Baseline in Streaming Data Analysis: Why, What, and How

When performing projection, prediction, and feature extraction with a time series, we often need to determine a baseline. This baseline typically captures the common or expected behavior, subtracting the baseline would allow the unusual and unexpected features in the data to be more easily revealed. In this brief note, we use a concrete example of extracting a baseline from a set of residential electricity usage data to illustrate the need for the baseline and potential approaches for derive this baseline.

Because electricity cannot be easily stored, electricity generation must be continuously matched to its consumption. A serious problem can occur when the peak demand exceeds the generation capacity, leading to a blackout during the time when consumers need the electricity the most. Since increasing generation capacity is expensive and takes a long time to implement, regulators are interested in devising pricing schemes that would discourage unnecessary consumption of consumers and businesses during peak demand periods.

In order to measure the effectiveness of a pricing policy on the peak demand periods, one can analyze electricity usage data generated from the monitoring system know as the advanced metering infrastructure (AMI). Figure 1 shows a set of average daily usage data from a large-scale study of residential electricity usage in a region of US where the usage peaks during the summer months. The data in this figure shows the usage from three consecutive summers marked as T-1, T, and T+1. The households in this study are divided into a number of different groups with

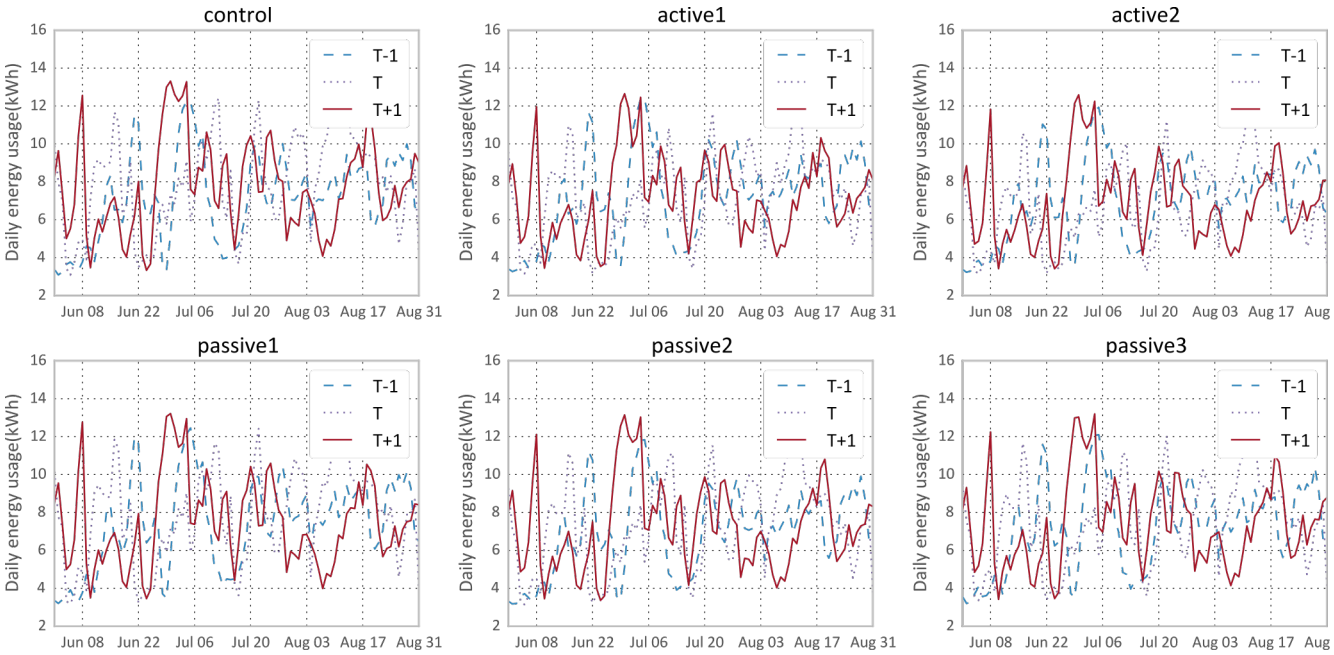

Figure 1: a sample set of electricity usage data from three consecutive nears (labeled as T-1, T, and T+1) in an large-scale experiment with carefully designed groups (labeled control, active1, active2, passive1, passive2, and passive3), however, usage curves are nearly identical because the common factors such as temperature and holidays dominate the typical electricity usage.

careful selection criteria indented to measure the impact of a number of different pricing schemes. However, as we can see from Figure 1, the electricity usage curves from different groups are very similar to each other. A key reason for this similarity is that residential electricity usage has a strong dependency on outside temperature during the summer time because the demand of air-conditioning during summer season. Additionally, national holidays and weekends also have significant impact on the electricity usage. In order to measure the impact of different pricing scheme, we need to develop a baseline to capture these common factors.

In general, a baseline is used to capture a set of features that are considered as expected or normal. In this particular case of residential electricity usage, we might want a baseline to represent an average household taking into account of outside temperature, day of the week and other common factors. In another application, the baseline might take on a very different set of features. For example, in an astronomy observation, the baseline for an image might need to take into account of cloud cover and brightness of the moon.

Although the work of constructing a baseline shares some similarities with other works on forecasting electricity demands and prices, there are a number of distinctive characteristics that necessitate us considering a different class of data analysis and data mining methods. The fundamental difference between a baseline model and a forecast model is that the baseline model needs to capture core behavior that might persist for a long time, while the forecast model typically aims at making a forecast for the next few cycles of a time series in question. In general, techniques that make long-term forecasts are based on highly aggregated time series with month or year as time steps, whereas those that work on time series with shorter time steps typically focus on making forecasts for the next day or the next few hours.

In the specific case that has motivated our work, the overall objective is to study the impacts of proposed pricing policies. The process of designing these pricing schemes, recruiting participants, implementing the pricing schemes, and monitoring the impacts has taken a few years. The baseline model here is expected to be established based on observed consumption prior to the implementation of the new pricing schemes, and applied to predict what consumer behavior would be without the pricing changes years into the future. This is challenging because the baseline model needs to not only capture intraday household electricity usage but also be applicable for years. Furthermore, when studying the impact of the proposed pricing scheme on electricity usage, we might notice that the impact of the scheme is weaker than the impact of other factors such as temperature, in which the baseline model for electricity usage must take these other factors into account.

In our work, we have examined a number of methods for developing the baseline models that could satisfy these requirements. From our study, we find that Gradient Tree Boosting (GTB) to be quite effective in extracting the baseline. Unlike previous approaches that focus on forecasts, our baseline model using GTB eliminates unnecessary effects of features such as temperature and captures core user behavior over the years.