Forecasting Peak Load Profiles for New High-Power Charging Stations

A K-Means Clustering and Random Forest Approach

Markus Gerholm

Master's thesis
2026:E58

Lund University

Faculty of Science
Centre for Mathematical Sciences
Mathematical Statistics

Centrum Scientiarum Mathematicarum

Master's Theses in Mathematical Sciences 2026:E58
ISSN 1404-6342

LUNFMS-3144-2026

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lu.se/

Abstract

The increasing demand for electric vehicle charging stations creates new challenges for electric grid planning. This thesis proposes a framework for forecasting realistic 24-hour peak load profiles for new direct current fast-charging stations without historical observations. The method combines K-means clustering of existing stations' hourly load patterns and random forest models that predict both the daily load shape and peak load magnitude of a new station. The resulting profile is scaled to create a low-risk predicted load curve and evaluated using leave-one-out cross-validation. The results show that the proposed profiles cover the majority of observed peak loads while remaining below theoretical maximum capacity for much of the day. Small charging stations are more difficult to predict, as they can more easily approach full capacity. The framework is designed to support grid planners in making more efficient and risk-aware decisions when assessing new charging station connections.

Popular scientific summary

Electric cars have gone from a futuristic idea to an everyday occurrence. Nowadays, it is hard to go about your day without seeing them charging their batteries. Shopping centers, gas stations and grocery stores, are but a few examples of where charging stations can be found. What happens when more and more people buy electric cars instead of fossil fuel cars? The need for charging stations increases. This increase in demand then brings the risk of straining the electric grid that provides the power to the charging stations. This thesis provides tools for grid planners, so that they can investigate whether a new charging station should be built with the current existing infrastructure or not.

Figure 1: Photo of electric vehicle charger: Kristoffer Andersson / Öresundskraft AB.

A safe approach to this problem, would be to assume that all charging stations are fully occupied and charge at maximum power, at all hours during the day, this is called the theoretical maximum capacity. When examining historical charging data, we see that this assumption is far from reality. The data consists of hourly measurements of power consumption from Direct Current fast-charging stations within Öresundskraft AB's electric grid. Öresundskraft AB is an energy company based in Helsingborg, southwest Sweden. The aim of this thesis, is to create more realistic peak charging station profiles, so that the electric grid can be used in a more efficient way.

The proposed method separates the problem into two parts: predicting the daily shape

of charging demand, and predicting how high the daily peak load may become. These two are then combined, in order to form a more realistic charging station profile. In order to predict a new charging station's daily shape, existing stations are grouped according to their charging patterns. Examples of charging patterns could be if a station is more popular during weekends or weekdays. A machine learning model is then used to predict the new charging station's most likely group charging pattern. A similar machine learning model is also used to predict the new charging station's daily peak load.

The results show that low-risk predicted load profiles usually cover most observed peak loads while still staying below the station's theoretical maximum capacity for much of the day. This means that the method can help grid planners use the electricity grid more efficiently than if they always assume full power at every hour. There is a clear trade-off: more cautious predictions reduce the risk of underestimating demand, while less cautious predictions allow more grid capacity to be used. Small charging stations, with only a few charging points, were harder to predict because they can easily come close to full capacity. The framework can be used as decision support when assessing whether a new fast-charging station can be connected to the grid or whether reinforcements may be needed.

Acknowledgements

This thesis was made possible by Öresundskraft AB. Öresundskraft AB provided the problem statement, as well as necessary software and permissions to data. The supervision of this study was conducted by Öresundskraft AB and Lund university, where the former handled more practical questions, and the latter handled more theoretical questions.

Thank you, August Millqvist, for all supervision regarding Öresundskraft AB, as well as teaching me the basics of electricity and the grid planners way of working.

Thank you, Erik Lindström, for all supervision from Lund university, especially for coming up with new ideas and identifying patterns when I could not.

Contents

Chapter 1

Introduction

On a global scale, electric vehicle (EV) adoption has accelerated rapidly. During 2024, 17 million electric cars were sold, this constitutes one fifth of all car sales during that year (International Energy Agency, 2025). To highlight the growth in EV adoption, International Energy Agency (2025) states that the increase in EV sales from 2023 to 2024 was 3.5 million cars, which surpasses the total EV sales during 2020. These numbers are expected to increase during 2025, where the number of EV sales are forecasted to surpass 20 million. In Europe, the share of sales consisting of electric cars in 2024, is similar to the global share, namely around 20%. Sweden has been among the leading European countries in vehicle electrification. According to Trafikanalys (2024), Sweden was an early adopter in the electrification of its vehicle fleet and, as of September 2024, had the third highest share of newly registered electric vehicles in Europe.

As the number of electric vehicles continue to increase, the need for new charging stations increase as well. In order to evaluate whether the existing electric grid has the capacity for a new charging station or not, grid planners need models that can predict future peak power consumption from a new charging station for each hour of the day. Data-driven methods that produce 24-hour peak load profiles, can be used to investigate if the existing grid can handle a new charging station's expected peak loads, or if it needs to be reinforced before approving the construction of the new charging station. Öresundskraft AB is a Swedish energy company based in southwestern Sweden. The company has four distribution areas for electricity in northwestern Skåne, including Ängelholm. Within the areas, sensors gather data that is stored, this data can then be utilized for data-driven models that assist grid planners in making more informed decisions. The data used in this study, consists of measurements from electric vehicle Direct Current (DC) fast-charging stations within Öresundskraft AB's electric grid. The locations of the charging stations are shown in Figure 1.1, below.

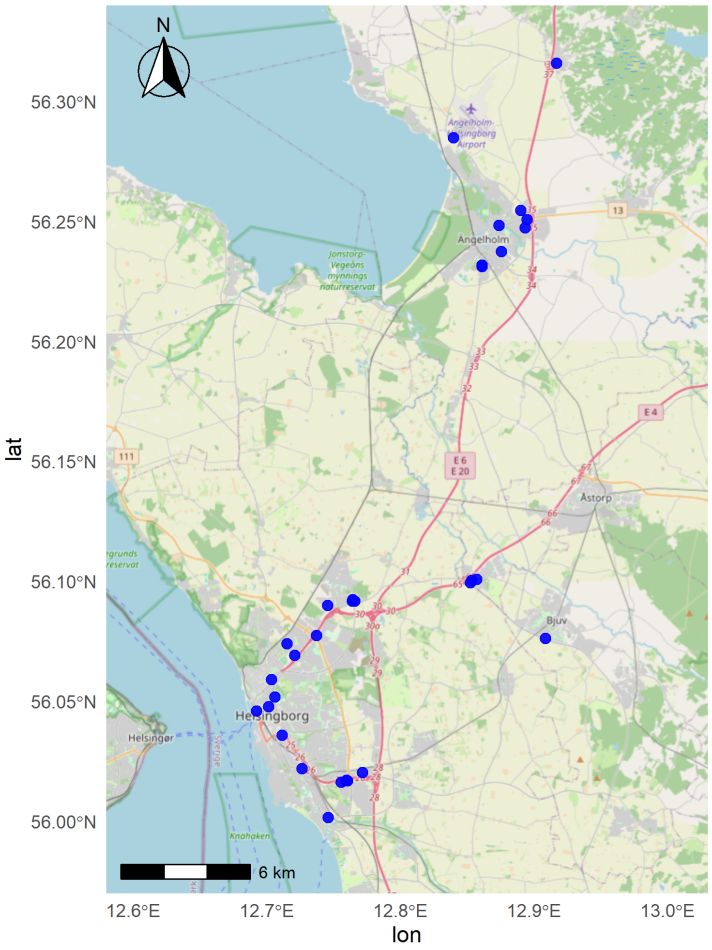

Figure 1.1: Map of charging station locations, where each blue dot represents one Direct Current charging station in Öresundskraft AB's electric grid, that is used within this study. Map constructed via Dunnington (2025).

Before stating the objective of this thesis in greater detail, previous studies within the electric vehicle research area are investigated during the next section.

1.1 Literature review

The modeling choices of electric vehicle charging, demand and grid impact can be distinguished by the data that is available, the assumptions of the underlying probability distributions and what objective the final model aims to achieve. In Lin et al. (2025), the modeling choices are split into three main branches:

1. Statistical models, such as linear regression, time-series models and generalized additive models.

2. Simulation methods, such as Monte Carlo simulation and scenario-based grid impact studies.
3. Data-driven models, such as random forests, recurrent neural networks and long short-term memory networks.

Statistical models. These models typically utilize variables such as arrival time, departure time and historical observations to estimate parameters of underlying probability distributions that describe dependencies between the variables. The estimated parameters are then used to form predictions of future EV charging. Typical models are time-series models such as the Autoregressive Integrated Moving Average Model (ARIMA) and the Trigonometric, Box-Cox transform, ARMA errors, Trend and Seasonal components model (TBATS) (De Livera et al., 2011; Lin et al., 2025). In Kim and Kim (2021), both previously mentioned models and more are used to predict charging demand of EVs on a national level, city level and single station level. Further research has been conducted regarding semi-parametric models. These models fall under the statistical branch but require less strict assumptions on dependency structure in data.

Semi-parametric models are more flexible than pure parametric models in that they allow for parametric components with fixed coefficients and non-parametric components. A reoccurring example in the literature is the Generalized Additive Model (GAM), which combines both frameworks by allowing the response variable to depend linearly on the feature variables in a parametric form, while also allowing other features to affect the response variable non-linearly through smooth functions such as splines (Wood, 2006). In Amara-Ouali et al. (2025), electric vehicle charging loads are predicted for different time horizons and with different types of models. Two of them are extensions of the previously mentioned GAM model. In the study, linear components include features such as relative humidity, wind speed and temperature, while non-linear components consist of calendar variables such as the time of day. The study finds that the extended Quantile GAM model predicts future quantiles in a more efficient way than the other models tested. Future demand for existing charging stations is however, not the only quantity that is of interest in the EV research area, the previously mentioned simulation based methods, are often relevant to broader scenarios and aims to help infrastructure planning.

Simulation methods. Studies within this branch often aim to evaluate grid impact for possible electrification scenarios. In Callanan et al. (2025), the potential grid impact is evaluated based on a simulated scenario of a car fleet consisting of only electric vehicles in the region of Skåne, southern Sweden. Data from travel statistics, charge usage and demographic data are used to construct synthetic load profiles that then are utilized to simulate grid impact. The study shows that aggregated EV charging may create substantial congestion, and that smart charging reduces the risk of overload during peak hours. Another simulation approach is presented in Betancur et al. (2021), where projected scenarios for year 2030 in Antioquia, Columbia is evaluated. The study differs from the previously mentioned study since available data is more limited, another difference is that the study simulates the impact on the grid based on a time horizon, while the previous study assumes a completely electrified car fleet. The study concludes that charging stations are needed in public areas in order to reduce future peak load hours, and that the capacity of the grid is sufficient and does not need further expansion. The commonalities of the mentioned studies is that they measure potential future impact and

risk on a large scale. Returning to the problem of estimating charging station demand, data-driven models offer an alternative to the previously mentioned statistical models.

Data-driven models. The models within this branch typically refer to machine learning models where the interpretability of statistical models are traded for relaxed assumptions regarding data. In Chang et al. (2021), aggregated fast-charging power demand is predicted by training a Long Short-Term Memory (LSTM) deep-learning model. The benefit of the deep-learning models is that there are no linearity assumptions on the data, and the model typically learns more complex behavior than a typical statistical model. The model is based on historical observations and learns patterns across time by investigating several charging stations' aggregated observations. The resulting model outperforms several other deep-learning models such as Recurrent Neural Network (RNN) models and Gated Recurrent Unit (GRU) models. The investigation of new charging station behavior is a relatively unexplored part of the EV research. The largest obstacle within this application is that no historical observations can be leveraged in order to form future demand predictions. In Zheng et al. (2026), this problem is approached by utilizing neighboring stations' observations as time-series variables that are used within a Convolutional Neural Network (CNN) model. The study predicts and models the number of charging orders of a new charging station based on neighboring charging stations and other features, such as price and temperature. Except for this study, no other research articles were found that covered new charging station behavior, especially in the application of grid planning. This implies that there is further need for frameworks that tackle the challenge of predicting new charging station behavior in many applications. This thesis aims to propose a framework where a new charging station's hourly shape and peak load can be estimated in order to further utilize an electric grid's capacity, as well as predict the size of expected peaks when building a new charging station.

1.2 Objective

As the previous section entails, there is a vast amount of models to choose from depending on the application of the model. The objective of this thesis has been the result of cooperation between the author and the Helsingborg based energy company Öresundskraft AB. Öresundskraft AB has provided necessary data from sensors in their electrical grid which has made the existence of this thesis possible. The purpose of this cooperation springs from the need for data-driven tools to better utilize the grid. Grid planners need sophisticated methods based on real measurements to estimate future peak loads for new Direct Current (DC) fast-charging stations. The suggested framework needs to be computationally efficient and dependent on solely static charging station metadata, such as *number of charging points, total capacity* and *geographical location*. For a newly installed charging station, there exists no historical measurements. Because of these requirements, a reasonable approach is a shallow data-driven framework that generates load profiles based on the information that the grid planner will have at an early planning stage. The models in this category are fast to implement and depending on model specification, can be trained to output peak load curves with the available information as input. The objective of this thesis is then to create a framework that generates a low-risk load profile that captures both shape and scale of peak loads of electric vehicle DC fast-charging to mitigate risk while still utilizing the electrical grid more efficiently than the current way of work. Similar approaches has been made, as seen

in Section 1.1, however, the vast majority of studies conducted in this area deals with either future demand forecasting of existing charging stations, or the creation of generic load profiles based on aggregated data from many sources. This thesis stands out in that the sole focus is to predict station-specific peak load profiles for new charging stations that has not yet been built, so that grid planners can evaluate the effect of approving the construction of a new charging station with certain meta-data features.

Chapter 2

Theory

Before proceeding to present the data material and chosen modeling steps, the necessary theory needed for understanding the proposed framework is presented in the same order as the implementation is presented in the subsequent Method chapter (see Chapter 3).

2.1 Hidden Markov Models

Hidden Markov models (HMMs) provide a probabilistic framework for modeling time series whose observed behavior is governed by an underlying sequence of unobserved states. Each hidden state corresponds to a latent regime with its own statistical properties, while the observed signal is treated as a realization generated from the active regime at each point in time. This makes HMMs useful in applications where a measured signal may switch between different regimes, and where the objective is to estimate both the hidden state sequence and the corresponding state-dependent distributions from the observed data. In this thesis, HMMs are only used in the context of estimating two hidden states for a given time series, which is why the theory from Rabiner (1989) will be presented in this context as well, the reason for this is motivated in the next chapter during Section 3.3.

2.1.1 Two-state Hidden Markov Model

In an HMM, the hidden state process is assumed to satisfy the Markov property, meaning that the current state depends only on the previous state, while each observation depends only on the current hidden state. The model then, consists of two stochastic processes: the unobserved state sequence and the observed sequence generated from it. All theory and notation during this subsection regarding HMMs springs from Rabiner (1989).

In the two-state case, let the hidden state at time t be denoted by

$$q_t \in \{S_1, S_2\}, \qquad t = 1, 2, \ldots, T.$$

Where S_1 and S_2 represent the two possible latent regimes. The observed sequence is denoted by

$$O = (O_1, O_2, \ldots, O_T).$$

Following the standard HMM formulation, the model is completely specified by three sets of parameters:

$$\lambda = (A, B, \pi),$$

where A is the state transition probability matrix, B is the observation probability model and π is the initial state distribution.

The transition probability matrix for a two-state HMM is

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix},$$

where

$$a_{ij} = P(q_{t+1} = S_j \mid q_t = S_i), \qquad i, j \in \{1, 2\}.$$

These probabilities describe how likely the model is to remain in the same state or switch to the other state from one time step to the next. Since they are probabilities, each row of A sums to one:

$$a_{11} + a_{12} = 1, \qquad a_{21} + a_{22} = 1.$$

The initial state probabilities are given by

$$\pi_i = P(q_1 = S_i), \qquad i = 1, 2,$$

with

$$\pi_1 + \pi_2 = 1.$$

These specify the probability that the hidden process starts in each of the two states.

The observation model specifies how the observed data are generated in each state. In Rabiner's formulation, this is represented by a set of state-dependent observation probabilities

$$B = \{b_j(O_t)\},$$

where

$$b_j(O_t) = P(O_t \mid q_t = S_j), \qquad j = 1, 2.$$

Therefore, conditional on the current hidden state, the observation at time t is generated according to the distribution associated with that state. In this way, each hidden state corresponds to a distinct statistical regime for the observed process.

For a two-state HMM, the probability of a particular hidden state sequence

$$Q = (q_1, q_2, \ldots, q_T)$$

is determined by the initial probabilities and the transition probabilities, while the probability of the observed sequence depends on the observation model. Although hidden Markov models are often introduced for discrete observation sequences, the framework also extends to continuous-valued observations by replacing discrete observation probabilities with probability density functions.

2.1.2 Viterbi Algorithm

After fitting a hidden Markov model, a natural question that arises is how to obtain the most likely hidden state sequence. That is, given the observed sequence together with the estimated transition probabilities and estimated parameters of the state-dependent probability density functions, what is the most probable sequence of the hidden states? The Viterbi algorithm, presented in Viterbi (1967), answers this question through dynamic programming. All theory during this subsection springs from the original source (Viterbi, 1967).

Let the observed sequence be denoted by

$$O = (O_1, O_2, \ldots, O_T),$$

and let the corresponding hidden state sequence be

$$Q = (q_1, q_2, \ldots, q_T).$$

Given a fitted HMM with parameter set $\lambda = (A, B, \pi)$, the objective of the Viterbi algorithm is to determine the state sequence

$$\hat{Q} = \arg\max_{Q} P(Q \mid O, \lambda).$$

Since maximizing $P(Q \mid O, \lambda)$ is equivalent to maximizing the joint probability $P(Q, O \mid \lambda)$, the problem may be written as

$$\hat{Q} = \arg\max_{Q} P(Q, O \mid \lambda).$$

Define

$$\delta_t(j) = \max_{q_1,\ldots,q_{t-1}} P(q_1, \ldots, q_{t-1}, q_t = S_j, O_1, \ldots, O_t \mid \lambda),$$

that is, $\delta_t(j)$ is the highest probability among all state paths of length t that end in state S_j. For each time point t and state S_j, the algorithm stores both this maximum probability and the state at time $t-1$ that achieves it.

The Viterbi recursion consists of three steps. First, the initialization step is given by

$$\delta_1(j) = \pi_j b_j(O_1), \qquad j = 1, \ldots, N,$$

where π_j is the initial probability of state S_j and $b_j(O_1)$ denotes the observation density or probability of O_1 in state S_j. Second, for $t = 2, \ldots, T$, the recursion step is

$$\delta_t(j) = \max_{i=1,\ldots,N} \left[\delta_{t-1}(i) a_{ij}\right] b_j(O_t), \qquad j = 1, \ldots, N,$$

where a_{ij} is the transition probability from state S_i to state S_j. At each step, the maximizing previous state is stored in order to reconstruct the optimal path. Finally, in the termination step, the likelihood of the optimal path is

$$P^* = \max_{j=1,\ldots,N} \delta_T(j),$$

and the final state is chosen as the state that attains this maximum. The most likely full state sequence is then obtained by backtracking through the stored maximizing states.

The efficiency of the Viterbi algorithm is evident when comparing the computational complexity between the algorithm and simply calculating all possible paths that yield the observed sequence. The method of calculating all paths that result in the observed sequence has a computational complexity of $O(N^T)$, while the Viterbi algorithm has the computational complexity of $O(N^2T)$. Assume that the observed sequence has the length of $T = 365$ and that the number of states are two ($N = 2$), this is on the order of 2^{365} compared to $2^2 \times 365$, which is a quite dramatic decrease in complexity when using the Viterbi algorithm.

2.2 K-Means Algorithm

Over the years, different algorithms have been developed to efficiently divide M points in N dimensions into a subset of K clusters where the clusters have similar attributes. The K-means algorithm aims to minimize the within-cluster sum of squares, resulting in reduced dimensionality and clusters that represent averages of similar points within each cluster. There are several different implementations of the algorithm, Forgy (1965); Lloyd (1982); MacQueen (1967) to name a few. The algorithm that has been used in this thesis is presented in Hartigan and Wong (1979), which is the default method in the base $\mathbb{R}$ package *stats* (R Core Team, 2025). The algorithm is explained thoroughly in Subsection 2.2.1, below.

2.2.1 Hartigan & Wong algorithm

Let the data matrix be $X \in \mathbb{R}^{M \times N}$, where row i is

$$x_i = (x_{i1}, x_{i2}, \ldots, x_{iN}), \qquad i = 1, 2, \ldots, M.$$

Let the current K cluster centers be $c_1, c_2, \ldots, c_K$, where $c_j \in \mathbb{R}^N$. In the initialization, K distinct rows are chosen randomly as initial centers. For each row x_i, the algorithm computes its squared Euclidean distance to each center,

$$d(x_i, c_j)^2 = \sum_{m=1}^{N} (x_{im} - c_{jm})^2,$$

and stores c^*_{i1}, the nearest center, and c^*_{i2}, the second-nearest center. Row x_i is initially assigned to cluster c^*_{i1}.

After the initial assignment, each center is recomputed as the mean of all rows currently assigned to that cluster. If cluster j contains the row index set

$$\mathcal{C}_j = \{i : x_i \text{ is currently assigned to } c_j\},$$

then the updated center is

$$c_j = \frac{1}{|\mathcal{C}_j|} \sum_{i \in \mathcal{C}_j} x_i.$$

Equivalently, for each column $m = 1, 2, \ldots, N$,

$$c_{jm} = \frac{1}{|\mathcal{C}_j|} \sum_{i \in \mathcal{C}_j} x_{im}.$$

Now define $n(c_j)$ to be the number of rows currently assigned to cluster j. For a given row x_i, suppose its current cluster is c_{i1}^*. The Hartigan-Wong algorithm evaluates the effect of removing x_i from its current cluster and adding it to another cluster through the quantities

$$R_1 = \frac{n(c_{i1}^*)d(x_i, c_{i1}^*)^2}{n(c_{i1}^*) - 1},$$
$$R_2 = \frac{n(c_j)d(x_i, c_j)^2}{n(c_j) + 1}, \qquad c_j \neq c_{i1}^*.$$

Here, R_1 is the adjusted cost of removing row x_i from its current cluster, while R_2 is the adjusted cost of adding x_i to a candidate cluster c_j. The algorithm compares these quantities to determine whether transferring x_i reduces the within-cluster sum of squares.

Optimal-transfer stage. In the optimal-transfer stage, each row x_i is considered in turn. If $n(c_{i1}^*) = 1$, then x_i is the only row in its cluster and cannot be transferred, since doing so would leave the cluster empty. Otherwise, the algorithm computes R_1 for the current cluster and evaluates R_2 for all candidate clusters $c_j \neq c_{i1}^*$.

The cluster that minimizes R_2 is denoted by

$$\hat{c}_i = \arg\min_{c_j \neq c_{i1}^*} \frac{n(c_j)d(x_i, c_j)^2}{n(c_j) + 1}.$$

If

$$\frac{n(\hat{c}_i)d(x_i, \hat{c}_i)^2}{n(\hat{c}_i) + 1} \geq \frac{n(c_{i1}^*)d(x_i, c_{i1}^*)^2}{n(c_{i1}^*) - 1},$$

then no transfer is performed and row x_i remains in cluster c_{i1}^*. In this case, the best alternative cluster is stored for later use. If instead

$$\frac{n(\hat{c}_i)d(x_i, \hat{c}_i)^2}{n(\hat{c}_i) + 1} < \frac{n(c_{i1}^*)d(x_i, c_{i1}^*)^2}{n(c_{i1}^*) - 1},$$

then row x_i is transferred from cluster c_{i1}^* to cluster $\hat{c}_i$.

Suppose that row x_i is moved from cluster a to cluster b, where

$$a = c_{i1}^*, \qquad b = \hat{c}_i.$$

The two affected centers are updated immediately as

$$c_a^{\text{new}} = \frac{n(a)c_a - x_i}{n(a) - 1}, \qquad c_b^{\text{new}} = \frac{n(b)c_b + x_i}{n(b) + 1}.$$

At the same time, the cluster sizes are updated as

$$n(a) \leftarrow n(a) - 1, \qquad n(b) \leftarrow n(b) + 1.$$

Therefore, after a transfer, the algorithm updates only the two affected centers rather than recomputing all centers again.

Quick-transfer stage. After one pass through all rows in the optimal-transfer stage, the algorithm enters the quick-transfer stage. In this stage, each row x_i is considered again, but instead of searching over all clusters, the algorithm compares only the current cluster c^*_{i1} with the stored alternative cluster c^*_{i2}.

For row x_i, the quantities

$$R_1 = \frac{n(c^*_{i1})d(x_i, c^*_{i1})^2}{n(c^*_{i1}) - 1} \qquad \text{and} \qquad R_2 = \frac{n(c^*_{i2})d(x_i, c^*_{i2})^2}{n(c^*_{i2}) + 1}$$

are computed. If

$$R_1 < R_2,$$

then row x_i remains in its current cluster. Otherwise, if

$$R_2 \leq R_1,$$

then x_i is transferred from cluster c^*_{i1} to cluster c^*_{i2}. After such a transfer, the two affected centers and their cluster sizes are updated immediately using the same formulas as in the optimal-transfer stage.

The quick-transfer stage is less expensive computationally because it checks only the most promising alternative cluster for each row. It therefore acts as a fast refinement step between the broader optimal-transfer passes.

Termination. The algorithm alternates between the optimal-transfer stage and the quick-transfer stage until no row transfer can further reduce the within-cluster sum of squares. At that point, the algorithm has reached a local optimum, meaning that no single row can be moved from its current cluster to another cluster in a way that decreases the objective function.

The final objective minimized by the algorithm is the total within-cluster sum of squares,

$$\text{WSS} = \sum_{j=1}^{K} \sum_{i \in \mathcal{C}_j} d(x_i, c_j)^2,$$

where $\mathcal{C}_j$ is the set of rows assigned to cluster j. The Hartigan-Wong algorithm does not guarantee the global minimum of this objective over all possible partitions, but it does produce a partition that is locally optimal with respect to single-row transfers.

2.3 Random Forests

In many situations, random forest models provide a flexible way of learning patterns in data without assuming linear dependency structures between the response variable and the feature variables. Classification And Regression Trees (CART) build the foundation of random forest models. A random forest is an ensemble of many decision trees, where the gain of averaging many trees' outputs is the decrease in risk of training an overfitted model that generalizes poorly to new data.

2.3.1 Classification And Regression Trees

A decision tree recursively partitions the predictor space into a set of disjoint regions. Let the predictor vector be $x = (x_1, x_2, \ldots, x_p)$, the tree begins with the full training sample and repeatedly splits it into smaller subsets using binary decision rules of the form: $x_i < s$, for some predictor x_i and splitting value s. Repeating this process grows the tree and produces a hierarchical tree structure with internal nonterminal nodes and terminal leaves which are the nodes with no descendants (Genuer and Poggi, 2020).

For which values of s that the nodes are split depends on the response variable y. If the response variable is continuous, the resulting tree is called a regression tree and the split is chosen as the value that minimizes the residual sum of squares within the resulting nodes t_L and t_R (left/right). This encourages nodes where data points have similar values instead of nodes where variance is high. Let the number of observations y_i within node t be $\#t$ and the mean of the observations be $\bar{y}_t$. Then the variance of node t is defined as

$$V(t) = \frac{1}{\#t} \sum_{i:x_i \in t} (y_i - \bar{y}_t)^2.$$

The aim is then to split the node where the resulting child nodes' variance is minimized, or, equivalently maximizing the quantity

$$V(t) - \left(\frac{\#t_L}{\#t} V(t_L) + \frac{\#t_R}{\#t} V(t_R) \right).$$

This process is then repeated until the size of a node has reached a minimum stopping size. For a new observation x, the prediction step is simply dropping x down the tree and then taking the average of the training observations within the final terminal leaf that x ends up in (Genuer and Poggi, 2020).

If y is categorical with categories $\{1, \ldots, C\}$, the resulting tree is called a classification tree. The splitting of the nodes are then chosen to maximize class purity instead. Let $\hat{p}_t^c$ denote the proportion of data points with category c in node t. Then the *Gini index* is defined as

$$\phi(t) = \sum_{c=1}^{C} \hat{p}_t^c (1 - \hat{p}_t^c),$$

and the aim is to maximize

$$\phi(t) - \left(\frac{\#t_L}{\#t} \phi(t_L) + \frac{\#t_R}{\#t} \phi(t_R) \right).$$

Similar to the continuous case, this maximizes homogeneity within the resulting child nodes after splitting node t. The classification step for a new observation x is then as previously, dropping the observation down the tree until the final terminal leaf, the maximum $\hat{p}_t^c$ within the terminal leaf is then the predicted category for that x. A problem with the CART framework is that the resulting trees are often prone to overfitting where the models begin to capture noise in the data instead of identifiable patters. An improvement of this is to grow many trees (typically > 500). In the regression case, the predictions among the trees are then averaged in order to reduce variance. The corresponding process for classification trees is choosing the category with the largest amount of votes by the individual classification trees (Genuer and Poggi, 2020).

2.3.2 Random Forests

The method of growing an ensemble of classification trees and choosing category with the majority of votes are called classifier forests and are part of the more general random forest modeling framework. In order to reduce the variance of the predictions, the standard is to sample many subsets of the training data with replacement and then to fit classification trees on each subset. For each node, a subset of the existing features are then sampled and used to perform the optimal split. The idea is then that many trees on different sets of data with different combinations of features for each split, reduces the correlation between the trees and reduces variance compared to a single fully grown classification tree that has the risk of being overfitted. An alternative implementation that will prove to be effective in Section 4.2, is to estimate category probabilities instead of counting the number of times each observation is assigned to a certain category. This is done by averaging the class proportions from each trees' terminal leaf across all trees, this then gives an idea of how certain the model is when classifying new data to existing categories and enables less strict decision rules (Breiman, 2001).

The continuous response variable analogue is called regression forests and follow the same overall procedure: bootstrap samples are drawn from the training data, a random subset of features are considered for each node, regression trees are grown on each subset of data and the prediction is formed by taking the average of all trees' individual predictions. In this case the resulting output is then of course a numeric value and not a majority voted category or vector of category probabilities (Breiman, 2001). The regression forest estimates the conditional mean of the response variable y given the observed data $X = x$, or $E(y|X = x)$. A useful way to express the regression forest prediction is through forest weights, following the notation in (Meinshausen, 2006), we have a single tree $T(\theta)$ where θ denotes the randomization used to grow the tree. For a new observation x, let $l(x, \theta)$ denote the terminal leaf that x ends up in after dropping it down the tree. Each training observation i then receives the single-tree weight

$$w_i(x, \theta) = \frac{\mathbb{1}_{\{X_i \in R_{l(x,\theta)}\}}}{\#\{j : X_j \in R_{l(x,\theta)}\}}, \tag{2.1}$$

where $\mathbb{1}_{\{\cdot\}}$ is the indicator function and $R_{l(x,\theta)}$ is the predictor-space region corresponding to that leaf. An observation then receives zero weight if it is not in the same terminal leaf as x, and otherwise receives equal weight among all observations in that leaf. The weights therefore sum to one within each tree. Averaging over all K trees gives the forest weights

$$w_i(x) = \frac{1}{K} \sum_{t=1}^{K} w_i(x, \theta_t). \tag{2.2}$$

Using the forest weights, the regression forest predictor can then be written as

$$\hat{E}(y|X = x) = \sum_{i=1}^{n} w_i(x) y_i.$$

This shows that a regression forest can be interpreted as an adaptive nearest-neighbor method, where the forest determines which training observations are most relevant for predicting y given x.

An extension of the regression forest is the quantile regression forest, which aims to estimate the full conditional distribution $F(y|X = x)$ rather than only its mean. This

extension enables the estimation of quantiles which for instance, is valuable when underestimation of future observations results in critical consequences. This will be made evident later on during Chapter 4. The key idea of quantile regression forests is to use the same forest structure and the same weights as regular regression forests (Equations 2.1 and 2.2), but instead of only averaging each terminal leaves' observations, the observations within the terminal leaf are also stored. The estimated conditional distribution is then formed as the weighted empirical distribution

$$\hat{F}(y|X = x) = \sum_{i=1}^{n} w_i(x)\mathbb{1}_{\{Y_i \leq y\}},$$

from which the conditional α-quantile is obtained as

$$\hat{Q}_\alpha(x) = \inf\{y : \hat{F}(y|X = x) \geq \alpha\}.$$

The extension by (Meinshausen, 2006) provides a method that outputs more information about the conditional response variable as several quantiles can be estimated instead of only relying on the conditional expectation. How all of the models presented in this theory chapter are implemented and to what purpose will be thoroughly discussed during the next methodology chapter.

Chapter 3

Methodology

Before delving into the specific steps of the methodology, a more general summary is presented which aims to give a broader context as to why these steps are taken and why the steps have been conducted in this specific order. In order to implement the proposed framework, electric vehicle fast-charging data is needed, details about the required data is presented more thoroughly in the subsequent Section 3.2. The data was provided by Öresundskraft AB, as mentioned previously in Section 1.2. Depending on the quality of the data, there is a risk of obtaining data that is not solely EV charging. In this thesis, the reason being that a gas station or a store shares measurement point with the EV charging, resulting in a mixed signal. This motivates Section 3.3, where a hidden Markov model is implemented with two states with the objective of finding the optimal hidden state sequence. After sufficient data preprocessing, hourly maxima of kilowatt-hours (kWh) are calculated, where the maximum values are grouped by charging station and weekday/weekend. In Section 3.4, the maxima are normalized and then used for clustering. The output is then K centroids, or average normalized shape profiles. In order to assign a new charging station's cluster membership, a classifier forest is trained on the meta-data and the cluster labels from the previous step. When the cluster membership is finalized, the last step is to predict the maximum scale that the centroid will achieve based on the meta-data. This is done by training three different quantile regression forests for three different quantiles, where the models are trained with the daily peak loads as response variables, and the meta-data as predictor variables. Finally, the model's performance is evaluated by performing the steps from sections 3.4, 3.5 and 3.6 repeatedly through Leave-One-Out Cross-Validation (LOOCV), where the model's ability to generalize to unseen data is tested through various metrics.

3.1 Software

Obtaining the data was done through a combination of Python, $\mathbb{R}$ and an external API. GPS points of fast-charging stations were provided through the NOBIL API, a database of public charging stations in Norway and Sweden (Skåland Webservice, 2025). The GPS points were then used to find fast-charging measure points in Öresundskraft AB's internal system. When the unique identifier for each measure point was obtained, the data was collected from Öresundskraft AB's private data base through Python and later on loaded into $\mathbb{R}$. The data preprocessing, modeling and evaluation was then completed using the

graphical user interface RStudio.

3.2 Data

The raw data obtained consists of measurements of hourly averages of kilowatt-hours (kWh) from 33 charging stations with a minimum count of two Direct Current (DC) charging points. The life-span of the stations differ, where the earliest date that has been used is 2025-01-01. All stations have data obtained up to 2026-03-30. Several stations have longer life-spans than the starting date, but to ensure that the size of the datasets being handled do not become too large, they are all capped from this date and onward. Roughly one third of the stations used are constructed more recently than 2025-01-01 and as a result, have fewer observations. Some charging behavior has been manually removed, an example is the initial behavior of newer charging stations', as there seems to be tests of the hardware before opening the station to the public. Another example is when the measure point in the electric grid suddenly starts to measure another signal as well, an example of this is depicted in Figure 3.1, where there suddenly is an offset and no measurements are exactly zero after the initial period.

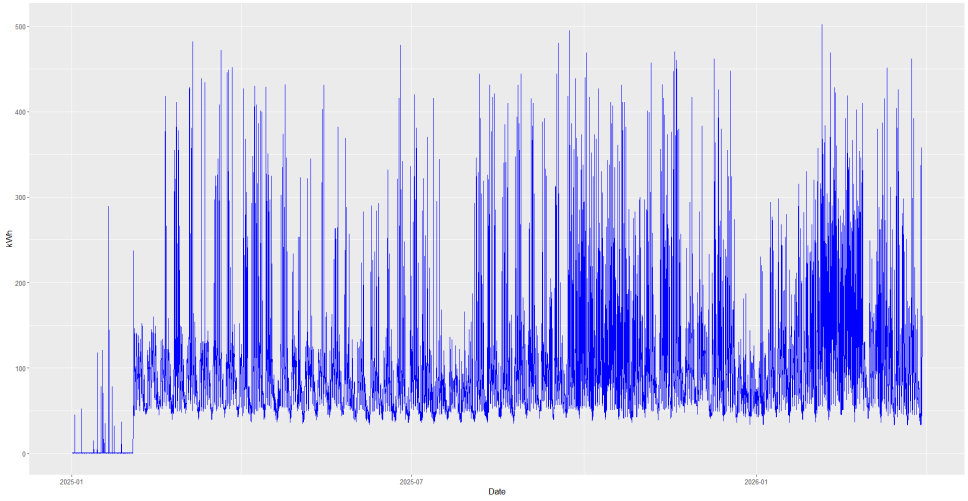

Figure 3.1: Example of mixed signal, where behavior of data suggests that the measured signal is not solely electric vehicle charging after the initial period.

Except for the raw observations, several station-specific features have been compiled in the station meta-data. A constrain is that the meta-data must only include static features that do not depend on raw observations. Examples of features from the meta-data are: *number of charging points, distance from nearest highway* and *number of neighboring stations within 500 m*. The $\mathbb{R}$ package *geosphere* (Hijmans, 2026) was used to calculate the distance from nearest highway and number of charging stations within 500 m. The package calculates the distance between two GPS points, where earth is assumed to be a perfect sphere.

Several calendar features were computed from the already existing date column, the two calendar features that ultimately proved useful was *hour of day* and *weekday/weekend*. In the next section, a subset of the hourly column also proved to be useful, which will be discussed thoroughly during the relevant section. No missing values that suggest

faulty measurements could be found during the preprocessing of the data. To further ease the reading of this thesis, from now on, the hourly average kilowatt-hours will be denoted by kWh, $y_{i,t}$ will denote the sequential observations of kWh for station i and timestamp t and $y_{i,d,h}$ will denote the observed hourly kWh for station i, on day d, at hour $h \in \{0, \ldots, 23\}$. Where the latter is used for aggregation into day and hour-specific quantities. Lastly, some transformations was performed during various stages of implementation of the methodology. Firstly, the separation of EV charging and non-EV charging during Section 3.3 utilizes $\ln(1 + y_{i,t})$, secondly, the normalization of the hourly maxima of $y_{i,d,h}$ in Section 3.4, and thirdly, the log odds of a utilization measure which is presented in Section 3.6. All of which will be motivated thoroughly during their relevant sections.

3.3 Separating Signals

As seen in Figure 3.1, there are cases where a measure point coupled with a charging station do not solely contain EV charging. In the figure, typical charging behavior is omitted during the start of the year, where charging behavior is characterized by sharp peaks in kWh, while it is zero when no charging occurs. After some time, a sudden offset of energy consumption is added, where the observations are nonzero for the remainder of the measured time interval. For a charging station to be used at this frequency, especially during the night, is highly unlikely. Since the objective is to create realistic peak load profiles for a new charging station, keeping an offset this large would influence the predicted load curves if not dealt with.

This reasoning motivates the action of trying to separate the charging signal from the non-charging signal. In order to do this, we first need a structured way of finding charging stations that omit this mixed behavior. The method suggested for screening for this behavior, is by calculating a baseline ratio between a low quantile and a high quantile. For a charging station with pure EV-charging, the low quantile should typically be close to zero and represent observations where no charging occurs. For a charging station with mixed signals, such as the station in Figure 3.1, the low quantile should be higher, resulting in a higher baseline ratio. This framework should then be combined with visual inspection of the candidate stations, as well as inspecting the actual values of the lower quantiles. This procedure is performed in the beginning of the Results chapter in Section 4.1.

After the screening is completed and the charging stations with mixed signals are found, the aim is to separate each station's signals. Since charging behavior typically does not occur independently between measured hours (an EV can actively charge during more than one observed hour), a model that allows for this structure is required. A two-state HMM (see Section 2.1) assumes that there are two underlying states, in this case, EV-charging and non EV-charging. The transition probabilities ensure that the probability of remaining in a state can differ from the probability of leaving it, which fits the described application where a charging session can span over more than one hourly measurement. In order to ease the model's identification of states, measurements where *hour_nightly* equals one are only used, where the boolean variable is defined as

$$hour\_nightly = \begin{cases} 1, & \text{if hour} \in \{0, 1, 2, 3, 4, 5\}, \\ 0, & \text{otherwise.} \end{cases}$$

The reason being that charging behavior is seldom occurring during the night, which makes nightly charging behavior stand out more compared to the non-charging consumption. If all hours of the day were used, there is a higher risk of wrongly separating the signals and later on removing actual charging behavior from the dataset.

In order to fit the two-state HMM model, the $\mathbb{R}$ package *depmixS4* (Visser and Speekenbrink, 2010) was used, where the number of states, assumed underlying distribution and response variable are used as input. As mentioned previously in Section 3.2, the measurements are transformed by $\ln(1 + y_{i,t})$, where $y_{i,t}$ is the observed kWh of timestamp t for station i. The reason for the transformation is to reduce skewness and stabilize variance, making a Gaussian approximation within each hidden state more plausible. The motivation behind using $1 + y_{i,t}$ instead of $y_{i,t}$, is simply because the natural logarithmic function is undefined for values at exactly zero, which can be the case for $y_{i,t}$ when no EVs are charging. Even though the screening aims to find stations with more than only a charging signal, the transformation ensures that no values are exactly zero before taking the natural logarithm.

After fitting one HMM model for each charging station that was selected from the previous screening, the Viterbi algorithm is used to obtain the most likely underlying state sequence, given the observed values and the fitted model. The Viterbi algorithm is implemented within the same $\mathbb{R}$ package, *depmixS4* (Visser and Speekenbrink, 2010), as before. In this thesis, the median of all nightly observations labeled by the non-charging state, is used to estimate the general non-charging behavior. This station-specific median is then subtracted from the complete dataset for each station that was chosen during the screening, this then aims to remove an estimate of the non-charging behavior. If any value of $y_{i,t}$ is negative after subtracting the median, it is replaced by zero, as negative charging patterns are not possible for these types of chargers. Different approaches for removing the non-charging behavior can be utilized during this part of the methodology, such as replacing the median with other quantiles or the mean. The median was chosen as it is robust against outliers that might be due to wrongful state labeling.

3.4 Clustering Charging Stations

Once the non-charging behavior has been dealt with, we are ready to move on to the clustering of the charging stations. The purpose of this section and Section 3.6, is to separate the problem of finding a model that estimates the shape of the load curve, and the scale of the load curve, as many models fail to capture both. Many different approaches were tested beforehand, where the goal was to obtain a new charging station's peak load profile. Separating the problem into several subproblems, simplified the overall complexity, and enables us to validate each subproblem independently. The suggested framework is then to separate the problem into two steps instead of one. We begin by focusing on the first step, namely estimating the normalized hourly shape

$$h_{j,w,h},$$

where $j \in \{1, \ldots, K\}$ denotes the cluster membership, $w \in \{\text{weekday}, \text{weekend}\}$ denotes day type, and $h \in \{0, \ldots, 23\}$ denotes hour of day.

Due to the erratic behavior of EV-charging, normalized kWh maxima were used to create low-risk features for clustering. If the normalization step is ignored, there is no guarantee

that the clustering algorithm will not cluster stations based on scale instead of shape, which is the opposite of this section's objective. An example station that omits the previously mentioned behavior is showcased in Figure 3.2, below. The station omits low charging frequency and relatively high peaks when charging does occur. This behavior is why the hourly maxima are taken over long time periods, as models handling raw sequential observations will most likely struggle due to the heteroskedastic nature of the data.

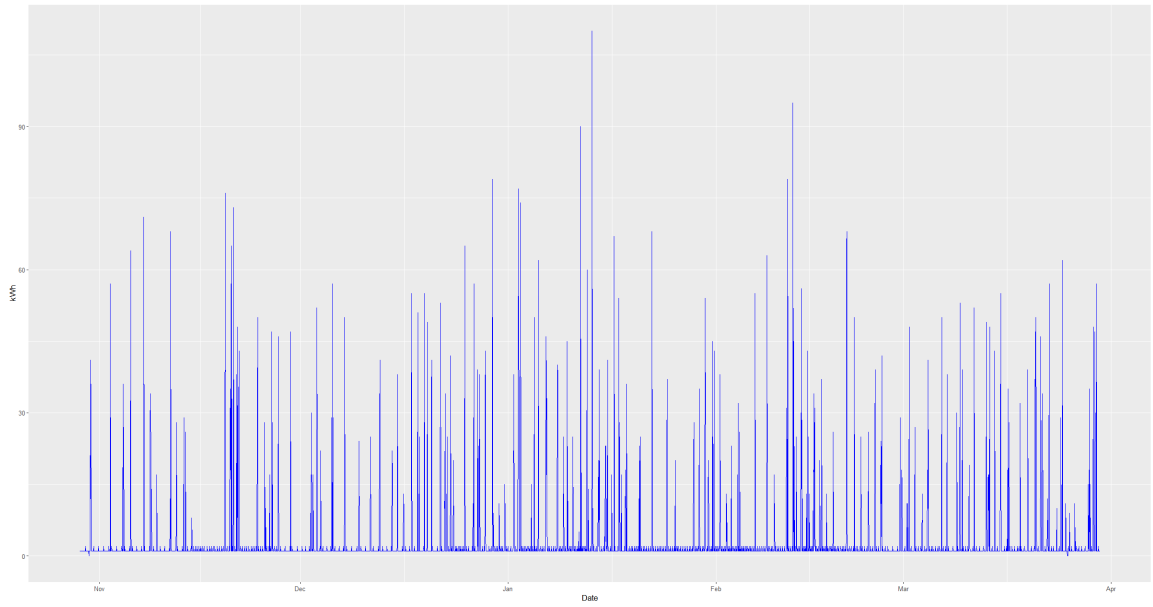

Figure 3.2: Example of electric vehicle charging behavior for one charging station across time, sharp peaks during daytime followed by low usage during nighttime is typical.

Each charging station's measurements of kWh were grouped by weekday and weekend. Then hourly maximum values were computed based on the data within the two groups. Since we are only interested in the peak hourly shape during this step, the hourly maxima were normalized. This was done by summing each station's weekday maxima and each station's weekend maxima, and then dividing each groups hourly maxima with the total sum of that groups hourly maxima. The resulting values are station-specific normalized weekday hourly maxima and normalized weekend hourly maxima that sum to one within each corresponding group. The weekday and weekend peak charging behavior differ significantly among the charging stations, this is the reason as to why the shape measures are separated by weekend and weekday, enabling the model to have different shapes depending on the type of day. The hourly maxima of kWh, divided by the station-specific theoretical max, separated by weekday and weekend, is showcased in Figure 3.3, below.

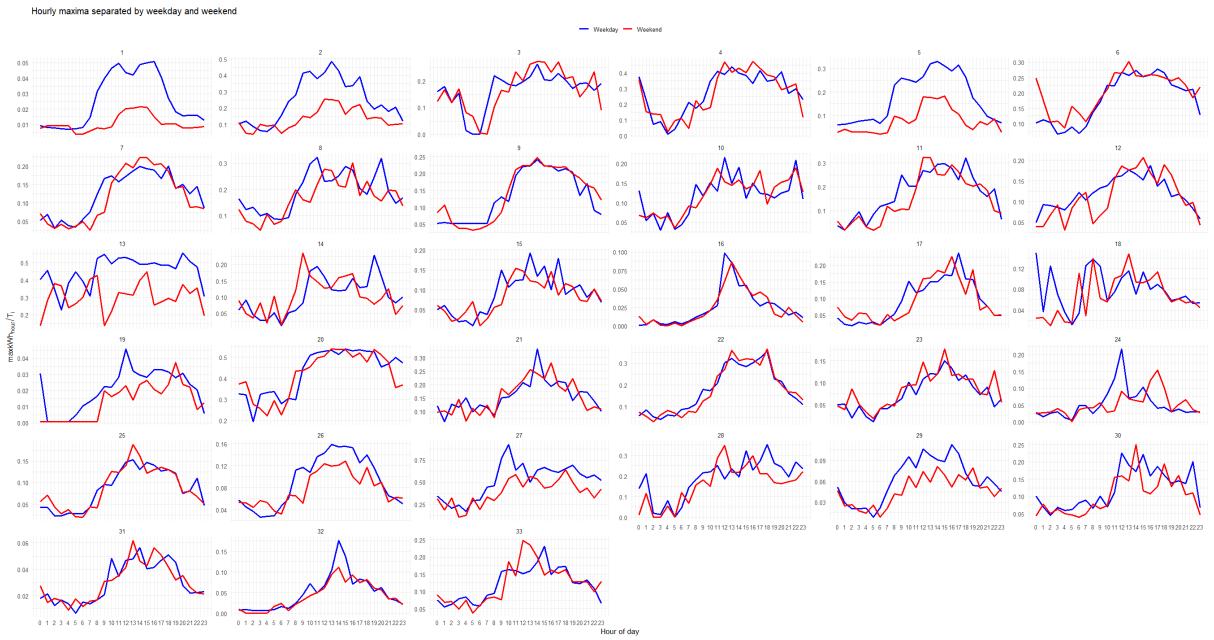

Figure 3.3: Hourly normalized maxima of kWh of each charging station within dataset, where the blue line represents weekday data, and the red line represents weekend data.

Now that the restructured data has been motivated, the next step is to run the K-means algorithm presented in Section 2.2. The input data is then a data matrix X with dimensions 33×48 where each row corresponds to one station's normalized hourly maxima. For a single station i, this is then

$$x_i = (wd_{i,0}, wd_{i,1}, \ldots, wd_{i,23}, we_{i,0}, we_{i,1}, \ldots, we_{i,23}), \tag{3.1}$$

where $wd_{i,0}$ corresponds to the normalized weekday maxima of hour 0 for station i and $we_{i,23}$ corresponds to the same measure computed from weekends and hour 23 etc. As previously mentioned, the $\mathbb{R}$ package *stats* (R Core Team, 2025) is used to implement the K-means algorithm. Required input is the data matrix X and the number of clusters K. How K is chosen will be made evident once the classifier forest in the next section has been presented.

3.5 Classification of Cluster Membership

For a given K, the previous step yields us K clusters. Each charging station is assigned to one of the clusters as a result of the algorithm. The centroids (average shape profile of all stations within one cluster) of each cluster is based on weekday and weekend shape. In a realistic situation, a new charging station will not have measurements of kWh that can be used to construct x_{new} in the same way as in Equation 3.1. Therefore, some form of classifier model is required, the goal is then for the model to assign cluster membership based on the new station's static meta-data (*number of charging points, GPS point etc*).

The classifier forest is a suitable choice due to the nonlinear and complex structure of the data. Charging behavior does not depend linearly on many of the covariates, for an example, two charging stations that are in the same proximity of a highway and that share many of the meta-data features might have very different charging frequency and scale of

peak loads. Another advantage with the classifier forest is the computational efficiency of the model, enabling many different models to be tested until the predicted output is satisfactory.

The classifier forest was trained through the $\mathbb{R}$ package *ranger* (Wright and Ziegler, 2017), the package allows for estimation of cluster membership probabilities, which was presented in Section 2.3.2. Estimating the probabilities of a new charging station's cluster membership has advantages over using a hard decision rule such as the majority vote of the decision trees. If the classifier forest's maximum predicted cluster membership probability is 40%, the hard decision rule assigns the new charging station to that cluster, without regards to the relatively low certainty of the assigned cluster. The proposed alternative, is to instead compute the cluster probabilities for the new charging station, and compute a weighted average of the cluster centroids. This alternative then, takes into account the uncertainty that the model has regarding the shape of the new charging station, in mathematical notation, this corresponds to

$$\hat{h}_{i,w,h} = \sum_{j=1}^{K} \hat{p}_{ij} h_{j,w,h}, \tag{3.2}$$

where $\hat{p}_{ij}$ is the predicted probability that station i belongs to cluster j. All of this is under the assumption that K is given, which is not the case as of now. In order to find a suitable number of clusters, K, two criteria are proposed.

1. The predicted shape (Equation 3.2), should minimize the error between the predicted hourly shape profile and the true hourly shape profile, either by Mean Absolute Error (MAE) or Root Mean Square Error (RMSE).
2. No cluster should contain one single charging station.

The first criterion is relatively straight forward, the predicted shape should be as close to the true shape as possible. The second criteria handles the risk of overfitting, if one cluster is based on one charging station's normalized hourly maxima, the resulting profile will most likely not generalize to new data, as it is tailored to one station's behavior and not an average of several similar stations. The proposed framework is then to compute a grid of K values and perform cross validation where a subset of the data is removed for testing. The remaining data is used for training, this means that for every fold, all values of K are used to construct clusters. For each iteration, the classifier forest is trained on the training split and used to predict the testing split. This yields us V number of folds per K, where for each K, we compute the MAE and RMSE between the V different predicted shapes and the true shapes. During this process, the average of the minimum cluster size is computed as well, in order to keep track of criteria number two. The chosen grid of K values in this thesis are $K_{grid} = \{1, 2, \ldots, 10\}$, and the number of folds are $V = 10$. Resulting in 1/10 of the data being held out for testing, while clustering and training the classifier forest on the remaining 9/10 of the data. This approach is beneficial since the number of clusters are chosen based on a situation that resembles reality, where out-of sample predictability and generalization to new data is prioritized.

3.6 Predicting Daily Peak Loads

Once the number of clusters have been chosen and the classifier model has been specified, the final modeling step is to predict the daily peak load. The objective of this section

is then to present a suitable model whose purpose is to scale the predicted weekday and weekend shape by the estimated daily peak load. Depending on objective, the response variable can be chosen to capture the desired quantity. For this thesis, underestimation could result in potentially critical consequences such as power failure, which is why a low-risk scale target is implemented. To begin with, each charging station's daily peak kWh measurements are extracted and denoted by

$$m_{i,d} = \max_{h \in \{0,\dots,23\}} y_{i,d,h}.$$

The same reasoning applies as in Section 3.4 and Figure 3.3, that is, that a model that captures differences between weekdays and weekends should be beneficial to the final modeling framework.

The next step is then to compute the maximum of the extracted daily peaks, separated by weekday and weekend. Let the set of days of type w for station i be defined as $\mathcal{D}_{i,w} = \{d : w(d) = w\}$. Now define the peak quantity by weekday and weekend as

$$M_{i,w} = \max_{d \in \mathcal{D}_{i,w}} m_{i,d}.$$

Reasonable alternatives to this would be to model the mean or any relevant quantile of the daily peaks instead, depending on risk aversion tied to the application. Using the raw maximum daily peak as response variable holds a significant disadvantage, namely that there is no guarantee that the predicted response will remain within realistic bounds. For a charging station with four charging points, each point having a maximum effect of 400 kW, the theoretical station maximum capacity is $4 \times 400 = 1600$ kWh. If the predicted peak load exceeds this hard limit, there is no point in using the suggested framework since one could simply use the hourly theoretical maximum capacity and still optimize the grid more sufficiently than what the modeling framework suggests. In order to ensure a bounded prediction, each station's maximum daily peak is divided by the station's theoretical maximum T_i. We call this quantity the maximum daily peak utility and define it as

$$u_{i,w} = \frac{M_{i,w}}{T_i}. \tag{3.3}$$

Under the assumption that the data consists of only active stations and that no station achieves it's theoretical maximum capacity, the following holds

$$u_{i,w} \in (0, 1).$$

In this thesis, this is true, but in other applications, some numeric safeguards are reasonable to apply, such as if statements that check wether the quantity is exactly zero or exactly one, and then adding or subtracting some small number to ensure that the following transformation is valid. Given that the maximum daily peak utility is bounded within $(0, 1)$, it can then be transformed through the *logit* function, the transformation was mentioned briefly in Section 3.2. The transformation maps a value within the interval $(0, 1)$ to the real line $\mathbb{R}$, the resulting variable and transformation are presented in Equation 3.4 below

$$z_{i,w} = \ln\left(\frac{u_{i,w}}{1 - u_{i,w}}\right). \tag{3.4}$$

The suggested approach is then to transform $u_{i,w}$ through Equation 3.4, and obtaining the final response variable $z_{i,w}$. Quantiles of $z_{i,w}$ are then estimated through the quantile

regression forest model, which was presented in Subsection 2.3.2 and implemented with the same $\mathbb{R}$ package as the classifier forest, namely *ranger* (Wright and Ziegler, 2017). The station meta-data that was used as predictors in the classifier forest in the previous section are now utilized as predictors for the quantile regression forest as well.

There are primarily two advantages with modeling quantiles of the conditional distribution instead of the conditional expectation. Firstly, due to the *equivariance to monotone transformations* property that quantiles possess, we can perform any monotone transformation of the response variable, predict the desired α quantile $\hat{Q}_\alpha(\cdot)$, then perform the inverse transformation in order to obtain the predicted quantile in the original scale (Koenker, 2005). Since the *logit* function is a monotone transformation, we can simply apply the inverse *logistic* function to the predicted quantile $\hat{Q}_\alpha(z_{i,w})$ in order to obtain the quantile prediction of the previously defined utility quantity (Equation 3.3)

$$\hat{Q}_\alpha(u_{i,w}) = \frac{e^{\hat{Q}_\alpha(z_{i,w})}}{1 + e^{\hat{Q}_\alpha(z_{i,w})}}. \tag{3.5}$$

Secondly, several quantiles can be estimated, yielding a better representation of the conditional distribution of the response variable rather than one single point estimate which would have been the case if modeling the conditional expectation. Performing the inverse transformation presented in Equation 3.5, we obtain the predicted quantile in utility scale. The final step is then to multiply the station-specific quantile by the total capacity T_i in order to obtain the final predicted peak load quantile in original kWh scale

$$\hat{Q}_\alpha(M_{i,w}) = T_i \hat{Q}_\alpha(u_{i,w}). \tag{3.6}$$

The chosen quantiles for this thesis are $\alpha = \{0.5, 0.75, 0.99\}$, but different quantiles can be utilized depending on objective. The motivation behind the chosen quantiles is to have the representative case, $\alpha = 0.5$, i.e the median, a high load case, $\alpha = 0.75$, and one extreme load case that is very low-risk if used when planning the electric grid, namely $\alpha = 0.99$. The number of regression trees used to estimate the quantiles are 1500, the reason for increasing the number from the typical 500 is that the estimated quantiles depend on an accurate approximation of the complete conditional distribution. For tail quantiles such as $\alpha = 0.99$, a relatively large number of trees are needed to reduce the sampling variability, otherwise we risk obtaining unstable and high variance estimates. The final modeling step is then to combine the predicted quantile $\hat{Q}_\alpha(M_{i,w})$ with the corresponding hourly shape. In order to ensure that the maximum value of the product between the predicted quantile and the predicted hourly shape corresponds to the predicted quantile, an effective transformation of the shape is to divide it by it's maximum value. At the shapes peak, the ratio will equal one and any scaling factor multiplied by it will consequently have it's value unchanged at the shapes peak. We therefore define the final predicted shape as

$$\hat{h}^\star_{i,w,h} = \frac{\hat{h}_{i,w,h}}{\max_{h' \in \{0,\dots,23\}} \hat{h}_{i,w,h'}}. \tag{3.7}$$

The final predicted quantile scaled hourly load curve is then the product of the two quantities, and is defined as

$$\hat{p}^{(\alpha)}_{i,w,h} = \hat{Q}_\alpha(M_{i,w}) \hat{h}^\star_{i,w,h}. \tag{3.8}$$

The complete modeling framework has now been presented and we proceed to the problem of evaluating the modeling framework's performance in Section 3.7, below.

3.7 Leave-One-Out Cross-Validation

In order to replicate a realistic situation where a new charging station's peak hourly load curve is to be predicted, we perform Leave-One-Out Cross-Validation (LOOCV). The procedure is then to remove one charging station, implement the presented framework, predict the unseen charging station's quantile scaled hourly load curve and then repeat the process until all charging stations' peak load profiles have been predicted. The complete prediction framework, repeated for each charging station, is presented in Algorithm 1, below.

Algorithm 1 LOOCV procedure for station-specific peak load profile prediction

for each station $i = 1, \ldots, N$ **do**

Remove station i from the data

Construct the training matrix $X^{(-i)}$ from the remaining stations, where each row contains the normalized weekday and weekend hourly maxima

Apply K-means to $X^{(-i)}$ and obtain cluster memberships and cluster shapes $h_{1,w,h}, \ldots, h_{K,w,h}$

Train a classifier forest on the remaining stations using static meta-data as predictors and cluster membership as response

Predict the cluster membership probabilities $\hat{p}_{i1}, \ldots, \hat{p}_{iK}$ for station i

Compute the predicted shape for station i as

$$\hat{h}_{i,w,h} = \sum_{j=1}^{K} \hat{p}_{ij} h_{j,w,h},$$

where $w \in \{\text{weekday}, \text{weekend}\}$ and $h \in \{0, \ldots, 23\}$

Normalize the predicted shape so that its maximum value equals one:

$$\hat{h}^{\star}_{i,w,h} = \frac{\hat{h}_{i,w,h}}{\max_{h' \in \{0,\ldots,23\}} \hat{h}_{i,w,h'}}$$

For each training station $r \neq i$, compute the daily peaks

$$m_{r,d} = \max_{h \in \{0,\ldots,23\}} y_{r,d,h}$$

and the maximum daily peak by day type

$$M_{r,w} = \max_{d \in \mathcal{D}_{r,w}} m_{r,d}$$

Compute the utility response vector

$$u_{r,w} = \frac{M_{r,w}}{T_r},$$

transform it by

$$z_{r,w} = \ln\left(\frac{u_{r,w}}{1 - u_{r,w}}\right)$$

Train quantile regression forests on the remaining stations using static meta-data as predictors and $z_{r,w}$ as response, separately for weekdays and weekends

Predict the target quantile $\hat{Q}_{\alpha}(z_{i,w})$ for station i

Apply the inverse logistic transformation:

$$\hat{Q}_{\alpha}(u_{i,w}) = \frac{e^{\hat{Q}_{\alpha}(z_{i,w})}}{1 + e^{\hat{Q}_{\alpha}(z_{i,w})}}$$

Rescale by total capacity:

$$\hat{Q}_{\alpha}(M_{i,w}) = T_i \hat{Q}_{\alpha}(u_{i,w})$$

Reconstruct the predicted quantile scaled hourly load curve:

$$\hat{p}^{(\alpha)}_{i,w,h} = \hat{Q}_{\alpha}(M_{i,w}) \hat{h}^{\star}_{i,w,h}$$

end for

In order to measure the performance of the predicted peak load curves, we introduce some relevant metrics that are calculated for each iteration of Algorithm 1. Firstly, two curve accuracy metrics, Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). These measure how close the predicted load curve is to the observed hourly maximum values ($p^{\text{obs}}_{i,w,h}$) for station i on average, either by the absolute difference or the squared difference. Here, $\mid W \mid$ is the number of elements within $W = \{\text{weekday, weekend}\}$ and 24 springs from the fact that there are 24 hours during a full day.

$$\text{MAE}_i^{(\alpha)} = \frac{1}{\mid W \mid 24} \sum_{w \in W} \sum_{h=0}^{23} \mid \hat{p}^{\alpha}_{i,w,h} - p^{\text{obs}}_{i,w,h} \mid \tag{3.9}$$

$$\text{RMSE}_i^{(\alpha)} = \left(\frac{1}{\mid W \mid 24} \sum_{w \in W} \sum_{h=0}^{23} (\hat{p}^{\alpha}_{i,w,h} - p^{\text{obs}}_{i,w,h})^2 \right)^{1/2}. \tag{3.10}$$

Secondly, a metric that measures the average gain when utilizing the predicted load curve in grid planning, compared to using the theoretical maximum T_i for every hour and day

$$\text{Gain}_i^{(\alpha)} = 1 - \left(\frac{1}{\mid W \mid 24} \sum_{w \in W} \sum_{h=0}^{23} \frac{\hat{p}^{\alpha}_{i,w,h}}{T_i} \right). \tag{3.11}$$

$\text{Gain}_i^{(\alpha)} = 0$, would mean that the predicted load curve is equal to the theoretical maximum for each hour and day type. $\text{Gain}_i^{(\alpha)} > 0$ represents an increase in grid optimization, since the grid is utilized more efficiently than the theoretical maximum.

Thirdly, a metric that measures the average rate of underestimation when using the predicted load curve in grid planning, where $\mathbb{1}_{\{\cdot\}}$ denotes the indicator function

$$\text{UnderRate}_i^{(\alpha)} = \frac{1}{\mid W \mid 24} \sum_{w \in W} \sum_{h=0}^{23} \mathbb{1}_{\{\hat{p}^{\alpha}_{i,w,h} < p^{\text{obs}}_{i,w,h}\}}. \tag{3.12}$$

$\text{UnderRate}_i^{(\alpha)} = 1$ would mean that the observed values exceed the predicted load curve in all cases, whereas $\text{UnderRate}_i^{(\alpha)} = 0$ would mean that no observed values exceed the predicted load curve. The results after implementing all steps within the proposed framework with real fast-charging EV data, are presented in Chapter 4.

Chapter 4

Results

During this chapter, all data regarding observations of kWh has been divided by the theoretical maximum station capacity T_i. This enables comparisons between charging stations and simplifies construction of figures, as the y-axis remains capped at 1.0 for each charging station. In order to ease comparisons, some figures are separated by color, which is included within the digital version of the thesis.

4.1 Separating Signals

We begin by investigating the 10th quantile and the 90th quantile of $\frac{kWh}{T_i}$ and the ratio between the two in Table 4.1, below. The station IDs are sorted based on highest baseline ratio ($\frac{\text{q10}}{\text{q90}}$) where the top ten charging stations are presented. We can see that the top five IDs have a higher baseline ratio than the remaining charging stations. In order to ensure that no mixed signals remain within the dataset, the sixth and seventh row is included as well. The resulting IDs that are investigated further is then 1, 2, 5, 13, 19, 24 and 26.

Table 4.1: Station-level quantiles and baseline ratio ($\frac{q10}{q90}$), ranked by baseline ratio, for the ten charging stations with the highest baseline ratio.

ID	q10	q90	$\frac{\text{q10}}{\text{q90}}$
19	0.000417	0.000833	0.5000
13	0.0613	0.198	0.3100
2	0.0919	0.301	0.3060
1	0.00671	0.0351	0.1910
5	0.0247	0.176	0.1410
26	0.00620	0.0603	0.1030
24	0.000833	0.0179	0.0465
30	0.00111	0.0333	0.0333
31	0.000417	0.0138	0.0303
18	0.000833	0.0315	0.0265

We then proceed to plot the tenth quantile for each hour of day of the candidate IDs in order to get a visual representation of the low quantile charging behavior. We see in Figure 4.1, below, that several IDs omit behavior that is non-typical for EV charging. Typically, the tenth quantile is very close to zero, as a charging station has periods of time where no charging is occuring. IDs 19 and 24 omit behavior that most likely is due to some artifact, measurement error or small power consumption related to the running

of the charging station since it is at a small constant level for large portions of the day. The remaining charging stations hover above zero and are therefore selected for the HMM that aims to separate the charging behavior from the other signal that might be present.

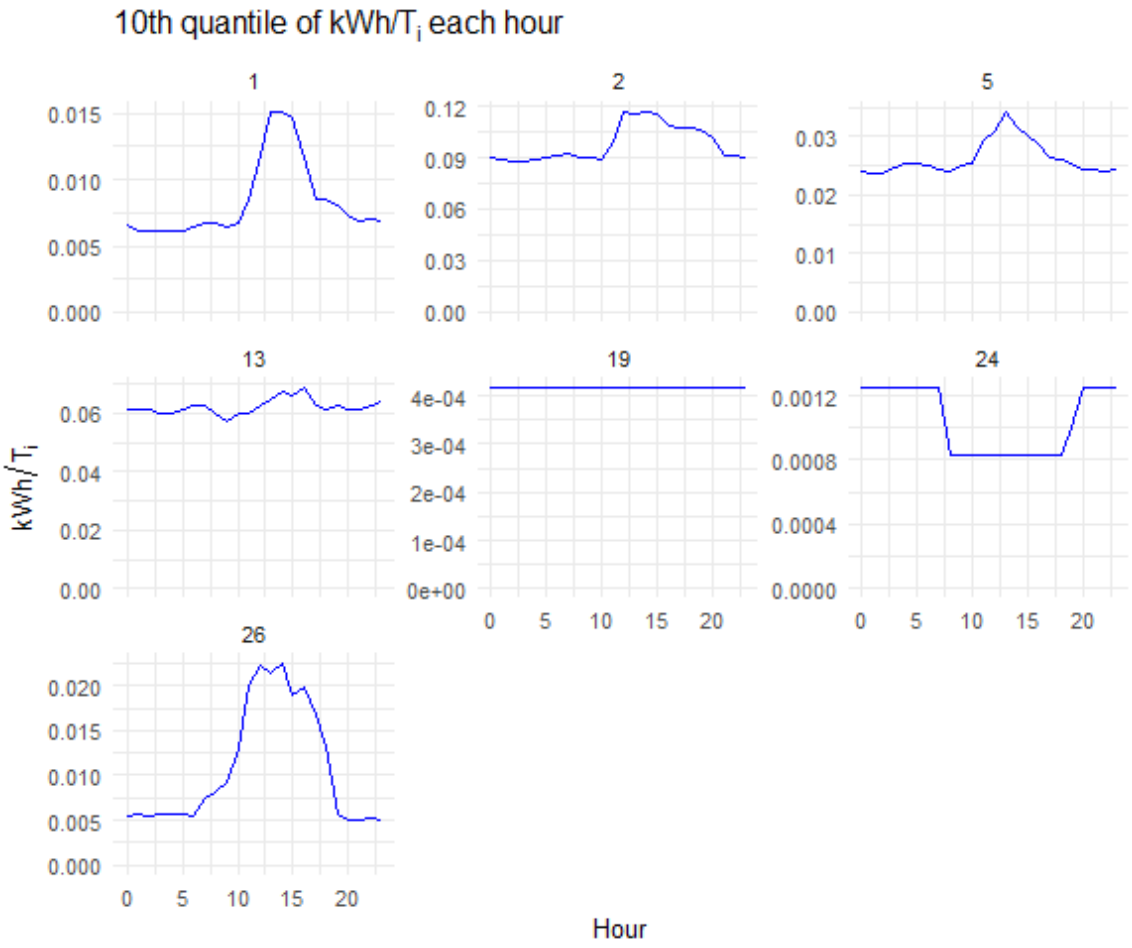

Figure 4.1: Candidate IDs normalized tenth quantile, across hour of day. IDs 19 and 24 omit power consumption close to zero, while the remaining IDs have greater lower quantile power consumption.

As mentioned in Section 3.3, a subset of the data is used to estimate the HMMs, namely only night time data that is transformed by the $\ln(1 + y_{i,t})$ transformation. After fitting a two-state HMM to each candidate IDs' night time data, the Viterbi algorithm is used to obtain the optimal hidden state sequences. The night time data for IDs 1, 2, 5, 13 and 26 are presented in Figure 4.2, where the optimal hidden state sequence is used for color coding the active state for each timestamp. IDs 5 and 13 omit a clearer separation between the states compared to the remaining IDs, especially when comparing to ID 1 that might capture charging behavior during the middle of the time series.

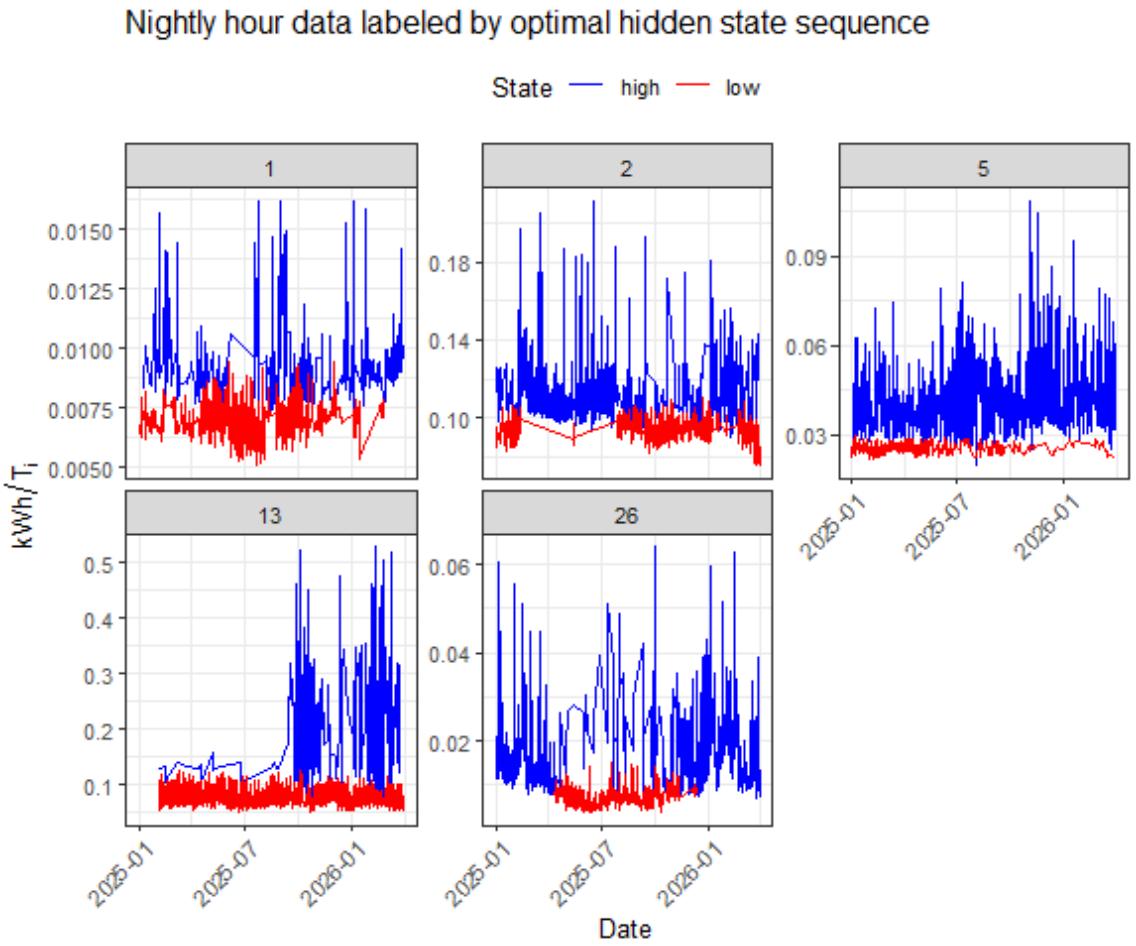

Figure 4.2: Candidate IDs' normalized kWh across time, filtered down to nighttime and separated by hidden Markov model's optimal state. Blue indicates likely charging behavior, whereas red indicates likely non-charging behavior.

The transformed night time observations within each state should be approximately Gaussian in order to satisfy the model assumptions of the HMM. The left panel in Figure 4.3 showcases that the Gaussianity assumption is fulfilled with greater accuracy for some IDs and states than others. The histogram omits some skewness and bimodal behavior for IDs 1, 2 and 3 within the *high* regime, which corresponds to EV charging. The Quantile-Quantile plot in the right panel, suggests that there are tails of the distributions that are either too heavy or too light in order to be completely Gaussian.

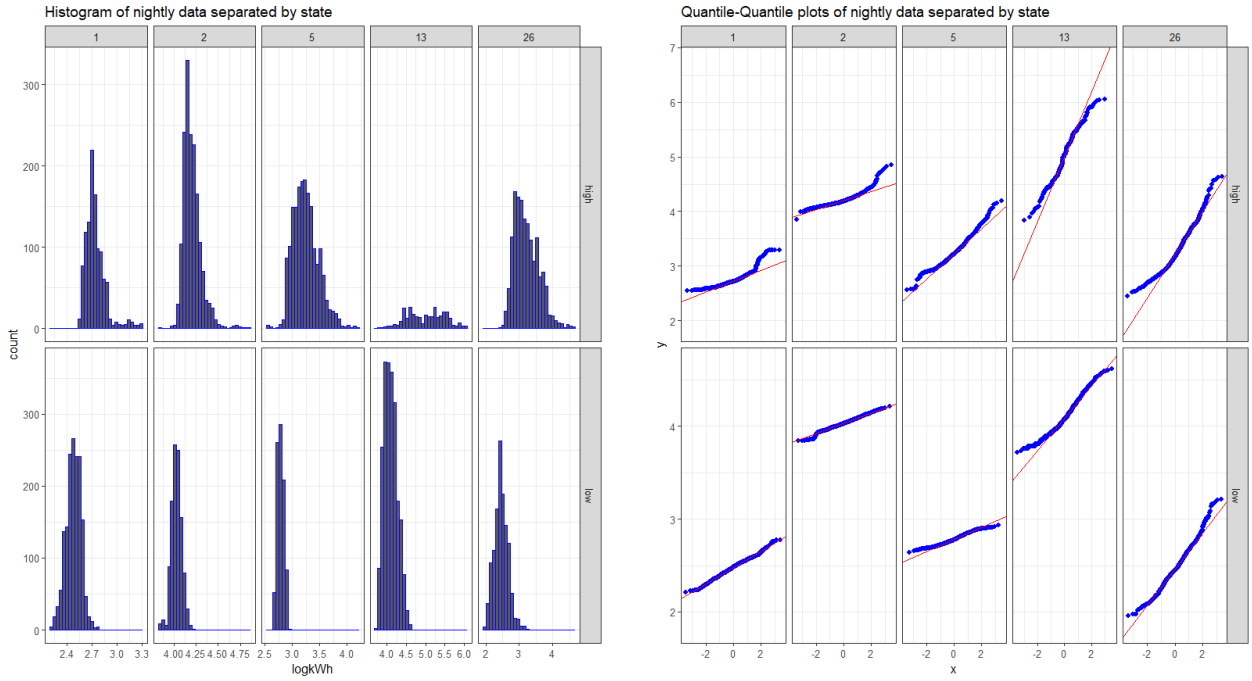

Figure 4.3: Histogram and Quantile-Quantile plots of candidate IDs' nighttime data separated by optimal state. The top row represents the charging behavior, while the bottom state represents non-charging behavior.

The estimated median from each station's *low* regime night time data is presented in Table 4.2, below. IDs 2 and 13 has the highest estimated median, which goes in line with the estimated hourly tenth quantile that omits the same behavior for the mentioned stations in Figure 4.1. The estimated medians are then used to remove the offset of the charging stations' full time series, which is shown in Figure 4.4.

Table 4.2: Estimated nighttime median consumption for selected stations.

ID	Median of $\frac{kWh}{T_i}$
1	0.00685
2	0.0920
5	0.0251
13	0.0725
26	0.00657

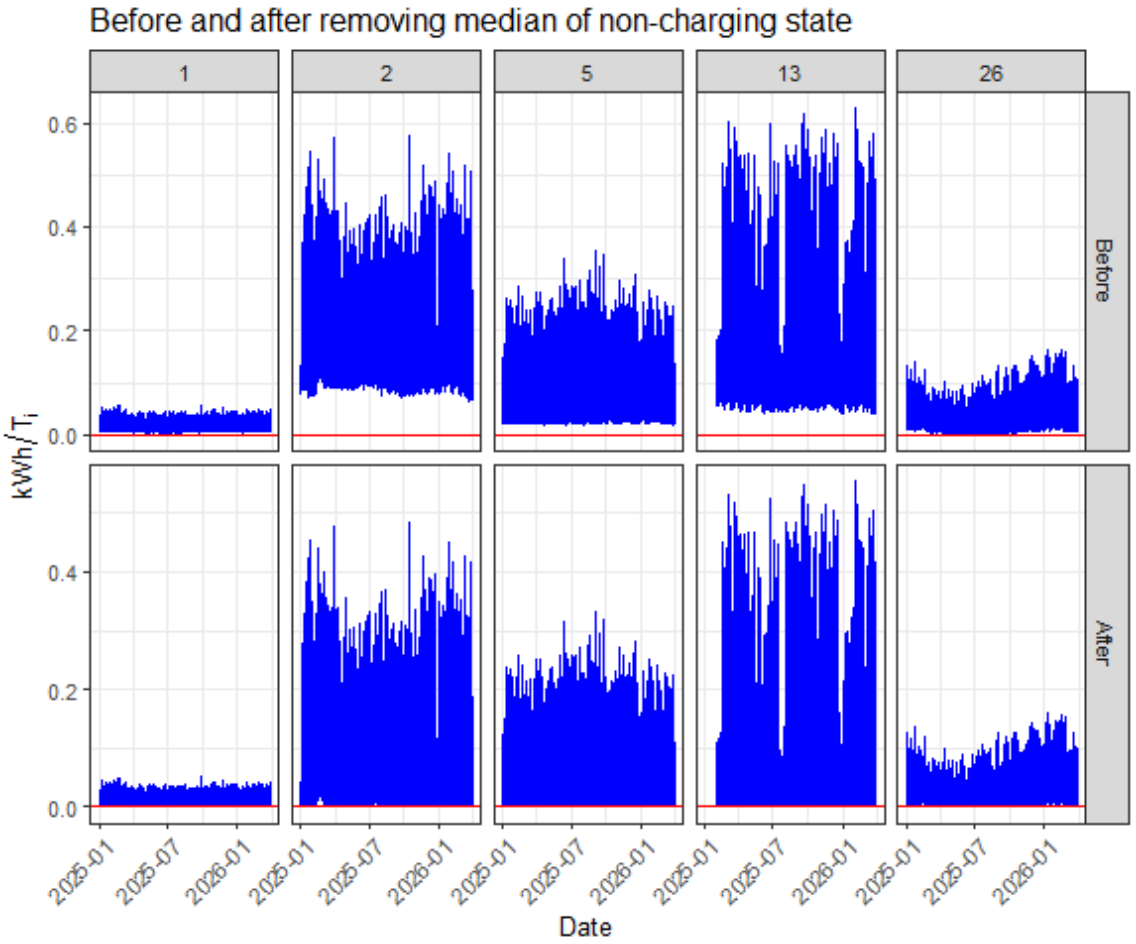

Figure 4.4: Candidate IDs' complete time series, before removing the nighttime median and after removing the nighttime median, from top to bottom in that order. The red line represents zero.

The time series have less of an offset from zero after subtracting the median, it is most evident for the previously mentioned IDs 2 and 13. Now that the potential non-charging behavior has been dealt with, the next step is to cluster the charging stations based on normalized hourly averages separated by weekday and weekend. The results of this procedure are presented in the next section.

4.2 Clustering Charging Stations

As mentioned in Section 3.5, finding the number of clusters K for the K-means algorithm is done by fulfilling two requirements. Firstly, that the MAE or RMSE is minimized between the predicted shape and the station's actual shape, and secondly, that the number of charging stations within each cluster exceeds one. The MAE and RMSE is calculated by taking the average of absolute differences and the average of the squared differences between the true shape and the predicted shape, and taking the root of the squared difference, where each calculated difference is the result of one fold during the cross-validation step. In Figure 4.5, the Mean Absolute Error and Root Mean Square Error between the true shape and the predicted shapes (Equation 3.2) is plotted for each cluster

size K. We see that the lowest MAE value is obtained when $K = 3$ and that the lowest RMSE value is obtained when $K = 10$.

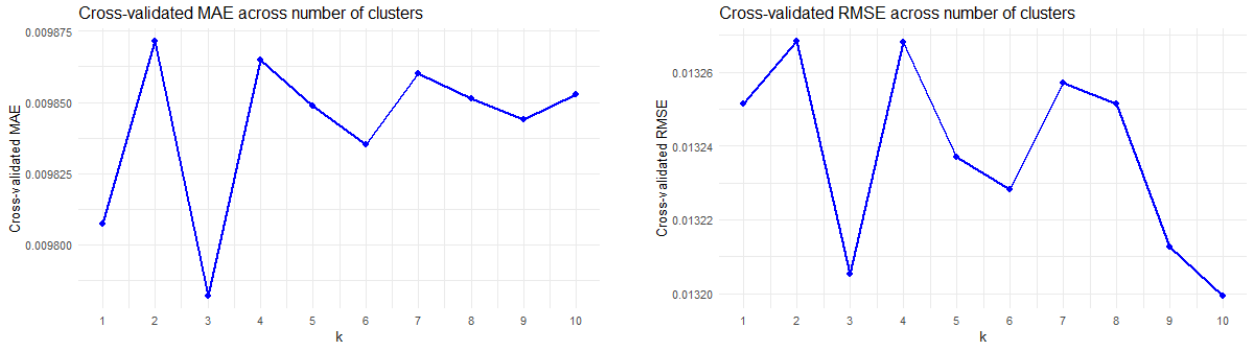

Figure 4.5: Mean Absolute Error and Root Mean Square Error for each cluster size K, calculated for each K based on the ten folds per cluster size.

Based on the first requirement, the number of clusters should be either three or ten. When examining the number of stations within each cluster when $K = 3$, there are no cases where clusters contained only one charging station, whereas the average number of clusters with a single charging station are 5.5 when $K = 10$. Therefore, $K = 3$ for the remainder of the thesis, where we could ensure that no cluster contained a single charging station. Since K is decided, we rerun the K-means algorithm on all of the data, in order to investigate the resulting cluster centroids when including all charging stations and the final value of K. The cluster centroids are showcased in Figure 4.6, below.

Figure 4.6: Final cluster centroids when running K-means algorithm on all charging stations' normalized hourly maxima separated by weekday and weekend. The weekday centroids correspond to the top row, whereas the weekend centroids correspond to the bottom row, $K = 3$.

We can see that the cluster shapes differ between weekdays and weekends. Cluster three

has a peak during earlier parts of the day during weekdays, while it omits a bimodal shape during weekends. We can see that the weekend shapes' peaks for cluster two and three are situated at earlier hours during the day, compared to the weekday shapes. The resulting cluster centroids are the components in the final model's weighted average shape, which enables many different combinations of the cluster centroids to fit the expected shape without assigning a new charging station to only one of the clusters. The number of charging stations within each cluster is presented in Table 4.3, below. We can see that cluster one and two has the majority of charging stations assigned to them. Cluster three only contains two charging stations, which explains the more jagged appearance in Figure 4.6, as a higher count of charging stations within a cluster tend to smoothen sharp peaks in the resulting centroid.

Table 4.3: Number of charging stations within each final cluster.

Cluster	Number of stations
1	19
2	12
3	2

The cluster centroids have been computed, and the next step is to train a model that predicts cluster membership probabilities of a new charging station, in order to generate a predicted shape, the results of this step is presented in Section 4.3

4.3 Classification of Cluster Membership

In the following sections during this chapter, the results are obtained from each LOOCV run, which was described fully in Algorithm 1. In the context of Table 4.4, this means that each charging station's cluster probabilities are predicted when running the K-means algorithm on the remaining charging stations with $K = 3$, and then training the forest classifier on the remaining charging stations' cluster labels. As the table suggests, for some charging stations, the cluster membership prediction of the model is relatively certain. For ID = 1, the cluster probability is maximized at 0.839. Other cases show the opposite, such as ID = 19, where the highest cluster probability is 0.501. The table shows why a less strict approach might be beneficial when the forest classifier is less confident in its predictions.

Table 4.4: Station-level cluster probabilities, where each column j correspond to the cluster membership probability of row i. The probabilities mentioned are in bold. Several cases where the maximum cluster probabilities are relatively low can be observed, such as IDs 21, 23, 25 and 30.

ID	Cluster 1	Cluster 2	Cluster 3
1	**0.008**	**0.153**	**0.839**
2	0.301	0.022	0.677
3	0.194	0.019	0.787
4	0.229	0.679	0.092
5	0.205	0.779	0.016
6	0.337	0.102	0.561
7	0.259	0.027	0.714
8	0.000	0.273	0.726
9	0.242	0.558	0.199
10	0.533	0.040	0.427
11	0.290	0.002	0.708
12	0.560	0.086	0.354
13	0.092	0.504	0.404
14	0.213	0.053	0.733
15	0.643	0.286	0.071
16	0.318	0.646	0.037
17	0.748	0.048	0.204
18	0.629	0.070	0.302
19	**0.399**	**0.101**	**0.501**
20	0.136	0.039	0.824
21	**0.565**	**0.058**	**0.377**
22	0.239	0.069	0.691
23	**0.528**	**0.435**	**0.037**
24	0.640	0.031	0.328
25	**0.289**	**0.114**	**0.597**
26	0.412	0.087	0.501
27	0.160	0.044	0.796
28	0.311	0.007	0.681
29	0.588	0.024	0.388
30	**0.409**	**0.099**	**0.492**
31	0.571	0.251	0.178
32	0.124	0.122	0.754
33	0.037	0.427	0.536

Something to note is that the resulting clusters when removing one charging station does not equal the final cluster centroids showcased in Figure 4.6. This is due to the fact that the final cluster centroids in Figure 4.6 are computed based on all data, while the cluster centroids are recomputed for each run when performing Algorithm 1 and removing the charging station that is being predicted. As a result, cluster labels in Table 4.4 does not necessarily indicate the same cluster between runs of the algorithm, as the labels are set randomly during the K-means algorithm. Aside from this, we can now move on to computing the weighted average described in Equation 3.2, and obtain the predicted weekday and weekend shapes for each charging station. The predicted shapes are presented in Figure 4.7, where they have been divided by their corresponding maximum value, in order to keep the peak of each predicted shape at one, as shown in Equation 3.7. We can see that the charging stations' predicted shapes differ by weekday and weekend, and that they often peak at different hours during the day. The majority of the predicted peak load EV charging seems to occur between hours 11 through 16.

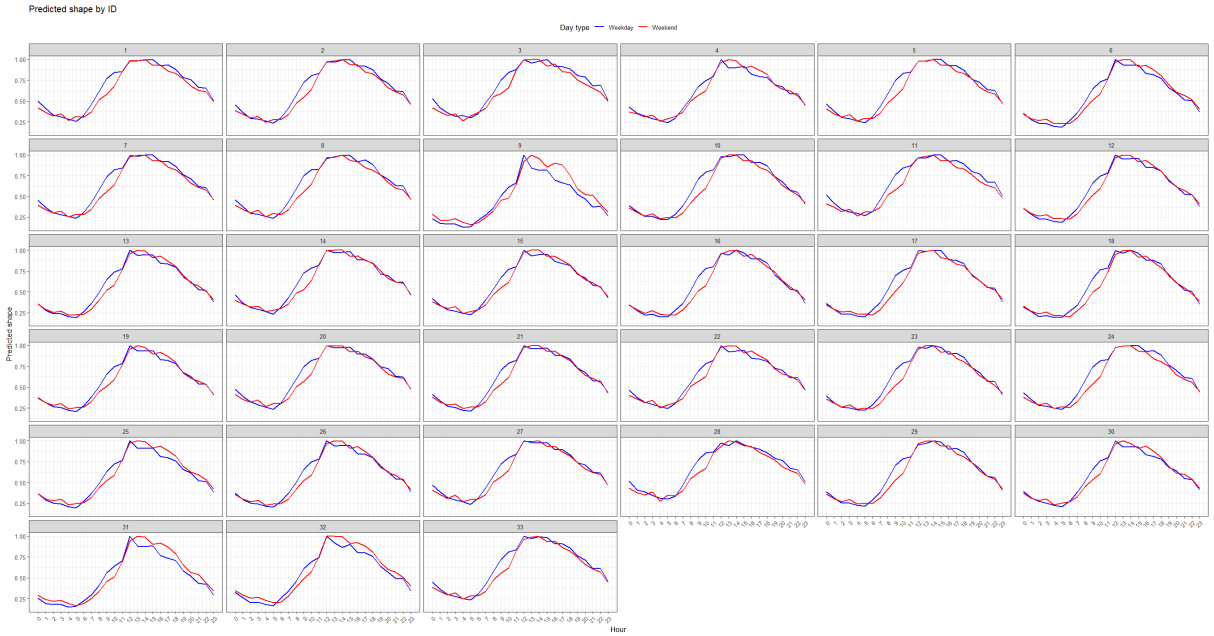

Figure 4.7: Predicted weekend and weekday hourly shapes for each charging station, where blue represents weekday, and red represents weekend. Each predicted shape is the result of one run of the leave-one-out cross-validation procedure.

At this point, we have removed potential non-charging behavior, computed three clusters of charging stations and constructed predicted weekend and weekday shapes for each run of Algorithm 1. To yield the final load profile, the quantiles of the daily maximum utility measure (Equation 3.3) is needed in order to scale the predicted shapes accordingly. The results of predicting daily peak loads are presented in Section 4.4, below.

4.4 Predicting Daily Peak Loads

As mentioned in Section 3.6, the daily peak loads are predicted by the quantile regression forest models that are trained during each iteration of Algorithm 1 with quantiles $\alpha = \{0.50, 0.75, 0.99\}$. The number of trees are set to 1500 for the quantile regression forest models. The predicted quantiles of the maximum utility $u_{i,w}$ are presented in Table 4.5, below. The resulting table contains all six predictions of the quantile regression forest models per algorithm run, since we have three quantiles of interest separated by weekday and weekend.

Table 4.5: Predicted quantile peak loads ($\hat{Q}_\alpha(u_{i,w})$) by station for weekday and weekend, where α is the quantile. Several stations obtain the same predicted quantile, where mentioned cases are in bold.

	Weekday			Weekend		
ID	$\alpha = 0.50$	$\alpha = 0.75$	$\alpha = 0.99$	$\alpha = 0.50$	$\alpha = 0.75$	$\alpha = 0.99$
1	0.159	0.216	0.555	0.148	0.190	0.448
2	**0.323**	**0.440**	**0.926**	**0.301**	**0.367**	**0.641**
3	**0.352**	**0.539**	**0.926**	**0.347**	**0.474**	**0.641**
4	**0.321**	**0.539**	**0.926**	**0.254**	**0.448**	**0.641**
5	**0.352**	**0.555**	**0.926**	**0.324**	**0.448**	**0.641**
6	0.176	0.200	0.441	0.154	0.208	0.448
7	0.187	0.237	0.356	0.188	0.227	0.347
8	0.321	0.352	0.555	0.254	0.324	0.475
9	0.153	0.216	0.555	0.111	0.188	0.474
10	0.321	0.352	0.555	0.248	0.302	0.641
11	0.330	0.352	0.555	0.274	0.301	0.540
12	0.200	0.278	0.440	0.224	0.303	0.447
13	0.237	0.330	0.926	0.182	0.274	0.641
14	0.216	0.321	0.926	0.248	0.324	0.641
15	0.229	0.321	0.926	0.235	0.301	0.641
16	0.229	0.242	0.484	0.227	0.248	0.474
17	0.187	0.217	0.484	0.188	0.224	0.448
18	0.153	0.200	0.555	0.177	0.208	0.448
19	0.216	0.242	0.555	0.188	0.248	0.448
20	0.352	0.555	0.926	0.347	0.474	0.641
21	0.323	0.355	0.926	0.324	0.367	0.641
22	0.226	0.264	0.555	0.250	0.281	0.641
23	0.200	0.237	0.539	0.224	0.248	0.540
24	0.159	0.242	0.555	0.177	0.250	0.540
25	0.159	0.217	0.555	0.148	0.177	0.448
26	0.151	0.217	0.926	0.154	0.190	0.641
27	0.440	0.484	0.555	0.347	0.474	0.540
28	0.330	0.539	0.926	0.301	0.448	0.641
29	0.153	0.200	0.555	0.148	0.190	0.448
30	0.264	0.440	0.926	0.274	0.448	0.641
31	0.226	0.242	0.555	0.208	0.248	0.540
32	0.216	0.237	0.440	0.177	0.248	0.474
33	0.226	0.323	0.561	0.250	0.303	0.641

We can see that several charging stations' predicted quantiles are similar in magnitude, IDs 2, 3, 4 and 5 have the same predicted value (after rounding to three decimals) for $\alpha = 0.99$ during both weekdays and weekends. The maximum daily predicted peak load is similar for several weekday predictions, and reaches a value of 0.926, where 1.000 is the theoretical maximum. At this point, we have computed each charging station's predicted shape $\hat{h}^{\star}_{i,w,h}$ (see Figure 4.7), and the predicted daily loads $\hat{Q}_\alpha(u_{i,w})$ (see Table 4.5). The final combined predicted peak load profiles $\hat{p}^{(\alpha)}_{i,w,h}$ of each charging station can then be obtained by taking the product of the two quantities. The resulting peak load profiles are presented and evaluated in Section 4.5, below.

4.5 Leave-One-Out Cross-Validation

We begin by investigating the predicted peak load profiles for weekdays where each station's hourly weekday maxima is marked by black crosses and the theoretical maximum is marked by a green dashed line at 1.00. The resulting figure is showcased in Figure 4.8, below. The predicted low-risk profile ($\alpha = 0.99$) exceeds the hourly maxima in all cases, except for the charging stations with IDs 9, 12, 13, 18, 20, and 27. The most evident case where the predicted profile fails to predict the peak load and shape, is ID 27, where

all but one observation exceeds the predicted low-risk load profile. The remaining predicted low-risk load profiles cover the complete range of hourly maxima, where some do so while leaving substantial room compared to the theoretical maximum at 1.00. In some cases, the low-risk profile exceeds the observed values by a relatively large magnitude, IDs: 1, 19 and 31 stand out since the low-risk quantile as well as the lower quantiles ($\alpha = \{0.50, 0.75\}$) exceed the observed hourly maxima in almost all cases. Aside from these cases, several other cases show that the lower quantiles follow the hourly maxima for many hours during the day, and where the low-risk quantile is relatively close to the hourly maxima. Cases that stand out in this regard are IDs: 4, 7, 9, 10, 12, 17, 18, 22, 23, 24, 25, 29, 31 and 32.

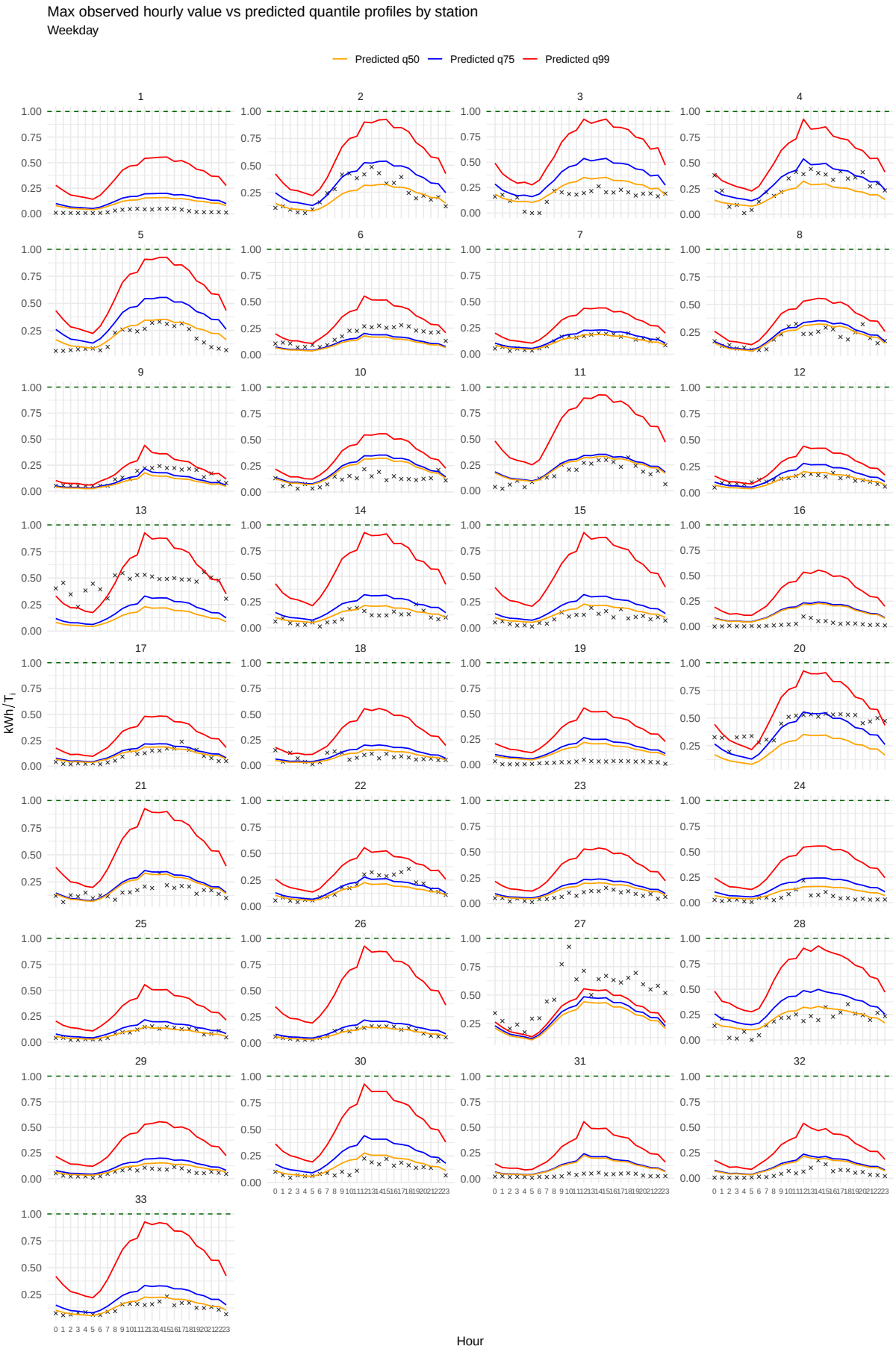

Figure 4.8: Predicted weekday peak loads for quantiles $\alpha = \{0.5, 0.75, 0.99\}$, where black crosses marks weekday hourly maxima and the dashed line marks the theoretical maximum.

The number of hourly weekday maxima that exceeds the predicted low-risk profile are presented in Table 4.6, below. The previously mentioned poor performance when predicting ID 27 is made evident, where all but one observation exceeds the low-risk load profile. The remaining charging stations have fewer exceedances compared to ID 27, where two IDs only contain one exceedance.

Table 4.6: Number of hourly weekday maxima per ID, that exceeds the predicted low-risk profile.

ID	Number of exceedances
9	1
12	2
13	9
18	1
20	4
27	23

The results of the predicted peak load profiles for weekends, where again, the black crosses marks hourly weekend maxima, and the green dashed line the theoretical maximum, are presented in Figure 4.9. Charging station IDs where hourly maxima exceed the predicted low-risk profile, consists of IDs 4, 6, 9, 12, 13, 18, 20 and 27. The two IDs with the highest count are 20 and 27, where the predicted low-risk load profile systematically underestimate the hourly maxima. The other charging station IDs' low-risk prediction cover the complete range of hourly weekend maxima. Cases where the low-risk quantile overestimates hourly maxima by a relatively large magnitude are less common, compared to the weekday case (Figure 4.8). The charging stations with low peak loads in the weekday case (IDs 1, 19 and 31) remain low in the weekend case, but the predicted quantiles' distance from the observed maxima are smaller. Cases where the lower quantiles as well as the low-risk quantile are relatively close to the hourly weekend maxima consist of the majority of charging stations, namely IDs 2, 3, 4, 6, 7, 8, 9,10, 11, 12, 14, 16, 17, 18, 21, 22, 23, 24, 25, 26, 28, 29, 30, 32 and 33. This suggests that the predicted weekend peak loads might be more accurate than the predicted weekday peak loads, where the trade-off is a higher risk of underestimating an hourly maxima.

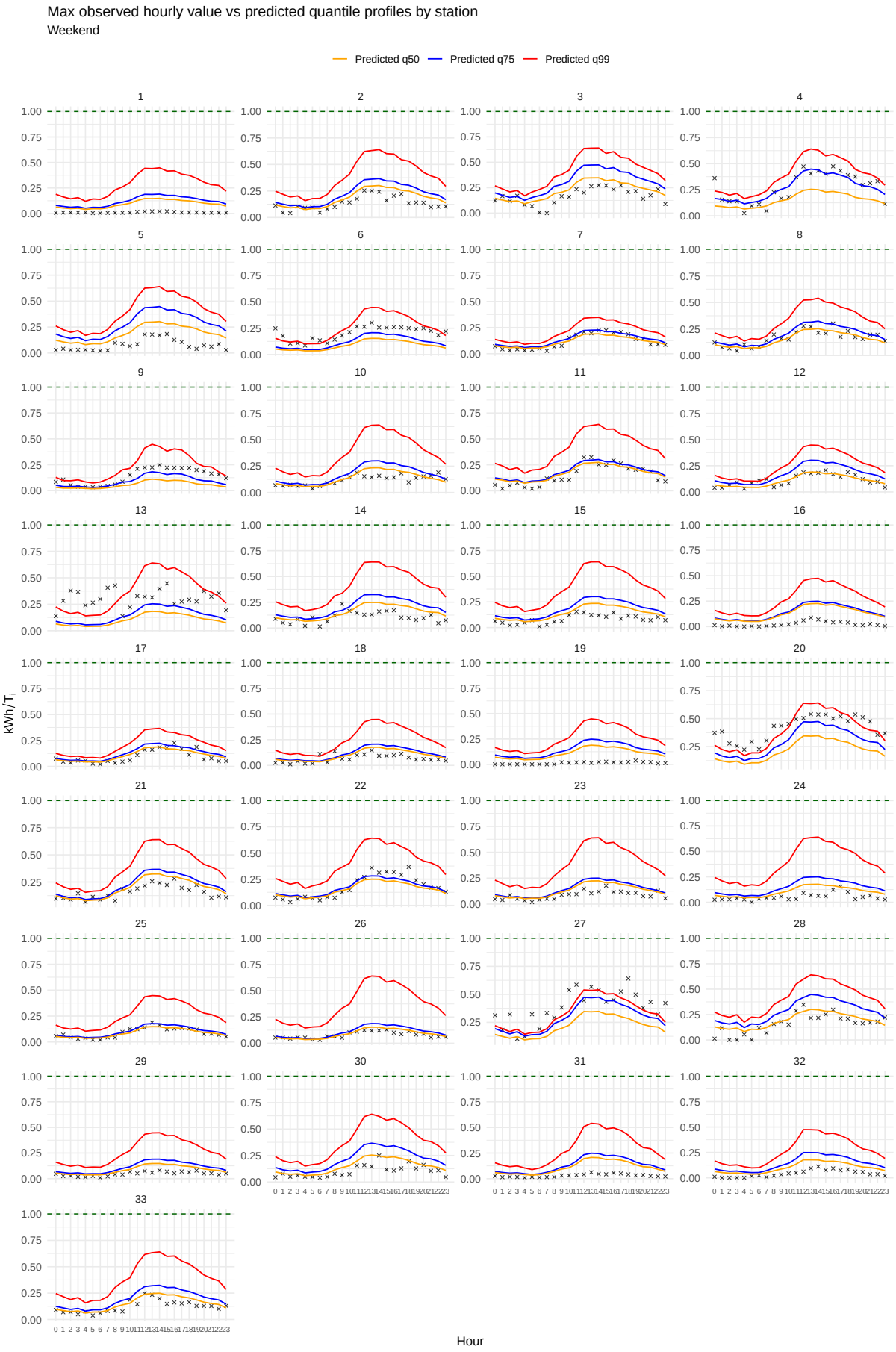

Figure 4.9: Predicted weekend peak loads for quantiles $\alpha = \{0.5, 0.75, 0.99\}$, where black crosses marks weekend hourly maxima and the dashed line marks the theoretical maximum.

The number of hourly weekend maxima exceedances are presented in Table 4.7, below. Similar to the weekday case in Table 4.6, ID 27 is underestimated during the majority of hours during a full day. A significant difference between the two, are that ID 20 has 15 exceedances in the weekend case, while only having four exceedances in the weekday case. The total number of charging stations with at least one exceedance also increase from six to eight, when comparing the weekday case to the weekend case. Something to note, is however, that four of the weekend cases consists of only one exceedance. The increase in number of stations with at least one exceedance, and the increase in some stations' exceedances, supports the claim that the weekend predicted low-risk profile incorporates a higher risk compared to the weekday predicted low-risk profile.

Table 4.7: Number of hourly weekend maxima per ID, that exceeds the predicted low-risk profile.

ID	Number of exceedances
4	1
6	5
9	1
12	1
13	9
18	1
20	15
27	16

Lastly, the metrics presented in Section 3.7, are calculated and presented for the station-specific low-risk load profile case, in Table 4.8. To briefly summarize, the Mean Absolute Error and Root Mean Square Error, measures station-specific average absolute difference and average squared difference (with the root taken of the final value) between predicted hourly maxima and observed hourly maxima. UnderRate measures the station-specific average rate of observed hourly maxima exceeding the predicted hourly maxima, where a value of zero equals no exceedances, and a number of one equals 100% exceedances. The Gain measures the station-specific average distance from the theoretical maximum and the predicted hourly maxima, where the theoretical maximum is 1.00 as all quantities are divided by the station-specific theoretical maximum T_i.

The risk and bias trade-off is made evident when inspecting the MAE, RMSE and UnderRate for IDs: 13, 20 and 27 in Table 4.8. The mentioned IDs have much lower values of MAE and RMSE compared to many of the remaining stations, but obtain a much higher value of UnderRate, meaning that a predicted curve close to the true values, often implies the risk of underestimating the hourly maxima. Some interesting cases are the station IDs where MAE and RMSE are relatively low, while keeping the UnderRate at zero, some examples are IDs: 2, 3, 8, 22 and 28. These predicted load profiles balance the risk well, since there are no exceedances that might result in damaging the electric grid, and while still optimizing the grid capacity better than the theoretical maximum. All of the predicted low-risk load profiles yields a better Gain when compared to using the theoretical maximum for all hours of the day. When inspecting Table 4.8, we can see that the minimum Gain is 0.48 for ID 3, and that the maximum Gain is 0.784 for ID 9, indicating that the grid is utilized more efficiently in all cases when using the predicted

low-risk load profile. However, some of the predicted low-risk load profiles, as mentioned, are systematically underestimated which should be kept in mind.

Table 4.8: Station-level LOOCV metrics for the predicted $\alpha = 0.99$ peak load profile. Rows where UnderRate exceeds zero by a relatively high magnitude are in bold.

ID	MAE	RMSE	Gain	UnderRate
1	492.0	529.0	0.674	0.000
2	180.0	197.0	0.509	0.000
3	106.0	119.0	0.482	0.000
4	29.7	34.5	0.527	0.021
5	223.0	240.0	0.502	0.000
6	221.0	270.0	0.714	0.104
7	252.0	277.0	0.753	0.000
8	98.2	112.0	0.657	0.000
9	483.0	594.0	0.784	0.042
10	215.0	246.0	0.643	0.000
11	209.0	229.0	0.489	0.000
12	459.0	534.0	0.744	0.062
13	**136.0**	**162.0**	**0.552**	**0.375**
14	359.0	403.0	0.510	0.000
15	411.0	455.0	0.527	0.000
16	1367.0	1506.0	0.702	0.000
17	421.0	474.0	0.751	0.000
18	509.0	599.0	0.715	0.042
19	659.0	721.0	0.709	0.000
20	**26.7**	**34.0**	**0.503**	**0.396**
21	191.0	223.0	0.525	0.000
22	174.0	184.0	0.630	0.000
23	381.0	421.0	0.648	0.000
24	757.0	826.0	0.632	0.000
25	482.0	530.0	0.712	0.000
26	592.0	666.0	0.545	0.000
27	**24.9**	**31.3**	**0.656**	**0.812**
28	102.0	113.0	0.485	0.000
29	394.0	437.0	0.699	0.000
30	312.0	351.0	0.541	0.000
31	1213.0	1382.0	0.717	0.000
32	793.0	868.0	0.719	0.000
33	347.0	392.0	0.511	0.000

The average metrics across all charging stations, separated by predicted quantile, is presented in Table 4.9, below. We can see that the median case ($\alpha = 0.50$) minimizes the average MAE and RMSE and maximizes the average Gain, while also maximizing the average UnderRate. When inspecting Figure 4.8 and Figure 4.9, the results are aligned with the resulting plots that show that many of the charging stations where the yellow line representing the median follows the black crosses closely. In many cases, the hourly maxima also exceed the yellow line resulting in a high average UnderRate. The $\alpha = 0.75$ case is naturally situated between the lower and upper cases, where mean MAE and mean RMSE is slightly increased, mean Gain is slightly lower and mean UnderRate is decreased. Lastly, the low-risk case, where $\alpha = 0.99$, then maximizes mean MAE and RMSE and minimizes mean Gain and UnderRate. The large increase in mean MAE and RMSE results in a much lower UnderRate compared to the other quantiles.

Table 4.9: Average LOOCV metrics across all stations by predicted quantile. As mean MAE and RMSE increase, mean Gain and UnderRate decreases.

Quantile	Mean MAE	Mean RMSE	Mean Gain	Mean UnderRate
$\alpha = 0.50$	106.0	121.0	0.856	0.3470
$\alpha = 0.75$	133.0	150.0	0.810	0.2170
$\alpha = 0.99$	382.0	429.0	0.620	0.0562

As seen in figures 4.8 and 4.9, ID 27 obtained the highest observed maximum utility by a large margin, especially during weekdays, and was systematically underestimated in the LOOCV results. In order to investigate this charging station's influence on all other charging stations' predicted peak load profiles, we excluded ID 27 and repeated the LOOCV procedure. The results are presented in Section 4.6, below.

4.6 Sensitivity Analysis

In order to simplify this procedure, the number of clusters were kept at three, and the only difference between this LOOCV procedure and the former, is that ID 27 was excluded entirely from the training and testing data. When comparing the weekday model in Figure 4.8 with the weekday model with ID 27 excluded (Figure 4.10), we can see that the magnitude of the peaks differ for many of the charging station's predicted low-risk peak load profiles. Examples that stand out are IDs 2, 3, 4, 5, 11, 14, 21, 26, 29, 30 and 33. All cases mentioned have lower peaks while still exceeding the observed hourly maxima in the majority of cases. However, IDs such as 13 and 20 have more observations outside the predicted low-risk quantile than when including ID 27. For ID 5, the 75th quantile and the 99th quantile result in the same prediction, causing only the blue line to be apparent.

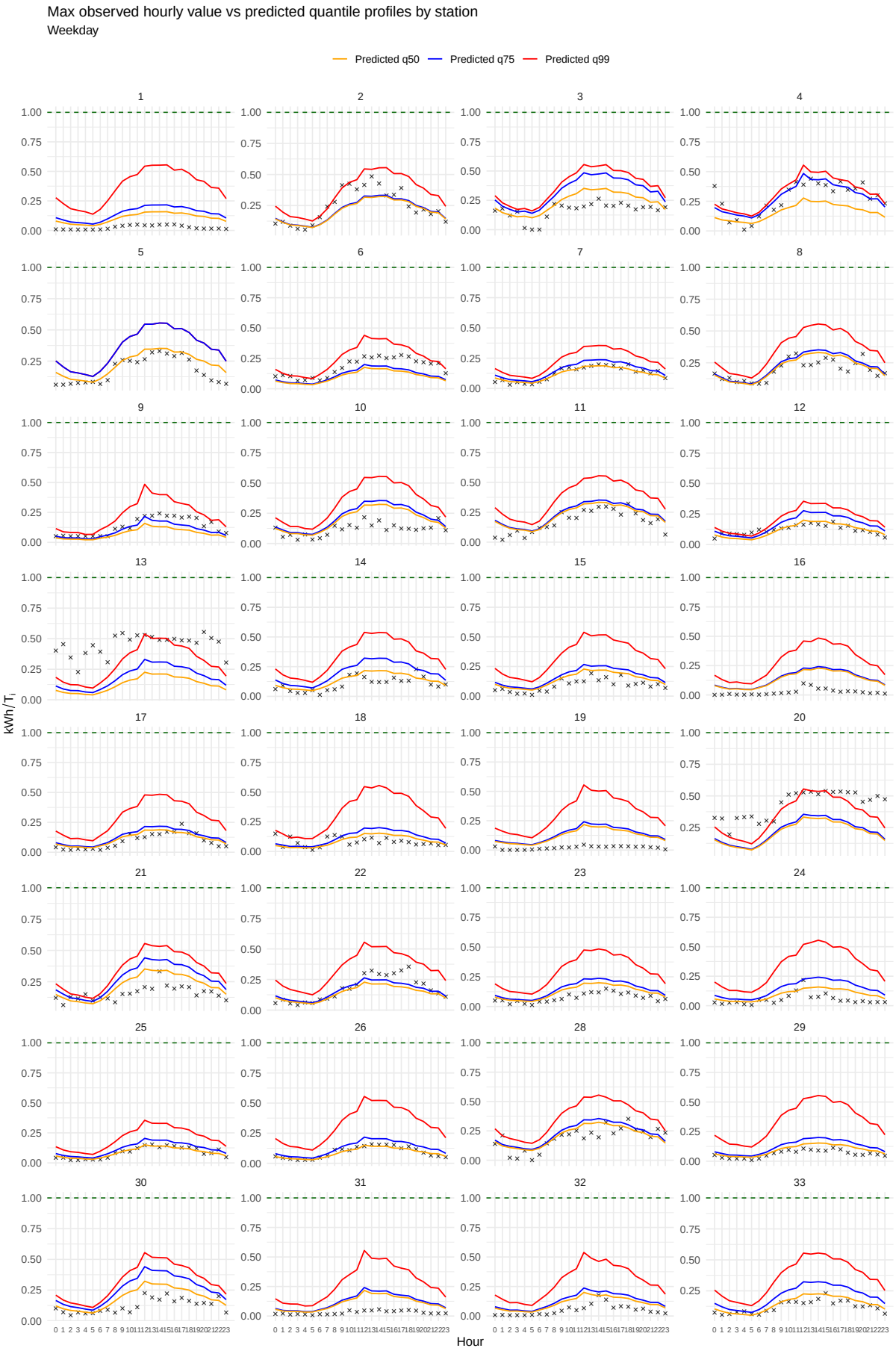

Figure 4.10: Predicted weekday peak loads for quantiles $\alpha = \{0.5, 0.75, 0.99\}$, when ID 27 is excluded, where black crosses marks weekday hourly maxima and the dashed line marks the theoretical maximum.

Moving on to the weekend case, there is a change in general magnitude, but the difference is less apparent when compared to the weekday case. The low-risk quantile in Figure 4.9, compared to the low-risk quantile when excluding ID 27 in Figure 4.11, shows similar results as in the weekday case regarding IDs 13 and 20. A greater amount of observed hourly maxima remain outside the predicted low-risk quantile. All other IDs, does however result in predicted quantiles that are closer to the hourly observed maxima while never underestimating the observed values during the most important hours of day, namely hours 10-17.

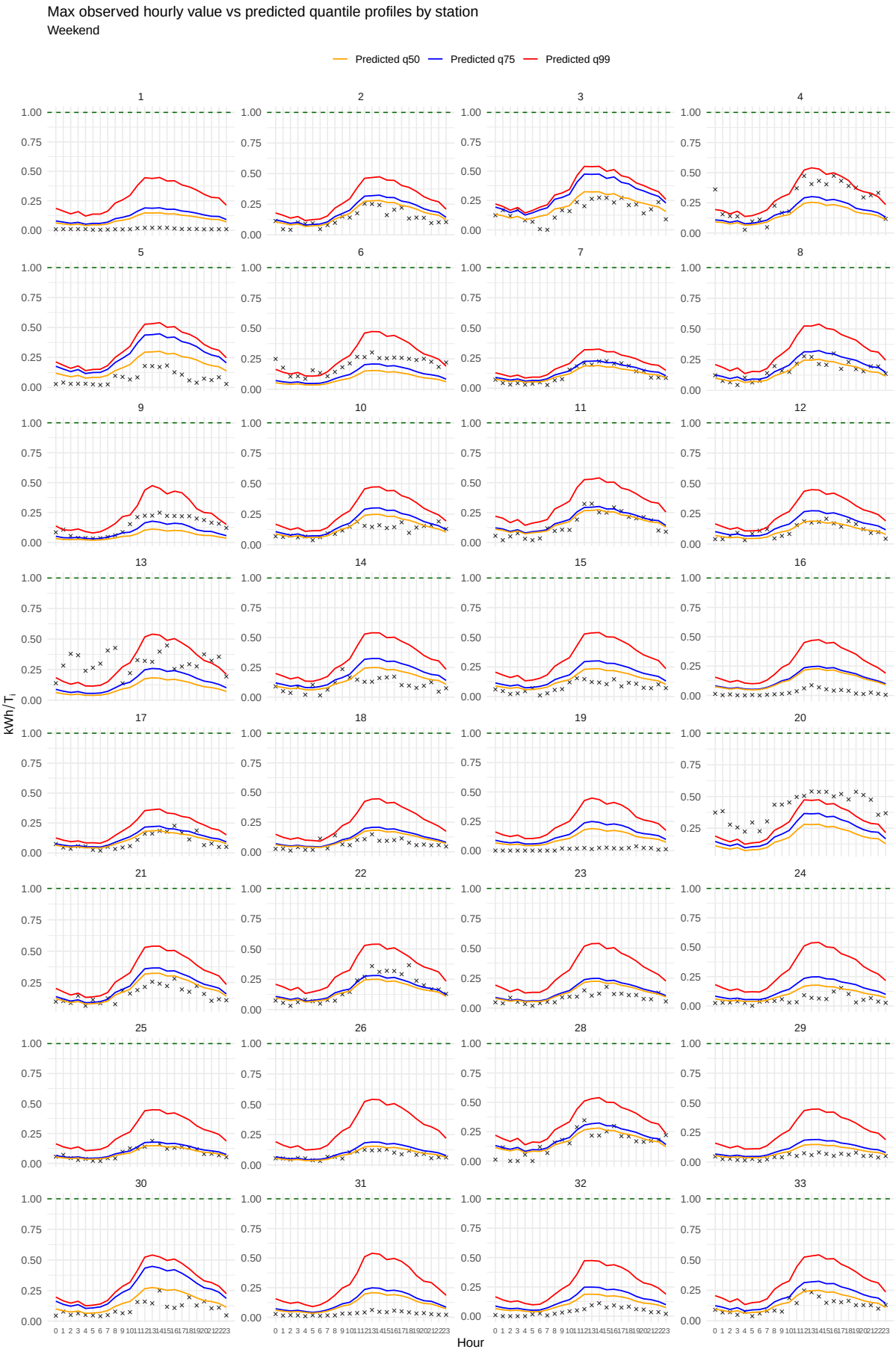

Figure 4.11: Predicted weekend peak loads for quantiles $\alpha = \{0.5, 0.75, 0.99\}$, when ID 27 is excluded, where black crosses marks weekend hourly maxima and the dashed line marks the theoretical maximum.

Lastly, the average metrics are presented for each quantile when excluding ID 27, in Table 4.10, below. When comparing the metrics to the previous case (Table 4.9), we see that the mean MAE and RMSE are barely changed for $\alpha = 0.50$ and $\alpha = 0.75$. The difference is greater when examining $\alpha = 0.99$, mean MAE and mean RMSE are substantially decreased, while mean Gain has increased somewhat. The improved performance for these metrics, does however reflect in the increased mean UnderRate, that increases from 0.0562 to 0.0638, which most likely is due to the worsened predictions for IDs 13 and 20.

Table 4.10: Average LOOCV metrics across all stations except for ID 27, by predicted quantile. Mean MAE and Mean RMSE are decreased for $\alpha = 0.99$, while mean UnderRate is increased from 0.0562 to 0.0638, when comparing to Table 4.9.

Quantile	Mean MAE	Mean RMSE	Mean Gain	Mean UnderRate
$\alpha = 0.50$	108.0	123.0	0.862	0.3310
$\alpha = 0.75$	133.0	151.0	0.826	0.2270
$\alpha = 0.99$	333.0	373.0	0.698	0.0638

Overall, the results indicate that the low-risk load profiles produce substantially lower UnderRate but at the cost of increased error and reduced grid optimization. Excluding ID 27 shows interesting results regarding the modeling frameworks sensitivity to outlier stations. As to why some charging stations' low-risk load profiles better balance the trade-off between Gain and UnderRate, compared to others, as well as potential explanations to previous sections' results, will be examined in more detail in the following discussion.

Chapter 5

Discussion

During this chapter, the discussion is split into three parts. The first section consists of the discussion of the results from Section 4.1, which mainly handles preprocessing of data, the second section discusses the final model and all relevant parts that it contains. Lastly, the third section consists of some topics that are outside the previous sections' scope, such as limitations, sensitivity analysis and ideas for future research.

5.1 Separating Signals

Starting off, the method for separating non-charging behavior from EV charging will be discussed. The method for screening for measure points including more than charging behavior seemed to be successful as the majority of the IDs investigated had an offset from zero. A weakness of the proposed method is however, that the regimes are estimated on nighttime data only. This approach does not measure any increase in non-charging power consumption that might occur during the day. An example might be a gas station that has more customers during the day, resulting in a higher level of power consumption since different devices are used to prepare items for sale etc. When using all of the investigated IDs' data, the hidden Markov model struggled to separate the data into two distinct states, likely due to the increased frequency of EV-charging, and potential increase in non-charging behavior, causing more frequent overlaps of the measured signals. When fitting the model on nighttime data, the regimes showcased in Figure 4.2, became more distinct, as charging is less frequent, and any non-charging behavior likely has less variance.

Using nighttime data holds another significant benefit except for easier state detection. If the HMM model were to be fitted on the complete set of data and charging behavior was classified as non-charging at a higher rate, this would then cause an increase in the estimated median of the non-charging state, which would in turn increase the risk of removing actual charging data from the charging stations' data history. The nighttime fitted HMM then, becomes the less biased choice as the estimated median will less likely remove actual charging data. As to choosing the quantity estimated from the non-charging nighttime regime, the median was chosen with the same reasoning, an average would be affected by potential outliers caused by wrong state classification. The removal of the median offset in Figure 4.4, showed that the chosen quantity did in fact remove the non-charging offset. There are, however, more novel approaches that are possible in this case,

perhaps a model that includes the autocorrelation between previous time steps would yield better results. The problem with such an approach is however that the regimes contain observations that are not continuous, and that the resulting model would be greatly influenced by any misclassifications of the two states.

Regarding the approximate Gaussian assumption for the two underlying states, Figure 4.3 showed that the resulting regimes for some stations satisfy the approximate normality assumption more often than others. In general, the Gaussianity is more apparent for data within the non-charging regime. This is likely due to the relatively stable power consumption that is displayed within this regime, where any outliers could be due to misclassifications of charging data. For the charging regime, data is less normally distributed, this is due to the less stable nature of charging occurrences. The charging regime distributions are either skewed, where tails of high values exist, or omit a bimodal behavior. This is likely due to the majority of nighttime charging consisting of a single EV charging, while in rare instances, more than one EV are charging at the same time, causing a heavy upper tail or bimodal behavior. In order to capture this, one could experiment with increasing the number of states, where a third state could represent charging behavior where the number of charging occurrences surpass one electric vehicle. A final remark that is worthy to note, is that the resulting optimal hidden state sequence is impossible to validate except for tools such as histograms and QQ-plots. This is due to the fact that some network stations within the electric grid, are connected to several sources of power consumption, while the measure point only gathers data at the network station level. In these cases, more detailed data granularity would be beneficial as the true signal could be extracted without having the need for filtering techniques.

5.2 Final Model

After fitting the K-means algorithm in Section 4.2, the cluster centroids in Figure 4.6 and the number of charging stations within each cluster in Table 4.3, showed that cluster three only contains two charging stations. As mentioned in the relevant section, the resulting centroid has a higher variance, since the average of many charging stations tend to smoothen sharp peaks. When examining the two charging stations within cluster three, no apparent patterns in the station meta-data is evident that explains why the charging stations differ from the remaining charging stations. They have differing amounts of charging points, average kW per point, distance to highway etc. When examining the charging stations within cluster one and two, stronger patterns emerge. Generally, cluster one contain charging stations with a lower count of charging points per station, where the station with the highest amount of charging points is 9. Cluster two, consists of charging stations with higher counts of charging points, where the highest count is 24. The resulting clusters then represent peak shape for smaller charging stations and peak shape for larger charging stations. In Figure 4.6, we see that cluster two has higher peaks during peak hours, while cluster one has higher peaks during less popular hours, such as early morning and night. This could be due to charging stations with many charging points having a higher frequency of charging during peak hours, since the charging station has a higher capacity of actively charging EVs. The two charging stations in cluster three, could then potentially be outliers compared to cluster one and cluster two, which would explain why two seemingly different charging stations would be clustered together.

Moving on to the predicted daily peak loads in Table 4.5, we saw that many of the low-

risk quantiles ($\alpha = 0.99$) were the same between charging stations. This is due to the structure of the random forest framework. Since the quantiles are estimated based on terminal leaves which consists of observations, a very high quantile will likely be the same for many charging stations, since their training data share the large majority of data which are used to form the predicted quantile. This is a weakness of the random forest models, they are unable to extrapolate outside of the range of the observed data. In this case, this would then mean that removing the charging station with maximum peak load utility ($u_{i,w}$ from Equation 3.3), results in predictions that are unable to achieve the same magnitude as the removed charging station. The issue of this will be made evident when examining the final model's results from Section 4.5.

In Figure 4.8, we saw the final model for weekdays. The most evident issue was for station ID 27, where all but one observation was outside the low-risk predicted peak load profile. The issue that was mentioned previously is the reason as to why this charging station's predicted load profile fails to capture the observed peak loads. When examining the maximum daily peak utility (Equation 3.3), defined here as the maximum station-specific daily kWh divided by the theoretical maximum capacity of that charging station, we see from Table 5.1 that ID 27 obtains the highest utility between all charging stations by a large margin. When omitting this charging station, the trained model has no extreme maximum utility observations except for ID 13 which only reaches a value of 0.555. A model that manages to extrapolate predicted values outside of the observations range would then be beneficial in this case.

Table 5.1: Maximum utility and first observed month by charging station ID, sorted by highest maximum utility. ID 27 achieves a higher magnitude compared to all other charging stations. Newly constructed charging stations tend to achieve lower magnitudes of maximum utility. Mentioned cases are marked with bold.

ID	First observed month	Maximum utility
27	**2025-01**	**0.92555000**
13	2025-02	0.55500000
20	2025-01	0.54016000
2	2025-01	0.48442667
4	2025-01	0.47430556
22	2025-01	0.36654667
28	2025-05	0.35180000
21	2025-02	0.33284000
5	2025-01	0.33048000
11	2025-01	0.32411667
8	2025-01	0.32316000
6	2025-01	0.30285714
3	2025-01	0.27408000
30	2025-01	0.25000000
9	2025-01	0.24833333
33	2025-01	0.24796610
17	2025-11	0.23740741
14	2025-05	0.23531330
7	2025-01	0.22388730
24	2025-11	0.21708333
10	2025-01	0.21645333
12	2025-01	0.20750958
15	2025-05	0.19223684
25	2025-02	0.18791667
23	2025-02	0.17724857
32	2025-01	0.17633136
26	2025-01	0.15851000
18	**2025-05**	**0.14916667**
29	**2025-04**	**0.11250000**
16	**2025-12**	**0.09867894**
31	**2025-03**	**0.06187500**
1	**2025-01**	**0.05087250**
19	**2025-10**	**0.04583333**

Other phenomena that was discussed during Section 4.5, was that of low hourly observed maxima. This behavior might spring from the fact that newer charging stations are less likely to be used for charging compared to more established charging stations. Within Table 5.1, the first observed month is presented as well. When examining the six IDs with the lowest observed maximum utility (1, 16, 18, 19, 29 and 31), all but one ID are set into production after 2025-01-01. The lower observed maxima could then be due to mobile applications for finding EV charging stations not having added the new stations

yet. Other reasons could be that individuals in the region have grown accustomed to utilizing certain charging stations that they know of, which would prioritize older well established charging stations. This is however, not true for ID 13 that has the second highest maximum utility of all charging stations, while being set into production during February 2025. Something to note, is also the fact that IDs 13, 20 and 27 all have a total of two charging points per charging station. These three charging stations obtain the highest maximum utility in Table 5.1, this is not surprising as two EVs charging simultaneously is not very rare, while ID 16 has a total number of 31 charging points. 31 EVs simultaneously charging at one charging station is however, a very rare occurrence. The mean maximum utility by number of charging points in Figure 5.1 further support this claim. We see that small charging stations generally tend to achieve higher values of maximum utility, on average, while mean maximum utility decreases as number of charging points increase, except for one charging station with 24 charging points.

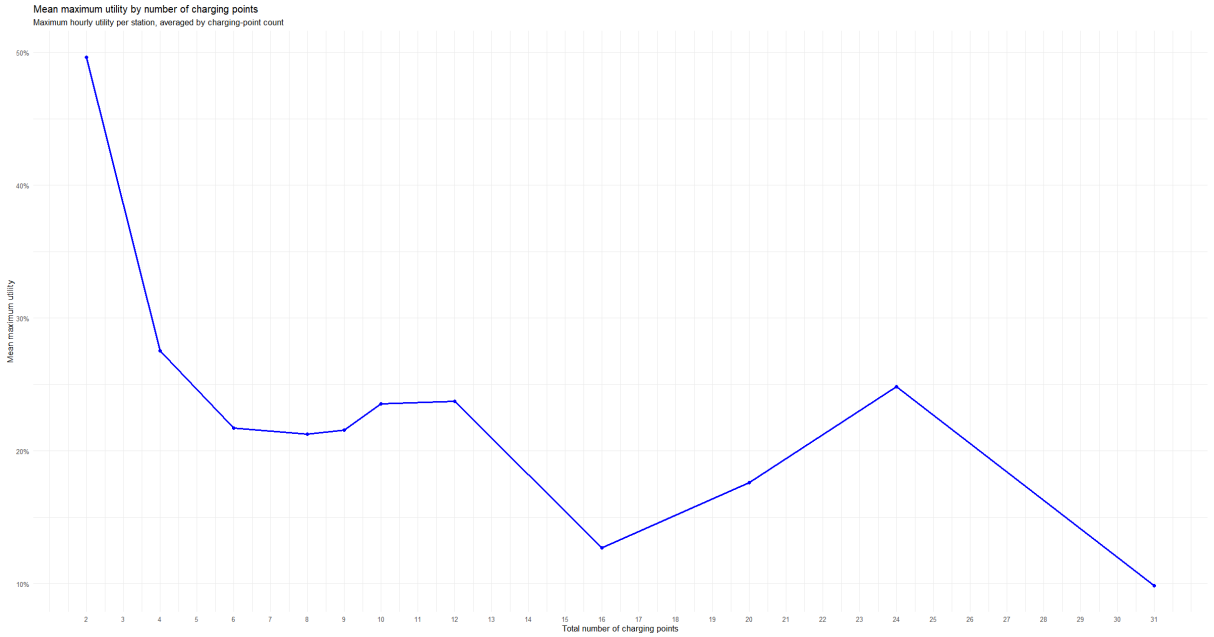

Figure 5.1: Mean utility of charging stations grouped by number of charging points, across number of charging points, where the blue dots represent the groups of charging stations. For charging stations with an unique number of charging points, the dot is simply the maximum utility from the previous table.

The final model for weekends, which was presented in Figure 4.9, generally had predicted low-risk quantiles closer to the observed values than the weekday case. This is due to the predicted daily peak loads in Table 4.5, that showed that the predicted weekend maxima were lower than the predicted weekday maxima. When examining the observed values for ID 27 in Figure 4.9, it is also lower than the corresponding weekday hourly maxima. This indicates that the range of observed values during weekends have lower upper bounds compared to weekdays. As to why some charging stations' predicted quantiles follow the observed hourly maxima better than others, is difficult to determine. When examining the station meta-data for station IDs such as 7, 8, 9, 10, 11 and 12, stations whom all have lower quantiles that follow the observed values well, no patterns emerge. The only attribute that holds for all mentioned charging stations, is that they belong to the group of charging stations that have been built before 2025-01-01, which is made evident when

examining Table 5.1. Since data is collected from 2025-01-01 onward, there are many stations that have existed for longer than the starting date, the reason for not including all data was mentioned previously in Section 3.2.

5.3 Further Discussion

An important caveat of the results presented in Chapter 4, is that the results are based of a sample size of 33 charging stations. If data from neighboring electric grids were to be utilized, many of the problems from the previous section would potentially be solved. If we had access to a greater number of charging stations, then station ID 27 would perhaps not be the only charging station that achieves a maximum utility of around 90%, this would further increase the accuracy of quantile estimation and reduce the risk of missing influential charging station behavior when performing LOOCV. Aside from the sample size, there also seems to be missing one or more predictors that explains charging behavior. As of now, ID 20 and 27 have the exact same, or very similar, station-specific meta-data except for the manufacturer of the chargers. Despite the similarities, their maximum utility differs by a large amount, as seen in Table 5.1. A predictor that explains frequency of charging would potentially benefit the performance of the model. Potential predictors that might result in such an improvement, would perhaps be historical charging prices from the predicted charging stations' manufacturers, traffic data or proximity to restaurants.

As mentioned previously, the random forest models can not extrapolate outside of the observed value's range. A potentially interesting approach would then be to test other models that do not have this issue. A model that can predict quantiles outside the data's range, is the Quantile Generalized Additive Model (QGAM), which was mentioned previously in the literature review during Section 1.1. The QGAM model was trained and tested throughout the same LOOCV procedure for predicting the daily peak loads, but failed to converge for very high quantiles ($\alpha = 0.99$), and when stable, increased the UnderRate significantly when compared to the quantile random forest model. Other models that generalize to unseen levels of data could however, potentially increase performance further.

The sensitivity analysis in Section 4.6, showed that ID 27 influences the remaining charging station's predicted quantiles. This is due to it's relatively high maximum utility, which was presented in Table 5.1. In order to recommend a safe implementation to Öresundskraft AB and any other future implementers of this framework, a reasonable approach would then be to investigate all charging stations with only 1-2 charging points first. Since these charging stations have a substantial risk of reaching total capacity of cars charging, a risk of very high maximum utility comes along with it. We saw, in Table 4.10, that the Gain, mean MAE and mean RMSE was improved when removing the outlier charging station. A reasonable approach would then be to assume the theoretical maximum, T_i, for all charging stations with only one or two charging points during peak hours. The next step would then be to remove any outlier charging stations that reach very high maximum utility from the training and testing data. Then proceed with the LOOCV procedure on the remaining charging stations' data, and use the resulting model for prediction of all new charging stations that do not have 1-2 charging points. When a new charging station is planned to have 1-2 charging points, always assume the theoretical maximum utility of 1.0 during peak hours. When utilizing this method, the risk is set to zero for high-risk

charging stations, since they can not exceed the theoretical maximum that the electric grid is planned for, while the grid is optimized for the remaining larger charging stations. The benefit with this approach, is then that risk is minimized for small charging stations, while the grid is optimized where there potentially are large gains, namely charging stations with many charging points.

Taken together, the results show that the proposed framework is able to construct station-specific low-risk load profiles that improve grid utilization compared to using the theoretical maximum capacity for all hours and all stations. At the same time, the discussion has highlighted several limitations that are important for interpreting the results. The separation of charging and non-charging behavior depends on assumptions that cannot be fully validated with the available data granularity, while the final model is limited by the small number of charging stations and by the inability of the quantile random forest to extrapolate beyond observed peak-utility values. These limitations are most evident for stations whose behavior differs substantially from the remaining sample, such as ID 27, and suggest that additional data, informative predictors, and alternative scale models could improve future versions of the framework. Despite these limitations, the results indicate that separating the prediction problem into shape and scale is a useful approach for estimating peak load profiles of new fast-charging stations. Lastly, a recommended approach for minimizing risk and maximizing grid capacity is presented. The main conclusions from this study, together with the practical implications of the proposed framework, are summarized in the following chapter.

Chapter 6

Conclusion

In this study, we have implemented a framework whose objective is to predict station-specific peak load profiles for new DC fast-charging stations. The problem setting requires that only static meta-data can be utilized as predictors, since a new charging station have no existing historical observations. The aim of the resulting predictions, is then to obtain a low-risk profile that improves grid utilization compared to using the theoretical maximum capacity of a new charging station. We have shown that estimating a new charging station's hourly peak load curve, can be separated into two more manageable subproblems.

The first being, predicting the shape of the curve, and the second being predicting the scale of the curve. Predicting the shape consists of clustering charging stations based on normalized hourly peak values, and then computing a weighted average of the cluster shapes, where the weights are the predicted cluster probabilities obtained by a classifier forest. The scale is then predicted by training a quantile regression forest on daily peak loads, where three different quantiles are estimated. The final predicted weekday and weekend peak load profiles are then the product of the predicted shape and scale.

The results when examining the final model's predicted peak load curves, showed that the low-risk quantile gave the lowest underestimation rate, and that the majority of observed hourly maxima falls beneath the predicted quantile. For lower quantiles, risk is higher, where there is a clear trade-off between potential underestimation and improved grid optimization.

Furthermore, the influence of outlier charging station behavior was evaluated. This led to a recommended approach of removing any such stations, while training the models on the remaining stations, in order to optimize grid capacity while simultaneously minimizing the risk of underestimation. Limitations of the study consists of sample size, outlier behavior and chosen models' inability to extrapolate outside of observed values' range. The proposed framework is to the author's knowledge, a new and unique approach to the problem of predicting future charging station behavior, when no historical observations from said station can be leveraged. Previous research shows that the area of new charging station behavior is a topic that needs to be explored further. As the number of electric vehicles increase, so does the demand for new charging stations as well, which in turn creates the need for data-driven tools specialized within this specific research area.

Bibliography

Amara-Ouali, Y., Hamrouche, B., Principato, G., and Goude, Y. (2025). Quantifying the uncertainty of electric vehicle charging with probabilistic load forecasting. *World Electric Vehicle Journal*, 16(2):88.

Betancur, D., Duarte, L. F., Revollo, J., Restrepo, C., Díez, A. E., Isaac, I. A., López, G. J., and González, J. W. (2021). Methodology to evaluate the impact of electric vehicles on electrical networks using monte carlo. *Energies*, 14(5):1300.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Callanan, A., Samuelsson, O., and Márquez-Fernández, F. J. (2025). A data-driven probabilistic framework for estimating grid impacts of ev charging at scale. *International Journal of Electrical Power and Energy Systems*, 172:111204.

Chang, M., Bae, S., Cha, G., and Yoo, J. (2021). Aggregated electric vehicle fast-charging power demand analysis and forecast based on lstm neural network. *Sustainability*, 13(24):13783.

De Livera, A. M., Hyndman, R. J., and Snyder, R. D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, 106(496):1513–1527.

Dunnington, D. (2025). *ggspatial: Spatial Data Framework for ggplot2*. R package version 1.1.10.

Forgy, E. W. (1965). Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications. In *Biometrics*, volume 21, pages 768–769.

Genuer, R. and Poggi, J.-M. (2020). *Random Forests with R*. Use R! Springer, Cham.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Applied Statistics*, 28:100–108.

Hijmans, R. J. (2026). *geosphere: Spherical Trigonometry*. R package version 1.6-8.

International Energy Agency (2025). Global EV Outlook 2025.

Kim, Y. and Kim, S. (2021). Forecasting charging demand of electric vehicles using time-series models. *Energies*, 14(5):1487.

Koenker, R. (2005). *Quantile Regression*, volume 38 of *Econometric Society Monographs*. Cambridge University Press, Cambridge.

Lin, X., Prabowo, A., Razzak, I., Xue, H., Amos, M., Behrens, S., and Salim, F. D. (2025).

Electric vehicle charging load modeling: A survey, trends, challenges and opportunities. *arXiv preprint arXiv:2511.03741*.

Lloyd, S. P. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28:128–137. Originally issued as a Bell Laboratories Technical Note in 1957.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Le Cam, L. M. and Neyman, J., editors, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, Berkeley, CA. University of California Press.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999.

R Core Team (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Skåland Webservice (2025). *NOBIL API: Version 3.0*. NOBIL / Enova. Revision 27.08.2025.

Trafikanalys (2024). Electrified Vehicles in Sweden.

Visser, I. and Speekenbrink, M. (2010). depmixS4: An R package for hidden markov models. *Journal of Statistical Software*, 36(7):1–21.

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC, Boca Raton.

Wright, M. N. and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17.

Zheng, J., Zhang, J., Jia, R., Song, P., and Zhao, S. (2026). Demand prediction of new electric vehicle charging stations: A deep learning approach. *Energies*, 19(2):378.