Predicting energy consumption of zero emission buses using route feature selection methods


Published in:
Transportation Research Part D: Transport and Environment

Document Version:
Publisher's PDF, also known as Version of record

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Publisher rights
Copyright 2024 the authors.
This is an open access article published under a Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution and reproduction in any medium, provided the author and source are cited.

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.

Open Access
This research has been made openly available by Queen's academics and its Open Research team. We would love to hear how access to this research benefits you. – Share your feedback with us: http://go.qub.ac.uk/oa-feedback

Download date: 28. Mar. 2024
Predicting energy consumption of zero emission buses using route feature selection methods

L.A.W. Blades\textsuperscript{a,}\textsuperscript{*}, T. Matthews\textsuperscript{a}, T.E. McGrath\textsuperscript{a}, J. Early\textsuperscript{a}, G. Cunningham\textsuperscript{a}, A. Harris\textsuperscript{b}

\textsuperscript{a} School of Mechanical and Aerospace Engineering, Queen’s University Belfast BT9 5AG, Northern Ireland, UK
\textsuperscript{b} Bamford Bus Company Ltd. Trading as Wrightbus, 201 Galgorm Road, Ballymena, Co. Antrim, BT42 1SA Northern Ireland, UK

**ARTICLE INFO**

**Keywords:**
- Zero emission transport
- Battery electric bus
- Vehicle modelling
- Feature selection methods
- Sustainable public transport
- Energy consumption prediction

**ABSTRACT**

This paper reports new insights into how the selection of route based characteristic parameters can influence the predicted energy consumption for next generation battery electric buses. 24 characteristic parameters have been studied to understand their relative importance on vehicle energy consumption and to develop new data-driven prediction models. The parameters are grouped into two scenarios, representative of the varying levels of route information available to a typical bus operator. A combination of feature selection methods was used to determine which characteristic parameters had the greatest influence on energy consumption. Regression based prediction models were developed, and models were then validated using standard and real vehicle drive cycles. The prediction models had a mean absolute percentage difference of 2.10\textsuperscript{-}10.67\%. This paper presents a novel methodology to estimate energy consumption of operating zero emission vehicles, which will support public transport operators, policy makers and energy suppliers in the decarbonisation of public transport.

1. Introduction

Global CO\textsubscript{2} emissions from transport accounted for 37\% of end-use sector emissions in 2021, with an annual average growth rate of nearly 1.7\% from 1990 to 2021, faster than any other end-use sectors (International Energy Agency, 2022a). The adoption of zero emission vehicles has been rapidly increasing globally, driven by legislative requirements to tackle air pollution and achieve net zero carbon targets. For example, the C40 Clean Bus Declaration has seen 36 cities (including London, Mexico City, Rio de Janeiro and Tokyo) already commit to procuring only zero emission vehicles from 2025, with a global reach of 140 million citizens and procurement potential of over 80,000 buses (C40 Cities, 2020).

The two main zero emission technologies that have emerged in the bus sector are battery electric buses (BEB) and hydrogen fuel cell electric buses (FCEB). Both of these technologies are zero emission at point of operation, meaning they do not emit greenhouse gas emissions at the tailpipe (referred to as tank-to-wheel TTW emissions). However the production of both the vehicle and the energy source (electricity/hydrogen) used to drive the power train (referred to as well-to-tank emission) will have related greenhouse gas emissions (Hensher et al., 2022). According to recent statistics, there are approximately 670,000 BEB operating globally projected to rise to up to 3 million in operation by 2030 (International Energy Agency, 2022b). There are an additional 4,738 FCEB estimated to be

\* Corresponding author.

E-mail address: L.Blades@qub.ac.uk (L.A.W. Blades).

https://doi.org/10.1016/j.trd.2024.104158

Received 14 September 2023; Received in revised form 19 February 2024; Accepted 10 March 2024
Available online 18 March 2024
1361-9209/Crown Copyright \copyright 2024 Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
operating across 16 countries, the majority of which (88%) are operating in China. It is estimated that these numbers will grow to approximately 45,000 FCEB operating in Europe, 1,200 FCEB operating in Japan and 11,600 FCEB and fuel cell coaches operating in China by 2030 (Samsun et al., 2022).

For operators, planners, local authorities and government bodies, accurate predictions of the likely operational energy demand of these new zero emission buses are essential to provide assurance that routes can be successfully serviced to meet timetable demands and to support planning and roll out of the necessary refuelling/recharging infrastructure. As the dominant zero emission technology, there has been a range of studies that have developed methods to estimate energy consumption of BEB configurations. These methods can be broadly categorised as rule of thumb approximations, kinematic/physics-based models and data-driven models (Abdelaty et al., 2021). Within the literature, a variety of statistical and machine learning techniques have been applied to kinematic/physics-based and data driven models. A number of characteristic parameters have been identified as having a significant impact on energy consumption and predictive models with varying degrees of accuracy developed. Significant characteristic parameters include road gradient (Pamula and Pamula, 2020; Abdelaty et al., 2021; Ma et al., 2021), numbers of stops per kilometre (Kivekas et al., 2018; Vepsäläinen et al., 2018), average speed (Lin, Lin and Ying, 2020; Wang et al., 2020) and driver aggressiveness (Kivekas et al., 2018; Vepsäläinen et al., 2018). Whilst previous studies have used a variety of methods to assess the significance and sensitivity of these parameters implementing these findings may be difficult for transport operators/planners, with limited understanding of the comparative accuracy of these different methods.

This paper demonstrates a novel methodology for estimating the energy consumption of BEBs, the most common zero emission technology for city-wide fleet assessment based on easily accessible route data. Two scenarios will be considered. In the first, full drive cycle and elevation profiles are available. The second is based on open-source timetable data and elevation/mapping data. It will be demonstrated how data-driven models could be implemented in future public transport planning where limited levels of data relating to route and operational requirements may exist. This paper will support operators considering fleet transitions and the likely energy consumption of zero emission vehicles within urban bus networks.

This work establishes a novel framework, based on statistical methods, which develops an approach that ranks variables robustly and flexibly. Energy prediction models are then tested to demonstrate the enhancement in energy prediction accuracy achieved with the new method.

The following sections present the methodology and its background in more details. Section 2 is a literature review of studies that have examined the significance of variables and used data-driven methods to predict operational energy consumption of battery electric and fuel cell vehicles. Section 3 outlines the underpinning vehicle models, synthetic drive cycle creation, calculation of drive cycle characteristic parameters and systematic regression framework applied. Within Section 4, the ranking of relative importance of the characteristic parameters and the developed prediction equations are presented and discussed. Finally, Section 5 completes the paper with concluding remarks, limitations of this work and recommendations for future work.

2. Literature review

In the early design stages of zero emission vehicles, one of three methods is typically used to estimate the energy consumption, as described by (Li et al., 2021); rule of thumb, physics-based models and data-driven models.

Rule of thumb estimations, where an average energy consumption (kWh/km) is combined with distance or average travel time to estimate route level consumption, have been used commonly within scheduling optimisation studies to simplify algorithms, environmental assessments and economic analysis to assess different technologies. These values may be extracted from publicly available sources such as manufacturer’s data or based on certification values generated via chassis dynamometer testing. In the UK bus sector, Zero Emission Bus (ZEB) accreditation is required to ensure public money supports proven technologies. As part of the accreditation process, vehicles are tested on the UK Bus Cycle with 50% seated passenger capacity at an ambient temperature of 10°C, and vehicle energy consumption is measured. Four double deck battery electric buses (BEB) have been certified under this accreditation scheme, with energy consumptions ranging from 0.68 kWh/km to 1.14 kWh/km (Zemo Partnership, 2023). Within the literature, average energy consumption of battery electric bus operation have converged on values of 1.0–1.3 kWh/km and have been used by (Xylia et al., 2019) to examine locations of charging infrastructure in Stockholm, (Logan, Nelson and Hastings, 2020) to estimate carbon dioxide emissions in the UK, and by (Hensher et al., 2022) to undertake environmental and economic impact assessments in Australia.

Both physics-based and data-driven approaches may offer higher accuracy than rule of thumb estimations as they can account for greater variability of real word driving conditions. However, they also require significant amounts of high-resolution data which may not be readily available and the appropriate skills to apply models/methods.

A recent review undertaken by (Lim et al., 2023) provides a detailed overview of the parameters that have been included across the literature for energy forecasting models and charging scheduling models for battery electric buses. (Abdelaty et al., 2021) broadly categorised these parameters into one of the four headings; (1) operational parameters (2) topological parameters, (3) vehicular parameters (bus mass, drag coefficient, rolling resistance) (4) external parameters (ambient temperature wind speed and auxiliary power).

A key input for vehicle modelling for physics-based models is the use of drive cycles (DC) which represent a set of data points of variables such as speed, route gradient and auxiliary power loads versus time. These represent real-world driving patterns of vehicles on specific routes. DCs have been used in the literature for a variety of purposes such as evaluating the energy economy of electric vehicles (EVs) (Brady and O’Mahony, 2016), an input for emissions models (An et al., 1997; Zhai, Frey and Rouphail, 2011; Ragione and Giovanni, 2016; Ahn et al., 2022) or fuel consumption (FC) models (Dhaou, 2011), a benchmark to compare the emissions of different vehicles (Tzamkiozis, Ntziachrists and Samaras, 2010) and for quickly evaluating powertrain variables in the design phase.
(minimising production and redesign costs) (Hereijgers et al., 2017). DCs provide additional benefits for electric vehicles (EVs) and hybrid electric vehicles (HEVs) such as aiding in the development of control strategies of HEVs (Wang and Lukic, 2011), training and testing state-of-charge (SOC) estimation models (Zahid et al., 2018), evaluating EV power management and design (Yang et al., 2018), helping predict EV battery lifetimes (Baure and Dubarry, 2019) and being an input for algorithms obtaining optimum power management for HEVs (Jeong et al., 2017). To enable comparisons between drive cycles, characteristic parameters (CPs) including maximum speed, average speed, average number of stops per km, and average gradient, may be calculated (Barlow et al., 2009) allowing for clustering of similar routes.

There are several studies that develop physics-based models that utilise drive cycle data and provide estimates of energy consumptions for battery electric buses. For example (Gao et al., 2017), using Autonomie software, estimated that electric bus energy consumption varied from 1.24 to 2.48 kWh/km operating on both real and standardized drive cycles. (Gallet, Massier and Hamacher, 2018) calculated the energy demand of electric buses operating in Singapore, reporting median values of 1.75 ± 0.41 kWh/km. As described by (Li et al., 2021) physics-based models encounter difficulties including incorporating factors such as departure day, time of day, and wet/dry weather conditions. Physics-based models are also restricted by the requirement for high-resolution speed profile data which ideally should be representative of a variety of driving styles, location (urban/rural) and congestion levels.

More recently, data-driven methods to estimate both energy consumption and identify significant features/predictor variables have been developed based on high-resolution data extracted from simulation models or data collected from vehicle operation. Early examples of data-driven methods applied to BEB energy consumption measured via synthetic drive cycle/simulation data (Kivekäs et al., 2018) or observed data (Vepsäläinen et al., 2018) identified driver aggressiveness, stop frequency and ambient temperature as significant influencing factors. (Kivekäs et al., 2018), extending the Autonomie models initially developed within (Launen and Lipman, 2016), undertook analysis of a range of bus propulsion technologies including battery electric to identify the factors impacting performance of a single deck vehicle and assessed sensitivity analysis of passenger loadings. Based on the 3000 synthetic drive cycles of a bus route in Espoo Finland, (Kivekäs et al., 2018) examined correlations, and using variance decomposition methods between the cycle characteristic parameters and energy consumption found that driver aggressiveness and stop frequency had the largest influence on energy consumption. (Vepsäläinen et al., 2018) applied linear stepwise regression with forward selection to measured data collected from operation of six single deck 12 m length BEB over the course of two years in Finland. They identified that the largest impact on operational energy consumption was due to external, topological and operational factors, namely direct current (DC) devices (highly correlated with ambient temperature), stops per km and driving aggressiveness with a total calculated R² value of 0.76. The inclusion of an additional seven factors did not significantly increase this R² value (to 0.81) and were excluded from the final regression model as they only marginally increased statistical power. It should be noted that in both studies, gradient/route elevation was not directly included as a study parameter, with (Vepsäläinen et al., 2018) noting it’s influence on the energy consumption rate of the vehicles operating certain routes (correlation value of 0.46 between energy consumption and direction of travel, proxy for elevation profile).

More recent studies that have considered driver behaviour have concluded that it is one of the least influential parameters of BEB energy consumption and predominately operational and topological factors are identified as being most significant. When included within analysis route elevation/gradient is highlighted within the literature as significant. Work undertaken by (Abdelaty et al., 2021) examined seven machine learning models, which were applied to a full factorial design of BEB operation to emulate all possible conditions examining 15 factors including vehicular, operational, topological and external parameters. Of the seven models considered, a multiple linear regression model and support vector machine techniques yielded the most accurate predictions achieving R² values of 0.943 and 0.946 respectively. They identified that road gradient and battery SOC were the most significant parameters in influencing energy consumption.

Similarly work undertaken (Pamula and Pamula, 2020; Ma et al., 2021) identified route elevation as a significant parameter influencing BEB energy consumption as well as factors associated with average speed and number of stops. (Pamula and Pamula, 2020) examined the application of both deep learning networks (DLN) and multiple linear regression (MLR) to estimate energy consumption based on measured data of over 3000 BEB trips (studied vehicles were single deck of various configurations; lengths of 8 m, 9 m, 12 m and 18 m; battery size of 160-240 kWh and motor size 160–240 kW). Four variables were considered and all found to be significant (distance between stops, travel time, elevation differences and weather) with the authors proposing that the error value (18 % (DLN) and 22 % (MLR)) were acceptable for practical use. (Ma et al., 2021) examined energy consumption of BEB by further developing an existing model by (Fiori, Ahn and Rakha, 2016) which estimated instantaneous energy consumption of battery electric cars based on quasi-static backwards model and reported an average error of 5.9 % compared with empirical data. The model which already included vehicular, operational, topological and external parameters was updated to incorporate stop level ridership and driving profiles collected from a network of 630 routes. They estimated BEB energy consumption as being 1.42 ± 0.32 kWh/km. A gradient boosting regression tree algorithm was used to rank factors which influenced energy consumption of both diesel and electric buses. Average journey speed, route grade and number of stops where identified as having the largest influence of energy consumption of electric buses.

In studies where gradient is not considered as a parameter within the analysis, typically average speed was identified as being of largest significance (Lin, Lin and Ying, 2020; Wang et al., 2020). (Wang et al., 2020) applied 13 machine learning methods to establish the magnitude of influence and correlation between a range of variables on the measured operational energy consumption of 99 BEBs (single deck, 133 kWh battery and 100 kW motor) of and identified that random forests have the lowest modelling errors in both training and testing datasets, with the average driving speed being essential to battery efficiency (evaluated by the distance travelled per unit battery energy 1 % SOC). (Lin, Lin and Ying, 2020) gathered data from the battery management system of single deck BEBs operating on urban routes and applied machine learning methods, decision trees and random forests to develop energy consumption analytical models and examine the impact of driving behaviours. Vehicle average speed, average motor speed and total voltage were

L.A.W. Blades et al.  Transportation Research Part D 130 (2024) 104158
identified as the features having significant impact on driving energy consumption. (Li et al., 2021) also utilised random forest and cooperative k nearest neighbour (KNN) methods for estimating energy consumption of single deck BEBs, based on real world data from 163,800 journeys operating over a five month period in China. Combining kinematic and external factors they created a model with accuracy measurements of Mean Absolute Error (MAE) of 1.281, and Root Mean Square Error (RMSE) 1.614, however the relative significance of different factors was not discussed within the work.

As discussed by (Abdelaty and Mohamed, 2022) the nature of both physics-based and data-driven models hinder widespread implementation by transit operators or planners. The necessity of significant skill to operate the complex physics-based simulation models and the difficulties in obtaining operating field test data from zero emission vehicles means that decision makers are unable to effectively use these methods to inform decision making.

Based on the literature review undertaken the following gaps have been identified:

Within the literature, a range of different regression models and feature selection techniques have been used to explore energy consumption prediction. However, there is still no consensus on which methods are most reliable.

Given the variety of methods applied there are conflicting reports in literature on the relative importance of characteristics in predicting energy consumption under different operational and topographical conditions.

Finally for transit operators and planners, there is need for a method that can provide reliable estimations of energy consumption for new vehicle types on new routes with increased confidence in the estimates provided.

In review of the literature, there is no consistent framework for selecting what features/CPs should be used to predict the energy consumption of a BEB. A number of previous studies have used all available CPs as inputs which can lead to overfitting and higher computational costs. Other previous studies have only used one feature selection method with no/little rationale given for choosing this feature selection method. It would be beneficial to (a) evaluate the effectiveness of the more traditional approaches to feature selection, (b) the consistency in results obtained as a function of method(s) applied and (c) to identify a robust strategy for data analysis which would not require expert user input and reduce the likelihood of biasing emerging.

Consequently, this paper presents a novel strategy for energy prediction of battery electric bus operation. The approach uses a combination of feature selection methods together. This leads to a more robust strategy for feature selection while reducing the need for significant user input or the collection of high-resolution data, whilst improving model prediction accuracy and reducing the likelihood of biasing emerging. The novel strategy also allows multiple regression/prediction models to be compared to each other. Furthermore, the best performing energy prediction models can be selected leading to improved model prediction accuracy. Utilising the characteristic parameters extracted from a database of 605 drive cycles of BEBs operating in urban areas, this framework initially conducts feature selection, ranking the characteristic parameters by their relative influence on energy consumption. These parameters are then used to train energy prediction models which are initially constructed with the parameter of largest influence with additional parameters included until the calculated co-efficient of determination exceeds a threshold value. The overall quality of the model is then assessed using both synthetic and real data logged from battery electric buses already in operation.

3. Methodology

This section provides an overview of the vehicle modelling method, the vehicle being modelled, and the drive cycles being considered. The extraction of characteristic parameters from a drive cycle database and the development of the novel feature selection method framework to determine the most influential characteristic parameters for use in energy consumption prediction modelling is also described.

3.1. Vehicle modelling

The battery electric vehicle simulation model described by (McGrath et al., 2022) was used in this study to predict the energy consumption of the generic BEB. (Stevens et al., 2017, 2019; Murtagh et al., 2019; Doyle et al., 2020; Blades et al., 2022; McGrath et al., 2022) all used fully validated versions of the model for bus applications and have outlined the approach and strategy to modelling as well presenting the governing equations. Although the model is not suitable for detailed vehicle control and dynamic response evaluation as the modelling methodology neglects transients, it does provide an accurate prediction of macro quantities such as

![Vehicle architecture for the BEB.](image-url)
electrical energy consumption for a specified drive cycle. Fig. 1 shows the high level vehicle architecture for the BEB.

The baseline vehicle model has been updated to describe a generic BEB that is realistic for today’s marketplace, and has been validated against ZEB accredited buses certified on the UK Bus Cycle (Zemo Partnership, 2023) as well as against data collected from BEBs operating in Belfast and London. Fig. 2 is a flow chart showing an overview of the BEB model, including the key model inputs and outputs.

### 3.2. Vehicle specification

The BEB considered in this study is of the same specification as that modelled by (McGrath et al., 2022). Whilst this study models a generic vehicle specification, this methodology can be applied for any vehicle specification. Operating conditions were assumed based on those used to certify vehicles under the ZEB accreditation scheme (Zemo Partnership, 2022). The mass of the vehicle was assumed to be constant throughout the simulations. The heating load was assumed to be 0 kW, allowing the prediction analysis to be used for any ambient conditions. The associated energy consumption due to heating can be idealised as a constant power draw and added post simulation/prediction (Harris et al., 2018), provided the vehicle is operating within the standard defined operating conditions of the battery charge and discharge curves. Within the standard operating limits, for the range of ambient temperatures defined by these limits the charge and discharge curves are assumed to be the same. If UK based operation is considered, with average ambient temperatures in the range of 1.13 – 19.62 °C (UK MET Office, 2024), and if the vehicle is within the SOC window of the battery, operation will be well within the standard operating limits of the battery. Therefore, for this methodology, the energy consumption will be the same if the heat load is included during or added on after the simulation/prediction. Within the auxiliary loads, 1.25 kW is attributed to the battery thermal management system (BTMS). It is assumed that ambient temperature changes only affect the load required by the HVAC system to heat/cool the cabin, as the BTMS maintains the battery within an optimal temperature range, and so battery output is not sensitive to the ambient temperature. Table 1 summarises some of the key vehicle parameters, components, and characteristics that are used to describe the modelled vehicle.

### 3.3. Synthetic drive cycle creation

As discussed previously, drive cycles may be generated via real world logging or the creation of synthetic drive cycles. Real drive cycles require data loggers to be fitted to the physical vehicles, making it a time-consuming and expensive process to produce a large collection of drive cycles, across a large number of routes. The creation of synthetic drive cycles allows for vehicles to be modelled on any bus route in the world, without the need to collect data from a physical bus operating on that route. In this study, synthetic drive cycles were used in the drive cycle database to ensure diversity and dissimilarity in driving cycles. The process of synthetic drive cycle creation for city bus operation used in this work is described in literature by (Blades et al., 2022).

In order to train a regression model for accurate energy consumption prediction, as large a dataset as possible is required. In this study, the dataset is made up of 605 synthetic drive cycles, specifically created for bus routes. The drive cycles represent city bus operation both in the UK and worldwide, covering urban and extra-urban driving, as well as inter-city bus routes. The drive cycles provide coverage of a wide range of topographies and route characteristics. This will allow for feature selection methods to fully explore the route characteristics that are most influential to energy consumption, and for prediction models to be created that cover a diverse range of drive cycles.

To determine the energy consumption of the specified BEB for each of the drive cycles within the drive cycle database, the BEB vehicle model was used. For each drive cycle, the vehicle was simulated as conducting ten back-to-back repetitions of the cycle. This allows for the average energy consumption to be calculated across the SOC range of the vehicle battery. This is important because as the SOC decreases, the available energy in the cell is reduced, leading to a reduction in the open circuit voltage. To deliver the same power, an increased current is required, which results in higher I²R losses, and therefore an increased energy consumption.

Fig. 3 shows the probability distribution of the average energy consumption for all the drive cycles and Table 2 shows the descriptive stats of the average energy consumption. The red curve is a normal distribution using the mean and standard deviation shown in Table 2. Fig. 3 shows that the probability distribution of the average energy consumption is like a normal distribution. Table 2 shows that the values of the average energy consumption range from −0.103 to 1.785 kWh/km. The negative energy consumption values occur when the energy harnessed from regenerative braking is higher than the energy consumed by the BEB. The low skewness value in Table 2 shows that the average energy consumption distribution can be considered symmetric, however, the high kurtosis

![Fig. 2. Schematic of battery electric vehicle simulation including model inputs and outputs (McGrath et al., 2022).](image-url)
value shows that the distribution is more leptokurtic than a normal distribution.

3.4. Characteristic parameters

The CPs selected are all features of the route that can be determined from the time, velocity and elevation profiles of the drive cycle. The equations used to calculate the CPs, with the exception of average road gradient and elevation change, are described in the report published by (Barlow et al., 2009) based on drive cycles used in the measurement of road vehicle emissions. For each of the 605 drive cycles within the database, 24 CPs are extracted, a list of which are shown in Table 3.

The average road gradient is calculated using Equation (1) and the elevation change is calculated using Equation (2)

\[
\text{AverageRoadGradient(\%)} = \frac{E_{\text{End}} - E_{\text{Start}}}{\text{dist}} \times 100
\]  

\[Elevation\text{change(m)} = E_{\text{max}} - E_{\text{min}}\]
The CPs that can be extracted for a particular route depend upon the information that is available, such as velocity and elevation profiles for the drive cycle, timetable data for the bus service, and mapping/elevation data for the terrain over which the vehicle is being operated. In this work the CPs are divided into two scenarios, which are then used to train regression models for energy consumption prediction. Scenario 1 assumes that full drive cycle velocity and elevation profiles are available, therefore all of the CPs listed in Table 3 can be extracted. Scenario 2 assumes that basic timetable data and elevation/mapping data are available, without the full drive cycle velocity and elevation profile. Therefore, the only CPs that can be extracted are total distance, total time (based on the timetabled time for the route), average speed, average road gradient and elevation change. Elevation data for this work has been extracted using the Bing Maps Elevations API, which returns the elevation for given latitude/longitude coordinates. As the bus route timetabled time for the route), average speed, average road gradient and elevation change. Elevation data for this work has been extracted using the Bing Maps Elevations API, which returns the elevation for given latitude/longitude coordinates. As the bus route (and therefore route coordinates) are known, the elevation data for the coordinates can be extracted from the API to determine characteristics such as the average road gradient and elevation change using Equation (1) and Equation (2) respectively. The purpose of choosing these two scenarios is to demonstrate how the availability of data effects the prediction results obtained from the regression models. Table 3 shows the CPs assumed for each of the two scenarios.

### Table 3

<table>
<thead>
<tr>
<th>Characteristic Parameters</th>
<th>Symbol</th>
<th>Mean</th>
<th>Max</th>
<th>Min</th>
<th>Scenarios</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Distance (m)</td>
<td>dist</td>
<td>16,230</td>
<td>175,413</td>
<td>3699</td>
<td>1, 2</td>
</tr>
<tr>
<td>Total Time (s)</td>
<td>T_total</td>
<td>3002</td>
<td>39,466</td>
<td>603</td>
<td>1, 2</td>
</tr>
<tr>
<td>Average Speed (km/h)</td>
<td>v_ave</td>
<td>19.3</td>
<td>41.5</td>
<td>8.5</td>
<td>1</td>
</tr>
<tr>
<td>Average Drive Speed (km/h)</td>
<td>v_ave</td>
<td>23.9</td>
<td>45.2</td>
<td>13.6</td>
<td>1</td>
</tr>
<tr>
<td>Maximum Speed (km/h)</td>
<td>v_max</td>
<td>50.8</td>
<td>86.7</td>
<td>28.8</td>
<td>1</td>
</tr>
<tr>
<td>Average Road Gradient</td>
<td>Δv</td>
<td>−0.01</td>
<td>1.88</td>
<td>−2.09</td>
<td>1, 2</td>
</tr>
<tr>
<td>Elevation Change (m)</td>
<td>Elev</td>
<td>55.3</td>
<td>355.6</td>
<td>8.8</td>
<td>1, 2</td>
</tr>
<tr>
<td>Number of Stops</td>
<td>Stop</td>
<td>38</td>
<td>434</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Number of Stops per km</td>
<td>Stop/km</td>
<td>2.7</td>
<td>7.0</td>
<td>0.2</td>
<td>1</td>
</tr>
<tr>
<td>Time Stationary (s)</td>
<td>T_stop</td>
<td>612</td>
<td>12,852</td>
<td>37</td>
<td>1</td>
</tr>
<tr>
<td>Time Driving (s)</td>
<td>T_drive</td>
<td>2390</td>
<td>26,614</td>
<td>495</td>
<td>1</td>
</tr>
<tr>
<td>Time Cruising (s)</td>
<td>T_cruise</td>
<td>234</td>
<td>2282</td>
<td>36</td>
<td>1</td>
</tr>
<tr>
<td>Time Accelerating (s)</td>
<td>T_acc</td>
<td>1087</td>
<td>12,474</td>
<td>178</td>
<td>1</td>
</tr>
<tr>
<td>Time Decelerating (s)</td>
<td>T_dec</td>
<td>1068</td>
<td>11,858</td>
<td>153</td>
<td>1</td>
</tr>
<tr>
<td>Average Positive Acceleration (m/s²)</td>
<td>a_pos</td>
<td>0.49</td>
<td>0.65</td>
<td>0.28</td>
<td>1</td>
</tr>
<tr>
<td>Average Negative Acceleration (m/s²)</td>
<td>a_neg</td>
<td>−0.50</td>
<td>−0.33</td>
<td>−0.71</td>
<td>1</td>
</tr>
<tr>
<td>Average Acceleration (m/s²)</td>
<td>a_ave</td>
<td>4.94E-06</td>
<td>2.99E-03</td>
<td>−1.03E-17</td>
<td>1</td>
</tr>
<tr>
<td>Average Acceleration (Threshold)(m/s²)</td>
<td>a_thres</td>
<td>−1.52E-05</td>
<td>3.73E-03</td>
<td>−2.90E-03</td>
<td>1</td>
</tr>
<tr>
<td>Accelerations per km</td>
<td>a_ave</td>
<td>141.3</td>
<td>242.5</td>
<td>56.8</td>
<td>1</td>
</tr>
<tr>
<td>Percentage Time Driving</td>
<td>%drive</td>
<td>79.8</td>
<td>94.8</td>
<td>52.5</td>
<td>1</td>
</tr>
<tr>
<td>Percentage Time Cruising</td>
<td>%cruise</td>
<td>8.6</td>
<td>36.9</td>
<td>2.5</td>
<td>1</td>
</tr>
<tr>
<td>Percentage Time Accelerating</td>
<td>%acc</td>
<td>36.0</td>
<td>43.2</td>
<td>21.6</td>
<td>1</td>
</tr>
<tr>
<td>Percentage Time Decelerating</td>
<td>%dec</td>
<td>35.2</td>
<td>42.9</td>
<td>18.5</td>
<td>1</td>
</tr>
<tr>
<td>Percentage Time Stationary</td>
<td>%stop</td>
<td>20.2</td>
<td>47.5</td>
<td>5.2</td>
<td>1</td>
</tr>
</tbody>
</table>

The CPs that can be extracted for a particular route depend upon the information that is available, such as velocity and elevation profiles for the drive cycle, timetable data for the bus service, and mapping/elevation data for the terrain over which the vehicle is being operated. In this work the CPs are divided into two scenarios, which are then used to train regression models for energy consumption prediction. Scenario 1 assumes that full drive cycle velocity and elevation profiles are available, therefore all of the CPs listed in Table 3 can be extracted. Scenario 2 assumes that basic timetable data and elevation/mapping data are available, without the full drive cycle velocity and elevation profile. Therefore, the only CPs that can be extracted are total distance, total time (based on the timetabled time for the route), average speed, average road gradient and elevation change. Elevation data for this work has been extracted using the Bing Maps Elevations API, which returns the elevation for given latitude/longitude coordinates. As the bus route (and therefore route coordinates) are known, the elevation data for the coordinates can be extracted from the API to determine characteristics such as the average road gradient and elevation change using Equation (1) and Equation (2) respectively. The purpose of choosing these two scenarios is to demonstrate how the availability of data effects the prediction results obtained from the regression models. Table 3 shows the CPs assumed for each of the two scenarios.

### 3.5. Feature selection framework

Five regression techniques were applied to the collated dataset of 605 drive cycles; linear regression, lasso regression, ridge regression, random forests and extra trees.

For the linear regression model the input variables/CPs were normalised so that the magnitude of the equation coefficients can be used to ascertain the influence that a CP has on the vehicle energy consumption. Unlike linear regression, lasso and ridge regression contain a model parameter, λ, which penalises coefficient terms with a high magnitude. By reducing the size of the model coefficients, model bias increases but variance decreases. Lasso regression forces weak variables (CPs that have little impact on the energy consumption) to have coefficients of zero magnitude. The CPs were ranked by steadily increasing the value and ascertaining when the co-efficient reached a value of zero first, whereas the CP whose co-efficient was the last to reach a value of zero was ranked first.

Unlike lasso regression, ridge regression will not force coefficients to a value of zero. Instead, the coefficient magnitudes are decreased until they converge at a given value. For ridge regression, λ was increased until the coefficients converged. Afterwards, the CPs were ranked according to their coefficient magnitudes.

Unlike the previously mentioned regression techniques, random forests and extra trees are capable of modelling non-linear relationships and therefore were also selected. Random forests act as an ensemble by using a large number of uncorrelated decision trees (Strobl et al., 2007) which reduces the bias and variance of predictions (Li et al., 2021). Each input for a random forests model has a property called information gain which assess how much impact each model input has on the model output. Information gain was used to rank the CPs. Random forests do not overfit as more trees are added and so model accuracy increases as trees are added (Breiman, 2001; Breiman and Cutler, 2004). However, the downside of adding trees is increased running time. In this analysis, 200 trees were used to create the random forest because this number exceeded the recommended 128 trees (Oshiro, Perez and Baranauskas, 2012), without the running time becoming unmanageable.
Finally, extra trees also act as an ensemble using a large number of uncorrelated decision trees. However, compared to random forests, extra trees is a faster algorithm. Also, compared to random forest, the use of extra trees leads to a decrease in variance but an increase in bias (Geurts, Ernst and Wehenkel, 2006). In this analysis, 200 trees were used for the same reasons that 200 trees were used to create the random forest.

Three wrapper methods were used alongside the regression methods to rank the CPs in terms of impact on vehicle energy consumption: recursive feature elimination (RFE), sequential forwards selection (SFS) and sequential backwards selection (SBS).

RFE initially creates a regression model using all the CPs. The CP with the smallest coefficient is removed and a new regression model is created. This process is repeated until only one CP remains. CPs are then ranked depending on what order they are removed from the process.

Within SFS, a number of regression models are initially created, each one containing one of the CPs as an input. Each of these models are scored in terms of the coefficient of determination ($R^2$), and the CP that performed the best is chosen. In the second step, the first CP is chosen in combination with all other CPs, with the combination of two CPs achieving the best results is chosen. This process happens iteratively, until all CPs are chosen. CPs are ranked depending on when they were added. SBS is the opposite of SFS, with one CP removed each time depending on the accuracy of the regression models.

![Fig. 4. Regression process algorithm.](image-url)
For SFS and SBS the scoring method was to take an average score of $R^2$ from a five-fold cross-validation. Both ridge and lasso regression had a grid search carried out at each step to ascertain which $\lambda$ provided the best accuracy for each model.

The use of wrapper methods inherently reduces issues associated with multicollinearity, removing redundant features that do not provide unique information to the model. This removes the need for a correlation analysis, where the strength of relationship between two variables is assessed and preselecting variables for inclusion in models. As the correlation process requires a level of user expertise, the application of wrapper methods within the feature selection framework ensures that the optimal model will be used given the available information.

To extract meaning from these results, the algorithm in Fig. 4 above was implemented. This algorithm first selected a feature selection method and started with the top ranked CP for that feature selection method used as the input. Afterwards, the relevant regression method was chosen and the data along with a cross validation of 5 was used to output an $R^2$ value. The relevant regression method depends on the feature selection method. For example, if the feature selection method was based on linear regression, then linear regression would be the relevant regression method. Afterwards, the top two ranked CPs were used as the input and an $R^2$ value was calculated using a cross validation of 5. This process continues until all CPs are used. Afterwards the next feature selection method is chosen, and the process is repeated for this method. This happens until all feature selection methods are assessed.

Afterwards, the optimum CP subset and regression models are selected. The optimum regression model should have a high $R^2$ value while having a low number of inputs. $R^2$ values of 0.7, 0.75, 0.8, 0.85, 0.9, 0.95 and 1 were examined. All the regression models were examined to see if they reached an $R^2$ of 1. If this condition is met, then the regression models that reached this value with the smallest number of inputs are selected. These regression models are compared to each other and if they all contain the same inputs then these inputs are deemed the optimum CP subset. This is because the more feature selection methods there are that choose the same CP subset the higher the consensus that the CP subset is the optimum subset. However, if an $R^2$ value of 1 is not met or if the regression models select different inputs from each other, then an $R^2$ value of 0.95 is considered and the process is repeated. If the conditions fail for an $R^2$ value of 0.95 then the other $R^2$ values mentioned before (0.7, 0.75, 0.8, 0.85 and 0.9) are also examined.

To summarise, the algorithm in Fig. 4 enables the optimum CP subset to be selected based on a high $R^2$ value being reached while having the lowest number of CPs possible. This is because a high $R^2$ value would be indicative of a higher model accuracy and a low number of CPs allows for better model interpretation and less chance of overfitting. The algorithm also allows for a range of feature selection methods to be compared and if there is consensus among the best performing feature selection methods on the optimum CP subset then the user has greater confidence in this chosen CP subset. This information can also be used to select the best performing regression models for energy consumption prediction.

3.6. Sensitivity analysis

Sensitivity analysis was carried out to ascertain how much influence each CP had on the energy consumption for the Linear, Lasso and Ridge regression models. Sobol analysis, which is a global variance-based sensitivity analysis method, was used on each of the three regression models. Sobol analysis decomposes the variance of the output of the model into fractions which can be attributed to inputs or sets of inputs. A direct variance-based measure of sensitivity, $S_i$, shows the contribution to the output variance of the main effect of input $X_i$. Equation (3) shows how $S_i$ is calculated, where $Y$ is the model output and $X_i$ is a model input.

$$S_i = \frac{\text{Var}(E(\frac{Y}{X_i}))}{\text{Var}(Y)}$$

(3)

$S_i$ is referred to as a first-order sensitivity index. Higher order interaction indices, $S_{ij}$, $S_{ik}$, etc. between model inputs can also be calculated. This would involve Equation (3) being turned into Equation (4) where $X_{ij}$ involves an interaction between input $X_i$ and $X_j$.

$$S_{ij} = \frac{\text{Var}(E(\frac{Y}{X_i X_j}))}{\text{Var}(Y)}$$

(4)

The sum of all the first order and higher order sensitivity indices would be equal to one. Sobol sampling (Sobol, 1967) and the Saltelli estimator (Saltelli et al., 2007) were used to estimate the indices in this work.

3.7. Prediction model validation

To assess the accuracy of the prediction models, six drive cycles which are not included in the database used within the regression analysis, are considered. For each of the drive cycles, the average energy consumption for the BEB is simulated using ten back-to-back repetitions of each cycle. The CPs are then extracted for each of the two scenarios described in section 3.4, and energy consumption predictions made using the equations generated from the regression analysis. The percentage difference between simulated average energy consumption from the vehicle models and predicted average energy consumption from the regression models are then calculated. Conclusions are drawn on the accuracy of the prediction model for each of the regression models trained assuming the two scenarios.

The six drive cycles selected to assess the prediction models include three industry standard drive cycles and three real drive cycles. The three industry standard drive cycles chosen are the Braunschweig City Driving Cycle, the Manhattan Bus Cycle, and the Orange.
County Bus (OC Bus) Cycle. These drive cycles were specifically created for chassis dynamometer testing of buses and aim to simulate the transient driving schedule of buses operating in urban areas and have been commonly used in literature to compare and determine energy consumption for vehicle modelling (Gao et al., 2017; Kivekas et al., 2018; Lajunen et al., 2018). The three real drive cycles are taken from the limited number of cycles that have been logged on board real BEB double deck buses operating in Belfast and London. The real drive cycle logged in Belfast is bus route 3A, whilst the real drive cycles logged in London are the Transport for London (TfL) bus routes 7 and U5. The real drive cycles consist of both velocity and elevation profiles, whilst the industry standard drive cycles consist of only a velocity profile. Therefore, the industry standard cycles are assumed to operate on a flat road with an elevation of 0 m throughout.

The validation method used will allow the prediction model accuracy to be assessed for use on real drive cycles. This in turn will allow a comparison to be made between the use of synthetic drive cycles and real drive cycles in vehicle modelling.

4. Results

4.1. BEB energy consumption prediction model

4.1.1. Scenario 1

For Scenario 1, all 24 of the CPs extracted from each drive cycle within the drive cycle database were ranked in terms of their influence on vehicle average energy consumption. For each of the 20 feature selection methods, the first step was to rank the 24 CPs from best to worst in terms of the influence that they have on the average energy consumption of the vehicle. The CPs were normalised for the reasons given in Section 3.5. Fig. 5 shows the median ranking across all of the feature selection methods for the top 10 CPs for the BEB. The values are numbered and colour coded for each feature selection method, with light blue showing the highest ranked CP and dark blue showing the lowest ranked CP. It can be seen that for 18 of the 20 feature selection methods, average road gradient had the greatest influence on the BEB energy consumption.

The regression algorithm in Fig. 4 was used to calculate the relevant $R^2$ values for each feature selection method. The $R^2$ values are shown in Fig. 6 with the column headings showing each feature selection method and the index showing how many CPs were used as inputs for each regression model. For example, the top row of Fig. 6 shows the $R^2$ values given when the top CP for each feature selection method was used as the input along with the relevant regression model for that feature selection method. Whereas the second row of Fig. 6 shows the $R^2$ values when the top two CPs for each selection method were used along with the relevant regression model. Fig. 6 shows that an $R^2$ value of 0.95 or 1 was never reached. Fig. 6 also shows that three feature selection methods (SFS Linear Regression, SFS Lasso Regression and SFS Ridge Regression) were able to reach an $R^2$ value of 0.9 with only three CPs. These three methods all chose the same three CPs, in the same order of influence; 1. Average Road Gradient, 2. Number of Stops per km, 3. Maximum Speed. The variance inflation factor (VIF) was calculated for these three CPs. Average road gradient, number of stops per km and maximum speed had VIF values of 1.00, 3.20 and 3.20, respectively. As all VIF values were below 5.00, it can be concluded that the multicollinearity between these three CPs was not high. This shows that the feature selection framework in this work has succeeded in selecting a subset of CPs with a low amount of multicollinearity.

Consequently, the optimum subset consisted of these three CPs and the optimum regression models can also be selected. As the three feature selection methods were based on Linear, Lasso and Ridge regression then the optimum regression models were Linear, Lasso and Ridge regression models that used the optimum subset of three CPs as their inputs. The Linear, Lasso and Ridge regression model equations for the calculation of energy consumption, $Y$ (in kWh/km), for the BEB modelled in this work are shown in Equation (5), Equation (6) & Equation (7) respectively.

**Linear Regression Model:**

$$Y = \beta_0 + \beta_1 \text{Average Road Gradient} + \beta_2 \text{Number of Stops per km} + \beta_3 \text{Maximum Speed}$$

**Lasso Regression Model:**

$$Y = \beta_0 + \sum_{i=1}^{3} \beta_i \text{CP}_i$$

**Ridge Regression Model:**

$$Y = \beta_0 + \sum_{i=1}^{3} \beta_i \text{CP}_i + \lambda \sum_{i=1}^{3} \beta_i^2$$

Fig. 5. BEB - Scenario 1 - Characteristic parameter ranking for each feature selection method.
Y = 0.579 + 46.0\nabla v_{av} + 0.0662 \text{Stop}_{km} + 0.00311 v_{max} \quad (5)

Lasso Regression Model:
Y = 0.579 + 44.7\nabla v_{av} + 0.0660 \text{Stop}_{km} + 0.00311 v_{max} \quad (6)

Ridge Regression Model:
Y = 0.579 + 44.8\nabla v_{av} + 0.0661 \text{Stop}_{km} + 0.00311 v_{max} \quad (7)

Fig. 6 shows the $R^2$ values using five-fold cross validation across the 605 synthetic cycles. The three optimum feature selection methods were SFS Linear Regression, SFS Lasso Regression and SFS Ridge Regression and the top three CP inputs were used for each method. This means that the Linear, Lasso and Ridge Regression models had five-fold cross validation $R^2$ values of 0.916, 0.917 and 0.917 respectively. However, these scores were derived using synthetic cycles and so it was decided to use real and standard cycles to validate the models as shown in Section 4.2.1.

The performance of the regression model was assessed by plotting the predicted energy consumption vs the simulated energy and P-P plots for the residuals (Supplementary Material). The regression model was shown to be accurately estimating energy consumption, with residuals normally distributed. Sensitivity analysis conducted for the Linear Regression model showed that the first order indices for the average road gradient, number of stops per km and maximum speed were 0.6122, 0.3065 and 0.0837, respectively. For the Lasso Regression model the first order indices for the average road gradient, number of stops per km and maximum speed were 0.5994, 0.3163, 0.0867, respectively. For the Ridge Regression model the first order indices for the average road gradient, number of stops per km and maximum speed were 0.6005, 0.3155, 0.0864, respectively. As all three regression models did not contain any interaction terms only first order indices were inspected.
4.1.2. Scenario 2

For Scenario 2, it was assumed that mapping/elevation data for each of the drive cycles was available as well as the timetable data, therefore, five CPs (Table 3) were extractable for the training of the energy consumption models. Fig. 7 shows the median ranking across all the feature selection methods for the five CPs. Like Scenario 1 the CPs were normalised before the CPs were ranked. The average road gradient was the CP with the greatest influence on the average energy consumption, according to the median ranking.

The regression algorithm in Fig. 4 was used to calculate the relevant R² values for each feature selection method. The R² values are shown in Fig. 8 with the column headings showing each feature selection method and the index showing how many CPs were used as inputs for each regression model. Fig. 8 shows that R² values of 0.9, 0.95 or 1 were never reached. Fig. 8 also shows that six feature selection methods (Random Forests (RFE and SBS) and Extra Trees (Extra Trees, RFE, SBS and SFS Extra Trees)) were able to reach an R² value of 0.85 with only two CPs. These six methods all chose the same two CPs, in the same order of influence; 1. Average Road Gradient, and 2. Average Speed. The variance inflation factor (VIF) was calculated for these two CPs. Both average road gradient and average speed had VIF values of 1.00. As all VIF values were below 5.00, it can be concluded that the multicollinearity between these two CPs was not high. Like Scenario 1, this shows that the feature selection framework in this work has succeeded in selecting a subset of CPs with a low amount of multicollinearity. Consequently, the optimum subset consisted of these two CPs and the optimum regression models can also be selected. As the six feature selection methods were based on Random Forests and Extra Trees then the optimum regression models were Random Forests and Extra Trees regression models that used the optimum subset of two CPs as their inputs. The Random Forests and Extra Trees prediction models cannot be defined by simple equations as they are each made up of 200 decision trees.

Fig. 8 shows the R² values using five-fold cross validation across the 605 synthetic cycles. The six optimum feature selection methods (Random Forests (RFE and SBS) and Extra Trees (Extra Trees, RFE, SBS and SFS Extra Trees)) and the top two CP inputs were used for each method. This means that the Random Forests model had a five-fold cross validation R² score ranging from 0.850 to 0.851 and the Extra Trees model had a five-fold cross validation R² score ranging from 0.852 to 0.856. However, these scores were derived using synthetic cycles and so it was decided to use real and standard cycles to validate the models as shown in Section 4.2.1.

As before the regression model was shown to be accurately estimating energy consumption, with residuals normally distributed (predicted vs actual energy consumption and residual P-P plots included in Supplementary Material).

For the Random Forests model the first order indices for the average road gradient and average speed were 0.6642 and 0.3021, respectively. There was also an interaction index between these two inputs of 0.0283. For the Extra Trees model the first order indices for the average road gradient and average speed were 0.6422 and 0.3021, respectively. There was also an interaction index between these two inputs of 0.0283.

4.2. Model validation results

4.2.1. Energy consumption prediction accuracy

Table 4 shows the predicted average energy consumption for each of the prediction models developed for the two CP Scenarios, as well as the simulated average energy consumption from the generic BEB double deck model, for each of the six drive cycles selected for model validation. To compare the prediction to the simulated results, percentage difference was calculated. The percentage difference, including the mean for each prediction model and Scenario are shown in Table 5 for the industry standard drive cycles, and in Table 6 for the real drive cycles. Values of positive percentage difference indicate predicted average energy consumptions higher than the simulated value, while negative percentage differences indicate predicted average energy consumptions lower than the simulated value.

As can be seen from Table 5, for the industry standard drive cycles, the mean absolute percentage difference across the two scenarios for all the regression models considered ranges from 2.12 % to 10.67 %. For Scenario 1, where the prediction models were trained using the most influential of the 24 available CPs, across the three standard drive cycles the percentage difference for the Linear Regression model (which proved to be most accurate) ranged from 0.33 % to 3.29 % (absolute values), with a mean of 2.12 %. For Scenario 2, where five CPs extracted from basic timetable data and elevation data were used to train the prediction models, the Extra Trees prediction model was the most accurate with a mean absolute percentage difference of 9.03 %.

Table 6 shows the percentage difference between the predicted and simulated average energy consumptions for the four real drive cycles considered in this study. The Lasso Regression prediction model was shown to perform most accurately for Scenario 1, with a mean absolute percentage difference of 2.10 % across the four drive cycles. For Scenario 2, both the prediction models showed mean absolute percentage differences less than 5 %, with the Extra Trees model performing most accurately (3.57 %).

<table>
<thead>
<tr>
<th>Linear Reg.</th>
<th>RFE (Lasso Reg.)</th>
<th>SBS (Lasso Reg.)</th>
<th>SFS (Lasso Reg.)</th>
<th>Ridge Reg.</th>
<th>RFE (Ridge Reg.)</th>
<th>SBS (Ridge Reg.)</th>
<th>SFS (Ridge Reg.)</th>
<th>Random Forests</th>
<th>RFE (Random Forests)</th>
<th>SBS (Random Forests)</th>
<th>SFS (Random Forests)</th>
<th>Extra Trees</th>
<th>RFE (Extra Trees)</th>
<th>SBS (Extra Trees)</th>
<th>SFS (Extra Trees)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.674</td>
<td>0.674</td>
<td>0.674</td>
<td>0.676</td>
<td>0.676</td>
<td>0.676</td>
<td>0.676</td>
<td>0.527</td>
<td>0.525</td>
<td>0.525</td>
<td>0.462</td>
<td>0.480</td>
<td>0.463</td>
<td>0.463</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.670</td>
<td>0.670</td>
<td>0.610</td>
<td>0.810</td>
<td>0.812</td>
<td>0.812</td>
<td>0.812</td>
<td>0.812</td>
<td>0.849</td>
<td>0.850</td>
<td>0.851</td>
<td>0.849</td>
<td>0.854</td>
<td>0.855</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.813</td>
<td>0.728</td>
<td>0.828</td>
<td>0.828</td>
<td>0.830</td>
<td>0.830</td>
<td>0.830</td>
<td>0.832</td>
<td>0.832</td>
<td>0.832</td>
<td>0.831</td>
<td>0.831</td>
<td>0.863</td>
<td>0.867</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.829</td>
<td>0.829</td>
<td>0.830</td>
<td>0.830</td>
<td>0.832</td>
<td>0.832</td>
<td>0.832</td>
<td>0.831</td>
<td>0.831</td>
<td>0.831</td>
<td>0.831</td>
<td>0.832</td>
<td>0.854</td>
<td>0.873</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.834</td>
<td>0.834</td>
<td>0.834</td>
<td>0.834</td>
<td>0.834</td>
<td>0.834</td>
<td>0.834</td>
<td>0.836</td>
<td>0.836</td>
<td>0.836</td>
<td>0.836</td>
<td>0.836</td>
<td>0.864</td>
<td>0.874</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 8. BEB – Scenario 2 – Evaluation of regression models using R² considering feature selection method used and number of inputs.
It can be seen from the results that for Scenario 1 the models produced perform very well for the prediction of average energy consumption for both the industry standard and real drive cycles. This is to be expected as the feature selection methods have the largest number of CPs to choose from to train the prediction models, with full velocity and elevation profiles available. For Scenario 2, whilst predictions for most of the drive cycles have a percentage difference of less than 10 % when compared to the simulations, the models produced are more accurate for the real drive cycles than for the industry standard cycles. This can be explained by the fact that the industry standard drive cycles do not have an elevation profile, and so the average road gradient is considered to be zero. The Random Forests and Extra Trees prediction models produced in this study consider only 2 CPs; the average road gradient and the average speed. In terms of average road gradient, the drive cycle database used to produce the prediction models is not representative of the industry drive cycles, as all of the synthetic cycles that make up the database have elevation profiles. Therefore, only 1 of the 2 CPs used in Scenario 2 in the prediction for the industry standard cycles are in line with the drive cycle database.

For the real drive cycles, the results show that the energy consumption predictions are very accurate for both Scenario 1 and Scenario 2. This shows that an accurate energy consumption prediction model for a generic battery electric bus has been produced by considering only basic timetable and elevation data, without the need for producing detailed drive cycles with a velocity profile. The results also show that the synthetic drive cycles produced and used for the drive cycle database are representative, and accurately describe the real drive cycles logged on-board operating battery electric vehicles. This study shows that, in the absence of a physical vehicle operating on a specified bus route, an energy prediction model is a low-cost, computationally inexpensive and fast alternative that can be developed using a database of synthetic drive cycles to accurately describe the energy requirements of a BEB.

4.2.2. Energy consumption comparison to literature

The average energy consumption results described so far in this study have all been in the absence of an HVAC load. In this section a heating load is assumed and added onto the simulated average energy consumptions so that the model results can be compared alongside those seen in literature and in industry. The power draw for the heating load was assumed to be 1.57 kW, based on an ambient temperature of 10 °C, derived from the same Recovery Heat Pump model considered by (Blades et al., 2022; McGrath et al., 2022).

The energy consumption associated with heating for each of the drive cycles is calculated by multiplying the heating load in kW by the time duration (in hours) of the drive cycle, and then dividing the result by the distance of the drive cycle (in km). Table 7 shows the simulated energy consumption with no heat load, the energy consumption associated with heat load, and the total energy consumption

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Prediction (kWh/km)</th>
<th>Scenario 1</th>
<th>Scenario 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Linear</td>
<td>Lasso</td>
</tr>
<tr>
<td>Braunschweig</td>
<td>0.9397</td>
<td>0.9366</td>
<td>0.9365</td>
</tr>
<tr>
<td>Manhattan</td>
<td>1.1401</td>
<td>1.1032</td>
<td>1.1025</td>
</tr>
<tr>
<td>Orange Co.</td>
<td>0.9507</td>
<td>0.9772</td>
<td>0.9770</td>
</tr>
<tr>
<td>Belfast 3A</td>
<td>0.8518</td>
<td>0.8220</td>
<td>0.8249</td>
</tr>
<tr>
<td>TfL U5</td>
<td>0.8162</td>
<td>0.8067</td>
<td>0.8078</td>
</tr>
<tr>
<td>TfL 7</td>
<td>0.9813</td>
<td>0.9601</td>
<td>0.9613</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 5</th>
<th>Percentage difference between regression models and standard duty cycles for BEB.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Scenario 1</td>
</tr>
<tr>
<td></td>
<td>Linear</td>
</tr>
<tr>
<td>Braunschweig</td>
<td>0.33</td>
</tr>
<tr>
<td>Manhattan</td>
<td>3.29</td>
</tr>
<tr>
<td>Orange Co.</td>
<td>-2.75</td>
</tr>
<tr>
<td>Mean (Absolute)</td>
<td>2.12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 6</th>
<th>Percentage difference between regression models and real duty cycles for BEB.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Scenario 1</td>
</tr>
<tr>
<td></td>
<td>Linear</td>
</tr>
<tr>
<td>Belfast 3A</td>
<td>3.56</td>
</tr>
<tr>
<td>TfL U5</td>
<td>1.17</td>
</tr>
<tr>
<td>TfL 7</td>
<td>2.18</td>
</tr>
<tr>
<td>Mean (Absolute)</td>
<td>2.30</td>
</tr>
</tbody>
</table>
including heat load for all of the real and industry standard drive cycles for the BEB.

From Table 7 the average energy consumption of the generic double deck BEB ranges from 0.8904 to 1.2829 kWh/km across all of the drive cycles when heating is considered. These are values are in general agreement with previous literature including those reported by (Ma et al., 2021) 1.42 ± 0.32 kWh/km, (Gao et al., 2017) 1.24–2.48 kWh/km and 1.75 ± 0.41 kWh/km (Gallet, Massier and Hamacher, 2018). These values can also be compared to the Zero Emission Bus certificates of the BEBs that are certified as Zero Emission Buses in the UK. For the double deck electric buses certified the energy consumptions over the UK Bus Cycle ranges from 0.68 kWh/km to 1.14 kWh/km. With the exception of the Manhattan Bus Cycle which consists of intensive city driving, the BEB operating on all of the standard and real drive cycles considered in this study lie within the range of energy consumption of the certified double deck BEBs. The simulated values of average energy consumption for the generic double deck BEB, with heating included, are in line with those achieved by double deck BEBs operating on our roads.

2.4.3. CP impact on energy consumption comparison to literature

Table 8 shows a comparison of CP impact on energy consumption of BEB operation estimated for Scenario 1 and Scenario 2 of this study against literature values. (Kivekas et al., 2018; Vepsäläinen et al., 2018; Abdelaty et al., 2021; Abdelaty and Mohamed, 2022) all used variance-based sensitivity analysis to analyse the impact CPs based on first order indices only as shown in Table 8. However, (Abdelaty et al., 2021) showed the total-effect index of their sampled-based sensitivity analysis and so their total-effect indices are shown instead. (Kivekas et al., 2018) performed sensitivity analyses for a synthetic cycle and 19 reference cycles. For the 19 reference cycles, none of the CPs had a high first order effect so only the synthetic cycle analysis is shown. (Lin, Lin and Ying, 2020; Ma et al., 2021) did not use variance-based sensitivity analysis but used the feature weights of their random forests and gradient boosted decision trees models, respectively.

For Scenario 1, Table 8 shows for all three regression models used in this study that average road gradient was the CP with the largest impact, with a Sobol index ranging from 0.5994 to 0.6122. Three studies in the literature also considered average road gradient. Two of these studies (Abdelaty et al., 2021; Abdelaty and Mohamed, 2022) had average road gradient as the largest impact, with Sobol indices of ~0.9500 and 0.8995, respectively. However, (Ma et al., 2021) had a feature weight of 0.2257 for average road gradient which was second behind average speed with a feature weight of 0.2527. Stops per km had the second biggest Sobol index, in the range of 0.3065 to 0.3155 which is larger than the Sobol indices reported by (Abdelaty et al., 2021; Abdelaty and Mohamed, 2022) of ~0.0500 and 0.0087, respectively. (Ma et al., 2021) did not consider number of stops per km but did look at its inverse which was average stop distance. (Ma et al., 2021) had a feature weight of 0.1063 for average stop distance, however, they also included number of stops in their analysis which would have a high correlation with average stop distance. If number of stops was removed from their model, then average stop distance would have a higher feature weight in their regression model. For Scenario 1 it was found that maximum speed had a Sobol index in the range of 0.0837 to 0.0867. No studies in the literature had scores for maximum speed.

Examining Scenario 2 the CP with the largest impact was again average road gradient with Sobol indices in the range of 0.6642 to 0.6726. The two regression models used for Scenario 2 both had a Sobol index of 0.3021 for the average speed. This is far greater than the Sobol indices given by (Abdelaty et al., 2021; Abdelaty and Mohamed, 2022) which were ~0.0200 and 0.0075, respectively. However, it was closer in value to the feature weight of 0.2527 shown in the study by (Ma et al., 2021). Contrarily, (Ma et al., 2021) had average velocity as having the biggest impact on energy consumption followed by average road gradient whereas the Scenario 2 analysis in this work found that average road gradient had the biggest impact followed by average velocity.

5. Conclusions

This paper develops and assesses prediction models for the energy consumption of a generic double deck BEB vehicle operating on any given drive cycle. The prediction models produced in this study have a mean absolute percentage difference accuracy between 2.10 % and 10.67 %.

This work has shown that there is no one best regression method for predicting energy consumption. Rather, the best regression method is influenced by the available input information. When all 24 CPs were available linear, lasso and ridge regression models were selected as the optimum models. However, when only basic timetable data and elevation/mapping data was available only 5 CPs could be used as inputs. Random forests and extra trees were selected as the optimum models when only these 5 CPs were available. The prediction models developed in this work can be used to ensure accurate predictions of the operational energy demand of battery electric buses is delivered to operators in a computationally inexpensive manner, without the need to run vehicles on the bus route.

Table 7

<table>
<thead>
<tr>
<th>Drive Cycle</th>
<th>Energy Consumption (kWh/km)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Simulation</td>
</tr>
<tr>
<td>Braunschweig</td>
<td>0.9397</td>
</tr>
<tr>
<td>Manhattan</td>
<td>1.1401</td>
</tr>
<tr>
<td>Orange Co.</td>
<td>0.9507</td>
</tr>
<tr>
<td>Belfast 3A</td>
<td>0.8518</td>
</tr>
<tr>
<td>TIL U5</td>
<td>0.8162</td>
</tr>
<tr>
<td>TIL 7</td>
<td>0.9813</td>
</tr>
</tbody>
</table>
Table 8
CP impact on energy consumption for Scenario 1 & 2 compared to literature values.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Driver Aggressiveness</td>
<td>-0.1000</td>
<td>-0.0500</td>
<td>-0.0500</td>
<td>X</td>
<td>X</td>
<td>0.0098</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Passenger Load</td>
<td>0.0664</td>
<td>X</td>
<td>-0.0800</td>
<td>X</td>
<td>-</td>
<td>0.0132</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>DCDC Convertor Pump</td>
<td>X</td>
<td>-0.1200</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Number of Stops per km</td>
<td>-</td>
<td>-0.0800</td>
<td>-0.0500</td>
<td>X</td>
<td>X</td>
<td>0.0087</td>
<td>0.3065</td>
<td>0.3163</td>
<td>0.3155</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Ambient Temperature</td>
<td>X</td>
<td>-0.0500</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Average Road Gradient</td>
<td>X</td>
<td>X</td>
<td>-0.9500</td>
<td>X</td>
<td>0.2257</td>
<td>0.8995</td>
<td>0.6122</td>
<td>0.5994</td>
<td>0.6005</td>
<td>0.6726</td>
<td>0.6642</td>
</tr>
<tr>
<td>Initial State of Charge</td>
<td>X</td>
<td>-</td>
<td>-0.1500</td>
<td>-</td>
<td>X</td>
<td>0.0284</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Road Condition HVAC</td>
<td>X</td>
<td>X</td>
<td>-0.1200</td>
<td>X</td>
<td>X</td>
<td>0.151</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Average Speed</td>
<td>-</td>
<td>-</td>
<td>-0.0200</td>
<td>-0.1400</td>
<td>0.2527</td>
<td>0.0075</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.3021</td>
<td>0.3021</td>
</tr>
<tr>
<td>Drag Coefficient</td>
<td>X</td>
<td>X</td>
<td>-0.0100</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Motor Speed</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>-0.1200</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Total Voltage</td>
<td>X</td>
<td>X</td>
<td>-0.1000</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Number of Stops</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>0.1896</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Average Stop Distance</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>0.1063</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Maximum Speed</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>0.0837</td>
<td>0.0867</td>
<td>0.0864</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

CP values graphically presented in a study shown with ‘~’. CP not considered within a study are shown as ‘X’.
Where a CP was considered but was not selected for the final sensitivity analysis or its impact is considered too low in a study, shown with a ‘-’.
This is essential to ensure that operators can make data-driven, informed decisions on when transitioning their bus fleets to zero emission technologies.

The vehicle modelled in this paper is a generic double deck BEB. The modelling in this paper has assumed a 320 kWh lithium manganese cobalt oxide (NMC) battery. Operation of the vehicle was simulated for 605 different synthetic drive cycles, and results were used to rank the influence of 24 different characteristic parameters (CPs) on average energy consumption for a number of feature selection methods. Prediction models were then trained based on two different scenarios, with Scenario 1 considering all CPs, and Scenario 2 considering timetable and elevation data. For both Scenario 1 and Scenario 2, the majority of the feature selection methods showed that the average road gradient was the most influential route based CP for energy consumption, which is similar to findings reported in other literature, with Sobol indices in the range of 0.5994 – 0.6726 across the two scenarios.

It was found that, generally, as more CPs are available in the dataset the prediction models perform better as there are a wider range of CPs to choose from. When considering the real drive cycles used to test the prediction models, Scenario 1 was shown to perform the best, with the Lasso Regression model for the BEB having a mean absolute percentage difference of 2.10 %. However, even though Scenario 2 only considered 5 CPs that can be extracted from basic timetable and mapping data, compared to 24 for Scenario 1, due to the presence of elevation data, accurate prediction were still achieved (3.57 % average absolute percentage difference for the Extra Trees model). Scenario 2 has shown that if public transport operators can provide only the start and end points of the drive cycle, as well the drive cycle distance and the timetabled time, average road gradient and average speed can be extracted for accurate energy consumption predictions to be made. The results show that the current methodology of using synthetic drive cycles for vehicle modelling is accurate for the prediction of real drive cycles, and that in the absence of real driving data, the creation of a synthetic drive cycle to represent the route is a suitable alternative.

It should be noted that there are a number of limitations associated with the prediction models developed in this study. For the prediction of industry standard drive cycles under Scenario 2, when more limited characteristics of the route were considered, whilst the prediction models developed are still quite accurate for the industry standard drive cycles (9.03 % average absolute percentage difference for the Extra Trees model), those developed were found to be less accurate than for the real drive cycles. This is due to the industry standard cycles having no elevation profile. Both the Extra Trees and Random Forests feature selection methods chose the average road gradient and the average speed to be the two CPs with the most influence on energy consumption. As all the synthetic drive cycles that make up the drive cycle database have an elevation profile, and with most having different route start and end stops, the majority have an average road gradient that doesn’t equal zero. Therefore, only one of the two CPs considered for the models developed for Scenario 2 are representative of the drive cycle database, and this leads to a loss in accuracy of the prediction models.

This study has been successful in developing an energy consumption prediction model for a BEB. A novel methodology has been applied in using multiple feature selection methods to select the CP subset that is optimal for the energy prediction model development. This method provides increased confidence that the optimal CP subset has been selected for energy consumption prediction, and this is shown in the low percentage difference between the simulated and predicted results for both real and industry standard drive cycles.

As well as extending the drive cycle database, future work should utilise these prediction models and estimation of vehicle operational capabilities to provide evidence and inform the deployment of charging infrastructure, bus fleet replacement, and specification of vehicle technology depending on route characteristics. Ideally further work should be undertaken to validate the outputs of the model and incorporate data fluctuations of passenger numbers, and ambient weather conditions. CPs related to driving style, such as average accelerator/brake pedal positions, should also be analysed for their influence on energy consumption. The same methodology could be used to develop a prediction model for the energy consumption of FCEBs.

CRediT authorship contribution statement

L.A.W. Blades: Conceptualization, Methodology, Writing – original draft. T. Matthews: Writing – original draft, Methodology, Investigation, Conceptualization. T.E. McGrath: Writing – original draft, Methodology, Investigation, Conceptualization. J. Early: Writing – review & editing, Supervision, Funding acquisition, Conceptualization. G. Cunningham: Writing – review & editing, Supervision, Funding acquisition, Conceptualization. A. Harris: Writing – review & editing, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to acknowledge the financial support for this work provided by the UK Engineering and Physical Sciences...
Research Council for the research project ‘Prosperity Partnership: Roadmaps to Zero Net Emissions in Urban Public Transport’ (Grant Number EP/S036695/1), Northern Ireland Department for the Economy (PhD stipend), and the Wright Technology & Research Centre at Queen’s University Belfast (W-Tech), as well as the technical support of Bamford Bus Company Ltd. Trading as Wrightbus.

References


