Energy 286 (2024) 129550

Contents lists available at ScienceDirect

Energy

journal homepage: www.elsevier.com/locate/energy

The impact of electromobility in public transport: An estimation of energy consumption using disaggregated data in Santiago, Chile

Franco Basso [a,b], Felipe Feijoo [a,*], Raúl Pezoa [c], Mauricio Varas [d], Brian Vidal [a]

[a] School of Industrial Engineering, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile
[b] Instituto Sistemas Complejos de Ingeniería, Chile
[c] Escuela de Ingeniería Industrial, Universidad Diego Portales, Santiago, Chile
[d] Centro de Investigación en Sustentabilidad y Gestión Estratégica de Recursos, Facultad de Ingeniería, Universidad del Desarrollo, Santiago, Chile

ARTICLE INFO

Keywords:
Public transport
Energy consumption
Machine learning
GPS
Electromobility

ABSTRACT

Electromobility in public transport has become a promising way to reduce environmental pollution. Several contributions have sought to estimate the energy consumption of buses in public transport. However, most of these efforts use measurements collected from controlled or simulated experiments, or that do not characterize the entire bus network. Unlike these studies, this article estimates the energy consumption of all the electric buses that circulate in the city of Santiago, Chile, during the studied period using full disaggregated GPS data and empirical measurements on some sensorized electric buses. The methodology considers a feature selection phase and the development of energy consumption prediction models using physics based and machine learning approaches. The performances of both models are compared with each other, and then, the best one is used to measure the impact of electromobility in the city. This analysis allows decision-makers to target investment by determining the buses with higher energy consumption savings in the face of budget constraints.

1. Introduction

Formerly known as Transantiago, RED, the public transport system of Santiago, Chile, has around 7400 buses in 2023, of which approximately 21% is electric. This high penetration rate has positioned Santiago as the city outside of China with the most significant number of electric buses in public transport [1]. The significant number of electric buses operating in RED aligns with local and international sustainable policies. Particularly, this is in line with the objectives set out in the latest National Determined Contributions (NDC) of Chile[1] and with the 2030 Agenda for Sustainable Development of the United Nations.

In the effort to climate change averting, new energy sources and improvements in vehicle design and information technology are necessary to reduce transport-related carbon emissions [2,3]. In this endeavor, the role of electric buses in public transit is important to take steps to reduce climate change [4]. Particularly, electric buses can reduce energy consumption and carbon dioxide emissions [5], diminishing the spewing of pollutants that harm the environment and people's health [6]. Furthermore, this technology is significantly more efficient than internal combustion diesel engines [7]. However, the adoption of electric buses shows some important drawbacks. Regarding economic aspects, the Total Cost of Ownership has been pointed out as one of the main barriers to their implementation [8]. Moreover, the operation of electric buses must overcome several challenges, including range limitation [9], lengthy charging times [10], and a proper distribution of a large number of charging spots [11]. For a recent discussion on the implementation challenges, we refer the reader to Aldenius et al. [12]. Additionally, a large deployment of electric buses (and public transport in general) would also require an assessment of the impacts on the electricity grid. Current distribution systems are not always designed to accommodate important increases of electricity load, with different charging patterns. Hence public charging infrastructure must also be planned accordingly between electric and transportation systems [3, 13]. A review of the impact of electric public transport in the power sector can be found in [14].

Nowadays, several cities are transitioning to greener bus fleets [15], and in Santiago, Chile, the number of electric buses in operation on RED will continue to increase in the coming years [16]. This raises the question of how a large share of electric buses impacts some key performance indicators of a public transport system. Literature shows that

* Corresponding author.
E-mail addresses: francobasso@gmail.com (F. Basso), felipe.feijoo@pucv.cl (F. Feijoo), raul.pezoa@udp.cl (R. Pezoa), mavaras@udd.cl (M. Varas), brian.vidal.a@mail.pucv.cl (B. Vidal).

[1] https://mma.gob.cl/wp-content/uploads/2020/07/Espanol-21-julio.pdf.

https://doi.org/10.1016/j.energy.2023.129550
Received 20 March 2023; Received in revised form 28 October 2023; Accepted 31 October 2023
Available online 9 November 2023

the overall energy consumption would diminish, but how and to what extent? Moreover, how do energy consumption savings vary with an increasing e-fleet considering actual driving conditions? In this paper, we tackle this issue by devising a novel methodology that estimates electric buses' energy consumption using two sources of information: GPS for all the buses in the network and empirical measurements on some sensorized electric buses. On this, highly disaggregated data provides greater precision in computing energy consumption, supporting better decision-making and control. Besides, by considering diverse driving patterns, our approach allows a better understanding of short-term energy consumption.

Different methodologies have been developed in the literature to measure the impact of electromobility, comparing the performance of conventional, electric, and hybrid buses. In general terms, these efforts conclude that, under normal conditions, new technologies reduce energy consumption, in addition to reducing CO_2 emissions compared to conventional vehicles. To support the above, the data from these works have been collected mainly in controlled or simulated environments. However, these protected conditions may limit the true understanding of energy consumption in real driving situations during the whole day of operation. Recently, studies have been published using sensorized bus data similar to ours with high granularity. However, these efforts have focused on the quality of the models rather than their application to the entire network to obtain public policy recommendations. This work helps bridge this gap by proposing an approach that uses disaggregated data measured in real driving situations throughout the entire bus network of Santiago's public transport system.

The contributions of this work are two-fold. First, two kinds of models are proposed that, taking advantage of the wealth of disaggregated data available, allow us to estimate the energy consumption of the electric buses in the system under different environmental and driving conditions: one based on physics that considers force and energy balance equations, and another empirically based, in which supervised machine learning techniques are used. Second, a case study is carried out in the city of Santiago, Chile, comparing all the proposed models in terms of goodness of fit and determining relevant variables. Finally, the impact of the introduction of electric buses on the energy consumption of the system is quantified.

The rest of this article is structured as follows. In Section 2, the literature is reviewed. Section 3 describes the data and the case study analyzed in this work. Section 4 presents the methodology for estimating energy consumption. In Section 5, the methodology is applied to the case of Santiago, Chile, including the analysis of the impact of increasing electromobility in the city. Finally, in Section 6, the conclusions and lines of future work are presented. For the sake of simplicity, Appendix shows all the abbreviations used in this paper.

2. Literature review

2.1. Influential factors for energy consumption

Growing pollution levels in major cities is becoming a delicate issue in cities worldwide. In this regard, although electric vehicles (EVs) offer several environmental advantages, there has been scarce research on the impact of EVs on public transit networks [14]. Indeed, most of the literature has focused on estimating the consumption of electric buses by using a fleet sample (e.g., [17–22]) or in planning problems, such as finding optimal charging schedules that reduce costs while ensuring operational continuity (see [23], for a review on electric bus planning and scheduling problems). Consequently, the emphasis of previous literature lies more on the accuracy of prediction techniques and the performance of optimization approaches rather than the actual influence of electric buses on the transportation network.

The literature shows that the power consumed by a bus depends on several factors. In particular, Sinhuber et al. [24] highlights the characteristics of the bus, speed and acceleration, while Abdelaty et al. [17] and Bartłomiejczyk and Kołacz [25] consider the frequency of stops and weather conditions, respectively, as relevant factors. Along these same lines, Pettersson et al. [26] employ stochastic models within a hierarchical structure to generate physical properties of road transport operations. This approach enables dynamic simulations that estimate energy usage. The authors find that the intensity parameter of the stochastic model for stops has an impact of 8.3% on energy consumption. Likewise, the characteristics of the route also have an influence on consumption. In this regard, Al-Ogaili et al. [27] show that altitude variation along the route is a first-order factor for the energy consumption of a bus. For the rest, the mass of passengers also significantly influences energy consumption [28]. The literature shows that the impact of these factors depends on the technology that the bus incorporates. In particular, Ma et al. [29] conclude that congestion and the number of stops have a different effect on the consumption of electric and diesel buses. However, in the case of electric buses, this impact is subject to a trade-off. Indeed, according to Liu et al. [30], an electric bus at its maximum capacity has a higher energy consumption of approximately 23% compared to the bus without considering the mass of the passengers. On the other hand, in the same study it is also commented that, when the mass of the vehicle is greater, the amount of energy that is stored in the battery through regenerative braking increases.

2.2. Methodologies for energy consumption estimation

Various methodologies allow us to estimate energy consumption, mainly highlighting two approaches. The first considers the calibration of physics-based models based on the dynamics of the vehicle and its powertrain [29,31–35], which, due to the large number of variables that influence energy consumption, can present considerable estimation errors [20]. The second approach considers instead the use of statistical learning models. These methods identify highly complex relationships between different variables and the measurement of consumption, generally through Machine Learning (ML) models [36–41]. Although this type of model has obtained good prediction results [42], its conclusions are strongly dependent on the available data set [32]. In this way, its generalization to different contexts is more difficult to perform.

Some efforts that use physical basis models are described below. Zhang et al. [31] estimate the fuel consumption of 75 public transport buses in Beijing, China, from emission measurements. In particular, the authors analyze the impact of driving conditions on consumption, finding, for example, that the average speed and the use of air conditioning have a great influence on the fuel consumed. Gallet et al. [33], meanwhile, proposes a longitudinal dynamics model to estimate the energy consumption of electric buses. The authors then conduct a case study on the complete bus network in Singapore, finding the estimated consumption per line and type of bus. More recently, Liu et al. [30] use a simulation tool, based on vehicle speed and torque, to estimate the impact of passenger mass on the energy consumption of public transportation buses in Minneapolis and Saint Paul, United States. The authors find that the mass of passengers produces a greater increase in the energy consumption of conventional buses, compared to electric buses, due to the regeneration capacity of the latter.

Statistical learning models have become a recent line of research. In this regard, Sun et al. [43] use Artificial Neural Networks models to predict fuel consumption in hybrid buses from disaggregated data. In particular, using data from two sensorized buses in Minneapolis, United States, the authors compare the predictive capacity of their models and show that the greater the temporal aggregation, the better the prediction quality. Abdelaty et al. [17] calibrate ML models to predict energy consumption of electric buses in Hamilton, Canada. The authors use a full-factorial experiment to define operational scenarios that allow simulating energy consumption, which is later incorporated as a dependent variable in the ML models. Finally, Li et al. [44]

Table 1
Examples of different approaches of energy consumption estimation in the literature.

Reference	Location	Method	Calibration data
Wu et al. [34]	California	Physical base model	169 trips (4 routes) over 5 months.
Zhao et al. [22]	Beijing	ML - Frequency item mining	10 electric buses for one year.
Gallet et al. [33]	Singapore	Physical base model	Non-data-driven model
Jiang et al. [21]	Beijing	ML - Gaussian processing regression	8 electric buses scheduled on the same route for one year.
Zeng et al. [36]	Toyota	ML - Support vector machine regression and Artificial Neural Network	7989 trips over one month.
Pamuła and Pamuła [39]	Jaworzno	ML - Multiple linear regression and deep learning network	12 electric buses over 4000 journeys for 10 months.
Sennefelder et al. [40]	Seville	ML - Multiple linear regression	30 conventional diesel-powered buses for 11 consecutive days.
Qin et al. [41]	Meihekou	ML - Support vector machine regression	3 identical electric buses under real conditions.
Chen et al. [45]	Chattanooga	ML - Long short-term memory and Artificial neural network	3 identical electric buses under real conditions.
Li et al. [44]	Shenzhen	ML - Stochastic and k nearest neighbor Random Forest	20 electric buses over five months.

propose a two-step methodology: Stochastic Random Forest and k-Nearest Neighbor. The authors use data from 20 buses over five months in Shenzhen, China, from which they show that the models have high predictive performance.

Table 1 summarizes examples of previous contributions that predict energy consumption using both physical base and statistical learning models.

In summary, taking the reviewed literature into account in both subsections, this article contributes to a recent line of research, proposing models that estimate the energy consumption of electric buses through ML models. Like some contributions from recent years, we use disaggregated data for a subsample of sensorized electric buses. Yet, contrary to most studies, we utilize these models to estimate energy consumption across an entire city's bus network, an unprecedented endeavor in the relevant literature. We aim to fill this gap by employing advanced quantitative methodologies for a comprehensive, city-wide evaluation of electromobility's impact. Our proposed methodology and its application could provide valuable insights to local decision-makers as, to the best of our knowledge, our paper is the first effort that estimates the energy consumption of buses in Chile for a full public transport network. These insights encompass tactical decisions, such as determining which bus lines to electrify first, and specific operational advice for drivers to promote more sustainable driving behaviors.

3. Data description and case study

RED works in an integrated way, that is, passengers can access any mode of transport with the same means of payment. In addition, the integrated modes are buses, Metro, and suburban trains, while payment is made through a smartcard called "Bip!" card [46]. Given the integration, it is possible to make combinations without additional cost between the modes of travel. The system covers the area known as Greater Santiago, made up of the 32 communes of the province of Santiago, in addition to the communes of San Bernardo and Puente Alto [47].

The study period examined in this research corresponds to September 22, 2021. Note that during this date, the COVID-19 lockdown started at midnight, implying that the buses operated normally during the daytime. Furthermore, during this period, most of the COVID-19 measures were already lifted, and the distribution of trip purposes quickly resumed patterns close to those observed before the pandemic [48].

For this period four databases are considered, namely: GPS of all the buses that circulated on the network that day, the physical characteristics of buses, network transactions, and empirical measurements of a sensorized sample of electric buses. The source of the first three bases is the *Directorio de Transporte Públíco Metropolitano* (DTPM),[2] which is a public body under the Ministry of Transport and Telecommunications that is in charge of RED operations. The source of the last base is the company TrackTec,[3] which develops and implements technological solutions through the monitoring of telemetry variables.

Table 2
Instance of GPS data provided by DTPM for the day September 22, 2021.

Timestamp	2021-09-22T12:09:47Z
License	BJFC-26
Latitude	−33.42045
Longitude	−70.61706
Route	T502 00R
Speed [km/h]	12.04

Table 3
Data sample with physical characteristics of the buses.

License	FLXV43	PLHK81
Emissions standard	ELECTRIC	EURO VI
Area [m^2]	7.67	7.50
Mass [kg]	12,420.0	19,203.71

Table 4
Quantity of license plates pér technology.

Emissions standard	Quantity	Percentage
EURO III WITH FILTER	2785	40%
EURO VI	1458	21%
EURO V	1320	19%
ELECTRIC	784	11%
EURO III	556	8%

3.1. GPS data

Each bus in the system emits a GPS signal every 30 s. Based on this, DTPM provides us with information on the fields shown in Table 2. This database contains 12,387,516 records, for a total of 6650 unique bus license plates, from 00:00:00 to 23:59:59 of the study period. Fig. 1 shows the distribution of the GPS records of the buses during the study period. It can be observed that, as expected, there is a large number of buses transiting in the central communes in comparison to the peripheral communes.

3.2. Physical characteristics of the buses

The DTPM provides us with specific information on the buses that circulated in September 2021. Specifically, as shown in Table 3, we obtain the emission standard of the engine, the mass of the bus and the cross-sectional area for each patent. Although there are values with high frequencies, the buses that circulate in Santiago are of different types. Finally, Table 4 shows the number of license plates and the

[2] https://www.dtpm.cl/.

[3] https://tracktec.cl/.

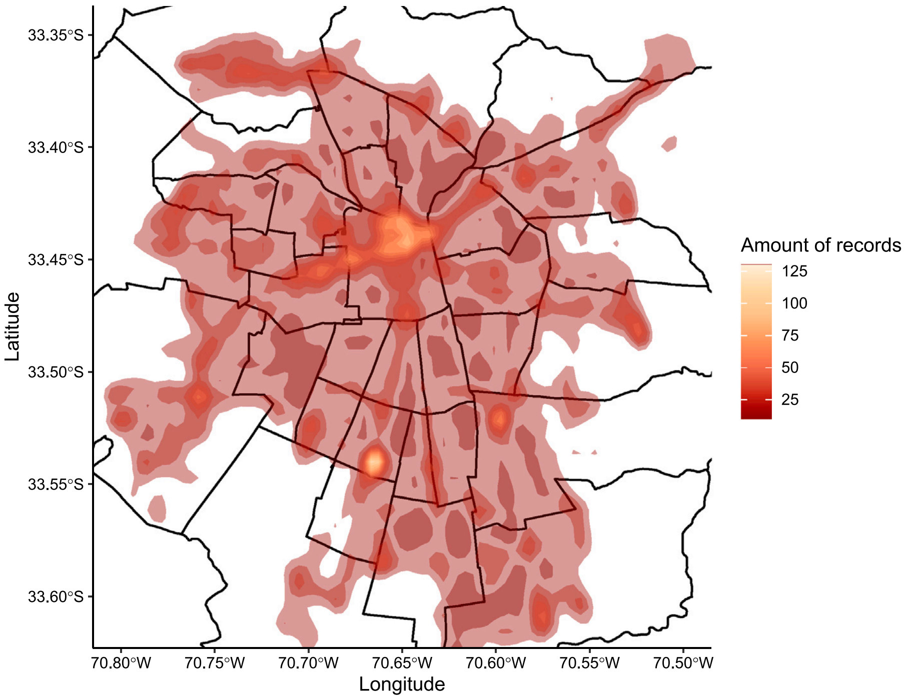

Fig. 1. Heatmap of the distribution of the buses during the study period.

Table 5
Fields registered in the transaction database.

Timestamp	22-09-2021 0:00	22-09-2021 5:42
Type	BUS	METRO - OT
Location	GCBC-29	Los Libertadores
Card ID	bd8e667d	38c12ec3
Operator	U2 - Su Bus	–
Route	T230	–
Direction	I	–

percentage of buses according to emission standards. It can be deduced from this that 89% of the patents correspond to diesel buses, while the remaining 11% correspond to electric buses.

3.3. Transactions

As previously mentioned, payment in the RED system is made through the "Bip!" Card. Through this card, only boarding information is recorded. This information is collected by an external company, and then sent to DTPM. The information related to where passengers alight is estimated by DTPM, using "Analysis of public transport data" (ADATRAP, in Spanish),[4] a software that implements the methodology proposed in [49]. Then, each entry in the database provided by DTPM corresponds to a transaction, which contains information according to the fields shown in Table 5. This database contains 2,989,099 transactions for the study period.

3.4. Sensorized bus data

The last database is provided by TrackTec. This is a private company that provides solutions through different types of sensors and monitoring systems. TrackTec works directly with some RED operators, measuring different operational variables of the buses. The foregoing seeks to help the decision-making of the operators to improve the indicators of quality of service to the user [50]. The database delivered by TrackTec contains 10,601 records, which are obtained from six electric buses in the study period. Table 6 shows an example of a record provided by the company

Table 6
Fields registered in the database provided by TrackTec.

License	FLXZ-27
Date	2021-09-22T08:23:42Z
Latitude	−33.508
Longitude	−70.779
Speed [km/h]	41
Odometer [km]	149 382.1
Accelerator %	63
Brake %	0
RPM [rpm]	1383
Vent. Level A/C	0
Set Temp. A/C [°C]	24
Temp. In [°C]	13
Temp. Ex [°C]	11
Av. State Pads %	60.8
Total generated power [kW]	−16 087.2
Total consumed power [kW]	63 461.9

The GPS devices installed by TrackTec on the buses under consideration emit a signal every minute. However, they also emit a signal when particular phenomena of interest to the company occur, such as acceleration above an assigned threshold. On the one hand, the database contains information identical to that obtained with GPS technologies such as the one described in Section 3.1. These variables are *License Plate*, *Date*, *Longitude*, *Latitude*, *Speed*, and *Odometer*. The TrackTec database also contains new information, less explored in the literature, that allows a better characterization of the bus operation. In particular, the *Accelerator* and *Brake* variables correspond to the percentage that the accelerator and brake pedal are pressed, respectively, while the *RPM* variable indicates the revolutions per minute of the engine. The variables *Vent. Level A/C* and *Set Temp. A/C* is attached to air conditioning. Specifically, the first indicates the intensity of the air conditioning, while the second corresponds to the selected temperature of the same. The variables *Temp. In* and *Temp. Ex* correspond to the internal and

[4] https://www.dtpm.cl/index.php/documentos/matrices-de-viaje.

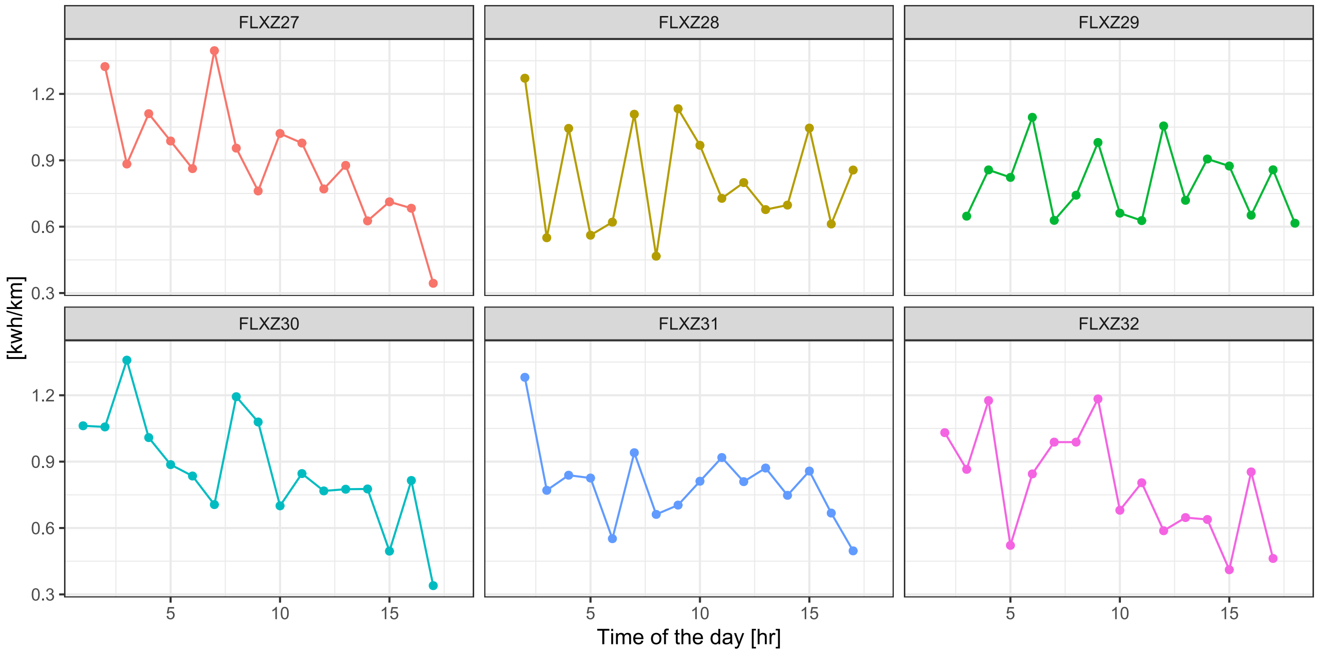

Fig. 2. Empirical energy efficiency of buses by hour.

external temperature measurements of the bus, respectively. Finally, the *Total Generated Power* and *Total Consumed Power* are calculated at each instant through the sum of the product of the voltage and the current that flows to and from the battery in a time interval.

With the data described above, it is possible to calibrate models that use sensor data as independent variables and empirical energy efficiency as a dependent variable. The latter is calculated as the quotient between energy consumption (calculated from the *Total Generated Power* and *Total Consumed Power*) and the distance traveled (calculated from the *Odometer*). Fig. 2 shows the variation of energy efficiency for each of the six buses in the study period.

4. Methodology

4.1. Physical bases model

In order to estimate energy consumption, the physical base models use a balance of the forces that the bus faces. This class of models, unlike models based on supervised learning methods, have the advantage of not requiring an empirical measurement of energy consumption for its evaluation. In this article, we use the model of Chen et al. [32]. According to this model, the energy consumption (CE_t) in an interval of T seconds is calculated according to Eq. (1), where P_t^{bat} is the power consumed by the battery to generate the movement.

$$CE_t = \int_{t-T}^{t} P_v^{bat}\, dv \tag{1}$$

Then, to calculate P_t^{bat}, Eq. (2) is used in which two cases are considered, depending on the sign of the sum of the rolling resistance force (F_t^{rr}), the aerodynamic force (F_t^{ad}), the road gradient force (F_t^{rg}) and the acceleration force (F_t^{acc}). If this sum is positive, the power consumed by the battery is equal to P_t, calculated according to Eq. (3), where the variables and parameters necessary for its computation are presented in Tables 7 and 8, respectively. If the sum of the forces is negative, the current is directed towards the battery, causing it to charge [51]. In this case, a parameter k_t associated with the percentage of power that is actually charged to the battery is considered. According to Zhang and Yao [52], the parameter k_t can be calculated as a piecewise function from the instantaneous speed (v_t) of the bus using Eq. (4).

Table 7
Physical base model variables.

Variable	Variable name
v_t	Instant speed of the bus
m_t	Instant mass of bus and passengers
α_t	Slope of route in degrees

$$P_t^{bat} = \begin{cases} P_t & \text{if } (F_t^{rr} + F_t^{ad} + F_t^{rg} + F_t^{acc}) \geq 0 \\ -k_t \cdot P_t & \text{if } (F_t^{rr} + F_t^{ad} + F_t^{rg} + F_t^{acc}) < 0 \end{cases} \tag{2}$$

$$\begin{aligned} P_t &= \frac{v_t(F_t^{rr} + F_t^{ad} + F_t^{rg} + F_t^{acc})}{\eta^{mo}\eta^{tr}} \\ &= \left(\frac{v_t}{\eta^{mo}\eta^{tr}}\right)\left(C^r m_t g \cos(\alpha_t) + \frac{\rho^a}{2} C^d A^f v_t^2 + m_t g \sin(\alpha_t) + m_t \delta \frac{\partial v_t}{\partial t}\right) \end{aligned} \tag{3}$$

$$k_t = \begin{cases} 0.5 \cdot \dfrac{v_t}{5} & \text{if } v_t < 5 \text{ [m/s]} \\ 0.5 + 0.3 \cdot \dfrac{v_t - 5}{20} & \text{if } v_t \geq 5 \text{ [m/s]} \end{cases} \tag{4}$$

The speed and slope of the route are calculated for each GPS record from the geographic coordinates. Finally, it should be considered that there are other parameters that are more difficult to estimate, so the values used are taken from the literature, which is shown in Table 8.

4.2. Statistical learning models

Taking into consideration the empirical measurements of energy consumption provided by TrackTec, ML and linear regression models are developed using the data from six electric buses monitored by the company.

4.2.1. Data processing

Part of the objectives of this work is to quantify the importance of different operational variables in the energy consumption of the network buses. For this reason, in addition to the data provided by

Table 8
Parameters of the physical basis mode.

Parameter	Name	Value	Source
η^{mo} (Electric)	Engine efficiency electric buses	0.85	Asamer et al. [53]
η^{tr} (Electric)	Transmission efficiency electric buses	0.97	Asamer et al. [53]
$\eta^{mo} \cdot \eta^{tr}$ (Diesel)	Total efficiency diesel buses	0.3	Wdaah and Müller [7]
C^{r}	Coefficient associated with rolling resistance	0.01	Gao et al. [54]
C^{d}	Aerodynamic drag coefficient	0.7	Lajunen et al. [55] and Gao et al. [54]
g	Acceleration of gravity	9.81	–
ρ^{a}	Air density	1.16	–
A^{f}	Cross-sectional area of the bus	–	DTPM
δ	Bus rotational inertia factor	1	Hjelkrem et al. [56]

Table 9
Variables available for each interval after aggregating the data.

Variable	Description	Unit
PMConsumption	Total energy consumption calculated from the physical base model	[kWh]
TotalDistance	Total distance traveled by the bus calculated from GPS data	[km]
AvgAng	Average angle of the incline of the route that the bus follows	[degrees]
SdAng	Standard deviation of the incline angle of the route followed by the energy bus	[degrees]
AvgK	Energy recovered from regenerative braking	%
SdK	Standard deviation of energy recovered from regenerative braking	%
AvgMassTotal	Average total mass of the bus (body and passengers)	[kg]
SdMassTotal	Standard deviation of the total mass of the bus (body and passengers)	[kg]
AvgVel.inst	Average instantaneous speed	[km/h]
SdVel.inst	Standard deviation instantaneous speed	[km/h]
AvgAce.inst	Average instantaneous acceleration	[m/s^2]
SdAce.inst	Standard deviation instantaneous acceleration	[m/s^2]
AvgPed.Ace	Average accelerator pedal utilization	%
SdPed.Ace	Standard deviation of the average use of the accelerator pedal	%
AvgRPM	Average revolutions per minute	[rpm]
SdRPM	Standard deviation revolutions per minute	[rpm]
AmountPed.Ace	Number of records in which the accelerator pedal is being used	–
AvgPed.Fre	Average brake pedal utilization	%
SdPed.Fre	Standard deviation of the average use of the brake pedal	%
AmountPed.Fre	Number of records in which the brake pedal is being used	–
AmountAc.on	Number of records in which the air conditioning is active	–
AvgT.Ac	Average air conditioning temperature	[°C]
SdT.Ac	Standard deviation temperature indicated in the air conditioner	[°C]
AvgT.in	Average temperature inside the bus	[°C]
SdT.in	Standard deviation mean temperature inside the bus	[°C]
AvgT.ex	Average temperature outside the bus	[°C]
SdT.ex	Standard deviation average temperature outside the bus	[°C]
AvgPast	Average condition of the pads	%
SdPast	Standard deviation mean status of pads	%
Hour	(Categorical) Time of the records of the specified interval	–

TrackTec, other variables are added that can be obtained from the data provided by the DTPM, namely: distance traveled, instantaneous speed, slope angle and instantaneous total mass.

We consider energy efficiency [kWh/km] as the response variable in time intervals of 5, 10, 15 and 30 min. The independent variables are shown in Table 9. On the other hand, the estimate of energy consumption made through the physical base model is incorporated as a predictor variable in Section 4.1. This is done in line with what was put forward in [45], where Vehicle Specific Power is included as a predictor variable given its interpretation linked to the power consumed.

4.2.2. Selection of variables

As described in the previous subsection, the training data considers a large number of variables. In the context of supervised learning models, if some of the variables are not relevant for the estimation of energy consumption, their use can result in models that present excessive variance, and therefore overfit the data of training [57]. Given this, in this study we perform a variable selection procedure. In this regard, the literature provides several methods to select variables (see, e.g., [58], for a comprehensive review and comparison of feature selection methods in the binary classification context). In this paper, we employ the Boruta algorithm, a tree-based method. This feature selection method has shown remarkable accuracy in previous literature, presenting low out-of-bag error rates and low computation times (e.g., [59–62]). Moreover, previous studies have found that the Boruta algorithm is one of the best-performing methods in low-dimensional contexts, such as ours, where the number of observations far exceeds the number of variables [63]. Boruta works as follows:

1. A new set of variables is generated from the random permutation of each of the original variables. These new variables are called *shadow features*.
2. A Random Forest model is fitted, using the original variables and the shadow features as predictors. With this model, the importance of each variable considered is found.
3. Iterate, checking if the original variables have a greater importance than the shadow features with the maximum importance. Then, those variables that are consistently less important than any of the shadow features are removed.

In the prediction models that are explained below, we only consider the variables that the Boruta model indicates as important.

4.2.3. Prediction models

Random forest

Random Forest models consider the calibration of multiple decision trees in parallel. The trees are generated using different training sets assembled by *bootstrapping* type sampling [64] so that in each iteration a number of variables are randomly selected that are used to fit a deep tree [65]. In the case of regression, for each new data x the estimate

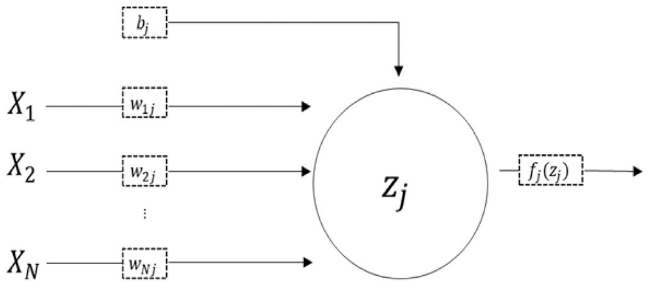

Fig. 3. Functioning of a neuron in a neural network.

is calculated as a simple average of the outputs of the trees as shown in Eq. (5), where T_b corresponds to tree number b and B is the total number of trees created:

$$\hat{f}^B(x) = \frac{1}{B} \cdot \sum_{b=1}^{B} T_b(x) \tag{5}$$

Support Vector Regression

A Support Vector Regressor (SVR) model seeks to construct a function $f(x) = w \cdot x + b$ that returns values close to the dependent variable y (within a margin of size ϵ), and that at the same time, is as regular as possible. The deviation that the function has with respect to the response variable is ϵ. In general, the optimization problem to be solved is shown in Eqs. (6a)–(6d) [66], considering $x, y, b, w \in \mathbb{R}^M$, where M is the number of predictor variables:

$$\text{min} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{6a}$$

$$\text{subject to} \quad (wx_i + b_i) - y_i \leq \epsilon + \xi_i \tag{6b}$$

$$y_i - (wx_i + b_i) \leq \epsilon + \xi_i^* \tag{6c}$$

$$\xi_i, \xi_i^* \geq 0 \tag{6d}$$

where $C > 0$ controls the penalty imposed on observations that fall outside the range, thus avoiding overfitting.

Additionally, SVR considers the use of Kernel functions. These functions take the data to a different, usually higher, dimensional space, allowing non-linear decision functions to be generated. The kernel functions that are used are the following:

1. Linear Kernel: $K(x, y) = (x^T y + C)$
2. Radial Kernel: $K(x, y) = \exp(-\gamma \|x - y\|^2)$
3. Polynomial Kernel: $K(x, y) = (\gamma x \cdot y + C)^q$

where γ is a hyperparameter to adjust.

Neural Networks

A Neural Network (NN) corresponds to a structure that seeks to replicate the behavior of the human brain by interconnecting a large number of neurons with each other from inputs and outputs. The general structure of a neuron is shown in Fig. 3.

Each neuron of the NN has an activation function f_j that transforms inputs into outputs. In each neuron, a weighted sum z_j of the input values is performed according to the weights assigned to each variable w_{ij} and the bias b_j. The weighted sum is calculated according to Eq. (7). The weights can be estimated through different methodologies, the most classic being *backpropagation* [67].

$$z_j = \sum_{i=1}^{N}[x_i \cdot w_{ij} + b_j] \tag{7}$$

Table 10
Variables included in each dataset.

Variable	Full dataset	Limited dataset
PMConsumption	✓	✓
TotalDistance	✓	✓
AvgAng	✓	✓
SdAng	✓	✓
AvgK	✓	✓
SdK	✓	✓
AvgMassTotal	✓	✓
SdMassTotal	✓	✓
AvgVel.inst	✓	✓
SdVel.inst	✓	✓
AvgAce.inst	✓	✓
SdAce.inst	✓	✓
Hour	✓	✓
AvgPed.Ace	✓	–
SdPed.Ace	✓	–
AvgRPM	✓	–
SdRPM	✓	–
AmountPed.Ace	✓	–
AvgPed.Fre	✓	–
SdPed.Fre	✓	–
AmountPed.Fre	✓	–
AmountAc.on	✓	–
AvgT.Ac	✓	–
SdT.Ac	✓	–
AvgT.in	✓	–
SdT.in	✓	–
AvgT.ex	✓	–
SdT.ex	✓	–
AvgPast	✓	–
SdPast	✓	–

4.2.4. Training and validation

A summary of the methodology developed in this study is presented in Fig. 4. From the aggregated database, the construction of which is described in Section 4.2.1, two data subsets are generated: (i) Training, which contains data from four buses (66.7%) and (ii) Testing, which contains data from the two remaining electric buses (33.3%). These subsets vary from a cross-validation that considers all the different combinations of bus layouts. Then, from the training set, the most relevant variables are selected using the Boruta algorithm. Subsequently, four models are trained: Linear Regression (LR), Random Forest (RF), Support Vector Regressor (SVR) and Neural Network (NN). Next, predictions are made from the test set and the results are evaluated. The metric used to compare the performance of the models is the Mean Absolute Percentage Error (MAPE), which is widely used in the works cited in the literature

It is important to remember that the variables shown in Table 9 require data from both DTPM and TrackTek. However, as mentioned, the data available for the entire network corresponds only to that of the DTPM. For the same reason, and depending on the circumstance, we will use two sets of variables. The first set called *full dataset* considers all the variables, while the second called *limited dataset* considers only those variables that can be calculated with data from the DTPM. We utilize the limited dataset when applying the most effective model across the entire network, simulating an all-electric bus scenario. These estimates are then juxtaposed with those derived from the physical base model (PM) to gauge potential energy consumption savings should the bus technology transition from diesel to electric. Table 10 shows the variables belonging to each set.

5. Computational results

In this section we apply the models described in Sections 4.1 and 4.2 to the data presented in Section 3. In particular, Section 5.1 presents the most relevant variables for estimating energy performance. In Section 5.2 the developed models are compared. In Section 5.3 we show the relevance of using our disaggregated approach. Finally, in

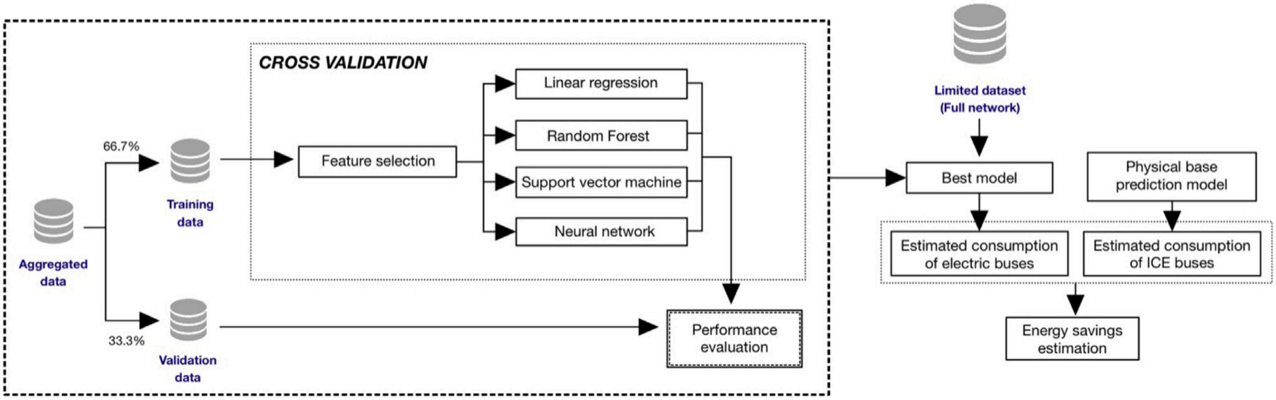

Fig. 4. A general overview of the proposed approach.

Table 11
Selection of variables at different levels of temporal aggregation.

Variables	5 min	10 min	15 min	30 min	Quantity
AvgT.Ac	✓	✓	✓	✓	4
SdPed.Ace	✓	✓	✓	✓	4
AvgVel.inst	✓	✓	✓	✓	4
PMConsumption	✓	✓	✓	✓	4
AvgPed.Fre	✓	✓	✓	✓	4
AvgAng	✓	✓	✓	✓	4
AvgT.in	✓	✓	✓	✓	4
AvgK	✓	✓	✓	✓	4
SdK	✓	✓	✓	✓	4
AvgMassTotal	✓	✓	✓	✓	4
AvgPed.Ace	✓	✓	✓	✓	4
AvgT.ex	✓	✓	✓	✓	4
TotalDistance	✓	✓	✓	–	3
SdVel.inst	✓	✓	✓	–	3
AvgRPM	✓	✓	✓	–	3
SdPed.Fre	✓	–	✓	–	2
AmountPed.Fre	✓	–	✓	–	2
SdMassTotal	–	✓	✓	–	2
SdAng	✓	–	–	–	1
SdPast	✓	–	–	–	1
SdRPM	✓	–	–	–	1
AvgAce.inst	✓	–	–	–	1
SdAce.inst	✓	–	–	–	1
SdT.ex	–	–	–	–	0
SdT.Ac	–	–	–	–	0
AmountAc.on	–	–	–	–	0
SdT.in	–	–	–	–	0
AmountPed.Ace	–	–	–	–	0
AvgPast	–	–	–	–	0

Section 5.4 we study the energy benefits of replacing diesel buses with electric buses, considering the entire network of Santiago, Chile.

5.1. Influential factors

In this subsection we analyze the most relevant factors to estimate energy efficiency. For this, we use the Boruta algorithm, described in Section 4.2.2, considering the explanatory variables described in Section 4.2.1. Table 11 shows, for each level of data aggregation, the variables accepted as predictive by the Boruta algorithm.

From the results in Table 11, it follows that a large part of the predictive variables are associated with operational aspects and driving style, in line with what is reported in Section 2. Some examples are the speed, RPM, path angle, pedal usage, average speed, average grade, and total bus mass. With this, it is evident that particular driving situations are fundamental factors to explain the energy efficiency of buses. On the other hand, the impact of the average temperature associated with air conditioning and the internal temperature of the bus stands out. In addition, the use of the accelerator pedal and its standard deviation are significant in all cases. This is in line with the previous studies presented in Section 2 regarding the importance of the frequency of stops and accelerations in energy consumption.

Lastly, it can be seen that the variable built from the physical base model (PMConsumption) also has a high importance, since the average is relevant for all aggregation levels. The latter is in line with what was obtained in previous studies (e.g., [45]).

Finally, it can be seen that, in general, the number of variables accepted as predictors decreases as the level of aggregation increases. This may be due to the fact that at high levels of aggregation, the difference in the measurements of the variables is smaller, resulting in homogeneous values between the records of the database. This further justifies the use of disaggregated data to explain the energy consumed by the buses.

5.2. Model performance comparison

Table 12 presents the resulting average Root Mean Square Error (RMSE) for each level of temporal aggregation of records, datasets, and for all the models developed. In addition, the standard deviation of the same indicator is presented in parentheses.

As expected, the results improve considerably when the record aggregation time is longer. This is achieved at the cost of less data disaggregation, so the model with the lowest RMSE does not necessarily correspond to the best of all. The best model should be considered that with a previously selected aggregation level.

In the Table 12 it can be seen that the LR and the linear kernel SVR always give the best results. The differences between both models are small and, in general, linear regression stands out when the data aggregation time is longer. When predicted in units of time of up to 15 min, the linear SVR explains the energy efficiency of electric buses equally or better.

On the other hand, it is possible to observe that the physically based model delivers worse results than the statistical learning models in all cases. This could be because the first one is based on physical equations that are not capable of explaining the particular situations that a bus faces in a day. However, this type of model allows estimating energy consumption for electric and conventional buses. This, in our case, is not possible to do from statistical learning models as we do not have conventional bus data with which to calibrate them.

It is important to emphasize the difference that exists in the results when using different datasets. As previously mentioned, the *full dataset* uses variables measured by TrackTec, so they are not available for the entire network. Given this, the model to be used must incorporate the variables considered in the *limited dataset*, which are available for all buses. The benefit of using the *full dataset* over the *limited dataset*, represented by the RMSE, is 0.01, 0.009, 0.008, and 0.019 for the aggregation levels of 5, 10, 15, and 30 min, respectively.

Table 12
RMSE of each model for different times.

Dataset	Model	5 [min]	10 [min]	15 [min]	30 [min]	Average
–	Physical basis models	0.402 (0.023)	0.369 (0.023)	0.334 (0.024)	0.332 (0.018)	0.359
Full dataset	Linear Regression	0.282 (0.011)	0.248 (0.022)	0.208 (0.022)	0.182 (0.025)	0.230
	Random Forest	0.287 (0.009)	0.254 (0.016)	0.219 (0.020)	0.193 (0.023)	0.241
	SVR (linear)	0.283 (0.013)	0.246 (0.021)	0.210 (0.021)	0.184 (0.026)	0.238
	SVR (radial)	0.369 (0.008)	0.322 (0.014)	0.290 (0.019)	0.251 (0.022)	0.244
	SVR (polynomial)	0.379 (0.010)	0.341 (0.013)	0.308 (0.017)	0.285 (0.021)	0.231
	Neural Network	0.290 (0.017)	0.255 (0.019)	0.218 (0.019)	0.195 (0.025)	0.308
Limited dataset	Linear Regression	0.292 (0.007)	0.255 (0.023)	0.216 (0.016)	0.201 (0.022)	0.328
	Random Forest	0.293 (0.005)	0.257 (0.018)	0.222 (0.016)	0.205 (0.024)	0.242
	SVR (linear)	0.292 (0.011)	0.255 (0.023)	0.216 (0.017)	0.204 (0.021)	0.289
	SVR (radial)	0.330 (0.008)	0.304 (0.014)	0.265 (0.014)	0.255 (0.020)	0.328
	SVR (polynomial)	0.379 (0.010)	0.341 (0.013)	0.308 (0.017)	0.285 (0.021)	0.240
	Neural Network	0.297 (0.011)	0.260 (0.021)	0.226 (0.018)	0.211 (0.022)	0.248

Table 13
Sensitivity analysis results of parameters of the physical base model.

Parameter	Avg difference	Difference in % [kWh/bus-h]
δ	[−0.07, 0.08]	[−2.13, 2.20]
$\eta^{mo} \cdot \eta^{tr}$	[−0.15, 0.22]	[−4.28, 6.42]
η^{mo}	[−0.02, 0.03]	[−0.23, 0.34]
η^{tr}	[−0.02, 0.03]	[−0.23, 0.34]
C^{d}	[−0.01, 0.01]	[−0.41, 0.41]
C^{r}	[−0.10, 0.03]	[−2.19, 2.26]

Finally, to test the robustness of our results, we conduct a sensitivity analysis in order to investigate how energy consumption changes when perturbing the PM input parameters. This analysis involves modifying the parameters by 20% and computing the difference with respect to the original values. As shown in Table 13, the average percentage difference remains below 7% in all cases, indicating a reasonable PM robustness.

5.3. The value of our approach

In this subsection we compare the actual value measured by TrackTec with three different estimates for energy performance. These three estimations are: (i) The estimation made by the physical base model, (ii) the estimation made with the best statistical learning model developed in this work (linear SVR kernel) and (iii) a theoretical value for energy efficiency, reported by the manufacturer of the electric buses considered in this work. The latter, being an average value, does not consider either the environmental situations that the bus faces or the improvements that Tracktec has implemented in its buses with the aim of reducing energy consumption.

The comparison is made based on two electric buses, since the other four available are used to calibrate the linear kernel SVR model. Fig. 5 shows the estimates of the energy efficiency estimated every 15 min for the entire study period, where PM indicates the physical model, while SVR represents the support vector regression.

Calculating the average percentage difference in each of the instants estimated and represented in Fig. 5, it is obtained that the average percentage error of the theoretical energy efficiency is 104%, for the physically based model it is 43% and that of the linear kernel SVR estimate is 22%. The distribution of the percentage error of the SVR predictions are shown in Fig. 6. The performance of our best model is in line with previous literature. For instance, Felipe et al. [38] reported mean errors of up to 19% when predicting consumption every 5 min using 43 variables. Likewise, Li et al. [18] found mean errors ranging from 14% to 20%. It is important to note that some studies in the literature considered higher levels of aggregation, resulting in lower mean errors due to error cancellation. For instance, Pamuła and Pamuła [39] observed an average error of approximately 7% for every trip they analyzed. Similarly, Qin et al. [41] reported average errors at the trip level around 14%.

These results confirm the benefit of using the models built with respect to the mentioned daily average value. The previous difference could be explained because the models capture particular microscopic aspects of the buses and their driving that cannot be observed in a theoretical value, since it does not consider the variables that are faced. in reality at each instant of operation. Unlike the theoretical value, the physical base model and the linear SVR manage to explain part of the behavior cycles of the empirical energy efficiency. In addition, the benefit of the statistical learning model over the physically based ones is observed. The latter is explained because the statistical learning model considers additional variables to those considered by the physical model.

5.4. The impact electromobility: the benefit of renewing conventional fleet by electric buses

In this subsection, we study the benefits of changing conventional buses to electric ones. In order to reduce energy consumption levels as much as possible, it is necessary to determine in the first instance the diesel buses with the greatest potential for energy savings during the day, which will be candidates to be replaced by new electric buses. For this, the physical base model is used to the GPS data of diesel buses to estimate the consumption of each of these buses. Then, for the purposes of the experiments that are presented below, we assume that a subset of the diesel buses are replaced by electric buses, and that therefore the way of driving does not vary when performing this change. Subsequently, using the best statistical learning model found in Section 5.2, that is SVR with linear kernel, the energy consumption of these buses is estimated assuming that they are electric. Finally, we subtract this value to the energy consumption obtained from the physical model so we can calculate a saving for each replaced bus considering different technologies.

Table 14 shows the different savings produced by carrying out this procedure, replacing 1, 10, 50, 100, 500, and 1000 buses and considering first the ones with higher energy consumption potential. The information is presented considering different hours of the day. As can be seen, consumption effectively decreases when electric buses are considered. In addition, the more conventional buses are considered for replacement, the greater the network savings. Also, the savings differ depending on the day's hour. This occurs because our methodology estimates vehicle-by-vehicle energy consumption during the study period. Moreover, as the analysis considers the buses with higher energy consumption potential first, the average hourly saving per bus diminishes when including more buses.

By estimating energy consumption with real and disaggregated data, it is possible to obtain results for each of the buses and at different times of the day. This is of great value with respect to the use of daily consumption averages that are not calculated based on real and particular driving situations. This type of disaggregated analysis benefits decision makers related to public transport, since it provides them with more

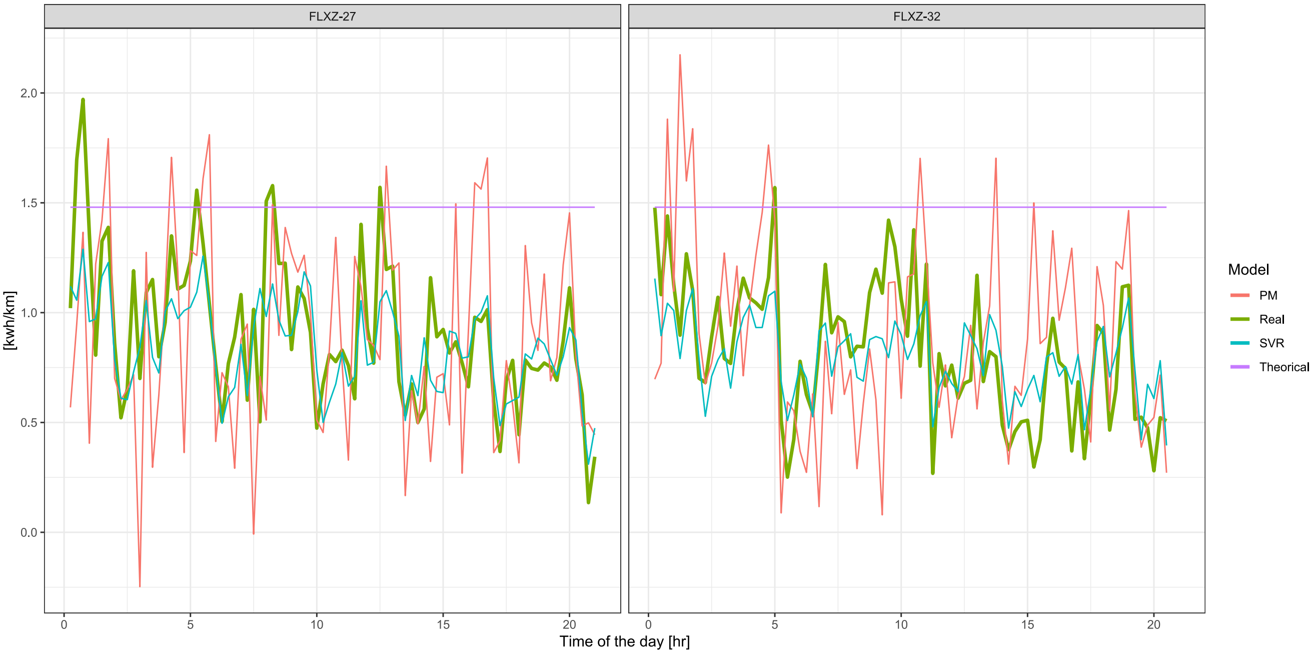

Fig. 5. Comparison of real energy performance with the best models.

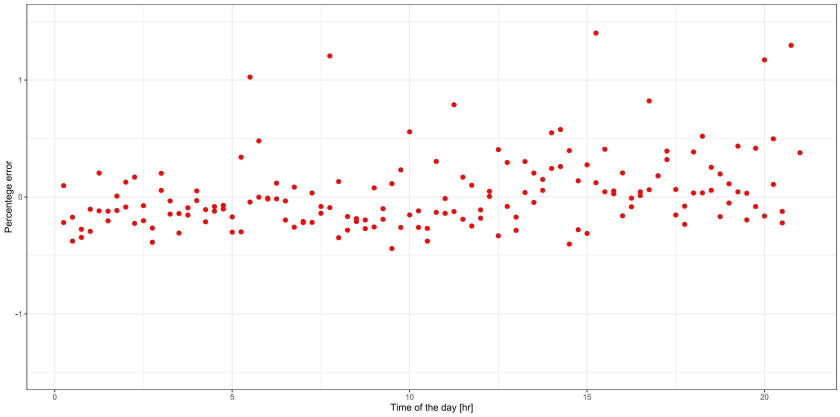

Fig. 6. Distribution of percentage error of SVR predictions.

Table 14
Savings in [kWh] by number of buses replaced, per hour.

Time of day	Quantity of buses considered					
	1 bus	10 buses	50 buses	100 buses	500 buses	1000 buses
6	92.65	375.53	2162.12	4028.90	14 526.06	23 549.22
7	25.59	431.28	2032.63	3933.08	15 350.36	25 607.80
8	24.73	473.68	1707.95	3163.72	14 757.24	25 154.34
9	10.10	336.48	1713.47	3206.16	14 805.45	25 390.76
10	91.65	255.99	1533.82	3133.29	14 390.46	24 518.90
11	34.85	193.01	1525.71	2910.95	13 046.24	22 175.73
12	33.35	538.59	1614.63	3100.23	12 138.79	20 885.16
13	11.82	370.41	1559.38	2780.99	12 111.00	20 938.50
14	97.38	377.97	1422.82	2762.02	12 865.64	23 097.32
15	23.38	231.12	1507.93	3312.28	13 596.26	23 213.19
16	39.86	586.97	1863.45	3534.50	13 730.38	23 107.77
17	0.06	377.73	1772.71	3061.09	13 765.16	23 168.32
18	67.10	302.11	1466.88	2797.85	13 347.41	22 103.96
19	52.84	371.49	1719.27	3365.47	14 469.17	24 708.20
20	28.56	468.63	1876.91	3768.89	14 415.28	24 735.23
21	29.13	577.05	2412.74	4042.06	13 470.03	22 398.08
22	113.05	375.93	1844.54	3454.07	10 826.14	17 720.47
Avg per bus and hour	45.7	39.1	35.0	33.2	27.2	23.1

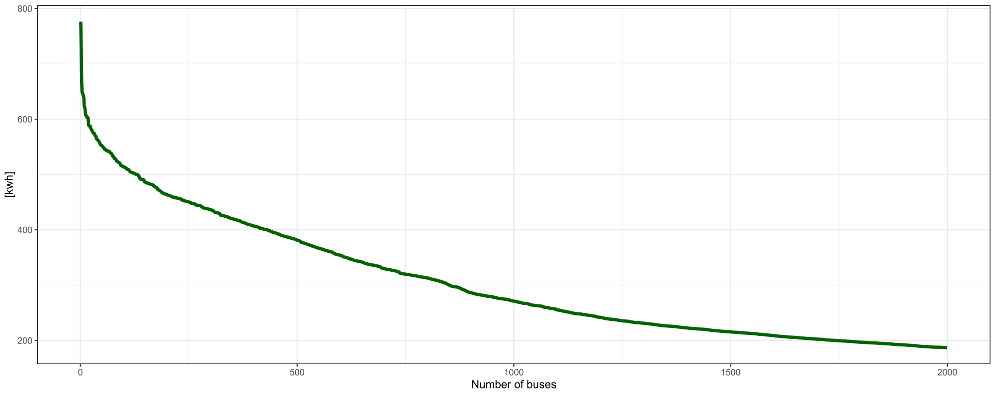

Fig. 7. Marginal savings per number of diesel buses replaced per day.

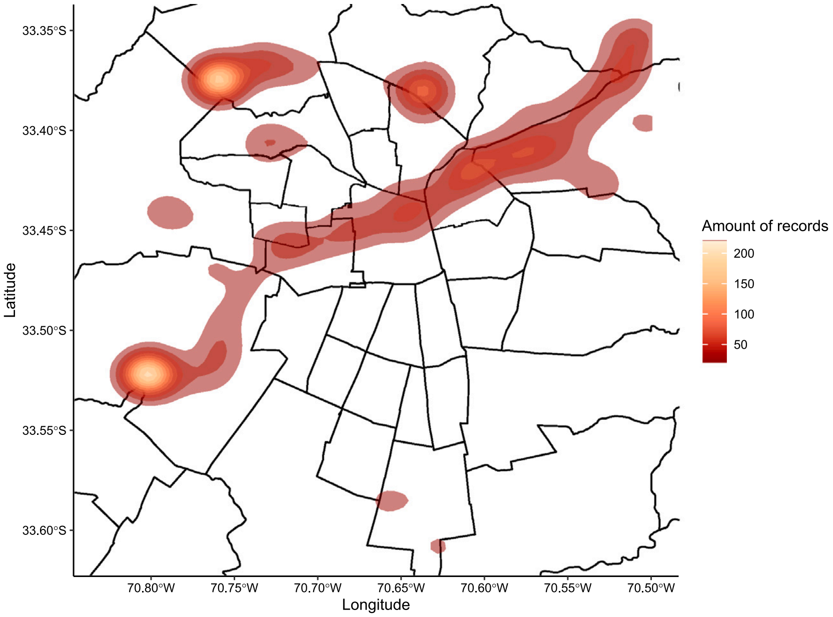

Fig. 8. Heatmap of the distribution of the 50 buses with the greatest potential for energy savings.

information on energy consumption that reduces part of the uncertainty that exists when evaluating possible public policies.

Fig. 7 shows the daily savings curve for different amounts of diesel buses replaced by electric buses. This analysis helps determine the minimum number of diesel buses to be replaced to attain an energy-saving goal. Additionally, these results make it possible to focus investment in the face of limitations that may arise that prevent the replacement of the entire fleet, including economic and technical feasibility.

On the other hand, Fig. 8 identifies the sectors of the region in which the 50 buses with the greatest potential for energy savings transit. This number of buses is considered because it is similar to some of the latest batches of new electric buses incorporated into the public transport network of Santiago, Chile. It can be seen that most of the buses with the greatest savings potential cross Santiago in the east–west direction. Both sectors have significant height differences between them, so the buses that make this type of route frequently face very marked positive and negative slopes. Taking into account the importance of this variable in energy consumption, the slopes of the route could explain, in part, why it is convenient to replace these buses with electric ones.

6. Concluding remarks

Public transport systems in the world have undergone a transition towards electro-mobility in recent years, incorporating more electric vehicles into their fleets. Motivated by the above, the number of works in the literature that seek to estimate the energy consumption of buses has increased considerably in recent years. However, and to the best of our knowledge, most of these efforts base their estimates on measurements collected from experiments under controlled conditions –and therefore may not reflect actual operating conditions–, or they perform the analysis for a limited subset of the system fleet. In this paper we seek to close this gap, estimating the energy consumption of all the buses in the public transport system of Santiago, Chile, using real and disaggregated data from the network. Our results allow us to analyze the real energy consumption of electric buses, and open the doors to the design, planning and operation of actions or policies that seek to reduce energy consumption from a holistic view of the network.

In this study, two data sources of the public transport bus system in Santiago, Chile are used to, on the one hand, estimate the energy consumption of electric buses, and on the other, identify relevant variables. The first base corresponds to GPS information of all the buses that operate in the system, which is provided by the DTPM. The second base, instead, corresponds to information from a subset of buses, which is provided by a sensorization company. These data include the use of the brake pedal and the use of air conditioning, among others. Based on this information, *machine learning* and physical base models are calibrated to estimate consumption across the entire network.

Finally, computational experiments are carried out comparing different instances and models, obtaining public policy recommendations.

The results indicate that using the sensorized data provides better results in the vast majority of cases compared to the use of only the data available for the entire network. However, not all buses are equipped with sensors to be able to measure these variables. Of the models that can be implemented for the entire network, since they only use variables available for all the buses, the linear regression (when the temporality is high) and the linear SVR stand out in most cases. This last model delivers a percentage error of 22%, standing out compared to the physical base model that presents an error of 43%, and the use of a theoretical value that has an error of 104%. This demonstrates the value of using real and disaggregated data capable of measuring particular situations that buses face in reality.

As the number of electric buses in Transantiago's network is set to rise, it is critical to understand the impact of increased electromobility on city public transport. Accurate energy consumption estimates can enhance fleet management and route design, helping decide which services to electrify first. Thus, the best model, namely SVR, is used to quantify the energy savings that occur when replacing 1, 10, 50, 100, 500, and 1000 diesel buses with electric buses, delivering a daily total of 776 [kWh], 6644 [kWh], 29 737 [kWh], 56 356 [kWh], 231 611 [kWh], and 392 473 [kWh], respectively. Moreover, we determine the areas of the region where the 50 buses with the highest potential for energy savings are most frequently found. Interestingly, the majority of buses with the greatest potential for energy savings traverse Santiago in an east–west direction. This route involves significant altitude variations, subjecting buses to steep inclines and declines. All these analyses allow decision-makers to target investment by determining the buses with higher energy consumption potential in the face of budget constraints.

One of the main limitations of this study is that the data used may not provide all the relevant information to estimate energy consumption. In particular, the empirical measurements are circumscribed to six electric buses that make similar routes to each other. In consideration of the above, the variability of the measured data is limited. In addition, the statistical learning models do not explain the consumption of buses that operate with diesel, since they were calibrated with real data from electric buses. On the other hand, the study period can also generate a bias in the estimates made, although it is to be expected that the variation in energy consumption between days of the week will not be as significant. Lastly, one of the assumptions made to assess the benefits of switching from conventional buses to electric buses is that the driving mode of the buses is the same. This might not always be true as different bus technologies might alter the driving style generating different data. Additionally, note that our models do not consider contextual variables, such as the conditions of the route, built environment, or traffic signals. Therefore, incorporating these types of variables could improve the performance of the models. Finally, we focus solely on the energy savings provided by electric buses when studying the benefits of replacing conventional buses. However, electric buses typically offer a higher level of service due to additional features, and thus, they can enhance the overall user experience. Consequently, when deciding which service lines to electrify first, these features and their impact on users should be taken into account.

Multiple lines of research can be pursued following this work to tackle relevant emerging problems in public transport electromobility. For example, investigating the use of estimates to develop performance indicators related to drivers, buses, and other relevant factors holds significant interest. This approach can enable the creation of targeted training plans or incentives for drivers based on their individual performance, thereby enhancing overall efficiency, safety, and quality of service in the transportation system. Moreover, utilizing estimates to modify the routes within the Operational Plan of the Santiago transport system is another important area of exploration. By incorporating energy consumption estimates into the planning process, it becomes possible to optimize routes with the objective of reducing energy consumption while maintaining service levels. This integration of energy considerations can contribute to building a more sustainable and environmentally friendly transportation system. Finally, energy consumption projections derived from the estimates can serve as inputs for operations research problems aimed at optimizing the location of load centers. This presents a considerable challenge in establishing a network of charging stations capable of meeting the high energy demands of electric buses or trains. It requires meticulous planning, investment, and coordination among various stakeholders, including transport authorities, energy providers, and infrastructure developers.

Table 15
List of abbreviations.

Abbreviation	Meaning
GPS	Global Positioning System
NDC	National Determined Contributions
CO2	Carbon dioxide
EV	Electric vehicle
ML	Machine learning
DTPM	Metropolitan public transport directory, by the acronym in Spanish
RPM	Revolutions per minute
A/C	Air conditioning
SVR	Support Vector Regression
NN	Neural Network
LR	Linear Regression
RF	Random Forest
MAPE	Mean absolute percentage error
PM	Physical base model
RMSE	Root mean square error

CRediT authorship contribution statement

Franco Basso: Conceptualization, Data curation, Writing – original draft, Investigation, Software, Methodology, Supervision, Writing – review & editing. **Felipe Feijoo:** Conceptualization, Methodology, Writing – review & editing. **Raúl Pezoa:** Conceptualization, Methodology, Writing – review & editing. **Mauricio Varas:** Conceptualization, Methodology, Writing – review & editing. **Brian Vidal:** Conceptualization, Data curation, Writing – original draft, Investigation, Software, Methodology, Supervision, Writing – review & editing.

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.

Data availability

Data will be made available on request.

Acknowledgments

Franco Basso gratefully acknowledges the financial support from the Complex Engineering Systems Institute, ISCI (grant ANID PIA/BASAL AFB180003). Felipe Feijoo acknowledges funding from ANID FONDECYT regular #1221894.

Appendix

See Table 15.

References

[1] Esanu V, Motroi A, Nuca I, Nuca I. Electrical buses: Development and implementation in Chisinau municipality, Moldova. In: 2019 international conference on electromechanical and energy systems (SIELMEN). IEEE; 2019, p. 1–5.

[2] Woodcock J, Banister D, Edwards P, Prentice AM, Roberts I. Energy and transport. Lancet 2007;370(9592):1078–88.

[3] González S, Feijoo F, Basso F, Subramanian V, Sankaranarayanan S, Das TK. Routing and charging facility location for EVs under nodal pricing of electricity: A bilevel model solved using special ordered set. IEEE Trans Smart Grid 2022;13(4):3059–68.

[4] Mahmoud M, Garnett R, Ferguson M, Kanaroglou P. Electric buses: A review of alternative powertrains. Renew Sustain Energy Rev 2016;62:673–84.

[5] Zhou B, Wu Y, Zhou B, Wang R, Ke W, Zhang S, Hao J. Real-world performance of battery electric buses and their life-cycle benefits with respect to energy consumption and carbon dioxide emissions. Energy 2016;96:603–13.

[6] Hung L-J, Tsai S-S, Chen P-S, Yang Y-H, Liou S-H, Wu T-N, Yang C-Y, et al. Traffic air pollution and risk of death from breast cancer in Taiwan: fine particulate matter (PM2. 5) as a proxy marker. Aerosol Air Qual Res 2012;12(2):275–82.

[7] Wdaah L, Müller S. Efficiency analysis of an electrification concept for a catering truck. In: 2016 IEEE transportation electrification conference and expo, Asia-Pacific (ITEC Asia-Pacific). IEEE; 2016, p. 837–42.

[8] Kühne R. Electric buses–An energy efficient urban transportation means. Energy 2010;35(12):4510–3.

[9] Feng W, Figliozzi M. An economic and technological analysis of the key factors affecting the competitiveness of electric commercial vehicles: A case study from the USA market. Transp Res C 2013;26:135–45.

[10] Miles J, Potter S. Developing a viable electric bus service: The Milton Keynes demonstration project. Res Transp Econ 2014;48:357–63.

[11] Kakuhama Y, Kato J, Fukuizumi Y, Watabe M, Fujinaga T, Tada T. Next-generation public transportation: Electric bus infrastructure project. Mitsubishi Heavy Ind Tech Rev 2011;48(1):1–4.

[12] Aldenius M, Mullen C, Pettersson-Löfstedt F. Electric buses in England and Sweden–overcoming barriers to introduction. Transp Res D 2022;104:103204.

[13] Subramanian V, Feijoo F, Sankaranarayanan S, Melendez K, Das TK. A bilevel conic optimization model for routing and charging of EV fleets serving long distance delivery networks. Energy 2022;251:123808.

[14] Clairand J-M, Guerra-Terán P, Serrano-Guerrero X, González-Rodríguez M, Escrivá-Escrivá G. Electric vehicles for public transportation in power systems: A review of methodologies. Energies 2019;12(16):3114.

[15] Rodrigues AL, Seixas SR. Battery-electric buses and their implementation barriers: Analysis and prospects for sustainability. Sustain Energy Technol Assess 2022;51:101896.

[16] Soler D, Ubilla L, Pudrencio G, Vial C, Lambert MJ, Pérez A, et al. Estrategia nacional de electromovilidad. 2022.

[17] Abdelaty H, Al-Obaidi A, Mohamed M, Farag HE. Machine learning prediction models for battery-electric bus energy consumption in transit. Transp Res D 2021;96:102868.

[18] Li P, Zhang Y, Zhang K, Jiang M. The effects of dynamic traffic conditions, route characteristics and environmental conditions on trip-based electricity consumption prediction of electric bus. Energy 2021;218:119437.

[19] Nan S, Tu R, Li T, Sun J, Chen H. From driving behavior to energy consumption: A novel method to predict the energy consumption of electric bus. Energy 2022;261:125188.

[20] Pamuła T, Pamuła D. Prediction of electric buses energy consumption from trip parameters using deep learning. Energies 2022;15(5):1747.

[21] Jiang J, Yu Y, Min H, Cao Q, Sun W, Zhang Z, Luo C. Trip-level energy consumption prediction model for electric bus combining Markov-based speed profile generation and Gaussian processing regression. Energy 2023;263:125866.

[22] Zhao L, Ke H, Huo W. A frequency item mining based energy consumption prediction method for electric bus. Energy 2023;263:125915.

[23] Perumal SS, Lusby RM, Larsen J. Electric bus planning & scheduling: A review of related problems and methodologies. European J Oper Res 2022;301(2):395–413.

[24] Sinhuber P, Rohlfs W, Sauer DU. Study on power and energy demand for sizing the energy storage systems for electrified local public transport buses. In: 2012 IEEE vehicle power and propulsion conference. IEEE; 2012, p. 315–20.

[25] Bartłomiejczyk M, Kołacz R. The reduction of auxiliaries power demand: The challenge for electromobility in public transportation. J Clean Prod 2020;252:119776.

[26] Pettersson P, Johannesson P, Jacobson B, Bruzelius F, Fast L, Berglund S. A statistical operating cycle description for prediction of road vehicles' energy consumption. Transp Res D 2019;73:205–29.

[27] Al-Ogaili AS, Ramasamy A, Hashim TJT, Al-Masri AN, Hoon Y, Jebur MN, Verayiah R, Marsadek M. Estimation of the energy consumption of battery driven electric buses by integrating digital elevation and longitudinal dynamic models: Malaysia as a case study. Appl Energy 2020;280:115873.

[28] Franca A. Electricity consumption and battery lifespan estimation for transit electric buses: drivetrain simulations and electrochemical modelling (Ph.D. thesis), University of Victoria; 2018.

[29] Ma X, Miao R, Wu X, Liu X. Examining influential factors on the energy consumption of electric and diesel buses: A data-driven analysis of large-scale public transit network in Beijing. Energy 2021;216:119196.

[30] Liu L, Kotz A, Salapaka A, Miller E, Northrop WF. Impact of time-varying passenger loading on conventional and electrified transit bus energy consumption. Transp Res Rec 2019;2673(10):632–40.

[31] Zhang S, Wu Y, Liu H, Huang R, Yang L, Li Z, Fu L, Hao J. Real-world fuel consumption and CO2 emissions of urban public buses in Beijing. Appl Energy 2014;113:1645–55.

[32] Chen Y, Wu G, Sun R, Dubey A, Laszka A, Pugliese P. A review and outlook of energy consumption estimation models for electric vehicles. 2020, arXiv preprint arXiv:2003.12873.

[33] Gallet M, Massier T, Hamacher T. Estimation of the energy demand of electric buses based on real-world data for large-scale public transport networks. Appl Energy 2018;230:344–56.

[34] Wu X, Freese D, Cabrera A, Kitch WA. Electric vehicles' energy consumption measurement and estimation. Transp Res D 2015;34:52–67.

[35] Luin B, Petelin S, Al-Mansour F. Microsimulation of electric vehicle energy consumption. Energy 2019;174:24–32.

[36] Zeng W, Miwa T, Morikawa T. Exploring trip fuel consumption by machine learning from GPS and CAN bus data. J East Asia Soc Transp Stud 2015;11:906–21.

[37] López-Martínez JM, Jiménez F, Páez-Ayuso FJ, Flores-Holgado MN, Arenas AN, Arenas-Ramirez B, Aparicio-Izquierdo F. Modelling the fuel consumption and pollutant emissions of the urban bus fleet of the city of Madrid. Transp Res D 2017;52:112–27.

[38] Felipe J, Amarillo JC, Naranjo JE, Serradilla F, Díaz A. Energy consumption estimation in electric vehicles considering driving style. In: 2015 IEEE 18th international conference on intelligent transportation systems. IEEE; 2015, p. 101–6.

[39] Pamuła T, Pamuła W. Estimation of the energy consumption of battery electric buses for public transport networks using real-world data and deep learning. Energies 2020;13(9):2340.

[40] Sennefelder RM, Martín-Clemente R, González-Carvajal R. Energy consumption prediction of electric city buses using multiple linear regression. Energies 2023;16(11):4365.

[41] Qin W, Wang L, Liu Y, Xu C. Energy consumption estimation of the electric bus based on grey wolf optimization algorithm and support vector machine regression. Sustainability 2021;13(9):4689.

[42] Lawson CE, Martí JM, Radivojevic T, Jonnalagadda SVR, Gentz R, Hillson NJ, Peisert S, Kim J, Simmons BA, Petzold CJ, et al. Machine learning for metabolic engineering: A review. Metab Eng 2021;63:34–60.

[43] Sun R, Chen Y, Dubey A, Pugliese P. Hybrid electric buses fuel consumption prediction based on real-world driving data. Transp Res D 2021;91:102637.

[44] Li P, Zhang Y, Zhang Y, Zhang K. Prediction of electric bus energy consumption with stochastic speed profile generation modelling and data driven method based on real-world big data. Appl Energy 2021;298:117204.

[45] Chen Y, Zhang Y, Sun R. Data-driven estimation of energy consumption for electric bus under real-world driving conditions. Transp Res D 2021;98:102969.

[46] Basso F, Frez J, Hernández H, Leiva V, Pezoa R, Varas M. Crowding on public transport using smart card data during the COVID-19 pandemic: New methodology and case study in Chile. Sustainable Cities Soc 2023;104712.

[47] Basso F, Frez J, Martínez L, Pezoa R, Varas M. Accessibility to opportunities based on public transport gps-monitored data: The case of Santiago, Chile. Travel Behav Soc 2020;21:140–53.

[48] Pezoa R, Basso F, Quilodrán P, Varas M. Estimation of trip purposes in public transport during the COVID-19 pandemic: The case of Santiago, Chile. J Transp Geogr 2023;109:103594.

[49] Munizaga MA, Palma C. Estimation of a disaggregate multimodal public transport origin–Destination matrix from passive smartcard data from Santiago, Chile. Transp Res C 2012;24:9–18.

[50] Frez J, Baloian N, Pino JA, Zurita G, Basso F. Planning of urban public transportation networks in a smart city. J UCS 2019;25(8):946–66.

[51] Maia R, Silva M, Araújo R, Nunes U. Electric vehicle simulator for energy consumption studies in electric mobility systems. In: 2011 IEEE forum on integrated and sustainable transportation systems. IEEE; 2011, p. 227–32.

[52] Zhang R, Yao E. Mesoscopic model framework for estimating electric vehicles' energy consumption. Sustainable Cities Soc 2019;47:101478.

[53] Asamer J, Graser A, Heilmann B, Ruthmair M. Sensitivity analysis for energy demand estimation of electric vehicles. Transp Res D 2016;46:182–99.

[54] Gao Z, Lin Z, LaClair TJ, Liu C, Li J-M, Birky AK, Ward J. Battery capacity and recharging needs for electric buses in city transit service. Energy 2017;122:588–600.

[55] Lajunen A, Kivekäs K, Baldi F, Vepsäläinen J, Tammi K. Different approaches to improve energy consumption of battery electric buses. In: 2018 IEEE vehicle power and propulsion conference (VPPC). IEEE; 2018, p. 1–6.

[56] Hjelkrem OA, Lervåg KY, Babri S, Lu C, Södersten C-J. A battery electric bus energy consumption model for strategic purposes: Validation of a proposed model structure with data from bus fleets in China and Norway. Transp Res D 2021;94:102804.

[57] James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning, vol. 112. Springer; 2013.

[58] Bag S, Gupta K, Deb S. A review and recommendations on variable selection methods in regression models for binary data. 2022, arXiv preprint arXiv:2201.06063.

[59] Rudnicki WR, Wrzesień M, Paja W. All relevant feature selection methods and applications. Feature Sel Data Pattern Recognit 2015;11–28.

[60] Sanchez-Pinto LN, Venable LR, Fahrenbach J, Churpek MM. Comparison of variable selection methods for clinical predictive modeling. Int J Med Inform 2018;116:10–7.

[61] Speiser JL, Miller ME, Tooze J, Ip E. A comparison of random forest variable selection methods for classification prediction modeling. Expert Syst Appl 2019;134:93–101.

[62] Basso F, Pezoa R, Varas M, Villalobos M. A deep learning approach for real-time crash prediction using vehicle-by-vehicle data. Accid Anal Prev 2021;162:106409.

[63] Degenhardt F, Seifert S, Szymczak S. Evaluation of variable selection methods for random forests and omics data sets. Brief Bioinform 2019;20(2):492–503.

[64] Breiman L. Bagging predictors. Mach Learn 1996;24(2):123–40.

[65] Hastie T, Tibshirani R, Friedman J. Random forests. In: The elements of statistical learning. Springer; 2009, p. 587–604.

[66] Trafalis TB, Ince H. Support vector machine for regression and applications to financial forecasting. In: Proceedings of the IEEE-INNS-ENNS international joint conference on neural networks. IJCNN 2000. Neural computing: new challenges and perspectives for the New Millennium, Vol. 6. IEEE; 2000, p. 348–53.

[67] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 1986;323(6088):533–6.