ADAML 2025 Project Work - Predicting House Energy - Intermediary Submission Week 2

Toni Koskinen
Veeti Rajaniemi
Amanda Valtanen

November 9, 2025

1 Introduction

In this work, the aim is to forecast daily electric power consumption for a house with previous energy consumption data.

2 Data description

The original time-series dataset contains approximately 2,000,000 measurements from a household between December 2006 and November 2010, with one measurement recorded per minute. In addition to date and timestamp, it includes 6 variables as follows:

- Global active power - minute-averaged active power (the target variable)
- Global reactive power - minute-averaged reactive power
- Voltage - minute-averaged voltage
- Global intensity - minute-averaged current intensity
- Sub-metering 1 - energy sub-metering corresponding to kitchen appliances
- Sub-metering 2 - energy sub-metering corresponding to the laundry room
- Sub-metering 3 - energy sub-metering corresponding to an electric water heater and an air-conditioner.

The dataset contains individual timestamps with missing values. Also, there are a few time periods when the measurements are missing for several straight days.

3 Data preprocessing and visualization

If the missing values were individual measurements or occurred over a short gap, maximum of 12 hours, they were approximated by simple linear interpolation. If the values were missing for a longer period, we used the value from one week before, as it correlates most with the missing value based on autocorrelation, which will be shown later.

As the goal of the work is to predict daily consumption, we decided to modify the original minute-level data into hourly measured values. Thus, the mean value of each variable during each hour is used.

The dataset after these modifications can be seen in Figure 1. Based only on these visualizations, we can see some kind of seasonality in almost each variable.

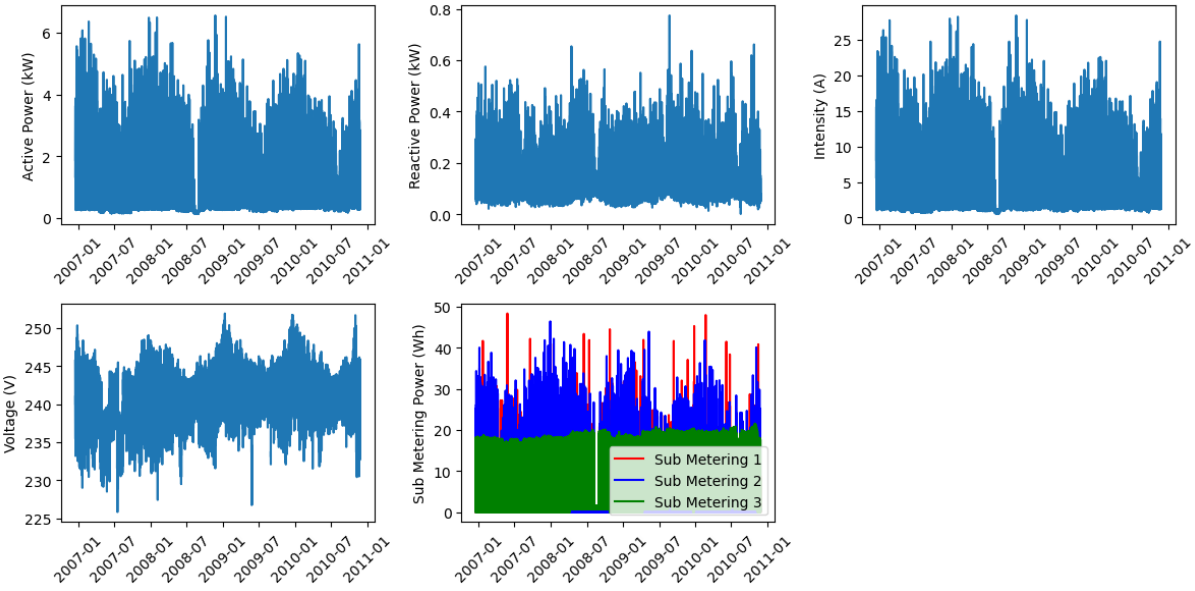

Figure 1: Variables after handling missing values and modifying data into hourly averages.

4 Time-series decomposition

Time-series decomposition refers to breaking down a time series into its different components in order to understand what causes variation of a specific variable. The components are trend, seasonality, and residual. The trend indicates long-term variation, seasonality indicates repeating seasonal variation, and residual indicates random variation that cannot be explained by trend or seasonality.

The time-series data of the target variable is used to analyze and understand its temporal behavior. Time-series decomposition was first applied on data containing hourly averages of target variable, using a season length of one day (24 hours). The result of this is shown in the Figure 2. The figure shows that the trend still includes seasonal variability, which could be interpreted as annual, because at the same time of year, the trend behaves repeatedly in the same way. For this reason, time-series decomposition is not satisfactory, and we must try to define the length of the season so that the trend does not include seasonal variation.

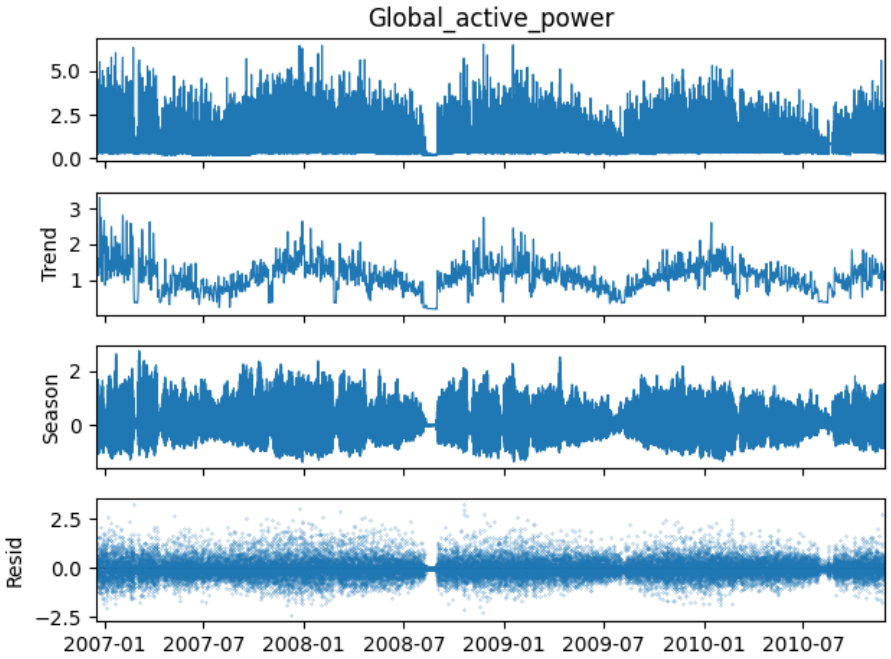

Figure 2: Time-series decomposition of the Global Active Power variable using hourly average values, and a day as a season length.

In the Figure 3, time-series decomposition is created for hourly average data, with the season length being one year (24 hours * 365). In Figure 4, daily averages have been taken from the data for the explained variable and a time-series decomposition has been performed using as a season length a year (365 days). Both annual season length distributions show a similar, slightly decreasing trend. Seasonality shows annual variation, with values of the target variable falling in summer and peaking in winter. The residuals appear to be quite evenly distributed around zero, although in the middle of the time series the residuals clearly show greater variance compared to the beginning and end.

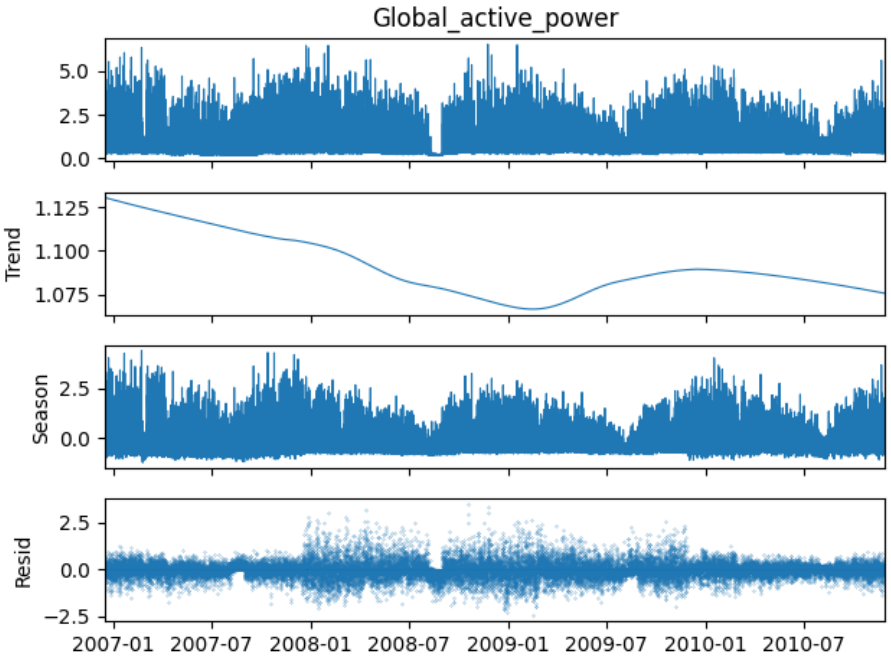

Figure 3: Time-series decomposition of the Global Active Power variable using hourly average values, and a year as a season length.

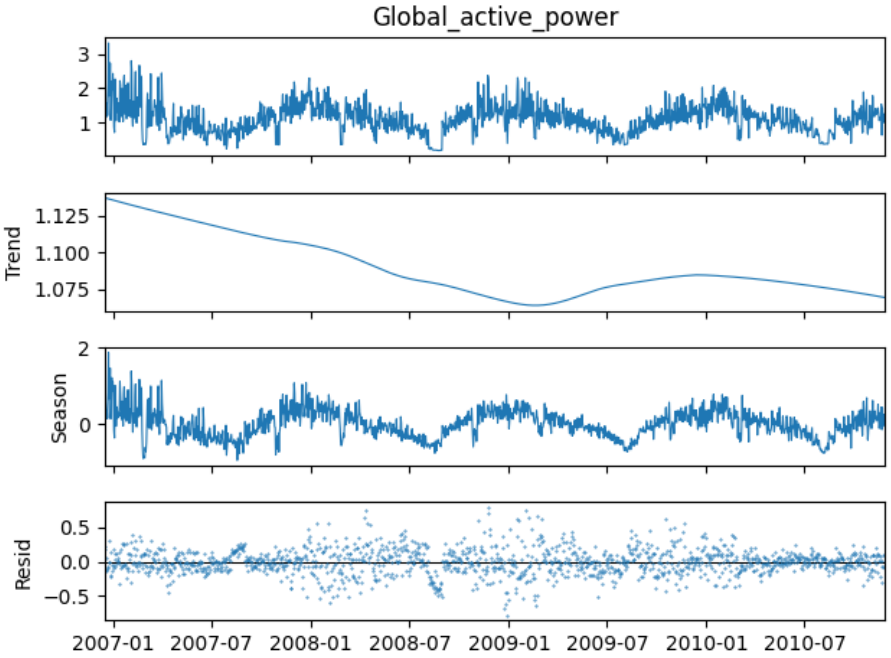

Figure 4: Time-series decomposition of the Global Active Power variable using daily average values, and a year as a season length.

5 Autocorrelation analysis

Autocorrelation function can be used to evaluate relationships between past and present observations. Significant correlation between these indicate that the time series is not a random walk.

Before using autocorrelation function, the stationarity of the observations was evaluated using Augmented Dickey-Fuller (ADF) test. For Global active power, the ADF Statistic of -11.12 and

p-value of $3 \cdot 10^{-20}$ indicated stationary observations, so no time series transformations were needed before applying the autocorrelation function.

The autocorrelation function plots for Global active power, using 48 and 168 lags are shown in Figure 5. Each lag represents delay of 1 hour, which means that 24 lags correspond to a one-day delay and 168 lags a one-week delay.

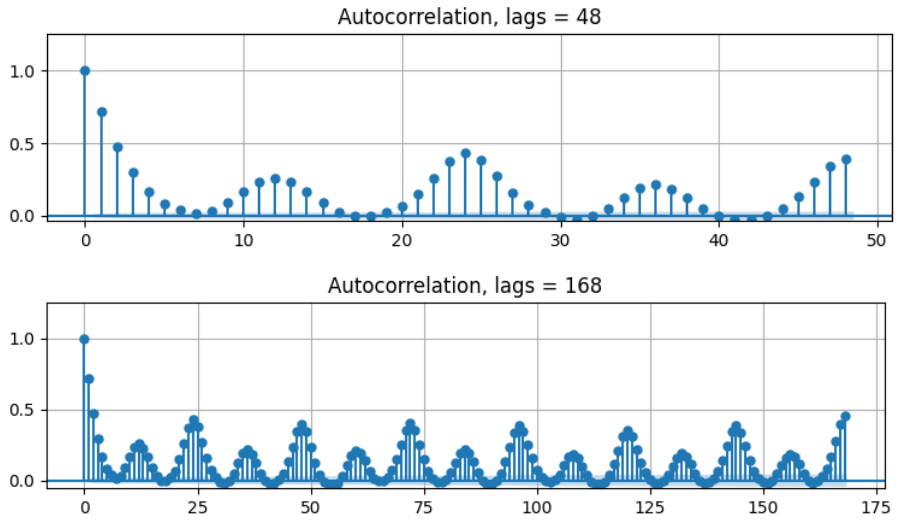

Figure 5: Autocorrelation function plots for Global active power, using 48 and 168 lags.

From the figure we could see that there were significant relationship between the past and the present observations, especially in daily and weekly periods. These significant relationships occurred also in explanatory variables.

6 Data partition plan for cross-validation

The data will be divided into several folds for cross-validation. We will start with a basic approach doing cross-validation on a rolling basis with expanding window. We will use 5 folds, and the size of each test set will be a fifth of one year (73 days). In practice, it means that we first use all data but the last year (meaning last 365 days of the dataset) as a training set, and the first fifth of the last year of the data is the first test set. During the second fold, we add the previous test set into our training data, and try to predict the second fifth of the last year. This process is repeated so that across five folds, the test periods together span the entire last year of the data.

With this approach, the validation process includes the full seasonality of one year. Of course, we can try to change the number of folds and the size of test sets later. After cross-validation is done, we can also try to predict values for the entire last year to see if the model performs well for a longer period. One option is to leave some data as fully unseen for our model so we can test it. These details will be decided later when the model type is known.