Algorithms in Data Science

ENS Challenge: Dynamic Profile Forecasting

Matthieu Briet - Tanguy Colleville - Antoine Pagneux
CentraleSupélec - 3A

February 13, 2022

Sous l'encadrement de :
Dr. Frederic Pennerath

CentraleSupélec

1 Introduction

ENEDIS is the main distribution system operator in France (95% of continental France, 36 million customers). The electricity market requires that production and consumption be assigned to an upstream-downstream balance operator, each electricity provider having at least one. These balance operators guarantee that supply and demand are balanced every half-hour 24/7. In order to assign energy, Enedis computes the electricity consumption balance every week for each balance operator. Part of the balancing computation involves some consumption or production modelling for several groups of customers. We would like to forecast 7 dynamic profile time-series, modelling the consumption shape of several mass-market customer groups (residential and small businesses with subscribed power up to 36 kVA) thanks to meteorological and calendar data, as well as any other real time dataset potentially correlated with consumption patterns. As you read through this document, you will see our state of the art containing a compilation of what we have seen in the scientific literature relevant to our mission, our intellectual pathway to our results and the model selected.

2 State of the Art

Processing times series is a very particular practice which requires precautions in order to carry out their regressions. In this short state of the art, we will focus on features engineering and the model.

2.1 Preprocessing

Some notions, although important to introduce, will only be quoted here for the sake of conciseness but they have been studied in depth.

- Autocorrelation plot is a measure of the correlation between the time series and the lagged version of itself.
- Partial correlation is a measure of the correlation between the time series and the lagged version of itself, after taking into account all other terms with a shorter lag.
- Stationnarity, we can test the hypothesis of a time series with the Dickey Fuller test whose conservative hypothesis is instationnarity.
- Ergodicity defines the ability to calculate statistical values from a realization of the stochastic process.
- Homoscedasticity, is when the variance of the stochastic errors of the regression is the same for each observation. We can test such a property with the Breusch-pagan test whose conservative assumption is homoscedasticity.
- Kurtosis, asymetric coefficent and Fisher-Pearson coefficient are based on moments of order 5 and characterizes the symmetry of the distribution and its ”heaviness”.
- Cointegration allows to detect the long term relationship between two or more time series.

There are common best practices in features engineering and time series forecasting, including rolled features and lag features.

Also, we can see and easily imagine that the courses of energy consumption follow seasonality. To detect them we can use Fourier transforms.

2.2 Modelling

There are many regression models, namely Gaussian process, linear regression with penalty, Generalized Linear Model,Generalized Additive Model. Markov processes or chains are particularly interesting tools to model time-related phenomena when they are characterized by an autocorrelation. Here

Concerning the statistical models, we had a large panel at our disposal to perform a regression. However, the time series and the dataset led to the exclusion of some of the statistical models mentioned. Our study of the state of the art and the data led us to select 3 types of models namely: SARIMAX, ARIMAX, and PROPHET.

Concerning deep learning, there are many solutions. However, we have selected some that are more conducive than others in the realization of such a challenge. CNNs can be used insofar as

the data present a structure along the time axis. Then we note that recurrent networks inevitably come first. It is possible to use CNNs combined with LSTMs CNN-LSTM. However, these models, although classical, are sources of error concerning the instationnarities and long term dependancies. This is why transformers with attention models seem to be efficient. However, Deep learning was not on the agenda, although it is omnipresent in the time series prediction literature.

2.3 Monitoring

To evaluate these models we can use the Root Mean Squared Error, an imposed metric, but we can also use criteria such as AIC and BIC. These two criteria must be minimized although they do not optimize the same things. In fact, The Akaike Information Criteria thus represents a compromise between bias (which decreases with the number of parameters) and parsimony, which is important according to Ockham's razor. Bayesian Information Criteria seeks to select the most likely model given the data.

3 Our intellectual pathway

The idea is to start from a very simple naive solution, whose weakness we have noted by a very low score, to finally orient ourselves towards more and more complex solutions. In this chapter we will develop some aspects that we considered very important when building iteratively our more and more sophisticated models.

- Dataset Analysis : Distribution Analysis, vizualisation.
- Modeling of the trend, seasonal variations (season, month, day).
- Checking for stationnarity, autocorrelogram, partial autocorrelogram.
- Study of correlation link between variables.

4 Dataset

In this section we would like to introduce a brief presentation of the datasets and the adjustments applied. The data we had available to train our models fell into 4 main categories:

1. The first one concerns meteorological data:
    - Smoothed achieved temperature (°C).
    - Smoothed normal temperature (°C).
    - Pseudo radiation.
2. The second one represents the measured data:
    - Power (W) injected by RTE.
    - Power (W) discharged at RTE.
    - Net extraction to other DSOs (W).
    - HV consumption remotely read on load curve (W).
    - Decentralised generation telemetered at load curve (W).
3. The third one represents the modelled data:
    - Modelled losses (W).
    - Total profiled consumption (W).
    - Profiled HV consumption (W).
    - Profiled SME-SMI consumption (W).
    - Professional consumption profiled (W).
    - Profiled residential consumption (W).
    - Profiled decentralised production (W).
    - Profiled photovoltaic generation (W).
    - Other profiled production (W).
4. The last one concerns sums calculated from the measured and modelled data:
    - Total consumption (W).

- Total decentralised production (W).
- Total wind power production (W).
- Total photovoltaic production (W).
- Total HV consumption (W).

Finally, and as an output of our models, we had to predict the following 7 time series in which we separate the residential from the professional and in which we then separate according to their contract with notably their subscribed power and their tariff option such as peak and off-peak hours:

- RES1_BASE: Residential customer profile, subscribed power up to 6kVA, without tariff option.
- RES11_BASE: Residential customer profile, subscribed power above 6kVA, without tariff option.
- RES2_HP: Residential customer profile, with peak/off-peak tariff option, during peak hours.
- RES2_HC: Residential customer profile, with On-Peak/Off-Peak tariff option, during off-peak hours.
- PRO1_BASE: Commercial customer profile, without tariff option.
- PRO2_HP: Commercial customer profile, with peak/off-peak tariff option, during peak hours.
- PRO2_HC: Commercial customer profile, with peak/off-peak tariff option, during off-peak hours.

The main difficulties with this challenge are firstly that our training data has a lot of missing values. For example, 76% of the values related to what we have to predict are missing, which does not necessarily facilitate learning. Secondly, the data we are manipulating are time series, in other words the order of the data makes sense. This does not allow us to separate our dataset into a training dataset and a test dataset to perform cross-validation as we could have done in a classical machine learning problem. In machine learning, it is admitted that cross-validation makes it possible to prevent over-fitting and to have a more robust model. In our case, we cannot randomly take samples from the available data and divide them into training and test sets. We need to keep the temporality. To do this, we start with a small subset of data to carry out an initial training, then we take the next subset of data directly to test our model. We repeat these steps several times, taking larger and larger training sets as follows in figure 1.

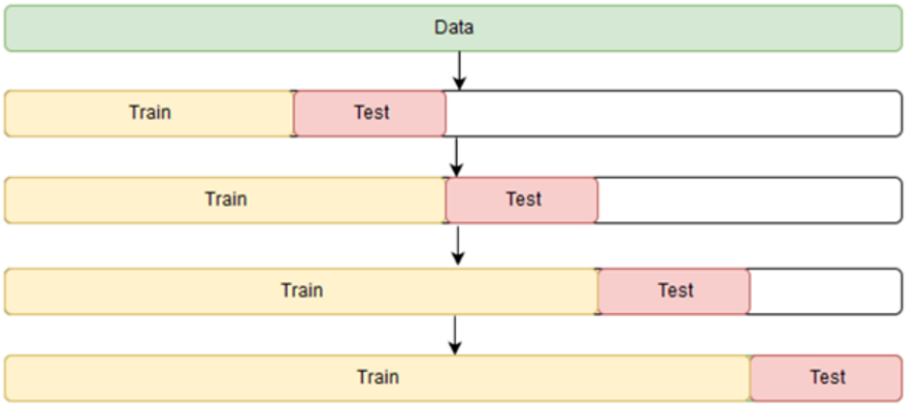


Figure 1: Cross-validation for time series.

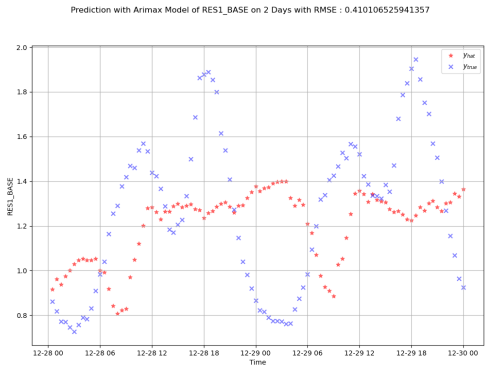
5 Our models

1. ARIMAX: Auto-Regressive Integrated Moving Average with exogenous variables is an extension of ARIMA models. The difference lies in the addition of exogenous variables X. An $ARIMAX(p, d, q)$ model is defined for a time serie y_t with exogenous variables X_t. p is the number of auto-regressive lags. d is the degree of differentiation and q is the number of moving-average lags. Such a model is defined this way on equation 1

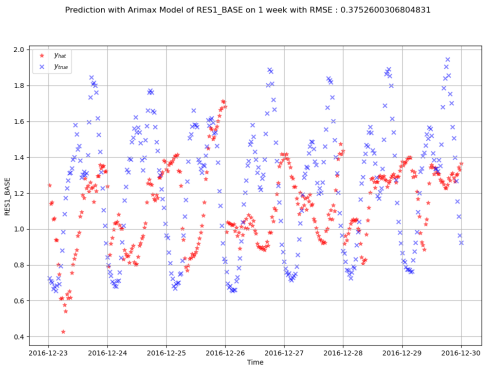
$$\Delta^d y_t = \sum_{i=1}^{p} \phi_i \Delta^d y_{t-1} + \sum_{j=1}^{q} \theta_{\epsilon_{t-j}} + \sum_{m=1}^{M} \beta_m X_{m,t} + \epsilon_t \tag{1}$$

with $\epsilon_t \sim \mathcal{N}(\mu, \sigma^2)$

The main results we got with this model are the following ones in figures 2a and 2b.



(a) Prediction of RES1_BASE on 2 last days of 2016.

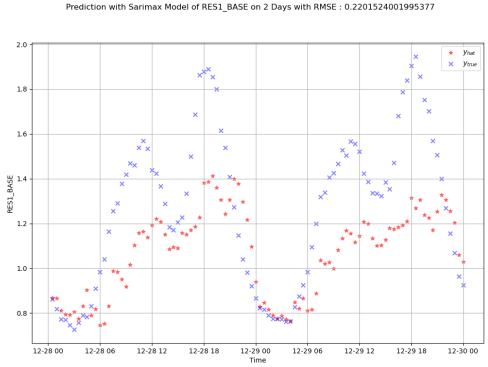


(b) Prediction of RES1_BASE on last week of 2016.

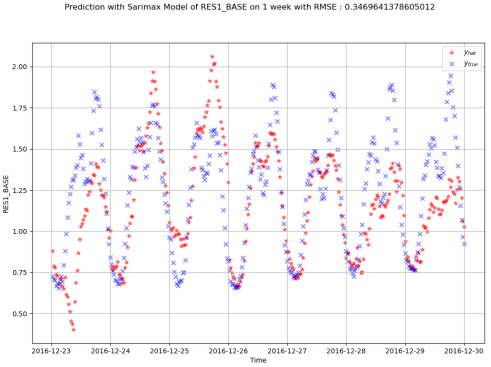
Figure 2: Prediction of $RES1\_BASE$ using ARIMAX model over different time periods.

Here we can see that the results follow a global "trend" but that they are very poor and do not fit the data perfectly.

2. SARIMAX: Seasonal Auto-Regressive Integrated Moving Average with exogenous variables is an extension of SARIMA models. The main difference with ARIMAX models is that this kind of model allows to deal with seasonality.This model adds (P, D, Q, s) parameters to ARIMAX models. (P, D, Q) correspond to the parameter of the AR, MA and the degree of differentiation of the seasonality and s corresponds to the number of points inside a seasonality. Because seasonality is important in our data we expect this kind of model to perform better. However, this model is extremely heavy and the computational cost is huge. The main results with a $SARIMAX\,(1, 0, 1)\,((1, 0, 1, 48)$ are the following on the figures 3a and 3b.



(a) Prediction on 2 last days of 2016.



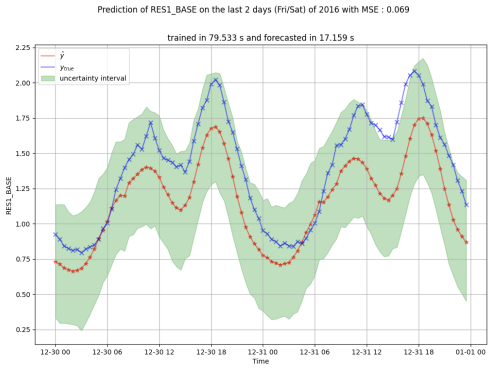
(b) Prediction on last week of 2016.

Figure 3: Prediction of $RES1\_BASE$ using SARIMAX model over different time periods.

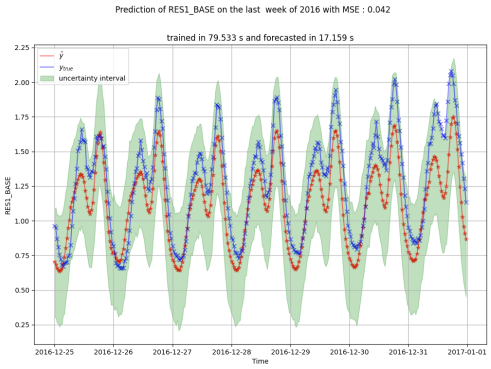
As expected, the SARIMAX model allows to make better forecasts. Indeed , we now see a good fit to the data on one week even if the result isn't perfect. On two days, the forecast is not that good and we are faced to the limitation of this model especially concerning computational time.

3. PROPHET: Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers

well. It decomposes the time series $y(t)$ into $y(t) = g(t) + s(t) + h(t) + \epsilon_t$ with $g(t)$ the trend aperiodic changes, $s(t)$ seasonality, $h(t)$ the holidays and ϵ_t residue. The results of the model are as follows on figures 4a and 4b.



(a) Prediction of RES1_BASE on 2 last days of 2016.



(b) Prediction of RES1_BASE on last week of 2016.

Figure 4: Prediction of $RES1\_BASE$ using PROPHET model over different time periods.

We can clearly see that prophet is doing well on its predictions, because it captures the day-night effect, but it still has trouble on the weekend effects. Therefore, with more time we could create a more important penalty on these times.

6 Conclusion

Since the problem was complicated, we have learned so much about modeling, time series features engineering and so on. The lecture was just an appetizer to go further on interesting point go get a better score with our model for this ENS challenge.

Glossary

AIC Akaike Information Criteria. 2

ARIMA Autoregressive integrated moving average. 3

BIC Bayesian Information Criteria. 2

FT Fourier transforms. 1

GAM Generalized Additive Model. 1

GLM Generalized Linear Model. 1

RMSE Root Mean Squared Error. 2