BEEDATA

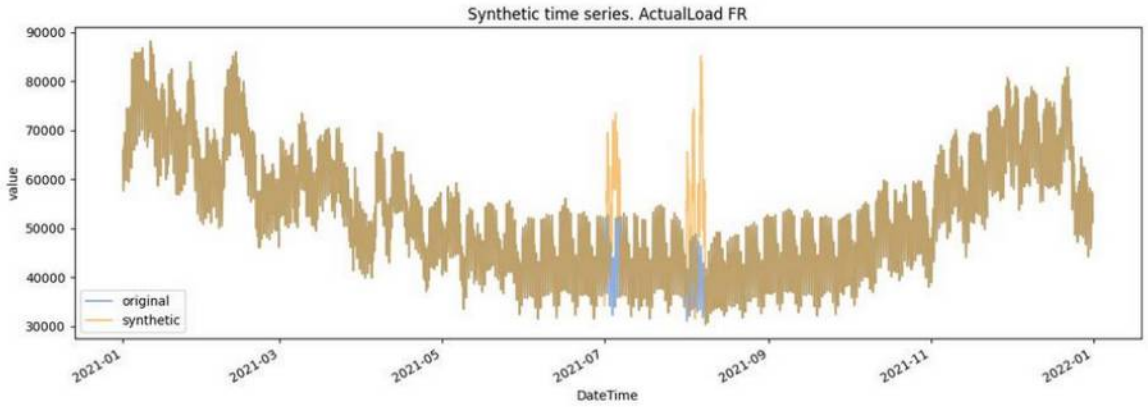
Scenario No 6.: Advanced Data Quality Analysis of Data Exchange Platforms

Tool for data quality analysis based on Identification and imputation of missing or erroneous readings

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 957739

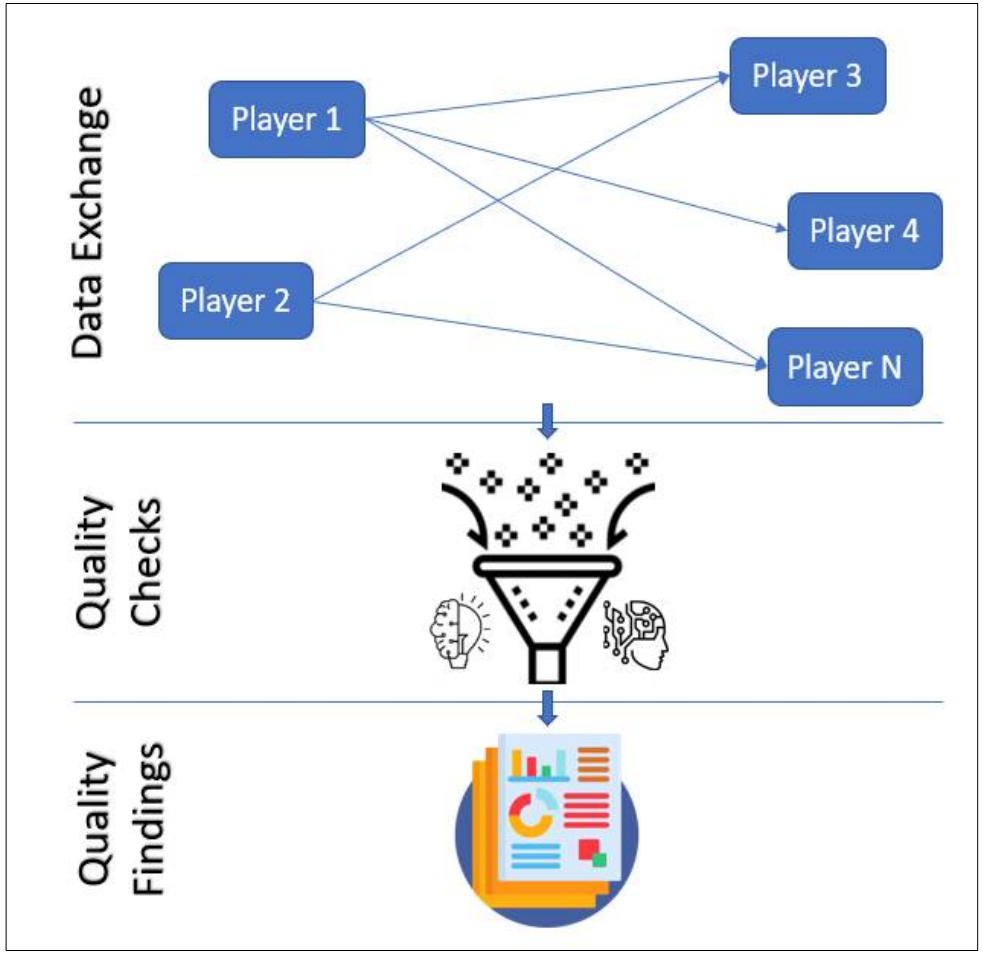
Context

- **Data and analysis** is an integral part of the everyday data exchanges among TSOs, DSOs and consumers.
- **Trust in the quality of data is vital**. Low quality data is propagated along the organization via erroneous data-driven decisions.
- **Data quality services** are focused on analyzing data in order to **detect, identify, quantify and fix issues in the provided data**. T
- This prject is focused on aggregated **data from ENTSO-E and smart grids data.**



Scope

1) **Detect outliers from time series** where standard methodologies are not sufficient
2) **Enhance the quality** of the data through detecting gaps in a given dataset.
3) **Apply advanced machine learning algorithms** on the data exchanged between different players
4) **To implement the tool of outliers detection and imputation** for two different data scenarios
   A) **Big data scenario:** Big number (>5K) of single high variable time series
   B) **Small data scenario:** Small number of time series data



Specific Objectives

1)To implement a set of flexible and open source big data algorithms to identify missing or outlier data in time series of data from exchanges among TSOs, DSOs and consumers

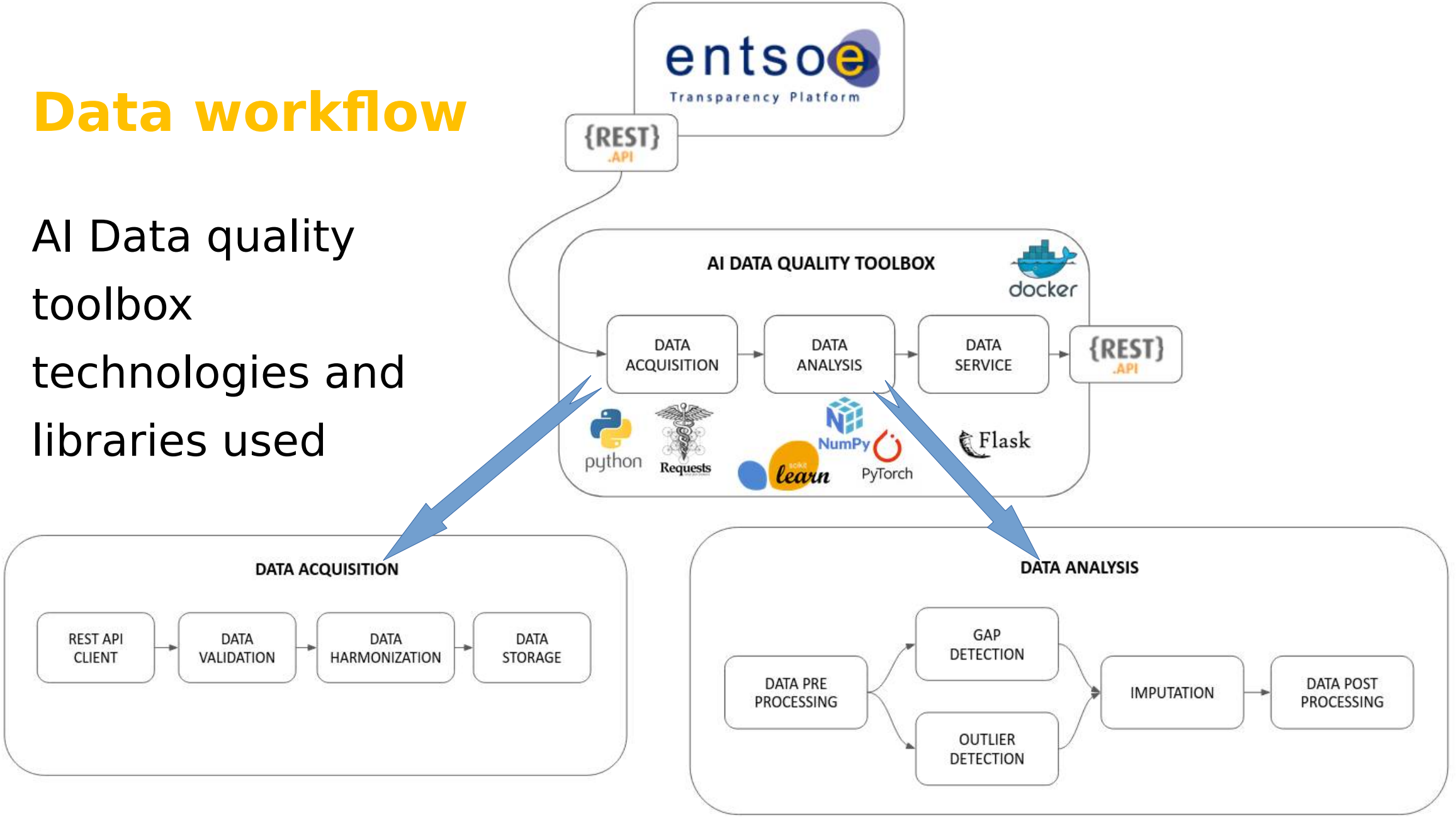
2)To design and implement a complementary big data algorithms to imputate and harmonize the missing or erroneous data collected

3)To validate these two algorithms integrated in a reference toolbox able to work both at large-scale and small scale pilots supporting the Onenet project

4)To promote and incentivise the widespread adoption of the big data toolbox To extent the use of the tool beyond the project consortium

Data workflow

AI Data quality toolbox technologies and libraries used



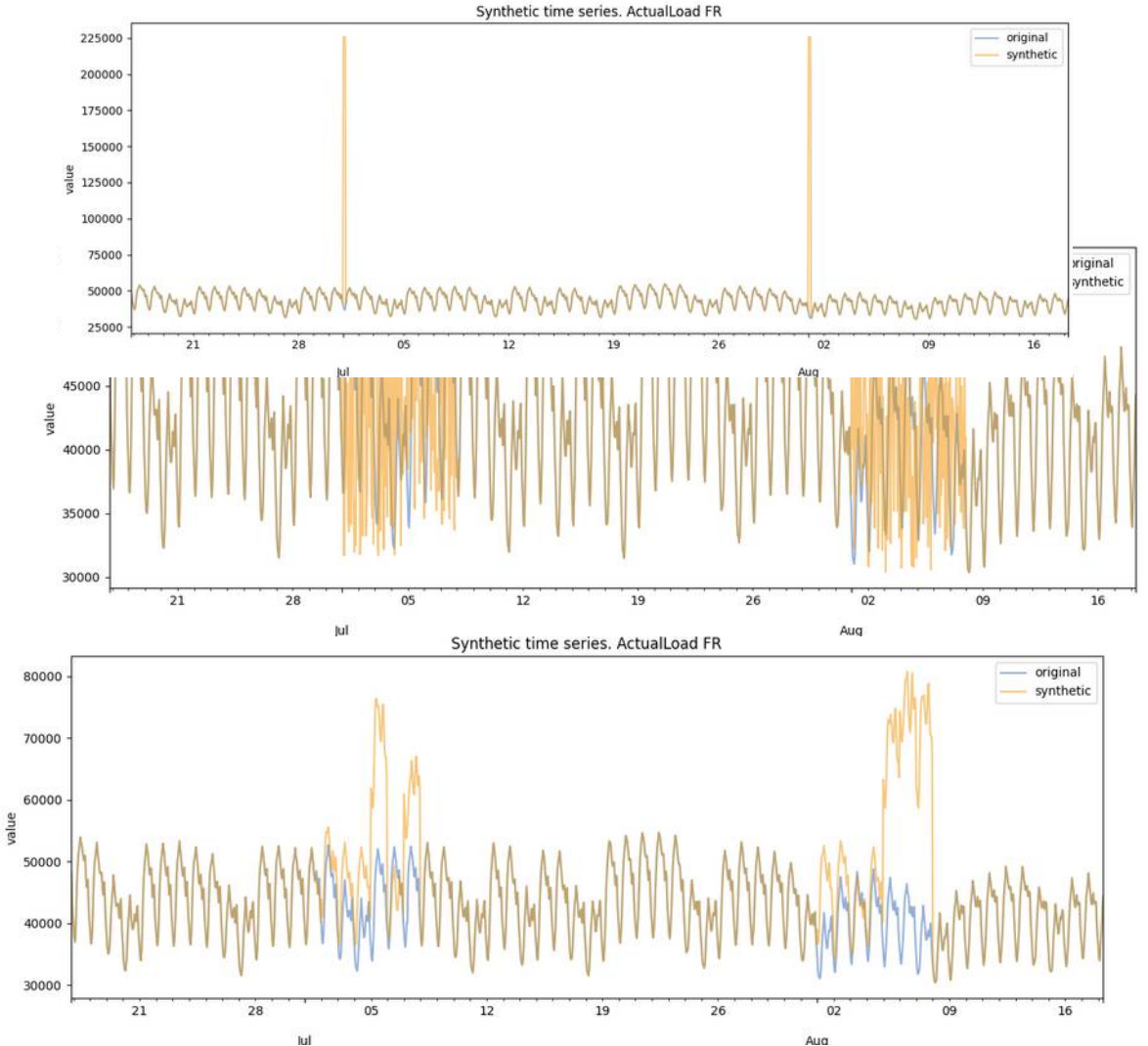
Outliers & gaps description

The data provider have no labelled data related to outliers so synthetic outlier scenarios are proposed.

Global outliers. A data point is considered a global outlier if its value is far outside the entirety of the data set in which it is found

Contextual outliers. Contextual outliers are data points whose value significantly deviates from other data within the same context

Collective outliers. A subset of data points within a data set is considered anomalous if those values as a collection deviate significantly from the entire data set, but the values of the individual data points are not themselves anomalous in either a contextual or global sense



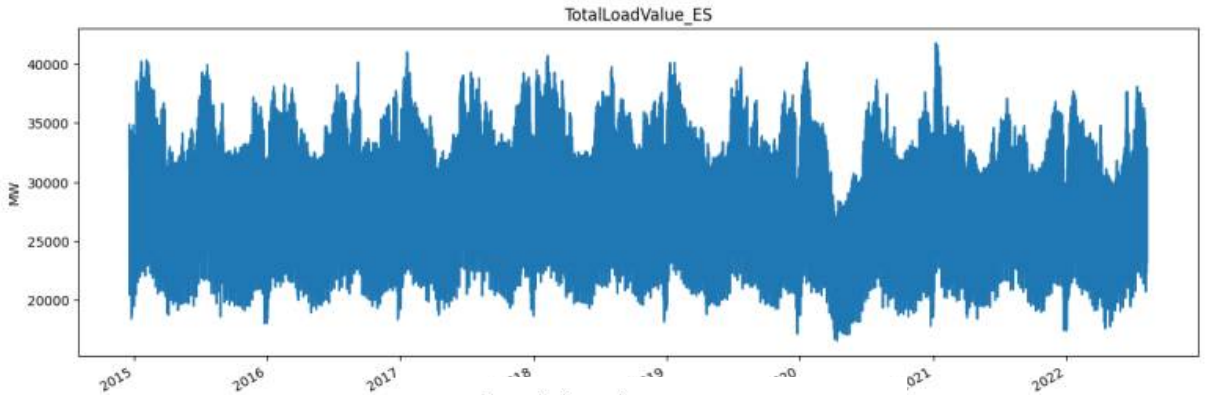
Data specification

Available data in ENSTO-E transparency platform is grouped in seven main topics:

- Load. Data about power consumption
- Generation. Energy production and production forecasts
- Transmission. Data about power transfers over borders between area
- Balancing. Data about Regulation energy used to keep the electrical transmission grid in balance
- Outages. Data about planned maintenances and failures inside the electrical transmission grid
- Congestion Management. Data about actions taken to relieve overloaded parts of the electrical transmission grid
- System Operations. Data about electricity transmission system operation

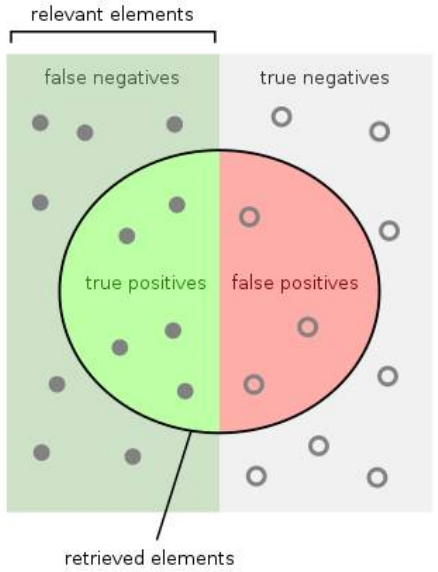
Data Pre processing

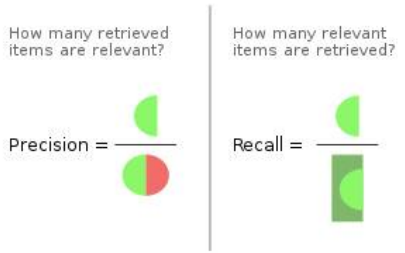
- Main goals of the preprocessing are:
- Apply physical constraints related to the type of data.
- Check values make sense considering the amount of consumers, installed power or transmission capacity
- Remove repeated entries
- Visual identification of time series intervals that should be handled apart



During the initial data exploration some potential outliers have been identified. The potential presence of non-labeled outliers in the training data can corrupt the results of the outlier detection methodology.

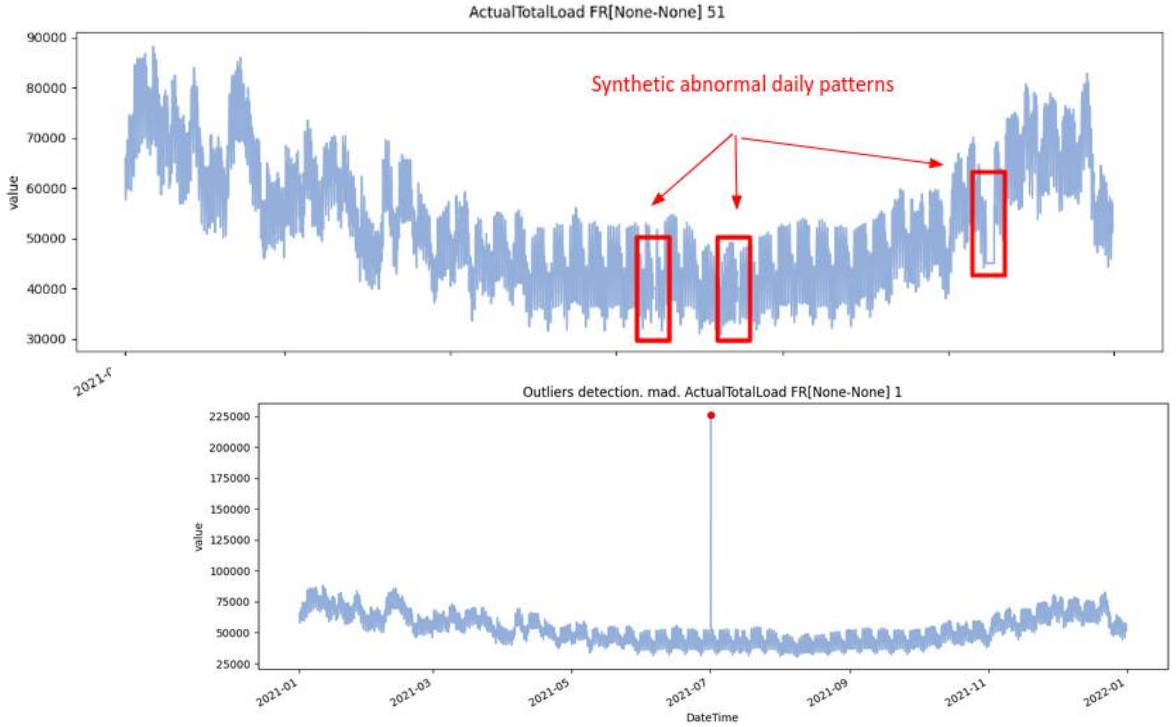
In the evaluation of the outlier detection method using the synthetic scenarios recall will be the main indicator and accuracy will be analyzed in each specific case.





Outliers detection models evaluation

Method	Type of error	Recall
LOF	Global spikes and plateaus	0.71
	Contextual	0.0
MAD	Global spikes and plateaus	0.75
	Contextual	0.0
SVM One-Class	Global spikes and plateaus	—
	Contextual	—
EnsembleIsolationForest	Global spikes and plateaus	0.68
	Contextual	0.15
PatternBased	Global spikes and plateaus	0.76
	Contextual	0.82 (*)



- Manual curation of positives is required in order to calculate precision so F1-score
- The EnsemleIsolationForest method requires more tuning in order to improve results in catching spike and global. Tuning of the model (windows length, threshold) is being reviewed to improve the performance.
- LOF and MAD method do not detect contextual outliers
- PatternBased method is the only outlier detection method that detects abnormal patterns in data

Imputation models evaluation

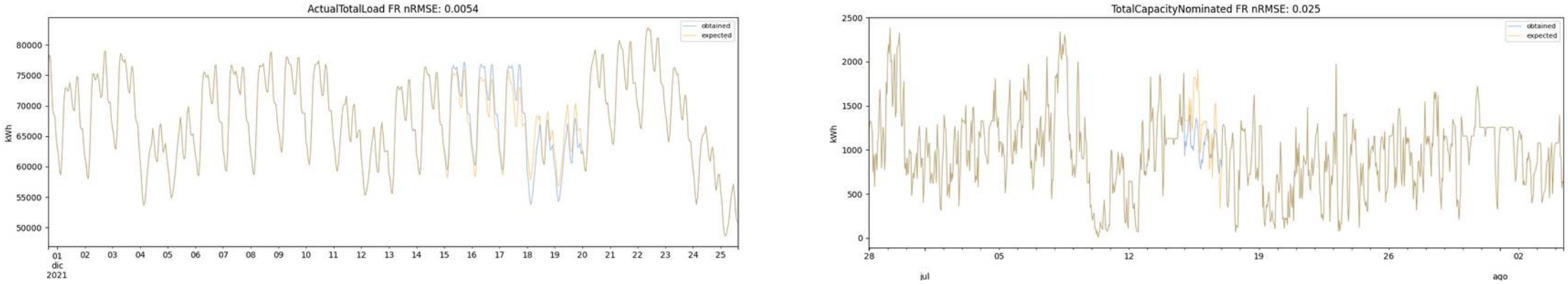


Figure 2.4.4- Imputation results in Actual Load time serie zoom (winter and summer)

- More imputation scenarios must be evaluated in order to provide final quality indicators. Even so, in the initial evaluated scenarios, the method performs well with cyclic behaviour time series and performs worse with non-cyclic behavior time series. This is the expected performing in imputation due the properties of the time series their self
- Imputation tuning (number of neighbours, lambda forgetting factor, etc.) is required per each time series because of different behaviour
- Hyper parameter optimization could be used to tune imputation.
- In cases when there are exogenous variables correlated with the time series it would be possible to consider them in the similarity analysis
- In case of known day-part cyclic behaviour in time series it would be possible to tune part of the day weighting in the similarity analysis