Statkraft assessment

Andrea Maioli

March 16, 2023

Contents

1 Introduction

2 Data formatting

I this section I will shortly describe how I've formatted and processed data starting from raw dataset.

The three main source of data are the datasets:

- **BPA**: daily timeseries of total system generation. It includes hydro generation, our target, but also other interesting values such as load, wind generation and fossil generation.

- **USACE Hydropower summaries**: monthly power summaries by hydro sites.
- **USGS Discharge**: daily timeseries of discahrge of each site of Columbia River projects

For simplicity I've decided to use only discharge timeseries of The Dallas Oregon.

Note that for **USGS and USACE** dataset has been implemented a fully automatised data scraper (`DataScraper` class) in `data_manager.py`.

Data have also been cleaned for example excluding all the samples where The Dellas discharge value is 0.

2.1 Scaling

The main issue of the hydro dataset was that it was not available the daily timeseries of the Columbia Project. It was needed a scaling of the Total System Generation (namely of the **total hydro** variable in BPA dataset).

To do that I've normalised the Total System hydro generation w.r.t. USACE monthly summary and multiplied the normalised timeseries by the USACE monthly summary values restricted on Columbia River perimeter.

Formally: let $M(d)$ be the month associated to valuedate d, S^C_m the power summary of Columbia River at month m, S^{Tot}_m the total power summary of month m, H^{Tot}_d the daily total power generation, H^C_d the daily Columbia River power generation. Then the following equation holds

$$H^C_d = H^{Tot}_d \cdot \frac{S^C_{M(d)}}{S^{Tot}_{M(d)}}$$

The timeseries H^C_d will be our target.

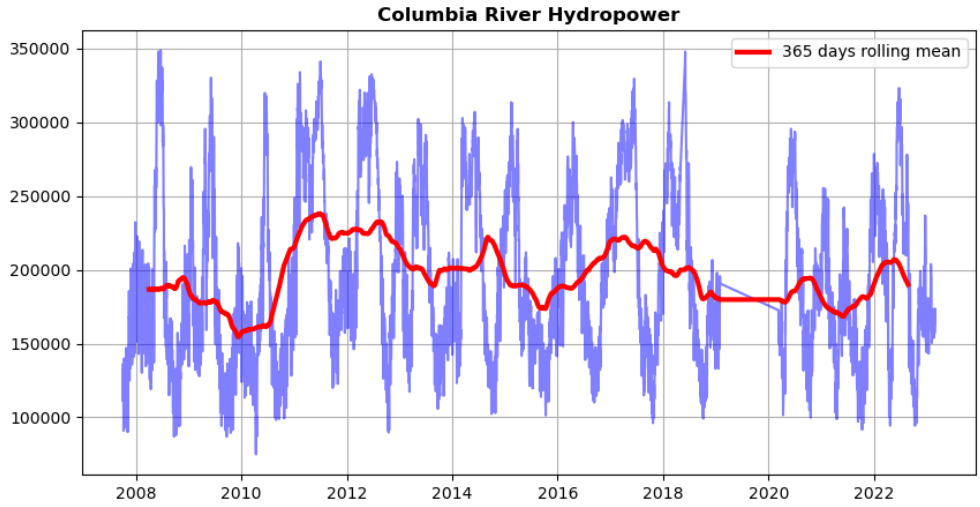
3 Data Analysis

3.1 Columbia generation properties

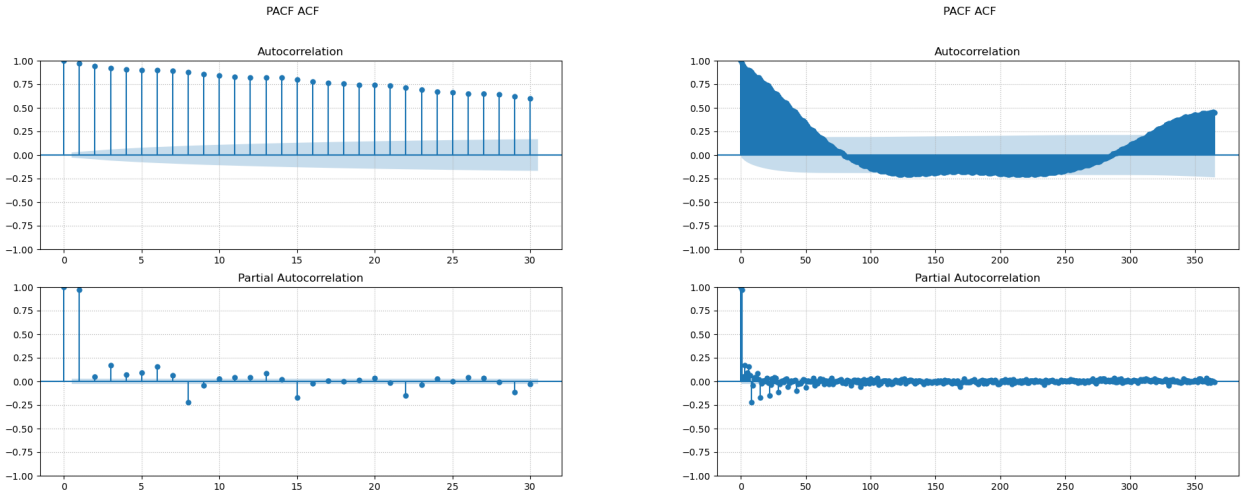
In this section I focus my attention to the properties of the target timeseries. I want to explore the following points:

- Trend
- Auto correlation
- Seasonality

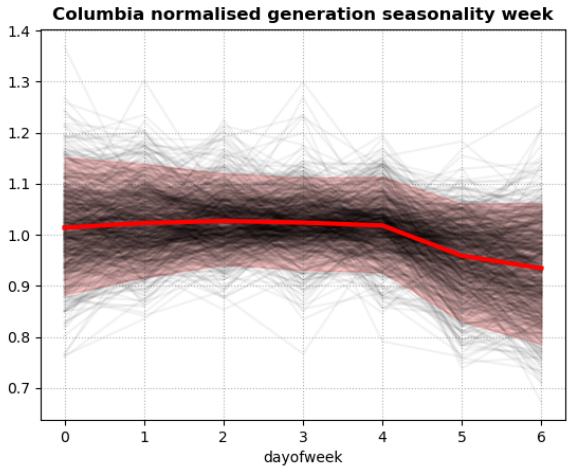
Target do not present global trend. For this reason I will not apply any detrending process.

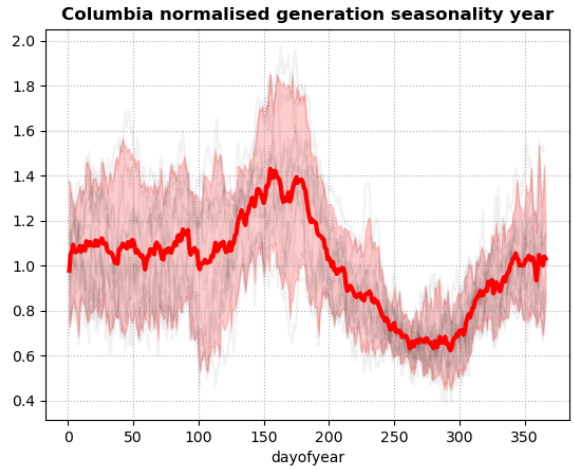


We study PACF/ACF plot for lags 30 and 365 since we want to estimate how important are lagged terms and how much the timeseries is autocorrelated.



I expect a yearly seasonality due to natural seasonality of meteo. If we visualize the normalized weeks we observe a weekly seasonality (this behaviour is confermed also by PACF value at lag 7, 14, ...).



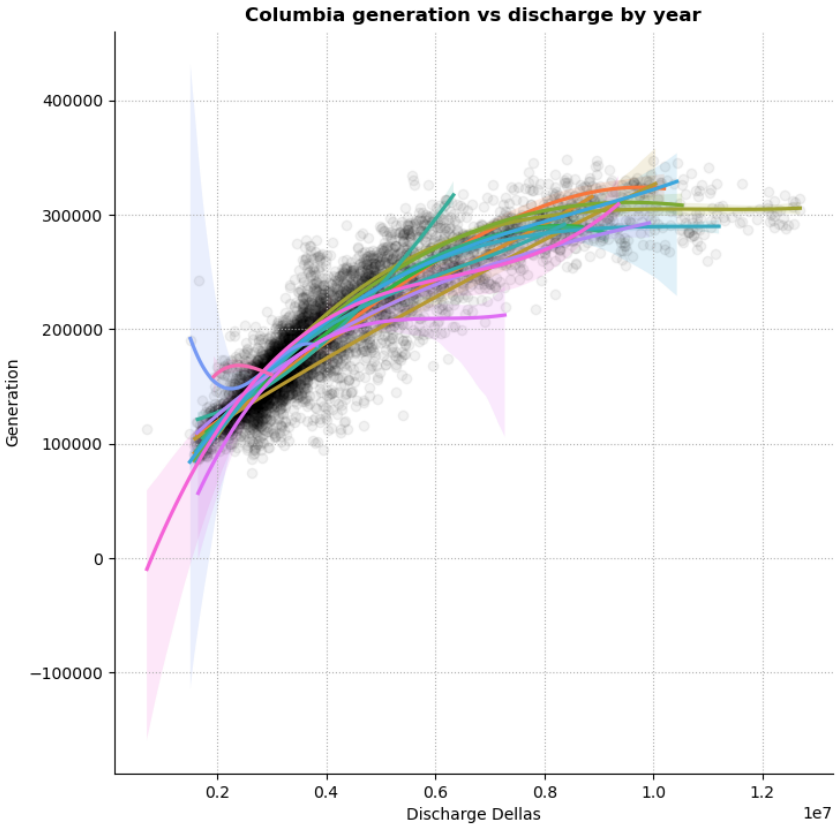


We conclude that calendar features should be included into the model since they explains important patterns. Operationally this means to use features `dayofyear` for yearly sasonality and `weekday` for weekly seasonality.

3.2 Features relations

3.2.1 Dellas discharge

The main explainer of this problem is of course the discharge. In the following plot I show the relation between Columbia generation and observed discharge at The Dellas.



In addition to the scatterplot you can see also lines of different colors. They are just the result of cubic spline fitted by year.

It is evident that the relation is not linear, in particular, after a certain level of discharge, the generation reach the saturation.

We deduce that the forecast model must be non linear.

From the plot I notice also that the yearly splines are not overlapping. This means that most problably I need also a feature to distinguish the year. This observation will be explained in details in the next section

3.2.2 Market relations

It is well known that power load, net interchange and wind, hydro, fossil power generation are strongly linked.

Since the network must be balanced the following equation must hold:

$$Load_d + Net_d = Wind_d + Solar_d + Fossil_d + Nuclear_d + Hydro_d$$

The previous equation is not the exact one. Infact not all the values depends directly on d but rather on the price at time d, $P(d)$. We have seen that in case of Hydro power generation also day of the week matter. This means that the variable is not fully meteo dependent but also prices dependent.

The previous equation can be written then as:

$$Load_d + Net_d = Wind_d + Solar_d + Fossil_{P(d)} + Nuclear_{P(d)} + Hydro_{d,P(d)}$$

This equation let us deduce one term once we have an estimate of all other terms.

Even if I will not exploit this relation, since I am not able to create a forecast for all the terms to deduce $Hydro_{d,P(d)}$, it is important to stress how usefull this relation can be also to forecast prices $P(d)$.

4 Forecast

Forecast generation will be divided into two part. In the first part I will create a static model, namely I will not use last observations of the target timeseries. In the second part I will try to take advantage of last observations to produce forecast for few steps ahead.

In this section I will use the following notation:

- G_t: Columbia generation, namely the target we have to forecast, at time t
- D_t: The Dellas discharge value at time t
- DoY_t, WD_t, Y_t: calendar feature at time t: DayofYear, WeekDay, Year.

I've decided to use the following data split:

- **Train set**: 2007 - 2017

- **Test set**: 2018

I've decided to exclude data starting from 2020 to avoid any effect of COVID and war. I want to focus on the phyisical transofrmation of discharge to power and avoid as much as possible exogenous effects.

4.1 Static models

I will shortly list here the models that I am going to use:

- Historical mean (benchmark)
- BDT
- GAM_{simple}
- GAM

4.1.1 Model definition

First of all I define a simple benchmark to beat, the $HistMean$ model, as :

$$HistMean_t = \frac{\sum_{\delta<t} G_{DoY(\delta)} \cdot \mathbb{1}_{\{DoY(\delta)=DoY(t)\}}}{|\{year < Y_t\}|}$$

It is nothing but the historical mean of G_t over same day of the year in the past.

From the previous analysis I already know that discharge feature is very important together with calendar. In this section I will do a strong assumption:

I will suppose that D_t is available at forecast time. This is possible if for example we suppose D_t itself to be a forecast of discharge for time t (see for example ECMWF GLOFAS model).

Under this assumption I define two GAM models, one with and one without calendar features. Using a R-like notation we can summarize the models as:

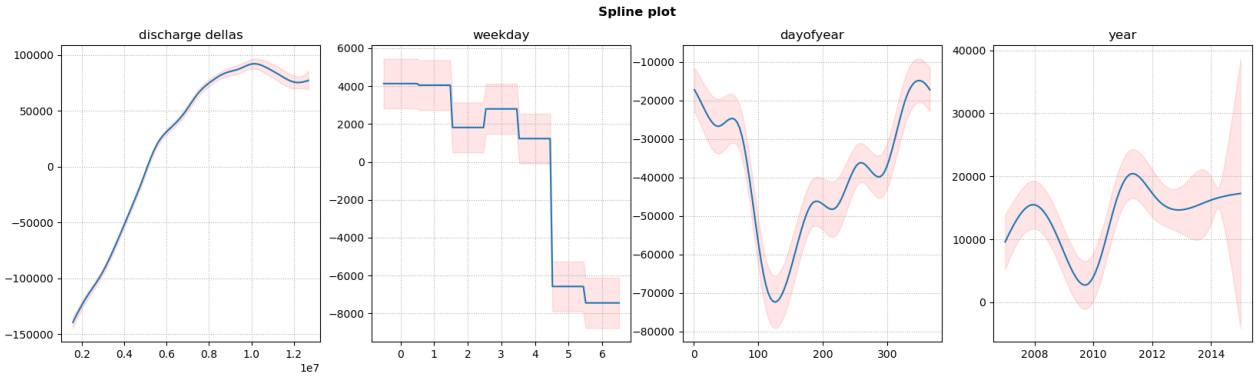
$$GAM_{simple} \sim s(D_t)$$

$$GAM \sim s(D_t) + f(WD_t) + s(DoY_t, basis = cp) + s(Y_t)$$

Just few notes:

- spline $s(DoY_t, basis = cp)$ is fitted using cyclic basis
- spline $f(WD_t)$ consist in a dummy like spline to manage categorical features
- Number of degree of splines for each term have be choosen in such a way is it greater than EDoF
- In both models I've adopted a Normal distribution assumption of the final output. Even if not formally correct (the output can not be negative) I've observed that transforming the target or using a Gamma distribution lead to worse results in terms of error.

The following plot show the splines of *GAM* fitted for each term toghether with a 95% confidence interval.



Before showing the results I want to summarize the reasons why I've decided to use GAM models:

- In the framework I've chosen (only one discharge feature) there is no need to manage big data and use model that can deal with them.
- These models are extremly interpretable and stables. We can even produce the output of each spline for each sample to have a fully explaination of the output it self.
- It is very easy, using the splines, to perform sensitivity analysis (in particular w.r.t. The Dellas discharge feature)
- By construction the model provide also a prediction interval in case one is interested in forecasting an interval rather than a deterministic value.

4.1.2 Results

The following table summarize the results:

Metric	HistMean	GAM simple	GAM	BDT
mae	28138	17784	12556	13154
mape [%]	13.4	8.59	6.6	6.4
bias	7083	4192	-4554	3797

Since GAM and BDT have comparable results I would use GAM as main model for the reasons I've listed above.

4.2 Autoregressive model

If this program is supposed to generate forecast every day it makes sense to take advantage of last observations. They are very usefull since aloud to calibrate predictions as soon as new target values are observed.

To do that I've decided to use simple linear model SARIMAX. Parameters of the models are obtained by gridsearch based on maximization of BIC/AIC values.

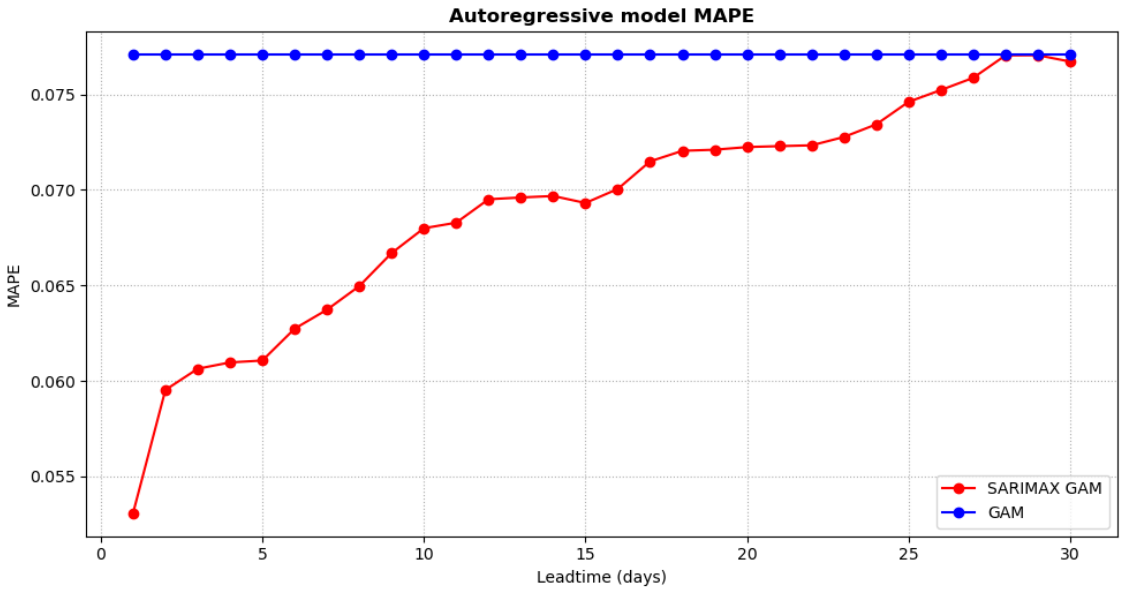
SARIMAX is applied to error of GAM model. The purpose is to split the physical component of the forecast, given by GAM model, from the autoregressive one.

$$
\begin{aligned}
G_t &= GAM_t + \epsilon_t \\
\hat{\epsilon}_t &= SARIMAX\left(\epsilon_{\delta<t}\right) \\
SARIMAX_t^{GAM} &= GAM_t + \hat{\epsilon}_t
\end{aligned}
$$

In production the forecast would be produced as follow:

1. Produce forecast GAM_t using prefitted model and D_t
2. Collect errors of GAM model in the past, calibrate SARIMAX on those observations.
3. Produce forecasst $\hat{\epsilon}_t$ using fitted SARIMAX
4. Define adjusted forecast as the sum: $GAM_t + \hat{\epsilon}_t$

This model should be used to generate forecast in the short term. From the performance we can see that, as leadtime increases, SARIMAX converges to baseline GAM.



5 Next steps

A lot of things can be done to improve the models of this assessment.

- As explained in the previous section all the Static models are unfair. The main regressor D_t is not available at time $t-1$. Namely is not possible to use that variable to produce values in the future. For this reason the first thing to do is to repeat the analysis (and eventually redesign the models) based on discharge **forecast** values.
- Discharge of different sites should be analysed to verify if they improve model performance.

- Since BDT and GAM have comparable performances, forecast can be improved using a **Stack blend** (for example another BDT) that use as regressors the output of the two models and mix them.
- Forecast hydro power generation isolating the corresponding term from the relation I've shown in **Market relations** section.