Three-step Machine Learning Approach for Location Recommendation and Demand Prediction of DC Charging Stations

Jenny Birzl
M.Sc. Management and Technology
Technical University of Munich
Munich, Germany
jenny.birzl@tum.de

Chun Hei Chung
M.Sc. Management and Technology
Technical University of Munich
Munich, Germany
tommych.chung@tum.de

Christina Hasert
M.Sc. Management and Technology
Technical University of Munich
Munich, Germany
christina.hasert@tum.de

Jee-Hyung Kim
M.Sc. Management and Technology
Technical University of Munich
Munich, Germany
ge37nuc@mytum.de

Jana Schneider
M.Sc. Management and Technology
Technical University of Munich
Munich, Germany
jana.schneider@tum.de

***Abstract*—Since an increasing number of charging stations will have to be placed in the future to meet the demand for electric mobility, the question arises as to where they should be optimally located. After selecting several categories of points of interest and extracting traffic data, we formulate two regression models by segmenting the research area into grids of different sizes and suggesting 1950 candidate locations within Germany. Subsequently, we evaluate these candidate locations by estimating the demand using a decision tree regressor model. The results of the candidates and demand prediction are then interactively visualized via Tableau, with stations ranking such that the best locations can be filtered and identified.**

***Index Terms*—electric charging station, location planning, regression, decision tree, demand prediction**

I. Introduction

With increased awareness of climate-related challenges in recent years and the fact that the transport sector accounts for 27% of the total CO2 emissions in the EU [1], many European countries have made considerable efforts to increase the share of electric vehicles. Also, the German government aims to reach 10 million registered electric vehicles by 2030 [2]. Establishing an appropriate charging infrastructure is, therefore, a crucial prerequisite to reaching this goal. However, recent studies have shown that, for instance, the expansion of charging stations in Germany falls short of expectations, leading to an estimated shortage of around 130,000 charging stations by 2030 [3]. This mismatch between demand and supply of charging stations raises the question of how charging station providers like E.ON can expand the network to fill the demand gap, and where the new stations should ideally be placed.

Which locations are considered as good strongly depends on the perspective of the different stakeholders. Grid operators, for example, aim to place charging stations such that the peak load is minimized and the load evenly distributed. Charging station owners prefer charging station locations that lead to high utilization of the charging network in order to maximize their profit. For electric vehicle drivers and local governments, the focus might rather be on placing charging stations in less populated areas in order to minimize range anxiety, which describes the fear that the energy storage is insufficient to cover the trip distance [4].

Generally speaking, locations can be seen as good if people are spending time at those locations anyway while parking their vehicle. For this reason, Wagner et al. examined the influence of points of interest (POI) on the usage behavior of charging stations. They found that POIs have indeed an influence on the charging station demand however the influence diminishes with increasing distance and depends on the duration of stay at the points of interest which need to be similar to the charging duration. When applying linear regression and assuming a distance of 500 meters, their model resulted in an adjusted R-squared of 0.1. The features of food, health, and museum were the most significant ones, whereas other features like finance, hair care or church were not significant at all. For example, withdrawing money from an ATM takes too little time compared to charging the car whereas going to a museum is a time-consuming activity [5].

A multitude of commonly considered qualitative and quantitative factors relating to optimizing charging stations locations have also been investigated extensively, despite diverging model settings and objectives. By sampling some recent researches on charging station location modelling process, such as articles by Wu [6], Karolemeas [7], Ahmad [8] and Csonka

[9], spatial factors utilized can be categorized as below:

- Traffic network: road network, traffic flow, pit stop rate, highway capacity, proximity to public transport, travel time
- Amenities/Points of Interest: parking facilities, gas stations, restrooms, restaurants, supermarkets, shops, pharmacies, hotels
- Infrastructure factors: energy network, geographical environment
- Operation: construction of charging station, annual maintenance and operation costs, labour cost
- Populations: population and household structure, intention to buy EV, work opportunities
- Land use: functional use of land, urbanized area, green space, land cost

Regarding methodology, previous literature has mainly focused on determining the best charging station locations and predicting the utilization of charging stations from the view of local government, by applying linear optimization and machine learning approaches. For example, Ip et al developed an approach that divides the road into segments. Afterward, a clustering method is applied to group the segments based on their utilization, and an optimization algorithm is used to determine the most suitable cluster [10]. Chen et al. suggest an approach that approximates the charging station demand by predicting the parking demand using linear regression. The results are then combined with a facility location problem [11]. Other authors that are using optimization approaches are He et al. and Zhang et al. applying a multi-period capacitated flow refueling location model to determine good locations of charging stations mainly focusing on coverage [12] [13] whereas Mao et al also consider the power grid's reliability in their optimization model [14].

This paper focuses on finding good charging station locations from the perspective of a charging station provider using a machine learning approach, with a coordinate-based focus of traffic network and amenities. Specifically, a case study of the charging station provider E.ON will be conducted considering Germany as the research area. The paper is structured according to the Cross Industry Standard Process for Data Mining (CRISP-DM), starting with a description of the case study from a business perspective. The subsequent sections provide an overview of the data and main data preparation steps, followed by a description of the applied machine learning model and respective evaluation. The last part covers the deployment as well as the interpretation and summary of the results.

II. Business Understanding

Since the strategy for placing charging stations highly depends on the stakeholder's view, in the following a closer look at E.ON's goals will be taken. According to E.ON's annual report (2021) [15], a goal of its digitization strategy is becoming one of the European leading operators of charging infrastructure by 2030, which in this business case, E.ON is looking to build 1000 new ultra-fast charging stations. These charging stations use the direct current charging (DC) technology, with a typical power rating of 150 kW for fast charging, and 300 kW for ultra-fast charging [16]. The main advantage of these stations is the short charging duration for a relative long range, taking less than 20 minutes to cover a range of more than 100 miles. From these given circumstances, we have summarized the following requirements in this business case:

Since E.ON's main goal in placing new charging stations is to maximize its profit, charging stations shall be placed at locations with high expected demand. As shown in several papers, the demand for charging stations mainly depends on geographical data e.g points of interest or competitors. Also, historical demand data of existing charging stations serve as a good reference for predicting the demand for new charging stations [4].

In terms of feature engineering, it can be concluded that long-distance traveling can be seen as a major use case of DC charging stations [17] and therefore the features of the model should account for this respective customer group. Ultra-fast charging stations are usually less used in urban areas but rather next to roads with higher traffic flows (highways or federal roads) or in close proximity to points of interest where the duration of stay matches the charging duration.

Moreover, suggested charging station locations shall account for coverage expansion. Since E.ON aims to build a large number of new charging stations (1,000 per year) and wants to be among the leading charging station provider in Germany, the newly built charging stations shall be evenly distributed throughout Germany. Since it is not assumed that E.ON pursues an explicit second mover strategy, or aggressive pricing strategy, the newly placed charging stations shall preserve a certain distance from competitors.

Besides, the solution approach shall be adjustable. Since new charging stations will be built on a running basis, the newly gathered data from the already built charging stations shall be included in the model for example to consider changes in the environment, and detect new demand patterns or network effects.

In terms of deployment, the resulting charging stations shall be presented transparently in order to give E.ON an overview of which factors influence the suitability of charging stations and allow them to analyze the locations in an interactive way.

In the subsequent sections, the approach and implemented model will be described in detail taking into account the above-mentioned assumptions.

III. Data Understanding and Preparation

A. E.ON Internal Data Understanding

The charging stations data set ("internal data") from E.ON forms the basis of our analysis and modeling processes. The following information has been provided in two data sheets, for daily and monthly aggregated data respectively from 1st January 2018 to 19th May 2022:

- Measure of demand (masked with prefix "evse"): Mean number of sessions, mean of duration per session, mean of energy consumption per session, sum duration of sessions, and sum of energy consumption.
- Location data: name, coordinates, city, and postal code.
- Charger data: charger ID, maximum power of charger, connector plug type, connector current type, first and last connection dates of charger, charger authentication modes, and operator of chargers.

Moreover, a unique reference for each record can be obtained by combining the charger ID (of EVSE equipment) and the date of record, for both daily and monthly data.

As our analysis is mainly driven by spatial data, two dataframes are created to aggregate data based on locations (by coordinates). Each of the 260 locations contains information such as the mean and standard deviation of the demand measures over time, type of charging equipment, and count of equipment provided in the stations. After creating the dataframes, we visualize the data to understand the relationship among these variables, especially the demand measures. Two methodologies are proposed for visualizing the relationship of the multi-dimension aggregated data. Firstly, the mean values between two or more variables are plotted between the demand measures, for example, the scatter plot in figure 1, the correlation heatmap in figure 2 and the density plot in figure 3. Secondly, the mean and standard deviation of each demand measure are also compared, as in 4.

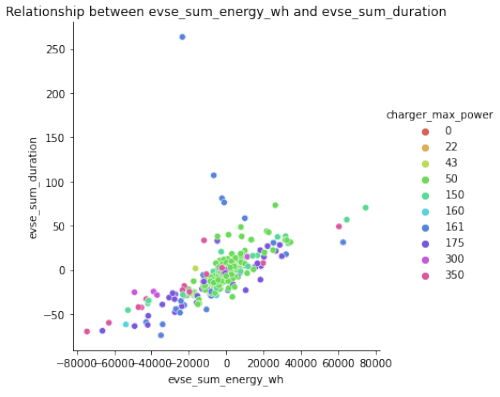

Fig. 1. Scatter plot between sum energy consumption and sum duration

From these visualizations, demand measures are mostly correlated with each other and they can be partially clustered by the charger max power, despite the existence of some

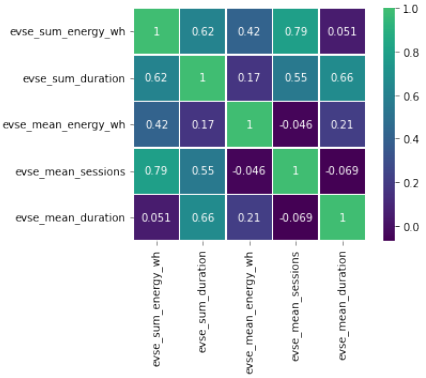

Fig. 2. Correlation heatmap between all demand measures

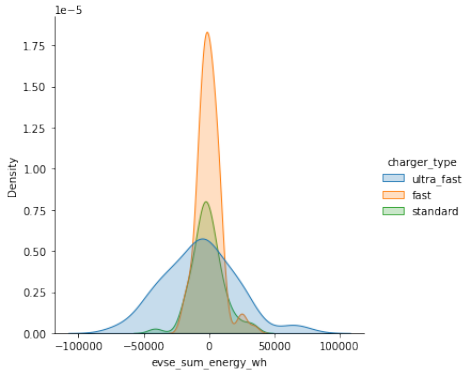

Fig. 3. Density plot of sum energy for each charger type

outliers. In terms of data dispersion, the demand for ultra-high and high power charging stations tends to be more dispersed over time, while for standard charging stations the standard deviations over time are more stable and concentrated. In this report, sum energy (evse sum energy wh) is chosen as the single demand indicator for each location, as this demand measure is closely correlated with the actual revenue of E.ON by operating these charging stations, and has the highest correlations with other demand measures. Other demand

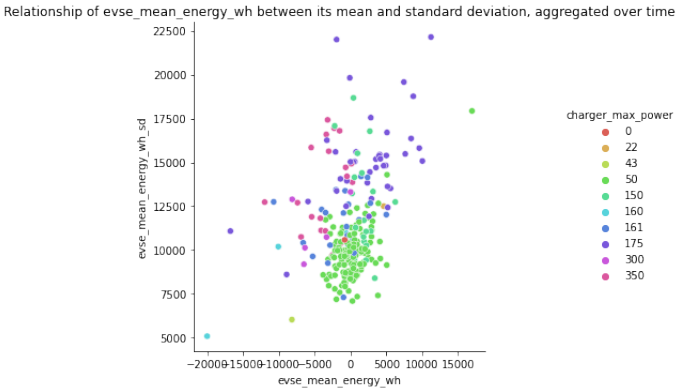

Fig. 4. Scatter plot between mean and standard deviation of sum energy over time

measures also do not provide as much information regarding overall demand, and any combination of these factors (e.g. by addition or multiplication) are not mathematically feasible due to negative values introduced by data masking. For example, the mean energy consumption concerns about usage per session only, and the mean number of sessions only indicates the frequency of customers using a charging station.

B. External Data Understanding

After a detailed understanding of E.ON's internal data, various external data are used to achieve a better result to predict the next charging stations to be built. After considering which features are important for a long-distance traveler at a charging station, as well as other factors that might influence the demand, the following five key elements have been identified: data of competitors, traffic data, parking lots, street points and points of interest.

In term of competitors, a list of all existing charging stations in Germany (as of May 2022) was downloaded from the official site of the Federal Network Agency (Bundesnetzagentur) [18] and for each location, the coordinates, the provider, and the maximum charging power were extracted. Subsequently, all the charging stations operated by E.ON as well as all alternating current (AC) charging stations were removed from the table, leading to a well-arranged list of over 6,000 competitor stations.

Another factor influencing the location of a new charging station is the existence of a parking lot. Since several data sources regarding parking were available, they were checked for completeness and it was decided to download the data from OpenStreetMap which results in around 57,000 parking lots in Germany [19]. In this context, it must be noted that OpenStreetMap is an open source platform and it is therefore not guaranteed that every existing parking space in Germany is actually listed.

In terms of road network, traffic flow data of highways and federal roads in Germany was downloaded from the website of the Federal Highway Research Institute, and only elements with the average number of vehicles per 24-hour period at the respective official counting points for the year 2021 [20] were selected. Since these vehicles counting points (around 900 points on highways and 800 on federal roads) are relatively well distributed across Germany, so that only a few road sections do not have a counting point (especially in Bavaria and northern Germany), the collected data can be used for the intended purposes and represent the traffic situation well. Assuming that a long-distance traveler would not want to take a long detour in order to be able to fast charge, the coordinates of all 9407 highway exits and more than 200,000 available points of the federal roads were also obtained from OpenStreetMap so that the proximity to a highway or federal road could be considered when identifying potential locations for new charging stations.

In terms of amenities, it was considered which points of interest might be attractive to a person while he or she is charging the vehicle. With the current technology, a charging process at 150 kW from 10 to 80 percent state of charge takes about 30 minutes, in the future rather less [21]. Under these circumstances, the coordinates of the following points of interest were downloaded from OpenStreetMap (in total more than 170,000 locations) and used as input factors of the prediction model as described in a later chapter: restaurants, cafes, fast food restaurants, supermarkets, toilets.

Apart from the external data used in this project, one might consider, for example, the number of registered electric vehicles per postal code, the population distribution, and the income distribution of Germany. While these can indicate the potential for the adoption of electric vehicles in a certain region, they cannot indicate the charging behavior of long-distance travelers who merely want to make a charging stop in that region. For this reason, our focus is only on collecting those external data that reflect the business case of long-distance travelers.

IV. Data Modelling

A. Motivation and Solution Structure

The aforementioned analysis of the business case has resulted in 3 major requirements for the machine learning model. Firstly, the charging stations should be placed close to a highway or federal road to obtain locations applicable for long-distance travelers, preferably close to some parking spots or other attractive amenities. Secondly, competing stations' data should be incorporated into the model to enrich the given data from E.ON. Yet, as it is difficult to acquire publicly available information about the competitors' demand level, another modeling focus is to find a way to fully utilize the existing competition's data and incorporate the demand data of E.ON's charging stations. Thirdly, the charging stations should be evenly distributed over Germany and they should not be placed in proximity to other existing stations, for instance, big cities where there is likely an abundant supply of charging stations already. In the following, we would like to outline the strategy to fulfill these three crucial business requirements, followed by the modeling details.

A top-down, zonal approach has been used to check which areas are lacking charging stations. The lack is measured by the difference between the estimated number of charging stations and the actual number within the respective area. We define our method as "boxes-approach", making use of the typical grid-based approach in spatial statistics. The whole research area of Germany is divided into a certain amount of grids, that buckets all coordinate data into different areas. Square is chosen as the polygon basis for geographical tessellation, as the coordinates of POIs can be conveniently

rounded off to the nearest grid centroids. A similar approach of grid segmentation has been widely used in different fields and works with dynamic data as well. Grid-based analysis also allows demand definition and data collection on specific areas [22]. Geohash is a notable example that splitting the whole world into grids of certain number of equal-size squares each time, which is a common method for weighted point sampling in geographic information systems [23].

The inherent difficulty for coordinate-based area segmentation, regardless of the tessellation polygon being used, is the non-linearity problem of mapping latitude and longitude onto the Earth, which is actually a sphere, so distance between consecutive longitude becomes smaller with increasing latitude. [24] However, as the latitude span of the research area Germany is limited to $[47.40, 54.90]$, the error by rounding off coordinates is minimal. In this study, the median latitude of Germany 51 degree has been taken as a reference, that the coordinates approximation and grids creation are based on that each degree latitude spans over the distance of $111km$ and each degree of longitude spans over the distance of $69km$. For this study, we have structured two levels of grids to serve different business objectives, namely 8km * 8km grids (big boxes) and 2km * 2km grids (small boxes). To ensure a geographically diverse solution over Germany, only one charging station would be placed per each big box area.

After identifying the areas with a lack of charging stations, the potential charging stations' locations are further narrowed down by dividing each big box area into a small box area of 2km * 2km. To achieve charging station locations suitable for long-distance traveling, only small boxes with a highway or federal road are considered viable for setting up a charging station. After the filtering, the best small box for each big box (given a lack of charging stations) can be obtained. A candidate location will be chosen within that best small box. After finding potential locations by box analysis, the demand level of each location is estimated by using a decision tree regressor in order to rank all potential candidate locations. As scenario analysis, the best charger type of each candidate location is also estimated.

B. Further Data Cleaning for Boxes Approach

Before implementing the boxes approach, the internal and external datasets are combined into a single dataset where the number of features within a box is counted. As some inconsistencies have been identified, further data cleaning has to take place, which will be explained in detail in the following.

Firstly, it was noticed that some coordinate groups do not contain traffic flow data because they do not have a flow counting point. In order to also obtain traffic flow data for areas that are not directly on, but close enough to a highway or federal road, the traffic flow amount is propagated by "copying" the flow observation points to nearby areas. For each of the flow points, four copies (at 2km direct north, east, south, east) are created, with an assumed traffic flow of 30 percent of the flow at the actual traffic flow point. As there are still some boxes without a flow point, the traffic flow in these areas is assumed to be at the lowest 10th percentile.

Secondly, it is observed that the count of some POIs contains very long right tails in the distribution, especially in the urban area where there are a large number of facilities within a square-bounded area. As one of the business objectives is to improve the coverage of E.ON charging stations, avoidance of overweighting urban areas has been done by setting a threshold to the count, which is the 90th percentile for the count data for each feature with some extreme, outlier counts of POIs. Whenever the count of a POI within the area exceeds the threshold, the count is truncated to the 90th percentile value only, as shown in figure 5.

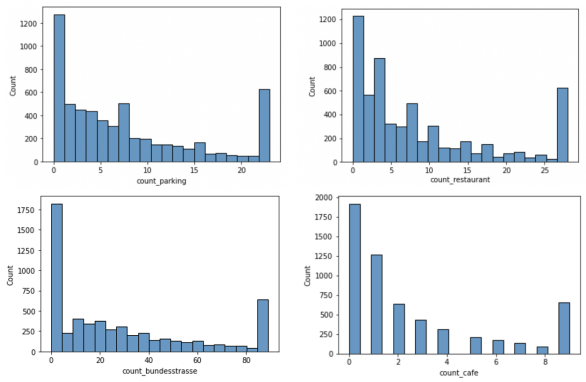

Fig. 5. Histograms of truncated counts of four POIs: parking spaces, restaurant, federal road network and cafes.

C. Big-small Boxes Approach

Fig. 6. Correlation between count of POIs among all big boxes

To start with the top-down, three-stage approach, we first identify the areas lacking charging stations using the "big-box" approach. A big box is defined by an 8km * 8km area. Therefore the whole research area, Germany, is segmented

into big boxes of 64 squared kilometers, which results in a total of 5,846 boxes, each with at least one point of interest or road network point. Then, the number of required charging stations in those areas is predicted by applying ordinary least squares linear (OLS) regression. This tries to answer the question that, given the number of points of interest and amount of traffic flow, how many stations a big box needs. The prediction is hence performed for each box. An obvious advantage of the linear regression model is its simplicity of interpretation for predicting a continuous response variable [25], and its disadvantage of being sensitive to outliers has been addressed by POIs count truncation explained previously.

For the POIs, we take into account the coordinates of food places (restaurants, cafés, and fast food shops), supermarkets, restrooms, parking lots and road network points of highways and federal roads. The occurrence of each category of POIs is counted for each box. The correlations among the counts for all boxes, as shown in figure 6, are all positive but not exceedingly large. The traffic flow data has been logarithmically transformed, as the data series becomes more normally distributed after the transformation. The distribution of traffic flow before and after the transformation is displayed in the figure 7. While the POIs and traffic flow serve as independent variables of the regression, the dependent variable comprises the number of DC charging stations. The number of DC charging stations consists of the number of E.ON's and competitions' charging stations and is counted for each box as well.

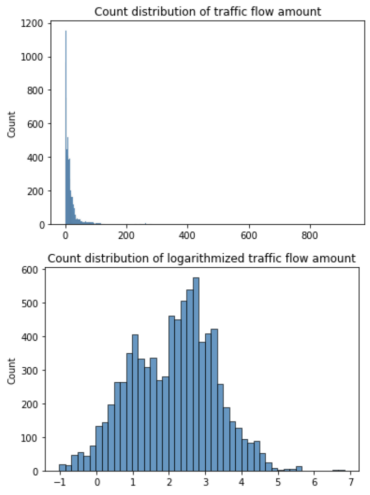

Fig. 7. Comparison of traffic flow data distribution before and after logarithmic transformation

After predicting the number of charging stations, we calculate the lack of charging stations for each big box (gap number of stations = estimated count - actual count). Therefore, the gap between the estimated number of stations and existing stations is computed. The prediction results are visualized in 8. It is evident that the areas lacking of charging stations are well distributed throughout Germany except for some gaps in the states of Sachsen and Sachsen-Anhalt, which are usually areas with low density of POIs. Nevertheless, this is reasonable as these regions have a lot of forests and natural reserve areas and therefore fewer points of interest. Furthermore, only very few highways go through that region. The boxes with a gap of charging stations higher than 0.2 are further taken into account as potential areas for new charging stations. Thereby, 1950 boxes out of the initial 5846 boxes remain.

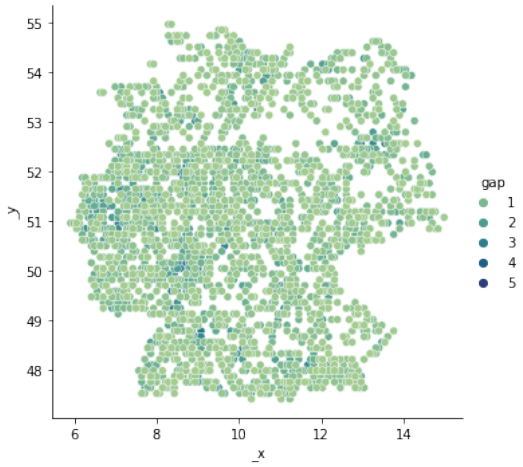

Fig. 8. Areas with estimated lack (gap) of charging stations

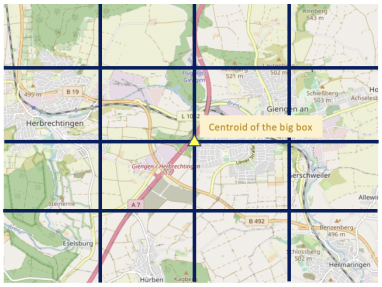

Fig. 9. Relationship between big and small boxes

For the small box analysis, we perform some data filtering processes to choose the best location of the charging station, assuming that only one new charging station will be placed per big box area. Each of those big boxes in the research area can be divided into at most 16 small boxes of 2 * 2 kilometers for location analysis, as shown in figure 9 where the centroid of a big box lies at the boundaries of some corresponding small boxes. Nevertheless, only small boxes with at least one highway or federal road network point are considered valid, to ensure that a new charging station can be placed next to a highway or federal road. Therefore, 19870 small boxes remain for further analysis within Germany. The best small box for each big box can then be estimated with help of logistic regression. The logistic regression tries to answer the

Count of stations	Count of small boxes	Percentage
0	17124	91.27%
1	1192	6.35%
2	290	1.55%
3 or above	156	0.83%

TABLE I
TABLE OF STATIONS COUNT OF SMALL BOXES

question that given the counts of POIs, where would a charging station likely be placed? Therefore, the predicted variable is binary on whether there is any DC charging station (E.ON or competitors) in a small box. Moreover, this alleviates the problem of class imbalance for counts of stations, in which there is a large portion of small boxes without a station as shown in table I.

The same categories of POIs used in the first regression are again applied and counted for each small box. Traffic flow has not been included in this model, as the number of flow observations per small box is not widespread and representative enough for analysis.

After performing the logistic regression, for any big box with a positive gap above 0.2, the best small box is determined. By predicting the likelihood of the existence of a charging station in the small boxes within a big box, the small box with the highest estimated probability is taken and considered as "best small box". A coordinate of a POI within the small box is randomly chosen as reference. By repeating this process for all best small boxes, a list of 1,950 candidates coordinates is obtained for further demand analysis. The actual positions of charging stations in real situation, however, would be more likely determined by other factors out of the scope of this project, such as electricity grid network, negotiation with local councils and the owners of the parking spaces.

D. Demand Prediction by Decision Tree

After selecting candidate points through box analysis, we predict the demand for all candidates. Since E.ON provided a daily aggregated data set and a monthly aggregated data set, it was necessary to select which data set to use. Daily aggregated data set has many data records, but a disadvantage is that it contains date-specific uniqueness, such as holiday effect or restrained traffic on the day due to car accidents nearby.

On the other hand, since the monthly data set is aggregated from the raw data on E.ON's side, it has fewer outliers, despite fewer data records considered. For a more accurate prediction, we utilize the monthly aggregated data to minimize the number of outliers. It is also crucial to select an accurate reference month to allow fair comparison across different calendar years while retaining as much information as possible. Data from the year 2022 is also a key criterion, as it shows the demand post Covid-19. In the dataset, only the complete monthly data from January 2022 to April 2022. Moreover, to minimize holiday effect with demand spikes for long-distance travelers, February is a good reference month when there are the least number of country-wide holidays, such as Easter or New year holiday. As the demand for charging has been steadily rising over the years due to the increasing adoption of electric vehicles, a dummy variable is created for each corresponding calendar year (except 2022 to avoid multi-collinearity), so that the year-to-year effect of the demand can be assessed, and subsequently adjusted for year 2022 in the demand prediction process.

As explained in section II, we have chosen the sum energy consumption as the main demand measure for this study, as it directly corresponds to the total demand. Therefore, the dependent variable in this decision tree model is the sum energy consumption as well. It is also important to select which independent variables to use for the model. In addition to the variables used in box analysis, the demand prediction requires the variables for the station itself provided by E.ON. The variable having the greatest influence on demand is charger max power, and it has a direct impact on infrastructure to be built at a potential candidate location. Therefore it is considered as the main parameter for demand prediction.

In addition, variables for the relationship between charging stations are also extracted using E.ON's data. It is possible to extract information on the demand of the nearest charging station at that time for each charging station, as the first connection of each existing charging station has been provided. Through this, it is possible to add mean sessions per month and sum energy consumption of the nearest active station as independent variables. In addition, the concept of point of interest used in box analysis was included. However, since box analysis targets a wider, boxed-area, it is not appropriate to utilize the variables the same way as in the box analysis. Therefore, the counts of POIs within a 500m radius, which a driver can walk to these POIs while charging, have been added as variables. On the other hand, the nearby charging stations where customers can drive to were counted at a 3km radius. The traffic flow of the nearest flow point was also considered, and since the range of the traffic flow is skewed, it was converted to log-scale before adding to the model.

In terms of model selection for demand estimation, an business objective is to have transparent models, that can conveniently tell the cause-and-effect of each variable. Examples of such simple models include linear regression, logistic regression and decision tree. Although other machine learning models such as XGboost, light GBM, and random forest have some obvious advantages, such as better prediction accuracy, the common difficulty on ensuring clearly explainable causal relationships outweighs any of these benefits, especially on the users' side. Furthermore, logistic regression model, as a binary classification model, cannot be

utilized for demand prediction to predict continuous values. Thus, simple regression models (including linear regression and polynomial regression) and decision tree model are the only two available choices.

The simple regression model is good for clearly showing the correlation between the dependent variable and the independent variables with the p-values, and for explaining simple data. However, this means that the more complex the structure of the data, the more difficult it is to produce good performance. In addition, since only a simple correlation between the dependent variable and the independent variable is used, it means that it is difficult to reflect the interaction effect between the independent variable. In the case of the polynomial regression, interaction between independent variables is also considered via simple correlations, and the model tends to generate too many variables, resulting in multicollinearity.

The decision tree regressor is a model predicting continuous variable using the decision tree model, which is a classification model discritizing all continuous variable. Accordingly, an obvious disadvantage is that the exact value cannot be predicted and only the section in which the dependent variable belongs can be known. The deeper the maximum depth of the decision tree, the more granular the dependent variable is. The predictive performance for the training dataset would improve, but subject to the risk of overfitting. On the other hand, if the maximum depth is too less, the model may underfit. Therefore, it is important to set an appropriate maximum depth. Nevertheless, decision tree has many advantages, such as considering various relationships between independent variables, and over-performing other simple regression models if the correct max depth is set. Therefore, it is appropriate to use a decision tree to recommend candidate maximum charger power, which will be described in detail in the result part. In our study, maximum depth of the decision tree regressor model is 7, the train-validation data split is 8:2.

V. Evaluation of Results

For the big-box analysis as shown in figure 10, based on 5846 observations (number of big boxes with at least one POI), the OLS regression model has achieved an overall R-squared value of 0.369. An adjusted R-squared of 0.368 indicates a proper amount of independent variables used in the model.

Most of the independent variables used in the model are considered significant based on the p-values, especially for road network-related variables that each has p-values of less than 0.001. In comparison, count of restaurants and cafes are considered not significant. By predicting number of stations needed per big box, and subsequently calculating the gap between actual and predicted number of stations (estimation error), one-third of big boxes have a positive gap of above 0.2, which signalizes that these areas are possibly under-served.

OLS Regression Results

Dep. Variable:	count_stations	**R-squared:**	0.369
Model:	OLS	**Adj. R-squared:**	0.368
Method:	Least Squares	**F-statistic:**	379.5
Date:	Mon, 08 Aug 2022	**Prob (F-statistic):**	0.00
Time:	07:51:31	**Log-Likelihood:**	-9698.6
No. Observations:	5846	**AIC:**	1.942e+04
Df Residuals:	5836	**BIC:**	1.948e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.2837	0.029	-9.891	0.000	-0.340	-0.227
count_cafe	0.0047	0.011	0.410	0.682	-0.018	0.027
count_fastfood	0.0473	0.009	5.100	0.000	0.029	0.065
count_parking	0.0103	0.003	3.084	0.002	0.004	0.017
count_restaurant	-0.0014	0.004	-0.313	0.755	-0.010	0.007
count_supermarket	0.0766	0.013	6.037	0.000	0.052	0.101
count_toilet	0.0682	0.010	7.091	0.000	0.049	0.087
count_bundesstrasse	0.0034	0.001	4.976	0.000	0.002	0.005
count_autobahn	0.1346	0.006	22.909	0.000	0.123	0.146
log_flow_amount	0.0638	0.017	3.784	0.000	0.031	0.097

Omnibus:	7149.069	**Durbin-Watson:**	1.543
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	2052155.444
Skew:	6.276	**Prob(JB):**	0.00
Kurtosis:	93.925	**Cond. No.**	78.0

Fig. 10. OLS regression result for the big-box analysis

For small-box analysis as shown in figure 11, the logistic regression has achieved a pseudo-R value of 0.1606, but most of the independent variables are considered significant individually. Similar to big-box analysis, road network-related variables are more significant than the POI-related variables.

Logit Regression Results

Dep. Variable:	binary_stations	**No. Observations:**	19870
Model:	Logit	**Df Residuals:**	19861
Method:	MLE	**Df Model:**	8
Date:	Mon, 08 Aug 2022	**Pseudo R-squ.:**	0.1606
Time:	07:51:34	**Log-Likelihood:**	-4800.6
converged:	True	**LL-Null:**	-5719.0
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.3132	0.042	-79.054	0.000	-3.395	-3.231
count_cafe	-0.0837	0.016	-5.364	0.000	-0.114	-0.053
count_fastfood	0.0712	0.012	5.895	0.000	0.048	0.095
count_parking	0.0456	0.008	5.745	0.000	0.030	0.061
count_restaurant	0.0023	0.008	0.292	0.770	-0.013	0.017
count_supermarket	0.1335	0.021	6.258	0.000	0.092	0.175
count_toilet	0.1418	0.019	7.647	0.000	0.105	0.178
count_bundesstrasse	0.0246	0.002	13.460	0.000	0.021	0.028
count_autobahn	0.4198	0.019	21.605	0.000	0.382	0.458

Fig. 11. Logistic regression result for small-box (2km*2km) analysis

The decision tree regressor model has a R-squared value of 0.8038, which significantly outperforms any linear regression model or polynomial regression models with a maximum R-squared value of 0.35. Decision tree regressor model also performs a lot better than simple regression models as the model copes better with the complexity of data.

Moreover, checking RMSE value of the validation data can also avoid the results being overfitted. RMSE value for the decision tree regressor model is 917,250, which is

slightly lower than 948294 from the simple regression model. However, the decision tree model is considered slightly overfitted, as the improvement was not huge. However, the decision tree is still a reasonable model for suggesting max charger power of the potential charging stations. As the result from regression denotes that max charger power and dependent variable has a negative correlation, it will always suggest the station with the lowest max charger power if a linear regression model is used. However, in a decision tree model, it was possible to get other max charger powers as recommendations by capturing non-linear characteristics of the data.

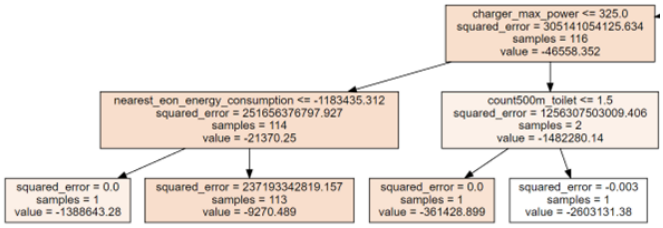

Fig. 12. Part of Decision Tree Result Image

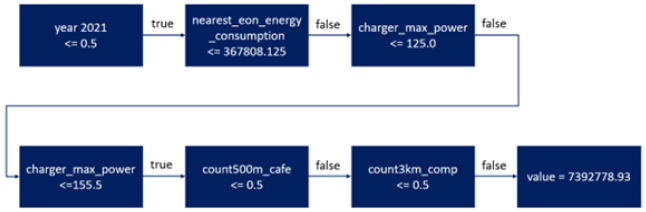

Fig. 13. Example of Decision Tree Decision-Making Process

Decision tree regressor model can also generate an image for its decision-making process. Figure 12 shows a part of the decision tree model result. When the branch expands to right it means true and if it expands to left it means false to the condition from the box. Therefore, users can easily identify how the model made the decision. Figure 13 is an example of decision-making process of the model. For the stations data, which are not from 2021, have nearest E.ON station which has higher than 367,808.125 of energy consumption. And if they have max charger power between 125 and 155.5, have cafe and competitors' station nearby will have energy consumption of 7392278.93.

The decision tree model also calculates the importance of features that are used in the model. The feature importance of our decision tree regressor model is denoted in table II. In line with E.ON on the internal data set, it is noted that the feature importance of charger max power is the highest, at 43.68%. The importance of the year 2021 was the second highest, although the importance for other years was low. However, removing other variables are not viable, as we can predict the demand for year 2022 by combining these Boolean values. Nearest E.ON station's sum energy consumption and session counts per month are also meaningful to the model, as expected. Moreover, traffic flow has 7.43% of importance, showing a clear impact on the model. For the count POIs within radius, cafes and charging stations are more meaningful than other POIs such as fast-food shops and restaurants.

Feature	**Feature Importance (%)**
Charger Max Power	43.68
Is Year 2021	13.07
Nearest E.ON: Sum Energy Consumption	10.71
Nearest E.ON: Mean Sessions	7.63
Traffic Flow (log value)	7.43
Count 500m: Cafe	6.45
Count 3km: Competitor Stations	3.23
Count 3km: E.ON Stations	2.97
Count 500m: Supermarket	1.70
Count 500m: Toilet	1.39
Count 500m: Parking	1.18
Count 500m: Fast-food	0.39
Count 500m: Restaurant	0.18
Is Year 2018	0.00
Is Year 2019	0.00
Is Year 2020	0.00

TABLE II
TABLE FOR FEATURE IMPORTANCE

VI. DEPLOYMENT

A. Visualization by Tableau Dashboard

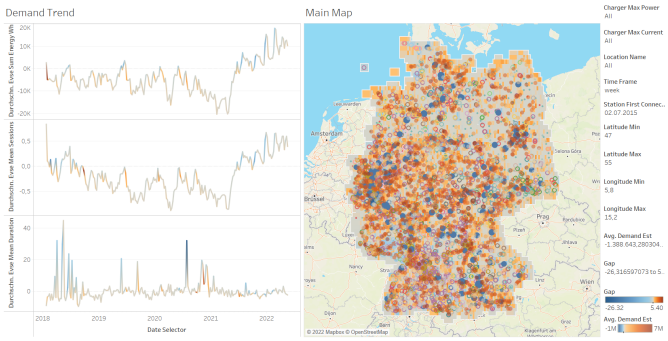

Fig. 14. Overview of Tableau Dashboard

Tableau is a business intelligence tool currently to visualize data analytics results via dashboarding. It is capable of making machine learning, smart data preparation and statistical analysis, making it a popular choice for many companies and institutions [26]. For our study, we have created a dashboard to project our results onto the map data and highlighted key statistics. The dashboard is divided into the demand trend part on the left and the main map on the right. These two parts are also configured to interact with each other through various dashboard actions.

The demand trend part shows the current change in demand for E.ON's charging stations. This demand trend part consists of three line charts, namely sum energy consumption at the top, mean sessions at the middle, and mean duration at the bottom. These three graphs can be further adjusted by the time frame parameters (from date-to date), as well as other parameters such as duration of Day, Week, Month, Quarter and Year, which these charts will show the aggregated values accordingly. In addition, users can check the trends of different demand measures by filtering parameters Max Charger Power, Max Charger Current and Location Name on the right. The demand measures of a particular E.ON station can be checked by tapping the corresponding blue circle on the map.

The main map on the right also contains various kinds of markers. Tableau supports the use of multiple layers when using geographic data, which we use to map the locations in the input data and the locations corresponding to the analysis results. Here is a list of layers included in the map:

- Charging stations: E.ON's stations in blue dots, competitors' stations in hollow circles, where different competitors are shown in different colours.
- Traffic volume observation points: red crosses for highway and blue-green crosses for federal roads.
- Gap number of stations according to big-box analysis: positive gap in orange and negative gap in blue. The stronger the tendency, the deeper the colour.
- Candidate locations: orange dots. As these locations are already selected by box analysis, only some predicted demands are negative. Points with deeper orange color on the dashboard represent a higher predicted demand.

Each of these map layers can be turned on and off individually on Tableau, such that a user can choose the markers needed in the main map. The tool tip of each layer is set to represent the necessary values separately, so that each item can also be shown separately or together according to the user's choice.

This dashboard can satisfy the need of the users in different use cases, due to various adjustable parameters, filters and layers. For example, if a user would like to find a big box with a large gap, the user can adjust the filter of the gap number of stations, so that only some areas within a specific demand range would be shown. Alternatively, to check the candidate points with the highest demand estimation, the user can set the range of the demand estimation filter to 7,000,000 or higher. A user can also zoom into a specific area for regional analysis by adjusting the latitude and longitude ranges. Therefore, this Tableau takes advantage of sophisticated statistical analysis, while maintaining the ease of usage of different cases.

B. Suggestions of Charging Stations and Business Insights

Having explained the model approach in detail and evaluated its performance in the previous chapters, this section now focuses on individual model outputs and discusses their meaningfulness with regard to the required business case.

The left part of figure 15 presents a screenshot of the results from Tableau, after filtering the boxes with a large station count gap (higher or equal 5.397) and those candidates with large demand estimation (higher or equal 7,392,779). While the filtered candidates with high demand, whose values were determined using the decision tree regressor, are also located in rural regions, the filtered boxes with high gap are mostly located near large cities. However, it can also be seen that the location of the boxes and candidate points often overlap, leading to comparable results, as the gap is still positive for rural candidates. Looking at the right side of figure 15, the recommended locations with a large gap of charging stations in Frankfurt am Main are close to the city, but not in the center. This example result shows that the model is orientated towards the existence of POIs, a low number of competitors and a high traffic flow, but at the same time makes sure that a highway or federal road is nearby, so that a long distance traveler does not have to drive into the city center to charge his or her car.

In summary, the locations proposed by our model for new DC charging stations appear to make sense not only at first look, but also at second glance, as all the required factors of the business case have been considered and met in most of the suggestions.

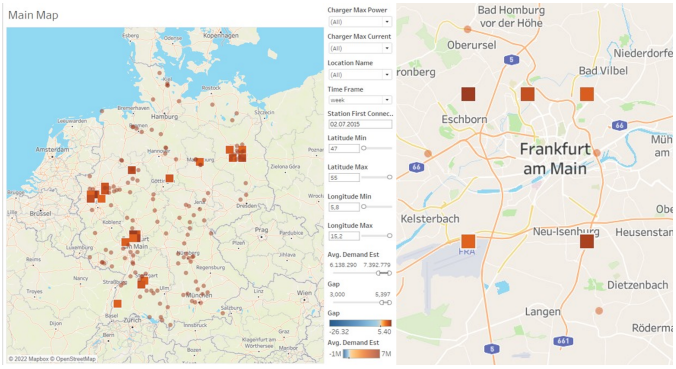

Fig. 15. Tableau Visualization for High Demand Candidates

VII. Conclusion

In this paper, a three-stage machine learning approach has been presented in order to determine good charging station locations in Germany from a charging station operator's view. The main focus is on incorporating geographical data (e.g points of interest and competitors) with real charging station demand from E.ON. The suggested new charging station

locations have been based on a threefold strategy: firstly estimating the gap of charging stations needed per 8km*8km area (big box) by OLS regression, secondly deciding the best 2km*2km area (small box) to place a charging station for each big box with positive gap by logistic regression, and thirdly estimating the demand of the charging stations and best charger type by decision tree analysis.

This approach can be further elaborated in the future. From the data perspective, additional features such as population structure of a region could have an influence on the machine learning model. Moreover, the quality of data can be improved by further data cleaning, for instance a more equally distributed distribution of road network points, as the current data extraction method will generate more network points given more curves on the road network. Traffic flow data can also be more holistically analyzed if data with more granularity, such as flow during different time of a day, can be obtained from the authority. Other information such as the trip origin-destination pairs of long-distance drivers would also improve the demand understanding.

Besides improving the data, the demand determination can also be further developed. The current model only suggests placing one charging station in a box. If the box has a large predicted gap of stations, signaling more underserved charging demand, more several charging stations could potentially be placed to serve the area. However, this will also induce the problem of suggested charging stations being too close to each other. Moreover, the decision tree model predicts discrete demand values. Continuous demand value would be a more suitable choice for profit estimation, given the true demand is known. However as E.ON's demand measures have been masked, binning demand into discrete values is still a viable option.

It is also essential to note that the actual installation of a new charging station is subject to more factors, such as preparedness of power infrastructure and connection, which these considerations are out of the scope of this project. Legal conditions and regulations, as well as the ownership of amenities and parking spaces would also play an important role in the actual location of a charging station.

After all, the ramp-up strategy of E.ON is also crucial for optimizing charging station locations for utilizing the network effects. As many competitors of E.ON have the similar ambition of building more charging stations in a fast pace, the desired density of stations and type of charging stations to be built depend significantly on the overall market situation and competitors' strategies, which are highly dynamic. Therefore, a holistic understanding of any newly obtained data and prompt update of the suitable data to the model would have an instrumental effect in supporting E.ON strategy.

REFERENCES

[1] E. E. Agency, "Greenhouse gas emissions from transport in europe." https://www.eea.europa.eu/data-and-maps/indicators/transport-emissions-of-greenhouse-gases/transport-emissions-of-greenhouse-gases-12. Accessed: 2022-08-08.

[2] W. of federal government, "Mehr ladestationen für elektroautos.." https://www.bundesregierung.de/breg-de/themen/klimaschutz/ladeinfrakstruktur-1692644), note = Accessed: 2022-08-08.

[3] H. Seitz, M. Hofmann, and M. Günther, "E-mobility check: Wie bereit ist deutschland?." https://www.strategyand.pwc.com/de/de/industrie-teams/automobil/e-mobility-check/strategyand-emobility-check.pdf), note = Accessed: 2022-08-08.

[4] 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," *International journal of energy research*, vol. 42, no. 9, pp. 3102–3120, 2018.

[5] S. Wagner, M. Götzinger, and D. Neumann, "Optimal location of charging stations in smart cities: A points of interest based approach," 2013.

[6] H. Wu and D. Niu, "Study on influence factors of electric vehicles charging station location based on ism and fmicmac," *Sustainability*, vol. 9, no. 4, p. 484, 2017.

[7] C. Karolemeas, S. Tsigdinos, P. G. Tzouras, A. Nikitas, and E. Bakogiannis, "Determining electric vehicle charging station location suitability: A qualitative study of greek stakeholders employing thematic analysis and analytical hierarchy process," *Sustainability*, vol. 13, no. 4, p. 2298, 2021.

[8] F. Ahmad, A. Iqbal, I. Ashraf, M. Marzband, *et al.*, "Optimal location of electric vehicle charging station and its impact on distribution network: A review," *Energy Reports*, vol. 8, pp. 2314–2333, 2022.

[9] B. Csonka and C. Csiszár, "Determination of charging infrastructure location for electric vehicles," *Transportation Research Procedia*, vol. 27, pp. 768–775, 2017.

[10] A. Ip, S. Fong, and E. Liu, "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)*, pp. 460–465, IEEE, 2010.

[11] T. D. Chen, K. M. Kockelman, and M. Khan, "Locating electric vehicle charging stations: Parking-based assignment method for seattle, washington," *Transportation research record*, vol. 2385, no. 1, pp. 28–36, 2013.

[12] Y. He, K. M. Kockelman, and K. A. Perrine, "Optimal locations of us fast charging stations for long-distance trip completion by battery electric vehicles," *Journal of cleaner production*, vol. 214, pp. 452–461, 2019.

[13] A. Zhang, J. E. Kang, and C. Kwon, "Incorporating demand dynamics in multi-period capacitated fast-charging location planning for electric vehicles," *Transportation Research Part B: Methodological*, vol. 103, pp. 5–29, 2017.

[14] D. Mao, J. Tan, and J. Wang, "Location planning of pev fast charging station: an integrated approach under traffic and power grid requirements," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 1, pp. 483–492, 2020.

[15] "E.on annual report for year 2021." https://www.eon.com/en/investor-relations/financial-publications/annual-report.html. Accessed: 2022-08-08.

[16] H. Lee and A. Clark, "Charging the future: Challenges and opportunities for electric vehicle adoption," 2018.

[17] R. Wolbertus and R. Van den Hoed, "Electric vehicle fast charging needs in cities and along corridors," *World Electric Vehicle Journal*, vol. 10, no. 2, p. 45, 2019.

[18] F. N. Agency, "List of all charging stations in germany." https://www.bundesnetzagentur.de/DE/Fachthemen/ElektrizitaetundGas/E-Mobilitaet/Ladesaeulenkarte/start.html;jsessionid=AEE11D653BA9AC942A384641E4C39D0C. Accessed: 2022-08-07.

[19] OpenStreetMap, "List of parking lots in germany." https://wiki.openstreetmap.org/wiki/DE:Tag:amenity%3Dparking. Accessed: 2022-08-07.

[20] F. H. R. Institute, "List containing the numbers of counted vehicles per counting point." https://www.bast.de/DE/Verkehrstechnik/Fachthemen/v2-verkehrszaehlung/Aktuell/zaehl_aktuell_node.html;jsessionid=

16FFE77EB1680ACFC59D27C62BA84F87.live11293?cms_map=1&cms_filter=true&cms_jahr=Jawe2021&cms_land=&cms_strTyp=A&cms_str=&cms_dtvKfz=&cms_dtvSv=. Accessed: 2022-08-07.

[21] Autozeitung, "List of charging duration depending on vehicle." https://www.autozeitung.de/ladezeiten-e-autos-199920.html. Accessed: 2022-08-09.

[22] A.-H. Sato, S. Nishimura, N. Makita, T. Namiki, and H. Tsubaki, "World grid square statistics and their application to data analytics,"

[23] K. Sims, G. Thakur, K. Sparks, M. Urban, A. Rose, and R. Stewart, "Dynamically-spaced geo-grid segmentation for weighted point sampling on a polygon map layer," tech. rep., Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States), 2018.

[24] "Geohash: Limitations when used for deciding proximity." https://en.wikipedia.org/wiki/Geohash. Accessed: 2022-08-08.

[25] A. Ostermann, Y. Fabel, K. Ouan, and H. Koo, "Forecasting charging point occupancy using supervised learning algorithms," *Energies*, vol. 15, no. 9, p. 3409, 2022.

[26] "What is tableau?." https://www.tableau.com/why-tableau/what-is-tableau. Accessed: 2022-08-09.