straight, and a score of two is added. Angles less than 162° are considered a left turn, and a score of one is added. In the case of segments with multiple roundabouts, the roundabout score is cumulative. This approach tries to numerically quantify the amount of time a bus spends on a roundabout.

3.7 Analysis

Analysis was performed on the resulting dataset to understand the effect of the features on the average speed and reliability of the bus at different times of the day. Due to geographic data's highly correlated nature, it was impossible to maintain all of the features and have a stable statistical analysis. Initially, Ordinary Least Squares (OLS) linear regression models were trained, however, the condition number of the models was high (22 and 440 for average speed and reliability models respectively). The condition number is diagnostic of multicollinearity issues (Kim, 2019) common in geographical data since linear dependencies cannot be avoided by experiment design (Brunsdon et al., 2012). Ideally, the condition number should be as low as possible and less than 20 is recommended (Brunsdon et al., 2012). Multicollinearity is problematic because it can cause models to have inaccurate regression coefficients, produce incorrect p-values and reduce model predictability making it sensitive to small feature changes (Altelbany, 2021). To avoid this, and improve confidence in the results, two changes were made. Firstly,

highly correlated features were removed. Correlation matrices and feature importance were used to determine the correlation between and relative impact of the features. Features that are deemed redundant due to being highly correlated with another feature and having minimal impact on the independent variable were removed. The removed features were ratio and the number of lanes for both reliability and average speed models, and secondary roads and residential roads for average speed and reliability models, respectively. Secondly, regularisation methods were used to overcome the shortcomings of OLS regression. Regularisation regression refers to forms of regression where the coefficient estimates are constrained with penalty terms to avoid overfitting. Elastic net regression is a combination of ridge regression which uses squared penalty terms and lasso regression which uses absolute penalty terms (Zou and Hastie, 2005). Elastic net regression was chosen as the superior regularisation method of handling data with multicollinearity issues (Altelbany, 2021). Elastic net models were developed with the remaining features to show their relative effect on the bus performance. The final models had condition numbers of 11.5 and 13 for average speed and reliability respectively. The accuracy of data extraction from OSM was evaluated for all features using visual examination in QGIS. Where possible, the ground truth for the features was determined using visual analysis of the bus routes in QGIS, such as turns and roundabouts. Bus lanes, traffic lights and pedestrian crossings were determined by visual inspection using Google Maps satellite imagery. It was not possible to verify the ground truth for the types of roads and land use features as they are subjective in many cases.

4. RESULTS AND DISCUSSION

Feature	Accuracy Extraction OSM	Accuracy Ground Truth
Traffic_Lights	100%	98%
Pedestrian_Crossings	100%	99%
Mini_Roundabouts	100%	99%
Road_Identification	100%	99%
Average_Speed_Limit	100%	N/A
Road_Primary	100%	N/A
Road_Secondary	100%	N/A
Road_Tertiary	100%	N/A
Road_Residential	100%	N/A
Bus_Lane	100%	79%
Bridge	100%	100%
Landuse_Residential	100%	N/A
Landuse_Industrial	100%	N/A
Landuse_Recreation	100%	N/A
Landuse_Retail	100%	N/A
Landuse_Transport	100%	N/A
Landuse_Education	99%	N/A
Landuse_Services	99%	N/A
Left_Turns	N/A	97.5%
Right_Turns	N/A	98%
Roundabouts	99%	98%

Table 2. Accuracy of feature extraction compared to OSM data and the ground truth

Table 2 details the results of the method compared to the OSM data and the ground truth. The method extracts the data present in OSM with almost perfect accuracy, and the results of the feature extraction method compared to the ground truth show over 97% accurate identification of all individual features except for bus lanes. In the case of the node features, the particular traffic light, pedestrian crossing, or mini-roundabout is either present in OSM or not. There are no cases of the traffic light being present in OSM and not being detected by the proximity query. Still, occasionally on narrow roads, a traffic light on an adjacent road is detected as a false positive. Also, there were two incidents of mini-roundabouts not being marked in OSM.

The next group of features are the ones that are stored in OSM as tags of the road - average speed limit, bus lanes and bridges. The ground truth for bridges is straightforward to assess and bridges are well recorded in the OSM data for Dublin. All bridges were correctly recorded with no false negatives. The worst recorded feature in our dataset was bus lanes. Bus lanes are often not recorded in OSM in Dublin, and as a result, the accuracy of this feature compared to the ground truth is only 79%. We suspect this is because the bus lanes in Dublin tend to be intermittent. This creates a barrier to entry in OSM because the way needs to be split into three ways and the details recorded separately. Analysis showed that the ground truth assessment showed 117 segments with a majority bus lane; only 28 were correctly recorded with 23 false positives and 89 false negatives. Land use features tended to be well recorded in OSM and extracted well using the search features. Some schools and other public services are not tagged as such in OSM, resulting in slightly lower accuracy for those features.

For changes of direction: left turns, right turns and roundabouts, assessing the method becomes non-trivial. Turns do not exist as entities in OSM data so no evaluation can be made on that basis. It can be difficult to determine in reality if a change in bearing is a turn or a curved road, so the ground evaluation of left turns is somewhat subjective. There are 95 left turns, 94 of those are recorded correctly, one was missed, and there are 12 false positives. The ground truth for right turns is more straightforward to quantify as they involve crossing a traffic lane. There are 75 right turns, and 71 of those are recorded correctly, four are missed, and there are six false positives. The false negatives were when the turn angle was insufficient to be recorded correctly. Roundabouts are recorded in OSM, and there are 64 segments with roundabouts in the ground truth data. Of these, there are 60 recorded correctly as roundabouts, and four are missed. There are no false positives. Of the missed roundabouts, all are not tagged as roundabouts in OSM. Of those recorded correctly as roundabouts, 55 have the correct directions indicated, and five do not. Untagged roundabouts are the single biggest cause of error in this method. They result in errors as roundabouts not being interpreted as roundabouts and also cause false-positive left turns.

A feature that was dropped from all models was the number of lanes feature, due to having low feature importance and being correlated with multiple other features. The number of lanes feature was analysed independently of the other features and an increased number of lanes generally decreases average speed and reliability, when only the urban segments were examined, the opposite is true. This result echoes the findings of Chioni et al. (2020), who found a positive correlation between the number of lanes and bus bunching in segments with heavy traffic and a negative correlation in segments in less-congested regions.

Feature	Weekday Early Morning	Weekday Morning Peak	Weekday Afternoon	Weekday Evening Peak	Weekday Late Evening	Weekend Morning	Weekend Afternoon	Weekend Evening
Intercept	30.154*	26.538*	24.825*	21.381*	28.451*	31.289*	27.456*	30.521*
Frequency	-0.273	-0.031	-0.278	-0.666*	-0.735	-0.364	-0.718	-0.741
Length	2.223*	2.458*	2.821*	2.841*	2.989*	2.652*	2.786*	2.89*
Traffic_Lights	-2.408*	-3.319*	-2.819*	-2.401*	-2.8*	-3.229*	-2.84*	-2.737*
Pedestrian_Crossings	-0.499	-0.935*	-0.919*	-1.013*	-0.889*	-0.694	-0.878*	-0.838*
Mini_Roundabouts	-0.124	-0.262	-0.521	-0.4	-0.561	-0.386	-0.455	-0.664
Average_Speed_Limit	1.334*	1.144*	1.347*	0.896*	1.629*	1.281*	1.286*	2.023*
Road_Primary	2.682	1.891	1.762	1.665	-0.069	2.539	1.427	0.018
Road_Tertiary	-2.698*	-1.035	-0.941	-0.473	-1.114	-1.499	-0.037	-1.708
Road_Residential	-0.614	-0.504	0.287	0.482	-0.727	-1.056	-0.305	-1.191
Bus_Lane	-3.617*	-2.148	-1.899	-2.305	-2.531	-2.149	-2.314	-2.092
Bridge	-4.322*	-3.062	-2.258	-2.254	-3.474*	-3.205	-3.057	-4.306*
Landuse_Residential	0.958	-1.114	-1.806	-0.926	-2.543*	-1.472	-2.95*	-1.902
Landuse_Industrial	-0.686	0.931	0.533	0.687	4.021*	1.749	3.395*	5.221*
Landuse_Recreation	0.481	0.952	1.362	0.769	0.749	0.492	0.406	0.861
Landuse_Retail	-5.223*	-4.402*	-4.761*	-4.432*	-4.422*	-5.417*	-4.893*	-4.679*
Landuse_Transport	-4.356*	-3.176	-2.435	-2.503	-3.938*	-4.477*	-4.077*	-4.383*
Landuse_Education	2.012	-0.479	-0.481	-0.356	-0.07	0.498	0.457	0.305
Landuse_Services	-1.579	-1.423	-1.627	-0.962	-0.96	-0.782	-1.499	-0.881
Left_Turns	-0.89	-0.735	-0.792*	-0.73*	-1.005*	-0.905*	-0.952*	-1.144*
Right_Turns	-1.368*	-1.195*	-1.038*	-1.003*	-1.121*	-1.504*	-1.217*	-1.067*
Roundabouts	-1.847*	-1.438*	-1.035*	-0.558	-1.062*	-1.724*	-1.106*	-1.195*
Adjusted R^2	0.342	0.446	0.490	0.510	0.508	0.453	0.496	0.492

$*$= statistically significant ($p < 0.025$)

Table 3. Average speed model coefficients by time period.

The results of the statistical analysis presented in Table 3 and Table 4 are generally intuitive and consistent with the literature. Reliability and average speed are related metrics with a correlation coefficient of 0.55. The frequency of the bus negatively impacts the average speed and reliability, especially during the weekday evening peak, late evening and at the weekend. This is perhaps surprising but is consistent with the literature (Cui et al., 2019). We suggest it may be due to increased bunching and headway variability if supply exceeds demand. The length of the route has a large positive impact on average speed and a small negative impact on reliability and is statistically significant at all times of day. These results are consistent with the literature (Lyan, 2021; Soza-Parra et al., 2021). Increased segment length improves speed as the bus spends proportionately less of the segment speeding up and slowing down but will result in a longer travel time with more opportunity for variability.

Traffic lights negatively affect average speed and reliability, as has been shown in the literature many times (Abkowitz and Engelstein, 1983; Feng et al., 2015). Similarly, pedestrian crossings always have a negative impact on the speed or reliability, but to a lesser extent than traffic lights as they are only activated when pedestrians are present. Pedestrian crossings are statistically significant during the busiest travel times. Mini-roundabouts have negative coefficients for average speed and mixed coefficients for reliability, likely because they are highly correlated with suburban areas; however, mini-roundabouts were never found to be statistically significant.

The average speed limit positively affects average speed, as expected (Lyan, 2021), and is statistically significant at all times. The average speed limit has a small mixed impact on reliability that is not statistically significant. Primary roads positively impact bus speed, except during peak evening travel time, and have a negative impact on bus reliability. Secondary roads have a large statistically significant negative impact on bus reliability. Tertiary roads have a negative impact on average speed, statistically significant during the early morning period, but they have a small mixed impact on reliability. Residential roads do not have a significant impact on average speed. Bus lanes have a negative impact on average speed on reliability. The analysis of bus lanes was done on the ground truth, not the values derived from the method, as the lack of bus lanes being included in OSM data meant no meaningful conclusion could be drawn from the extracted data. It may seem counterintuitive that bus lanes do not increase speed and reliability. Still, it is consistent with the existing literature (Soza-Parra et al., 2021; Chioni et al., 2020). The most likely explanation is that since bus lanes tend to be put in places with heavy traffic when compared to segments without bus lanes, there is a negative impact. Another possible reason for that may be that in Dublin, the bus lanes tend to be short and intermittent, and in many places in the city, and they are quite narrow, which has been shown to have a less positive impact on bus speed (Arasan and Vedagiri, 2010). Bridges were found to have a significant negative impact on average speed but a minimal mixed impact on reliability. This is likely because bridges often are narrow points on the road. This narrowing causes a consistent slowdown, impacting speed but not reliability.

Feature	Weekday Early Morning	Weekday Morning Peak	Weekday Afternoon	Weekday Evening Peak	Weekday Late Evening	Weekend Morning	Weekend Afternoon	Weekend Evening
Intercept	0.125*	0.078*	0.068*	0.068*	0.088*	0.099*	0.077*	0.105*
Frequency	-0.008*	-0.0	-0.001	-0.001	-0.004*	-0.003	-0.004*	-0.005*
Length	-0.013*	-0.008*	-0.006*	-0.007*	-0.008*	-0.009*	-0.008*	-0.011*
Traffic_Lights	-0.018*	-0.011*	-0.008*	-0.008*	-0.01*	-0.012*	-0.01*	-0.012*
Pedestrian_Crossings	-0.003	-0.002	-0.003*	-0.003*	-0.002	-0.002	-0.002	-0.002
Mini_Roundabouts	0.003	0.001	0.0	-0.001	-0.001	-0.001	0.0	-0.0
Average_Speed_Limit	-0.001	0.003	0.0	-0.003	0.0	-0.002	-0.0	0.003
Road_Primary	-0.021	-0.012	-0.005	-0.013	-0.015	-0.014	-0.014	-0.023
Road_Tertiary	0.004	-0.005	-0.004	-0.001	0.0	-0.005	0.005	0.004
Road_Secondary	-0.021	-0.014*	-0.015*	-0.02*	-0.016*	-0.019*	-0.011*	-0.015*
Bus_Lane	-0.013	-0.004	-0.003	-0.006	-0.006	-0.007	-0.006	-0.01
Bridge	-0.003	-0.006	-0.001	-0.001	-0.006	0.001	-0.002	-0.008
Landuse_Residential	-0.014	-0.007	-0.002	0.004	-0.006	-0.006	-0.007	-0.009
Landuse_Industrial	-0.006	0.004	0.0	-0.005	0.012*	0.004	0.01*	0.027*
Landuse_Recreation	-0.002	0.001	0.005*	0.002	-0.001	-0.002	-0.0	-0.003
Landuse_Retail	-0.02*	-0.012*	-0.013*	-0.016*	-0.016*	-0.018*	-0.014*	-0.017*
Landuse_Transport	-0.012	-0.007	0.0	-0.005	-0.011	-0.011	-0.012	-0.011
Landuse_Education	0.019*	-0.003	-0.005	0.003	0.0	0.001	0.001	0.0
Landuse_Services	-0.018	-0.005	-0.003	-0.003	-0.008	-0.008	-0.006	-0.009
Left_Turns	-0.002	-0.0	-0.0	-0.0	-0.001	0.0	-0.0	-0.002
Right_Turns	-0.004	-0.002	-0.002	-0.004*	-0.003	-0.004	-0.003	-0.002
Roundabouts	-0.011*	-0.006*	-0.002	-0.002	-0.003	-0.005*	-0.003	-0.004
Adjusted R^2	0.291	0.281	0.388	0.454	0.393	0.318	0.372	0.318

$*$= statistically significant ($p < 0.025$)

Table 4. Reliability model coefficients by time period.

Retail land use has the largest impact on average speed and reliability, and it is always negative and statistically significant. Other land-use features show mixed impact on the average speed and reliability corresponding to the normal movement of people. Industrial land has a large significant positive impact on both reliability and average speed during the late evenings and weekends, corresponding with times when the local businesses are closed. Similarly, recreation land use significantly impacts average speed and reliability during the weekday afternoons when many people are at school or work. Educational land use was not significant for average speed and positively impacted reliability during the weekday early mornings.

Both left and right turns have large significant negative effects on average speed, with right turns (against traffic) having more than double the impact of left turns (with traffic). Right turns only significantly impact reliability during the evening peak travel period, and left turns have no significant impact on reliability. These findings make sense when one considers a bus slowing down to navigate a left turn safely, but this speed reduction is consistent and does not impact reliability. Similarly, a right turn always impacts the bus' average speed but only impacts reliability at peak times when the volume of oncoming traffic will influence how long the bus must wait before turning. Roundabouts significantly negatively impact the average speed at all times of the day except evening peak but only significantly impacts reliability in the early part of the day. Generally, the coefficients for both metrics are bigger during off-peak periods, indicating that perhaps the effect of roundabouts is only seen in free-moving traffic, and the effect is lost as traffic increases.

The average adjusted R^2 values in the models in this study, 0.47 and 0.35 for average speed and reliability respectively, are in line with the existing literature. The adjusted R^2 values vary widely depending on the metric and features used and are not reported in all studies because they are not always deemed important in understanding the relationship between the dependent and independent variables (Kaewunruen et al., 2021; El-Geneidy et al., 2011). Consistently lower adjusted R^2 values are seen in studies looking at reliability versus average speed: El-Geneidy et al. (2011) found adjusted R^2 between 0.07 and 0.59. Chioni et al. (2020) had adjusted R^2 of 0.11 when looking at bus bunching with OLS linear regression but that improved to 0.57 when Geographically Weighted Regression was used. Studies that include passenger and traffic information as features tend to report higher adjusted R^2 values. Feng et al. (2015) and Hu and Shalaby (2017) included passenger boarding and traffic features in their segment models and achieved adjusted R^2 values as high as 0.78 and 0.75 respectively.

5. CONCLUSION

The proposed framework returns results consistent with the literature, is comparable to a manual approach in accuracy, and can automatically analyse many bus routes quickly compared to a manual survey. The method is independent of the geographical region and bus metrics. The methods for detecting turns and roundabouts are successful. A large number of features were examined in this study in contrast to previous works, and it is the largest segment dataset to be validated against the ground

truth. Retail land use, right turns, left turns, traffic lights, and roundabouts emerged as the biggest negative factors for bus average speed and reliability. The average speed limit and the length of the segment have large positive impacts on average speed. The biggest limitation of this approach is the dependence on the quality of the local OSM data, especially the tendency for contributors not to record bus lanes and roundabouts.

Planned further work includes validation of the land-use features against the official Ordnance Survey data in Ireland, and the development of methods to detect unmarked roundabouts in OSM to improve the quality of OSM data and this method. Additional metrics such as headway variability, timetable adherence or bus bunching can be analysed with this method and the method can be applied to whole bus routes, not just segments. Additional features could also be included, such as on-street parking. This work can be extended to improve the prediction of bus journey times on existing and planned routes and bus journey time simulations and improve scheduling and bus networks' overall reliability.

ACKNOWLEDGEMENTS

This publication has emanated from research supported in part by a grant from Science Foundation Ireland under Grant number 18/CRT/6183. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

Abkowitz, M., Engelstein, I., 1983. FACTORS AFFECTING RUNNING TIME ON TRANSIT ROUTES. *Transportation Research part A: General*, 17(2), 107–113.

Alfa, A. S., Menzies, W. B., Purcha, J., McPherson, R., 1988. A regression model for bus running times in suburban areas of winnipeg. *Journal of Advanced Transportation*, 21(3), 227–237.

Alghanim, A., Jilani, M., Bertolotto, M., McArdle, G., 2021. Leveraging Road Characteristics and Contributor Behaviour for Assessing Road Type Quality in OSM. *ISPRS International Journal of Geo-Information*, 10(7), 436.

Almeida, F., Lobo, A., Couto, A., Ferreira, J. P., Ferreira, S., 2022. Urban factors influencing the vehicle speed of public transport. *Transportation Research Procedia*, 62, 318–324.

Altelbany, S., 2021. Evaluation of Ridge, Elastic Net and Lasso Regression Methods in Precedence of Multicollinearity Problem: A Simulation Study. *Journal of Applied Economics and Business Studies*, 5(1), 131–142.

Arasan, V. T., Vedagiri, P., 2010. Study of the impact of exclusive bus lane under highly heterogeneous traffic condition. *Public Transport*, 2(1), 135–155.

Bertolotto, M., McArdle, G., Schoen-Phelan, B., 2020. Volunteered and crowdsourced geographic information: the OpenStreetMap project. *Journal of Spatial Information Science*, 20, 65–70.

Brunsdon, C., Charlton, M., Harris, P., 2012. Living with Collinearity in Local Regression Models. *In Proceedings of the Tenth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*.

Chioni, E., Iliopoulou, C., Milioti, C., Kepaptsoglou, K., 2020. Factors affecting bus bunching at the stop level: A geographically weighted regression approach. *International Journal of Transportation Science and Technology*, 9(3), 207–217.

Cui, H., Xie, K., Hu, B., Lin, H., Zhang, R., 2019. Analysis of Bus Speed Using a Multivariate Conditional Autoregressive Model: Contributing Factors and Spatiotemporal Correlation. *Journal of Transportation Engineering, Part A: Systems*, 145(4), 04019009.

Dastjerdi, A. M., Kaplan, S., de Abreu e Silva, J., Anker Nielsen, O., Camara Pereira, F., 2019. Use intention of mobility-management travel apps: The role of users goals, technophile attitude and community trust. *Transportation Research Part A: Policy and Practice*, 126, 114–135.

El-Geneidy, A. M., Horning, J., Krizek, K. J., 2011. Analyzing transit service reliability using detailed data from automatic vehicular locator systems. *Journal of Advanced Transportation*, 45(1), 66–79.

Feng, W., Figliozzi, M., Bertini, R., 2015. Quantifying the joint impacts of stop locations, signalized intersections, and traffic conditions on bus travel time. *Public Transport*, 7, 391–408.

Grinberger, A. Y., Minghini, M., Yeboah, G., Juhász, L., Mooney, P., 2022. Bridges and Barriers: An Exploration of Engagements of the Research Community with the OpenStreetMap Community. *ISPRS International Journal of Geo-Information*, 11, 54–83.

Hu, W. X., Shalaby, A., 2017. Use of Automated Vehicle Location Data for Route- and Segment-Level Analyses of Bus Route Reliability and Speed. *Transportation Research Record*, 2649(1), 9–19.

Kaewunruen, S., Sresakoolchai, J., Sun, H., 2021. Causal analysis of bus travel time reliability in Birmingham, UK. *Results in Engineering*, 12, 100280.

Kim, J. H., 2019. Multicollinearity and misleading statistical results. *Korean Journal of Anesthesiology*, 72(6), 558–569.

Lyan, G., 2021. Bus Commercial Speed Impact Factors, A Network Scale Analysis Using Massive Data. hal-03356103.

Mazloumi, E., Rose, G., Currie, G., Sarvi, M., 2011. An Integrated Framework to Predict Bus Travel Time and Its Variability Using Traffic Flow Data. *Journal of Intelligent Transportation Systems*, 15(2), 75–90.

Rabiei-Dastjerdi, H., McArdle, G., Ballatore, A., 2020. Urban Consumption Patterns: OpenStreetMap Quality for Social Science Research. *In Proceedings International Conference on Geographical Information Systems Theory, Applications and Management*.

Soza-Parra, J., Muñoz, J. C., Raveau, S., 2021. Factors that affect the evolution of headway variability along an urban bus service. *Transportmetrica B: Transport Dynamics*, 9(1), 479–490.

Sterman, B. P., Schofer, J. L., 1976. Factors Affecting Reliability of Urban Bus Services. *Transportation Engineering Journal of ASCE*, 102(1), 147–159.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.