MEMORIA

SALE OFFERS PREDICTION IN THE SPANISH DAY-AHEAD ELECTRICITY MARKET

LUIS MUÑOZ DEL POZO

DATA SCIENCE MASTER, 26th edition (MAD) - KSCHOOL

July, 2021

1. OBJECTIVE

The main objective of this TFM is to find out if there is a way to predict with **ML techniques**, and only **public information** the sale offers that **combined cycle power plants** and **hydraulic plants** send to the Nominated Electricity Market Operator (OMIE) every day in the Spanish day-ahead electricity market.

If this prediction is feasible, it would be a great tool for the electrical companies to optimize their sale bids.

Excellent papers related to electricity demand forecast (per day, per technology, considering weather conditions, etc.) can be easily found, but the kind of study that is presented in this TFM is not frequent at all.

2. INTRODUCTION

In order to explain how this research work has been developed, it is important to know a bit how the **Spanish Electricity Market** works.

Like all the liberalized markets, the Spanish electricity market is based on supply and demand fundamentals, matching, for every hour, the sale and purchase electricity power offers between the agents that sell electricity and agents that buy it.

There are two main figures in this market: **System Operator (REE)** and **Market Operator** (**OMIE**). REE is in charge of managing the technical part of the system, and OMIE manages the economical part.

OMIE manages the auctions between sale and purchase offers and is in charge of selecting the units most economically suitable to generate the electricity.

There are two market **categories**: **long term markets** and **spot markets**. Long term markets are financial markets where energy price is arranged between two parties (not considered in this TFM), and **spot markets** are markets where the energy is sold and bought at the same time.

In the case of the Spanish electricity market, there are three types of spot markets:

- Bilateral contracts
- Daily markets:
    - **Day-ahead market**
    - Intraday markets

- Operator System markets

The most important market is the **Day-ahead Market**, where approximately 70% of the energy is negotiated. Obviously, this is the market this work is focused on.

The day-ahead market aims to carry out electrical energy transactions by submitting selling and purchase offers for electrical energy on behalf of the market agents for the twenty-four hours of the following day.

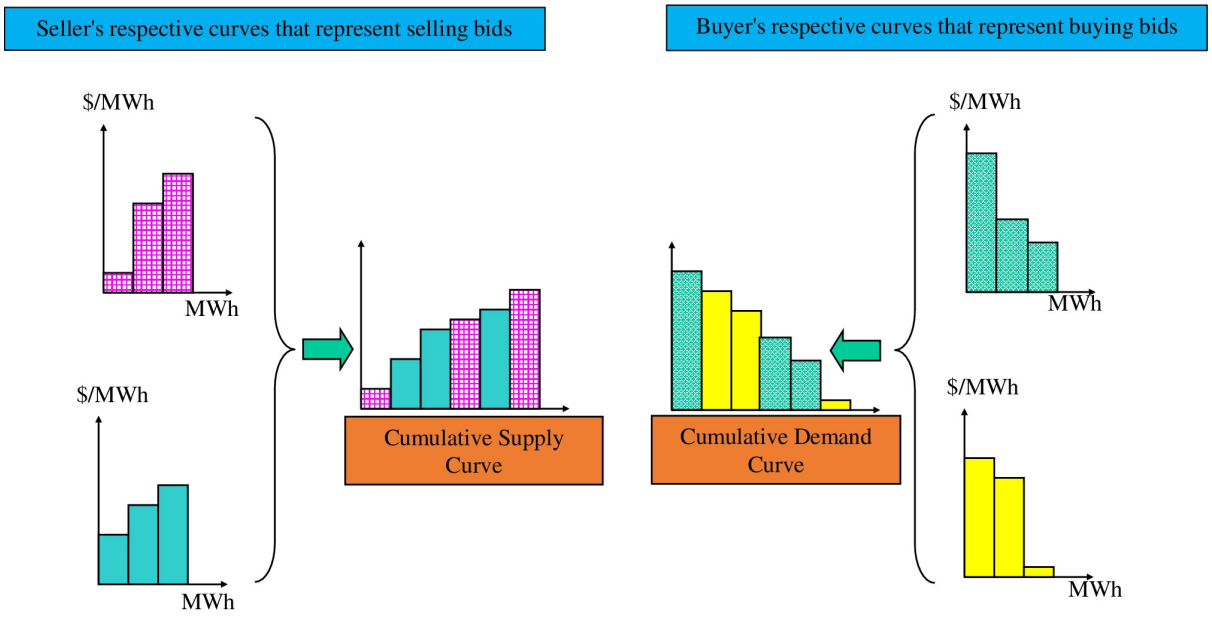
The buying and selling agents present their offers to the day-ahead market through OMIE. The mechanism described for setting electricity prices is called **market coupling**. The results from the day-ahead market are not the final ones, as it is necessary to take into account all the technical limitations related to the transportation network. As such, results from the day-ahead market may be altered slightly as a result of the analysis of technical limitations done by the System Operator (REE).

The market coupling sorts all the sale offers by ascending price order and all the purchase offers by descending price order. The final price of the energy is the one where both curves cross, and this price will be the same for all the energy traded in the auction, and for all the units. This last price is called **MCP** (**Market Clearing Price**) or **marginal price**.

Let us take a deeper look at this market coupling methodology. To do so, it is important to take into account the following:

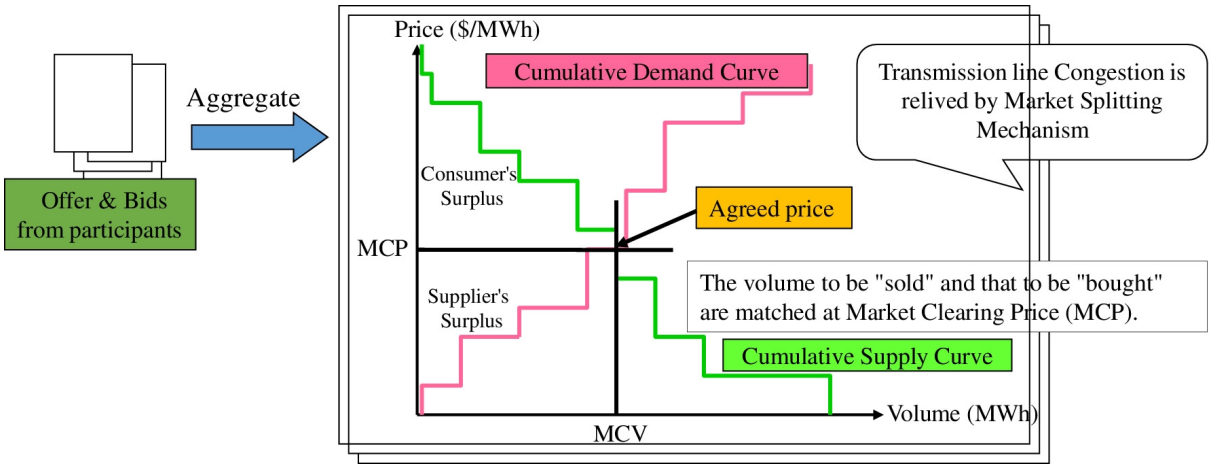
- All the hourly individual offers for all the agents are divided in blocks that must be sorted by price (ascending/descending for sale/purchase offers).
- When sorting individual bids in the aggregate process, the different unit blocks will be disordered.

The following picture represens how hourly sale and purchase offers are presented in the auction, and the way they are grouped and sorted in the aggregate process previous to the market coupling.



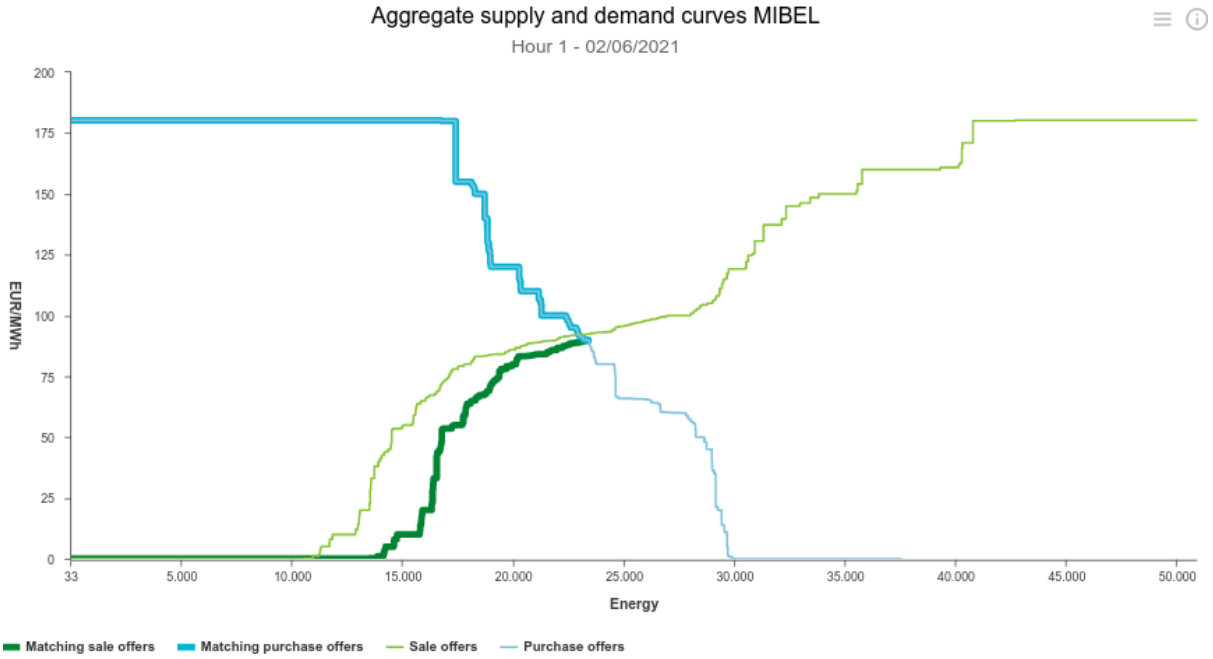
Offer and bid submission of sellers and buyers (from https://onlinelibrary.wiley.com/doi/full/10.1002/2050-7038.12360)

Considering the **cumulative sale offer curve** and **purchase offer curve**, the method to determine the **MCP** is shown in the picture below.



Market clearing, consumer's surplus, and supplier's surplus (from https://onlinelibrary.wiley.com/doi/full/10.1002/2050-7038.12360)

In the following picture from OMIE a real coupling process is shown.



More information about the operation of the day-ahead market can be found in this link: https://www.omie.es/sites/default/files/inline-files/day_ahead_market.pdf

Economic theory states that this kind of market and this kind of market coupling method is the most economically efficient due to the fact that the different sale agents (units) offer depends on their opportunity cost.

The opportunity cost is very low for **renewable units** (wind, solar, biomass, etc.), and **nuclear units**, so most of them sell their energy at 0 €/MWh. For that reason, they are not considered in the scope of this TFM, since the bid prediction for most of them is obvious (always the maximum unit power per hour at 0 €/MWh)

The next cost step would be **coal fired units** (almost negligible nowadays, since most of them have been dismantled recently), and **combined cycle units** (burning natural gas). In combined cycle power plants, the opportunity cost should be, normally, very similar to the operation cost of the unit, that mainly is the **natural gas cost**, and if the price is high, the **CO_2 price** is too.

Finally, in the highest cost step, we can find the **hydraulic units** that sell energy at a high price, since they can store water and wait till the price is high enough to maximize their profits.

For the reasons explained above, the technologies studied in this TFM are **combined cycle** and **hydraulic**.

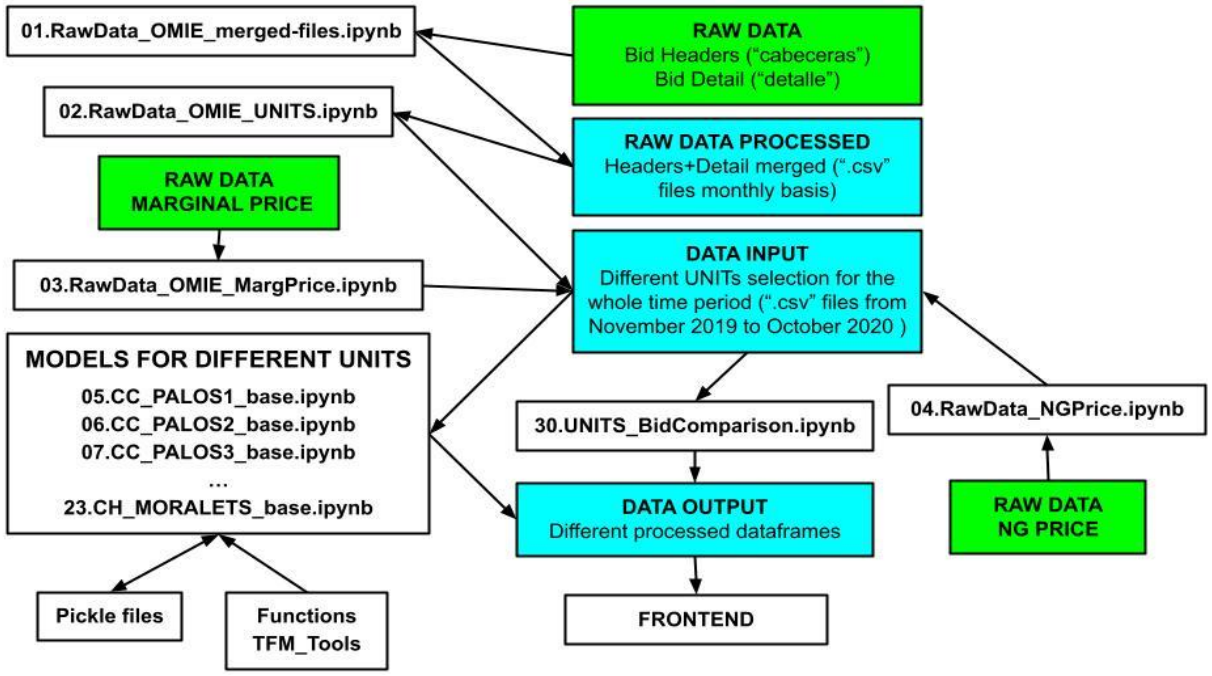
3. RAW DATA DESCRIPTION

The raw data used in this TFM are mainly the day-ahead market sale offer information, and the external information that was considered important to fulfill the purpose of the offer prediction, such as electricity day-ahead market hourly prices (MCP), and natural gas prices.

The main raw data mainly used in this TFM comes from **OMIE** public files (https://www.omie.es/en/file-access-list), specifically the sale offer information.

The **time period** selected for this TFM is 12 months, from 2019-11-01 to 2020-10-31.

A scheme of the raw data and the notebooks where they are processed and used is shown in the figure below.



In the following sections, a detailed description of the raw data and a complete explanation about the data retrieving process is presented.

3.1. DAY-AHEAD MARKET SALE OFFERS

The day-ahead market bid information (for all units and all hours) is published by OMIE after a confidentiality period of 90 days. The information is divided into two daily text files (header and detail). The format of the files is ".1", meaning that it is a tailor made format.

NOTE: An important fact is that there is no information in the OMIE website about the meaning of the fields or the way these files are organized or sorted, so it is not easy to figure out the meaning of all of the data provided in them.

The information provided in the OMIE bid set of files is the summarized below:

- **Header of bids for Day-ahead Market**. In these files, information related to the header of the bids is provided. The information provided is divided in the following features:

    - **Bid Code**: Code generated every day for each agent.
    - **Number of Version**: Number of the bid version.
    - **Year**: Year of the bid.
    - **Month**: Month of the bid.
    - **Day**: Day of the bid.
    - **Bid Unit**: Code of the unit/agent.
    - **Unit Description**: Brief description of the unit/agent (name, and technology type).
    - **Sell/Buy Bid**: Type of bid.
    - **Max Power**: Maximum Power of the unit/agent in MW.
    - More information not used in this TFM like Maximum power increasing, Maximum start-up power, Maximum shut-off power, Interconexion code, etc.

    So, the information in these files just match the "Bid_Code" information with the date, type of bid (sell or buy), and the unit code and description.

    There are approximately 1500 bid agents in the Spanish market, so these files have this number of rows.

    An example of the information provided (after being processed) is presented below.

Bid_Code	Num_Version	Bid_Unit	Unit_Description	Sell_Buy	Pot_max	Year	Month	Day
1696149	6	EDPC2	EDP COMERCIAL COMPRA (PORT)	CNO	6000.0	2020	09	01
1717319	3	EONUC01	EONUR CONSUMO CLIENTES TUR	CNO	400.0	2020	09	01
1811311	7	IPG	C.H. IP GENERACION	VNO	84.0	2020	09	01
426609	12	IPB	C.H.B.IP BOMBEO	CNO	99.0	2020	09	01
2532852	28	NRENVD1	NRENO-VENTA	VNO	1.7	2020	09	01

Example of Header of bids for Day-ahead Market ("Cabeceras" files)

- **Day-ahead Market bids detail**. In these files, all the information related to the bids is provided. The information is divided as follows:

    - **Bid Code**: Code generated every day for each unit/agent.
    - **Number of Version**: Number of the bid version.
    - **Year**: Year of the bid.
    - **Month**: Month of the bid.
    - **Day**: Day of the bid.
    - **Period**: Hour of the bid.
    - **Block**: Number of divisions of the bid for each date and hour.
    - **Energy**: Electrical energy in MWh offered for the corresponding block, hour, and date.
    - **Price**: Price per energy in €/MWh offered for the corresponding block, hour, and date.

    Each daily file contains approximately 300000 rows, since there are 1500 agents, 24 hours and a number of blocks per bid that vary from 1 to 15.

    An example of the information provided by OMIE (after being processed) is presented below.

Bid_Code	Num_Version	Period	Block	Price	Energy	Year	Month	Day
1696149	6	22	1	0.010	0.1	2020	09	01
1717319	3	1	1	0.000	1.0	2020	09	01
1717319	3	2	1	0.000	1.0	2020	09	01
1717319	3	3	1	0.000	1.0	2020	09	01
1717319	3	4	1	0.000	1.0	2020	09	01

Example of Detail of bids for Day-ahead Market ("Detalles" files)

In order to understand and retrieve the information from these OMIE text files, a notebook called **01_RawData_OMIE_merged_files.ipynb** is created.

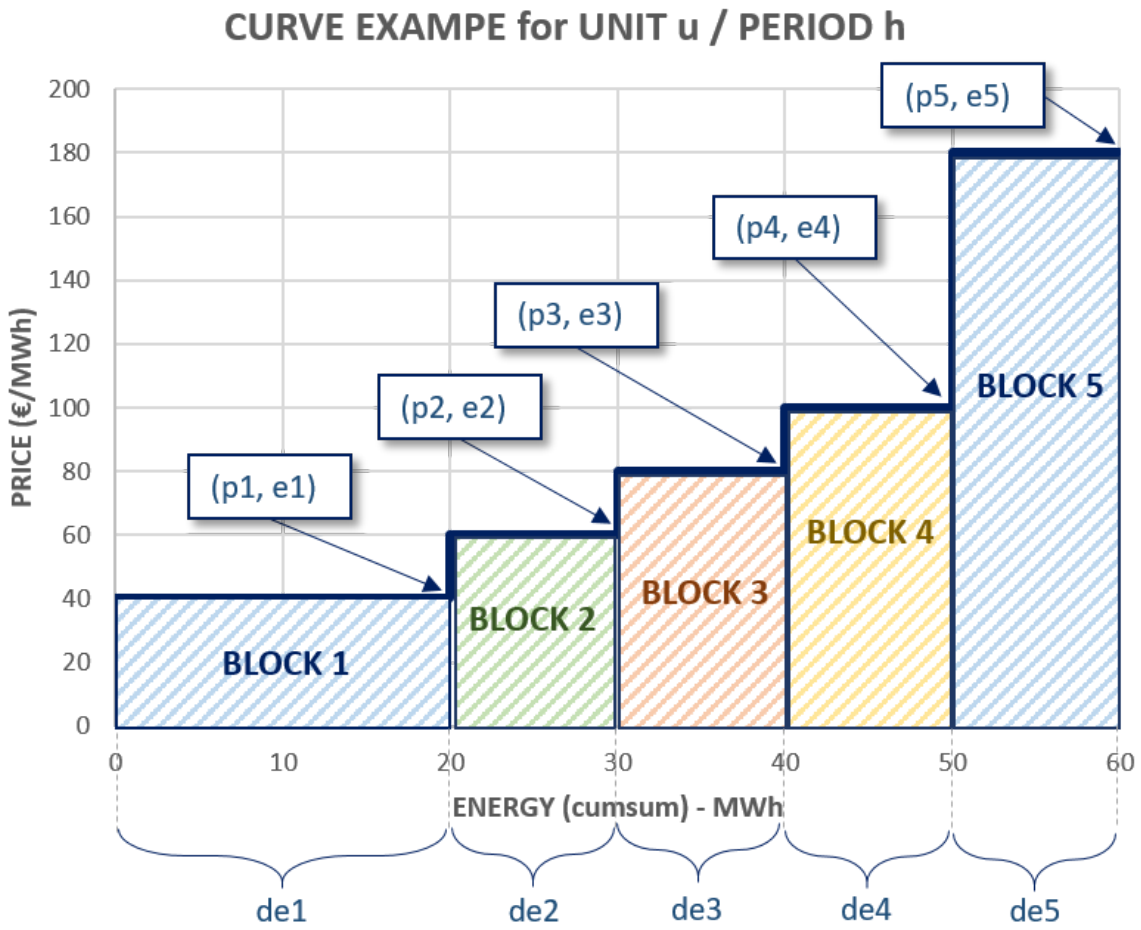
In this notebook, the OMIE text files are read (after downloading locally the information from the OMIE webpage) and the format is studied in order to store the information as a proper dataframe. The output of this notebook are several ".csv" files. In the case of this TFM, 12 files are created, one per month, as they are quite big to manage with a normal laptop on an yearly basis.

Those new ".csv" files are stored locally in order to use them in another notebook (**02_RawData_OMIE_UNITS.ipynb**) where they are filtered by a selected unit and concatenated to have a complete time set of data (12 months in this case). The output of this notebook are several datasets (locally stored as ".csv" files) with the bids for a chosen set of units for the selected time period. These new files will be the main input of the notebooks where the bids are studied and models are built.

An example of a sale offer for PALOS1 dataframe (stored locally as OMIE_PALOS1_112019_102020.csv is shown in the following table and picture.

	Bid_Code	Num_Version	Bid_Unit	Unit_Description	Sell_Buy	Pot_max	Year	Month	Day	Period	Block	Price	Energy
0	6128191	2	PALOS1	C.C. PALOS 1	VNO	394.1	2020	1	1	1	12	180.30	394.1
1	6128191	2	PALOS1	C.C. PALOS 1	VNO	394.1	2020	1	1	2	12	180.30	394.1
2	6128191	2	PALOS1	C.C. PALOS 1	VNO	394.1	2020	1	1	3	1	1.13	50.0
3	6128191	2	PALOS1	C.C. PALOS 1	VNO	394.1	2020	1	1	3	12	180.30	344.1
4	6128191	2	PALOS1	C.C. PALOS 1	VNO	394.1	2020	1	1	4	1	1.13	60.0

An example of a generic sale offer for a unit “u”, for a date “d”, and an hour (period) “h” stored in the “.csv” files as the one indicated above can be seen in the following picture.



It is important to point out that “**Price**” and “**Energy**” features in the OMIE bid files correspond to p1, p2, ..., **p5** (bid prices) and de1, de2, …, **de5** (bid delta energy) in the figure above. Points e1, e2, …, e5 must be calculated as the cumulative sum of the “Energy” field to plot the curve with the raw data from OMIE.

3.2. DAY-AHEAD MARKET HOURLY PRICES

The information of the day-ahead market hourly prices (MCP or **marginal price**) was considered useful to build the **ML models**, so it was retrieved from OMIE: Day-ahead market hourly prices in Spain

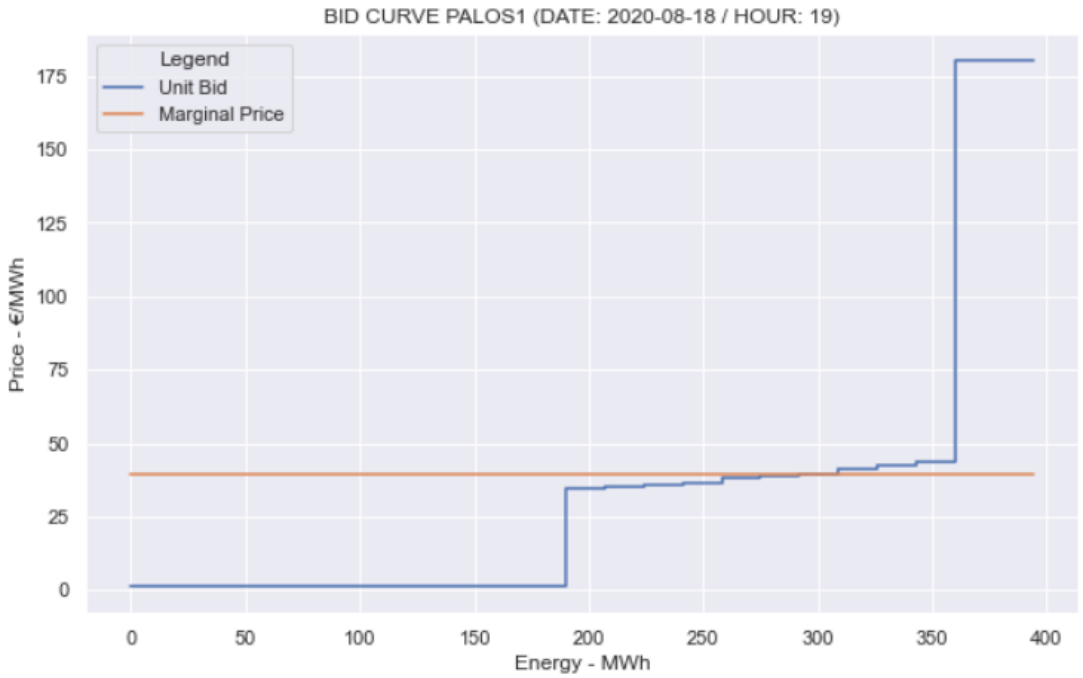
The way this information was read and processed is similar to the process done with day-ahead market bids explained in section 3.1: the retrieving and processing is done in a notebook called **03_RawData_OMIE_MargPrice.ipynb**, where a dataframe with the information needed is created and stored locally in a ".csv" file. This file will be used in the notebooks where ML models are built.

An example of the marginal price (in €/MWh) information provided by OMIE is presented below (after been processed):

Year	Month	Day	Period	Marg_Price
2019	1	1	1	66.88
2019	1	1	2	66.88
2019	1	1	3	66.00
2019	1	1	4	63.64
2019	1	1	5	58.85

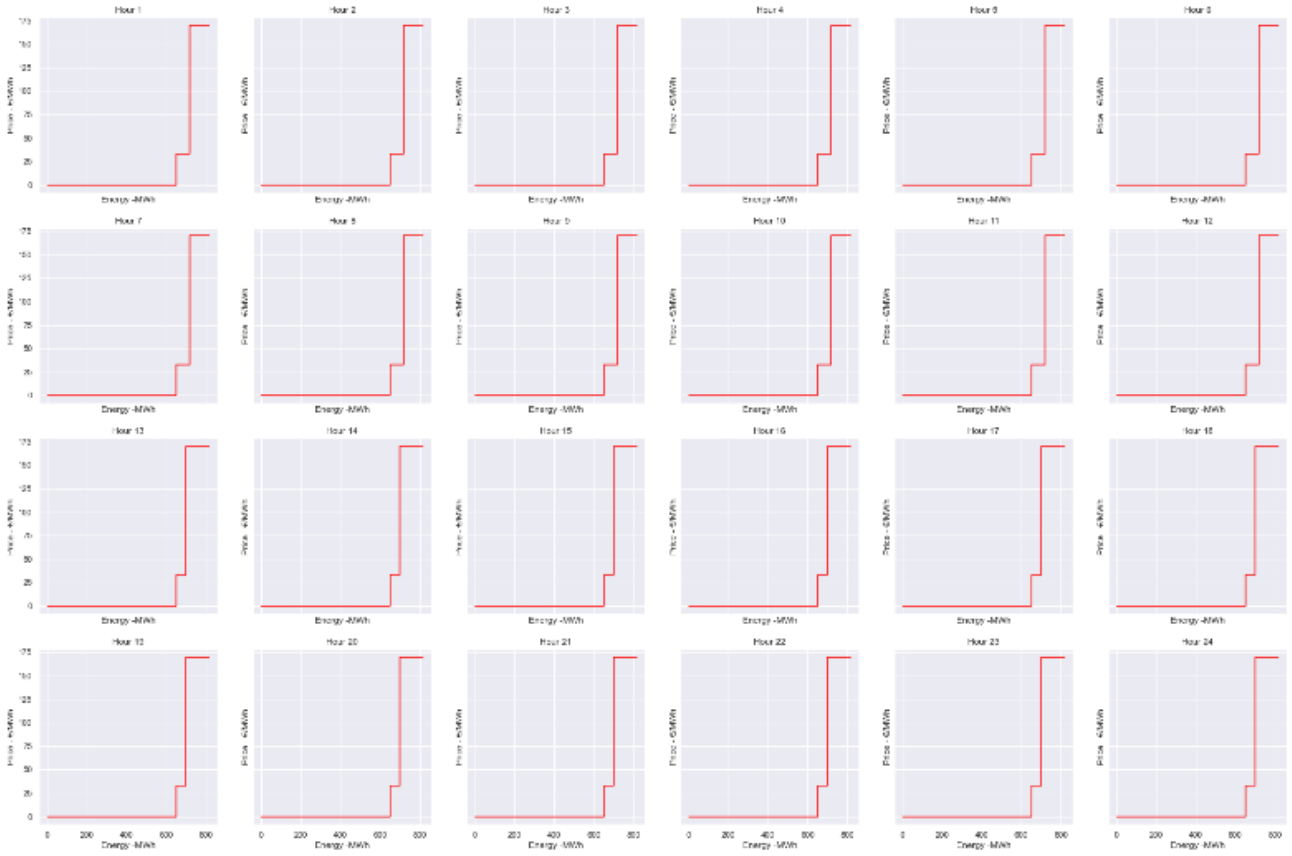
Example of Day-ahead market hourly prices.

In the following picture, an example of a combined cycle unit bid curve of C.C. PALOS 1 can be shown with the marginal price coupled for that date and period. In this example, this unit sold around 310MWh (out of 394.1MWh) at a price of approx. 40 €/MWh.



Bid curve example for C.C. PALOS 1.

In the following picture, an example of a combined cycle unit 24 bid curves of C.C. ESCARÓN 3 can be shown.



Daily 24 bid curves example for C.C. ESCATRÓN 3.

3.3. NATURAL GAS PRICES

The natural gas prices have been taken from the following source as a ".csv" file: https://www.macrotrends.net/2478/natural-gas-prices-historical-chart. **"Value"** field is the natural price in US$.

An example of the information provided is presented below.

date	value
2008-04-14	10.03
2008-04-15	10.16
2008-04-16	10.11
2008-04-17	10.27
2008-04-18	10.08

Example of natural gas prices.

The information is processed in a notebook called **04_RawData_NGPrice.ipynb**.

4. METHODOLOGY

The main decisions regarding the model selection can be summarized as follows:

- This TFM has been developed in **PYTHON version 3.8.5** and in a **UNIX** environment (some notebooks use UNIX commands to retrieve local files information).

- The problem has been considered as a **multi-output regression** problem, meaning that the **energy** and **price** are predicted at the same time (both are the targets of the problem). Regarding the energy feature, two different approaches have been checked:

    - Using **block energy** or delta energy (the value that is considered in the bids, as it is shown in section 3.1.) as the energy feature.

    - Using **total energy** or cumulative energy in the bids as the energy feature.

    Considering that the results of these two approaches are very similar, the **block energy** approach was chosen as it is the best to evaluate the **block-area metric** created to compare the quality of the models (explained below).

- Bid raw data have been modified to have a complete and more structured data set. The main modifications are:

    - **Transforming all days to a 24-hour day**. During a year, one day is a 23-hour day and another a 25-hour day, so these days must be transformed to be able to shift the data properly to consider time lagged features.

    - **Completing all bids with the maximum number of blocks**. Note that not all the bids for a same unit have the same number of blocks, so bids are transformed to complete the bids to consider the maximum number of blocks. Depending on the unit, the way to perform the transformation is different. Two different types of bidding have been detected, so two types of transformation have been developed:

        - **Type A**: Bids where the number of the block that is in all the bids is equal to the number of the maximum number of blocks for all bids.

        - **Type B**: Bids where the number of the block that is in all the bids is equal to 1.

- The following **features** have been selected as models´ input:

    - Weekday
    - Period (hour)
    - Block
    - Energy lagged 24h
    - Energy lagged 168h
    - Price lagged 24h
    - Price lagged 168h
    - Marginal price lagged 24h
    - Natural gas price lagged 24h (only for checking purposes in C.C. PALOS 1)

    Note that CO_2 price has not been considered as model input since the trend through the time period considered was very stable and, as ML models did not improve including natural gas price as an external feature, it was inferred that CO_2 price was not going to significantly improve the results.

- **ARIMA models** have not been considered since they cannot be considered as a time series problem.

- After trying some linear regression and k-neighbours models, without any success, only decision-trees based models have been considered: **Random Forest** and **XG-Boost**.

    Note that Random Forest can handle multi-output regressions, but XG-Boost cannot, so two **wrapper methods** have been checked: **direct** and **chained**. The results with both methods are very similar, but chained is chosen since it seems to be more suitable for this purpose.

- Since the first results are not good, instead of using a single model with **Block** and **Period** as inputs, two different approaches are considered:

    - **ML models for every Period** (24 models with **Block** as feature)
    - **ML models for every Block** (number of models equal to the maximum number of blocks with **Period** as feature)

    Results considering both approaches are similar.

- An **area based metric** is created to compare model predictions. Instead of comparing MAE or RMSE for Energy and Price (it does not seem to be the best metric), MAE and RMSE for the block areas (Energy times Price) is calculated.

The methodology followed in this TFM to find out if it is possible to predict the sale bids can be summarized as follows:

- The **Combined Cycle PALOS 1** is chosen as an example to find out how the bids are sent by the sale agents, and to create the models to predict the bids. The searching process is:

    - **Retrieving**, **exploring** and **plotting** the information from the ".csv" file created in previous notebooks as explained in section 3.

    - **Data wrangling**, where dataframes are created to be used as inputs in the different ML models, considering splitting in **train**, **validation** and **test sets**.

    - **Different ML models** and different approaches (depending on the number of models, feature time lagging, and features selection) have been checked, including adding new features to improve the behavior of the models.

    - A **naive model** is defined to check if ML models improve it. The naive model considered to predict a bid is just using the bid of the previous day (24h lagged bid).

    - A **new metric** has been created in order to check properly the quality of the models.

    - **Result summary** and **model comparison**. In this stage the best approach is chosen to be used with other units to check that the same conclusions are obtained.

  During this process, a notebook (**10_CC_PALOS1_base.ipynb**) has been created considering all different model approaches explained above. As the running time of the models can be long (approx. 10-15 minutes) it was decided to store them in ".pkl" files (**.\Pickle_Models**).

  Tool functions called **TFM_PredCurve_Tools.py** have been created to be used in the main unit notebooks.

  Dataframes with the information of the bids and the model results are stored locally (**.\Data_Output**) in ".csv" files. These files are used in other notebooks and in the frontend.

- Once **C.C. PALOS 1** results are obtained and studied, the same process is done for different Combined Cycle and Hydraulic units in order to find out if the same conclusions are reached. By doing so, different ways to send bids have been discovered (mostly depending on the unit company), so the functions created to prepare the data, etc. are improved in order for them to be used for all types of units. Since 12 units have been studied (besides C.C. PALOS 1), 12 notebooks have been created (**11_CC_PALOS2_base.ipynb**, **12_CC_PALOS3_base.ipynb**, etc.)

Dataframes with the information of the bids and the model results are stored locally (**.\Data_Output**) in ".csv" files. These files are used in other notebooks and in the frontend.

- Finally, a notebook to compare and plot all of the units is created (**30_UNIT_Bid_Comparison.ipynb**).

- The information from the bid curves and some comparisons can be seen in the **FRONTEND** developed with **STREAMLIT** in a notebook called **40_FRONTEND.ipynb**, where 5 .py files are created. The FRONTEND user manual can be found in ANNEX 1 of this document.

5. SUMMARY OF MAIN RESULTS

In this section a summary of the main results are presented for all of the studied units. Further details can be seen in the corresponding notebook.

5.1. C.C. PALOS 1

As mentioned above, PALOS 1 has been the “guinea pig” for this study so, different types of models have been checked after considering the final one.

In the following table, the results for the base case for the different models can be seen:

	MAE_Energy	MAE_Price	MAE_E+P	MAE_Area	RMSE_Energy	RMSE_Price	RMSE_E+P	RMSE_Area
Naive	0.16	0.52	0.34	19.16	3.42	7.07	5.56	519.64
RF_basic	0.15	0.88	0.51	24.03	3.29	7.17	5.58	511.75
RF_grid	0.15	0.59	0.37	20.14	3.27	7.00	5.47	508.38
XGB_dir_basic	0.19	0.67	0.43	25.96	3.43	7.05	5.55	529.99
XGB_chain_basic	0.19	0.65	0.42	25.31	3.43	7.11	5.58	526.59
XGB_chain_grid	0.15	0.67	0.41	21.32	3.29	7.10	5.53	511.68

Comparison ML model results (models divided by Period)

Considering the table above, **RANDOM FOREST** considering **grid search** optimization for hyperparameters is the best ML model. The bad news is that this model is not much better than the naive model.

In the following table, models divided by Period are compared with models divided by Blocks:

	MAE_Energy	MAE_Price	MAE_E+P	MAE_Area	RMSE_Energy	RMSE_Price	RMSE_E+P	RMSE_Area
Naive	0.16	0.52	0.34	19.16	3.42	7.07	5.56	519.64
RF_grid	0.15	0.59	0.37	20.14	3.27	7.00	5.47	508.38
RF_grid_BLOCK	0.18	0.65	0.42	24.09	3.46	7.00	5.52	532.44

Comparison RANDOM FOREST model results (models divided by Period vs. Block)

It can be seen that the results are very similar.

In the following table, models divided by Period are compared with a single model considering both Period and Block input features:

	MAE_Energy	MAE_Price	MAE_E+P	MAE_Area	RMSE_Energy	RMSE_Price	RMSE_E+P	RMSE_Area
Naive	0.16	0.52	0.34	19.16	3.42	7.07	5.56	519.64
RF_grid	0.15	0.59	0.37	20.14	3.27	7.00	5.47	508.38
RF_grid_TOTAL	0.18	0.64	0.41	23.48	3.28	7.05	5.50	509.06

Comparison RANDOM FOREST model results (models divided by Period vs. single model)

Finally, in the following table, natural gas price (lagged 24 hours) has been included into the models, and the results have been compared with the base approach:

	MAE_Energy	MAE_Price	MAE_E+P	MAE_Area	RMSE_Energy	RMSE_Price	RMSE_E+P	RMSE_Area
Naive	0.16	0.52	0.34	19.16	3.42	7.07	5.56	519.64
RF_grid	0.15	0.59	0.37	20.14	3.27	7.00	5.47	508.38
RF_grid_NG	0.18	0.64	0.41	23.48	3.28	7.05	5.50	509.06

Comparison RANDOM FOREST model results (with and without natural gas price)

It can be seen that the new feature does not improve the results, so it was decided not to include it in the base models.

The main conclusion is that naive and ML models considered are very similar, giving reasonably good results, with MAPE for ‘Area’ prediction approx. 0.8% for both models.

5.2. C.C. PALOS 2

In the following table, the results for the base case (RANDOM FOREST models divided by Period without natural gas price) are shown:

	MAE_Energy	MAE_Price	MAE_E+P	MAE_Area	RMSE_Energy	RMSE_Price	RMSE_E+P	RMSE_Area
Naive	5.48	12.52	9.00	554.28	30.24	42.71	37.00	4,888.87
RF_grid	6.15	13.28	9.72	669.21	27.35	40.40	34.50	4,328.35

The main conclusion is that both models are very similar, so it cannot be said that the ML model improves the naive model prediction results.

The prediction for both models are very poor: MAPE for ‘Area’ prediction higher than 20% for both models.

5.3. C.C. PALOS 3

In the following table, the results for the base case (RANDOM FOREST models divided by Period without natural gas price) are shown:

	MAE_Energy	MAE_Price	MAE_E+P	MAE_Area	RMSE_Energy	RMSE_Price	RMSE_E+P	RMSE_Area
Naive	0.16	0.52	0.34	19.05	3.37	7.11	5.56	510.75
RF_grid	0.14	0.58	0.36	18.50	2.97	7.05	5.41	456.68

The main conclusion is that naive and ML models considered are very similar, giving good results, with MAPE for ‘Area’ prediction lower than 0.8% for both models.

5.4. C.C. SAGUNTO 1

In the following table, the results for the base case (RANDOM FOREST models divided by Period without natural gas price) are shown:

	MAE_Energy	MAE_Price	MAE_E+P	MAE_Area	RMSE_Energy	RMSE_Price	RMSE_E+P	RMSE_Area
Naive	0.63	1.54	1.09	67.01	7.03	14.09	11.13	1,124.54
RF_grid	1.23	2.46	1.84	146.07	6.22	13.25	10.35	960.18

The main conclusion is that both models are very similar: naive model is better considering MAE metric and the other has better predictions considering RMSE metric, so it cannot be said that the ML model improves the naive model prediction results.

Both models do not give bad predictions: MAPE for ‘Area’ prediction around 5% RANDOM FOREST, and 2.3% naive model.

5.5. C.C. ESCATRÓN 3

In the following table, the results for the base case (RANDOM FOREST models divided by Period without natural gas price) are shown:

	MAE_Energy	MAE_Price	MAE_E+P	MAE_Area	RMSE_Energy	RMSE_Price	RMSE_E+P	RMSE_Area
Naive	91.48	14.66	53.07	5,378.67	190.97	33.88	137.14	12,517.04
RF_grid	103.88	19.65	61.77	7,387.64	164.93	32.25	118.83	12,016.31

The main conclusion is that both models are very similar, so it cannot be said that the ML model improves the naive model prediction results. Both models give very bad results, with MAPE for ‘Area’ prediction higher than 50% and 80%, (naive model and RANDOM FOREST respectively).

5.6. C.C. ALGECIRAS 3

In the following table, the results for the base case (RANDOM FOREST models divided by Period without natural gas price) are shown:

	MAE_Energy	MAE_Price	MAE_E+P	MAE_Area	RMSE_Energy	RMSE_Price	RMSE_E+P	RMSE_Area
Naive	78.64	10.49	44.57	5,164.95	181.64	25.52	129.70	12,529.77
RF_grid	100.89	17.06	58.97	7,859.69	162.22	23.18	115.87	12,382.91

The main conclusion is that both models are very similar: naive model is better considering MAE metric and the other has better predictions considering RMSE metric, so it cannot be said that the ML model improves the naive model prediction results.

Both models give very bad results, with MAPE for ‘Area’ prediction higher than 50% and 80%, (naive model and RANDOM FOREST respectively).

5.7. C.C. ARCOS 1

In the following table, the results for the base case (RANDOM FOREST models divided by Period without natural gas price) are shown:

	MAE_Energy	MAE_Price	MAE_E+P	MAE_Area	RMSE_Energy	RMSE_Price	RMSE_E+P	RMSE_Area
Naive	8.62	8.48	8.55	1,365.16	32.57	25.59	29.29	6,954.24
RF_grid	10.91	9.46	10.19	1,592.96	30.29	23.41	27.07	6,639.93

The main conclusion is that both models are very similar, so it cannot be said that the ML model improves the naive model prediction results. The results for both models are very poor: MAPE for ‘Area’ prediction of 26% and 31% (naive model and RANDOM FOREST respectively).

5.8. C.C. COLÓN 4

In the following table, the results for the base case (RANDOM FOREST models divided by Period without natural gas price) are shown:

	MAE_Energy	MAE_Price	MAE_E+P	MAE_Area	RMSE_Energy	RMSE_Price	RMSE_E+P	RMSE_Area
Naive	4.32	1.21	2.77	740.91	30.22	12.66	23.17	5,511.01
RF_grid	11.64	4.69	8.16	2,083.09	30.65	13.28	23.62	5,652.65

In this case, naive model results are better than the RANDOM FOREST considering MAE and RMSE metric. MAPE for ‘Area’ prediction for naive model is 7.4%, while approx. 21% for RANDOM FOREST model.

5.9. C.C. CASTELNOU

In the following table, the results for the base case (RANDOM FOREST models divided by Period without natural gas price) are shown:

	MAE_Energy	MAE_Price	MAE_E+P	MAE_Area	RMSE_Energy	RMSE_Price	RMSE_E+P	RMSE_Area
Naive	0.00	0.05	0.02	8.64	0.00	0.97	0.69	175.13
RF_grid	2.90	1.00	1.95	241.91	12.14	2.29	8.74	897.94

In this case, naive model results are better than the RANDOM FOREST considering MAE and RMSE metric. Particularly, MAPE for 'Area' prediction for naive model is 0.4% (a really good prediction!), and 11.4% for RANDOM FOREST model. The ML model works bad for this unit because the plant was unavailable more that 60% of the time period, so the model could not be fitted with enough data.

5.10. C.H. AGUAYO GENERACIÓN

In the following table, the results for the base case (RANDOM FOREST models divided by Period without natural gas price) are shown:

	MAE_Energy	MAE_Price	MAE_E+P	MAE_Area	RMSE_Energy	RMSE_Price	RMSE_E+P	RMSE_Area
Naive	4.65	5.65	5.15	732.10	17.30	13.02	15.31	2,072.35
RF_grid	7.43	7.74	7.59	915.85	17.93	13.57	15.90	2,106.65

In this case, naive model results are better than the RANDOM FOREST considering MAE and RMSE metric.

The results for both models are poor: MAPE for 'Area' prediction of 15% and 19% (naive model and RANDOM FOREST respectively).

5.11. C.H. LA MUELA GENERACIÓN

In the following table, the results for the base case (RANDOM FOREST models divided by Period without natural gas price) are shown:

	MAE_Energy	MAE_Price	MAE_E+P	MAE_Area	RMSE_Energy	RMSE_Price	RMSE_E+P	RMSE_Area
Naive	2.62	4.35	3.49	580.60	32.48	14.74	25.22	5,304.45
RF_grid	4.40	5.92	5.16	783.20	24.82	14.58	20.36	3,912.13

The main conclusion is that both models are very similar, so it cannot be said that the ML model improves the naive model prediction results.

The results for both models are poor: MAPE for 'Area' prediction of approx. 13% and 17% (naive model and RANDOM FOREST respectively).

5.12. C.H. TAJO ENCANTADA

In the following table, the results for the base case (RANDOM FOREST models divided by Period without natural gas price) are shown:

	MAE_Energy	MAE_Price	MAE_E+P	MAE_Area	RMSE_Energy	RMSE_Price	RMSE_E+P	RMSE_Area
Naive	10.79	3.03	6.91	703.20	31.16	6.35	22.48	1,623.36
RF_grid	14.62	4.68	9.65	910.30	28.90	6.80	20.99	1,500.31

The main conclusion is that both models are very similar: naive model is better considering MAE metric and the other has better predictions considering RMSE metric, so it cannot be said that the ML model improves the naive model prediction results.

The results for both models are bad: MAPE for ‘Area’ prediction of approx. 23% and 30% (naive model and RANDOM FOREST respectively).

5.13. C.H. MORALETS GENERACIÓN

In the following table, the results for the base case (RANDOM FOREST models divided by Period without natural gas price) are shown:

	MAE_Energy	MAE_Price	MAE_E+P	MAE_Area	RMSE_Energy	RMSE_Price	RMSE_E+P	RMSE_Area
Naive	3.58	2.26	2.92	278.37	15.25	4.58	11.26	821.85
RF_grid	7.12	3.40	5.26	431.74	13.50	4.73	10.12	725.46

The main conclusion is that both models are very similar: naive model is better considering MAE metric and the other has better predictions considering RMSE metric, so it cannot be said that the ML model improves the naive model prediction results.

The results for both models are poor: MAPE for ‘Area’ prediction of approx. 12% and 19% (naive model and RANDOM FOREST respectively).

6. CONCLUSIONS

The main conclusion of this research project is that the prediction of the day-ahead electricity market sale bids for individual units using **ML techniques**, with only **public information** for **combined cycles** and **hydraulic units**, does not significantly improve the most basic prediction (bid prediction equals to the 24h time lagged bid).

Several reasons to understand this conclusion are presented:

- Individual unit sale bids do not seem to be elaborated by electrical companies following the opportunity cost (that should be approximately equal to the operative cost for combined cycle units) that the economic theory, and the market rules supposes. In fact, for combined cycle units, a relationship could not be found between natural gas price and the bids, but one was found between the marginal price lagged 24h and the bids. So, it seems that unit bids are set considering strategic company policies rather than a real opportunity cost calculation for every unit. The fact that the Spanish electricity market is dominated by only three big companies, which sell (and buy) the majority of the electrical energy could be behind the observed behaviour of the bids.

- ML techniques could be useful for predicting sale bids if more variables related to company strategic decisions could be included, such as natural gas price long term contracts, real unit performance (maximum power and efficiency at full and partial loads), contracts with electric buyer companies, relationship between units with the same owner, etc. Obviously, this type of information is not public, and it is almost impossible to find it using public resources.

- It is possible that ML techniques are not adequate to be used in this problem, and more complex techniques with specific logic and assumptions could be more suitable for achieving this purpose. It can be said that this problem could be similar to predicting ambient temperature without using meteorological models: it can be done, but possibly, a naive model will work better.

- More external features could be considered such as the prediction of the energy generation at 0 €/MWh price, or electricity demand prediction. Most of them are not easy to find in public repositories, and others do not seem to be the “magic clue” to solve the problem. Other external features that could improve the ML models are features related to the owner of the units, as it can be seen in notebook **“30_UNIT_Bid_Comparison.ipynb”**, since it is clear that each electrical company has its own way to make the bids (similar to all the units that owns). In order to do it, a not easy deep study of all the unit bids of each electrical company is needed.

- The basic prediction (naive model) works surprisingly well for some of the units studied (MAPE for “Area” predictions lower than 1%). That means that the behaviour of these offers is usually very stable, and that the abrupt changes in the bids are not frequent, nor easy to predict considering only public information. Predictions for units with variable behaviour were very poor, but even in this case, the basic model works better or equal than ML models.

ANNEX 1 - FRONTEND - USER MANUAL

The FRONTEND has been created with STREAMLIT in a notebook called **40_FRONTEND.ipynb**. In it, five different ".py" files have been developed:

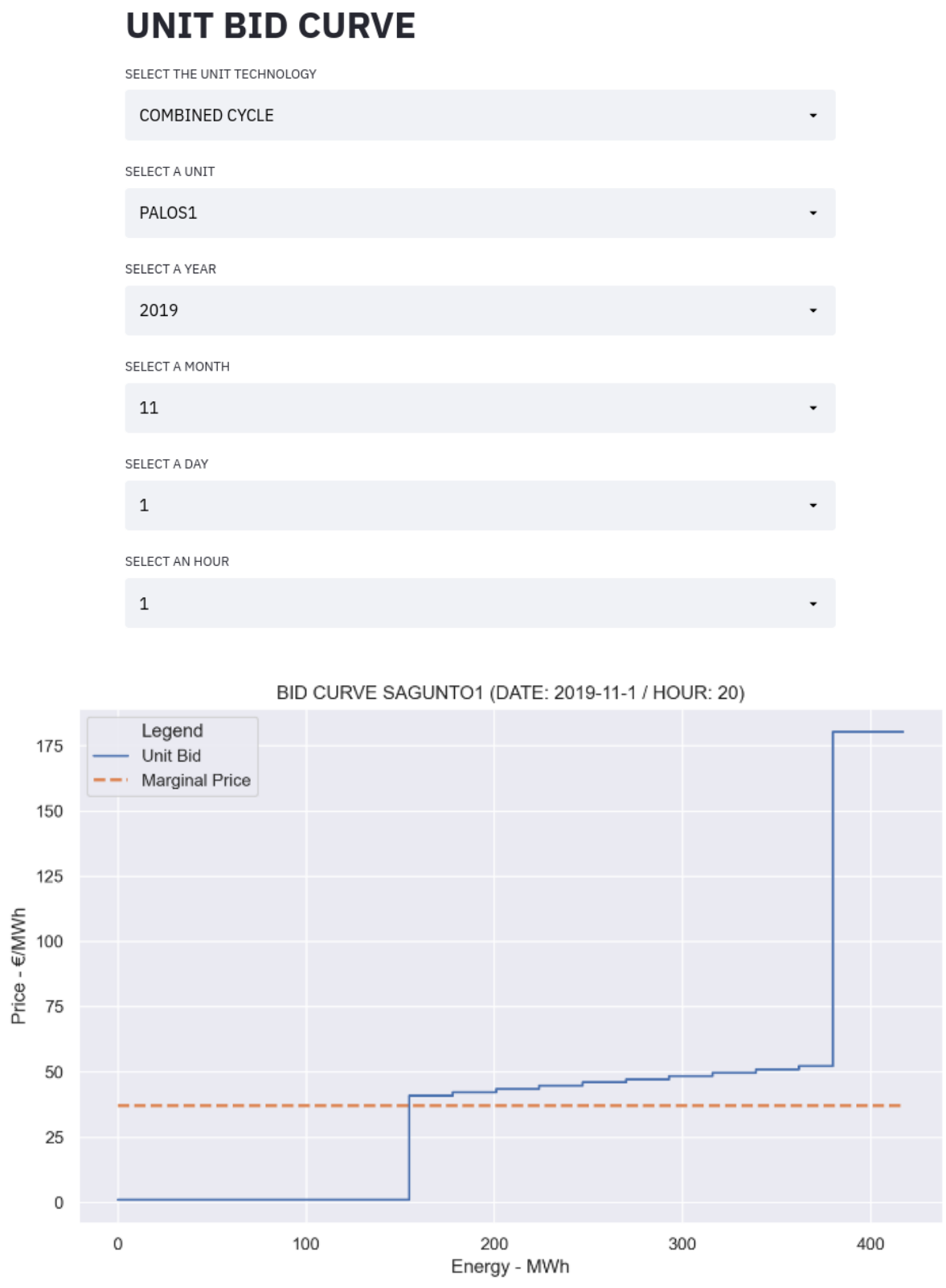
- **41_FRONTEND_BidCurve_Units.py**
- **42_FRONTEND_24h_BidCurve_Units.py**
- **43_FRONTEND_Bid_Units_timeperiod.py**
- **44_FRONTEND_BidCurve_Comparison.py**
- **45_FRONTEND_BidCurve_CompCumm.py**

To use them the following command must be run in the terminal:

> streamlit run *file_name.py*

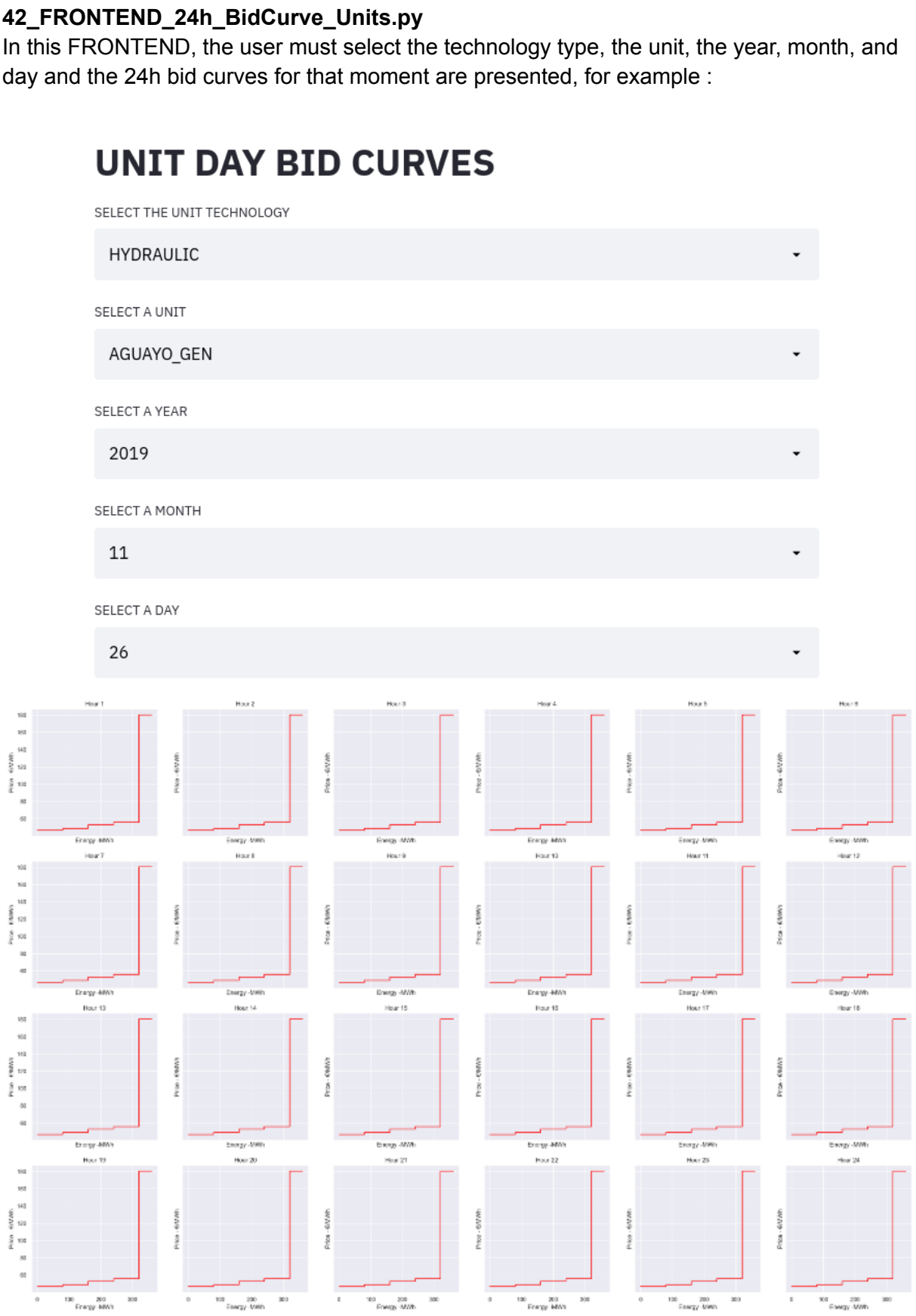
41_FRONTEND_BidCurve_Units.py

In this FRONTEND, the user must select the technology type, the unit, the year, month, day, and hour and the bid curve with the marginal price for that moment is presented, for example:



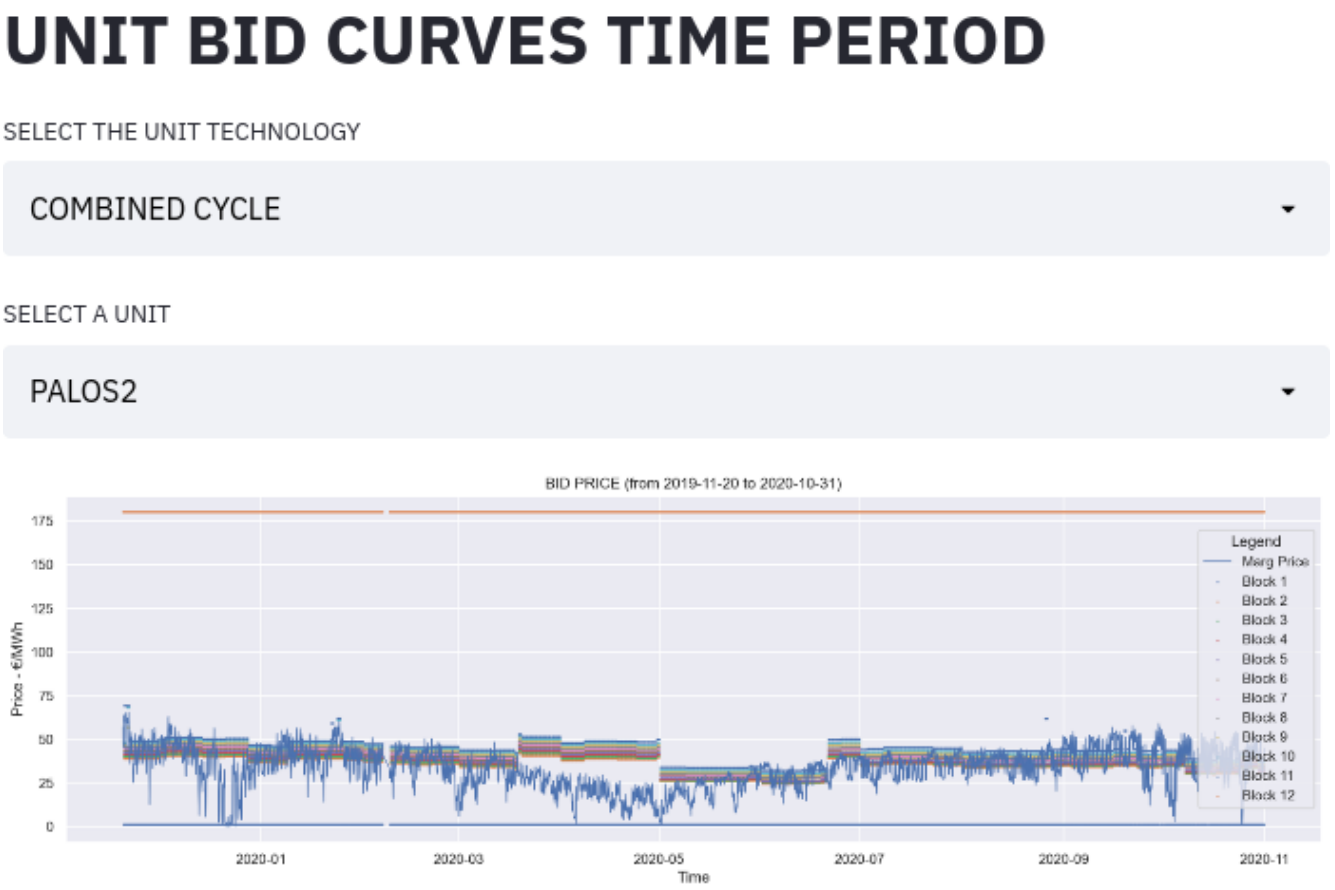
42_FRONTEND_24h_BidCurve_Units.py

In this FRONTEND, the user must select the technology type, the unit, the year, month, and day and the 24h bid curves for that moment are presented, for example :



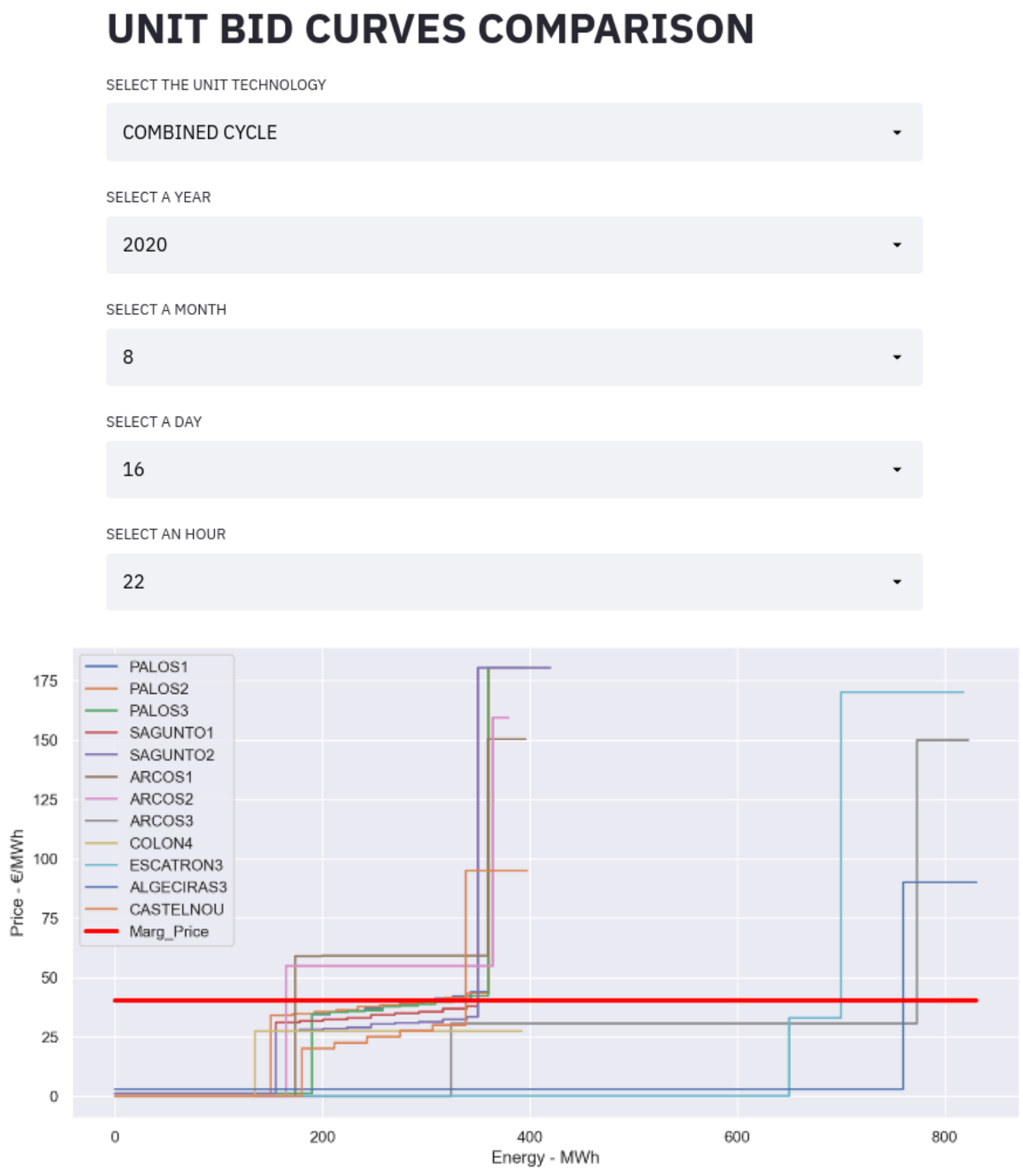
43_FRONTEND_Bid_Units_timeperiod.py

In this FRONTEND, the user must select the technology type, and the unit and the bid trend, together with the marginal price, are presented for the whole time period considered in this TFM, for example :



44_FRONTEND_BidCurve_Comparison.py

In this FRONTEND, the user must select the technology type, the unit, the year, month, day, and hour, and the bid curves for all the units considered in this TFM, together with the marginal price for that moment are presented, for example:



45_FRONTEND_BidCurve_CompCumm.py

In this FRONTEND, the user must select the technology type, the unit, the year, month, day, and hour, and the bid curves for all the units considered in this TFM, together with the marginal price for that moment are presented as an aggregated energy curve, for example: