Received: 1 September 2017 | Revised: 21 November 2017 | Accepted: 28 November 2017

DOI: 10.1002/er.3978

SPECIAL ISSUE PAPER

WILEY INTERNATIONAL JOURNAL OF ENERGY RESEARCH

A data-driven statistical approach for extending electric vehicle charging infrastructure

Dario Pevec[1] | Jurica Babic[1] | Martin A. Kayser[2] | Arthur Carvalho[3] | Yashar Ghiassi-Farrokhfal[2] | Vedran Podobnik[1]

[1]Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia

[2]Rotterdam School of Management, Erasmus University, Rotterdam, The Netherlands

[3]Farmer School of Business, Miami University, Oxford, USA

Correspondence
Vedran Podobnik, Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia.
Email: vedran.podobnik@fer.hr

Funding information
Managing Trust and Coordinating Interactions in Smart Networks of People, Machines and Organizations, Grant/Award Number: UIP-11-2013-8813 and DOK-2015-10-1777

Summary

Current trends suggest that there is a substantial increase in the overall usage of electric vehicles (EVs). This, in turn, is causing drastic changes in the transportation industry and, more broadly, in business, policy making, and society. One concrete challenge brought by the increase in the number of EVs is a higher demand for charging stations. This paper presents a methodology to address the challenge of EV charging station deployment. The proposed methodology combines multiple sources of heterogeneous real-world data for the sake of deriving insights that can be of a great value to decision makers in the field, such as EV charging infrastructure providers and/or local governments. Our starting point is the business data, ie, data describing charging infrastructure, historical data about charging transactions, and information about competitors in the market. Another type of data used are geographical data, such as places of interest located around chargers (eg, hospitals, restaurants, and shops) and driving distances between available chargers. The merged data from different sources are used to predict charging station utilization when EV charging infrastructure and/or contextual data change, eg, when another charging station or a place of interest is created. On the basis of such predictions, we suggest where to deploy new charging stations. We foresee that the proposed methodology can be used by EV charging infrastructure providers and/or local governments as a decision support tool that prescribes an optimal area to place a new charging station while keeping a desired level of utilization of the charging stations. We showcase the proposed methodology with an illustrative example involving the Dutch EV charging infrastructure through the period from 2013 to 2016. Specifically, we prescribe the optimal location for new ELaadNL charging stations based on different objectives such as maximizing the overall charging network utilization and/or increasing the number of chargers in scarcely populated areas.

KEYWORDS
charging infrastructure, data science, electric vehicles, green transportation, energy informatics

1 | INTRODUCTION

CO_2 emissions are considered one of the prime factors behind climate change. The transportation sector, in particular, is one of the main contributors to CO_2 emissions.[1] This clearly implies that a possible solution to reduce such emissions is to invest in green transportation. We explore a specific type of green transportation, ie, electric

vehicles (EVs),[2] which produce zero tailpipe emissions, thus being perceived as sustainable when they are sourced through renewable sources, such as solar energy, wind energy, or biomass energy cite(richardson2013electric, dwarakanath2017solar). Even when EVs are charged via the main energy grid, the net CO_2 emission is still likely to be reduced when compared to conventional cars since some fraction of the total energy supply is provided by renewable energy sources. Developed countries throughout the world are already making a significant progress towards switching from internal combustion engine vehicles to EVs. For example, the EV penetration in the Netherlands has increased ninefold in the 4-year period between 2013 and 2016.[3]

To sustain such a growth in EV sales, it is important to understand the underlying preferences and behavior related to current and future EV owners. For example, Adnan et al[4] developed a theoretical framework for modeling EV adoption behavior with the ultimate goal of understanding the driving factors related to EV adoption. In particular, the authors found that one of the key factors that negatively impacts current and future EV owners is *range anxiety*, which is defined as the fear of running out of electricity before reaching an available charging station (CS).[5] There are 2 possible solutions for lowering range anxiety: to increase the capacity of EV batteries and/or to populate the EV charging station network with new chargers. This paper focuses on the latter approach and poses a solution on how the EV charging station network should be extended in a smart way with the aim of not just lowering range anxiety but also optimizing additional business and/or societal objectives. The rationale behind our approach lies in the fact that the surroundings near the location of a new charger plays an important role when determining the impact resulting from the charger network expansion.

The above setting leads us to formulate the following research question: "Where should an EV charging infrastructure provider place a new charging station?". The answer to this question could be different based on the perspective of different stakeholders. First, from the grid operator's point of view, it is important to place new charging stations in a way to minimize peak load and distribute the load evenly. Second, from the charging station owner's point of view, it is important to maximize the total utilization of the charging network to maximize profit. Finally, from EV owners' and local governments' points of view, placing new charging stations in less populated areas to decrease range anxiety could be more important. In this paper, we explore the last 2 stakeholders' perspectives.

That being said, this research is interdisciplinary in nature, touching on the areas of green transportation, energy informatics, and data science. First, *green transportation* is a generic term for zero-emission vehicles (eg, cars, trains, or buses). Second, energy informatics uses information and communication technologies to analyze and improve energy systems.[6] Finally, *data science* is a relatively new engineering area that provides methods and tools not only for statistical analysis of (big) datasets but also for highly accurate predictive modeling. In this paper, we show how data science tools, aligned with the key goals behind energy informatics, can be used for analyzing and improving a specific green transportation challenge, namely, the deployment of EV charging infrastructure (EVCI).

The contribution of this research is a novel methodology for recommending the optimal location of a new charging station to an EVCI provider or a local government, as we describe in Section 3. The key features of the proposed methodology are the following:

- Introducing preprocessing techniques characterized by a novel clustering method and multiple variable transformations, which in turn are used to generate representative synthetic data trace (see Sections 4 and 5);
- Providing a powerful predictive model built on the preprocessed data for estimating the utilization of EV charging stations. This statistical model is characterized by a high predictive accuracy and low computational demand once the same is deployed (see Section 6.2).

Additionally, our research confirms findings from other researchers in the field regarding the aggregate energy demand pattern of EV owners (Section 6.1). Moreover, we highlight the importance of using driving-distance metrics rather than aerial-distance metrics when modeling transportation problems (Section 5.1).

The proposed methodology relies on multiple sources of heterogeneous real-world data: (1) *business data* such as historical data about EV charging transactions and information about competitors in the market and (2) *geographical data* such as places of interest (PoIs) located around chargers and driving distance between available chargers. Based on business and geographical data, the trained predictive algorithm is able to accurately predict charging utilization, which in turn is used when prescribing the optimal location for a new charging station regarding the charging infrastructure operator's and/or local government's underlying objectives. This methodology is general enough that it can be used by various charging infrastructure providers to extend their infrastructure while maintaining high infrastructure utilization. Furthermore, prescriptive insights based on synthetic data traces can be used by policy makers (eg, local governments) to impose certain rules with respect to the required number of chargers aiming at promoting EVs and increasing their numbers on the road.

2 | LITERATURE REVIEW

In this section, we position our work against the relevant literature.

2.1 | EV owners' charging behavior

To build a powerful predictive model for estimating the utilization of EV chargers in different contexts, we first have to study EV owners' charging behavior to validate the underlying dataset, eg, to ensure that some values are within the expected boundaries as well as to understand regularities and patterns in the dataset. There are many studies that deal with the problem of detecting EV owners' charging patterns. For example, Develder et al [7] investigated EV owners' charging patterns using 2 different real-world datasets belonging to different EVCI providers (ElaadNL and iMove). Based on the clustering of arrival and departure times of EVs to/from charging stations, charging sessions were classified into 3 categories: (1) parking to charge; (2) charging near home; and (3) charging near work. We adopt a similar classification in our work, where chargers are classified as *near work* or *near home* (see Section 6.1). Develder et al[7] also discovered different charging patterns for charging sessions happening during weekdays and weekends. While some interesting observations were made by Develder et al,[7] the influence of key variables on charging behavior, such as the number of available charging stations in a region or the points of interest around charging stations, was not investigated. We close this gap with our work by studying those key variables.

Different than previous studies that used data from public charging stations, the research by Franke and Krems [8] used data from private charging stations. To understand the charging patterns of EV owners, Franke and Krems[8] considered 2 important factors: the state-of-charge of EV batteries and driving range. Although this research was focused on understanding the charging patterns of EV owners, we note that the same is significantly different than our research since our ultimate goal is to apply the acquired knowledge about charging patterns to build a predictive model that can be used to extend a charging infrastructure.

The research by Taylor et al[9] estimated charging patterns based on travel data from conventional vehicles with the ultimate goal of assessing their impact on power distribution systems. Although this work developed a valuable methodology, we note that the same did not rely on EV data, and it was focused only on plug-in hybrid EVs, a technology different than what we consider in this paper.

2.2 | EV charging station deployment

The other research stream highly relevant to our study deals with the problem of deploying a charging station. To solve the charging station deployment problem, researchers have used several different methods. As we detail next, most of the relevant papers use methods based on mathematical programming, machine learning, and computer simulations.

He et al[10] suggested a game-theoretic approach to find the optimal place for the deployment of a new charging station in a regional road network (ie, roads between cities and/or metropolitan areas). On the basis of the previous research by Tuttle and Kockelman[11] and Lin and Greene, [12] He et al[10] assumed that EV owners favor the path to the destination that has available charging stations. Given that assumption, the authors also modelled the equilibrium between energy supply and demand. Our work circumvents a limitation with the model by He et al[10] in that our model is able to suggest the precise number of chargers to be placed in a certain charging station zone without relying on strong assumptions related to EV owners' behavior.

Chen et al[13] also applied mathematical programming techniques to solve an optimization problem related to the deployment of new charging stations. Chen et al[13] dealt with such a problem by taking the perspective of car parking. In particular, based on data from Washington state's parking lots, those authors first determined parking location and the duration of different trips. This information was then used to predict the zone-level parking demand and the trip-level parking durations. Thereafter, an optimization technique was applied to optimize the deployment of a new charging station with the goal of minimizing the cost to drive and the driving distance between 2 highly important zones. We note that the above research did not use data that are exclusively EV related and, therefore, the obtained results are likely less reliable when applied to fulfill the goal of deploying charging stations when compared to research based on EV-related data.

The research by Sadeghi-Barzani et al[14] also used mathematical programming methods to solve an optimization problem of deploying a new charging station in a region. Different than our work, the model by Sadeghi-Barzani et al[14] focused on fast chargers in urban areas. Specifically, these authors assumed a fixed value of 3 km as the distance between charging stations and, based on that distance, they tried to optimize the placement of charging stations while minimizing costs.

Liu et al[15] developed an interesting mathematical model for the optimal planning of EVCI. The major difference when compared to our study is the objective function these authors used, ie, the minimization of the total cost associated with the charging station deployment. Finally, the

study by Mak et al[16] developed a methodology based on mathematical approaches for the optimal infrastructure planning. The main focus of their research was on the charging stations support for battery swapping, an aspect that is not considered in our model.

Instead of using optimization techniques, the studies by Ip et al[17] and Andrenacci et al[18] used machine learning methods to find the optimal location for a new charging station. Ip et al[17] proposed a method that starts by dividing roads into segments. Thereafter, each segment is placed into a 2-dimensional space based on its utilization. Next, a clustering method is applied to group the road segments based on their utilization intensity. After clusters become available, an optimization function is used to decide the most suitable cluster for a new charging station. The main problem with the proposed solution is that it assumes that the data provided by various sensors deployed on the road are always available, which in practice is not always true. Unlike the research by Ip et al,[17] the research by Andrenacci et al[18] did not attempt to provide the exact location of a new charging station, instead it suggests the region where a new charging station should be deployed. This paper uses data from 6% of the privately owned vehicles in Rome, with the assumption that all vehicles are electric. All trips that ended in the urban area of Rome were clustered into subregions. A charging infrastructure was then associated with the center of each cluster. Thereafter, the total energy spent on each trip is calculated for each subregion, and the number of charging stations that could satisfy the demand for restoring that energy is the number of charging stations that should be deployed. While the quality of the data and clustering of the Rome urban region into subregions are of high value to researchers, the obtained results are rather questionable since they are based on the assumption that the underlying fleet is all electric, which is not true in practice, thus disregarding behavior specific to EV owners.

The third widely used method to determine the optimal location of a charging station are *computer simulations*. Sweda and Klabjan[19] developed an agent-based model to simulate a complex EV environment. Their model uses real-world data that includes sales, prices of EVs, and reference prices for gasoline and alternative fuels. Driving patterns and the state of charge at arrival at a charging station are, however, randomized. The final model analyzes existing infrastructure, and it does not include the proposal for a new charging station. The model by Lu and Hua,[20] on the other hand, can propose the location for a new charging station with the goal of optimizing usage (ie, where the charging station will be used the most) and the size (ie, number of plugs) of the charging station. Their model is based on queuing theory and is an extension of an earlier work by Capar et al.[21] While the ultimate goal is similar to ours, the main differences between our work and the work by Lu and Hua[20] lie in the underlying method and the lack of EV-related data in the latter work. Finally, Babic et al[22] also combined queuing theory and computer simulations, namely, agent-based simulations, to find an effective investment strategy with respect to the number and speed of chargers within EV-enabled parking lots (EVPLs). In contrast to that work, which takes into account the underlying dynamics among the EVPL, EVs, and the electricity market to solve the EV charger infrastructure sizing problem, the focus of our work is placed on identifying the optimal location of a new charging station.

Besides the aforementioned research based on mathematical programming, machine learning, and computer simulations, researchers have also applied other techniques as well when dealing with the EV charging station deployment problem, such as the approaches in the studies by Dong et al[23] and Guo and Zhao.[24] Dong et al [23] first analyzed the charging and traveling behavior of EV users. Thereafter, a genetic algorithm was applied to decide on the optimal location for a new charging station. What makes our work different is the fact that we are using real-world EV-related data from EV charging stations, while the study by Dong et al[23] used GPS data from conventional vehicles. Guo and Zhao[24] developed an interesting approach for solving the challenge of charging station deployment taking into account economical, environmental, and social criteria. Their methodology includes a fuzzy TOPSIS (Technique for Order Performance by Similarity to Ideal Solution) method to find the optimal location among all possible locations. Unlike our work, the work by Guo and Zhao[24] did not use EV-related real-world data.

3 | METHODOLOGY

The methodology proposed in this research is based on the 4 recursive steps described in Figure 1 and explained in the following subsections. The first 2 steps of the methodology (ie, data collection and data preprocessing) are used to generate a dataset appropriate for further statistical and predictive analysis. The third step is data analysis, followed by the last step, which provides an answer to our research question. Note that the methodology follows ideas set by Pevec et al[25] and the good principles from the well-known CRISP-DM.[26] In particular, *business understanding* and *data understanding* are the first 2 key steps from CRISP-DM, which showcase the importance of data understanding before any predictive analysis is performed.

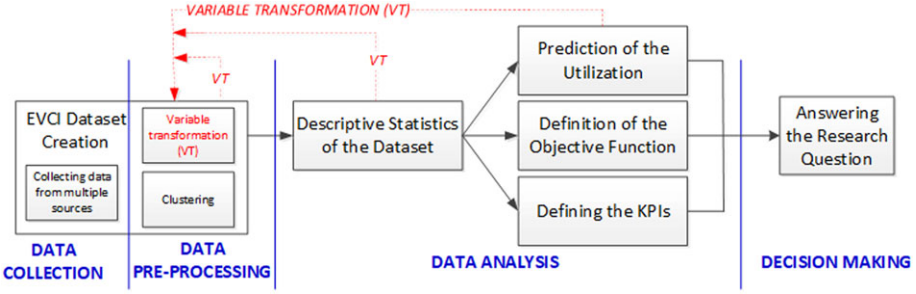

FIGURE 1 Methodology for extending an electric vehicle charging infrastructure (EVCI) [Colour figure can be viewed at wileyonlinelibrary.com]

3.1 | Data collection and preprocessing

The dataset we suggest using for training a predictive model, here called the EVCI dataset, comes from different sources. In particular, we use business data, such as historical data about EV charging transactions and information about competitors in the market, as well as geographical data, such as PoIs surrounding charging stations and driving distance between available chargers, when building such a dataset.

In our case study, the core business dataset is provided by ELaadNL,[27] one of the biggest charging infrastructure providers in the Netherlands (see Section 4.1). We extended the initial dataset with information from multiple additional sources to build a predictive model, as explained in Section 5. This process required experimenting with different variable representations before choosing the best set of variables that can describe the dataset.

3.2 | Data analysis and decision making

After going through the data preprocessing step, the resulting dataset is used for descriptive and predictive analysis, as we explain next.

3.2.1 | Descriptive statistics of the EVCI dataset

Exploring the statistics of the dataset is one of the most important steps in the proposed methodology. This step not only provides insights about the dataset that are necessary when manipulating variables for the sake of more accurate predictions but also results in some interesting insights about EV owners' charging patterns. Section 6 elaborates upon this step with various examples of charging patterns and utilizations for different time intervals, while Section 5.2 describes the variable transformations that are resulting from this step.

3.2.2 | Prediction of charging utilization

Prediction models are built based on the EVCI dataset. The use case for the model is to predict charging utilization when the number of chargers in a charging zone changes, although the model can also predict utilization when other parameters change, eg, the number of competing charging stations or the number of specific PoIs. Section 6 describes the models that we used, together with an analysis of their strengths and weaknesses.

3.2.3 | Defining the key performance indicators (KPI)

A charging station is a place where EV owners can park and charge their vehicles. Each charging station can be equipped with only one charger plug (CP) and only one parking spot, as in Figure 2A, or with multiple charger plugs and parking spots, as in Figure 2B, where the number of plugs is equal to the number of available parking spots. In this work, we consider the *charging utilization* KPI (U_{ch}), which is computed per geographical segmentation called *zones*. For example, the charging utilization in a certain zone is the likelihood that any charging plug in that zone is being used at any arbitrary time. We assume that all charging plugs in the same charging zone are equally likely to be busy (ie, that there is a car being charged there).

Let Z be the total number of zones. For any given zone z, for $z \in \{1, \dots, Z\}$, N_z and N_{tot} denotes, respectively, the number of charging plugs in that charging zone and the total number of charging plugs. Clearly, we have that

$$\sum_{z=1}^{Z} N_z = N_{tot}. \tag{1}$$

The *charging utilization* at zone z with a total number of N_z charging plugs is defined as the likelihood that, at any time, a car is being charged by one of the charging plugs in that charging zone. Formally,

$$U_{ch}(z, N_z) = \frac{\sum_{n=1}^{N_z} \sum_{t=1}^{T} I_{ch}(n, z, t)}{N_z T}, \tag{2}$$

where T is the length of the time horizon under study and $I_{ch}(n, z, t)$ is the charging indicator function, which equals

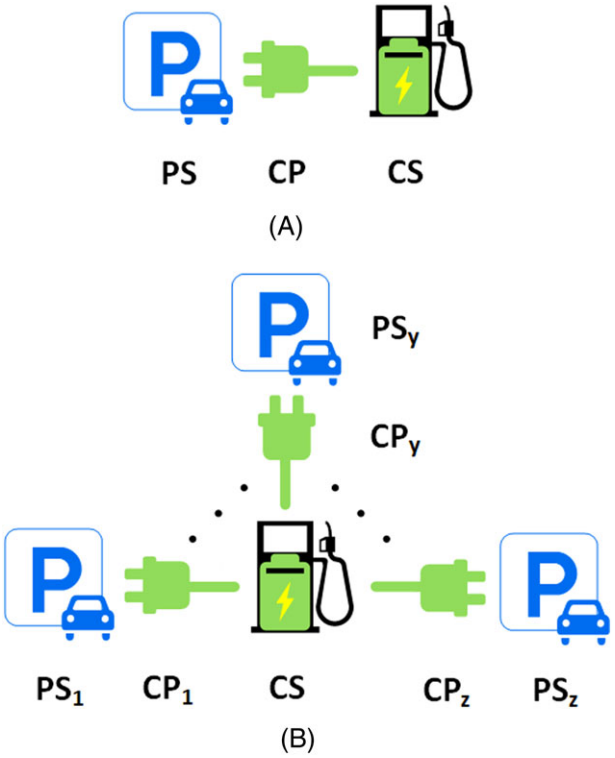

FIGURE 2 A, charging station with one plug. B, charging station with multiple plugs [Colour figure can be viewed at wileyonlinelibrary.com]

to 1 if an EV is charging at charging plug n, zone z, during time t, and 0 otherwise. In other words, $U_{ch}(z, N_z)$ is the likelihood that any charging plug is used in a certain zone over the entire time horizon (T).

3.2.4 | Defining optimization problems for extending EVCI

After building a predictive model, one must next determine how such a model will be used to address the question of the optimal placement of new charging stations. We capture this choice via defining optimization problems. Suppose that the original setting has N_{tot} charging plugs in total. A potential investment will increase the number of charging plugs to $N_{tot} + M_{tot}$. We can split the new M_{tot} chargers among different charging zones in many different ways or permutations. Let $M_z \geq 0$ be the number of new charging plugs in zone z. Then, any feasible vector of values $(M_1, M_2, \ldots, M_Z)$ must satisfy

$$\sum_{z=1}^{Z} M_z = M_{tot}. \tag{3}$$

Note that after the installation of new chargers, the total number of charging plugs in any zone z will be $N_z + M_z$. Define $\mathcal{M}$ to be the set of all feasible tuples $(M_1, M_2, \ldots, M_Z)$ of additional charging plugs in all zones, ie,

$$\mathcal{M} = \left\{ (M_1, M_2, \ldots, M_Z) \middle| \sum_{z=1}^{Z} M_z = M_{tot} \right\}. \tag{4}$$

In other words, $\mathcal{M}$ includes all possible ways the additional M_{tot} charging plugs can be split among all zones. This raises the question: What is the best permutation or tuple among all feasible permutations? or what is the optimal tuple among all feasible tuples in $\mathcal{M}$? Mathematically speaking, this is equivalent to finding an optimal tuple $(M_1, M_2, \ldots, M_Z) \in \mathcal{M}$, denoted by $\mathcal{M}^{opt}$. Clearly, this depends on the underlying optimization problems. Here, we define 3 important optimization problems as follows:

P1: (Utilization maximization): Find the optimal zones to place the new charging plugs so that the new setting has the maximum total utilization. Mathematically,

$$\mathcal{M}^{opt} = \underset{(M_1, M_2, \ldots, M_Z) \in \mathcal{M}}{\operatorname{argmax}} (f_1(M_1, M_2, \ldots, M_Z)), \tag{5}$$

where

$$f_1(M_1, M_2, \ldots, M_Z) = \frac{\sum_{z=1}^{Z} U(z, N_z + M_z)}{\sum_{z=1}^{Z} U(z, N_z)}, \tag{6}$$

and U represents the charging utilization U_{ch} defined in Equation 2. The numerator of f_1 is the total utilization after adding the new charging stations to the system, and the denominator is the original utilization before the upgrading. Clearly, $f_1 \leq 1$, because adding new charging stations decreases the utilizations of the charging stations in their vicinity and, hence, the overall utilization of the fleet when its averaged out over all charging stations.

Adding charging stations in such a way is certainly desirable from a charging station owner's point of view since this leads to profit maximization due to having charging stations in places where they will likely be mostly utilized.

P2: (Unpopulated area first): Another way to formulate the optimization problem is to consider the EV owners' and local governments' points of view. Their primary interest is in the number of available charging stations. One way to capture this view is by defining an optimization problem that tries to increase the number of charging stations in unpopulated area. Formally,

$$\mathcal{M}^{opt} = \underset{(M_1, M_2, \ldots, M_Z) \in \mathcal{M}}{\operatorname{argmax}} (f_2(M_1, M_2, \ldots, M_Z)), \tag{7}$$

$$f_2(M_1, M_2, \dots, M_Z) = \frac{\min_{z\in\{1,\dots,Z\}}(N_z + M_z)}{\bar{N}}, \quad (8)$$

where $\bar{N} = \frac{\sum_{z=1}^{Z}(N_z+M_z)}{Z}$ is the average number of charging stations in each cluster after adding the new installations. Clearly, $f_2 \leq 1$, because the minimum of a set is always less than or equal to its average. This objective function ensures that the new charging plugs are installed in areas with the least number of existing charging plugs, hence giving prioritizing unpopulated areas.

P3: (Hybrid solution): Since the first optimization problem generally favors charging station owners and the second optimization problem favors EV owners and local governments, the third approach aims at combining those 2 potentially conflicting objectives, thus keeping a fair balance between the 2 stakeholders. Mathematically,

$$\mathcal{M}^{opt} = \underset{(M_1,M_2,\dots,M_Z)\in\mathcal{M}}{\text{argmax}} \quad (\alpha f_1 + \beta f_2), \quad (9)$$

where α and β are weights that define the importance of the objective functions f_1 and f_2.

3.2.5 | Answering the research question

Answering the research question at hand equates to finding an optimal $\mathcal{M}^{opt}$ given the underlying optimization problem. This is clearly a decision making process that is based on the previously described steps of the proposed methodology. Moreover, it requires the development of a predictive model that determines how the utilization in Equation 2 varies for any feasible permutation in $\mathcal{M}$ and when the number of charging plugs in a zone z varies. To develop this model, one needs to perform statistical and predictive analysis on the underlying dataset taking into account the defined KPIs. In what follows, we explain how we develop such a model based on our dataset.

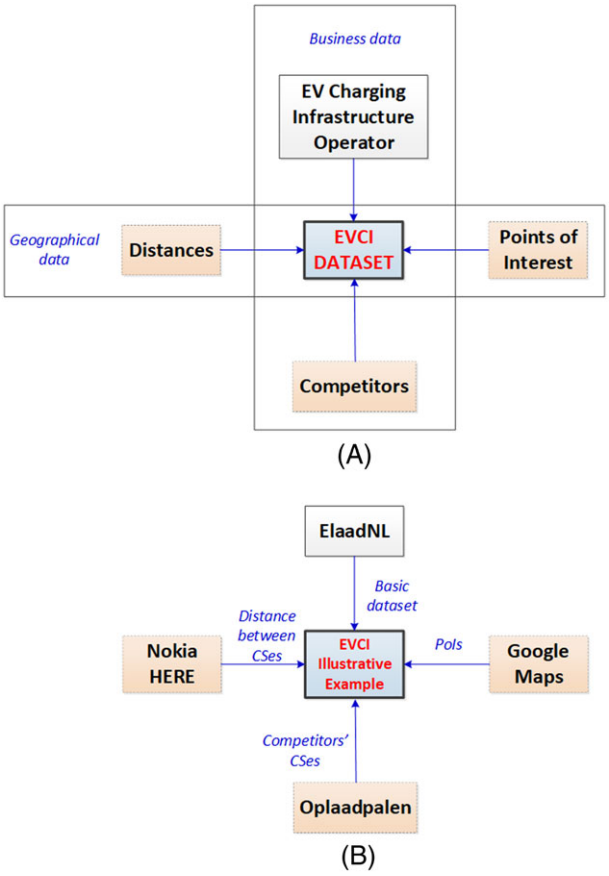

FIGURE 3 A, Components of the generalized electric vehicle charging infrastructure (EVCI) dataset. B, Components of the EVCI dataset in our case study [Colour figure can be viewed at wileyonlinelibrary.com]

4 | DATA COLLECTION

In this section, we describe the processes of generating the EVCI dataset from various different sources, having the EVCI operator's dataset as the core dataset. The main variables that must be present in the core dataset are geographical location of charging stations, number of plugs associated with each charging station, EV user identification, start/end time of charging transactions, and charging time.

As depicted in Figure 3A, the core dataset is extended with geographical data, ie, data containing information about *PoIs*, *distance between charging stations*, and with information about the number of *competing charging stations*. The expanded dataset gives further insights into the environment of each charging station in the core dataset, and consequently, it enables better analysis based on richer contextual insights. In the following subsections, we elaborate on the methods for gathering each contextual information by focusing on our case study, as depicted in Figure 3B.

4.1 | EVCI operator

The core dataset in our case study was provided by one of the largest EVCI providers in the Netherlands, ELaadNL. The dataset consists of all charging transactions for ELaadNL charging stations for 4 consecutive years (2013 to 2016). For our purposes, the relevant variables in the dataset are:

- *TransactionID* - unique numeric identification of a transaction (eg, 1391709);
- *ChargePoint* - identification of a charging station (eg, AL100);

TABLE 1 Number of transactions and charging plugs in the ELaadNL dataset per year

Year	Number of Transactions	Number of Chargers
2013	165 641	2676
2014	342 419	2687
2015	391 375	2728
2016	548 317	2922
CAGR	34.89%	2.22%

Abbreviation: CAGR, compound annual growth rate.

- *Connector* - numeric value representing the number of chargers available in a charging station (eg, 1);
- *StartCard/StopCard* - identification of the user's ID card used in the beginning and end of a charging session (eg, 0488D392 213180);
- *UTCTransactionStart/Stop* - time the charging session started/stopped (eg, 2016-03-07 16:54:10);
- *ConnectedTime/ChargeTime/IdleTime* - the time (in hours) the vehicle was, respectively, connected to the charger, charging, or idle (eg, 4.3834, 3.5003, and 0.8831); and
- *Lat/Lon* - latitude and longitude coordinates of a charging station (eg, 53.19865 and 5.792520).

Each transaction is defined by all actions from the time when an EV owner plugs the EV to the charger (*UTCTransactionStart*) until the EV is unplugged (*UTCTransactionStop*). A charging session starts when an EV owner initiates the transaction with his/her charging card. As soon as the transaction starts, the car begins to charge. After the car is charged to a desired level, it stays in idle mode (ie, not charging) connected to a charger until the EV owner ends the transaction with his/her charging card.

On the basis of transaction variables, we generate an hourly time-based version of the original dataset. This dataset transformation facilitates the study of temporal charging behaviors for different time intervals. Moreover, it introduces more instances for the machine learning algorithms. For example, if a transaction starts at 5 AM and ends at 1 PM, that single line in the core dataset is replaced with 8 rows corresponding to each hour in which the transaction was active.

The number of charging points maintained by ELaadNL and the number of transactions per year are shown in Table 1. Interestingly, the number of transactions is growing at a much faster rate than the number of charging points, as one can see from the calculated CAGR (compound annual growth rate) values.

4.2 | Distances

The distance between charging stations is needed for performing a distance-based clustering of charging stations, which results in the charging zones. We determine the clusters (zones) based on the *driving distance* between charging stations. To do so, we use the Nokia HERE API.[28] Based on the Nokia HERE API and geographical coordinates of the charging stations, the *NxN* driving distance matrix was calculated. All the distances in the matrix are in kilometers. Besides the driving-distance matrix, the *aerial distance* matrix was also created using the haversine formula.[29] Both distance measures are evaluated, and the results are reported in Section 5.1.

4.3 | Places of interest

Places of interest are also added to the core dataset. Places of interest are distributed across 13 categories as follows: *community*, *drinks*, *entertainment*, *finance*, *food*, *health*, *office*, *religion*, *school*, *shop*, *shop essentials*, *sport*, and *transport*. The PoIs were identified for each charging station by considering an area of 500 m radius since PoIs located at that distance have significant influence on charging utilization.[30] Since 2 charging stations can be deployed close to each other (ie, distance lower than 500 m), our algorithm for identifying PoIs takes that into consideration so that no duplicate PoIs are in the EVCI dataset. In other words, even if one PoI is in between 2 charging stations that have the distance between themselves lower than 500 m, that PoI will not have a duplicate entry in the dataset. Places of interest were located and identified with the Google Maps API.[31] As we detail in Section 6.2, PoIs turned out to be interesting features of the dataset since the utilization in a charging zone is highly dependent on the number of PoIs in that zone.

4.4 | Competitors

Covering around 15% of the total EVCI in the country, ELaadNL is not the only EVCI provider in the Netherlands. To account for this, the number of charging stations owned by the competitors in each charging zone was also added to the dataset. Such information is highly important for the prediction of charging utilization since the more alternatives an user has, the higher the probability he/she will choose one of the competitor's charging stations. Information about all charging stations in the Netherlands is available through the Oplaadpalen API.[32] In the EVCI dataset, the number of competing charging stations is treated as a separate PoI category.

5 | DATASET PREPROCESSING

In this step, we explain how the EVCI dataset was created with information about the charging zones through

clustering and how variables were changed based on observations from other steps of the proposed methodology.

5.1 | Clustering

It is well known that the accuracy of predictive algorithms is not only dependent on the underlying machine learning technique, but also on how the variables are represented. In our solution, predictive accuracy is highly affected by how charging stations are clustered together into charging zones. For that reason, an appropriate selection of a clustering method and its parameter values is required. Figure 4 illustrates the definition of charging zones, *eg*, Z_1 consists of all charging stations, *eg*, CS_1, CS_a, and CS_b that are relatively close to each other.

Clustering is generally used to group entities together based on similarities among them. In our case, entities are the charging stations and similarity is the distance between them. We employ a *hierarchical clustering* approach that builds a dendogram based on that distance. Hierarchical clustering has proven to be very versatile since it enables easy clustering by any pre-defined distance with just one dendogram being built.[33] For example, if one wants the clustering distance to be 3 km, the dendogram can then be cut at the corresponding level. The function that builds the clusters in this research is based on the minimum distance between the elements of clusters, *ie*, we use the single linkage function.

Clusters in this research are based on a fixed 3 km distance. The reason for this is that we assume that an average EV owner has a similar range anxiety as an owner of a fossil-fueled vehicle, meaning that he/she is willing to travel an additional 5 to 10 minutes to reach the next available charging station.[34] Combining this assumption with the fact that the average driving speed is between 20 and 35 km/h in populated areas,[35] allow us to infer that the average distance that an EV owner is willing to travel to reach another available charging station is around 3 km.

We use an innovative approach to compute mutual distances. The most widely used method to compute mutual distances in this field is to calculate the aerial distance between 2 geographical coordinates, eg, haversine distance.[17] We, however, complement this information by computing the *driving distance* between each 2 charging stations acquired via Nokia HERE API as their mutual distance. Aerial distance–based clustering could result in unrealistic scenarios, eg, grouping together charging stations around a bay that have small aerial distance, but long driving distance, or even grouping charging stations that are not connected through land. For example, in the Netherlands, 2 charging stations located near the Oud Valkeveen and the Bikbergen have the aerial distance of 1.6 km, while the driving distance is 2.2 km. On the other hand, the charging station near Oud Valkeveen has only 4 km aerial distance from the charging station located near the Kromslootpark, while the actual driving distance is 16.4 km. Naturally, the driving distance approach is more suitable for the aforementioned applications. After calculating the average distance between all ElaadNL charging stations in the Netherlands for both aerial and driving approaches, it was proven that, on average, the driving distance improves precision by 31% over traditional aerial-based clustering, ie, aerial-distance results in 31% unrealistically shorter distances than driving distance. An illustrative comparison of aerial- and driving-based clusters is depicted in Figure 5, where it can be noticed that the number of clusters is higher when the driving-distance approach is used. As a reminder, the

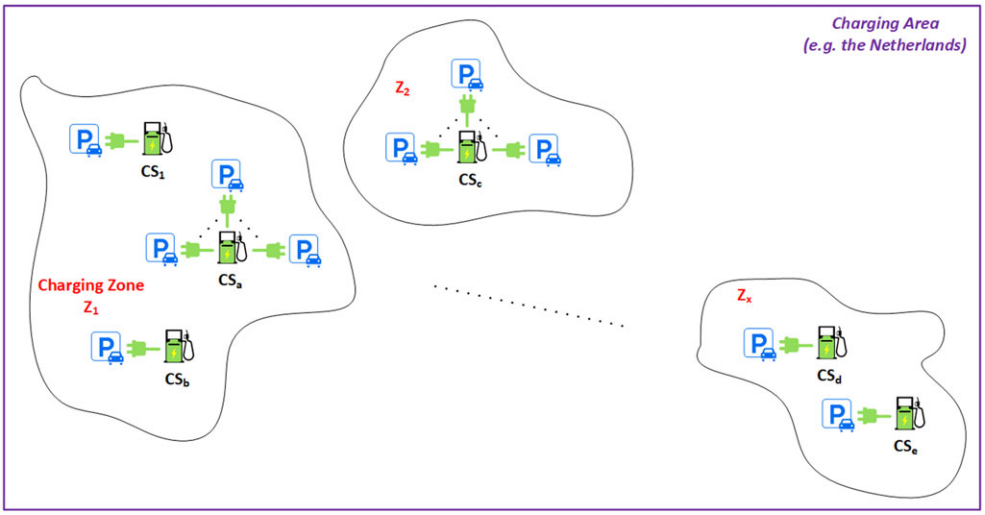

FIGURE 4 Description of charging zones, where a charging zone consists of one or more charging stations [Colour figure can be viewed at wileyonlinelibrary.com]

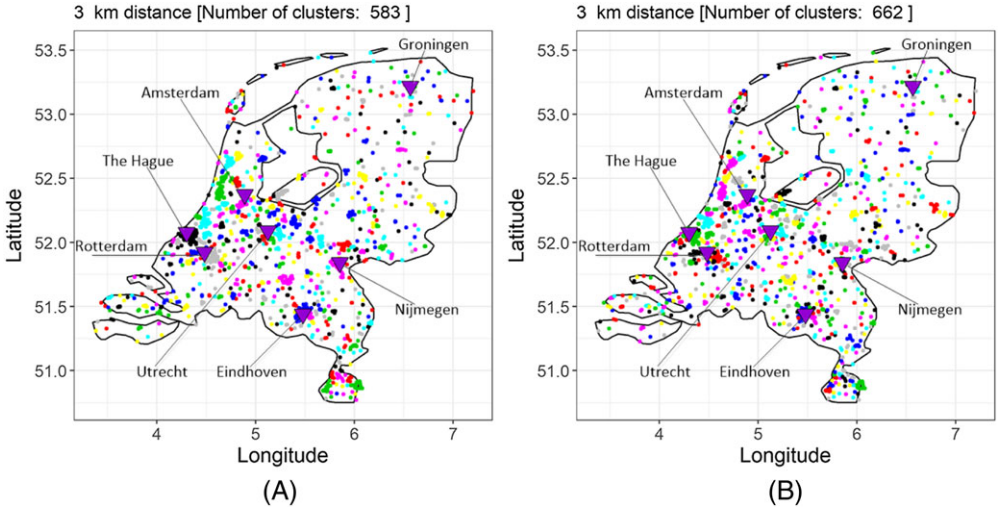

FIGURE 5 Difference between A, aerial-distance and B, driving-distance–based clustering [Colour figure can be viewed at wileyonlinelibrary.com]

clustering step in the EVCI dataset generation identifies the charging zone each charging station in the EVCI dataset belongs to.

5.2 | Variable transformation

Once the EVCI dataset is created, one further needs to select the variables that should be used in the modeling/data analysis phase. A proper representation of such variables helps increasing the accuracy of the machine learning algorithm (eg, multiple linear regression (MLR) and XGBoost) and potentially reducing the computational time for making a prediction. One of the key challenges in this step is to define a representative variable for the PoIs. Clearly, a good representation of PoIs must reflect how they affect utilization in a certain zone. We define 3 candidates as good representative variables for PoIs: (V1) the absolute value, ie, the number of a certain PoI category in a zone; (V2) the relative share of PoIs in a certain zone taking into account the total number of all PoIs (see Table 2); and (V3) the existence of PoIs, ie, 1 if the PoI exist, 0 otherwise (see Table 2). Our methodology uses the summarized value of the first representation of PoIs for prediction since the correlation between the PoIs is very high, as explained in Section 6.2.

TABLE 2 Example of different PoI representation

PoI Category	1	2	3	4	5	6	7
V1	0	1	22	12	11	0	3
V2	0.00	0.03	0.44	0.24	0.22	0.00	0.07
V3	0	1	1	1	1	0	1

abbreviation: PoIs, places of interest; V1, absolute number of PoIs; V2, relative number of PoIs; and V3, existence of PoI category.

Another important factor that needs to be coded into variables is the temporal charging behavior. We define multiple hierarchical variables to capture the temporal behavior in our analysis. The first one is the time of the day. Both charging and parking utilizations, explained in Section 6.1, are highly dependent on the time of the day. The EVCI dataset consists of start and end times for transactions (ie, hour and minutes). To explore how utilizations depend on the time of the day, original variables that describe start and end times of transactions were transformed into a variable called the *category of the day* that can take on only 4 different values:

- morning (from 5 AM to 12 noon);
- afternoon (from 12 noon to 6 PM);
- evening (from 6 PM to 12 midnight); and
- night (from 12 midnight to 5 AM).

Next, we capture a long-term temporal behavior, namely, the *day of the week*, having the possible values "Monday", "Tuesday", … , "Sunday". This is translated into a variable called *isWeekend*, with possible values "YES" and "NO". As the utilization greatly varies between weekdays and weekends (see Figure 6), we used that insight to reduce the number of categorical values from 7 to 2. This translation process highlights the recursive nature of our methodology (see Figure 1): the peculiarities related to charging and parking utilization patterns, eg, the significance of including information on whether a day is a weekday or a weekend, were identified only after the descriptive analysis was performed, and not during the initial variable transformation step.

Lastly, since all categorical variables are represented with numbers, we used the so called *one-hot-encoding* to avoid interpreting them as numerical variables. This

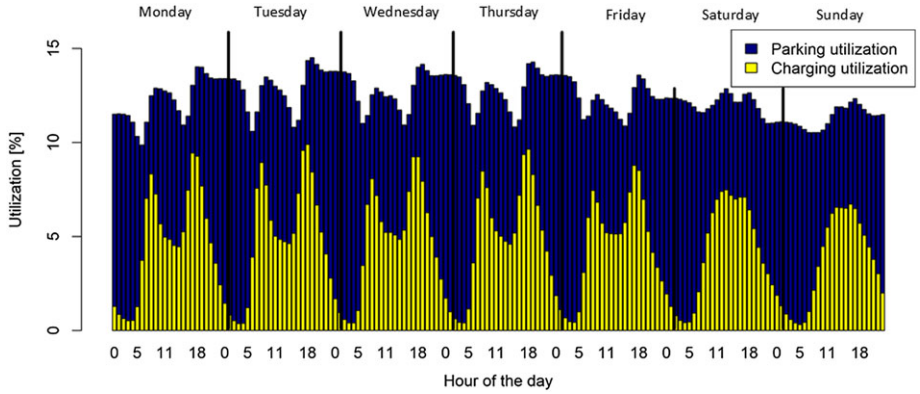

FIGURE 6 Average charging and parking utilization per hour of the day of the week, for the year 2016 [Colour figure can be viewed at wileyonlinelibrary.com]

means that all categorical variables were converted to binary vectors of length equal to the number of different categories. For example, the variable *isWeekend* was converted into a vector with 2 variables, namely, *isWeekend0* and *isWeekend1*, with values 0 and 1 that are complementary.

6 | DATA ANALYSIS

Within the scope of this work, we can differentiate between 2 main analyses: (1) descriptive analysis of the EVCI dataset and (2) predictive modeling of charging utilization. The first step was performed to better understand the underlying dataset, whereas the second step results in a predictive model used to answer our research question. The following subsections describe each of steps and its results in detail.

Henceforth, when we refer to utilization in this section as a function of the time of the day or day of the week (or of the year), we mean the mean utilization in that time or day averaged over all time horizon and over all charging plugs in all zones, which means for the case of charging utilization:

$$U_{ch}(\mathcal{T}) = \sum_{z=1}^{Z} \left(\frac{\sum_{n=1}^{N_z} \sum_{t \in \mathcal{T}} I_{ch}(n,z,t)}{N_{tot}|\mathcal{T}|} \right), \qquad (10)$$

where $\mathcal{T}$ represents a set of time/day values and $|\mathcal{T}|$ is the total number of occasions that a certain time or day occurs in the entire time horizon of the dataset. For example, the charging utilization at 12 PM is the utilization of any charging plug averaged over all the N charging plugs and over the utilization at 12 PM during all days in our dataset. In this case, $\mathcal{T}$ is the set of all days in our dataset, and $|\mathcal{T}|$ is the total number of times 12 PM happens (ie, the total number of days).

6.1 | Descriptive statistics of the EVCI dataset

Understanding the dataset is an important step towards effectively using proper machine learning algorithms to tackle our research question. For this reason, a descriptive statistical analysis of the dataset was performed, which led to interesting conclusions about the utilization of charging stations and parking spaces.

As described in the Section 4, the EVCI dataset consists of charging transaction data concerning 4 consecutive years (ie, from 2013 to 2016). Figure 7 describes the yearly utilization of charging stations and parking spaces. It can be observed that the utilizations of both charging stations and parking spaces increase over the years, which is expected as a result of technological advancements and increased consumer knowledge about and adoption of EVs. The average charging and parking utilizations from the EVCI dataset are provided in Table 3.

In Figure 7, a consistent drop in utilizations is noticeable around July and August. One can argue that such a drop in utilizations corresponds to the period of the year when individuals usually go on vacation. Also, Figure 7 shows that both utilizations are lower in the last quarter of the year 2016 than during the same time in 2015. This might be because of the expansion of the charging station infrastructure of the competitors of ELaadNL. At the end of 2016, ELaadNL had around 15% of the EVCI in the Netherlands. This share is unknown for previous years.

Figure 8 illustrates the utilization for both charging stations and parking spots for each hour of the day for the year of 2016. It can be observed that the charging utilization has 2 peaks during a day, mainly around 8 AM and 5 PM. These correspond to times when drivers usually arrive at workplaces and at home coming from work. With

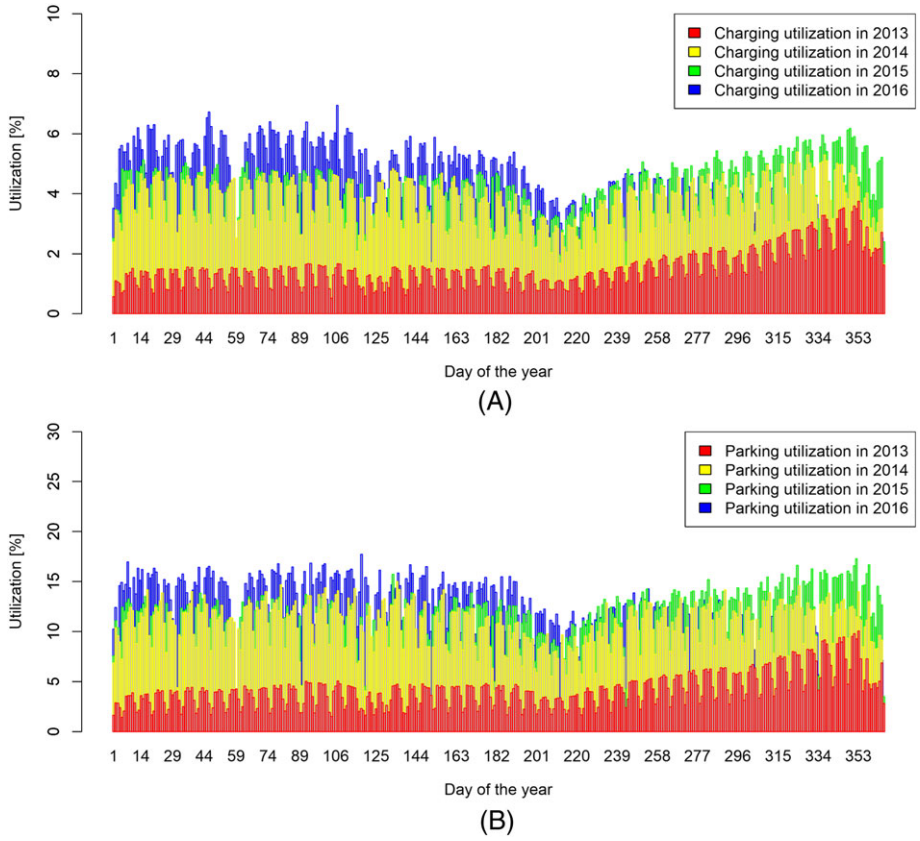

FIGURE 7 Comparison of A, charging and B, parking average utilization over the years 2013-2016 [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Average charging and parking utilizations through the period 2013 to 2016, as calculated from the ELaadNL dataset

Year	Average Charging Utilization, %	Average Parking Utilization, %
2013	1.50	4.12
2014	3.72	11.20
2015	4.10	11.82
2016	4.62	12.71

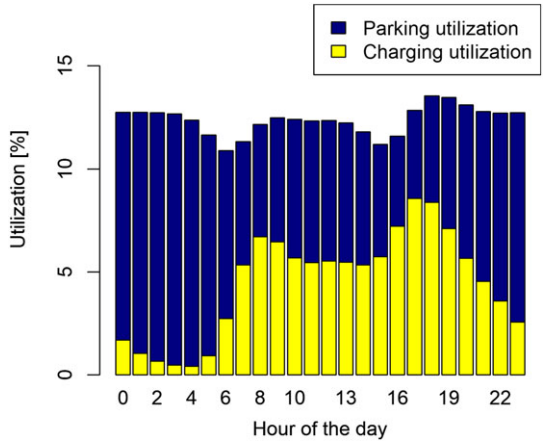

FIGURE 8 Average charging and parking utilization per hour of the day for the year of 2016 [Colour figure can be viewed at wileyonlinelibrary.com]

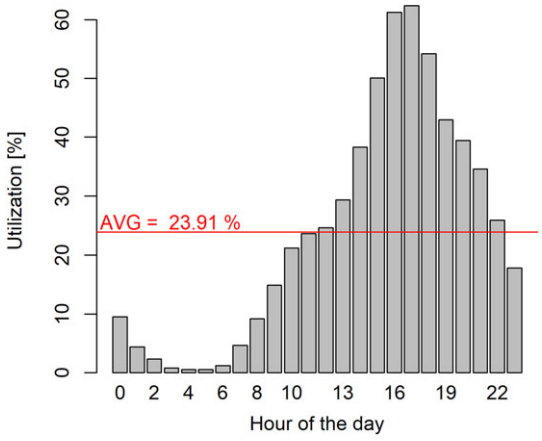

FIGURE 9 Average hourly charging utilization of charging stations located near home for the year of 2016 [Colour figure can be viewed at wileyonlinelibrary.com]

this information, charging stations can be classified as charge-near-work and charge-near-home, depending on the time the charging station utilization reaches its peak. For the sake of illustration, Figure 9 depicts the average utilization of charging stations near home. Figure 8 also compares the utilization of charging stations and parking spots. The utilization of parking spots is approximately 2 times greater than the utilization of charging stations. Also, the

utilization of parking spots has 2 noticeable drops right before peaks in charging utilization. One can argue that this phenomenon corresponds to the time when drivers are driving to work from home and to home coming from work, thus leaving parking spaces unoccupied.

Another interesting fact is that hourly utilization greatly differs between weekday and weekend, which can potentially be explained by the assumption that cars are used more often during working days, eg, for the sake of commuting to work. Figure 10 depicts the utilization of charging stations per hour of the day on weekends. It can be observed that there is only one peak in the utilization around 3 PM. For the weekday utilization, the pattern is the same as in Figure 8.

Yet another interesting observation about EV owners' charging behavior can be observed in Figure 11, which shows that drivers are more likely to charge their EVs

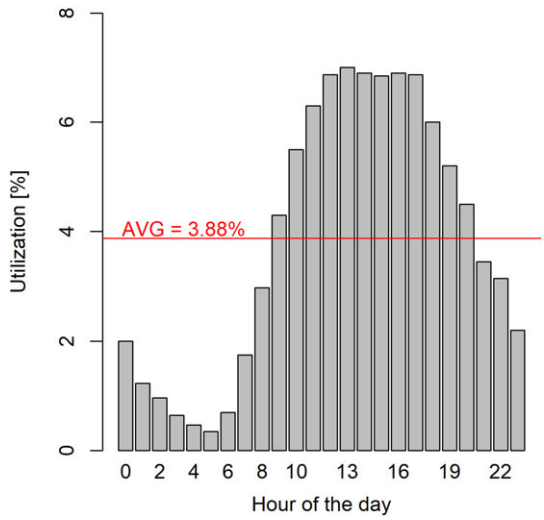

FIGURE 10 Average hourly charging utilization on weekends for the year of 2016 [Colour figure can be viewed at wileyonlinelibrary.com]

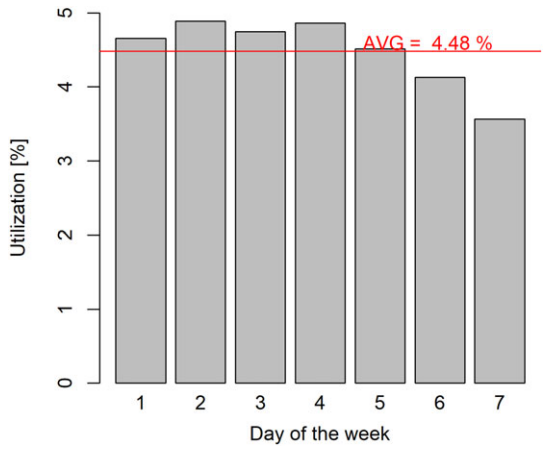

FIGURE 11 Average daily charging utilization for the year of 2016 [Colour figure can be viewed at wileyonlinelibrary.com]

during weekdays than weekends. The utilization reaches its peak around midweek and then starts falling almost linearly.

Charging stations utilization, as can be seen in Figure 7, varies considerably depending on the weather season. Figure 12 depicts daily utilization for each quarter of the year of 2016. The utilization of charging stations is the highest in the first quarter (average of 5.49%), while the utilization is the lowest in the last quarter of the year (3.54%).

The EVCI dataset for the year 2016 includes 1765 charging stations, some of them having more than one charging plug (CP). The total number of charging plugs in the dataset for the year of 2016 is 2922. The top 500 CPs (which make around 17% of the dataset) are involved in 65% of all the charging transactions throughout the year. The rest of the CPs have a negative impact on the overall average utilization of CPs in the Netherlands. That can be seen in the histogram in Figure 13.

Figure 14 compares the average utilization of the top 10, 100, 200, 400, 500, 1000, and all CPs. As the number of charging plugs increases, the charging utilization becomes lower as expected from the previously described fact about charging plugs. The top 10 charging plugs have an average utilization around 27%, while the parking spots associated with them have parking utilizations over 50%. The most utilized chargers are located near big cities.

6.2 | Prediction of Charging Utilizations

We argue that accurately predicting the charging utilization is one of the most important steps towards optimally placing new charging stations. In the scope of this research, 2 different machine learning algorithms are applied to the EVCI dataset. First, to understand how each variable affects charging utilization, an MLR model is trained. When it comes to the predictive model itself, the *XGBoost* algorithm is used due to its versatility in handling nonlinear relationships between variables, which often results in higher accuracy.

We first built multiple MLR models by considering our EVCI dataset under the following time resolutions:

- Day of the year (eg, every day in a year);
- Day of the week (eg, every Monday of a year);
- Hour of the day of the week (eg, 8 AM for every Monday in a year);
- Hour of the day (eg, 8 AM for each day through a year); and
- Category of the day (ie, morning, afternoon, evening, and night) of the week in a month of a year.

Our goal when building all these models was to observe how each variable influences charging utilization

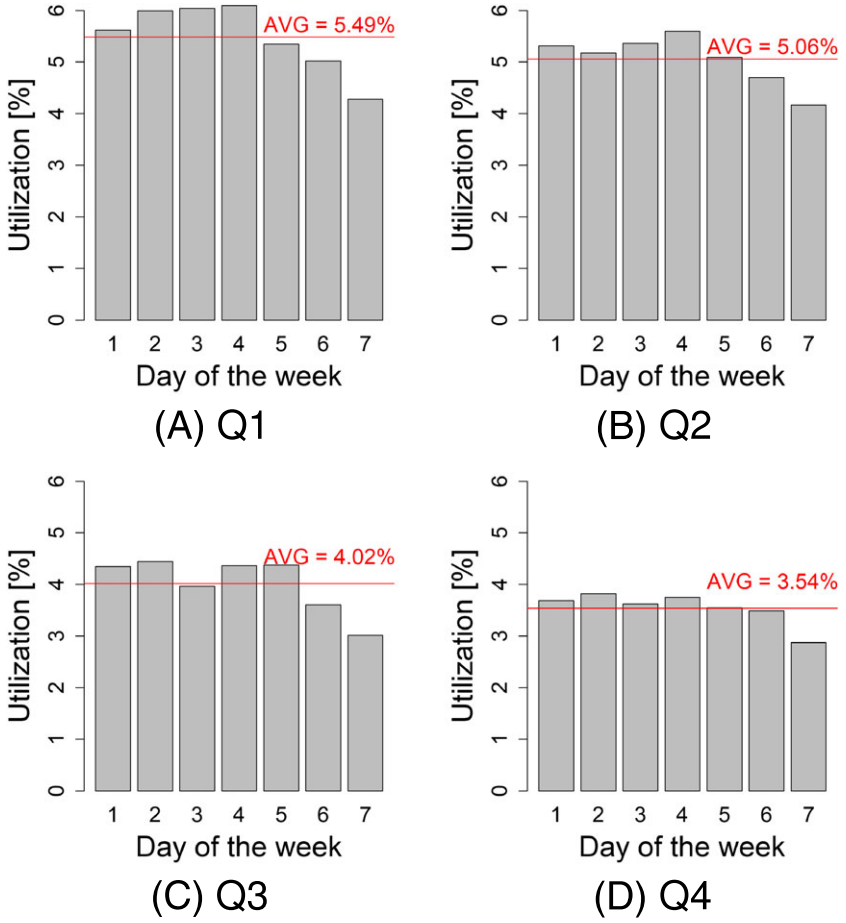

FIGURE 12 Comparison of average charging utilization for each quarter of the year of 2016 [Colour figure can be viewed at wileyonlinelibrary.com]

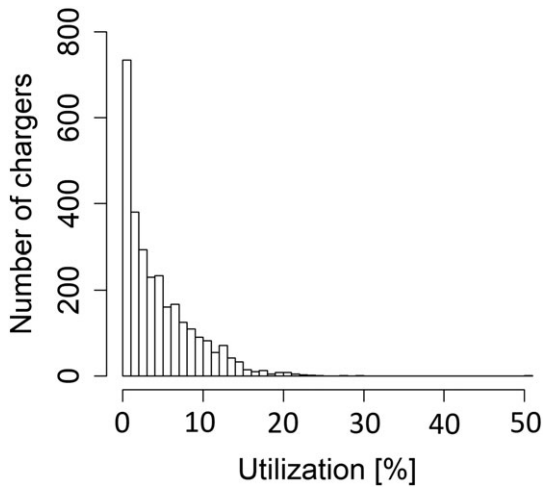

FIGURE 13 Distribution of chargers per utilization level for the year of 2016.

depending on the selected time interval. All scenarios have 17 common variables, ie, 13 + 1 PoI categories, charging zone ID, the number of different EV owners' cards, and the number of charging plugs in the zone. Clearly, the scenarios have different variables for specifying the time interval. After testing PoI categories for correlation, it was discovered that they highly correlate with each other, eg, the PoI category "community" has a coefficient of correlation not less than 0.75 with every other PoI category. Solving this problem was accomplished by summing all PoI categories (except for the variable about competitors' chargers) into one variable: *sumPoi*. After this summation of variables, there were 5 variables shared across all models, namely, the number of competing charging stations, charging zone IDs, the number of different EV owners' cards, the total number of PoIs, and the number of chargers in the charging zone. Recall that charging zone ID is a categorical variable and, as explained in Section 5.2, we used the one-hot-encoding technique to extend that variable into a vector of variables. The same is true for the variables describing the time intervals.

The results of the MLR model for the most complex time interval (ie, hour of the weekend and weekdays for each month) are provided in Table 4. As can be seen from the *p*-values, all variables are relevant, which is understandable after a careful inspection of the dataset. Table 4 also shows the accuracy of the MLR algorithm. This algorithm explains around 30% of the variance in the dependent

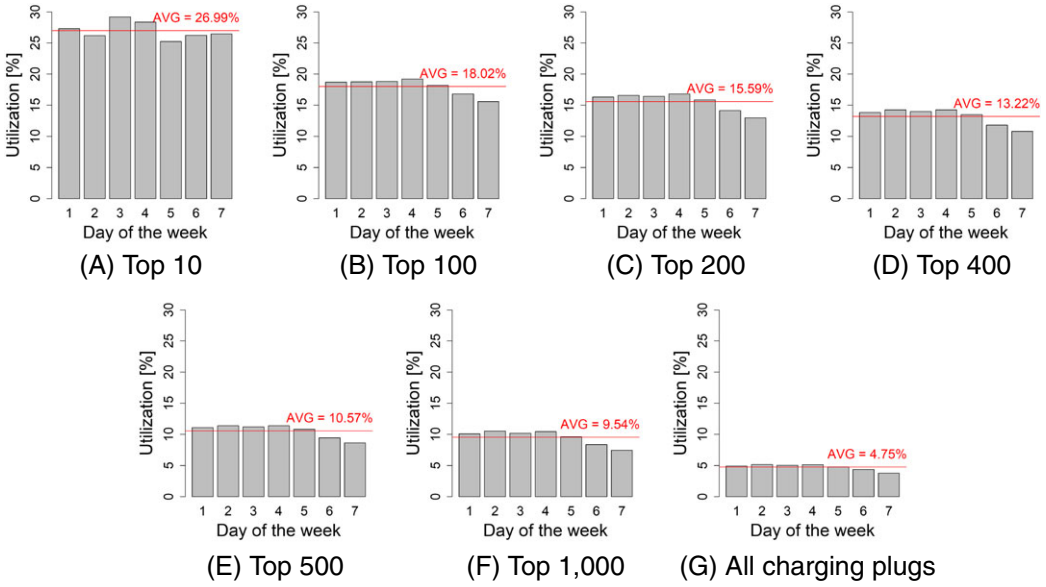

FIGURE 14 Comparison of average utilization of top charging plugs for the year of 2016. [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 4 Result of multiple linear regression prediction algorithm

formula = hourly actual charging time ~ hour day + is weekend + month + number of cards + number of chargers + zone + competitor chargers	
Variable	*p*-value
hour day i, i ∈ [1,23]	*** (hour day 18, **)
month j, j ∈ [1,11]	*** (months 1 and 2, *)
zone k, k ∈ [1,661]	***
is weekend	***
number of cards	***
number of chargers	***
competitor chargers	***
sumPoi	***
Significance codes 0 "***", 0.001 "**", 0.01 "*"	
Residual standard error 0.06495	
Multiple R-squared 0.30650	

variable, which is not good enough to make dependable predictions of charging utilization. The influence of a single variable on utilization can be derived from the estimated coefficients. For example, the variable that describes the number of plugs in a certain area has a negative influence on utilization, ie, for each new added charger, if nothing else changes, the utilization will decrease in expectation by $1.16 * 10^{-3}$. Contrary to the number of plugs, the variable that represents the total number of PoIs in an area has a positive impact on the charging utilization, ie, if another PoI is added to a specific area without changes to other variables, then the utilization is expected to increase by $5.47 * 10^{-6}$.

To predict the charging utilization, the XGBoost algorithm was used since it is currently one of the most popular and accurate machine learning algorithms. [36] This algorithm is based on building decision trees, and it allows for a great fine-tuning through its parameters. To validate and calculate the accuracy of the XGBoost algorithm without inducing bias, the dataset was split into 3 parts: *training dataset*, *validation dataset*, and *test dataset* in a *60:30:10* ratio. The predictive model was built using the hour of the weekday/weekend timespan and was validated on the validation dataset until satisfactory accuracy was achieved. After achieving a high accuracy on the validation set, the test dataset was used to calculate the final model's error.

The XGBoost model is compared against a *baseline model*. The baseline model is the statistical model that returns the historical average value of utilization for a specific cluster at a specific hour of the weekday/weekend. The aforementioned predictive model is natural in this

TABLE 5 Comparison of error measures for the XGBoost algorithm and the baseline model

Measure	XGBoost	Baseline Model
Mean absolute error	0.03184	0.04699
Root mean square error	0.05122	0.07057

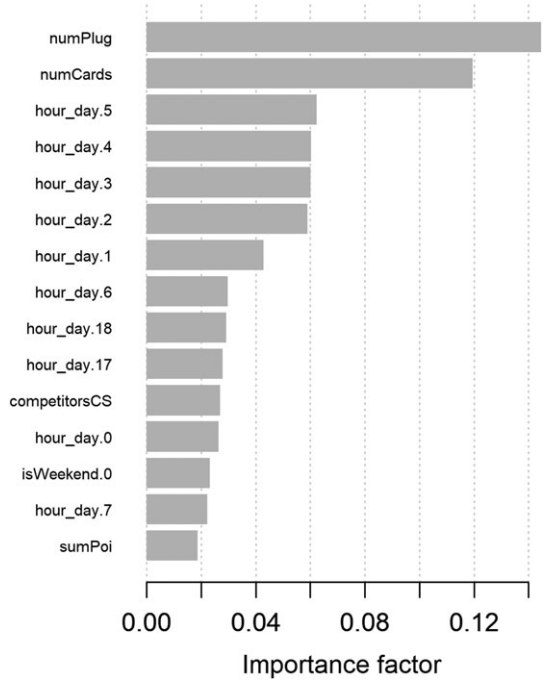

FIGURE 15 Variable importance for the built XGBoost algorithm

scenario because many of the patterns in charging utilization are defined in terms of the time when the charging session occurred (see the previous subsection). In Table 5, a comparison between the prediction errors for the XGBoost algorithm and the baseline algorithm is provided. As it can be seen in that table, the XGBoost has a low error of only 5% (respectively, 3%) for the root mean square error (respectively, the mean absolute error), thus beating the baseline model in the terms of accuracy. One can argue that the reason why the XGBoost model is more accurate is due to the same relying not only on variables from the core dataset (business data) but also on variables that originate from the other sources (eg, POIs from geographical data). Figure 15 depicts the variable importance for the top 15 variables used in the XGBoost model. The variable importance is calculated based on the number of times the decision trees in the XGBoost model were split based on each specific variable. As expected, the number of plugs in a zone is the most influential variable in the model.

6.3 | Illustrative example: extending the ELaadNL EVCI

Since the core dataset (business data) in our case study was provided by ELaadNL, the EVCI provider operating in the whole Netherlands, we illustrate next how to use our model when prescribing a zone to add a new charger, ie, where to deploy a new charging station with one charging plug in the ELaadNL infrastructure. We will showcase our approach with 3 different scenarios: prescribing the optimal location based on the (1) maximization of utilization; (2) increase of charging stations in unpopulated areas; and (3) a hybrid approach between the first 2 approaches. These are the optimization problems we defined in Section 3.2.4.

The first scenario is based on the optimization problem in Equation 5 and the goal of maximizing the total utilization of the EVCI operated by ELaadNL. Specifically, for each zone in the dataset, we run our predictive model to estimate charging utilization after adding one more charger to that zone. Figure 16A reveals that, for this scenario, a new charger should be deployed to the *cluster 525*, ie, in a 3 km radius from the place marked on the map. This is located in a fairly populated part of the Netherlands, being close to 3 big cities: Rotterdam, The Hague, and Amsterdam. This region currently has only 4 charging plugs operated by ELaadNL, thus having a great potential for the addition of new chargers. Besides ELaadNL's charging stations, there are also 9 charging stations from other EV charging station infrastructure providers in the *cluster 525*. If another ELaadNL's charging station is deployed, the average utilization of ELaadNL's charging infrastructure in the *cluster 525* will have a decrease of only 0.000125%, which is not an unusual result since, under the reasonable assumption that the number of charging cards correspond to the number of EVs, that charging zone has around 90 EVs. This relatively high number of EVs, together with a small number of chargers, results in high charging demand, which means that adding another charger will have a low negative impact on the average utilization per charger in the same zone, while having a positive impact on the aggregate utilization for all chargers in the same zone.

After illustrating how a charging infrastructure provider can be informed on where to place a charging station so as to maximize the total charging utilization, we now investigate a different point of view where one wants to place a new charging station in an area that has few stations. In this second scenario, which is defined by Equation 7, the utilization of charging stations in a cluster is completely ignored. Note that this scenario can also be presented via the hybrid function defined in Equation 9 by using the values $\alpha = 0$ and $\beta = 1$. The optimal solution in this scenario is to place the charger in a location in the northern part of the Netherlands (*cluster 633*), in a 3 km radius from the location marked on the map depicted in Figure 16B. That area, close to Groningen, will have a decrease of 4% in the average utilization after the deployment of a new

FIGURE 16 Recommended location for the new ELaadNL charging station based on A, maximizing the overall utilization; B, populating charger unpopulated areas; and C, hybrid approach between first 2 approaches [Colour figure can be viewed at wileyonlinelibrary.com]

charging station, which is a significant decrease in comparison to the first scenario. This result, however, is not surprising since there are only 5 different EV owners' cards in that area, which we assume to be the number of EVs. Currently, this location has only one ELaadNL maintained charging plug as well as one competing charging station.

The first 2 scenarios have potentially conflicting objectives. One way to address this problem is to use the hybrid objective function, defined in Equation 9, which assigns certain weights to each of the 2 previously mentioned objectives (ie, utilization maximization and finding the zone with the lowest number of charging stations). Determining the exact values for the parameters α and β is a complex challenge that must involve multiple stakeholders, eg, charging infrastructure providers, EV owners, and local governments. For example, a decision maker who is more interested in making the charging infrastructure more utilized will define the value of α grater than β, whereas a decision maker who wants to promote EVs by adding chargers in areas with low number of charging stations will do the opposite. The precise estimation of the values of α and β is beyond the scope of this paper. Instead, we now present the third scenario where the first and the second objective functions are equally important, ie, $\alpha = 0.5$ and $\beta = 0.5$. In practice, this may represent a case where there is a fair trade-off between having a well-utilized charging system and deploying charging stations in areas that have the lowest number of charging stations. Based on predictions from our predictive model, the prescribed location for a new charging station is village De Hoef (*cluster 391*), ie, 3 km radius from the place marked on the map depicted in Figure 16C. This area has 2 charging plugs maintained by ELaadNL and a drop in average utilization of 0.002% when a new ElaadNL charger is added. Besides the ELaadNL charging stations, there is only one competing station. This can be explained by the fact that this village has only 8 different charging cards. Moreover, the same is not positioned around a major highway, which results in a lack of investments in charging stations in that location.

7 | CONCLUSION

This research provides a methodology for extending electric vehicle charging infrastructure (EVCI) that is based on machine learning predictive techniques. To answer the research question formed as "Where should an EVCI provider place a new charging station?" it is not just enough to choose and apply a machine learning algorithm. Instead, we suggest in this paper how a dataset that contains charging sessions of EV owners should be enriched with contextual information (eg, PoIs and driving distances between charging stations) to provide more informative recommendations as to where deploy a new charging station.

Using real-world data, the developed methodology is able to recommend the optimal location for a new charging station with respect to, for example, the minimal average charging utilization drop in a charging zone, which is the same as to say that it maximizes the overall aggregate utilization in a charging zone once the new charger is deployed. Besides proposing the location for a new charging station, this methodology also sheds light on the utilization patterns of charging stations as well as EV owners' charging behavior. Hence, the proposed methodology for extending the EVCI can be used by EVCI providers as a decision support tool that prescribes the optimal area to place a new charging station, or by governments as a policy development tool, which enables them to measure the impact of incentives (such as the deployment of

new charging stations) targeting the increase of EVs on the road.

As future work, we plan to extend the proposed methodology and develop a more comprehensive framework for managing an EVCI using big data analysis.[37] Based on the prediction of charging patterns after a new charging station is installed, the proposed framework can be used to address a large range of open questions in this field such as where to reallocate an existing charging station, when it is desirable to remove an existing charging station, or for planing an infrastructure according to the level of EV penetration. Besides the development of the framework, our future work include improvements on the predictive algorithm to increase its accuracy and proposing a well-grounded manner of determining the values for α and β in the hybrid optimization function (see Equation 9). Ultimately, the information provided by the proposed framework would be of great value when it comes to the 3 pillars of sustainability: (1) *people* will have lower range anxiety because the EV charging station infrastructure is optimally deployed; (2) *profit* can be achieved by EVCI providers by optimizing their investment strategies; and (3) *planet* would implicitly benefit as well through an increase of EV sales due to the likely reduction in CO_2 emissions.

ACKNOWLEDGEMENTS

The authors acknowledge the support of the research project "Managing Trust and Coordinating Interactions in Smart Networks of People, Machines and Organizations", funded by the Croatian Science Foundation under grants UIP-11-2013-8813 and DOK-2015-10-1777. The Croatian Centre of Research Excellence for "Data Science and Advanced Cooperative Systems" supported the authors as well. Furthermore, we are thankful to ElaadNL for providing us the data describing their charging infrastructure as well as for their useful comments on the earlier versions of the manuscript and to Oplaadpalen for providing us with API for accessing their data about all chargers in the Netherlands.

ORCID

Dario Pevec http://orcid.org/0000-0002-6926-0914
Jurica Babic http://orcid.org/0000-0001-8241-4116
Arthur Carvalho http://orcid.org/0000-0002-5381-3588
Yashar Ghiassi-Farrokhfal http://orcid.org/0000-0001-6365-1001
Vedran Podobnik https://orcid.org/0000-0001-7070-9283

REFERENCES

1. Stern. The stern review on the economic effects of climate change. *Popul Dev Rev*. 2006;32(4):793-798. http://onlinelibrary.wiley.com/doi/10.1111/j.1728-4457.2006.00153.x/abstract.
2. Saber AY, Venayagamoorthy GK. Plug-in vehicles and renewable energy sources for cost and emission reductions. *IEEE Trans Ind Electron*. 2011;58(4):1229-1238.
3. Cijfers elektrisch vervoer. http://www.rvo.nl/onderwerpen/duurzaam-ondernemen/energie-en-milieu-innovaties/elektrisch-rijden/stand-van-zaken/cijfers. Accessed June 4, 2017.
4. Adnan N, Nordin SM, Rahman I, Vasant PM, Noor A. A comprehensive review on theoretical framework-based electric vehicle consumer adoption research. *Int J Energy Res*. 2017;41(3):317-335.
5. Neubauer J, Wood E. The impact of range anxiety and home, workplace, and public charging infrastructure on simulated battery electric vehicle lifetime utility. *J Power Sources*. 2014;257:12-20.
6. Watson RT, Boudreau MC, Chen AJ. Information systems and environmentally sustainable development: energy informatics and new directions for the IS Community. *Mis Quarterly*. 2010;34(1):23-38. WOS:000275074600002.
7. Develder C, Sadeghianpourhamami N, Strobbe M, Refa N. Quantifying flexibility in EV charging as DR potential: analysis of two real-world data sets. In: 2016 IEEE International Conference on Smart Grid Communications (SmartGridComm). IEEE; 2016:600-605.
8. Franke T, Krems JF. Understanding charging behaviour of electric vehicle users. *Transp Res Part F Traffic Psychol Behav*. 2013;21:75-89.
9. Taylor J, Maitra A, Alexander M, Brooks D, Duvall M. Evaluation of the impact of plug-in electric vehicle loading on distribution system operations. In: Power & Energy Society General Meeting, 2009. PES'09. IEEE. IEEE; 2009:1-6.
10. He F, Wu D, Yin Y, Guan Y. Optimal deployment of public charging stations for plug-in hybrid electric vehicles. *Transp Res B Methodol*. 2013;47:87-101.
11. Tuttle DP, Kockelman KM. Electrified vehicle technology trends, infrastructure implications, and cost comparisons. *J Transp Res Forum*. 2012;51.
12. Lin Z, Greene D. Promoting the market for plug-in hybrid and battery electric vehicles: role of recharge availability. *TTR J*. 2011;2252:49-56.
13. Chen TD, Kockelman KM, Khan M, et al. The electric vehicle charging station location problem: a parking-based assignment method for seattle, Vol. 340; 2013:13-1254.
14. Sadeghi-Barzani P, Rajabi-Ghahnavieh A, Kazemi-Karegar H. Optimal fast charging station placing and sizing. *Appl Energy*. 2014;125:289-299.
15. Liu Z, Wen F, Ledwich G. Optimal planning of electric-vehicle charging stations in distribution systems. *IEEE Trans Power Delivery*. 2013;28(1):102-110.
16. Mak HY, Rong Y, Shen ZJM. Infrastructure planning for electric vehicles with battery swapping. *Manage Sci*. 2013;59(7):1557-1575.
17. Ip A, Fong S, Liu E. Optimization for allocating bev recharging stations in urban areas by using hierarchical clustering. In: 2010 6th International Conference on Advanced Information Management and Service (IMS). IEEE; 2010:460-465.
18. Andrenacci N, Ragona R, Valenti G. A demand-side approach to the optimal deployment of electric vehicle charging stations in metropolitan areas. *Appl Energy*. 2016;182:39-46.

19. Sweda T, Klabjan D. An agent-based decision support system for electric vehicle charging infrastructure deployment. In: Vehicle Power and Propulsion Conference (VPPC), 2011 IEEE. IEEE; 2011:1-5.
20. Lu F, Hua G. A location-sizing model for electric vehicle charging station deployment based on queuing theory. In: 2015 International Conference on Logistics, Informatics and Service Sciences (LISS). IEEE; 2015:1-5.
21. Capar I, Kuby M, Leon VJ, Tsai YJ. An arc cover–path-cover formulation and strategic analysis of alternative-fuel station locations. *Eur J Oper Res*. 2013;227(1):142-151.
22. Babic J, Carvalho A, Ketter W, Podobnik V. Electricity Trading Agent for EV-enabled Parking Lots. *Agent-Mediated Electronic Commerce. Designing Trading Strategies and Mechanisms for Electronic Markets*. Cham: Springer; 2017:35-49.
23. Dong J, Liu C, Lin Z. Charging infrastructure planning for promoting battery electric vehicles: an activity-based approach using multiday travel data. *Transp Res Part C: Emerg Technol*. 2014;38:44-55.
24. Guo S, Zhao H. Optimal site selection of electric vehicle charging station by using fuzzy TOPSIS based on sustainability perspective. *Appl Energy*. 2015;158:390-402.
25. Pevec D, Kayser MA, Babic J, Carvalho A, Ghiassi-Farrokhfal Y, Podobnik V. A computational framework for managing electric vehicle charging infrastructure. In: Proceedings of the 9th International Exergy, Energy and Environment Symposium; 2017.
26. Chapman P, Clinton J, Kerber R, et al. CRISP-DM 1.0 step-by-step data mining guide. 2000.
27. ElaadNL. Smart charging: leven van de wind en rijden op de zon!. https://www.elaad.nl/. Accessed June 2, 2017.
28. Here javascript & rest apis. https://developer.here.com. Accessed June 24, 2017
29. Robusto CC. The cosine-haversine formula. *Am Math Monthly*. 1957;64(1):38-40.
30. Wagner S, Götzinger M, Neumann D. Optimal location of charging stations in smart cities: a points of interest based approach; 2013.
31. Google maps: Google. https://www.google.com/maps. Accessed May 20, 2017
32. Oplaadpalen / oplaadpunten / auto opladen. https://www.oplaadpalen.nl/. Accessed June 24, 2017
33. Mingoti SA, Lima JO. Comparing som neural network with fuzzy c-means, k-means and traditional hierarchical clustering algorithms. *Eur J Oper Res*. 2006;174(3):1742-1759.
34. Langer A, McRae S. Fueling alternatives: evidence from naturalistic driving data. Working paper Tech. Rep.; 2013.
35. Average speed in europe's 15 most congested cities | statistic. https://www.statista.com/statistics/264703/average-speed-in-europes-15-most-congested-cities/. Accessed July 1, 2017
36. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining ACM; 2016:785-794.
37. Ketter W, Peters M, Collins J, Gupta A. Competitive benchmarking: an IS research approach to address wicked problems with big data and analytics. *Manage Inf Syst Q*. December 2016;40(4):1057-1080. http://aisel.aisnet.org/misq/vol40/iss4/14.

How to cite this article: Pevec D, Babic J, Kayser MA, Carvalho A, Ghiassi-Farrokhfal Y, Podobnik V. A data-driven statistical approach for extending electric vehicle charging infrastructure. *Int J Energy Res*. 2018;42:3102–3120. https://doi.org/10.1002/er.3978