IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 11, 2021, accepted March 31, 2021, date of publication April 5, 2021, date of current version April 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3071180

Analysis of Energy Consumption at Slow Charging Infrastructure for Electric Vehicles

MILAN STRAKA[1,2], RUI CARVALHO[3,4,5], GIJS VAN DER POEL[6], AND L'UBOŠ BUZNA[1,7]

[1]Department of Mathematical Methods and Operations Research, University of Žilina, 010 26 Žilina, Slovakia
[2]Division of Information and Communication Technologies, University Science Park, University of Žilina, 010 26 Žilina, Slovakia
[3]Department of Engineering, Durham University, Durham DH1 3LE, U.K.
[4]Durham Energy Institute, Durham University, Durham DH1 3LE, U.K.
[5]Institute for Data Science, Durham University, Durham DH1 3LE, U.K.
[6]ElaadNL, 6812 AR Arnhem (GL), The Netherlands
[7]Department of International Research Projects-ERAdiate+, University of Žilina, 010 26 Žilina, Slovakia

Corresponding author: Milan Straka (milan.straka@fri.uniza.sk)

This work was supported in part by the research project "Data Analysis Methods and Decisions Support Tools for Service Systems Supporting Electric Vehicles" under Grant VEGA 1/0089/19, in part by the "Allocation of Limited Resources to Public Service Systems with Conflicting Quality Criteria" under Grant APVV-19-0441, in part by the Slovak Research and Development Agency under Contract SK-IL-RD-18–005, in part by the Operational Program Integrated Infrastructure 2014–2020 "Innovative Solutions for Propulsion, Power, and Safety Components of Transport Vehicles" through the European Regional Development Fund under Grant ITMS313011V334, and in part by the Operational Program Research and Innovation in the frame of the project: ICT Products for Intelligent Systems Communication, through the European Regional Development Fund under Grant ITMS2014+313011T413.

ABSTRACT Here, we develop a data-centric approach to analyse which activities, functions, and characteristics of the environment surrounding the slow charging infrastructure impact the distribution of the electricity consumed at slow charging infrastructure. We analysed the probability distribution of energy consumption and its relation to indicators characterising charging events to gain basic insights. The energy consumption can be satisfactorily modelled by a transformed beta distribution and the number of charging transactions is the driving factor among the characteristics constituting the energy consumption. We collected geospatial datasets and prepared a large number of candidate features modelling the spatial context in which the charging infrastructure operates. Using statistical methods, we identified and interpreted a relatively small subset of the most influential features correlated with energy consumption. The majority of these features are related to the economic prosperity of residents. Residents and businesses with high (low) income, situated nearby charging infrastructure, are linked to a positive (negative) impact on energy consumption. Similarly, charging infrastructure located close to expensive newly built housing shows higher energy consumption. The largest adverse impact has the high concentration of residents receiving social assistance. By applying the methodology to a specific charging infrastructure class, e.g. determined by the used rollout strategy, we differentiated the selected features. Business types, working sector of residents and public venues in the proximity are linked to higher consumption of energy at charging infrastructure deployed strategically. Characteristics linked with the age structure of the population are linked to the energy consumption at charging infrastructure placed based on the demand. Data collection and data processing are among the most time-consuming activities. The paper provides valuable insights into which data to collect and use as features when developing prediction models to inform charging infrastructure deployment and planning of power grids.

INDEX TERMS Electric vehicles, charging infrastructure, energy consumption, variable selection.

I. INTRODUCTION

The European Union (EU) is moving towards commitments adopted under the Paris Agreement by aiming at domestic CO_2 cuts of at least 40% below 1990 levels by 2030 [1]. To tackle this challenge, the deployment of plug-in hybrid electric vehicles (PHEVs), battery electric vehicles (BEVs) and fuel cell electric vehicles (FCEVs) appears to be inevitable [2, p. 91]. Electric mobility is growing at a rapid speed. In 2018, the number of new electric car sales almost

The associate editor coordinating the review of this manuscript and approving it for publication was Rui Xiong.

doubled compared to 2017, and the global electric car fleet exceeded 5.1 million [3, p. 33]. The world's largest electric car market is the People's Republic of China, followed by the EU and the United States (US). The global leaders in terms of electric car market share are Norway, Sweden, and the Netherlands, which is also dominant in terms of the charging infrastructure density, not only in Europe but globally [3, p. 4]. In 2018, the global EV fleet consumed 58 TWh of electricity, which is comparable to Switzerland's total electricity demand in 2017 [3, p. 9]. The majority of outlooks envisions growing trend. In 2030, the global electric car sales are expected to reach 23 million, the stock will exceed 130 million vehicles (excluding two/three-wheelers), and electricity demand from EVs is estimated to reach almost 640 TWh [3, p. 6]. In this scenario, slow chargers, which can provide flexibility services to power systems [4], are estimated to account for more than 60% of the total electricity consumed globally to charge electric vehicles (EVs). Consequently, the number of new applications for prediction algorithms in the domain of EV charging is steadily increasing.

A. LITERATURE REVIEW

Although electric vehicles have a longer history than fossil fuel vehicles, their mass adoption has started only recently [5]. With the growing number of EVs on roads in the last few years, a lot of research in the EV domain has based on data-centric (i.e. machine learning or data science) approaches. Several recent review papers [6]–[8] provide a comprehensive overview of data sources and summarise data science (machine learning) literature in the domain of EV charging.

Applications of data science methods to EV charging already include the whole spectrum of supervised learning methods (e.g. K-nearest neighbour [9], linear regression [10], decision trees and their aggregations [11], support vector regression [12], etc.), unsupervised learning methods (e.g. clustering [13], Gaussian mixture models [14], and kernel density estimator [15]), and deep learning [16]. Deterministic models, providing scalar predictions, dominate, while the probabilistic models have not received the same attention [17]. Problems addressed by data science methods in EV domain range from forecasts of EV sales [18], EV battery-related problems (e.g. prediction of the state of charge [19] and prediction of battery cycle life before capacity degradation [20]), EV load analysis (e.g. EV load prediction [21], detection of households charging EV [22]) to charging infrastructure planning [23]–[25], planning of power grids [26] and predictions of usage patterns [27].

Studies related to charging infrastructure focus on temporal and spatial aspects of EV charging. Typically, temporal aspects refer to the utilisation of charging infrastructure. Predictability of energy consumption at the level of single chargers was investigated in [28], finding potentially useful results only for some chargers. Reference [29] identified and evaluated the time-series seasonal auto-regressive integrated moving average (ARIMA) models of EV load aggregated over 2400 chargers. The long-term models (for two years) were found decidedly less accurate than the near-term models (for the most recent 60 weekdays and 24 weekend days). Comparison of ARIMA models with decision trees, considering some exogenous features, concluded that former models are a better choice for forecasting aggregated EV charging loads [21]. Commonly used machine learning algorithms (K-nearest neighbour, pattern sequence-based forecasting, support vector regression and random forest), yielded very similar prediction errors when applied to forecasts of EV charging load based on customer profile and charger measurements [30]. In [31], the day-ahead EV charging load is forecasted as EV charging occurrence-time and the ''no charge'' day respectively, by several widely used machine learning algorithms. The best performance achieved the hybrid model combining random forest, naive Bayes and XGBoost. In [22] authors used a data-driven approach to identify households charging EVs. Utilising the historical electricity consumption data, including the kurtosis of the residential electricity load, a random forest classifier reached a prediction accuracy over 90%.

Spatial analyses of charging infrastructure utilisation are less developed in the literature than temporal. A set of key performance indicators characterising utilisation of chargers was defined and used to compare two rollout strategies: demand-driven and strategic rollout [32]. No rollout strategy is favourable over the other on all metrics, and the difference between strategies reduces as the EV adoption progresses. A preliminary exploratory analysis of spatial patterns formed by energy consumption on charging stations was presented in [33]. A study found a heterogeneous pattern, observing higher energy intensity in a small number of urban areas and 50% of the energy supplied comes from 19.6% of chargers.

Location features, i.e. features that characterise the close vicinity of charging infrastructures, are frequently used to improve EV charging predictions. Reference [23] employs XGBoost model and five features describing the location and parameters of the charging infrastructure to predict its utilisation. It also demonstrates how the model support decisions on locating the charging infrastructure at the level of zones with a radius of 3 km. Only two location features were used, the number of points of interest and the number of competitive charging stations. Regression models (linear regression, XGBoost, artificial neural networks) were used to predict the departures of vehicles to improve smart charging heuristic [10]. The set of considered features consisted of five categorical and four numerical features, among them three location features (floor of the car park, car park and charge point) were used. The multinomial logistic regression was used by [34] to determine key factors explaining heterogeneity in the charging duration of categorised charging sessions. The time-of-day-related variables and the type of charging station have the most substantial effect. Some location-related features such as type of the urban area, the density of chargers and parking possibilities were considered in this study.

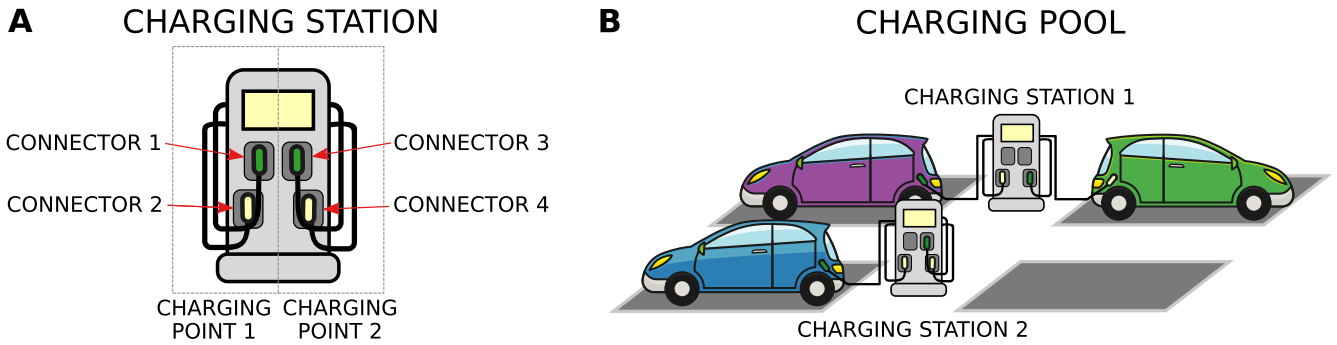

FIGURE 1. Schematics illustrating the terminology suggested in [37] to denote the main components of the charging infrastructure for electric vehicles.

The ability of machine learning methods (random forest, gradient boosting and XGBoost) to predict the idle time (i.e. the time when an EV is connected to a charger without charging) was evaluated by [35]. The best model is XGBoost, reaching R^2 score of 0.603. The most influential features are time-of-day-related features and the total energy supplied. Only one location feature, namely the type of the closest road segment, was considered. Using the location features, authors in [36] built prediction models for the popularity of charging infrastructure (i.e. the number of unique users). In summary, the research gaps can be summarized as follows:

- Location features used by prediction models in the EV charging domain vary from model to model.
- Researchers select location features intuitively, without having a clear idea about their significance.
- The number of location features used in prediction models is typically very small.

B. OUR CONTRIBUTION

Recently, several prediction models appeared in the literature [10], [23], [34]–[36] explaining performance indicators of charging infrastructure also from location features. The candidate set of influence factors that can be described by location features is large. In general, the decision which data to collect when developing a model is difficult, and data collection and data processing are the most time-consuming activities. Hence, providing insights which data can be used to derive useful features is highly beneficial. This paper brings the following contributions:

- Our primary contribution is in extending the available knowledge of location factors, potentially affecting the energy consumption at public charging infrastructure. From the collected data, we extracted more than 120 location features. We employ available statistical methods, specifically designed to explore a large set of potentially influential features, and we identified a relatively small relevant subset. We provide interpretations of significant location features.
- As our secondary contribution, we analyse the distribution of energy consumption and reveal how it is shaped by indicators characterising the charging events.
- Our last contribution is methodological. In the analysis, we consider the influence of multicollinearity and statistical stability of selected regression coefficients to improve the reliability of results. These two problems are very well known in statistics, however, they are often overlooked in field studies that employ regression approaches combined with variable selection.

II. MATERIALS AND METHODS

A. TERMINOLOGY

Charging of electric vehicles is a new field and various terminologies can be found in the literature. We follow [37], where a *connector* is defined as a physical interface between an electric vehicle and charging infrastructure through which electricity is delivered. Due to the incompatibilities of connector types used by different car manufacturers, several connectors might be available at a charging point, however, no more than one connector can be active at a time. A *charging point* is an energy delivery device equipped with one or more connectors. The charging point can charge an EV with a power which is less or equal to the maximum power, given in kW, referred to as *the charging capacity*. A *charging station* is composed of one or multiple charging points. A user identification interface and all human-machine interfaces are attributed to the charging station and are shared by all charging points. A *charging pool* is one or a collection of charging stations including the adjacent parking lots. Components of the charging infrastructure are visualised in Figure 1. A charging transaction starts by plugging a connector into an EV at a *start time* and ends by unplugging the EV at an *end time*. The difference between end time and start time is *the connection time*. A part of the connection time when EV was charging is referred to as *the charging time*. *The idle time* is the connection time minus the charging time. *An RFID card* identifies EV driver and is used to initiate and terminate charging session.

B. EVnetNL DATASET

The EVnetNL dataset has been provided to us for research purposes by ElaadNL, a Dutch research organisation involved in the development, deployment and operation of EV

charging infrastructure. The data is organised into two tables. The table *Transactions* contains 1 060 763 rows, each characterising an individual charging event, with columns such as an identifier of a charging point, GPS coordinates of charging station, start time, end time, connection time, idle time, charging time, the number of the used RFID cards, consumed energy and unique identifier of the charging event. The second table, *Meterreadings*, has 32 440 911 rows, each corresponding to a meter reading that is taken each 15 minutes if a vehicle is connected. A meter reading is described by the transaction's identifier, the charging point's identifier, UTC timestamp and value of the meter.

The EVnetNL dataset covers 1747 charging stations equipped with 2893 charging points (identified by unique labels) operated by the ElaadNL in the period from January 2012 until March 2016. Some charging stations are located close to each other (e.g. in one parking lot). Consequently, the urban context of charging stations cannot be distinguished. Therefore, we considered the charging pools as the main object of the analysis. Locations of charging pools we estimated by aggregating the charging stations. For each charging station, we identified a set of neighbouring stations located within the radius of 50 meters. For every pair of stations distant more than 50 meters from each other, we found an empty intersection of neighbouring stations' sets. Therefore, we selected one representative charging station in each set of neighbouring stations, and all stations in the set were merged and formed a charging pool.

In the analysis of energy consumption, we consider only the period from January 1, 2015, until December 31, 2015, as this is the latest available complete year, when the number of charging pools was reasonably stable (see Figure 2A). As shown in Figure 2B, the number of active users, estimated from the number of RFID cards in use, exceeded the value 15 000 and it has been relatively stable throughout the year 2015 as well. We obtained a set of 1604 charging pools (accommodating 1660 charging stations) operational in 2015 while being distributed across the entire area of the Netherlands (see Figure 2C). In Figure 2D we analyse spatial representativeness of the EVnetNL dataset by calculating the ratio between the number of EVnetNL charging pools operational in 2015 and the number of charging pools in the Charging pools 2015 dataset (for more information about the dataset, please refer to the Section S1 I of the SI file) for cells of a regular square grid. The ratio takes higher values in the east and south of the Netherlands. The EVnetNL dataset covers smaller cities better, while in large cities, such as Amsterdam and Rotterdam, only a small percentage of charging stations is covered.

We excluded 40 charging transactions for which either meter values or charging start and stop times were inconsistent across Meterreadings and Transactions tables. To eliminate charging pools with sparse usage patterns, we excluded from the analyses 218 pools with either less than 30 charging transactions in 2015 or less than 1 kW charging capacity. To minimise the effects of the transition period that follows the introduction of a new charging pool, we consider only charging pools that have been in use before January 1, 2015 (we excluded 87 pools established in 2015 and later). After these rearrangements, we retained 369 550 transactions taking place on 1 386 EVnetNL charging pools. The large majority of charging pools possess 1 or 2 charging points and deliver power ranging up to 12.5 kW. Often, the fast charging is declared when the power is exceeding the value of 22 kW [37], hence, all considered charging pools are used for slow charging (see Figure 3A-B).

C. GEOSPATIAL DATASETS

To characterise the area and human activities taking place in the vicinity of charging pools distributed across the territory of the Netherlands, we collected potentially relevant publicly available geospatial datasets illustrated in Figure 4.

The geospatial datasets describe locations on the Earth's surface by geometric objects (points, polylines, polygons) and associate geometric objects with alphanumeric attributes. We inspected all available attributes, and if multiple datasets contained the same or very similar data, we considered a complete source or the source providing data with higher resolution. Moreover, we excluded attributes that lead to multicollinearity (e.g. from the triplet of attributes the total population, the population of men and the population of women, we considered only the first attribute as the populations of men and women tend to be very similar, hence, constitute one half of the total population). As we focus on the spatial distribution of energy consumption at the long time scale (we analyse the annual spatial distribution of energy consumption among charging pools), we excluded from the analysis factors that affect the charging behaviour at short time scales such as air temperature or other weather characteristics. Such factors would be more appropriate for temporal analysis of energy consumption.

In what follows, we briefly describe each geospatial dataset. The complete list of selected attributes for each of 9 datasets is given in the SI file, Tables S2- S10. We do not apply any criteria other than described in the Supplementary information file (Section S2) to decide which attributes are selected for the analysis. The used methodology addresses potential data problems and selects relevant features.

1) POPULATION CORES

The population cores are continuous spatial units with at least 25 homes or 50 registered residents [39]. The dataset associates the population cores with the information about the households (e.g. size and composition), the cardinality of population age groups, and family and civil status of residents. It also includes the information about the employment rate of residents, type of their occupation and characteristics of real-estate properties. The spatial resolution of this dataset is rather low (typically, a population core corresponds to a municipality) and we use it to investigate whether aggregate characteristics of municipalities can explain the usage pattern

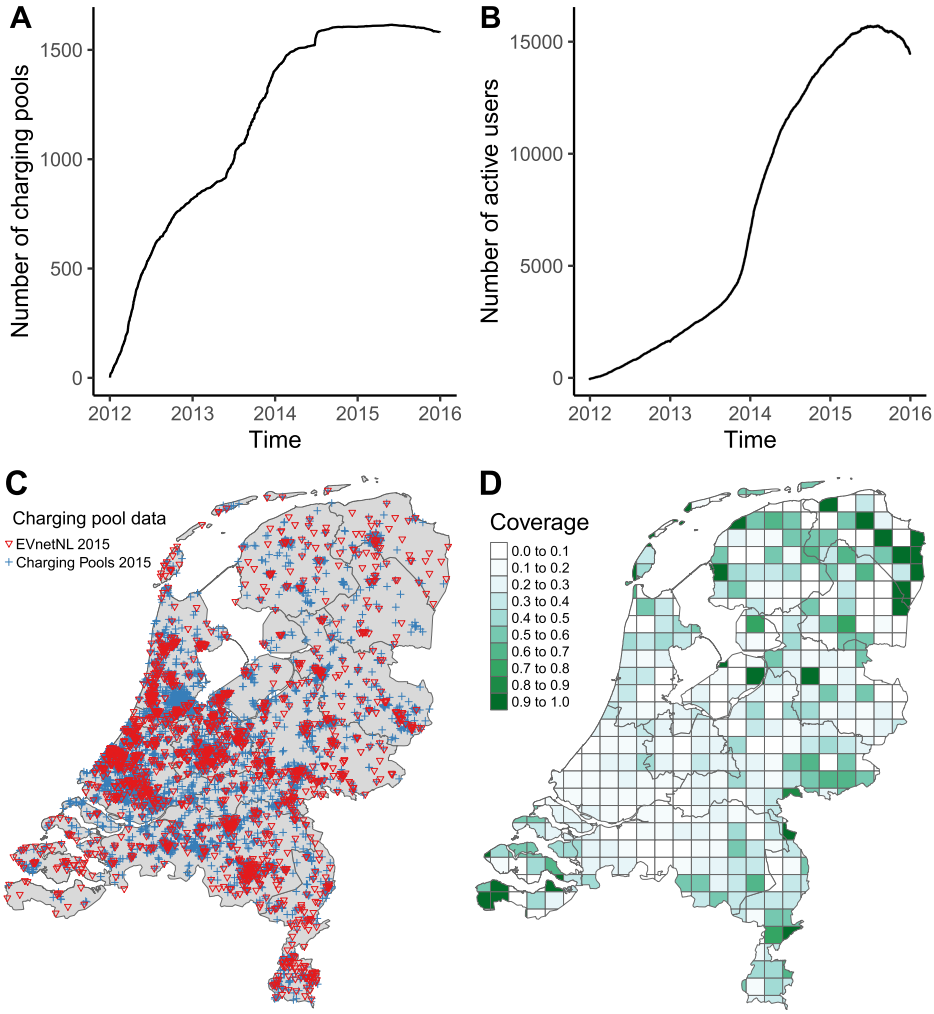

FIGURE 2. **A The number of charging pools in active use as a function of time. To estimate the number of active charging pools, the presence of each charging pool is bounded by the start time of the first charging transaction and the end time of the last charging transaction recorded in the EVnetNL dataset. B The number of active users as a function of time (estimated from the first and the last recorded use of RFID cards). The number of active users is peaking in 2015 as in this period many RFID cards are used only a few times. C A map of the Netherlands showing the geographical locations of EVnetNL charging pools operational in the year 2015 (triangles) together with the charging pools from the dataset Charging pools 2015 (crosses). In the Netherlands, 17 786 slow charging points were operational in 2015, according to [38]. In the Charging pools 2015 dataset, we identified 8 366 unique positions of charging pools. Considering the distribution of charging points at charging pools observed in the EVnetNL dataset, we estimate that the Charging pools 2015 dataset covers about 78.3% of all charging pools. D The spatial representativeness of the EVnetNL dataset estimated by calculating the ratio between the number of stations in the EVnetNL and in the Charging pools 2015 datasets for square cells of a regular grid.**

of charging pools. From this dataset, we selected **45** attributes for further analysis (see Table S2 of the SI file).

2) NEIGHBOURHOODS

The neighbourhoods are spatial units that are approximately uniform when considering the type of built-up area or the socio-economic indicators [40]. The neighbourhoods are used by the Dutch national statistical office to collect, maintain and distribute the statistical data. In addition to the population data, the Neighbourhoods dataset maintains information about the number of address points within the neighbourhoods, level of the income of the residents, number of registered private and company vehicles and so on. From the available attributes, we selected **63** attributes listed in Table S3 of the SI file.

3) ENERGY CONSUMPTION

The Energy consumption dataset contains records of the annual natural gas and electricity consumption in residential houses and industrial facilities, together with the number of buildings equipped with metering devices [41]. The spatial resolution is identical with the Neighbourhoods dataset. For

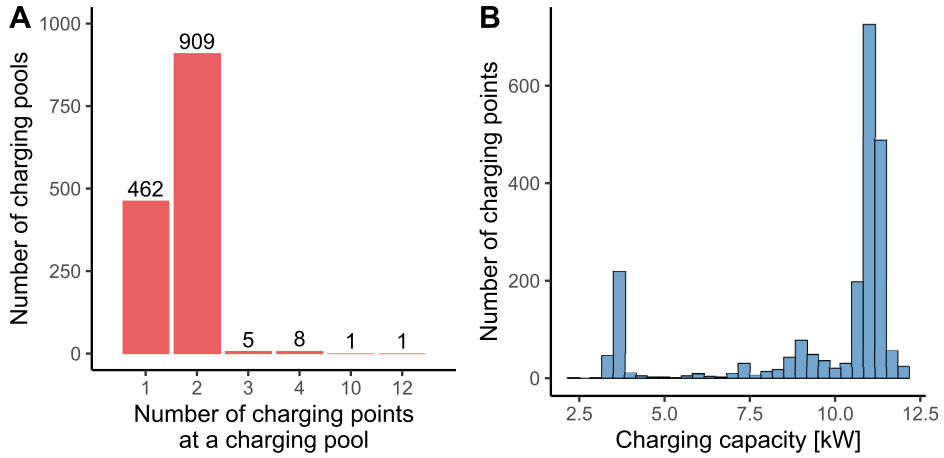

FIGURE 3. A The number of charging pools in use on January 1, 2015 with a given number of charging points in the EVnetNL dataset. B Histogram of charging capacities of charging points estimated from the meter reading values.

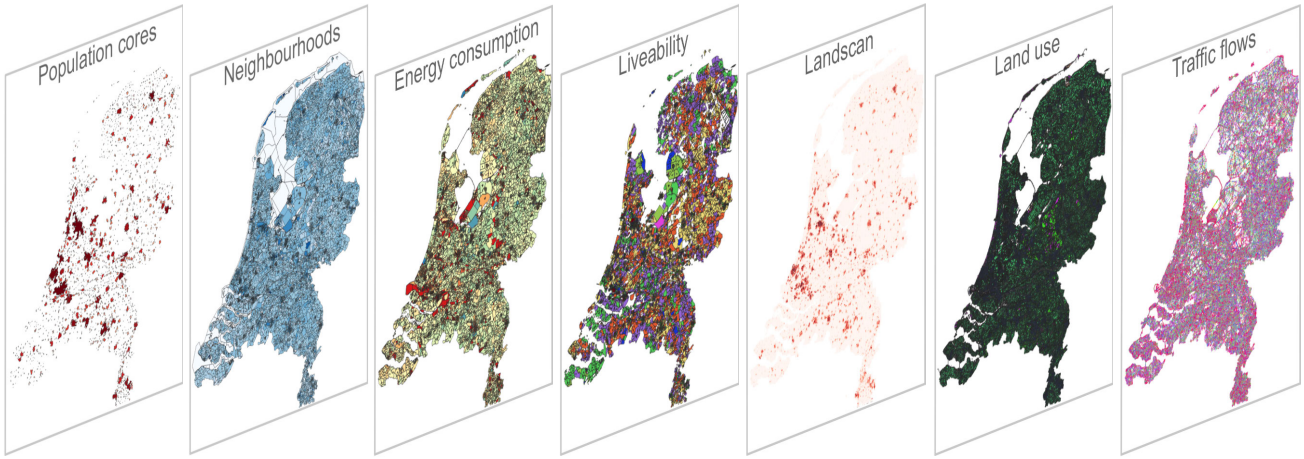

FIGURE 4. Overview of geospatial datasets that have been compiled to characterise the geographical area and human activities in the vicinity of charging pools.

further analyses, we selected **12** attributes that are detailed in Table S5 of the SI file.

4) LIVEABILITY

The Liveability dataset has been introduced by the Dutch Ministry of the Interior and Kingdom Relations to monitor the quality of living in Dutch neighbourhoods [42]. We use the liveability index 2016, that was revised in 2015. The liveability is quantified by a composite index and by five specific indices evaluating categories such as housing, socio-economic background of residents, services, safety, and the environment. Hence, from the liveability dataset, we extracted **6** attributes listed in Table S6 of the SI file.

5) LANDSCAN

We use the LandScan 2015 [43] high resolution population raster grid estimating the 24-hour average of population count with a spatial resolution of approximately 1 km × 1 km. In contrast to the Population cores or the Neighbourhoods datasets that capture the residential population only, the Landscan considers the mobility of residents.

6) LAND USE

The Land use dataset describes the occupation of land in the Netherlands by polygons [44]. Each polygon is assigned an attribute value taking one out of predefined classes of land use. Examples of land use categories are traffic areas,building sites, recreational areas and business areas. Complete list of **25** categories is given in Table S4 of the SI file.

7) TRAFFIC FLOWS

To model the impact of traffic on the usage of charging pools, we consider the traffic flows dataset [45]. The dataset is organised around a high resolution model of the road network and the traffic flow information is added in the form of attributes that are associated with the road segments.

The description of **9** attributes that have been selected to compile features is given in Table S7 of the SI file.

8) OpenStreetMap

The OpenStreetMap (OSM) is one of the most successful free maps [46]. From the OpenStreetMap of the Netherlands, we extracted all points of interest (POIs) considering 2 km $\times$ 2 km squared areas centred at the positions of the EVnetNL charging pools. We identified 593 different POI types, some of them appearing in only very few instances. For this reason, we associated manually POI types with one of the **15** categories listed in Table S9 of the SI file. The POIs, organised in these 15 categories, were used to model the venues in the proximity of charging pools that are often visited by EV drivers.

9) CHARGING POOLS 2015

Aiming at estimating the positions of all available charging pools present in the Netherlands by the end of the year 2015, the Charging pools 2015 dataset was compiled from the EVnetNL, OpenChargeMap [47] and OplaadPalen [48] datasets. Utilising the date when a charging station was added to the dataset, we extracted from the OpenChargeMap and OplaadPalen positions of all charging stations that were available by the end of 2015. As with the EVnetNL dataset, we estimated the position of charging pools from the positions of charging stations, while utilising the information about the geographical proximity. In the first step, we added to the Charging pools 2015 all EVnetNL charging pools available in 2015. In the second step, we added one-by-one to the Charging pools 2015 dataset the charging stations from the OpenChargeMap and OplaadPalen datasets, if their position was more than 50 meters distant from already added pools. This way, we obtained positions of 8 366 charging pools (see Figure 2C).

D. DATA PRE-PROCESSING

To prepare the data for the analyses, we applied the pre-processing procedure composed of three stages: the missing value handling, the extraction of features and the analysis of potential data modelling problems, described in the following subsections.

1) MISSING VALUES HANDLING

Values of some attributes associated with geometric objects in geospatial datasets are missing. By visualising the geometric objects with missing attribute values on the map, we identified two main sources of problems. Some geometric objects with missing data represent water landscape. In such a case, we excluded the geometric objects from datasets. The second source of specific problems with missing data are cities of Baarle-Nassau and Baarle-Hertog with peculiar borders between the Netherlands and Belgium. In this case, we ignored missing values as the EVnetNL dataset does not feature any charging pool in this area. We applied a two-step approach to handle missing values (see section S2 A of the SI file). In the first step, when it was possible, we estimated the missing values from the available data. Further analysis revealed that the attribute values in the geospatial data tend to be missing in areas with low intensity of human activities (e.g. low population, low density of buildings, low electricity consumption, etc.). Therefore, in the second step, we applied simple rules that set the missing values of some attributes to zero (or lowest possible value) in areas with low intensity of human activities. Remaining missing values are addressed after generating features characterising the vicinity of charging pools.

2) PREPARATION OF FEATURES

We modelled each charging pool as a single point defined by GPS coordinates. We used a buffer, circular area centred at the position of charging pool and having the radius r, to model the vicinity of charging pools. Values of features were calculated from GIS polygon data, considering spatial intersections between the area of buffers and GIS geometric objects, while assuming a uniform spatial distribution of considered quantities over the area represented by GIS geometric objects. For POI data (OpenStreetMap and Charging Stations 2015), the features are defined two ways: as the distance from the pool to the closest point of interest and as the density of points of interest within the buffer area. We set the buffer radius to $r = 350$ meters. For more details, please refer to Section S2 B of the SI file.

To ensure that the handling of missing values cannot influence the results of the analysis significantly, we applied to each feature the following rule: The feature is used in the analysis if areas with missing values of the attribute take less than 15% of the buffer areas, otherwise it is excluded. To make sure that features are not built based on GIS data with a large proportion of missing values, if there was less than 1.5% of feature values missing after applying the estimations, missing values were imputed by a median value, otherwise, the feature was discarded. Finally, we obtained 195 features.

3) ANALYSIS OF POTENTIAL DATA MODELLING PROBLEMS

Adopting the notation from [49, p. 5], features are organised in a matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$, where the column $\boldsymbol{x}_j$ corresponds to the feature $j = 1, \ldots, p$ and x_{ij} is the i-th observation of the feature j, for $i = 1, \ldots, n$, where n is the number of observations (charging pools). The response vector $\boldsymbol{y}$ represents the energy consumption in kWh on charging pools during the year 2015, i.e. y_i is the energy consumption of the charging pool i.

When a dataset is fit with a regression model, many problems may occur. Several steps, recommended in the literature [50]–[52], were applied to the feature matrix $\boldsymbol{X}$ and the response vector $\boldsymbol{y}$, to explore and address potential problems. The sequence of steps is illustrated in Figure 5.

First, uninformative features, i.e. those with more than 95% of zero values, were excluded. Dependencies between features can cause serious problems when interpreting results

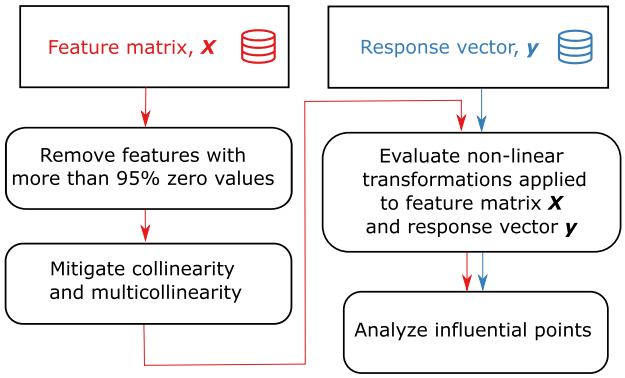

FIGURE 5. Schematic illustrating the workflow applied to the feature matrix X and response vector y to analyse potential data modelling problems.

obtained by regression methods, referred to as collinearity and multicollinearity. Despite the potential to bias the interpretation of results, these problems tend to be overlooked in the data analysis [53]. We identified groups of features with absolute value of the Pearson correlation coefficient between each pair of features in the group greater than 0.95 [51, p. 47]. For each group, we chose one feature as a representative of the group. The selected feature was included, while other group members were excluded from further analysis. Identified groups and selected representative features are listed in Table S11 of the SI file. To mitigate multicollinearity, we applied a procedure composed of two steps: in the first step, the values of the variance inflation factor (VIF) [50] were calculated. In the second step, we identified feature with the maximum value of the VIF. If the maximum VIF was greater than or equal to value 10, the corresponding feature was excluded from further analysis and the procedure was repeated, otherwise, the procedure was terminated. The list of excluded features is provided in Section S2 C of the SI file. These steps reduced the multicollinearity to the level recommended in the literature, while quantified not only by the values of the VIF but also by values of measures derived from eigenvalues of the correlation matrix [54, p. 252].

To explore whether a nonlinear function of $\boldsymbol{X}$ could explain response vector $\boldsymbol{y}$ better than the linear function, we transformed original features $j = 1, \ldots, p$, except binary features, using functions $\sqrt{\boldsymbol{x}_j}$, $\boldsymbol{x}_j^2$ and $\log(\boldsymbol{x}_j + 1)$ and the feature matrix $\boldsymbol{X}$ was extended by transformed features (in this section, functions are applied to vectors element-wise). By combining the ordinary least squares (OLS) with 10-fold cross-validation [50], the extended feature matrix was fit to the response vector $\boldsymbol{y}$ and the mean squared error (MSE) was evaluated on the test data. From the results, we concluded that the basic non-linear transformations do not improve the fit sufficiently to compensate for additional model complexity and for making the model more likely to overfit the data. Hence, we do not apply any non-linear transformation to the feature matrix $\boldsymbol{X}$. Furthermore, we evaluated several transformations of the response vector: $\sqrt{\boldsymbol{y}}$, $\boldsymbol{y}^2$, $\log(\boldsymbol{y})$ and the Box-Cox transformation [51, p. 32]. By analysing the residual plots (see Figure S3 of the SI file) we concluded that the transformation $\log(\boldsymbol{y})$ significantly improves the fit and we apply it in the regression analyses when fitting the energy consumption with the feature matrix, presented in Section III.

Finally, we obtained 1259 observations and 119 features derived from geospatial datasets characterising the vicinity of charging pools and 5 features derived from EVnetNL dataset specifying the location and basic characteristics of charging pools. We have published the feature matrix $\boldsymbol{X}$ with EVnetNL features together with the response vector on the Mendeley Data repository (`http://dx.doi.org/10.17632/kdx92hmgkx`).

E. DATA ANALYSES METHODS

This paper aims at explaining consumed energy on charging pools using features derived from the geospatial datasets. This task can be formulated as a regression problem combined with the statistical inference, indicating the statistical strength of features. The flow chart in Figure 6 illustrates the data analysis procedure. From the geospatial datasets, we extracted a relatively large number of features. To facilitate the interpretation of results, features that have a higher potential to explain the response variable shall be selected. We assume that from all features, only a relatively small number plays an important role. To minimize the uncertainty that incorrect features are selected, we apply the bootstrap method [49]

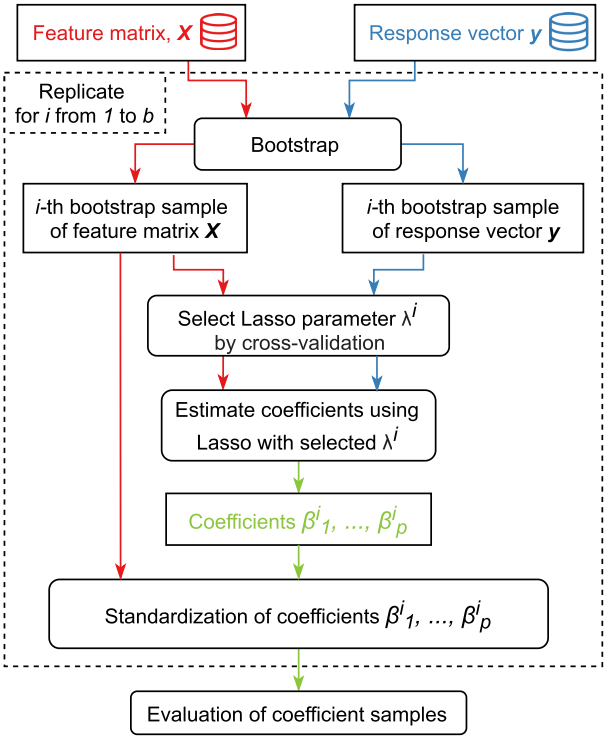

FIGURE 6. A schematic illustrating the workflow followed by the data analysis. The rectangles represent input or output data, and rectangles with rounded corners represent the execution of methods. The items inside the large dashed rectangle are replicated b times via bootstrap.

to create b samples of the dataset. To each bootstrap sample we apply the Lasso method [55], which combines the parameter fitting with the variable selection functionality. The consistency of the selected regression coefficient across bootstrap samples is evaluated, and the conclusions about the significance of features are derived.

1) THE LASSO METHOD

Considering the input data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, for some $\lambda \geq 0$ the Lasso method solves the optimisation problem

$$\underset{\beta_0,\beta}{\text{minimise}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}, \qquad (1)$$

where the scalar β_0 (intercept) and vector β (regression coefficients) are optimisation variables. The first term corresponds to the least squares objective function and its role is to ensure a good fit between the linear regression model $\eta(\boldsymbol{x}_i) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j$ and response value y_i, while the second term regularises the estimated values of regression coefficients in a way that leads to a variable selection (i.e. some regression coefficients are set to zero). The parameter λ determines the trade-off between goodness of the fit and strength of the variable selection. Often, the factor $\frac{1}{2n}$ in (1) is replaced, with $1/2$ or 1. Although this corresponds to a simple reparametrisation of λ, the factor $\frac{1}{2n}$ makes λ values comparable across different sample sizes, which is useful for cross-validation [49]. The solution of problem (1), $\hat{\beta}_0$ and $\hat{\beta}$, constitutes the estimate of the model parameters.

The Lasso method tends to have some difficulties with the identification of relevant features on datasets with highly correlated features [49, p. 55]. The Elastic Net method, i.e. adding the term $\lambda_2 \sum_{j=1}^{p} |\beta_j|^2$ (with $\lambda_2 \geq 0$) to (1), may help to identify correlated features [49, p. 56]. In numerical experiments, the Elastic Net method was tested as well, however, it selected largely the same set of features as the Lasso method.

2) MODEL SELECTION

The parameter λ in (1) controls the complexity of the model. A smaller value of λ results in a larger number of non-zero regression coefficients and allows the model to adapt more closely to data, however, it can lead to overfitting. On the contrary, a larger value of λ leads to a sparser and more interpretable model with the risk of preventing the Lasso from capturing the main signal in the data. Hence, the value of λ should be carefully chosen. To estimate a suitable value of λ, we evaluated a range of values using log spacing. For each value, we used the k-fold cross-validation [49] to evaluate the MSE of the model. Finally, we found the λ^{CV} for which the minimum MSE was achieved and we selected the corresponding $\hat{\beta}_0^{CV}$ and $\hat{\beta}^{CV}$.

3) STATISTICAL INFERENCE

Traditional methods, such as OLS, determine the statistical strength of features by evaluating p-values. The results obtained by the OLS regression on features selected by the Lasso method cannot be fully used in post-selection analysis as the exclusion of some features causes a bias [49, p. 155]. The adaptive nature of the Lasso method makes the problem of estimating p-values difficult–both conceptually and analytically. Reference [49, p. 139] describes three basic statistical inference approaches applicable together with the Lasso method: Bayesian Lasso, non-parametric bootstrap and parametric bootstrap. As we do not know the distribution of regression coefficients and considering the computational complexity of the Bayesian Lasso, we decided to use non-parametric bootstrap method. The bootstrap was recommended as a suitable method for the assessment of the stability of selected regression coefficients in [53], even though there it has not been used in the analysis while taking a risk of presenting an unreliable set of significant coefficients. The bootstrap is a generic tool for assessing the statistical properties of complex estimators. First, the dataset is sampled. Second, the k-fold cross-validation is applied to each sample, to find λ^{CV}, $\hat{\beta}_0^{CV}$ and $\hat{\beta}^{CV}$. The frequency by which the regression coefficients take the value of zero in bootstrap samples captures the uncertainty regarding the selection of the corresponding feature [49, p. 153]. A small frequency of zero values or their absence increases the confidence that a feature should be selected. The consistency of the coefficients' sign across the bootstrap samples can be used as a measure of uncertainty regarding the interpretation of the corresponding feature. The higher is the occurrence of positive (negative) signs of regression coefficients, the higher is the evidence for the positive (negative) correlation between the feature and the response variable. Thus, in the numerical experiments, we evaluate the frequency of zero values of regression coefficients and their probability distributions.

III. RESULTS

A. SOFTWARE LIBRARIES AND SETTINGS

We prepared and modelled the data using the R language. We processed the GIS data with *sf*, *raster* and *osmar* packages. The distribution of energy consumption was fitted with the *fitdistrplus* package. The Lasso method, including model selection, is implemented in the *glmnet* package. For the k-fold cross-validation we used $k = 10$. We considered the values 10^i, for i in the interval from -4 to 0 in steps of 0.02, when applying the cross-validation to explore the values of the parameter λ in the Lasso method. When studying the stability of coefficients selected by the Lasso method, we used the bootstrap method with $b = 10\,000$ realisations.

B. METRICS OF CONSUMED ENERGY AT CHARGING POOLS

Charging pools differ in the maximum capacity and in the number of charging points (see panels A and B of Figure 3), potentially leading to differences in the consumed energy. As illustrated in Figure 7A, charging pools with a higher number of charging points tend to have slightly higher energy

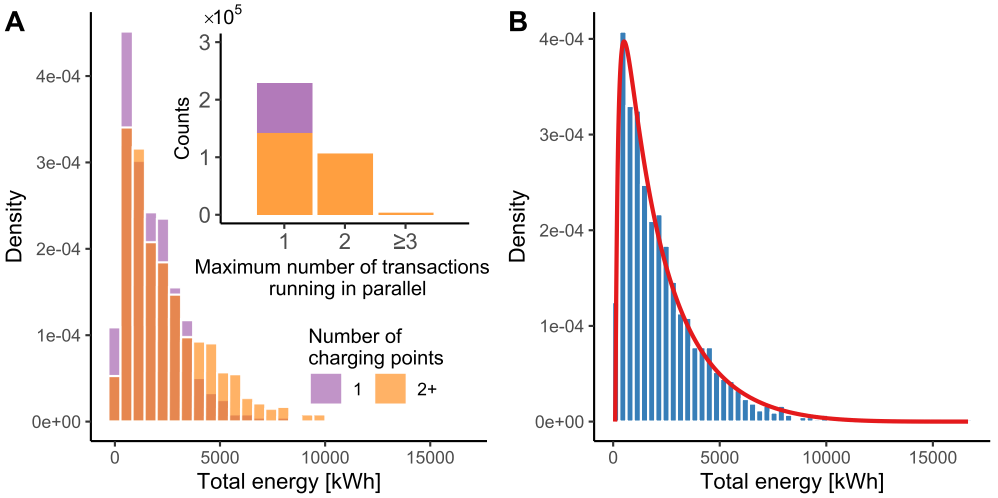

FIGURE 7. The empirical probability distribution of energy consumed at charging pools. A We show two separate probability distributions, one for charging pools with only one charging point and one for charging pools with more than one charging point. The inset shows the stacked bar plot of the number of EVs charged in parallel with other EVs on a charging pool. B The empirical probability distribution of the total energy consumed at charging pools. The red line represents the fit obtained using Eq. (2).

TABLE 1. Response vectors characterising the energy consumed at a charging pool. The coefficient of determination, R^2, was obtained by fitting the logarithmic transformation of the response vector (see Figure S3 of the SI file) with the feature matrix by applying the ordinary least squares method.

Response variable characterising a charging pool	R^2
Consumed energy	0.380
Consumed energy per charging point	0.367
Maximum energy consumed at a charging point	0.360
Consumed energy per one unit of charging capacity	0.333

consumption. Moreover, the number of cases when two transactions run on a charging pool in parallel is not negligible (see inset of Figure 7A). To analyse to what extent it is important to account for these effects, we define several simple metrics of the consumed energy at charging pools in Table 1. To gain some basic understanding, whether some measures can be better explained with the feature matrix, we ran the OLS regression. In Table 1 we report the values of the coefficient of determination, R^2. The highest R^2 we obtained for the consumed energy at a charging pool. High similarity in R^2 values we attribute to the high correlations between response vectors. The Pearson correlation coefficient for all pairs of response vectors ranges from 0.86 to 0.98. Hence, for further analyses, we chose the consumed energy at a charging pool as the response vector $\boldsymbol{y}$.

C. DISTRIBUTION OF CONSUMED ENERGY OVER CHARGING POOLS

The total electric energy consumption is 2.84 GWh. To introduce a simple model describing how is the consumption distributed over charging pools, we investigate the probability distribution. The density histogram indicates that the distribution is rather heterogeneous and positively skewed (see Figure 7B). We observe high consumption on a small number of charging pools and small consumption is observed for a large group of charging pools. To model the distribution of energy consumed at charging pools, we considered original data and five transformations ($\boldsymbol{y}^2$, $\boldsymbol{y}^3$, $\sqrt{\boldsymbol{y}}$, $\sqrt[3]{\boldsymbol{y}}$ and $log(\boldsymbol{y})$) and parametrised density functions of three known probabilistic distributions (Weibull, beta and gamma). We used the Kolmogorov-Smirnov goodness of fit test together with the inspection of the P-P and the Q-Q plots [56] to conclude that the transformation $\sqrt[3]{\boldsymbol{y}}$ combined with the beta distribution provides the best fit to the data. The fitting procedure is detailed in Section S3 of the SI file. The functional form of the probability density function is derived in Appendix and takes the form

$$f(y, \alpha, \beta) = \frac{\left(\frac{\sqrt[3]{y}-y_{min}}{y_{max}-y_{min}}\right)^{\alpha-1}\left(1-\frac{\sqrt[3]{y}-y_{min}}{y_{max}-y_{min}}\right)^{\beta-1}}{B(\alpha,\beta)2(y_{max}-y_{min})y^{\frac{2}{3}}}. \quad (2)$$

The symbol $B(\alpha, \beta)$ denotes the beta function and the estimates of parameters take the following values: $\alpha = 2.58$, $\beta = 4.53$, $y_{min} = 91.55$ kWh and $y_{max} = 16\,649.40$ kWh.

D. EXPLAINING THE ENERGY CONSUMPTION FROM OTHER CHARGING POOL PERFORMANCE INDICATORS

We can gain interesting insights by analysing the relationship between energy consumption and other charging pool indicators constituting the energy consumption. The energy consumption at a charging pool i can be decomposed into the product of three other indicators, i.e.

$$y_i = n_i t_i p_i, \quad (3)$$

where n_i is the number of charging transactions taking place at a charging pool i, t_i is the average charging time per transaction at a charging pool i and p_i is the average charging power

at a charging pool i. We organised these quantities for all charging pools as vectors $\boldsymbol{n}, \boldsymbol{t}$ and $\boldsymbol{p}$. To assess the role of these three factors, in the heterogeneity of the consumed energy across charging pools, we explored six models presented in Table 2. Models are based on Eq. (3), where one indicator or product of two indicators, represented by the regression coefficient k, is considered invariant across charging pools.

TABLE 2. Simple regression models of the energy consumed at charging pools. In columns are presented the estimates of the regression coefficient $\hat{k}$; the corresponding R^2 value obtained by the OLS regression; the mean value of the quantity that is represented by the regression coefficient k calculated over charging pools (mean); the corresponding standard deviation (stdev) and the coefficient of variation (cv) calculated from the mean and the standard deviation. The symbol $\circ$ denotes the component-wise multiplication (the Hadamard product) of vectors.

Model	$\hat{k}$	R^2	mean	stdev	cv
$\boldsymbol{y} = k\boldsymbol{n}$	8.10	0.90	8.69	3.65	0.42
$\boldsymbol{y} = k\boldsymbol{t}$	895.55	0.59	897.02	764.74	0.85
$\boldsymbol{y} = k\boldsymbol{p}$	632.67	0.56	678.46	589.20	0.87
$\boldsymbol{y} = k(\boldsymbol{t} \circ \boldsymbol{p})$	245.25	0.59	269.05	228.06	0.85
$\boldsymbol{y} = k(\boldsymbol{n} \circ \boldsymbol{p})$	2.49	0.95	2.52	0.63	0.25
$\boldsymbol{y} = k(\boldsymbol{n} \circ \boldsymbol{t})$	3.25	0.94	3.45	0.92	0.27

We obtained the estimate $\hat{k}$ of the coefficient k from the data by using the OLS method. Among the models that explain the consumed energy from one indicator, the highest value of R^2 is obtained for the model explaining the consumed energy from the number of charging transactions. Similarly, models explaining energy consumption from pairs of indicators that include the number of transactions have high values of R^2. Hence, the major factor associated with the heterogeneity in the consumed energy across charging pools is the number of transactions. The fluctuations in the charging patterns (i.e. average charging time and average charging power) play a much smaller role.

In Table 2, we calculated, mean, standard deviation and the coefficient of variation over all charging pools for quantities which are represented by the coefficient k. For example, in the model $\boldsymbol{y} = k\boldsymbol{n}$, the coefficient k that replaces in Eq. (3) the expression $t_i p_i$ can be interpreted as the average of consumed energy per transaction. Hence, the average consumed energy per transaction is 8.69 kWh, the charging time per transaction takes on average 2.52 hours and the average power reaches 3.45 kW. These numbers indicate that the majority of charged EVs are plug-in hybrids. The charging capacity, which is for the majority of charging points around 11 kW (see Figure 3), is significantly underused. The coefficient of variation values is high when the number of transactions is included in the analysed quantity, confirming that the largest variance is associated with the number of transactions.

As the number of charging transactions is closely associated with the consumed energy at charging pools, it is crucial to understand the way EV drivers decide which charging opportunity they choose. We hypothesise that some exogenous factors, characterising the environment surrounding the charging pools affect the consumed energy at charging pools, and we explore it in the next section.

E. EXPLAINING THE ENERGY CONSUMPTION FROM GEOSPATIAL DATA

We applied the methodology described in Section II-E to the feature matrix derived from geospatial datasets, first without considering the features derived from the EVnetNL dataset (see Section S2 B), and to the *log*-transformation of the response vector $\boldsymbol{y}$, representing the consumed energy at charging pools. To facilitate the mutual comparison of regression coefficients, we standardised each coefficient by multiplying it with the standard deviation of the bootstrap sample of the feature. Thus, the standardised coefficient can be considered as an estimate of the change in the response variable, when the feature increases by one standard deviation. The larger is the absolute value of the median of standardised regression coefficient samples, the stronger is the feature's potential to influence the response variable [57, p. 372]. Hence, the absolute value of the median of bootstrap realisations can be considered as a measure of the feature strengths. The different signs of regression coefficients, across bootstrap realisations, can be attributed to the low significance or to the simultaneous selection of correlated features [49, p. 144]. Hence, the more consistent are the values of standardised regression coefficient across bootstrap samples, the more significant is the feature corresponding to a regression coefficient. As a rule of thumb, we consider as significant those features where the number of bootstrap samples with zero coefficient value is less than 5%, and the number of samples with the opposite coefficient sign to the sign of the majority of the sample is negligible. To provide a broader view on results, we show in Figure 8 the empirical distributions of standardised regression coefficients that reached the value of zero in maximum 10% of the bootstrap realisations. The statistical inference is sometimes omitted in energy studies, and only a single run of a variable selection method is evaluated [53], [58]. The inspection of Tukey's boxplots justifies the use of the bootstrap or in more general, the use of a statistical inference method. The majority of presented coefficients obtained a zero value in few samples or exceptionally the opposite sign, that could lead to omitting these coefficients or interpreting them in a wrong way if evaluating only a single run of the Lasso method.

We show the regression coefficients in descending order of median value in Figure 8. To clarify the results, we organised significant features with the positive sign of the median in four groups:

- **Physical environment(+):** *terrain for social and cultural facilities* (LC_7); *open wet natural terrain and water* (LC_{24}); *density of roads*; *parkland* (LC_{15}); *terrain for public facilities* (LC_6).
- **Population(+):** *employed residents working in the wholesale, ICT and finance sector [%]* (PC_{32}); *employed residents working in non-commercial sector [%]* (PC_{33}); *unregistered couples with children* (PC_{16}); *Morocco immigrants [%]* (N_{22}); *employed residents working in power, waterworks, horeca*

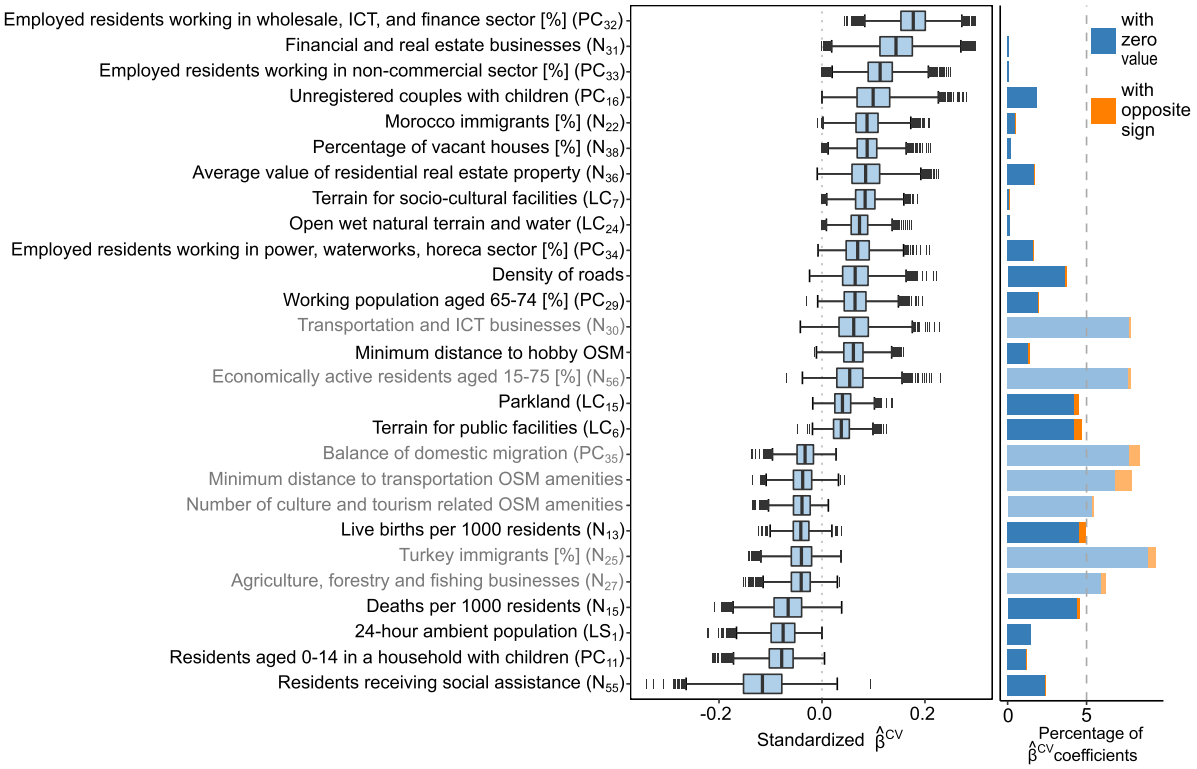

FIGURE 8. The empirical distributions of standardised regression coefficients obtained by the Lasso method and the 10-fold cross-validation applied to 10 000 samples of bootstrapped data. We show only features where the value of the regression coefficient was set to zero in less than 10% of the samples. Coefficients are descendingly ordered, from the largest to the smallest median value. The left panel presents the Tukey's box plot of coefficients. On the right, the stacked bar plot shows the percentage of samples when the regression coefficient $\hat{\beta}^{CV}$ was set to zero and the number of samples where it reached the opposite sign as the sign of the median. We consider as significant those features where the number of samples with zero coefficient value is less than 5% and the number of samples with opposite sign is small. The dashed line indicates the 5% threshold value. Features that are not considered significant we display using faded colours. Full descriptions of coefficients can be found in tables S2 - S8 of the SI file by using the code in brackets.

(hotel/ restaurant/café) sector [%] (PC_{34}); *working population aged 65 - 74 [%]* (PC_{29}).

- **Services and businesses(+):** *financial and real estate businesses* (N_{31}); *minimum distance to hobby OSM*.
- **Buildings(+):** *percentage of vacant houses* [%] (N_{38}); *average value of residential real estate property* (N_{36}).

Similarly, we organised significant features with the negative value of the median into a group:

- **Population(−):** *residents receiving social assistance* (N_{55}); *residents aged 0 - 14 living in a household with children* (PC_{11}); *24-hour ambient population* (LS_1); *deaths per 1000 residents* (N_{15}); *live births per 1000 residents* (N_{13}).

The largest number of significant features we found in the population group, which is partly because this group contains most of the features. Many significant features point to a single factor. The largest group of features, PC_{32}, PC_{29}, N_{36}, (N_{55}), indicate that high (low) income and wealth (poverty) are positively (negatively) linked with the amount of consumed energy at charging pools. Most likely, this is due to the high prices of EVs, making them more affordable for better-situated residents and businesses. Probably for the same reasons, some significant features (PC_{11}, N_{13}) are linked with children or youth and to the elderly or retired population (N_{15}), i.e. social groups that are typically less wealthy. The high *percentage of vacant houses* (N_{38}) and the high *average value of residential real estate property* (N_{36}) are associated with high energy consumption at charging pools. It could correspond to newly built and yet not entirely inhabited areas, with a higher standard of living expressed in higher real estate values. Note, if a feature representing a minimum distance to an object has a positive coefficient, then the energy consumption increases with increasing the distance from a given object and vice versa. Hence, the proximity of *hobby related points of interest* is negatively associated with the energy consumption at charging pools. Furthermore, the *density of roads* is also among features that are positively linked with the energy consumption at charging pools, indicating that good access to charging pools contributes to energy consumption.

We included to the feature matrix some features (*number of charging points*, *maximum power*, *latitude* and *longitude* and the *rollout strategy*) derived from the EVnetNL dataset (see Section S2 B) and repeated the analysis. The results are shown in Figure S7 of the SI file. In general, significant features are similar as in Figure 8; however, the number

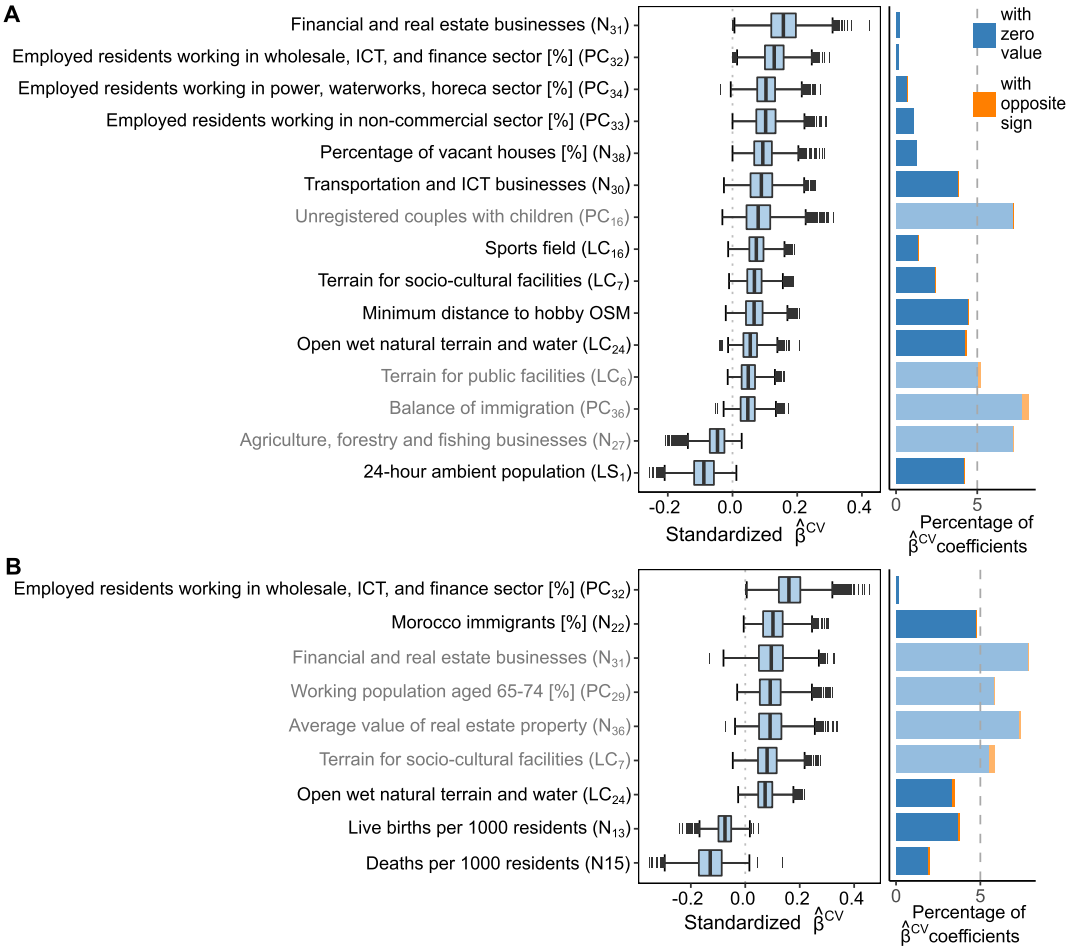

FIGURE 9. **The empirical distributions of standardised regression coefficients obtained by the Lasso method combined with the 10-fold cross-validation applied to 10 000 samples of bootstrapped data. A Strategic rollout of charging pools. B Demand-driven rollout of charging pools. We show only features where the value of the regression coefficient was set to zero in less than 10% of the samples. Regression coefficients are descendingly ordered from the largest to the smallest median value. The left panel presents the Tukey's box plot of coefficients. On the right, the stacked bar plot shows the percentage of samples where the regression coefficient $\hat{\beta}^{CV}$ was set to zero and the number of samples where it reached the opposite sign as the sign of the median. We consider as significant those features where the number of samples with zero coefficient value is less than 5% and the number of samples with opposite sign is small. The dashed line indicates the 5% threshold value. Features that are not considered significant we display using faded colours. Full descriptions of coefficients can be found in tables S2 - S8 of the SI file by using the code in brackets.**

of significant features is slightly smaller. The significance of some features from groups Physical environment(+) and Population(−) was reduced. All EVnetNL features are significant, while the *maximum charging capacity*, the *number of charging points* and *longitude* have high strength, indicating that parameters of charging pools potentially influence the energy consumption. We attribute the reduction in the number of significant features to the replacement of some features by EVnetNL features. For example, previously significant feature, *open wet natural terrain and water* (LC_{24}), seems to be expressed via the *longitude*. The negative influence of *longitude* can be explained by the geography of the Netherlands, whereas the western part of the country is more urbanised and we find here the majority of large Dutch cities. At the same time, there is a lot of surface water as the western part of the country is largely situated below the sea level.

F. INFLUENCE OF THE ROLLOUT STRATEGY ON THE ENERGY CONSUMPTION

The majority of charging pools has been located using one out of two (strategic or demand-driven) rollout strategies [32]. The strategically located charging pools are placed near public facilities, where the EV charging is intuitively expected. The demand-driven charging pools are built upon the request from EV users, typically near to their homes. In this section, we investigate whether the rollout strategy makes a difference in factors associated with energy consumption. Information about the rollout strategy is available in the EVnetNL dataset, and we used it to split charging pools into two groups. We applied the Lasso method separately to each group considering the feature matrix without the features derived from the EVnetNL dataset. The selected features in Figure 9 coincide to a large extent with factors selected

for the complete dataset (see Figure 8), however, now we can observe differentiation of factors according to the location strategy.

The energy consumed at strategically located charging pools, (Figure 9A), is positively linked to the working sector of residents and the physical environment, i.e. to certain types of venues adjacent to the charging pool. Working sectors of residents represented by PC_{32}, PC_{33} and PC_{34} indicate the prevalent businesses in municipalities, positively associated with energy consumption. Moreover, selected features for strategical rollout refer to some businesses and some locations (*sports fields*, *socio-cultural facilities*) that could be associated with occasional charging.

For the charging pools with the demand-driven rollout, (Figure 9B), the negative coefficients of *deaths per 1000 residents* and *live births per 1000 residents* indicate that areas with higher natality and mortality are negatively linked with the energy consumption, pointing out to areas inhabited by socially weaker groups of specific age categories.

We have tested several other stratifications, e.g. based on the *number of charging points*, the proportion of the residential area within the charging pool's buffer, the administrative division of the Netherlands into provinces and the number of residents of municipalities. Except for the last criterion, we obtained only a very small number of selected features. Dividing charging pools into two groups based on the municipality population, considering a threshold of 50 000 residents, we obtained approximately the same size of groups. Interestingly, we observe that charging pools located in municipalities with more than 50 000 residents consume more energy on average by 48% than charging pools in the other group of municipalities. We find the higher number of significant features, linked to municipality population characteristics, *financial and real estate businesses* and the physical environment, for the charging pools located in municipalities with a smaller population (see Figure S8 in the SI file).

IV. CONCLUSIONS AND DISCUSSION

We analysed the explainability of consumed energy at charging pools from several points of view. Main conclusions derived from the data analysis are the following:

- The energy consumption can be satisfactorily modelled by a transformed beta distribution.
- The number of charging transactions is the driving factor among the characteristics constituting energy consumption.
- The economic prosperity appears to be behind a large group of regression variables selected for the mathematical description of the relationship between energy consumption and locational factors derived from available geospatial datasets. For example, residents and businesses with high (low) income, situated in the charging pool vicinity, are linked to a positive (negative) impact on energy consumption. Similarly, charging pools located close to expensive newly built housing show higher energy consumption. The western part of the Netherlands with four major large cities is positively linked to energy consumption as well. Considering the standardised values of regression coefficients, certain working sectors of municipalities' residents and the *number of financial and real estate businesses* have a large positive impact on energy consumption. The largest adverse impact have *residents receiving social assistance*.
- The stratification of charging pools by the rollout strategy leads to the split of selected regression coefficients. Business types, working sector of residents and public venues in the proximity are linked to higher consumption of energy at charging pools deployed strategically. Population characteristics, e.g. *live births* and *deaths per 1000 residents* are linked to the energy consumption at charging pools placed based on the demand.

Our results extend the knowledge base about the energy consumption at charging pools and provide the advice which location features to focus on when building a predictive model. Data collection and data processing are among the most time-consuming activities. Hence, these results can be highly beneficial. Furthermore, this paper is opening several possibilities for future applications.

The methodology and our findings can be used to fine-tune rollout strategies for deployment of charging pools. A rollout strategy optimised for a specific group of charging pools, e.g. charging pools used for work-charging, can be designed by applying the presented methodology to this specific group of charging pools and identified characteristics can be used to select locations of new charging pools in a way which corresponds better to energy constraints. Information on selected regression variables can be used to build prediction models in a more targeted way. The presented results are applicable to the Netherlands, however, the proposed methodology can also be used elsewhere. It is likely that utilising the knowledge on location features that are relevant for charging behaviour observed in the Netherlands elsewhere, would be more efficient than selecting features randomly or based on the intuition.

The selected regression coefficients can be associated with three different spatial scales. Some describe close vicinity of charging pools, e.g. *the number of financial and real estate businesses* (N_{31}), others are attributed to the municipality, e.g. features derived from the Population cores datasets such as the *percentage of employed residents in the municipality working in a non-commercial sector* (PC_{33}). The last group of regression coefficients has the potential to characterise the location of a charging pool at the country level, e.g. *latitude* and *longitude* (see Figure S7 in the SI file). Hence, while properly considering the spatial level of selected regression coefficients, a hierarchical or a customised rollout strategy could be designed considering specific geographic scales.

Apart from enhancing the strategies for charging pools deployment, our study can help to improve energy demand models for power grid capacity planning by better considering the energy demand at charging pools and the proposed methodology could be also applied to other service systems, e.g. to stations for shared electric cars, scooters or bicycles.

A. LIMITATIONS AND FURTHER RESEARCH

Several important limitations are inherited from the used statistical methods. Apart from notorious limitation ''correlation does not imply causation'', which requires careful consideration of all results, we wish to point out limitation to our results' interpretability due to the presence of multicollinearity in the input data. To obtain statistically stable results, we reduced the level of multicollinearity to the level recommended in the literature. However, the removal of some features hampers the interpretation of our results. Similarly, it is likely that some relevant data representing important determinants of energy consumption are missing in the analyses, e.g. mobility behaviour of the population or visitation patterns of venues located in the vicinity of charging pools, as we were not able to collect them. Further research could focus on a more complex characterisation of the usage of charging pools, specific groups of charging pools, e.g. determined based on similarities in usage patterns, or exploration of possibilities to extend the proposed methodology and findings to other geographic areas.

APPENDIX. FITTING THE ENERGY CONSUMPTION WITH THE TRANSFORMED BETA PROBABILITY DENSITY FUNCTION

Considering the Kolmogorov-Smirnov goodness-of-fit test together with the inspection of the P-P and the Q-Q plots [56], we concluded that the most satisfactory fit of the consumed energy at charging pools is obtained by transforming the data with the function $g(\boldsymbol{y}) = \sqrt[3]{\boldsymbol{y}}$ (the function is applied to the vector element-wise) and using the beta distribution (see section S3 of the SI file). The beta distribution is defined on the interval $\langle 0, 1\rangle$. After the rescaling, the beta distribution is suitable to represent a random variable between a minimum value y_{min} and a maximum value y_{max}. Hence, with the beta distribution, we model the random variable

$$Z = \frac{\sqrt[3]{Y} - y_{min}}{y_{max} - y_{min}}, \tag{4}$$

where the random variable Y is modelling the energy consumption. Using Eq. (4), we establish the relation between the distribution function of Y and Z as

$$\begin{aligned} F_Y(y) &= P(Y < y) \\ &= P((Z(y_{max} - y_{min}) + y_{min})^3 < y) \\ &= P\Big(Z < \frac{\sqrt[3]{y} - y_{min}}{y_{max} - y_{min}}\Big) = F_Z\Big(\frac{\sqrt[3]{y} - y_{min}}{y_{max} - y_{min}}\Big). \end{aligned} \tag{5}$$

Consequently, the density function characterizing the random variable Y is

$$\begin{aligned} f_Y(y) &= \big[F_Y(y)\big]' = \left[F_Z\Big(\frac{\sqrt[3]{y} - y_{min}}{y_{max} - y_{min}}\Big)\right]' \\ &= F_Z'\Big(\frac{\sqrt[3]{y} - y_{min}}{y_{max} - y_{min}}\Big)\left[\frac{\sqrt[3]{y} - y_{min}}{y_{max} - y_{min}}\right]' \\ &= f_Z\Big(\frac{\sqrt[3]{y} - y_{min}}{y_{max} - y_{min}}\Big)\frac{1}{3(y_{max} - y_{min})y^{\frac{2}{3}}}. \end{aligned} \tag{6}$$

Since $f_Z(z)$ is the probability density function of the beta distribution, we get the following density function for the energy consumption

$$f_Y(y, \alpha, \beta) = \frac{\Big(\frac{\sqrt[3]{y} - y_{min}}{y_{max} - y_{min}}\Big)^{\alpha-1}\Big(1 - \frac{\sqrt[3]{y} - y_{min}}{y_{max} - y_{min}}\Big)^{\beta-1}}{B(\alpha, \beta)3(y_{max} - y_{min})y^{\frac{2}{3}}}, \tag{7}$$

where $B(\alpha, \beta)$ is the beta function with parameters α and β.

REFERENCES

[1] EC. (2019). *2030 Climate & Energy Framework*. Accessed: Aug. 11, 2019. [Online]. Available: https://ec.europa.eu/clima/policies/strategies/2030_en

[2] M. Alonso, B. Ciuffo, F. Ardente, J.-P. Aurambout, G. Baldini, R. Braun, P. Christidis, A. Christodoulou, A. Duboz, S. Felici, J. Rey, A. Georgakaki, K. Gkoumas, M. Grosso, M. Portela, A. Julea, J. Krause, B. Martens, F. Mathieux, and I. Vandecasteele, ''The future of road transport: Implications of automated, connected, low-carbon and shared mobility,'' Joint Res. Centre (Eur. Commission), Ispra, Italy, Tech. Rep. 29748 EN, Jun. 2019. [Online]. Available: https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/future-road-transport

[3] P. Cazzola, M. Gorner, J. Tattini, R. Schuitmaker, S. Scheffer, L. D'Amore, H. Signollet, L. Paoli, J. Teter, and T. Bunsen, *Global EV Outlook 2019: Scaling Up Transition to Electric Mobility*. Paris, France: International Energy Agency, Jun. 2019, doi: 10.1787/35fb60bd-en.

[4] M. Shepero and J. Munkhammar, ''Spatial Markov chain model for electric vehicle charging in cities using geographical information system (GIS) data,'' *Appl. Energy*, vol. 231, pp. 1089–1099, Dec. 2018.

[5] C. C. Chan, ''The rise & fall of electric vehicles in 1828–1930: Lessons learned [scanning our past],'' *Proc. IEEE*, vol. 101, no. 1, pp. 206–212, Jan. 2013.

[6] D. Pevec, J. Babic, and V. Podobnik, ''Electric vehicles: A data science perspective review,'' *Electronics*, vol. 8, no. 10, p. 1190, Oct. 2019.

[7] S. Shahriar, A. R. Al-Ali, A. H. Osman, S. Dhou, and M. Nijim, ''Machine learning approaches for EV charging behavior: A review,'' *IEEE Access*, vol. 8, pp. 168980–168993, 2020.

[8] Y. Amara-Ouali, Y. Goude, P. Massart, J.-M. Poggi, and H. Yan, ''A review of electric vehicle load open data and models,'' Univ. Paris-Saclay, Orsay, France, Working Paper hal-03028375f, Nov. 2020. [Online]. Available: https://hal.archives-ouvertes.fr/

[9] M. Majidpour, C. Qiu, P. Chu, R. Gadh, and H. R. Pota, ''Fast prediction for sparse time series: Demand forecast of EV charging stations for cell phone applications,'' *IEEE Trans. Ind. Informat.*, vol. 11, no. 1, pp. 242–250, Feb. 2015.

[10] O. Frendo, N. Gaertner, and H. Stuckenschmidt, ''Improving smart charging prioritization by predicting electric vehicle departure time,'' *IEEE Trans. Intell. Transp. Syst.*, early access, Apr. 30, 2020, doi: 10.1109/TITS.2020.2988648.

[11] O. Frendo, J. Graf, N. Gaertner, and H. Stuckenschmidt, ''Data-driven smart charging for heterogeneous electric vehicle fleets,'' *Energy AI*, vol. 1, Aug. 2020, Art. no. 100007.

[12] Y.-W. Chung, B. Khaki, T. Li, C. Chu, and R. Gadh, ''Ensemble machine learning-based algorithm for electric vehicle user behavior prediction,'' *Appl. Energy*, vol. 254, Nov. 2019, Art. no. 113732.

[13] S. Venticinque and S. Nacchia, ''Learning and prediction of E-car charging requirements for flexible loads shifting,'' in *Internet and Distributed Computing Systems*, R. Montella, A. Ciaramella, G. Fortino, A. Guerrieri, and A. Liotta, Eds. Cham, Switzerland: Springer, 2019, pp. 284–293.

[14] J. R. Helmus, M. H. Lees, and R. van den Hoed, "A data driven typology of electric vehicle user types and charging sessions," *Transp. Res. C, Emerg. Technol.*, vol. 115, Jun. 2020, Art. no. 102637.
[15] B. Khaki, Y.-W. Chung, C. Chu, and R. Gadh, "Nonparametric user behavior prediction for distributed EV charging scheduling," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, Aug. 2018, pp. 1–5.
[16] S. Wang, L. Du, J. Ye, and D. Zhao, "A deep generative model for non-intrusive identification of EV charging profiles," *IEEE Trans. Smart Grid*, vol. 11, no. 6, pp. 4916–4927, Nov. 2020.
[17] L. Buzna, P. De Falco, G. Ferruzzi, S. Khormali, D. Proto, N. Refa, M. Straka, and G. van der Poel, "An ensemble methodology for hierarchical probabilistic electric vehicle load forecasting at regular charging stations," *Appl. Energy*, vol. 283, Feb. 2021, Art. no. 116337.
[18] Y. Zhang, M. Zhong, N. Geng, and Y. Jiang, "Forecasting electric vehicles sales with univariate and multivariate time series models: The case of China," *PLoS ONE*, vol. 12, no. 5, pp. 1–15, May 2017.
[19] F. Zhao, P. Li, Y. Li, and Y. Li, "The li-ion battery state of charge prediction of electric vehicle using deep neural network," in *Proc. Chin. Control Decis. Conf. (CCDC)*, Jun. 2019, pp. 773–777.
[20] K. A. Severson, P. M. Attia, N. Jin, N. Perkins, B. Jiang, Z. Yang, M. H. Chen, M. Aykol, P. K. Herring, D. Fraggedakis, M. Z. Bazant, S. J. Harris, W. C. Chueh, and R. D. Braatz, "Data-driven prediction of battery cycle life before capacity degradation," *Nature Energy*, vol. 4, no. 5, pp. 383–391, May 2019.
[21] L. Buzna, P. De Falco, S. Khormali, D. Proto, and M. Straka, "Electric vehicle load forecasting: A comparison between time series and machine learning approaches," in *Proc. 1st Int. Conf. Energy Transition Medit. Area (SyNERGY MED)*, May 2019, pp. 1–5.
[22] A. Verma, A. Asadi, K. Yang, A. Maitra, and H. Asgeirsson, "Analyzing household charging patterns of plug-in electric vehicles (PEVs): A data mining approach," *Comput. Ind. Eng.*, vol. 128, pp. 964–973, Feb. 2019.
[23] D. Pevec, J. Babic, M. A. Kayser, A. Carvalho, Y. Ghiassi-Farrokhfal, and V. Podobnik, "A data-driven statistical approach for extending electric vehicle charging infrastructure," *Int. J. Energy Res.*, vol. 42, no. 9, pp. 3102–3120, Jul. 2018.
[24] M. Brenna, G. C. Lazaroiu, M. Roscia, and S. Saadatmandi, "Dynamic model for the EV's charging infrastructure planning through finite element method," *IEEE Access*, vol. 8, pp. 102399–102408, 2020.
[25] Y. Yang, Y. Zhang, and X. Meng, "A data-driven approach for optimizing the EV charging stations network," *IEEE Access*, vol. 8, pp. 118572–118592, 2020.
[26] M. Z. Zeb, K. Imran, A. Khattak, A. K. Janjua, A. Pal, M. Nadeem, J. Zhang, and S. Khan, "Optimal placement of electric vehicle charging stations in the active distribution network," *IEEE Access*, vol. 8, pp. 68124–68134, 2020.
[27] C. Develder, N. Sadeghianpourhamami, M. Strobbe, and N. Refa, "Quantifying flexibility in EV charging as DR potential: Analysis of two real-world data sets," in *Proc. IEEE Int. Conf. Smart Grid Commun. (SmartGridComm)*, Nov. 2016, pp. 600–605.
[28] C. Bikcora, N. Refa, L. Verheijen, and S. Weiland, "Prediction of availability and charging rate at charging stations for electric vehicles," in *Proc. Int. Conf. Probabilistic Methods Appl. to Power Syst. (PMAPS)*, Oct. 2016, pp. 1–6.
[29] H. M. Louie, "Time-series modeling of aggregated electric vehicle charging station load," *Electr. Power Compon. Syst.*, vol. 45, no. 14, pp. 1498–1511, Aug. 2017.
[30] M. Majidpour, C. Qiu, P. Chu, H. R. Pota, and R. Gadh, "Forecasting the EV charging load based on customer profile or station measurement?" *Appl. Energy*, vol. 163, pp. 134–141, Feb. 2016.
[31] S. Ai, A. Chakravorty, and C. Rong, "Household EV charging demand prediction using machine and ensemble learning," in *Proc. IEEE Int. Conf. Energy Internet (ICEI)*, May 2018, pp. 163–168.
[32] J. R. Helmus, J. C. Spoelstra, N. Refa, M. Lees, and R. van den Hoed, "Assessment of public charging infrastructure push and pull rollout strategies: The case of The Netherlands," *Energy Policy*, vol. 121, pp. 35–47, Oct. 2018.
[33] A. Lucas, G. Prettico, M. Flammini, E. Kotsakis, G. Fulli, and M. Masera, "Indicator-based methodology for assessing EV charging infrastructure using exploratory data analysis," *Energies*, vol. 11, no. 7, p. 1869, Jul. 2018.
[34] R. Wolbertus, M. Kroesen, R. van den Hoed, and C. Chorus, "Fully charged: An empirical study into the factors that influence connection times at EV-charging stations," *Energy Policy*, vol. 123, pp. 1–7, Dec. 2018.
[35] A. Lucas, R. Barranco, and N. Refa, "EV idle time estimation on charging infrastructure, comparing supervised machine learning regressions," *Energies*, vol. 12, no. 2, p. 269, Jan. 2019.
[36] M. Straka, P. De Falco, G. Ferruzzi, D. Proto, G. Van Der Poel, S. Khormali, and L. Buzna, "Predicting popularity of electric vehicle charging infrastructure in urban context," *IEEE Access*, vol. 8, pp. 11315–11327, 2020.
[37] Netherlands Enterprise Agency. (2019). *Electric Vehicle Charging—Definitions and Explanation*. Accessed: Jun. 19, 2019. [Online]. Available: https://www.nklnederland.nl/uploads/files/Electric_Vehicle_Charging_-_Definitions_and_Explanation_-_january_2019.pdf
[38] *European Alternative Fuels Observatory*. Accessed: Feb. 15, 2019. [Online]. Available: https://www.eafo.eu/
[39] C. Netherlands. *Population Cores in the Netherlands*. Accessed: Nov. 17, 2017. [Online]. Available: https://www.cbs.nl/nl-nl/achtergrond/2014/13/bevolkingskernen-in-nederland-2011
[40] C. Netherlands. *Neighbourhoods Dataset 2015*. Accessed: Aug. 20, 2018. [Online]. Available: https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/wijk-en-buurtkaart-2015
[41] N. EnergieAtlas. *Energy Atlas*. Accessed: Oct. 16, 2018. [Online]. Available: https://www.pdok.nl/introductie/-/article/cbs-aardgas-en-elektriciteitsleveri-1
[42] M. V. B. Z. E. Koninkrijksrelaties. *Liveability Meter*. Accessed: Oct. 15, 2018. [Online]. Available: https://data.overheid.nl/dataset/leefbaarometer-2-0—meting-2016
[43] O. R. N. Laboratory, *Landscan Datasets*. Accessed: May 20, 2018. [Online]. Available: https://landscan.ornl.gov/landscan-datasets
[44] C. Netherlands. *CBS Land Cover*. Accessed: Nov. 14, 2017. [Online]. Available: https://www.pdok.nl/introductie/-/article/statistics-netherlands-land-use-2015
[45] N. I. P. for Health and the Environment. *Traffic Flows Database, The Database Was Provided for Research Purposes by the National Institute for Public Health and Environment*. Accessed: Jan. 7, 2019. [Online]. Available: http://www.rivm.nl/
[46] *OpenStreetMap*. Accessed: Feb. 13, 2019. [Online]. Available: https://www.openstreetmap.org
[47] *OpenChargeMap*. Accessed: Jan. 10, 2019. [Online]. Available: https://openchargemap.org/
[48] *OplaadPalen*. Accessed: Feb. 20, 2019. [Online]. Available: https://www.oplaadpalen.nl/
[49] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning With Sparsity: The Lasso and Generalizations*. London, U.K.: Chapman & Hall, 2015.
[50] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*. Springer, 2014.
[51] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. New York, NY, USA: Springer, 2013.
[52] A. Abdelmonem, M. Susanne, and V. A. Clark, *Practical Multivariate Analysis*. Boca Raton, FL, USA: CRC Press, 2011.
[53] D. Hsu, "Identifying key variables and interactions in statistical models of building energy consumption using regularization," *Energy*, vol. 83, pp. 144–155, Apr. 2015.
[54] S. Chatterjee and A. S. Hadi, *Regression Analysis by Example*. Hoboken, NJ, USA: Wiley, 2015.
[55] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. B, Methodol.*, vol. 58, no. 1, pp. 267–288, Jan. 1996.
[56] W. J. Warren-Hicks and A. Hart, *Application of Uncertainty Analysis to Ecological Risks of Pesticides*. Boca Raton, FL, USA: CRC Press, 2010.
[57] A. Siegel, *Practical Business Statistics*. New York, NY, USA: Academic, 2016.
[58] J. Ma and J. C. P. Cheng, "Estimation of the building energy use intensity in the urban scale by integrating GIS and big data technology," *Appl. Energy*, vol. 183, pp. 182–192, Dec. 2016.

MILAN STRAKA received the Ph.D. degree from the University of Žilina, in 2020. He is currently a Postdoctoral Researcher and with focus on developing data science methodologies for transport and energy applications, mainly in the electromobility field.

RUI CARVALHO was a Senior Research Associate with the Statistical Laboratory, University of Cambridge. Prior to that, he was a Research Fellow, and a Researcher Co-Investigator with the School of Mathematical Sciences, Queen Mary University of London. He is currently an Assistant Professor with the Department of Engineering, University of Durham, U.K. His research interests include the integration of methods of convex optimization and statistical machine learning. He has been a member of the EPSRC Peer Review College and also a Mid Career Fellow of the Durham Energy Institute.

GIJS VAN DER POEL received the B.Sc. degree in urban planning from the University of Amsterdam and the M.Sc. degree in economic growth, innovation and spatial dynamics from the University of Lund. He is currently a Market and Data Analyst with ElaadNL. He has been working on geographic models regarding electric mobility ever since. His interests include rollout strategies, spatial optimization, and usage analytics regarding public charging infrastructure.

L'UBOŠ BUZNA received the Ph.D. degree from the University of Žilina, in 2003. In the past, he worked as a Postdoctoral Researcher with several institutions, such as the University of Barcelona from 2014 to 2015, ETH Zurich from 2007 to 2009, and TU Dresden from 2005to 2007. He is currently a Professor of applied informatics with the University of Žilina. His research interests include the development of optimization algorithms applied to transportation and complex systems.