

Scalable probabilistic estimates of electric vehicle charging given observed driver behavior

Siobhan Powell ^{a,*}, Gustavo Vianna Cezar ^{b,c}, Ram Rajagopal ^{c,*}

^a Department of Mechanical Engineering, Stanford University, United States of America

^b Applied Energy Division, SLAC National Accelerator Laboratory, United States of America

^c Department of Civil and Environmental Engineering, Stanford University, United States of America

ARTICLE INFO

Keywords:

Electric vehicle
Charging behavior
Graphical model
Clustering
Long-term planning
Large-scale

ABSTRACT

To prepare for rapid growth in global electric vehicle adoption, grid and policy planners depend on detailed forecasts of future charging demand. In this paper we propose a novel holistic, scalable, probabilistic framework to produce large-scale estimates of electric vehicle charging load for long-term planning that capture real drivers' charging patterns. Our framework captures the uncertainty and stochasticity in charging demand by taking a graphical modeling approach. It has three core elements: driver groups, charging segment choices, and charging session time and energy requirements. The framework uses hierarchical clustering to group drivers by their charging histories, capturing their heterogeneous behaviors and preferences across different segments or types of charging. The framework uses probabilistic mixture models for each driver group's sessions to identify the unique charging behaviors observed within each segment. We illustrate its application with a large data set from California, profiling the charging patterns and unique driver clusters it identifies. Using the model knobs representing drivers' battery capacities, behavior, and segment access we present scenarios for California's charging demand in 2030 with 8 million passenger electric vehicles. Peak charging demand ranged from 3.3 to 8.7 GW across scenarios. Each was calculated in under 45 s on a laptop computer.

1. Introduction

Transportation accounts for a large and growing fraction of direct emissions around the world. Electrification is key to decarbonizing the sector and the global fleet of passenger plug-in electric vehicles (EVs) is expected to increase from 7 million today to over 400 million by 2040 [1].

Supporting charging at those levels will challenge the electricity grid [2], requiring equipment upgrades in the distribution grid [3–5] and adjustments to transmission-level capacity expansion planning [6–8]. The connection point between the grid and vehicles is another critical problem. A projected 300 million charging connectors will be needed to support the fleet in 2040 [1]. The convenience and availability of charging infrastructure at home, at workplaces, and in public impacts drivers' decisions to switch to electric vehicles [9,10] and much work has focused on the optimal placement of public stations to meet demand [11,12].

Long-term planning to face each of these challenges requires estimates of the future charging demand. Modeling driver behavior, including where, when, how, how much and how often drivers will charge, is critical to building accurate estimates. Scalability is also

critical to long-term planning models based on high adoption targets and is a key area of research [13].

Driver behavior comprises both travel and charging. Many studies characterize travel by clustering drivers by their travel histories [14–16] or building statistical models to generate new trip chains [17–20]. Many large-scale models in this category simulate travel for a given region from detailed survey data [21,22], but this agent-based approach is expensive and requires simulating a large number of variables for each driver. Travel-based approaches are also challenged to accurately model drivers' charging decisions from their travel.

Studies of real driver charging data have found many unexpected behaviors as drivers mix usage of multiple charging segments [10], including strong habits of regular users [23], tendencies toward charging close to home [24,25], and tendencies toward shorter more regular sessions to maintain high state of charge [26]. Charging choices are also influenced by driver characteristics including socioeconomic factors [10,27]. Many studies characterize individual charging decisions using logistic regression, e.g. [28–30], but this approach is computationally expensive to include at large scale in travel-based charging simulators.

* Corresponding authors.

E-mail addresses: siobhan.powell@stanford.edu (S. Powell), ramr@stanford.edu (R. Rajagopal).

Data-driven, statistical characterizations of charging data can reveal important behaviors [15,31]. A recent sessions-centered clustering approach revealed a wide typology of driver types and behaviors within public charging in the Netherlands [32]. This set of clustering approaches, however, has yet to be incorporated in large-scale models of charging demand.

As a result, current large-scale models either lack focus on driver behavior or use deterministic models for decisions which are inherently probabilistic. Recent attempts using probabilistic methods [31,33] have improved costs but depend heavily on modeler assumptions about drivers' use of different segments.

To address these challenges, in this paper we propose a novel holistic, scalable, probabilistic modeling framework for large-scale scenarios of charging demand which makes driver behavior core to its design. The methodology uses a graphical model with three core elements: driver groups, charging segment probabilities, and charging session behaviors. We propose a driver-centered clustering that highlights the role of multiple charging segments as a key feature of driver behavior in the United States. Using data-driven unsupervised methods, the model identifies distinct patterns in the charging data, and the probabilistic approach captures the inherent uncertainty in estimating future charging demand. The framework can be used to rapidly generate large-scale estimates of charging demand, and adjusting the distributions in the graphical model lets the modeler design a wide range of scenarios.

In this paper we use the methodology to present a case-study for California using a large data set of over 4 million real charging sessions with 30 thousand drivers covering workplace, public, and residential charging. We present a set of scenarios for California's charging demand in 2030 at 8 million passenger electric vehicles.

The main contributions of the modeling methodology are:

- the novel driver-behavior-focused method for clustering charging which reveals insights into driver charging patterns with mixed use of charging infrastructure;
- the data-driven, unsupervised, charging data-based approach to modeling driver patterns, which replaces modeled decisions used in other large-scale models with real driver behaviors;
- its modularity, which reveals novel behavior-based knobs for scenario design by separating driver groups and isolating charging patterns; and
- its probabilistic approach which provides scalability, reducing the simulation time for charging scenarios with millions of drivers, and which represents the inherent stochasticity in charging decisions across segments.

To allow others to explore different scenarios, and to facilitate adoption of the methodology by the modeling community, with this paper we are publishing our code open-source [34].

The remainder of the paper is outlined as follows: in Section 2 we present the framework and methods; in Section 3 we calibrate the framework with a data set from California, study the driver groups it identifies, and validate the model; in Section 4 we use the framework to generate sets of future charging scenarios; in Section 5 we discuss the results and approach; and in Section 6 we conclude.

2. Methodology

2.1. Graphical model

Our graphical model is depicted in Fig. 1 as a probability tree. Graphical modeling describes a large class of probabilistic models wherein independence and relationships between variables are encoded in a graphical representation [35]. In a graphical model, each variable is contained in one node and is dependent on the variables in its parent nodes: the joint distribution of all variables is the product of these conditional distributions [36]. Hidden Markov models and neural networks are two common examples.

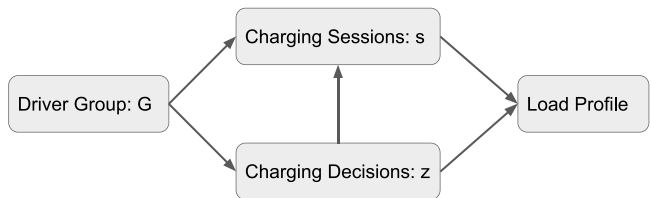


Fig. 1. Probability tree depicting our graphical model, connecting the three main elements: the driver groups, decisions between charging segments, charging session parameters. The resulting load profile is a function of the charging sessions and the charging rate for each segment. Controlled charging could be implemented through this framework by adding a step to adjust how the load profile is calculated from the uncontrolled charging session parameters; in this paper we focus on uncontrolled charging.

Our model arranges all drivers into groups and each driver has a group label, G . The distribution over driver groups, denoted $P(G)$, describes what fraction of the total number of drivers are in each group G .

Current technologies can be divided into Level 1 (L1) or wall-plug charging, Level 2 (L2) slow charging at charging stations, and Level 3 or DC Fast Charging (DCFC) at designated fast charging stations. These are typically 1.2 kW, 6.6 kW, and 50 kW or 150 kW in data from the United States. We consider six charging segments: L1 and L2 charging at a single family residence, L2 charging at a multi-unit dwelling, L2 charging at a workplace, L2 charging at a public charging station, and DCFC at a public charging station. Each segment is denoted by z , and the driver's decision to charge in each segment on a given day is modeled with a probability distribution conditional on the driver's group, $P(z|G)$. For each driver group and charging segment we build a model of the parameters which define each charging session: arrival time, energy, and duration. These are denoted by the vector variable s and the model for $P(s|z, G)$ is further described below.

The graphical model combines these three elements to prescribe the probability distribution for each charging session which occurs within the framework:

$$P(s, z, G) = P(s|z, G)P(z|G)P(G). \quad (1)$$

Once the session parameters, s , are known, the load profile can be calculated deterministically. The load profile is a time series in one minute intervals of the vehicle's charging demand. The demand is zero before the vehicle arrives and after the total session energy has been delivered. We assume charging occurs at the maximum rate for the given type of station. At this point controlled charging could be applied to influence the load profile, either through altering the start time of the session as in the case of residential timers or through modulating the charging rate between the drivers arrival and departure time, as in the case of workplace charging.

Rather than fit the model all at once with a form of the Expectation-Maximization algorithm, we chose to fit the driver groups and charging sessions elements separately. This approach let us customize the feature vector used to identify the driver groups, including important meta-data on the drivers such as their battery capacity and frequency of charging. Separating the steps also makes the framework more modular so it can easily incorporate components from other researchers.

The model focuses on the temporal aspect of charging demand. When applied to drivers from a given region, it estimates the charging sessions and aggregate charging load profile of those drivers without modeling where exactly their sessions take place. This result is suited to large-scale and generation-level grid planning and to large aggregations of drivers; planning for the specific locations of charging stations or for some distribution-level grid impacts would require further modeling to represent the spatial aspect of charging demand.

2.2. Driver clusters

We propose that drivers can be divided into distinct groups by their behaviors. These groups capture the heterogeneity in driver behavior better than models assuming the same preferences for all drivers, and reduce the dimensionality relative to the full spectrum of individuals.

We use clustering to group the drivers for our model. Clustering is a form of unsupervised machine learning which works by grouping similar samples [37]. To describe each driver as a sample in the clustering we create a feature vector based on the drivers charging history. This approach depends on having detailed data on the drivers charging history including the times, energy, rate, and location of each of their charging sessions over some length of time.

Let superscript d index the drivers such that x^d denotes the feature vector for driver d . To construct x^d , for each segment, z , we include: the total number of sessions the driver had in that segment, their mean session energy, duration, and start time when charging in that segment, and the fraction of their sessions in that segment which occurred on a weekend rather than a weekday. The feature vector also includes two driver properties: the battery capacity, B , of their vehicle and the number of unique zip codes, Z , in which they charge. This design captures both the drivers' choices between segments and their behavior within each segment; critical features which drive the drivers' ultimate load profile, grid impact, and infrastructure needs.

With our notation, s^d is the set of all sessions by driver d , and session $s_{i,z,w}^d$ denotes that the i th session by driver d occurred in segment z and on day type w . The session vector for session i includes the arrival time, a_i , energy, e_i , and duration, d_i :

$$s_i = \begin{bmatrix} a_i \\ e_i \\ d_i \end{bmatrix}. \quad (2)$$

Let $n_z^d = |s_z^d|$ be the number of sessions by driver d in segment z and let $n_{z,we}^d = |s_{z,u=we}^d|$ represent the number of sessions by driver d in segment z which occur specifically on weekends rather than weekdays. The fraction of driver d 's sessions in segment z which fall on a weekend rather than weekday can be calculated as

$$\alpha_{z,we}^d = \frac{n_{z,we}^d}{n_z^d}. \quad (3)$$

For each segment z we define a vector m_z^d describing the mean behavior of driver d when charging in that segment:

$$m_z^d = \begin{bmatrix} \frac{1}{n_z^d} \sum_i s_{i,z}^d \\ \alpha_{z,we}^d \end{bmatrix} \quad (4)$$

Then the total feature vector for each driver can be constructed as

$$x^d = \begin{bmatrix} m_{z_1}^d \\ \vdots \\ m_{z_M}^d \\ B^d \\ Z^d \end{bmatrix}, \quad (5)$$

where z_1 to z_M denote the different charging segments. The feature vectors are normalized before applying the clustering algorithm.

We apply a type of hierarchical clustering called agglomerative clustering which groups drivers from the bottom-up: initialization places each driver in their own cluster, then at each step the algorithm combines two clusters [38]. The decision of which two to combine is made so as to minimize the increase in the within-cluster sum of squares. Using the Euclidean distance, the objective is

$$\sum_{l=1}^L \sum_{x^d \in C_l} \|x^d - \overline{x_{C_l}}\|_2^2, \quad (6)$$

where C_l is the set of points in cluster l , $\overline{x_{C_l}}$ is the centroid of cluster l , and there are L clusters. This is known as Ward's method or

the minimum variance method [38]. We implement this algorithm in Python using the AgglomerativeClustering feature in the scikit-learn package [39].

The output is a tree of clusters which can be depicted using a dendrogram. The root node of the full dendrogram comprises the entire data set, the leaf nodes are individual data points, and the intermediate nodes illustrate how clusters are related: the clustering result is recorded by cutting the dendrogram at a particular level [38]. The lower down in the dendrogram two clusters are merged, the more similar they are [40]. The dendrogram can be interpreted in the context of graphical modeling: where G_i and G_j were two clusters joined at one step in the agglomerative algorithm to form the larger cluster G_k , the division of drivers into the two smaller clusters gives the conditional probability of being in cluster G_i or G_j given that a driver is in the larger cluster, G_k . The total sum of squares metric is also used to determine the optimal number of clusters, or the cut-off level for the dendrogram.

The driver groups defined by these clusters are foundational to our graphical model, and changing the distribution $P(G)$ is key to generating different scenarios.

2.3. Charging behaviors

The frequency with which drivers charge is highly variable across both different drivers and charging segments. We capture this with $P(z|G)$, the probability that a driver in group G will have a charging session in segment z on a given day. We calculate a separate distribution for each segment and for weekends and weekdays. Many driver groups have sessions in multiple segments so the different segments weekday probabilities, for example, $P_{wd}(z_{work} = 1|G)$ and $P_{wd}(z_{home} = 1|G)$, need not sum to one. From each driver's history we calculate their charging frequency as the ratio between the number of times in the year they charged in segment z on a weekday or a weekend, and the total number of weekdays or weekend days in the time period.

Given that a driver in group G has a charging session in segment z , we build a model for the parameters defining the session $P(s|z, G)$. Mixture models have been used in previous works to model charging parameter distributions, and Gaussian Mixture Models in particular have been applied to capture the joint distribution of multiple charging parameters [15,32,33].

A Gaussian Mixture Model (GMM) represents a distribution with a weighted sum of multiple multivariate Gaussians [41]. Where each component, k , is described by a mean, μ_k , and covariance matrix, Σ_k , where ϕ denotes the density function of a multivariate Gaussian distribution, and where π_k gives the weight of component k in the mixture, the likelihood of the GMM for sessions s is

$$\mathcal{L}(\pi, \mu, \Sigma) = \prod_{i=1}^n \sum_{k=1}^K \pi_k \phi(s_i | \mu_k, \Sigma_k). \quad (7)$$

GMMs are another form of unsupervised machine learning and can be viewed as a density based clustering method [37]. We implement this algorithm in Python using the GaussianMixture feature in the scikit-learn package [39]. This method is less sensitive to noise and outliers than non-density based clustering methods: with very stochastic charging data, this helps it identify the meaningful patterns behind the distributions [38]. To select the number of components we used the Akaike information criterion (AIC), a measure of model fit that trades off accuracy with simplicity.

We propose that the mixture components each describe unique behavior patterns in the charging sessions data. The mixture of these behaviors can be viewed in the context of graphical modeling: the sessions' parameters follow a Gaussian distribution, conditioned on their behavior pattern as represented by their component in the mixture. In that framing, the probability density of session s_i is

$$P(s_i) = P(s_i|k)P(k), \quad (8)$$

where the class-conditional probability density, $P(s_i|k)$, is that of the multivariate Gaussian, $\phi(s_i|\mu_k, \Sigma_k)$, and the probability of s_i coming from component k , $P(k)$, is the mixture weight, π_k , both from Eq. (7) [38].

Both distributions in this model are specific to the segment and driver group. Similarly to the distribution over driver groups, $P(G)$, changing this distribution over the sessions model components, $P(k)$, lets us adjust the frequency of specific charging behaviors for each driver group and segment to create new scenarios.

Gaussian Mixture Models are easier to interpret, scalable to large amounts of data and fast to generate sessions from a fitted model. The framework can also support other models of $P(s|z, G)$ such as Beta Mixture Models and non-parametric copula functions depending on the data set and application [25,30].

The final element in modeling charging behaviors is the impact of controlled charging. Control can be reflected in the model either by adjusting this model, $P(s|z, G)$, to reflect shifted sessions or start times adjusted by timers, or by adjusting how the load profile is calculated from the session parameters, s . That will not be a focus of the results presented in this paper as it is further analyzed in our other work [33], but its inclusion is supported within the model framework.

2.4. Simulation

The methodology outputs a model that can be used to simulate new, large-scale scenarios. Let D represent the total number of drivers in the simulation. First, we calculate the number of drivers to simulate in each group using the probability distribution over driver groups, $P(G)$, to generate the charging demand profile for a typical weekday. The number of drivers in group i is

$$D_i = P(G = G_i) * D. \quad (9)$$

Second, for each driver group, we calculate the number of sessions to generate for that day in each charging segment. To calculate the number of sessions, $S_{i,z}$, for driver group i and segment z , we sample D_i values from $P(z|G_i)$. Third, for each driver group and charging segment, we generate $S_{i,z}$ sessions using the mixture model of sessions parameters, $P(s|z, G_i)$. Fourth, we use the simulated parameters to calculate the uncontrolled charging load profile for each session. Finally, we aggregate the load profiles across all sessions and all driver groups to calculate the total uncontrolled charging demand.

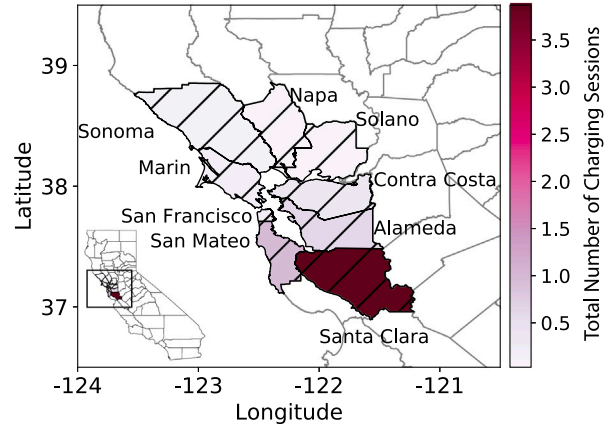
3. Experiments

3.1. Data

We apply the methodology to a large data set of real charging sessions collected by a charging station provider in California. The data was anonymized and each driver given a unique driver ID. For each of their sessions, in addition to the start time, end time, charging rate, and energy delivered, the data set also includes the date, county, and type of charging station where the session took place. The data includes charging sessions from both plug-in hybrid electric vehicles and battery electric vehicles. The data was cleaned to remove sessions with impossible energy or power values, sessions that lasted less than one minute or delivered no energy, and sessions with missing driver data or missing start times.

We analyze sessions which occurred in 2019 in California in the San Francisco Bay Area. Some data from the month of September is missing due to a data collection issue, so the time period of the study is 11 months. The majority of the sessions, 65.7%, were located in the largest county, Santa Clara, followed by 14.5% in San Mateo county and 8.8% in Alameda county. The distribution of the datasets' sessions over counties in the Bay Area is illustrated in the top panel of Fig. 2. We focus our analysis on drivers who charge regularly in the data set, with at least one session on average every two weeks. Since the data

Charging Sessions by County, Total Data



Charging Sessions by Category, Subset of Drivers

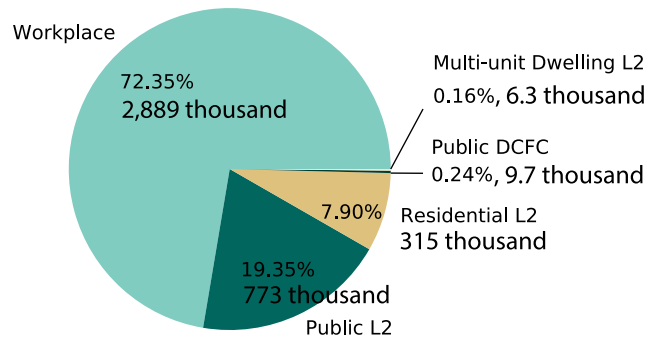


Fig. 2. The distribution of the data set over counties and types of charging. The inset map in the top panel shows the location of the San Francisco Bay Area in Northern California.

provider is not the only option for charging in the area, drivers who appear infrequently in our data set are likely sourcing the majority of their charging from another network. This leaves us with 38 thousand drivers with 3.99 million sessions.

Of the 3.99 million sessions, 72.4% occur at workplaces and over 32 thousand of the drivers have had at least one workplace charging session. 17.3% of the sessions occur at public charging stations and nearly 23 thousand of the drivers charge at a public station at least once in the data set. Nearly ten thousand of those sessions are at fast charging stations. 7.9% of the sessions occur at single family residences and 2.0 thousand drivers charge in that category. Finally, 2.5% of the sessions occur at multifamily sites and 3.7 thousand drivers charge in that category. This distribution is illustrated in the bottom panel of Fig. 2. This distribution is unique to the data set and reflects the charging network company's focus on workplace and public charging. The true distribution of charging today would include more residential charging and is discussed in Section 4.

We segment the data into five categories of charging: residential L2, multifamily L2, workplace L2, public L2, and public DCFC. This data set does not include L1 charging at residences as charging station providers primarily support L2 or higher, and additional sensing would be needed to capture L1 charging sessions.

3.2. Clustering

We calculate the feature vector detailed in Section 2.2 for each of the 38 thousand drivers based on their charging history. The results of the clustering are presented in a dendrogram in Fig. 3 and in

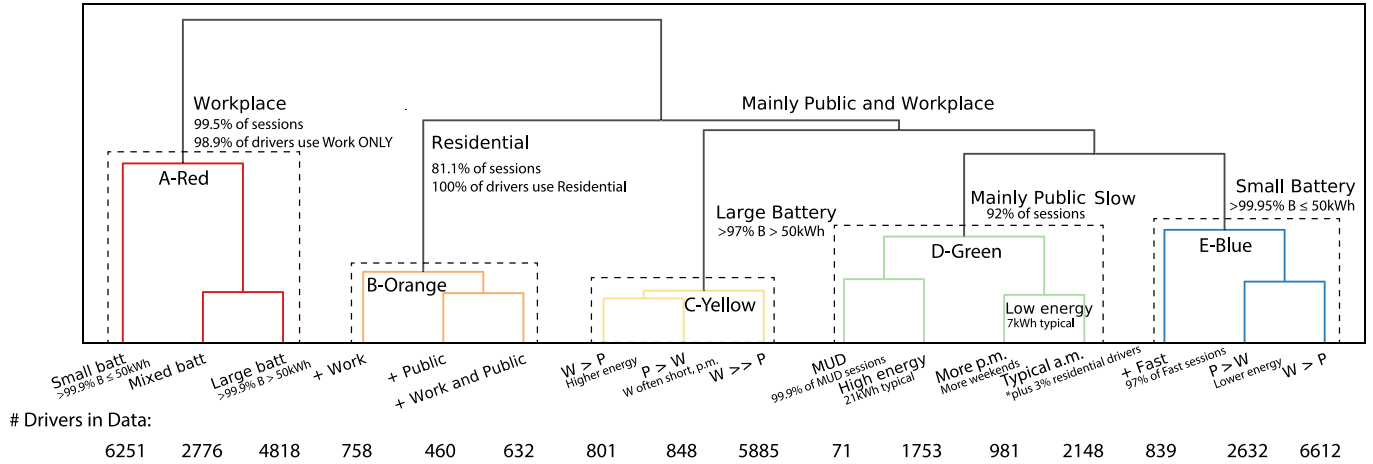


Fig. 3. This dendrogram illustrates the result of the hierarchical clustering with 16 clusters. Some statistics are annotated to support the labeling interpreting the clustering results. The abbreviation ‘batt’ refers to battery capacity. The symbol > indicates the relative frequency for driver groups that charge in multiple segments. The main divisions of the dendrogram into five “superclusters” are boxed, colored, and labeled as A-Red, B-Orange, C-Yellow, D-Green, and E-Blue. The number of drivers assigned to each cluster in the original data set is noted at the base of the figure.

load profiles in Fig. 4. An elbow curve of cluster inertia was used to select 16 as an optimal number of clusters. By comparing the drivers on the two sides of each fork in the dendrogram, we can interpret which differences between the drivers caused the clustering result. The agglomerative clustering algorithm is not rule-based so these labels are based on those observations.

Fig. 3 uses colors and the letters A, B, C, D, and E to highlight the division of drivers into five “superclusters”: A-Red, drivers who use workplace charging almost exclusively; B-Orange, drivers who use a significant amount of residential L2 charging; C-Yellow and E-Blue, drivers with typically large and small batteries, respectively, who use both public and workplace charging; and D-Green, drivers who use predominantly public L2 charging. Each supercluster is divided into smaller clusters as labeled in the figure, totaling the 16 unique driver groups.

The figure labels each cluster within each set to further interpret the key differences between them. Two small clusters stand out from their larger branches: the tenth cluster contains the small number of drivers who frequently use multifamily or multi-unit dwelling stations; and the fourteenth cluster contains the small number of drivers in the data set who account for the majority of public fast charging sessions. Another exception occurs in cluster thirteen where a very small cluster of residential drivers with early a.m. start times are grouped in with a larger cluster of public charging drivers. These would likely be a separate cluster if the total number of clusters were increased. Annotations on Fig. 3 give further details on the statistics behind these interpretations.

Our first observation is that the drivers preference between segments is very important. We also observe that battery size has a large impact on the clustering, driven by its connection to two other parameters in the feature vector, charging frequency and session energy.

Each cluster of drivers has unique characteristics, combining charging in multiple segments with different frequencies and patterns of behavior. To illustrate the profile of each cluster we present simulated load profiles for a typical weekday in Fig. 4. Each cluster is modeled on thousands of sessions so a single driver’s load profile would not represent the full distribution for the group. To capture the profile of the group as a whole, these plots show the aggregate load of 10,000 simulated drivers, normalized to keep units in terms of an individual driver.

We observe that many of the clusters comprise substantial amounts of charging from multiple segments. For example, drivers in cluster 6 charge at work, at home at single family residences, and at public charging stations. For another example, drivers in cluster 14 charge at work, at public L2 charging stations, and at public DCFC charging

stations. We also observe that the total energy, the integral of the load profile, is different between each cluster; each has different typical session energies and probability of charging on a given day.

3.3. Charging behaviors

The distributions of charging sessions behaviors were modeled with Gaussian Mixture Models, and AIC was used to select the optimal number of components between 4 and 8. Each component of the mixture can be interpreted as a unique charging behavior. For cluster 6, drivers who frequently use both workplace and single family residential charging, two mixture models are illustrated in Fig. 5.

The GMMs tend to generate smoother profiles than are observed in the raw data, as can be seen by comparing the simulated profile in Fig. 5 against the series sampled directly from the data, however the method can capture sharp discontinuities, like the timers starting at 11pm. If the number of mixture components were increased, it would be more difficult to interpret the charging behaviors, but the model would be able to better capture these smaller behavioral differences and irregularities in the profiles.

Looking at the specific behaviors in this example highlighted by the components of the mixture models, we make several observations. For residential charging, start times in the evening, either uncontrolled or with timers, are the most common behaviors and account for the two largest components with combined weight 0.68. Contrastingly in workplace charging, the behaviors are more evenly split over components with relatively similar weights. We do not see the sharp profile of a timer in the workplace charging components, but each has a different timing and magnitude of its peak. The peak is determined by both the weight in the mixture and the typical session parameters in that behavior: notably, the third and fourth components have the same weight and hence the same number of sessions, but those in the fourth require significantly more energy.

3.4. Validation

In addition to comparing the individual GMMs, we validate the model by comparing a simulated and typical day in the data. Fig. 6 illustrates the load in the system from the 38 thousand drivers for an average weekday and weekend day calculated by taking the mean over all the weekday and weekend aggregate profiles in the historical data. The comparable simulation, shown on the right, is constructed using the model for 38 thousand drivers with the base distribution over driver groups, $P(G)$. The root mean squared error (RMSE) between the profiles

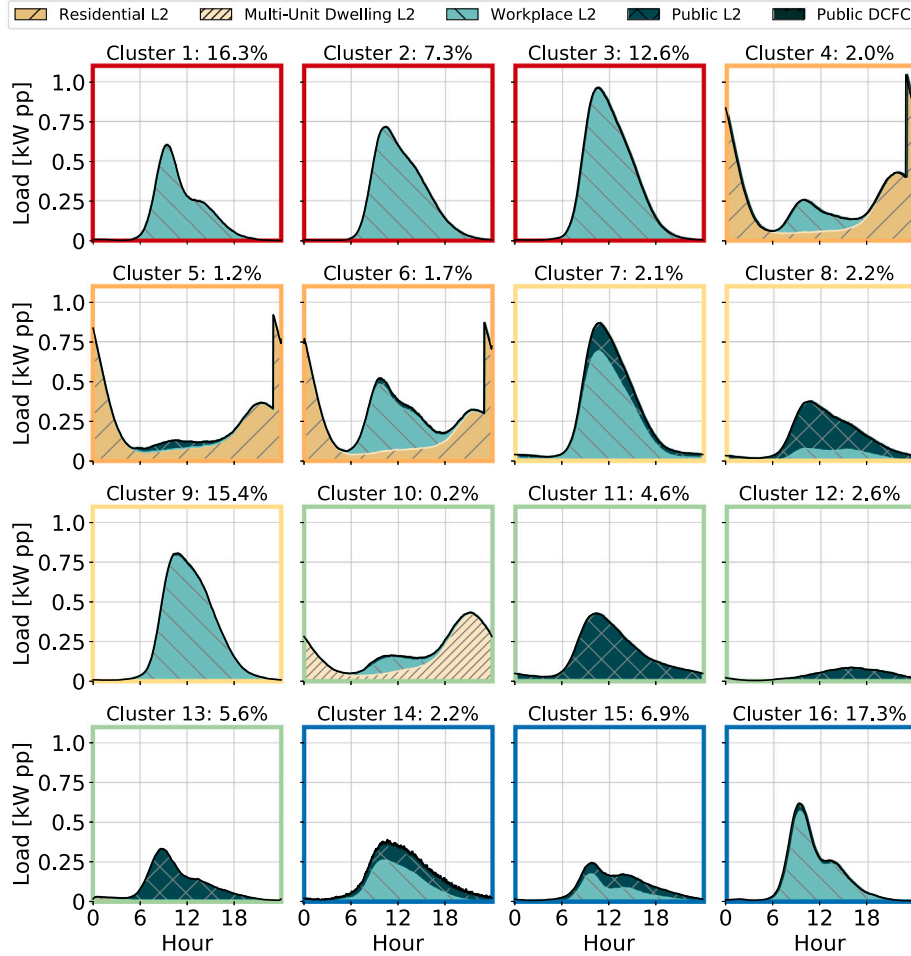


Fig. 4. This figure shows the typical weekday load profile for each driver group. The border of each plot is colored according to the superclusters defined in Fig. 3: clusters 1–3 in A-Red, clusters 4–6 in B-Orange, clusters 7–9 in C-Yellow, clusters 10–13 in D-Green, and clusters 14–16 in E-Blue. The title in each subplot notes weight of each cluster, $P(G)$. A charging rate of 150 kW is used for fast charging, Public DCFC, to match near-term projections.

is 1519.9 kW for the weekday model, or 5.72% of the peak value in the average weekday profile, and 306.2 kW for the weekend model, or 11.56% of the average weekend profile.

Studying the different driver clusters and GMMs we find that some are better modeled than others. Those for which we have less data tend to have larger errors as our picture of their charging behaviors is less complete and the mixture models fit to their charging sessions have a smaller set of training data to use for fitting. Error is introduced when we limit the maximum number of components for each GMM. We also do not account for holidays and seasonal variation; our implementation of $P(z|G)$ assumes each driver is equally as likely to charge on any weekday of the year. By including all weekdays, for example, our calculation of $P(z|G)$ underestimates the drivers’ likelihood of charging at work on a normal workday that is not a holiday. The weekend model error is higher because our data set has many fewer charging sessions on weekend days than on weekdays, as shown in Fig. 6.

The goal of our approach is to build a tool for testing assumptions and exploring changes in behavior. There is significant uncertainty involved in designing scenarios: this modeling error is small in comparison, and demonstrates the model can accurately capture the overall profile and trends in the charging data.

4. Results

The model framework breaks down the load into driver clusters and charging session behaviors: by changing the distribution over each of these building blocks, we can quickly generate a wide range of

scenarios for future charging demand. To illustrate this process we will present a set of scenarios for California’s passenger vehicle charging load in 2030 with 8 million individual EV drivers. Based on its projected trajectory toward having zero-emissions vehicles comprise 100% of light-duty vehicles sales by 2035, the state is planning for 8 million EVs in 2030; an increase over its previous target of 5 million [22]. The following scenarios are presented to demonstrate the model’s flexibility; it is not possible to show all possible scenarios in this paper. The code has been published open-source so that others can use the model to run scenarios of their own [34].

The first adjustment we make is to correct the base load: in California today, the majority of charging occurs at single family residences [21], whereas our base distribution as illustrated in Fig. 6 is dominated by workplace charging. This fraction is likely to decrease by 2030 as more investments follow in public and workplace charging, and as more drivers who do not have residential charging convert to EVs. The latest projections for California anticipate 67% of drivers in 2030 will have access to charging at their residence, split 90-10 between drivers in single family residences and multi-unit dwellings [22]. We make this adjustment by changing the distribution $P(G)$, increasing the weight of supercluster B-Orange which includes residential charging and the 10th cluster which includes charging at multi-unit dwellings. The other clusters have their collective weight in $P(G)$ decreased to 33%, while maintaining the same proportions.

We also adjust the fraction of drivers using residential timers. In our data set, 28% of the weekday residential charging sessions are exactly aligned with the lowest price periods in rate schedules offered

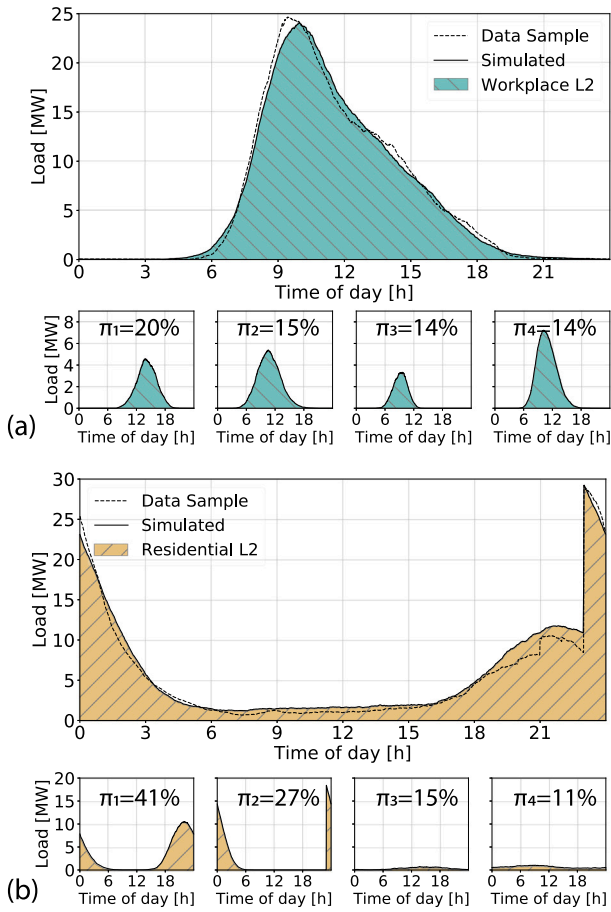


Fig. 5. Taking cluster 3 as an example, this figure shows the weekday load profile for 10 thousand samples simulated for the (a) workplace and (b) residential segments. Under each we illustrate the contribution from each of the four mixture components with the largest weights; the residential mixture had one component not shown with weight 0.057, and the workplace mixture had four components not shown with weights 0.131, 0.126, 0.102, and 0.009. The same number of sessions were randomly sampled from the raw data for drivers in that cluster: the black line shows that result for comparison.

to EV owners by the local utility. The adoption of residential time-of-use rates is increasing, however, and projections for 2030 anticipate 67% of residential customers will participate with residential timers [22].

That leverages one of the main model knobs: changing the distribution over the mixture components in a driver groups sessions model, $P(s|z, G)$, reweights different behaviors. To implement this change, for each of the three clusters with residential charging, we isolate which mixture component corresponds to a timer, either at 11pm or midnight, and increase the weight of those behaviors to 67%. This base case is shown in part (a) of Fig. 7.

Further exploring this flexibility in the model, we present a second scenario: if there were no residential time-of-use rates and all charging were uncontrolled, how would the load profile change? Working from the same base case, we again isolate which mixture components correspond to timers and shift their weights to the other components with evening start times. In Fig. 5 with cluster 6, that would mean shifting all the sessions from the second illustrated component to the first. The result is shown in part (b) of Fig. 7.

The residential segment need not be the only one where behaviors change: we create a third scenario, adding to the previous, where workplace charging behaviors change. If workplaces incentivized employees to begin charging after the noon-hour to spread out their load, how would that impact the load profile? To implement this we adjust the distribution over the sessions behavior components in the

two residential clusters from supercluster B-Orange which also have workplace charging, clusters 4 and 6, since together they comprise over 50% of the drivers in this scenario. Their workplace GMM components with afternoon start times are increased to a weight of 50% in each mixture, and the other components downweighted in proportion. The impact is shown in part (c) of Fig. 7.

Finally, we can hypothesize that the residential segment might change under widespread working-from-home. In each of the three clusters with residential charging there are mixture components representing morning, midday, and afternoon home charging behaviors. We implement work-from-home charging by redistributing some weight from evening to daytime behaviors in clusters 4, 5, and 6. The result is shown in part (d) of Fig. 7.

Comparing the four scenarios in Fig. 7 we see that assumptions about driver behavior lead to large magnitude changes in the resulting load profile: the addition of timers due to a residential time-of-use rate schedule leads to a spike of more than 7 GW with this number of drivers. The change to the workplace segment is also significant and shows how a simple shift in behavior could create a flatter load profile throughout the day. The change to daytime residential charging represents a more substantial change in behavior, but shows how flexible the load profile of charging could be in future scenarios.

A large number of scenarios can be simulated by adjusting the weights of each groups' charging behaviors. These four were selected to demonstrate the modeling flexibility this affords.

Driver groups are the other building block of our model and the distribution over driver groups, $P(G)$, will certainly change as we head toward 2030 scenarios. We present a second set of scenarios to illustrate how this can be represented with our model.

Whether a driver can charge at their residence has a strong impact on their charging load profile. We consider two cases to illustrate this effect: a low and a high home charging scenario. In single family residences, parking proximity and high installation costs for chargers prevent many from accessing home charging. It is harder still for renters and residents of multi-unit dwellings to access home charging. With deeper penetration of EVs in 2030, California will move beyond early adopters and could see such a scenario. For the low scenario, we reduce the fraction of drivers with access to home charging from the 67% base case to just 50%. The remaining 50% are distributed according to our data set, so this also represents a high workplace charging scenario. This is shown in parts (a) and (b) of Fig. 8. Part (a) shows the weekday profile and part (b) shows the weekend profile; with high dependence on workplace charging, we see that most charging demand falls on weekdays.

At the same time, policy interventions to support installation of charging stations at residences, especially multi-unit dwellings, could target increased access to home charging. How would those impact the load? For the high scenario, we hold the fraction of single family residential charging from the base case, but increase the fraction of drivers using multi-unit dwelling charging to 30%. This is shown in part (c) of Fig. 8.

The distribution over driver clusters need not be determined only by the use of different segments: if only the drivers in the dataset with vehicles with large battery capacities were represented, how would the load profile be changed? To implement this final scenario we make several adjustments to the low home charging high workplace charging scenario: within the supercluster A-Red we keep only cluster 3, and we shift all drivers using workplace and public charging from the small battery branch, supercluster E-Blue, to the large battery branch, supercluster C-Yellow. This is shown in part (d) of Fig. 8.

Studying the results in Fig. 8 we observe significant differences between scenarios. The multi-unit dwelling charging segment is very similar to the uncontrolled single family residential charging and the scenario in Fig. 7 part (a). This scenario, in Fig. 8 part (c), however, has the highest peak: the load spikes by more than 7 GW to reach 8.725 GW at 11pm. The scenarios in parts (a) and (d) of Fig. 8 with

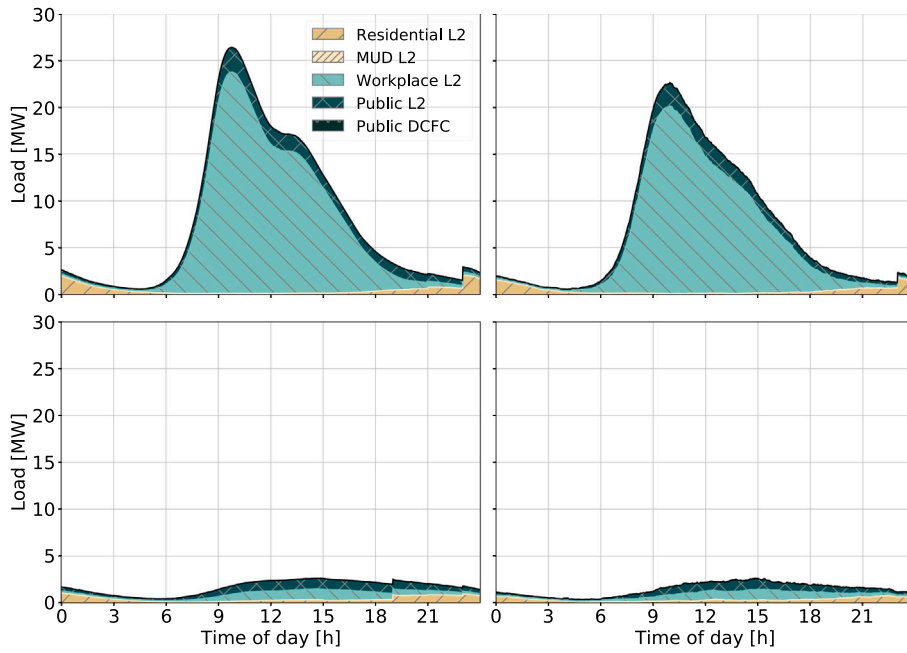


Fig. 6. Validation of the model comparing an average day calculated directly from historical data (left) with a simulated day calculated by the model (right). The top panel shows an average and simulated weekday; the bottom panel shows an average and simulated weekend day. The weekday and weekend profile are shown with the same y -axis scale to illustrate the split between weekday and weekend charging sessions in the data. Since this data set was collected by a charging station provider focused on workplace charging, the average day (left) includes more workplace charging than any other segment and as a result includes far fewer weekend charging sessions. Here a charging rate of 50 kW is used for fast charging, Public DCFC, to match the historical data.

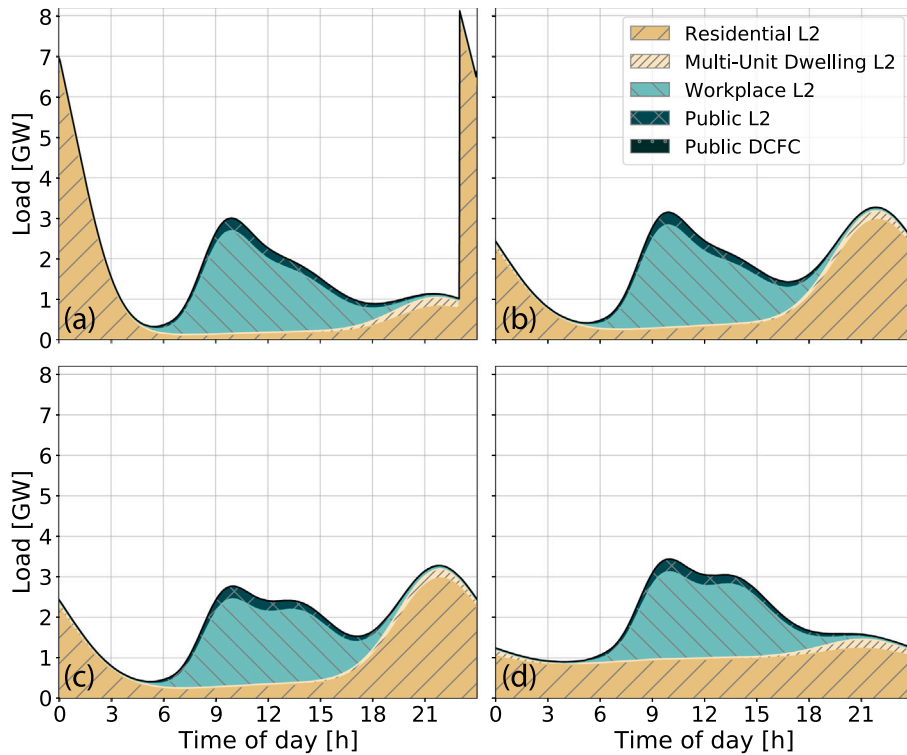


Fig. 7. This figure illustrates changing the distribution over behaviors in the sessions model, $P(s|z, G)$. The four scenarios: (a) a base case with residential timers; (b) uncontrolled residential charging; (c) the previous scenario with more workplace charging behaviors shifted to afternoon start times; and (d) shifting some evening residential charging to the morning and afternoon to simulate work-from-home charging possibilities, in addition to the afternoon workplace charging change.

more workplace charging present an interesting double peak, though the peak load in the morning still does not match the timer peak from residential charging in the evening. The scenario in part (d) of Fig. 8 with only large battery size drivers adds slight but significant

differences: in addition to higher energy consumption from workplace charging, there is a higher prevalence of charging in the afternoon.

Each of these scenarios with 8 million drivers was generated within less than 45 s on a laptop computer with the posted code. On the

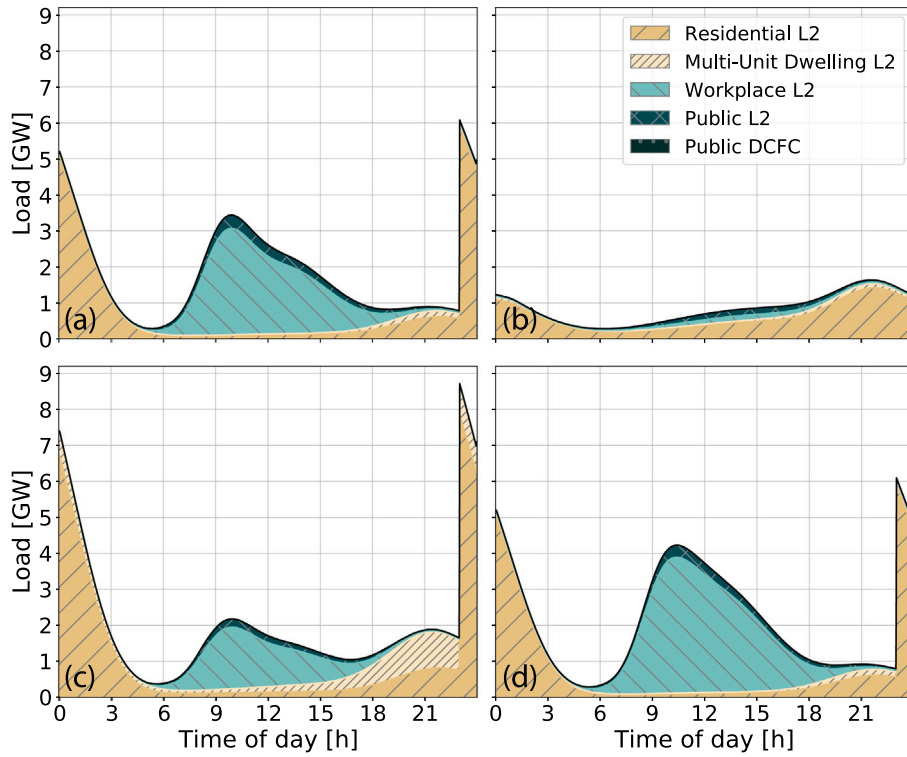


Fig. 8. This figure illustrates changing the distribution over driver clusters, $P(G)$. The three scenarios: (a) weekday and (b) weekend profiles for the low home charging access, high workplace charging scenario; (c) high home charging access with support for multi-unit dwellings; and (d) augmenting scenario (a) to consider only driver groups with large battery capacity vehicles.

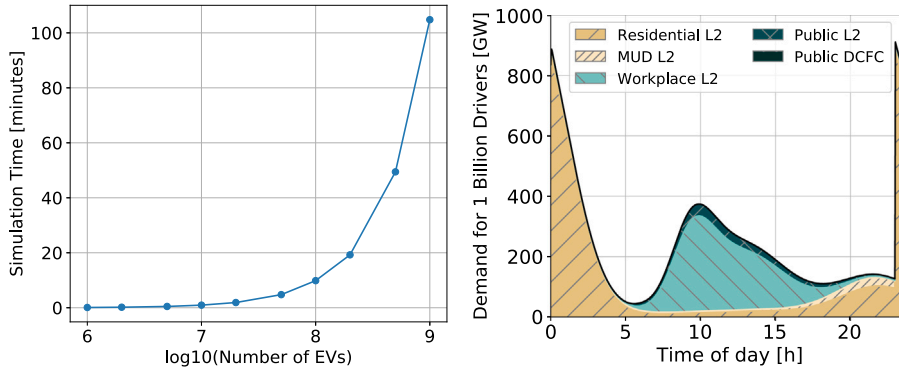


Fig. 9. This figure illustrates the model's speed: on the left, the simulation time taken to generate a typical weekday profile for the base case scenario with increasing numbers of EVs (x-axis shown on a log scale); on the right, the profile generated for 1 billion vehicles.

same computer, we tested how long it would take to simulate a typical weekday for the base case scenario with more than 8 million drivers. We observed that the simulation time increased approximately linearly with the number of drivers. The log-scale plot shown in Fig. 9 illustrates the result. The model could simulate the base case load profile, including all the individual charging sessions, for one billion drivers in just over 104 min. With one billion drivers, the daytime peak charging demand reaches 374 GW and the evening peak at 11pm due to residential timers reaches 910 GW. A sample of the times are reported in Table 1.

5. Discussion

The scenarios in Section 4 illustrated the impact of changing assumptions when generating scenarios to plan for the future demand from charging. The magnitude and timing of the peak, the implications for charging and grid infrastructure from switching between the

Table 1

The simulation time to generate a typical weekday profile under the basecase scenario increases approximately linearly with the number of vehicles from 6 s for one million vehicles to just over 104 min for one billion vehicles.

| Number of vehicles | Simulation time [s] |
|--------------------|---------------------|
| 1×10^6 | 5.8 |
| 5×10^6 | 28.3 |
| 1×10^7 | 57.0 |
| 5×10^7 | 286.3 |
| 1×10^8 | 591.6 |
| 5×10^8 | 2966.2 |
| 1×10^9 | 6287.4 |

segments, and the load shapes and ramping are all affected. As we plan further into the future, the magnitude of these differences only increases.

The results also illustrate the immense flexibility in the charging load. That includes traditional notions of flexibility like timers and automated control: a spike on the order of 7 GW is significant in a grid like California's where the typical daily peak is on the order of 25 GW. It also includes longer term notions of flexibility: investments made today in making types of charging infrastructure other than residential more widely available could reshape the daily load profile by 2030. The spikes introduced by timers operating at this scale, however, could pose severe challenges to the grid in terms of both ramping at the generation level and damage to equipment at the distribution level. The timer behavior occurs in response to electricity rates: future rates considering the impact of automated control should be designed to better distribute this demand over multiple start times.

The graphical model design of the framework enabled the ease of defining these scenarios by separating the charging into distinct elements: driver groups, segments, and behaviors. The distribution over behaviors could be adjusted separately within each driver group, and some groups could be adjusted without disturbing others. Representing each charging element with a statistical model also enabled scale and speed: the framework referred only to models of the distributions, not individual travel patterns or histories, so each scenario of 8 million drivers was generated within less than 45 s on a laptop computer. Scenarios with 5 million drivers were generated within approximately 30 s. The simulation time results shown in Fig. 9 and Table 1 demonstrate that the model could be used for much larger simulations. Simulating charging for larger areas like the entire United States would require substantial scenario design and would likely use different behavior distributions in different regions, but the total set of sessions and load profiles could be simulated using our methodology within well under an hour.

This flexible framework can be applied to many other datasets or use-cases and can easily integrate other components. Connections to other models will yield valuable results, as this framework can be used as post-processing to identify patterns, behaviors, and driver types within simulated data sets. It could also be coupled with other models of driver decision making, for example, to adjust the decision to charge to include additional options. In the absence of data for certain elements or driver groups, for example, in locations with very few EV drivers today, modelers could directly adjust the distributions and estimates to apply this tool, then improve the elements and behavior models as more data becomes available.

6. Conclusion

We proposed a novel framework based on graphical modeling for identifying patterns in the charging patterns of EV drivers and building estimates for future charging demand. This probabilistic approach is well suited to the uncertainty and inherent stochasticity in modeling EV charging. It is also highly scalable, enabling the generation of scenarios for 8 million EVs in under 45 s on a standard laptop computer. Scaling to beyond 2030 and beyond California, the model could simulate scenarios for 100 million vehicles in 10 min and for 1 billion vehicles in just over 104 min.

We applied this framework to a large data set of real charging sessions from California. Validation of the model comparing simulation results with average days in the historical data found good performance on weekdays with root mean squared error equal to 5.72% of peak demand. Performance on weekends for which the model had less training data was slightly lower, with 11.56% error. We used the model to identify unique driver groups and charging behaviors and used those results to generate a set of illustrative scenarios. We showed that the different driver groups are very distinct: changing their distribution has a substantial impact on any load scenario and they are critical to representing the heterogeneity of drivers in the data. The scenario with high access to home charging and high use of automated timers in residential charging yielded the highest peak

charging demand of 8.725 GW. By comparison, scenarios without timer control and with behavioral changes to better spread charging throughout the day limited peak charging demand to under 4 GW. This illustrates the substantial impact of assumptions about access to charging infrastructure, automated control, and driver behaviors on the final load, highlighting the importance of scenario-based planning.

There are many other scenarios to be run and the set presented here is not comprehensive. It is instead meant to demonstrate how the model can be used to consider different assumptions involved in long-term planning for EVs. The tool can be used to test other scenarios, and the approach can be directly applied to other data sets and regions in the world.

Future work will focus on studying trends of the change in driver groups and charging behavior as well as understanding what factors influence how particular drivers charge and fall into the model. Another key challenge is introduced by the number of knobs and variables in the framework: determining bounds on reasonable adjustments to the parameters for different case-studies will improve the utility of the scenarios. Another important problem is improving the accuracy of the model for critical subsets of days like holidays and for underrepresented behaviors using additional data sources. Such data would enable the model to incorporate day of the week, season, holiday, or region to the clustering hierarchy or build additional submodels. Augmenting or considering alternative statistical models to GMM could be useful to capture charging rate fluctuations during sessions and improve the accuracy of representation. Finally, future work to couple the model with data on driving patterns could extend the results to support station planning and represent the spatial aspect of charging demand.

To make this tool most useful to the community we have published the code to enable other researchers to try new scenarios, adjust assumptions, and interact with the framework themselves [34]. Running large scale scenarios is critical to planning to support charging at deep levels of EV penetration; by improving that planning, this tool can help accelerate the transition to an electrified, decarbonized future.

CRedit authorship contribution statement

Siobhan Powell: Conceptualization, Methodology, Software, Investigation, Writing – original draft. **Gustavo Vianna Cezar:** Data curation, Writing – review & editing, Supervision. **Ram Rajagopal:** Conceptualization, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank Noel Crisostomo and Matt Alexander of the California Energy Commission and Eric Wood of the National Renewable Energy Laboratory for their insights, guidance, and support of this project. The authors would also like to thank their colleagues in the Stanford Sustainable Systems Lab (S3L) at Stanford University led by Prof. Ram Rajagopal, the Grid Integration Systems Mobility group (GISMo) at SLAC National Accelerator Laboratory, and Prof. Charles Kolstad, Prof. Ines Azevedo, Lesley Ryan, Liang Min and the EV50 program at Stanford for their continued support. The authors would like to thank ChargePoint for the use of their data in this project. This work was funded by the California Energy Commission under grant EPC-16-057, by the Bits & Watts Initiative of Stanford University through the EV50 project, and by the National Science Foundation through a CAREER award (#1554178). SLAC National Accelerator Laboratory is operated for the US Department of Energy by Stanford University under Contract DE-AC02-76SF00515.

Table A.2

The mean log-likelihood score achieved by each probability distribution when applied to the total set of weekday start times, energies, and durations within each charging segment. The gamma and log normal distribution could not be applied to a small number of cases where there were many zero values, for example indicating an exact midnight session start time.

| | Residential L2 | | | MUD L2 | | | Work L2 | | |
|------------|----------------|--------|-------|--------|--------|--------|---------|--------|--------|
| | S | E | D | S | E | D | S | E | D |
| Normal | -3.42 | -3.86 | -3.38 | -3.02 | -4.07 | -3.37 | -2.6 | -3.77 | -2.5 |
| GMM 5 | -1.38 | -3.53 | -3.01 | -2.88 | -3.68 | -2.72 | -2.45 | -3.57 | -2.26 |
| Log Normal | - | -3.58 | -3.28 | -3.5 | -3.71 | -2.79 | - | -3.62 | -2.35 |
| Gamma | - | -3.52 | -3.14 | -3.24 | -3.7 | -2.82 | - | -3.57 | -2.3 |
| Gamma 5 | - | -3.49 | -2.2 | -2.88 | -3.66 | -2.65 | - | -3.55 | -2.25 |
| Laplace | -3.41 | -3.75 | -3.2 | -3.06 | -3.98 | -3.17 | -2.64 | -3.73 | -2.44 |
| Beta | -2.99 | -3.52 | -3.14 | -2.92 | -3.76 | -2.81 | -2.68 | -3.57 | -2.3 |
| Prob Table | -7.95 | -11.76 | -9.42 | -10.44 | -11.08 | -10.19 | -10.61 | -14.67 | -10.41 |

| | Public L2 | | | Public DCFC | | |
|------------|-----------|--------|--------|-------------|-------|-------|
| | S | E | D | S | E | D |
| Normal | -2.91 | -3.69 | -2.83 | -3.03 | -3.68 | -1.51 |
| GMM 5 | -2.81 | -3.38 | -2.3 | -2.98 | -3.45 | -0.49 |
| Log Normal | -2.94 | -3.41 | -2.34 | -3.39 | -3.52 | -0.52 |
| Gamma | -2.89 | -3.37 | -2.33 | -3.18 | -3.46 | -0.64 |
| Gamma 5 | -2.8 | -3.35 | -2.28 | -2.98 | -3.44 | -0.46 |
| Laplace | -3.01 | -3.58 | -2.56 | -3.12 | -3.6 | -0.83 |
| Beta | -2.88 | -3.37 | -2.35 | -2.99 | -3.46 | -0.63 |
| Prob Table | -10.9 | -13.11 | -10.35 | -8.7 | -8.1 | -7.96 |

Appendix

Table A.2 compares the log-likelihood result for different probability distributions used to fit the sessions parameters: start time (S), energy (E), and duration (D). For these tests we pooled the weekday sessions across all driver groups. The probability table and gamma mixture model were implemented in Python with pomegranate [42], the Gaussian mixture model was implemented with scikit learn [39], and the remaining distributions were implemented with scipy.stats [43].

Several distributions show comparable results, and other implementations of the methodology could use different statistical models. We used Gaussian Mixture Models because they let us model the start, energy, and duration together as a multivariate distribution and because they are numerically stable and fast to implement. Critically, as a mixture model and form of clustering, they added interpretability and let us breakdown the distributions into distinct behaviors which build in knobs for designing future scenarios.

References

- [1] McKerracher C, Izadi-Najafabadi A, ODonovan A, Albanese N, Soulopolous N, Doherty D, et al. Electric vehicle outlook 2020. Tech. rep., BloombergNEF; 2020.
- [2] Kapustin NO, Grushevenko DA. Long-term electric vehicles outlook and their potential impact on electric grid. *Energy Policy* 2020;137:111103.
- [3] Muratori M. Impact of uncoordinated plug-in electric vehicle charging on residential power demand. *Nat Energy* 2018;3(3):193–201.
- [4] Coignard J, MacDougall P, Stadtmueller F, Vrettos E. Will electric vehicles drive distribution grid upgrades?: The case of california. *IEEE Electr Mag* 2019;7(2):46–56.
- [5] Brinkel N, Schram W, AlSkaif T, Lampropoulos I, van Sark W. Should we reinforce the grid? Cost and emission optimization of electric vehicle charging under different transformer limits. *Appl Energy* 2020;276:115285.
- [6] Graabak I, Wu Q, Warland L, Liu Z. Optimal planning of the nordic transmission system with 100% electric vehicle penetration of passenger cars by 2050. *Energy* 2016;107:648–60.
- [7] Schill W-P, Gerbaulet C. Power system impacts of electric vehicles in Germany: Charging with coal or renewables? *Appl Energy* 2015;156:185–96.
- [8] Kintner-Meyer M, Davis S, Sridhar S, Bhatnagar D, Mahserejian S, Ghosal M. Electric vehicles at scale – phase I analysis: High EV adoption impacts on the western U.S. power grid. Tech. rep., Richland, WA (United States: Pacific Northwest National Laboratory (PNNL); 2020.
- [9] Neubauer J, Wood E. The impact of range anxiety and home, workplace, and public charging infrastructure on simulated battery electric vehicle lifetime utility. *J Power Sources* 2014;257:12–20.
- [10] Lee JH, Chakraborty D, Hardman SJ, Tal G. Exploring electric vehicle charging patterns: Mixed usage of charging infrastructure. *Transp Res D* 2020;79:102249.
- [11] Andrenacci N, Ragona R, Valenti G. A demand-side approach to the optimal deployment of electric vehicle charging stations in metropolitan areas. *Appl Energy* 2016;182:39–46.
- [12] Shepero M, Munkhammar J. Spatial Markov chain model for electric vehicle charging in cities using geographical information system (GIS) data. *Appl Energy* 2018;231:1089–99.
- [13] Muratori M, Jadun P, Bush B, Bielen D, Vimmerstedt L, Gonder J, et al. Future integrated mobility-energy systems: A modeling perspective. *Renew Sustain Energy Rev* 2020;119:109541.
- [14] Crozier C, Morstyn T, McCulloch M. A stochastic model for uncontrolled charging of electric vehicles using cluster analysis. 2019, arXiv preprint arXiv:1907.09458.
- [15] Sadeghianpourhamami N, Refa N, Strobbe M, Develder C. Quantitive analysis of electric vehicle flexibility: A data-driven approach. *Int J Electr Power Energy Syst* 2018;95:451–62.
- [16] Sodenkamp M, Wenig J, Thiesse F, Staake T. Who can drive electric? Segmentation of car drivers based on longitudinal GPS travel data. *Energy Policy* 2019;130:111–29.
- [17] Tang D, Wang P. Probabilistic modeling of nodal charging demand based on spatial-temporal dynamics of moving electric vehicles. *IEEE Trans Smart Grid* 2015;7(2):627–36.
- [18] Wang D, Gao J, Li P, Wang B, Zhang C, Saxena S. Modeling of plug-in electric vehicle travel patterns and charging load based on trip chain generation. *J Power Sources* 2017;359:468–79.
- [19] Wang Y, Infield D. Markov chain Monte Carlo simulation of electric vehicle use for network integration studies. *Int J Electr Power Energy Syst* 2018;99:85–94.
- [20] Li M, Lenzen M, Keck F, McBain B, Rey-Lescure O, Li B, et al. GIS-based probabilistic modeling of BEV charging load for Australia. *IEEE Trans Smart Grid* 2018;10(4):3525–34.
- [21] Wood EW, Rames CL, Muratori M. New EVSE analytical tools/models: Electric vehicle infrastructure projection tool (EVI-pro). Tech. rep., National Renewable Energy Lab.(NREL), Golden, CO (United States); 2018.
- [22] Crisostomo N, Krell W, Lu J, Ramesh R, Pratt K, Smith C, et al. Assembly bill 2127 electric vehicle charging infrastructure assessment (Staff report): Analyzing charging needs to support zero-emission vehicles in 2030. Tech. rep., California Energy Commission; Jan. 2021.
- [23] Kim S, Yang D, Rasouli S, Timmermans H. Heterogeneous hazard model of PEV users charging intervals: Analysis of four year charging transactions data. *Transp Res C* 2017;82:248–60.
- [24] Nicholas M, Tal G, Ji W. Lessons from in-use fast charging data: Why are drivers staying close to home. *Inst Transp Stud* 2017.
- [25] Flammini MG, Prettico G, Julea A, Fulli G, Mazza A, Chicco G. Statistical characterisation of the real transaction data gathered from electric vehicle charging stations. *Electr Power Syst Res* 2019;166:136–50.
- [26] Weldon P, Morrissey P, Brady J, O'Mahony M. An investigation into usage patterns of electric vehicles in Ireland. *Transp Res D* 2016;43:207–25.
- [27] Zhang J, Yan J, Liu Y, Zhang H, Lv G. Daily electric vehicle charging load profiles considering demographics of vehicle users. *Appl Energy* 2020;274:115063.
- [28] Xu M, Meng Q, Liu K, Yamamoto T. Joint charging mode and location choice model for battery electric vehicle users. *Transp Res B* 2017;103:68–86.
- [29] Pan L, Yao E, MacKenzie D. Modeling EV charging choice considering risk attitudes and attribute non-attendance. *Transp Res C* 2019;102:60–72.

- [30] Brady J, O'Mahony M. Modelling charging profiles of electric vehicles based on real-world electric vehicle charging data. *Sustainable Cities Soc* 2016;26:203–16.
- [31] Quirós-Tortós J, Navarro-Espinosa A, Ochoa LF, Butler T. Statistical representation of EV charging: Real data analysis and applications. In: 2018 power systems computation conference. (PSCC), IEEE; 2018, p. 1–7.
- [32] Helmus JR, Lees MH, van den Hoed R. A data driven typology of electric vehicle user types and charging sessions. *Transp Res C* 2020;115:102637.
- [33] Powell S, Cezar GV, Apostolaki-Iosifidou E, Rajagopal R. Large-scale scenarios of electric vehicle charging with a data-driven model of control. 2021, arXiv preprint arXiv:2105.12234.
- [34] Powell S, Cezar GV, Rajagopal R. Siobhanpowell/speech: First release. Zenodo; 2021, <http://dx.doi.org/10.5281/zenodo.5593509>, URL <https://github.com/SiobhanPowell/speech>.
- [35] Jordan MI, Sejnowski TJ, Poggio TA, et al. *Graphical models: Foundations of neural computation*. MIT Press; 2001.
- [36] Howard RA, Matheson JE. Influence diagrams. *Decis Anal* 2005;2(3):127–43.
- [37] Friedman J, Hastie T, Tibshirani R. *The elements of statistical learning*. In: Springer series in statistics vol. 1. New York. 2001.
- [38] Xu R, Wunsch D. *Clustering*, vol. 10. John Wiley & Sons; 2008.
- [39] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. *Scikit-learn: Machine learning in Python*. *J Mach Learn Res* 2011;12:2825–30.
- [40] James G, Witten D, Hastie T, Tibshirani R. *An introduction to statistical learning*, vol. 112. Springer; 2013.
- [41] McNicholas P. *Mixture model-based classification*. 1st ed.. Boca Raton, Florida: CRC Press, Taylor & Francis Group; 2017.
- [42] Schreiber J. *Pomegranate: fast and flexible probabilistic modeling in Python*. *J Mach Learn Res* 2017;18(1):5992–7.
- [43] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. *SciPy 1.0: fundamental algorithms for scientific computing in Python*. *Nature Methods* 2020;17(3):261–72.