ACN-Data: Analysis and Applications of an Open EV Charging Dataset

Zachary J. Lee
EE, Caltech
zlee@caltech.edu

Tongxin Li
CMS, Caltech
tongxin@caltech.edu

Steven H. Low
CMS, EE, Caltech
slow@caltech.edu

ABSTRACT

We are releasing **ACN-Data**, a dynamic dataset of workplace EV charging which currently includes over 30,000 sessions with more added daily. In this paper we describe the dataset, as well as some interesting user behavior it exhibits. To demonstrate the usefulness of the dataset, we present three examples, learning and predicting user behavior using Gaussian mixture models, optimally sizing on-site solar generation for adaptive electric vehicle charging, and using workplace charging to smooth the net demand Duck Curve.

CCS CONCEPTS

• **Mathematics of computing** → **Exploratory data analysis**; • **Hardware** → **Smart grid**; *Power networks*; Impact on the environment; • **General and reference** → *Empirical studies*;

KEYWORDS

Electric vehicle charging, open dataset, user behavior prediction, workplace charging, on-site solar generation, duck curve

ACM Reference format:
Zachary J. Lee, Tongxin Li, and Steven H. Low. 2019. ACN-Data: Analysis and Applications of an Open EV Charging Dataset. In *Proceedings of Proceedings of the Tenth ACM International Conference on Future Energy Systems, Phoenix, AZ, USA, June 25–28, 2019 (e-Energy '19),* 12 pages.
https://doi.org/10.1145/3307772.3328313

1 INTRODUCTION

Electric vehicles (EVs) have the potential to drastically reduce the carbon-footprint of the transportation sector. However, the growth of EVs in recent years has raised the question of how to best charge these massive loads. This has motivated studies on the impact of EV charging on the grid [8–10, 18, 28, 29]. Other researchers have focused on the potential of EVs as a controllable load, proposing algorithms to reduce demand variability [4, 20, 30], minimize costs when subject to time-varying prices [11, 20, 30], take advantage of intermittent renewable resources [1, 12, 22, 31, 33], or meet charging demands using limited infrastructure capacity [21, 23, 26]. While some of these studies, for example [10, 18, 20–23, 26, 32], have had access to real EV data to analyze their proposed algorithms, many others have had to rely on distributions derived from data collected from internal combustion engine (ICE) vehicles [9, 11, 12, 29–31] or assumed behaviors [1, 4, 8, 28, 33]. In addition, since all of these studies utilize different data sources, it can be difficult to compare one algorithm or approach against another.

There is thus a need for real EV charging data in the community in order to evaluate algorithms and study driver behavior. To meet this need we are releasing **ACN-Data**, a publicly-available dataset for EV charging research. This dataset currently consists of over 30,000 charging sessions collected from two workplace charging sites in California managed by PowerFlex, a smart EV charging startup. We hope that this dataset will be useful to the research community for evaluating new algorithms, understanding users' behavior, and informing the design of the next generation of smart EV charging systems.

A unique aspect of this dataset is that it is continually updated with sessions being uploaded daily.[1] This allows researchers to quickly gather data and understand trends in EV charging as they happen. While our current dataset includes only two sites, the startup with which we are partnering to collect data, PowerFlex, currently has over 60 sites around the country, and we plan to expand our database to new sites in the future. The dataset is available at `https://ev.caltech.edu/dataset`.

ACN-Data is not the first EV charging dataset studied by the academic community. A dataset collected by a Dutch smart EV charging provider ElaadNL has been used by [13] to examine capacity for demand response, [15] to build statistical models, and [25] to develop indicators which can be used to plan, and evaluate EV charging infrastructure. Similarly, [5] uses data collected from a Chinese charging system to investigate correlations in charging session parameters and [7, 32] use data collected from a charging network in Los Angeles to predict user behavior and evaluate proposed scheduling algorithms. Another study, My Electric Avenue [10], collected data from residential EV charging to examine its effect on the distribution grid. However, it can be difficult for researchers to access these datasets. The Pecan Street Dataport [19] is an exception, as it is publicly available for academic research. However, this dataset only includes residential EV charging data, so it is complementary to **ACN-Data**'s focus on workplace charging.

Our major contributions are as follows. 1) We describe the **ACN-Data** dataset and provide context which is important for its use. 2) We use this dataset to understand user behavior and flexibility in large-scale workplace charging. 3) We demonstrate that Gaussian mixture models can be used to learn the underlying distribution of charging session parameters at both the population and the individual level. 4) We show that prediction of session duration and energy demand of each user upon their arrival using the learned

[1] To protect user privacy we institute a two-week delay before data is released.

e-Energy '19, June 25–28, 2019, Phoenix, AZ, USA

ACM ISBN 978-1-4503-6671-7/19/06…$15.00
https://doi.org/10.1145/3307772.3328313

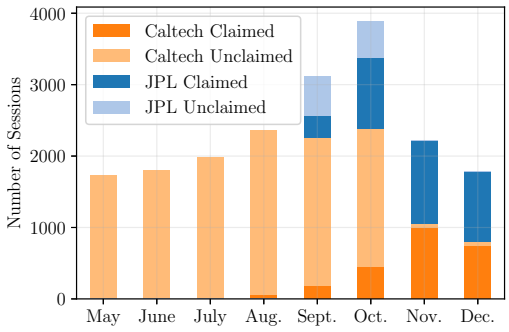

Figure 1: Number of charging sessions collected per month at each site for claimed and unclaimed sessions.

distribution is much more reliable than user input. 5) We use the dataset to formulate a novel data-driven approach to optimal sizing of on-site solar systems for adaptive EV charging. 6) We illustrate, using models derived from this dataset, the potential of adaptive EV charging to smooth the Duck Curve of net demand.

2 THE ACN-DATA DATASET

In this section we describe the dataset and how it is collected. More details on the charging facility and adaptive algorithm can be found in [21, 23].

2.1 Adaptive Charging Network (ACN)

ACN-Data was collected from two Adaptive Charging Networks located in California. The ACN on the Caltech campus is in a parking garage and has 54 EVSEs (Electric Vehicle Supply Equipment or charging stations) along with a 50 kW dc fast charger. The Caltech ACN is open to the public and is often used by non-Caltech drivers. Since the parking garage is near the campus gym, many drivers charge their EVs while working out in the morning or evening. JPL's ACN includes 52 EVSEs in a parking garage. In contrast with Caltech, access to the JPL campus is restricted and only employees are able to use the charging system. The JPL site is representative of workplace charging while Caltech is a hybrid between workplace and public use charging. EV penetration is also quite high at JPL. This leads to high utilization of the EVSEs as well as an ad-hoc program where drivers move their EVs after they have finished charging to free up plugs for other drivers. In both cases, to reduce capital costs, infrastructure elements such as transformers have been oversubscribed. The current architecture of the ACN for Caltech is described in [23] though both systems have a similar structure.

An adaptive scheduling algorithm is used to deliver each driver's requested energy prior to her stated departure time without exceeding the infrastructure capacity. We now describe an offline version of the algorithm that assumes full knowledge of all EV arrival times, departure times, and energy demands in advance. The formulation is used in later sections of this paper. The algorithm that is implemented in the actual ACNs is an online version in the form of model predictive control as shown in [23].

Let $\mathcal{V}$ be the set of all EVs over an optimization horizon $\mathcal{T} := \{1, \dots, T\}$. Each EV $i \in \mathcal{V}$ is described by a tuple $(a_i, e_i, d_i, \bar{r}_i)$ where a_i is the EV's *arrival time* relative to the start of the optimization horizon, e_i is its *energy demand*, d_i is the *duration* of the session, and $\bar{r}_i$ is the *maximum charging rate* for EV i. The charging rates for each EV in each period solve the following problem:

$$\textbf{SCH}(\mathcal{V}, U, \mathcal{R}): \quad \min_{\hat{r}} U(\hat{r}) \tag{1a}$$

$$\text{s.t. } \hat{r} \in \mathcal{R} \tag{1b}$$

where the optimization variable $\hat{r} := (\hat{r}_i(1), \dots, \hat{r}_i(T), i \in \mathcal{V})$ defines the scheduled charging rates of each EV over the optimization horizon $\mathcal{T}$. The utility function $U(r)$ encodes the operator's objectives and the feasible set $\mathcal{R}$ the various constraints.

To illustrate, we use the objective

$$U(r) := \sum_{\substack{t \in \mathcal{T} \\ i \in \mathcal{V}}} (t - T)\, r_i(t)$$

to encourage EVs to finish charging as quickly as possible, freeing up capacity for future arrivals. This objective, along with the regularization terms described in [23], is currently used in the ACNs at Caltech and JPL. Our feasible set $\mathcal{R}$ takes the form

$$0 \le r_i(t) \le \bar{r}_i \qquad a_i \le t < a_i + d_i, i \in \mathcal{V} \tag{2a}$$

$$r_i(t) = 0 \qquad t < a_i, t \ge a_i + d_i, i \in \mathcal{V} \tag{2b}$$

$$\sum_{t=a_i}^{d_i - 1} r_i(t) \le e_i \qquad i \in \mathcal{V} \tag{2c}$$

$$f_j(r_1(t), \dots, r_N(t)) \le I_j(t) \qquad t \in \mathcal{T}, j \in \mathcal{I} \tag{2d}$$

Constraints (2a) ensure that charging rates are nonnegative and below their maximum $\bar{r}_i(t)$; (2b) ensure that an EV does not charge before its arrival or after its departure time; (2c) limits the total energy delivered to EV i to at most e_i; and (2d) enforce a set of given infrastructure limits $I_j(t)$ indexed by $j \in \mathcal{I}$.

Since the utility function is strictly decreasing in every element of r, if it is feasible to meet all EV's energy demands, then constraint (2c) will be tight. In general, it is possible that the energy delivered may not reach the user's requested energy due to their battery becoming full or congestion in the system.

2.2 Data Collected

The ACN framework allows us to collect detailed data about each charging session which occurs in the system. Table 1 describes some of the relevant data fields we collect. To obtain data directly from users, we use a mobile application. The driver first scans a QR code on the EVSE which allows us to associate the driver with a particular charging session. The driver is then able to input their estimated departure time and requested energy. We refer to this as user input data. When a user does not use the mobile application, default values for energy requested and duration are assumed and no user identifier is attached to the session. We refer to sessions with an associated user input as claimed and those without as unclaimed.

In this paper we will focus on the 3-tuple (a_i, d_i, e_i) in the collected data for both user input and the actual measured behavior. Figure 1 displays the number of sessions collected from each site per month as well as whether these sessions were tagged with a

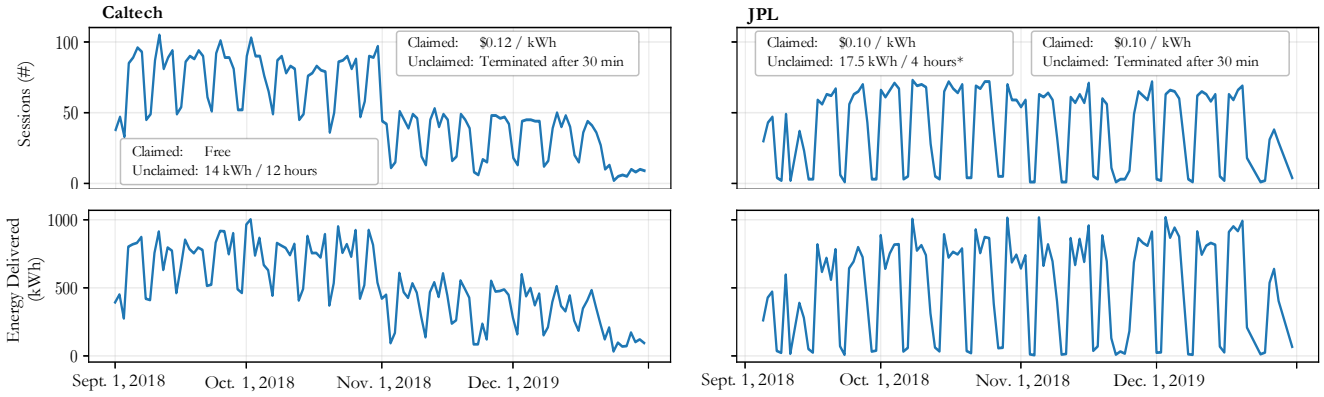

Figure 2: System utilization for the Caltech (left) and JPL (right) for the period from Sept. 1, 2018 to Jan. 1, 2018. Policy changes for claimed and unclaimed sessions are also shown. For claimed sessions, users specify their energy demand and session duration. For unclaimed sessions, default parameters are used, which have changed over time. Prior to Sept. 1, unclaimed sessions at Caltech received 42, 21, or 14 kWh over 12 hours depending on the specific EVSE. Unclaimed sessions have always been free. **While JPL has always required payment, some EVSEs were not able to be claimed prior to Nov. 1, so generous default parameters were instituted.*

Table 1: Selected data fields in ACN-Data.

Field	Description
connectionTime	Time when the user plugs in.
doneChargingTime	Time of the last non-zero charging rate.
disconnectTime	Time when the user unplugs.
kWhDelivered	Measured Energy Delivered
siteID	Identifier of the site where the session took place.
stationID	Unique identifier of the EVSE.
sessionID	Unique identifier for the session.
timezone	Timezone for the site.
pilotSignal	Time series of pilot signals during the session.
chargingCurrent	Time series of actual charging current of the EV.
userID*	Unique identifier of the user.
requestedDeparture*	Estimated time of departure.
kWhRequested*	Estimated energy demand.

*Field not available for every session.

user's input, i.e. claimed. Claimed sessions are useful for studying individual user behavior.

3 UNDERSTANDING USER BEHAVIOR

3.1 System Utilization

Figure 2 shows system utilization, specifically the number of sessions served and amount of energy delivered each day, from September through December 2018, together with pricing information and default parameters for unclaimed sessions.

3.1.1 Weekday vs. weekend charging. Figure 2 shows that both sites display a cyclic usage pattern with much higher utilization during weekdays than on weekends, as expected for workplace charging. Furthermore, Caltech, being a university and an open campus, has non-trivial usage on weekends. In contrast, JPL, as a closed campus, has next to no charging on weekends and holidays.

3.1.2 Free vs. paid charging. The data confirms the difference between paid and free charging facilities. During the first 2.5 years of operation the Caltech ACN was free for drivers. However, beginning Nov. 1, 2018, a fee of $0.12 / kWh was imposed. We can see this date clearly from Figure 2, as both the number of session per day and daily energy delivered decreased significantly. Because of an issue with site configuration, approximately half of the EVSEs at JPL were free prior to Nov. 1, 2018. However at JPL we do not see a large decrease in utilization in terms of number of sessions or energy delivered after Nov. 1. This is likely because demand for charging is high enough to overshadow any price sensitivity.

3.2 Arrivals and Departures

Figure 3 shows the distributions of arrivals to and departures for both sites. For Caltech we plot the distributions for weekends and weekdays, as well as for free and paid charging separately.

3.2.1 Effects of free vs. paid charging. From the figure the shape of the distributions are similar before and after paid charging was implemented. We note two key differences, however, in weekday charging between free and paid periods. First, the second peak around 6 pm vanishes. We attribute this to a decrease in community usage of the Caltech ACN after its cost became comparable to at-home charging. Second, the peak in arrivals (departures) around 8 am (5 pm) increases. This is expected as instituting paid charging has reduced community usage in the evening which leads to a higher proportion of users displaying standard work schedules.

3.2.2 Weekday distribution. The figure shows that the weekday arrival distribution has a morning peak at both sites. For conventional charging systems, these peaks necessitate a larger infrastructure capacity and lead to higher demand charges. In addition, as EVs adoption grows, these morning spikes in demand could prove challenging for utilities as well. As expected, departures are analogous to arrivals. They begin to increase as the workday ends, with

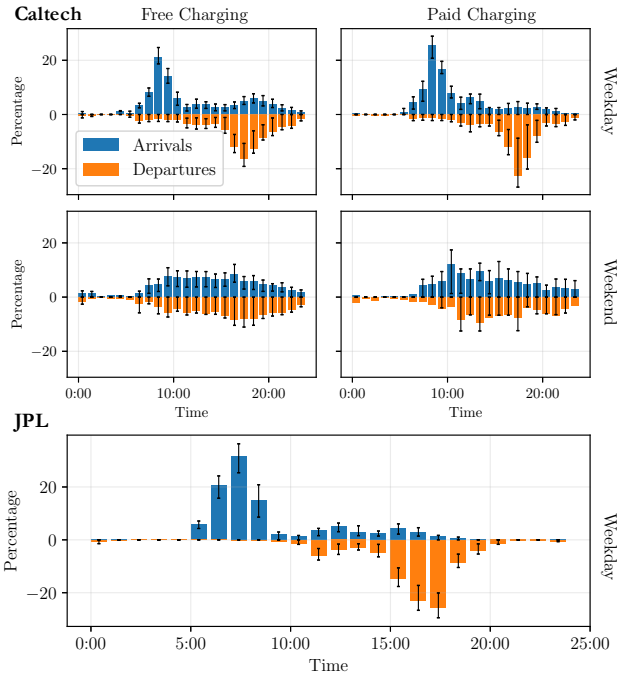

Figure 3: Distribution of arrivals and departures in each ACN. Bars denote the mean distribution over the period of May 1, 2018 to Jan. 1, 2019 and whiskers denote the first and third quartiles. For Caltech we differentiate between paid and free periods and weekdays versus weekends. For JPL since free charging was only offered at a subset of EVSEs and weekend usages is extremely low, we plot only a single distribution for weekdays.

peaks in the period 5-6 pm at both Caltech and JPL. Departures at JPL tend to begin earlier, which is consistent with the earlier arrival times while departures at Caltech tend to stretch into the night owing to the heterogeneity of individual schedules as well as later arrivals.

3.2.3 Weekend distribution. Since the Caltech ACN is open to the public and is located on a university campus, it receives use on the weekends. From Figure 3 arrivals and departures are much more uniform on weekends for both the unpaid and paid periods. This uniformity is probably due to the aggregation of many highly heterogeneous weekend schedules.

3.3 Driver and System Flexibility

3.3.1 Driver laxity. There are different notions of driver laxity, and we use the following definition that has been applied to EV charging [26]. The initial laxity of an EV charging session i is defined as

$$\mathrm{LAX}(i) = d_i - \frac{e_i}{\bar{r}_i}$$

$\mathrm{LAX}(i) = 0$ means that EV i must be charged at its maximum rate $\bar{r}_i$ over the entire duration d_i of its session in order to meet its energy demand e_i. A higher value of $\mathrm{LAX}(i)$ means there is more flexibility

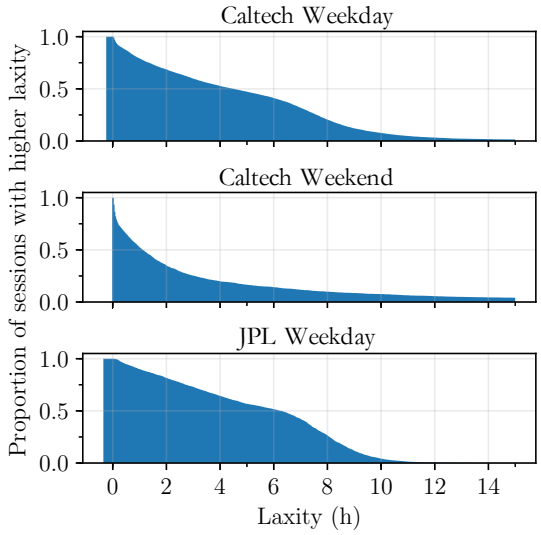

Figure 4: Empirical complementary cumulative distribution of laxity in each ACN.

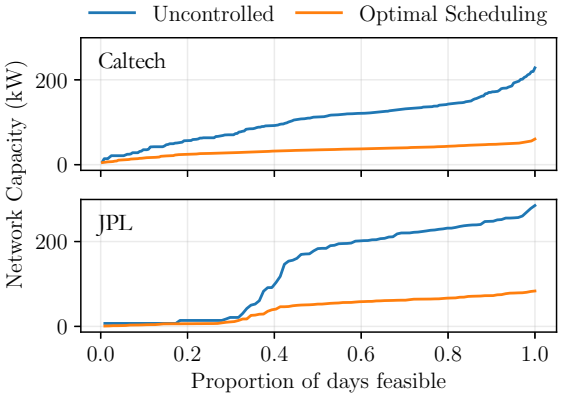

Figure 5: Capacity required to fulfill a given proportion of the days in our dataset for Caltech (top) and JPL (bottom).

in satisfying its energy demand. Figure 4 shows the distribution of initial laxities in our dataset. It confirms that, for weekdays, most EVs display high laxity. On weekends laxity tends to be lower as drivers tend to want to get charged and get on with their day.

3.3.2 Minimum system capacity. One way to quantify the aggregate flexibility of a group of EVs is the minimum system capacity needed to meet all their charging demands. A smaller system capacity requires a lower capital investment and operating cost for a charging operator. To calculate the minimum system capacity we solve $\mathbf{SCH}(\mathcal{V}, U_{\mathrm{cap}}, \hat{\mathcal{R}})$ for the optimal charging rates $\hat{r}^*$, for each day in our dataset where $\mathcal{V}$ is the set of all EVs using the charging system in a day,

$$U_{\mathrm{cap}}(r) := \max_t \sum_{i \in \mathcal{V}} r_i(t)$$

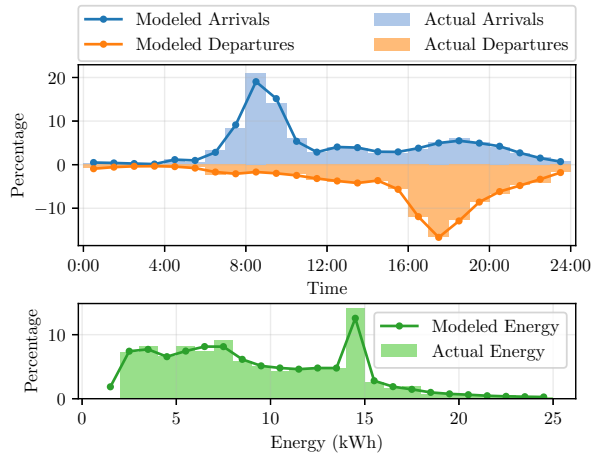

Figure 6: Comparison of model distributions with actual data for Caltech during training period.

and $\hat{\mathcal{R}}$ is equivalent to (2) except that (2c) is strengthened to equality.[2] For simplicity, we do not consider any infrastructure constraints (2d) in $\hat{\mathcal{R}}$. The distribution of the minimum system capacity $U_{\text{cap}}(\hat{r}^*)$ per day in our dataset is plotted in Figure 5. It shows that we would have been able to meet the demand for 100% of days in our dataset with just 60 kW of capacity for Caltech and 84 kW for JPL. Meanwhile conventional uncontrolled systems of the same capacity would only be able to meet demand on 22% and 38% of days respectively. For reference Caltech has an actual system capacity of 150 kW and JPL has 195 kW.

4 LEARNING USER BEHAVIOR

In this section, we illustrate how to learn the underlying joint distribution of arrival time, session duration, and energy delivered using Gaussian mixture models (GMMs) (e.g., [14, 24]). We then use these GMMs to predict user behavior (Section 5), optimally size onsite solar for adaptive EV charging (Section 6), and control EV charging to smooth the duck curve (Section 7).

4.1 Problem Formulation

We utilize the GMM as a second-order approximation to the underlying distribution. Our dataset can be modeled as follows to fit a GMM. Consider a dataset X consisting of N charging sessions. The data for each session $i = 1, \ldots, N$, is represented by a triple $x_i = (a_i, d_i, e_i)$ in $\mathbb{R}^3$ where a_i denotes the arrival time, d_i denotes the duration and e_i is the total energy (in kWh) delivered. The data point X_i (we use capital letters for random variables) are independently and identically distributed (i.i.d.) according to some unknown distribution. In practice, each driver in a workplace environment exhibits only a few regular patterns. For example, on weekdays, a driver may typically arrive at 8 am and leave around 6 pm, though her actual arrival and departure times may be randomly perturbed around their typical values. On weekends, driver behavior may change such that the same driver may come around noon. We hence assume that drivers have finitely many behavior profiles. Therefore, let K be the number of typical profiles denoted by $\mu_1, \ldots, \mu_K$.[3] Each data point X_i can be regarded a corrupted version of a typical profile with a certain probability. Define a latent variable $Y_i \equiv k$ if and only if X_i is corrupted from μ_k. Moreover, by the i.i.d. assumption, each incoming EV has an identical probability π_k taking μ_k, *i.e.*, $\pi_k := \mathbb{P}(Y_i = k)$ for $i = 1, \ldots, N,\ k = 1, \ldots, K$. Conditioned on $Y_i = k$, the difference $X_i - \mu_k$ that the profile X_i deviates from the typical profile μ_k can be regarded as Gaussian noise. In this manner, assuming $Y_i = k$, we let $X_i \sim \mathcal{N}(\mu_k, \Sigma_k)$ be a Gaussian random variable with mean μ_k and covariance matrix Σ_k. To estimate the underlying distribution and approximate it as a mixture of Gaussians, it suffices to estimate the parameters $\theta = (\pi_k, \mu_k, \Sigma_k)_{k=1}^K$. The probability density of observing a data point x can then be approximated using the learned GMM as

$$p(x|\theta) = \sum_{k=1}^{K} \pi_k \frac{\exp\left(-||x-\mu_k||^2_{\Sigma_k^{-1}}/2\right)}{\sqrt{(2\pi)^3 \det(\Sigma_k)}}$$

4.2 Population and Individual-level GMMs

We train GMMs based on a training dataset $\mathrm{X}_{\text{Train}}$ and predict the charging duration and energy delivered for drivers in a set $\mathcal{U}$. The results are tested on a corresponding testing dataset X_{Test}. As illustrated in Figure 1, the training data collected at both Caltech and JPL can be divided into two parts: user-claimed data X_{C} and unclaimed data X_{U}.

This motivates us to study two different approaches. The first approach generates a population-level GMM (P-GMM) based on the overall training data $\mathrm{X}_{\text{Train}} = \mathrm{X}_{\text{C}} \bigcup \mathrm{X}_{\text{U}}$. However, users can have distinctive charging behaviors. To achieve better prediction accuracy, we take advantage of the user-claimed data and predict the charging duration and energy delivered for each individual user. In the second approach, the claimed data can be partitioned into a collection of smaller datasets consisting of the charging information of each user in $\mathcal{U}$. We write $\mathrm{X}_{\text{C}} = \bigcup_{j \in \mathcal{U}} \mathrm{X}_j$. We can then train individual-level GMMs (I-GMM) for each user $j \in \mathcal{U}$ by fine tuning the weights of the components of the P-GMM with data from each of the users to arrive at a final model for each of them.

4.3 Distribution Learned by P-GMM

To evaluate how well our learned population-level GMM fits the underlying distribution, we gather 100,000 samples from a P-GMM trained on data from Caltech collected prior to Sep. 1, 2018. We then plot in Figure 6 the distribution of these samples along with the empirical distribution from our training set. We choose to plot departure time instead of duration directly as this demonstrates that our model has learned not only the distribution of session duration but also the correlation between arrival time and duration. We see that in all cases, our learned distribution matches the empirical

[2] This is necessary because U_{cap} is not strictly decreasing in r. We are only able to strengthen (2c) to equality when it is feasible to meet all energy demands which is the case here since we use actual delivered energy.

[3] We assume the number K of components is known. In our experiments in Section 4, grid search [27] and cross-validation is used to find the best number of components.

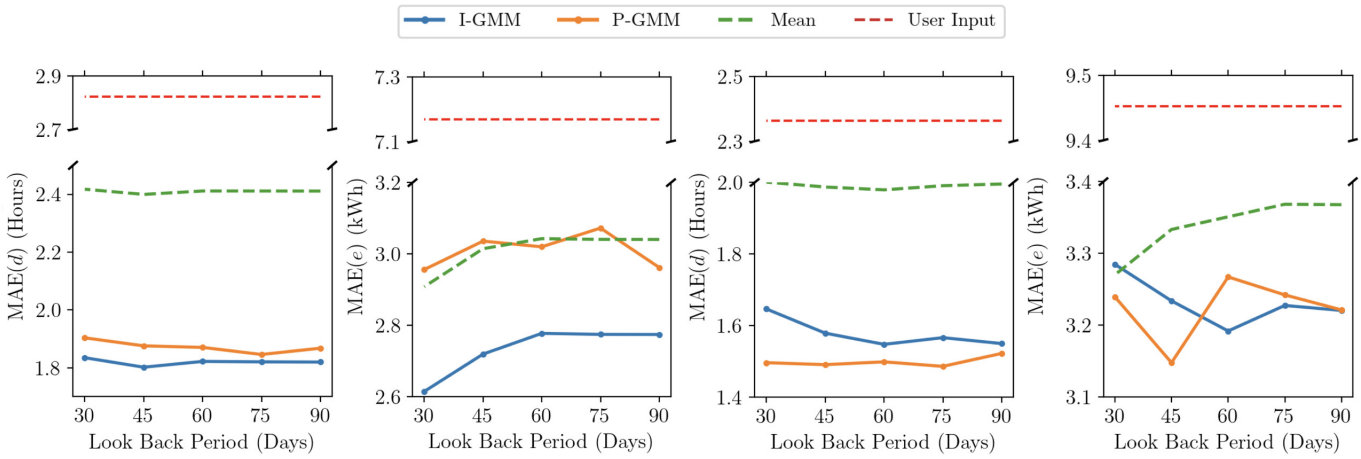

Figure 7: Prediction errors for Caltech (left two columns) and JPL (right two columns) for training dataset sizes ranging from 30 days to 90 days in the past. As a benchmark, we consider simply taking the mean of each user's prior behavior. For comparison, we also include the errors of user inputs. The results are measured by the mean absolute error (MAE) defined in (5).

distribution well. We next present three applications of the **ACN-Data** dataset and the learned distribution in Sections 5 to 7.

5 PREDICTING USER BEHAVIOR

In this section, we use the GMM that we have learned from the **ACN-Data** dataset to predict a user's departure time and the associated energy consumption based on their known arrival time. Despite recent advances in arrival time based prediction via kernel density estimation [5, 7, 32], simple empirical predictions are commonly used in practical EV charging systems. For example, the ACNs from which this data was collected use user inputs directly in the scheduling problem [23], while other charging systems simply take the average of the past behavior as a prediction. Our data, however, shows that user input can be quite unreliable, partially because of a lack of incentives for users to provide accurate predictions. We demonstrate that the predictions can be more precise using simple probabilistic models.

5.1 Calculating Arrival Time-based Predictions

Let $\mathcal{U}$ denote the set of users. Suppose a convergent solution $\theta^{(j)} = (\pi_k^{(j)}, \mu_k^{(j)}, \Sigma_k^{(j)})_{k=1}^K$ is obtained for user $j \in \mathcal{U}$ where $\mu_k^{(j)} := (a_k^{(j)}, d_k^{(j)}, e_k^{(j)})$ and the user's arrival time is known *a priori* as $\alpha^{(j)}$. For the sake of completeness, we present the following formulas used for predicting the duration $\delta^{(j)}$ and energy to be delivered $\varepsilon^{(j)}$ as conditional Gaussians of the user $j \in \mathcal{U}$:

$$\delta^{(j)} = \sum_{k=1}^{K} \overline{\pi}_k^{(j)} \left(d_k^{(j)} + (\alpha^{(j)} - a_k^{(j)}) \frac{\Sigma_k^{(j)}(1,2)}{\Sigma_k^{(j)}(1,1)} \right) \tag{3}$$

$$\varepsilon^{(j)} = \sum_{k=1}^{K} \overline{\pi}_k^{(j)} \left(e_k^{(j)} + (\alpha^{(j)} - a_k^{(j)}) \frac{\Sigma_k^{(j)}(1,3)}{\Sigma_k^{(j)}(1,1)} \right) \tag{4}$$

where $\Sigma_k^{(j)}(1,1)$, $\Sigma_k^{(j)}(1,2)$ and $\Sigma_k^{(j)}(1,3)$ are the first, second and third entries in the first column (or row) of the covariance matrix $\Sigma_k^{(j)}$ respectively. Denoting by $p(\cdot|\mu, \sigma^2)$ the probability density for a normal distribution with mean μ and variance σ^2, the modified weights conditioned on arrival time in (3) and (4) above are

$$\overline{\pi}_k := \frac{p\left(\alpha^{(j)} | a_k^{(j)}, \Sigma_k^{(j)}(1,1)\right)}{\sum_{k=1}^{K} p\left(\alpha^{(j)} | a_k^{(j)}, \Sigma_k^{(j)}(1,1)\right)}$$

5.2 Error Metrics

We consider both absolute error and percentage error when evaluating duration and energy predictions.

5.2.1 Mean absolute error. Recall that $\mathcal{U}$ is the set of all users in a testing dataset $\mathrm{X}_{\mathrm{Test}}$. Let $\mathcal{A}_j$ denote the set of charging sessions for user $j \in \mathcal{U}$. The Mean Absolute Error (MAE) is defined in (5) to assess the overall deviation of the duration and energy consumption. For a testing dataset $\mathrm{X}_{\mathrm{Test}} = \{(a_{i,j}, d_{i,j}, e_{i,j})\}_{j \in \mathcal{U}, i \in \mathcal{A}_j}$, the corresponding MAEs for duration and energy are represented by MAE(d) and MAE(e) with

$$\mathrm{MAE}(x) := \sum_{j \in \mathcal{U}} \frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{A}_j} \frac{1}{|\mathcal{A}|_j} \left| x_{i,j} - \widehat{x}_{i,j} \right| \tag{5}$$

where $\widehat{x}_{i,j}$ is the estimate of $x_{i,j}$ and $x = d$ or e.

5.2.2 Symmetric mean absolute percentage error. The Symmetric Mean Absolute Percentage Error (SMAPE) in (6) is commonly used (for example, see [7]) to avoid skewing the overall error by the data points wherein the duration and energy consumption take small values. The corresponding SMAPEs for duration and energy are represented by SMAPE(d) and SMAPE(e) with

$$\mathrm{SMAPE}(x) := \sum_{j \in \mathcal{U}} \frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{A}_j} \frac{1}{|\mathcal{A}|_j} \left| \frac{x_{i,j} - \widehat{x}_{i,j}}{x_{i,j} + \widehat{x}_{i,j}} \right| \times 100\% \tag{6}$$

Table 2: SMAPE**s for Caltech and JPL datasets.**

Caltech	I-GMM	P-GMM	Mean	User Input
SMAPE(d)%	15.8543	16.6313	20.4432	25.8093
SMAPE(e)%	14.4273	17.2927	15.9275	27.5523

JPL	I-GMM	P-GMM	Mean	User Input
SMAPE(d)%	12.2500	12.5079	15.8985	18.5994
SMAPE(e)%	12.7318	13.6863	13.3014	26.8769

5.3 Results and Discussion

5.3.1 Experimental setup. In Figure 7, we report MAE(d) and MAE(e) for I-GMM and P-GMM on Caltech dataset as a function of the look back period which defines the length of the training set. Users with larger than 20 sessions during Nov. 1, 2018 and Jan. 1, 2019 are included in $\mathcal{U}$ and tested. Note that the size of the training data may not be proportional to the length of periods since in general there is less claimed session data early in the dataset as shown in Figure 3. The 30-day testing data is collected from Dec. 1, 2018 to Jan. 1, 2019. We study the behavior of prediction accuracy with different training data sizes by training the GMMs with data collected from five time intervals ending on Nov. 30, 2018 and starting on Sep. 1, 2018, Sep. 15, 2018, Oct. 1, 2018, Oct. 15, 2018 and Nov. 1, 2018 respectively. The GMM components are initialized using k-means clustering as implemented by the Scikit Learn GMM package [27]. Since it is not deterministic, we repeat this initialization 25 times and keep the model with the highest log-likelihood on the training dataset. Grid search and cross validation [27] are used to find the best number of components for each GMM.

5.3.2 Observations. As observed from Figure 7, for the JPL dataset with testing data obtained from Dec. 1, 2018 to Jan. 1, 2019, the 60-day training data gives the best overall performance. This coincides with our intuition that user behavior changes over time and there is a trade-off between data quality and size. The Caltech dataset also displays this trade-off; however, the best performance was found for only a 30-day training set. This is likely because there was a transition from free to paid charging on Nov. 1, which meant that data prior to that date had very different properties.

Hence, for the JPL dataset, we fix the training data as the one collected from Oct. 1, 2018 to Dec. 1, 2018 and show the scatterings of SMAPEs for each session in the testing data (from Dec. 1, 2018 to Jan. 1, 2019) in Figure 8. The SMAPEs are concentrated on small values with a few outliers and high-quality duration prediction has a positive correlation with high-quality energy prediction. As a comparison, user input SMAPEs, shown as Xs, are much worse.

Table 2 shows the average SMAPEs for the various methods tested. For Caltech and JPL, we display the results using the 30 and 60-day training data respectively. For reference we also calculate the error of two additional ways to predict user parameters: 1) we use the mean of the training data X_j as our prediction for each user, 2) we treat the user input data directly as the prediction. Note that to account for stochasticity in the GMM training process, the results in Figure 7 and Table 2 are obtained via 50 Monte Carlo simulations.

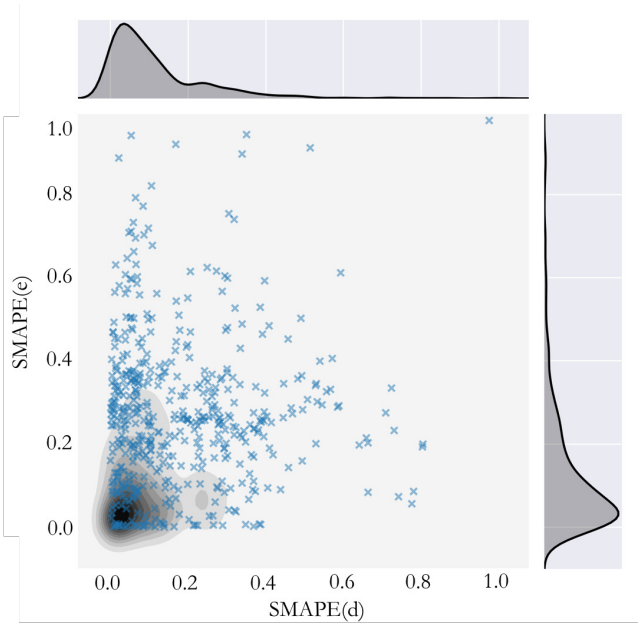

Figure 8: Correlation between SMAPE(d) **and** SMAPE(e) **and their marginal distributions for the JPL dataset. Kernel density estimation is used to approximate the joint distribution of the SMAPEs for predicted duration and energy which is shown as grey shading. The blue crosses represent the corresponding user input SMAPEs (for I-GMM) with respect to each charging session in the testing data set** X_{Test}**.**

5.3.3 Implications. EV users need incentives to provide more accurate predictions. As shown in Figures 7, Figure 8 and Table 2, user input data conspicuously gives the worst overall prediction. However, in some commercial EV charging companies, *e.g.,* PowerFlex, user input data is used as the direct input for the scheduling and pricing algorithms. Therefore, significant improvements can be made in the future by leveraging tools from statistics and machine learning to better predict user behaviors, *e.g.,* using GMMs. In addition, we find that when predicting user behavior there is a trade-off between between training data quantity and quality caused by changing user behavior over time which must be considered.

6 WORKPLACE SOLAR CHARGING

While many studies have addressed the problem of scheduling EV charging to utilize on-site solar generation, it is generally assumed that solar capacity is fixed [1, 22, 31, 33]. In this section, we use the **ACN-Data** dataset, as well as our learned distributions, to address the separate problem of how historical data can be used to optimally size on-site solar generation in order to minimize the cost of workplace EV charging.

6.1 Problem Formulation

Assuming that at operation time EVs will be charged optimally, we can formulate the problem of finding the optimal solar capacity, α, of an on-site solar installation as minimizing the following cost

function, given a set $\mathcal{V}$ of EVs:

$$c^*(\alpha, \mathcal{V}) := \mathbf{SCH}(\mathcal{V}, U_{\text{sol}}, \hat{\mathcal{R}}) \tag{7}$$

where the total charging cost, U_{sol} is defined as:

$$\begin{aligned} U_{\text{sol}} := c_s \sum_{t \in \mathcal{T}} \alpha s(t) + \sum_{t \in \mathcal{T}} c_e(t) \left[\sum_{i \in \mathcal{V}} r_i(t) - \alpha s(t) \right]^+ \\ + \Delta \cdot \max_{t \in \mathcal{T}} \sum_{i \in \mathcal{V}} \left[\sum_{i \in \mathcal{V}} r_i(t) - \alpha s(t) \right]^+ \end{aligned} \tag{8}$$

and $\hat{\mathcal{R}}$ is defined in Section 3.3.2. In (8), $s(t)$ is a solar generation profile found by normalizing the time series of solar production over a period by the capacity of the system which generated it. c_s is the levelized cost of energy (LCOE) for the solar array which accounts for capital and operating costs. LCOE is calculated assuming that all solar generated is utilized, and therefore we are charged for all solar production regardless of if it is used to charge EVs or not.[4] $c_e(t)$ is the possibly time-varying electricity cost from the grid and Δ is the demand charge levied by the utility based on peak usage throughout a billing period. Here $[a]^+ := \max\{a, 0\}$.

Given that EV charging will be scheduled optimally according to (7), the optimal solar capacity to install is then:

$$\alpha^* := \arg\min_{\alpha} \ \mathbb{E}_{\mathcal{V}}[c^*(\alpha, \mathcal{V})]$$

Hence the optimal solar capacity α^* minimizes the expected total cost where the expectation is taken over the random set $\mathcal{V}$ of EV arrivals. To compute α^* we estimate the expected cost using the empirical mean of scenarios $\{\mathcal{V}_1, ..., \mathcal{V}_S\}$ sampled from the learned distribution:

$$\mathbb{E}_{\mathcal{V}}[c^*(\alpha, \mathcal{V})] \approx \frac{1}{S} \sum_{j=1}^{S} c^*(\alpha, \mathcal{V}_j) \tag{9}$$

6.2 Case Studies

6.2.1 Computing optimal solar capacity α^*. We first compute the optimal solar capacity using our purposed method for a single month, then evaluate its performance based on actual data.

Consider the Caltech ACN during the month of September. We adopt the Southern California Edison (SCE) time-of-use (TOU) rate schedule [6] for separately metered EV charging systems between 20-500 kW which is shown in Table 3 (SCE considers September a summer month). We use a LCOE for solar of \$0.08 / kWh which corresponds to NREL's SunShot 2020 goal for commercial PV systems and which was met in 2017 [17]. To produce a realistic solar profile $s(t)$, we use NREL's SAM tool and the National Solar Radiation Database (NSRD) to estimate solar output for a typical meteorological year at Caltech [16]. For this study we set the length of each discrete time interval in the optimization to be 15 min.

We generate 100 EV charging scenarios using the learned GMMs, each 1 month long. The GMMs are trained, one for weekdays and one for weekends, using data up to September 1 as described in Section 5. Since the GMMs do not model the number of arrivals each day, we fit a separate Gaussian to predict the number of arrivals using data from the previous month. Once again we have one model for weekdays and one for weekends. We then generate scenarios by first taking a draw from the appropriate Gaussian to estimate the number of arrivals on the given day, and then gathering that many samples from the corresponding GMM. We repeat this procedure until we have accumulated 100 months worth of generated data. The optimal solar capacity α^* that minimizes the average total cost across these 100 EV scenarios is then computed from (7)–(9). The result is denoted "Planning" in Table 4. When the optimal solar capacity $\alpha^* = 76$ kW is installed, *on average* we expect 57.5% of EV demand to be met by on-site solar generation and the resulting total cost to be \$2,156 a month.

We can then evaluate how this α^* chosen at planning time performs using real charging sessions collected from the Caltech ACN in Sept. 2018 and solar data from Sept. 2017 provided through the NSRD, i.e., we evaluate (7) –(8) with $\alpha^* = 76$ kW. The result is denoted "Evaluation" in Table 4. Compared with no on-site solar ($\alpha = 0$ in (7) –(8)), solar generation could have saved Caltech \$1,092 for that month.

To appreciate how well this performs, suppose we optimize the solar capacity α with respect to the real September EV data from Caltech and NSRD solar data from Sept. 2017. To do this, we find the optimal α by minimizing (7) over α with $\mathcal{V}$ being the real data. We then evaluate the performance of this optimally sized array using (7) –(8). The result, denoted by "Optimal" in Table 4, represents a lower bound on the total cost of charging with on-site solar. Compared with the performance of $\alpha^* = 76$ kW on the real September data, the solar capacity and solar coverage are slightly higher, however the total cost and savings differ by only \$3 for that month. This suggests that the capacity determined at planning time performs very well on the real EV data.[5]

6.2.2 Optimal capacity over a year. The previous example only covers a single month in order to illustrate our methodology. We now estimate the optimal solar capacity over an entire year. To do so we first generate a set of scenarios $\{\mathcal{V}_1, ..., \mathcal{V}_{240}\}$ where each scenario is 1 month long using the procedure outlined in

Table 3: SCE TOU Rate Schedule for EV Charging

	Summer Rates	Winter Rates
On-Peak	\$0.25 / kWh	\$0.08 / kWh
Mid-Peak	\$0.09 / kWh	\$0.08 / kWh
Off-Peak	\$0.06 / kWh	\$0.07 / kWh
Demand Charge	\$15.48 / kW / month	

Table 4: Evaluating Planned Solar Capacity for Caltech

	Data	Solar Capacity	Percent Solar	Total Cost	Savings
Planning	Synthetic	76 kW	57.5%	\$2,156	-
Evaluation	Real Sept	76 kW	50.0%	\$2,447	**\$1,092**
Optimal	Real Sept	81 kW	52.3%	\$2,444	**\$1,095**

[4] In many cases excess solar generation could be used by other loads or sold back to the grid. However, in this example we consider the simple case where the PV system is connected only to the EVSEs and no net metering is offered.

[5] The small change in cost (0.12%) for a relative large change in capacity (6.2%) is indicative of the shallow slope in the cost function near the minimum which is helpful for robustness.

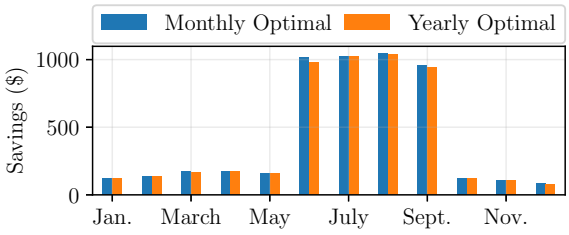

Figure 9: Savings achieved versus no on-site solar for each month of the year. We consider two scenarios, one where the solar capacity is allocated optimally for each month and a second where the solar capacity is fixed based on an optimization over the full year.

Section 6.2.1. This is equivalent to 20 years of simulated data. Since solar generation and time-of-use prices depend on the month, we adjust $s(t)$ and $c_e(t)$ for each scenario based on its corresponding month. We then solve (7)–(9) to find the optimal solar capacity over the year and estimate the expected savings. The result for Caltech is denoted by "Yearly Optimal" in Figure 9. It shows that most savings are concentrated during the summer months when both TOU rates and solar production are high. Over the year we expect a total savings of $5,043 of which $3,984 is from June through Sept.

For reference, we compute the optimal α^* for each month individually (as in Section 6.2.1) and estimate the corresonding savings. The result, denoted by "Monthly Optimal" in Figure 9, represents an upper bound on expected savings because it is not practical to change the solar installation month-to-month. It shows that the yearly optimal solar capacity α^* can achieve an expected savings that are close to their upper bounds.

6.2.3 Sensitivity to LCOE. We next illustrate the sensitivity of these benefits to the LCOE of solar, which will continue to fall in the coming years with a goal of $0.04 / kWh for commercial PV systems by 2030 [17]. We first generate a 20-year collection of scenarios for JPL as we did for Caltech. For each LCOE we solve (7)–(9) to find the optimal α^* over the year for both Caltech and JPL. The expected benefits are shown in Figure 10.

These results confirm that as solar prices decrease, there is an increase in the optimal solar capacity, percent of charging demand met by solar, and operators' savings for both sites. At very low solar costs, we can meet a very high percentage of total charging demand using solar alone, especially at JPL where users' schedules align well with solar production. This results in substantial cost reductions for site operators as well as significant reduction in the environmental impact of EV charging. Thus as the LCOE for solar decreases, we expect that on-site generation will play an important role in reducing the cost and environmental footprint of workplace EV charging.

The optimal capacities α^* at the two sites differ due to differences in their usage profiles. While α^* becomes nonzero at Caltech at a LCOE of $0.17 / kWh, it remains zero at JPL until the cost drops below $0.12 kWh. We suspect the reason for this is that JPL users tend to arrive earlier than Caltech users, allowing them to be scheduled to avoid on-peak rates during the summer which reduces the

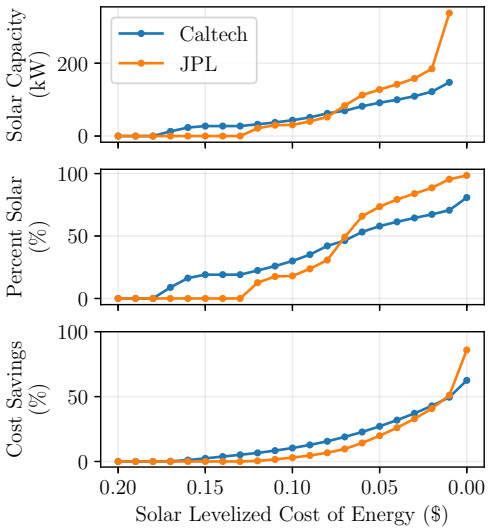

Figure 10: Effect of solar levelized cost of energy and site on optimal solar capacity, percent of demand met by solar energy, and cost savings over the no solar case.

marginal benefit of on-site solar when LCOE is high. In addition, since JPL has almost zero utilization on weekends and receives no compensation for solar generated during that time, it makes less sense to install solar there when LCOE is high.

7 SMOOTHING THE DUCK CURVE

While on-site renewable energy has many advantages, not all facilities will have the ability to install large PV arrays. For these sites, EVs must be charged using energy from the grid. The effect that large numbers of EVs will have on the grid is highly dependent on the usage patterns and flexibility of individual users. For this reason, having access to real data for these studies is important. In this section we use the **ACN-Data** dataset to illustrate the potential of controlling a large number of EVs to help alleviate the steep ramping conditions caused by the Duck Curve [2].

7.1 Problem Formulation

We formulate the problem of minimizing ramping as

$$\textbf{SCH}(\mathcal{V}, U_{\text{ramp}}, \hat{\mathcal{R}}) \tag{10}$$

where we denote the objective by

$$U_{\text{ramp}} := \sum_{t \in \mathcal{T} \setminus \{0\}} \left(\hat{d}(t) - \hat{d}(t-1) \right)^2 \tag{11}$$

and

$$\hat{d}(t) := \sum_{i \in \mathcal{V}} r_i(t) + d(t) \tag{12}$$

Here $d(t)$ is the net demand placed on the grid after non-dispatchable renewable energy is subtracted from the total demand. While we acknowledge that many other flexible loads such as water heaters, appliances, pool pumps, etc. are currently included in $d(t)$ and could be used to aid in smoothing the Duck Curve, we focus our attention

on the contribution of electric vehicles and thus treat these loads as fixed.

7.2 Case Studies

7.2.1 Qualitative results. In order to demonstrate the potential of workplace charging to smooth the Duck Curve, we consider a net demand curve for Dec. 11, 2018 from CAISO [3]. We consider three levels of EV penetration in California based on the current number of EVs in California (350,000) and the state's goals for 2025 (1.5 million) and 2030 (5 million). For this case study, we make the optimistic assumption that all of these vehicles would be available for workplace charging. While this assumption is unrealistic, it bounds the potential benefit from above at each level of penetration. Once again we set the length of each discrete time interval in the optimization to be 15 min.

To reduce the computational burden in solving (10) for millions of EVs, we use a representative sample of n EVs drawn from our learned distribution and scale down the net demand curve $\tilde{d}(t)$ from CAISO by the ratio of n to the desired number of EVs, denoted by N. Define $d(t) := (n/N)\tilde{d}(t)$. We then solve (10)-(12) for $d(t)$ and this representative sample. Finally, we scale the optimal net demand curve, $\hat{d}^*(t)$, by N/n to arrive at a final curve in the original units. For this experiment, $n = 1,000$.

Figure 11 plots the resulting optimal net demand curve $\hat{d}^*(t)$, for both the Caltech data and the JPL data. Even with only 350,000 EVs, we see a non-trivial smoothing of the net demand curve. With 1.5 million EVs under control, we see a significant filling of the "belly" of the duck as well as a reduction in the morning and afternoon ramping requirements. By the time we reach 5 million EVs under control, we can see an almost complete smoothing of the duck in the JPL case. We note however that for the Caltech distribution, 5 million EVs lead to a noticeable increase in peak demand. This is because we use the distribution for free charging which includes a significant number of short sessions that begin around 5-7 pm, thus requiring us to charge these EVs during the peak of background demand. This demonstrates benefits of concentrating EV charging during normal working hours, for which the JPL distribution is representative.

7.2.2 Quantitative results. To examine quantitatively how much we are able to smooth the Duck Curve, we vary the number of EVs under control from 10,000 to 10 million for each distribution. We optimally schedule each group of EVs using (10) and measure the resulting maximum up and down ramps as well as the peak demand. The results are shown in Figure 11. Surprisingly, we find that with as few as 2 million EVs under control, we can cut up and down ramping requirements by nearly 50% with only a 0.6% increase in peak demand when using the JPL distribution. This is encouraging as JPL is closest to what we expect for workplace charging.

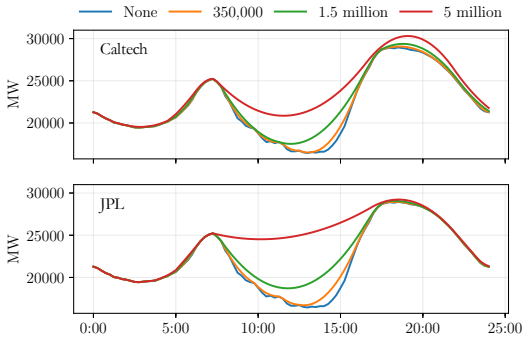

Figure 11: Net demand curves after optimal smoothing. EV penetration levels are based on the curent population of EVs in California and California's goals for 2025 and 2030. Upper panel: the Caltech data. Lower panel: the JPL data.

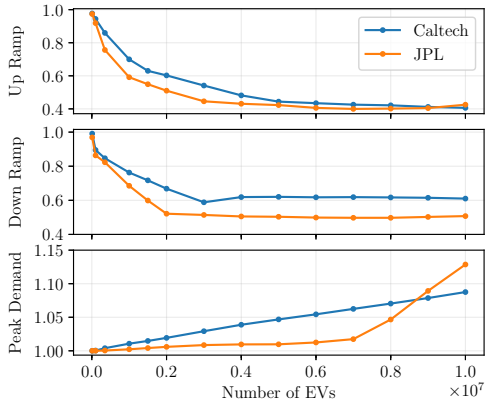

Figure 12: 15 minute maximum ramping rates and peak demand relative to the baseline without EVs.

8 CONCLUSION AND FUTURE DIRECTIONS

In this paper, we have presented **ACN-Data**, a unique publicly accessible dataset of workplace EV charging. We have presented some of the interesting driver behavior exhibited in the dataset and have described how to learn the underlying distribution using GMMs. We have demonstrated that predicting session duration and energy based on learned distribution from actual behavior in the past is more reliable than asking users to predict these parameters directly. We have also shown that with optimal solar sizing, about 50% of EV charging demand can be economically met by on-site solar generation at current prices, leading to savings of approximately $1,000/month in the summer. Finally, we have demonstrated that, with as few as 2 million EVs under control, we can limit the up and down ramp necessary to track net demand in California by nearly 50% with only a 0.6% increase in peak demand. In the future we hope that **ACN-Data** will facilitate work in many areas of EV research beyond those demonstrated here.

ACKNOWLEDGMENTS

This dataset would not be possible without the combined efforts of PowerFlex Systems, Caltech Facilities, and JPL Facilities. Specifically we would like to thank Ted Lee, Cheng Jin and George Lee of PowerFlex who were instrumental to collecting this dataset as well as Caltech students Sophia Coplin and Garret Sullivan who worked on cleaning the dataset and developing tools to make it

more accessible.. This material is based upon work supported by the NSF Graduate Research Fellowship (DGE-1745301) and NSF grants CCF-1637598, ECCS-1619352, CNS-1545096, and CPS-1739355.

REFERENCES

[1] Omid Ardakanian, Catherine Rosenberg, and S. Keshav. 2014. Quantifying the benefits of extending electric vehicle charging deadlines with solar generation. In *2014 IEEE International Conference on Smart Grid Communications (SmartGridComm)*. IEEE, Venice, Italy, 620–625. https://doi.org/10.1109/SmartGridComm.2014.7007716

[2] CAISO. 2016. What the duck curve tell us about managing a green grid. (2016). https://www.caiso.com/Documents/FlexibleResourcesHelpRenewables_FastFacts.pdf

[3] CAISO. 2019. Today's Outlook. (Jan. 2019). http://www.caiso.com/TodaysOutlook/Pages/default.aspx

[4] Niangjun Chen, Chee Wei Tan, and Tony Q. S. Quek. 2014. Electric Vehicle Charging in Smart Grid: Optimality and Valley-Filling Algorithms. *IEEE Journal of Selected Topics in Signal Processing* 8, 6 (Dec. 2014), 1073–1083. https://doi.org/10.1109/JSTSP.2014.2334275

[5] Zhong Chen, Ziqi Zhang, Jiaqing Zhao, Bowen Wu, and Xueliang Huang. 2018. An Analysis of the Charging Characteristics of Electric Vehicles Based on Measured Data and Its Application. *IEEE Access* 6 (2018), 24475–24487.

[6] Caroline Choi, Megan Scott-Kakures, and Akbar Jazayeri. 2017. General Service Time-Of-Use Electric Vehicle Charging - Demand Metered. (Aug. 2017). https://www1.sce.com/NR/sc3/tm2/pdf/ce141-12.pdf

[7] Yu-Wei Chung, Behnam Khaki, Chicheng Chu, and Rajit Gadh. 2018. Electric Vehicle User Behavior Prediction Using Hybrid Kernel Density Estimator. In *2018 IEEE International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*. IEEE, 1–6.

[8] K. Clement-Nyns, E. Haesen, and J. Driesen. 2010. The Impact of Charging Plug-In Hybrid Electric Vehicles on a Residential Distribution Grid. *IEEE Transactions on Power Systems* 25, 1 (Feb. 2010), 371–380. https://doi.org/10.1109/TPWRS.2009.2036481

[9] Jonathan Coignard, Samveg Saxena, Jeffery Greenblatt, and Dai Wang. 2018. Clean vehicles as an enabler for a clean electricity grid. *Environmental Research Letters* 13, 5 (2018), 054031.

[10] J.D. Cross and R. Hartshorn. 2016. My Electric Avenue: Integrating Electric Vehicles into the Electrical Networks. Institution of Engineering and Technology, 12 (6 .)–12 (6 .). https://doi.org/10.1049/cp.2016.0972

[11] Julian de Hoog, Tansu Alpcan, Marcus Brazil, Doreen Anne Thomas, and Iven Mareels. 2015. Optimal Charging of Electric Vehicles Taking Distribution Network Constraints Into Account. *IEEE Transactions on Power Systems* 30, 1 (Jan. 2015), 365–375. https://doi.org/10.1109/TPWRS.2014.2318293

[12] Paul Denholm, Michael Kuss, and Robert M. Margolis. 2013. Co-benefits of large scale plug-in hybrid electric vehicle and solar PV deployment. *Journal of Power Sources* 236 (Aug. 2013), 350–356. https://doi.org/10.1016/j.jpowsour.2012.10.007

[13] Chris Develder, Nasrin Sadeghianpourhamami, Matthias Strobbe, and Nazir Refa. 2016. Quantifying flexibility in EV charging as DR potential: Analysis of two real-world data sets. In *2016 IEEE International Conference on Smart Grid Communications (SmartGridComm)*. IEEE, Sydney, Australia, 600–605. https://doi.org/10.1109/SmartGridComm.2016.7778827

[14] Emil Eirola and Amaury Lendasse. 2013. Gaussian mixture models for time series modelling, forecasting, and interpolation. In *International Symposium on Intelligent Data Analysis*. Springer, 162–173.

[15] Marco Giacomo Flammini, Giuseppe Prettico, Andreea Julea, Gianluca Fulli, Andrea Mazza, and Gianfranco Chicco. 2019. Statistical characterisation of the real transaction data gathered from electric vehicle charging stations. *Electric Power Systems Research* 166 (Jan. 2019), 136–150. https://doi.org/10.1016/j.epsr.2018.09.022

[16] Janine M. Freeman, Nicholas A. DiOrio, Nathan J. Blair, Ty W. Neises, Michael J. Wagner, Paul Gilman, and Steven Janzou. [n. d.]. System Advisor Model (SAM) General Description (Version 2017.9.5). ([n. d.]).

[17] Ran Fu, David Feldman, and Robert Margolis. 2018. *U.S. Solar Photovoltaic System Cost Benchmark: Q1 2018.* Technical Report.

[18] Qiuming Gong, Shawn Midlam-Mohler, Vincenzo Marano, and Giorgio Rizzoni. 2012. Study of PEV Charging on Residential Distribution Transformer Life. *IEEE Transactions on Smart Grid* 3, 1 (March 2012), 404–412. https://doi.org/10.1109/TSG.2011.2163650

[19] Pecan Street Inc. 2019. Pecan Street Dataport. (2019). https://dataport.cloud

[20] Behnam Khaki, Chicheng Chu, and Rajit Gadh. 2018. A Hierarchical ADMM Based Framework for EV Charging Scheduling. In *2018 IEEE/PES Transmission and Distribution Conference and Exposition (T&D)*. IEEE, Denver, CO, USA, 1–9. https://doi.org/10.1109/TDC.2018.8440531

[21] G. Lee, T. Lee, Z. Low, S. H. Low, and C. Ortega. 2016. Adaptive Charging Network for Electric Vehicles. In *2016 IEEE Global Conference on Signal and Information Processing*.

[22] Stephen Lee, Srinivasan Iyengar, David Irwin, and Prashant Shenoy. 2016. Shared solar-powered EV charging stations: Feasibility and benefits. In *2016 Seventh International Green and Sustainable Computing Conference (IGSC)*. IEEE, Hangzhou, China, 1–8. https://doi.org/10.1109/IGCC.2016.7892600

[23] Zachary J. Lee, Daniel Chang, Cheng Jin, George S. Lee, Rand Lee, Ted Lee, and Steven H. Low. 2018. Large-Scale Adaptive Electric Vehicle Charging. In *IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids*.

[24] Bruce G Lindsay. 1995. Mixture models: theory, geometry and applications. In *NSF-CBMS regional conference series in probability and statistics*. JSTOR, i–163.

[25] Alexandre Lucas, Giuseppe Prettico, Marco Flammini, Evangelos Kotsakis, Gianluca Fulli, and Marcelo Masera. 2018. Indicator-Based Methodology for Assessing EV Charging Infrastructure Using Exploratory Data Analysis. *Energies* 11, 7 (July 2018), 1869. https://doi.org/10.3390/en11071869

[26] Yorie Nakahira, Niangjun Chen, Lijun Chen, and Steven H. Low. 2017. Smoothed Least-laxity-first Algorithm for EV Charging. ACM Press, 242–251. https://doi.org/10.1145/3077839.3077864

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[28] G.A. Putrus, P. Suwanapingkarl, D. Johnston, E.C. Bentley, and M. Narayana. 2009. Impact of electric vehicles on power distribution networks. In *2009 IEEE Vehicle Power and Propulsion Conference*. IEEE, Dearborn, MI, 827–831. https://doi.org/10.1109/VPPC.2009.5289760

[29] Arvind Ramanujam, Pandeeswari Sankaranarayanan, Arun Vasan, Rajesh Jayaprakash, Venkatesh Sarangan, and Anand Sivasubramaniam. 2017. Qantifying The Impact of Electric Vehicles On The Electric Grid: A Simulation Based Case-Study. In *Proceedings of the Eighth International Conference on Future Energy Systems - e-Energy '17*. ACM Press, Shatin, Hong Kong, 228–233. https://doi.org/10.1145/3077839.3077854

[30] Jose Rivera, Christoph Goebel, and Hans-Arno Jacobsen. 2017. Distributed Convex Optimization for Electric Vehicle Aggregators. *IEEE Transactions on Smart Grid* 8, 4 (July 2017), 1852–1863. https://doi.org/10.1109/TSG.2015.2509030

[31] Alexander Schuller, Christoph M. Flath, and Sebastian Gottwalt. 2015. Quantifying load flexibility of electric vehicles for renewable energy integration. *Applied Energy* 151 (Aug. 2015), 335–344. https://doi.org/10.1016/j.apenergy.2015.04.004

[32] Bin Wang, Yubo Wang, Hamidreza Nazaripouya, Charlie Qiu, Chi-cheng Chu, and Rajit Gadh. 2016. Predictive Scheduling Framework for Electric Vehicles Considering Uncertainties of User Behaviors. *IEEE Internet of Things Journal* (2016), 1–1. https://doi.org/10.1109/JIOT.2016.2617314

[33] Di Wu, Haibo Zeng, Chao Lu, and Benoit Boulet. 2017. Two-Stage Energy Management for Office Buildings With Workplace EV Charging and Renewable Energy. *IEEE Transactions on Transportation Electrification* 3, 1 (March 2017), 225–237. https://doi.org/10.1109/TTE.2017.2659626