Signature Discovery

1 Introduction

One of the key challenges facing power companies and motivating the adaption of advanced metering infrastructure (AMI), is the rapidly increasing market share of electric vehicles (EV). Tackling the large power surges associated with EV charging is going to require deep insight into the nature of these charging cycles.

A natural question to ask is, what does a typical charging cycle look like? I attempt here to answer this question based on data from the research and technology organization Pecan Street.[1]

2 Data Preparation

The data I have used comes from the Pecan Street Dataport database (www.dataport.cloud). Pecan Street is a nonprofit research institute focused on the utility sectors and has data from over 1400 households freely available for academic use.

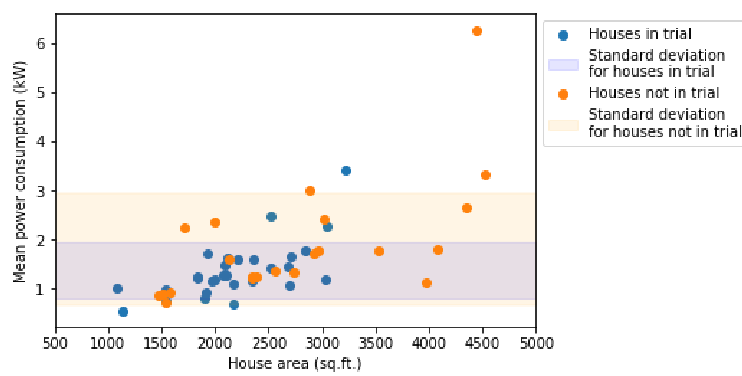
2.1 Raw Data

For this experiment I collected power data from the EV-charger outlet, as well as the household net power drain on the grid (power use minus power generation) at 1-minute resolution. 55 houses had records of this data from January 1st to December 31st, 2017. Out of these houses, 12 houses had incomplete records, and were initially discarded for convenience, and later used for testing.

Out of these 55 houses, four were town homes, sharing at least one wall with a neighbor. The rest were free standing single-family homes. 30 of the houses were enrolled in a CCET electricity pricing and customer behavior trial in 2013 and 2014. Out of these, three were in the text message group and received text messages on days with a high peak load asking them to reduce their power consumption. Four were in the UT text message group and received text messages asking them to restrict usage of specific appliances on peak days. Eighteen households had access to information about their hypothetical electricity cost with an hourly rate. Sixteen of these received financial incentives for reducing their power consumption according to this rate. The rest were in the control group.

It is unlikely that a study concluded three years before had much to say on the behavior of the consumers. The differences observed in figure 1 can probably be attributed to the larger share of big houses among those which did not participate in the trial.

Figure 1



Before using the data, I downsampled it to a 15-minute resolution, taking the mean over each 15-minute interval. This in order to match the likely resolution of Norwegian smart meters[2] and reduce the computational effort necessary.

2.2 Extracting Charging Signatures

To get data with which I could train a model, I created a library of EV-charging signatures. This was done by collecting all instances of at least three consecutive measurements from the EV-charger of a house with a value of at least 1 kW. To preserve the shape of the signal, the signatures were also padded at the beginning with the last preceding measurement, and at the end with the first succeeding measurement.

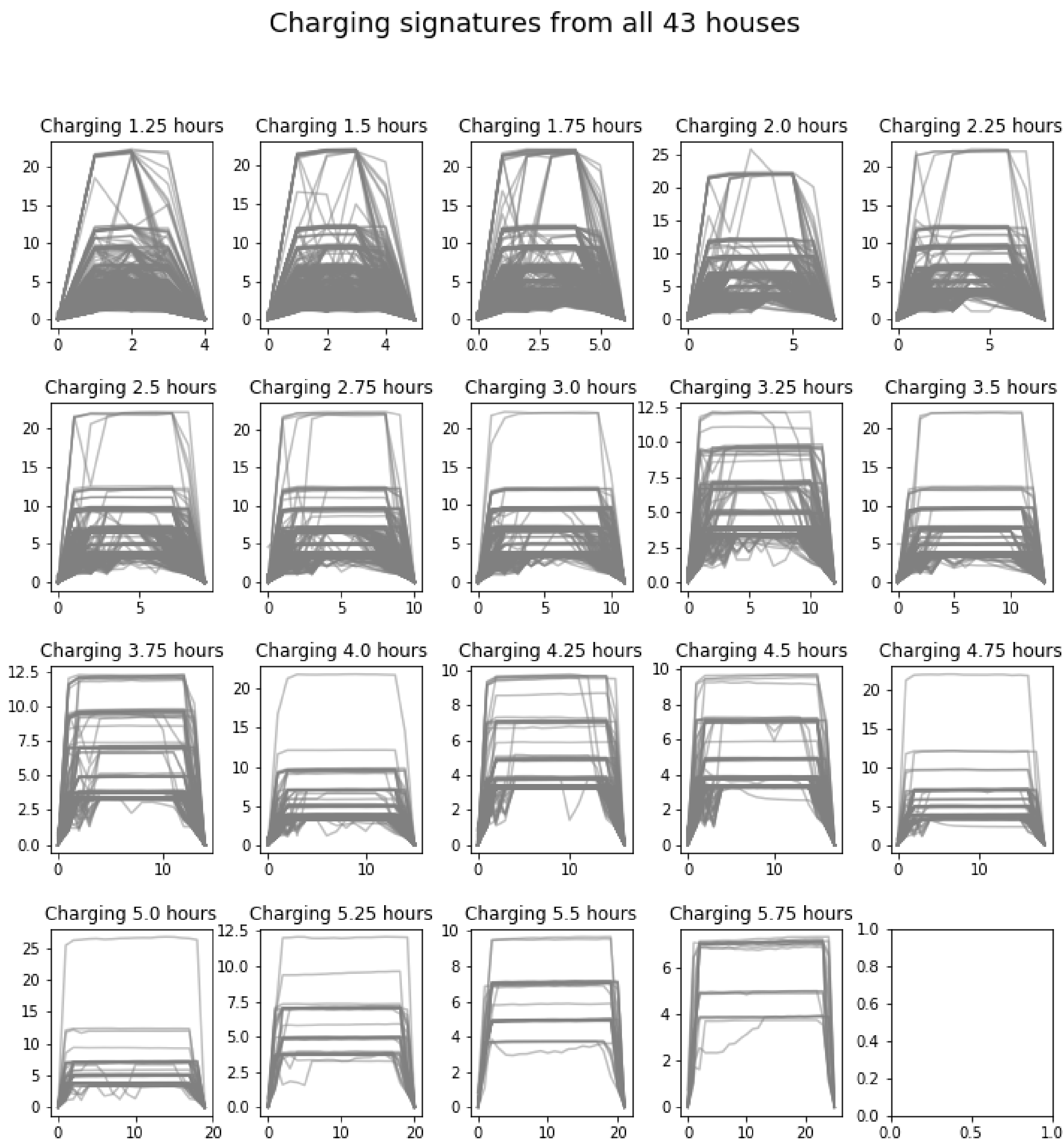
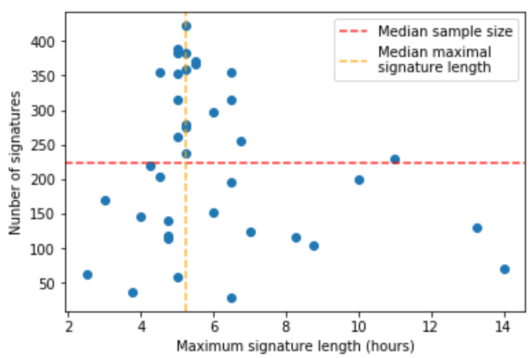


Figure 2: When grouped by length, the charging signatures across all houses display groupings in typical charging patterns, mostly defined by their maximum power drain. The y-axes here are the effect measured at the EV charger outtake in kW, while the x-axes are timesteps of 15 minutes.

Figure 3 compares the data drawn from each house by number of signatures collected versus the longest signatures collected. Figure 2 shows all collected charging signatures plotted by duration.

Figure 3:



3 Templates

Figure 3 conveys a decent idea about the shape of all the individual charging signatures in the data. Interestingly, the signals do appear to come in regular forms. For many applications, such as signal discovery by matched filters, using all the discovered charging signatures is infeasible. A typical, representative pattern, or an aggregate signature for each charging duration are therefore needed.

The raw signatures are first sorted by their lengths to be more easily comparable. For the purposes of finding typical charging signatures, every signature of a length of which there are less than 100 signatures is discarded to reduce the variance of the results. Signatures of the three shortest charging durations are also discarded, as they are very prone to noise.

A way to discern discrete categories into which the signatures fall is required. Figure 4 shows the average maximum power drain per charging event for each house plotted as a horizontal line. From this perspective there appears to be seven clear groupings of signatures. Figure 5 shows the maximum power drain per individual signature. This figure is less easily interpretable but appears to reveal at least four or five distinct groups.

Figure 4:

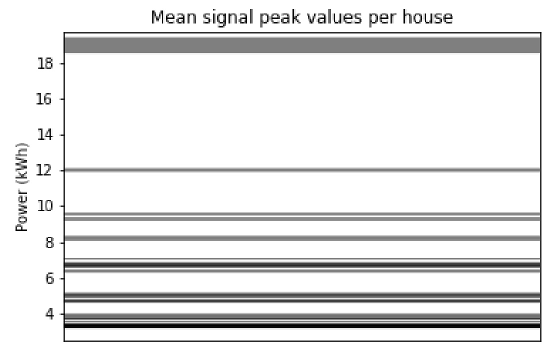
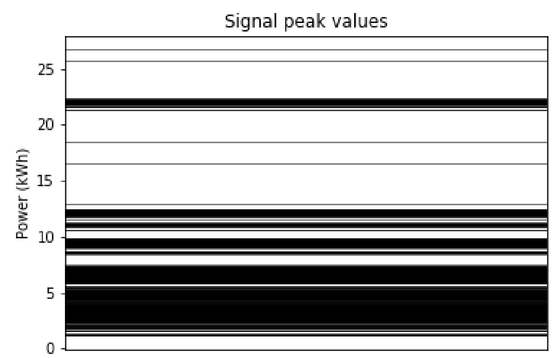


Figure 5:



3.1 Clustering

Clustering methods are a class of generally unsupervised machine learning techniques, which attempt to discover natural formations, or clusters, in the data. Clustering methods are thus well suited for the problem of discovering typical charging signatures.

Among the most common clustering methods is the k-means algorithm. K-means fits k centroids to the data by first creating centroids randomly, then iteratively assigning each data point to the centroid closest to it (by Euclidean distance) and moving each centroid towards the center of mass of the points currently assigned to it. More formally, at every time step, each data point x_p is assigned a cluster C_i belonging to a centroid c_i by:

$$C_i = \left\{ x_p : \left\| x_p - c_i \right\|^2 \leq \left\| x_p - c_j \right\|^2 \forall j, 1 \leq j \leq k \right\}$$

Then every centroid is moved by:

$$c_i^{(t+1)} = \frac{1}{|C_i|} \sum_{x_j \in C_i} x_j$$

Finally, when the centroids' movement converges, or a maximum number of iterations is reached, the algorithm stops.

3.2 Choosing k

The biggest problem with the k-means algorithm for this purpose is selecting the number of clusters to fit. Figure 4 hints at seven clusters, but a more thorough investigation is clearly needed, preferably fully automated and generalizable. This problem is of course well examined in literature, but perhaps with a known, regular data structure it is possible to do better than an educated guess.

3.2.1 Silhouette score

To determine what is a good clustering, a metric is required. One such metric is the silhouette score. The silhouette score measures the relationship for each data point between intra-cluster distance and the distance to the data points in the closest neighboring cluster. Specifically, the silhouette score $s[i]$ for a data point i is defined as:

$$s[\mathrm{i}] = \frac{\mathrm{b[i]} - \mathrm{a[i]}}{\max(\mathrm{b[i]}, \mathrm{a[i]})}$$

Where $b[i]$ is the average distance from data point i to every data point in the closest cluster which i does not belong to, and $a[i]$ is the average distance from i to every data point in the same cluster as i. This means that a data point with a score close to 1 is perfectly clustered, while one with a negative score is incorrectly clustered.

There are two common ways to use this metric: By examining the average silhouette score of all clustered points, and by examining at the silhouette scores of the individual data points within each cluster. The latter method is useful when choosing the number of clusters because it reveals clusters which are particularly well or poorly suited, for example by comparing the maximum silhouette score of a cluster to the global average. In the present case, both analyses are useful.

Another factor which influences our choice of k is that since we are looking specifically for typical nuances in the signatures, it makes sense to choose more clusters rather than fewer.

Figure 6 shows the average silhouette score of all data points for each signature length, with the mean per choice of k in black and the mean of these means shown as a horizontal red line. Since we are interested in discovering more, rather than fewer typical signatures, a good choice of k from this graph could be at the "elbow". This is where increasing the number of clusters no longer corresponds with a drop in silhouette score, implying that there are no more sensible partitions of the data after this.

Figure 6:

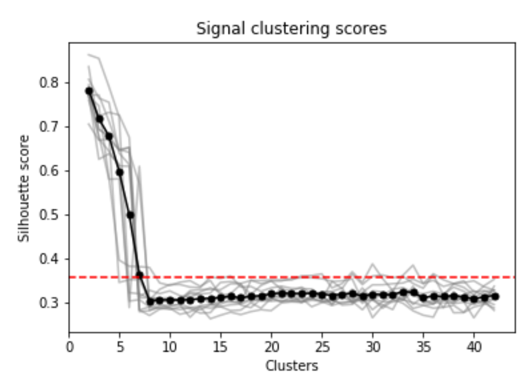


Figure 7 shows the proportion of "poor" clusters – that is, clusters where all the points fall beneath the global average silhouette score – per choice of k for each signature length. Figure 8 shows the distribution of "perfect" clusterings – that is, clusterings in which no cluster falls completely below the global average – per choice of k. Together, these two graphs show that a k around seven yield the lowest number of poor clusters. This coincides well with the intuition gained from figure 6 and 4.

Figure 7:

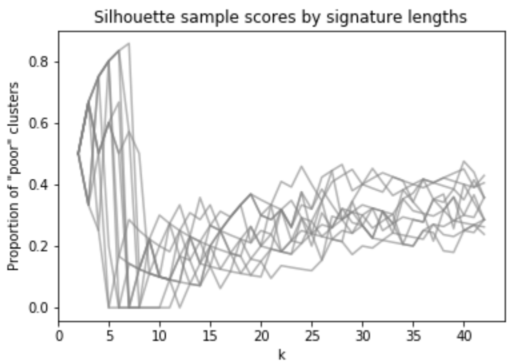
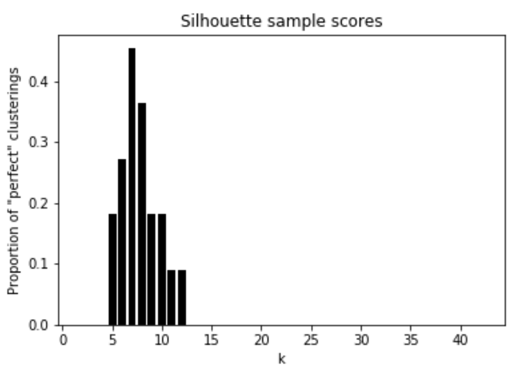


Figure 8:

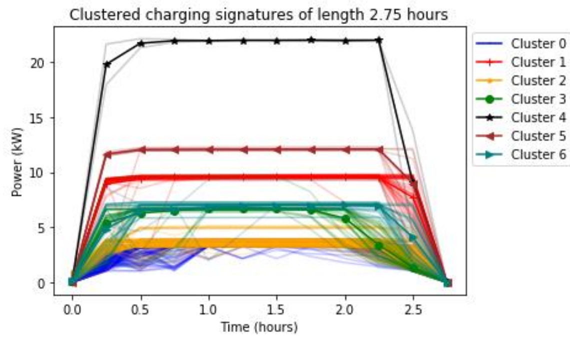


A pragmatic approach to choosing k for our purposes (crucially, choosing a single k for multiple signature lengths) can thus consist of two measures: The highest k with silhouette score above some deviation from the mean for a sufficiently large range of k, and the k with the largest number of "perfect" clusterings. In this case, the two measures coincide and yield $k = 7$.

3.3 Choosing Templates

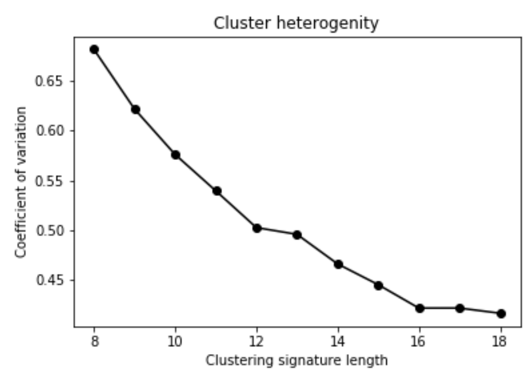
Figure 9 plots signatures of length 3 hours by cluster with $k = 7$, showing the cluster means as heavy lines with markers.

Figure 9:



This shows that the clustering seems reasonable at least for this sample signature length. Figure 10 plots the average coefficient of variation (standard deviation normalized by mean) for each signature length, showing that longer signatures lead to more coherent clusters.

Figure 10:

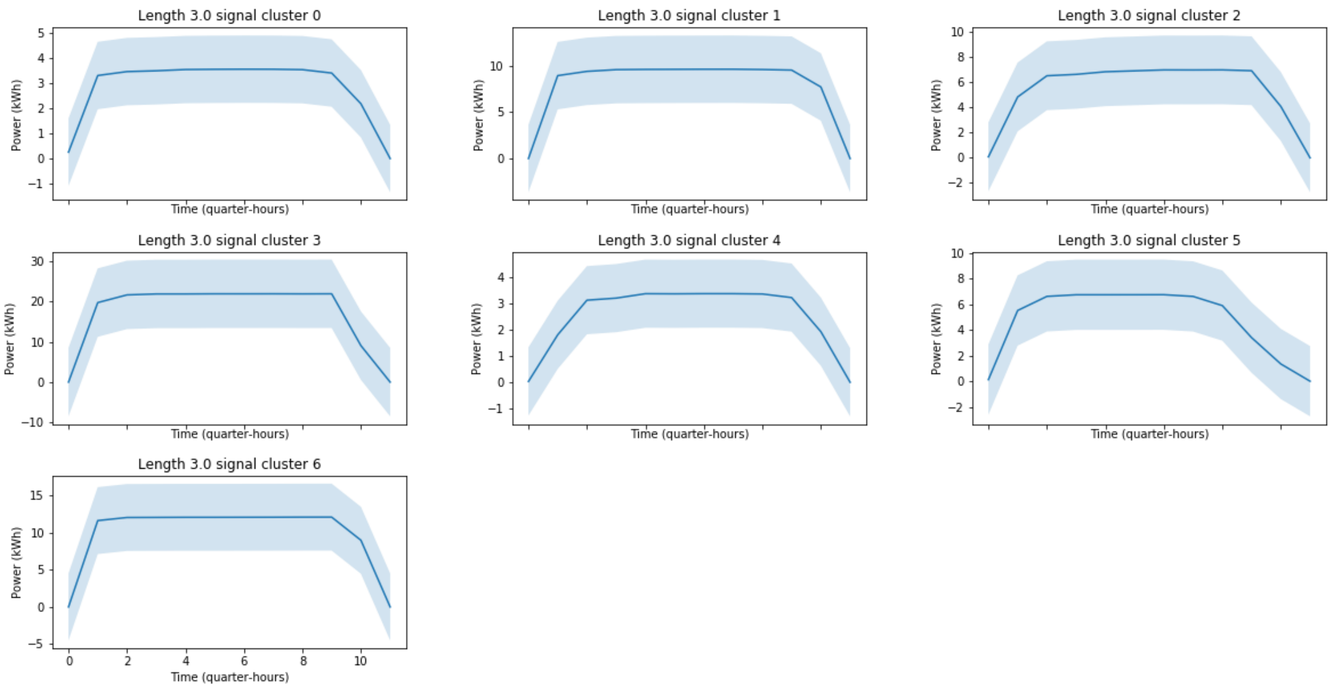


4 Conclusion

Since these clusterings ideally capture the distinct structures in the data, their means provide good insight into the shape of typical charging signatures. Choosing k remains the primary bottleneck, and while this example led to an unequivocal choice, automating the selection for less obvious cases may provide a challenge.

Figure 11 shows the final results for the sample signature length of three hours, with the templates (cluster means) in dark blue and one standard deviation in light blue.

Figure 11: Sample resulting signature templates. Cluster 3 and 5 display an exponential decline seen clearly in the signatures of some households at higher resolution. The rest appear to be clustered by mean power drain.



5 Bibliography

[1] “Dataport from Pecan Street.” [Online]. Available: https://dataport.cloud/.

[2] Norsk vassdrags- og energidirektorat, “NVE - slik fungerer måleren.” .