Beedata Mid-term report

D2

Authors:

Aleix Badia (Beedata Analytics)

Helena Boltà (Beedata Analytics)

Daniel Pérez (Beedata Analytics)

Verified by the appointed Reviewers	ENTSOE/Dmitry Belichenko ENTSOE/ Paulius Buktus, 29.12.2022
Approved by Project Coordinator	Fraunhofer / Padraic McKeever, 12.01.2023

Dissemination Level	Public	

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

Issue Record

Planned delivery date	07-10-2022
Actual date of delivery	12-01-2023

Version	Date	Author(s)	Notes
0.1	06-10-2022	Aleix Badia Helena Boltà Daniel Pérez	Technical version
0.2	28-11-2022	Aleix Badia	Integrating the reviewer's feedback _revision submitted
0.3	19-12-2022	Aleix Badia	Integrating the reviewer's feedback _revision submitted
0.4	28-12-2022	Aleix Badia	Integrating the reviewer's feedback and finalizing the document

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

About OneNet

The project OneNet (One Network for Europe) will provide a seamless integration of all the actors in the electricity network across Europe to create the conditions for a synergistic operation that optimizes the overall energy system while creating an open and fair market structure.

OneNet is funded through the EU's eighth Framework Programme Horizon 2020, "TSO – DSO Consumer: Large-scale demonstrations of innovative grid services through demand response, storage and small-scale (RES) generation" and responds to the call "Building a low-carbon, climate resilient future (LC)".

As the electrical grid moves from being a fully centralized to a highly decentralized system, grid operators have to adapt to this changing environment and adjust their current business model to accommodate faster reactions and adaptive flexibility. This is an unprecedented challenge requiring an unprecedented solution. The project brings together a consortium of over 70 partners, including key IT players, leading research institutions and the two most relevant associations for grid operators.

The key elements of the project are:

1. Definition of a common market design for Europe: this means standardized products and key parameters for grid services which aim at the coordination of all actors, from grid operators to customers;
2. Definition of a Common IT Architecture and Common IT Interfaces: this means not trying to create a single IT platform for all the products but enabling an open architecture of interactions among several platforms so that anybody can join any market across Europe; and
3. Large-scale demonstrators to implement and showcase the scalable solutions developed throughout the project. These demonstrators are organized in four clusters coming to include countries in every region of Europe and testing innovative use cases never validated before.

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

Table of Contents

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

List of Abbreviations and Acronyms

Acronym	Meaning
AI	Artificial Intelligence
CH	Switzerland
DoA	Description of the Action
DSO	Distribution System Operators
ENTSO-E	European Network of Transmission System Operators for Electricity
ES	Spain
FR	France
GMM	Gaussian Mixture Model
IQR	Interquartile range
kNN	K-Nearest Neighbor
LOF	Local Outlier Factor
LSTM	Long Short-Term Memory
MAD	Median Absolute Deviation
SVM	Super Vector Machine
TSO	Transmission System Operator
UCTE	Union for the Coordination of Transmission of Electricity

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

Executive Summary

Data and analysis is increasingly becoming an integral part of the everyday electricity system and more specific in data exchanges among Transmission System Operators (TSO), Distribution System Operators (DSO) and consumers. With a growing emphasis on data-led decision making across different organizations, trust in the quality of data is vital. Low quality data is propagated along the organization via erroneous data-driven decisions. A common error-prone use case would be forecasting. Fitting forecasting models with erroneous data would lead to predicting erroneous scenarios. With the AI data quality toolbox developed by beedata in this scenario 6 of the OneNet Project, we expect to improve the quality of the data managed by the data provider.

Within this deliverable the work carried for the project executed by Beedata will be showed. In this specific project, the use cases are focused on aggregated data from the European Network of Transmission System Operators for Electricity (ENTSO-E) [4], association of grid operators in Europe. Based on these use cases, the automatic data downloader per ENTSO E data provider has been implemented, together with the initial data exploration, to have a preliminary idea of the kind and amount of data available. A synthetic outlier generator library has been then implemented to create synthetic scenarios required to evaluate the detection and imputation methods, since there were no outliers data previously labelled.

In task 3 the big data outlier detection algorithm and library has been implemented, and results have been obtained in a preliminary version. The Initial evaluation and tuning of the big data outlier detection algorithm is ongoing. In order to finish with a first version of the whole process, the big data imputation algorithm and library has been implemented, and an initial evaluation and tuning of the big data imputation algorithm has been done. Preliminary conclusions of this two algorithms implementation are finally presented.

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

1 Introduction

Data quality services are focused on analyzing data in order to detect, identify, quantify and fix issues in the provided data. Type and source of issues are multiple and diverse. In this specific project, the use cases are focused on aggregated data from the European Network of Transmission System Operators for Electricity (ENTSO-E) [4], association of grid operators in Europe, and complementary on smart grids data from other use cases.

The implementation of a data quality service requires a dataset to proper design and test the detection, identification, quantification and fixing algorithms.

In this deliverable, data requirements are introduced to:

- Describe data scenario attributes
- Identify uses cases and provider
- Describe data scenarios

All these data scenarios are going to be used as input and test data during the implementation of the data quality analysis toolbox.

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

2 Overall objectives and work carried out

2.1 Explanation of the work carried out and overview of the progress

The work is progressing well, with all due deliverables for the period submitted on time. Major objectives of the project have been achieved to the extent of the planned within the scope of the reporting period. The progress in all expected impacts is fully on-track. There are no significant deviations from plan either in time, resource use, or achievement of milestones and results.

See the main results already been achieved:

Technical highlights

- Automatic data downloader per ENTSO E data provider has been implemented
- Initial data exploration has been done to have a preliminary idea of the kind and amount of data available
- A synthetic outlier generator library has been implemented to create synthetic scenarios required to evaluate the detection and imputation methods
- Big data outlier detection algorithm and library has been implemented
- Initial evaluation and tuning of the big data outlier detection algorithm is ongoing
- Big data imputation algorithm and library has been implemented
- Initial evaluation and tuning of the big data imputation algorithm has been done

For the next period the highlights are:

- Final evaluation and tuning of the big data outlier detection algorithm
- Final evaluation and tuning of the big data imputation algorithm
- Implementation of the small data outlier detection algorithm
- Implementation of the small data outlier detection library
- Evaluation and tuning of the small data outlier detection algorithm
- Implementation of the small data imputation algorithm
- Implementation of the small data imputation library
- Evaluation and tuning of the small data imputation algorithm

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

2.1.1 Objectives

Achievement highlights:

Objective 1	To implement a set of flexible and open source big data algorithms to collect, and identification of missing or outlier data in time series of data from exchanges among TSOs, DSOs and consumers, while ensuring data protection and security.
	Data is automatically downloaded from the data provider in order to implement an automatic outlier detection system. Already implemented an automatic download library to be used in the OneNet pipeline. Implemented the big data outlier detection algorithm and library. Both the data downloader and big data outlier detection libraries are brought together in the first version of the OneNet toolbox to be used to automatically download, detect and impute outliers The achievement of this objective is fully on-track, with all work planned for the period successfully executed. DoA planned finalization: M4.
Objective 2	To design and implement complementary big data algorithms to impute and harmonize the missing or erroneous data collected as a basis for full interoperability between databases and tools.
	Implemented the big data imputation algorithm and library. The big data imputation library is part of the first version of the OneNet toolbox to be used to automatically to download, detect and impute outliers The achievement of this objective is fully on-track, with all work planned for the period successfully executed. DoA planned finalization: M5
Objective 3	To validate these two algorithms integrated in a reference toolbox able to work both at large-scale and small scale pilots supporting different multi-party business cases, in the different pilots and scenarios of the OneNet project
	Synthetic outliers generator has been implemented as was agreed with patterns due missing outlier labeled data. The big data outlier detection and imputation has been evaluated under the synthetic outlier scenarios introduced in the first deliverable The achievement of this objective is fully on-track, with all work planned for the period successfully executed. DoA planned finalization: M4 and M5
Objective 4	To promote and incentivise the widespread adoption of the big data toolbox To extent the use of the tool beyond the project consortium
	PENDING task. We need a strategy defined with the OneNet Consortium members, in order to activate the tasks related to this objective, starting on November 2022

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

2.2 Explanation of the work carried in T2: Database selection, description, classification of data sources, and technical requirements

All sub tasks related to task 2 have been finished and are explained within the Deliverable D1: Datasets description and technical requirements.

2.3 Explanation of the work carried in T3. Build outlier detection method - Big data and small data model

2.3.1 Overall T3 progress towards objectives

Objectives of the task

The development of the time series data outliers detection algorithm that will allow the data coming from different sources to be aligned and treated together in an agile and robust way.

2.3.1.1 Task 3.1 - Data Pre processing [M1]

Objectives of the sub-task

A preprocessing procedure which removes invalid data in terms of specific properties from the times series is implemented. The specific properties to consider are: i) physical constraints i.e., load consumption must be positive, generation must be lower than installed power; ii) time constraints.:i.e. repeated data points that must be removed or merged. Preprocessing is also done to obtain sample weights in case some samples are less important for outlier detection. Weighting criteria is based on calendar and specific type of outliers or application. Dataset is then split between train and test subsets.

Summary of progress towards objectives and description of the work performed

Data downloading and exploration. Automatic ENTSO-E data downloading has been implemented and time series exploration has been done. Client has been implemented to download data from the ENSTO-E transparency platform which 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 areas
- Balancing. Data about Regulation energy used to keep the electrical transmission grid in balance

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

- 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

The specific data selected, downloaded and processed is the one described in table 2.3.1.

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

Table 2.3.1 - Data selected from ENTSO-E transparency platform

Name	Description	Document Type	Process Type	Business Type	Data unit	Time resolution	Country	Amount of series
Actual Total Load [6.1.A]	Actual total load per bidding zone per market time unit, the total load being defined as equal to the sum of power generated by plants on both TSO/DSO network	A65	A16		MW	15 minutes 60 minutes	ES DE FR	3
Aggregated Generation per Type [16.1.B&C]	Actual aggregated Net generation output (MW) per market time unit and per production type.	A75	A16		MW	60 minutes	ES DE FR	3 x 3 Nuclear Solar Wind
Total Capacity Nominated [12.1.B]	For every market time unit and per direction between bidding zones the total capacity nominated (MW) from capacity allocated via explicit allocations only.	A26		B08	MW	60 minutes	FR	2 x bidding zone
Forecasted Day-ahead Transfer Capacities [11.1]	The forecasted NTC (MW) per direction between bidding zones, including technical profiles. only in NTC allocation method	A61			MW	60 minutes	FR	2 x bidding zone

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

Data exploration has been done in order to identify which are the best time series to be used in the development. Time series have been downloaded and analyzed.

- Total Load Value. See Figure 2.3.1, 2.3.2
- Actual Generation Output for different technologies. See Figure 2.3.3, 2.3.4, 2.3.5
- Total Capacity Nominated. See Figure 2.3.6, 2.3.7
- Forecast Transfer Capacity. See Figure 2.3.8, 2.3.9

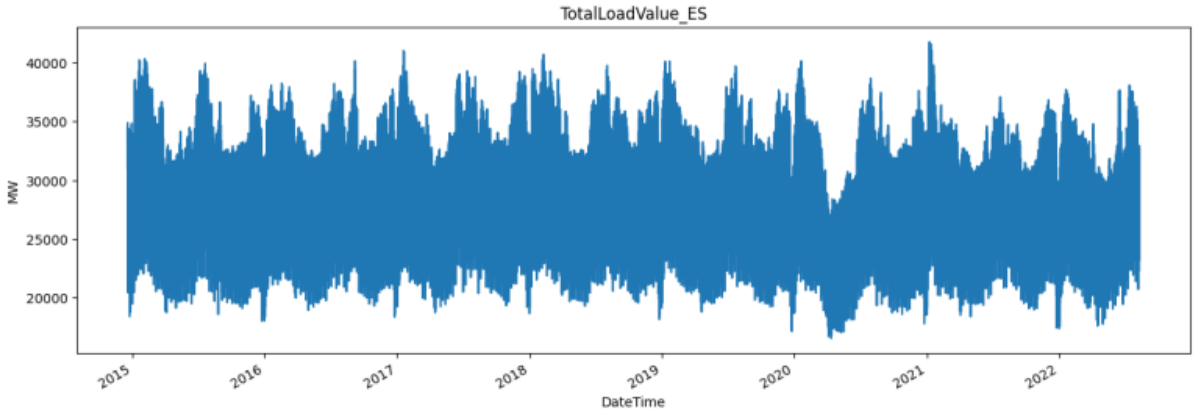


Figure 2.3.1 - Total Load Value ES

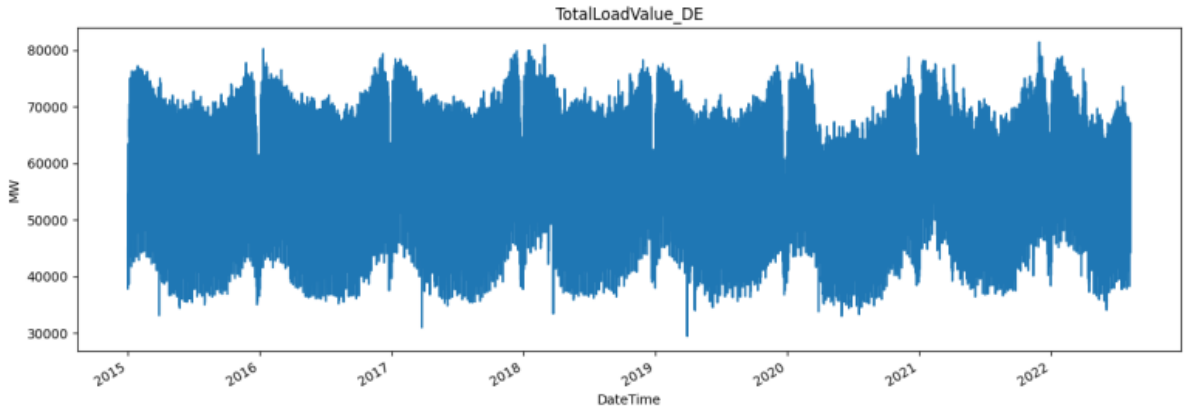


Figure 2.3.2. Total Load Value DE

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

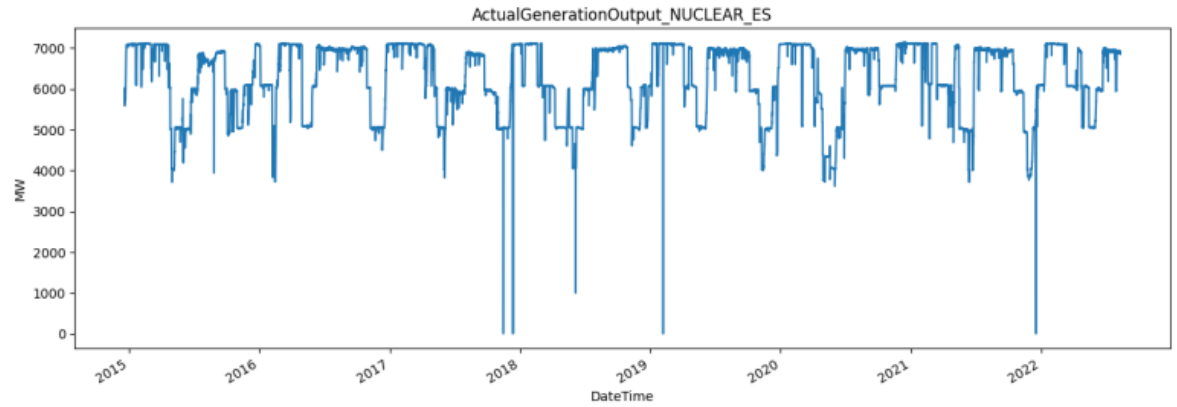


Figure 2.3.3. Actual Generation Output - Nuclear - ES

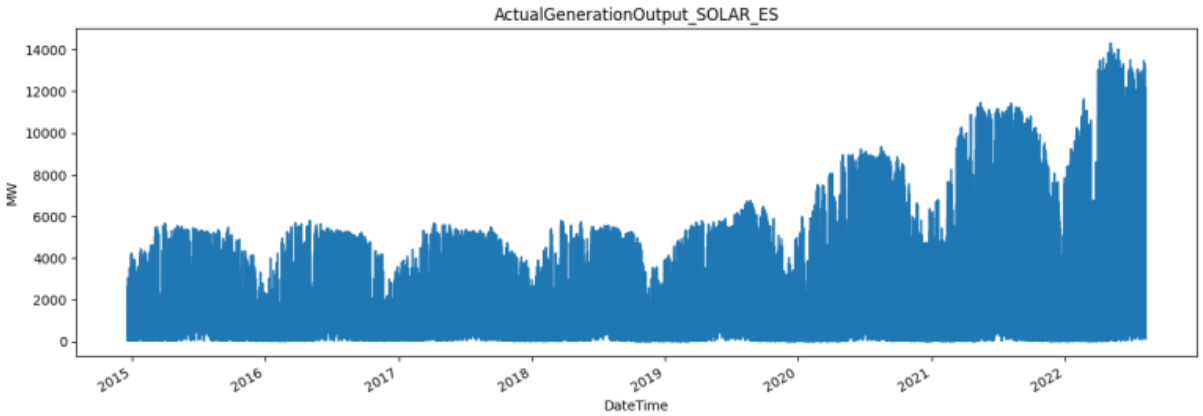


Figure 2.3.4. Actual Generation Output - Solar - ES

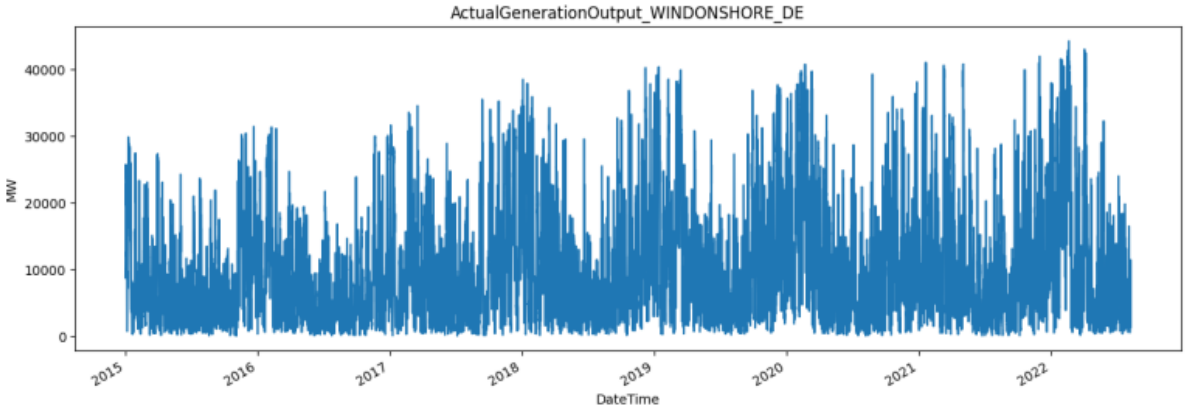


Figure 2.3.5. Actual Generation Output - Wind Onshore - ES

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

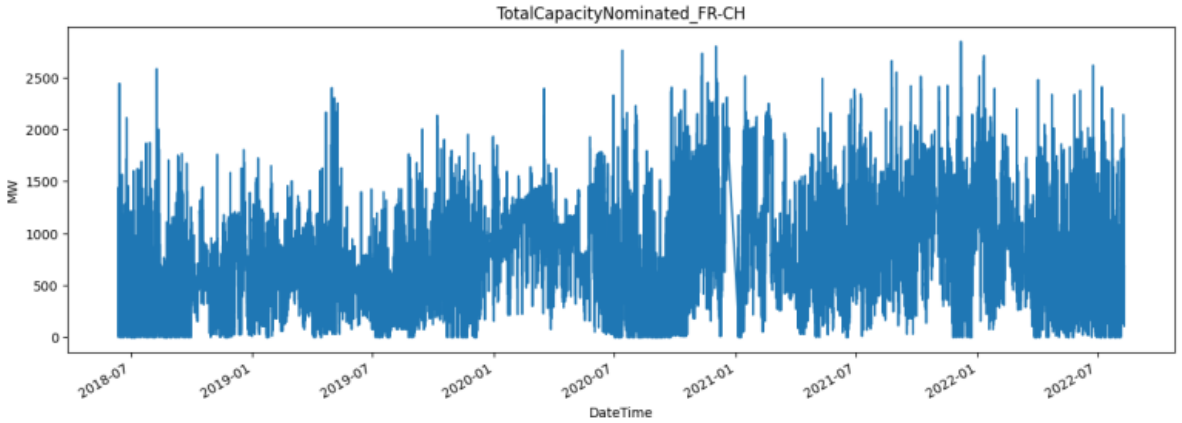


Figure 2.3.6. Total Capacity Nominated - FR-CH

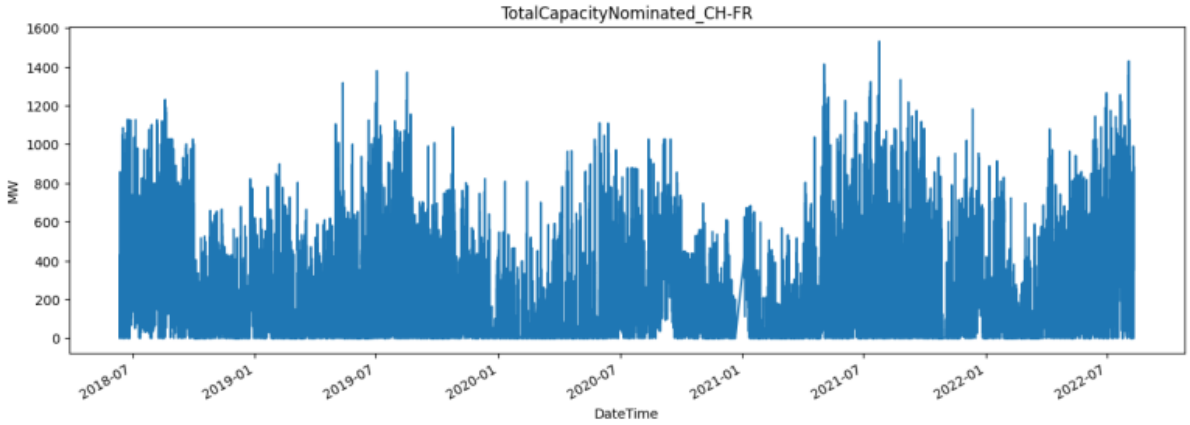


Figure 2.3.7. Total Capacity Nominated - CH - FR

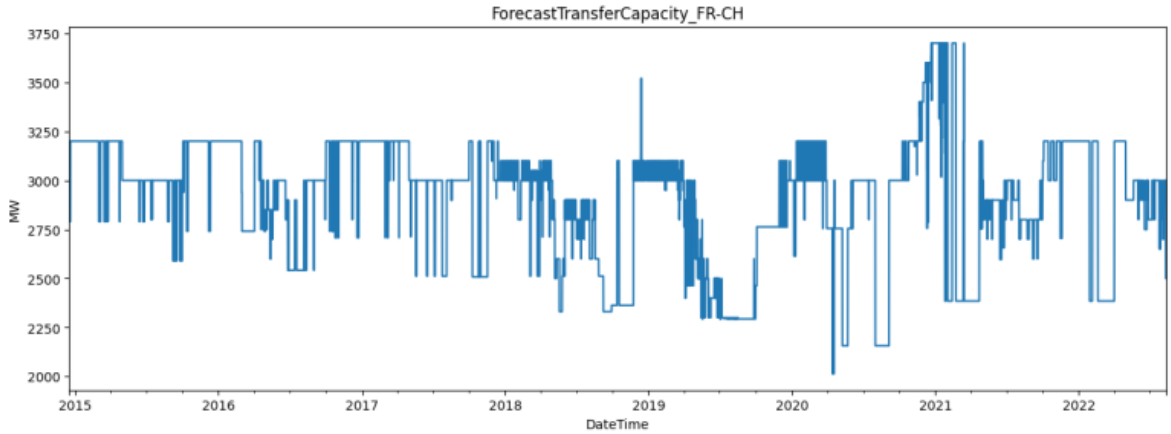


Figure 2.3.8. Forecast Transfer Capacity FR - CH

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

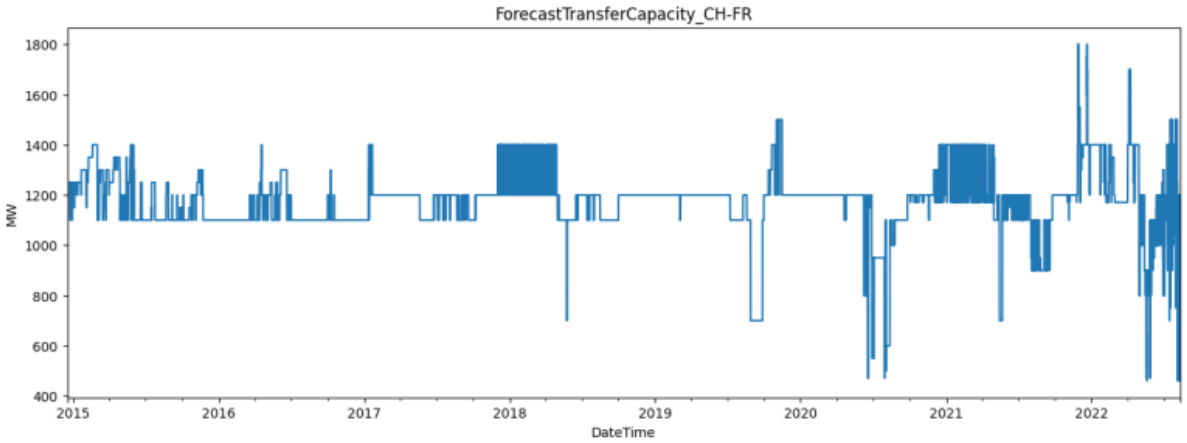


Figure 2.3.9. Forecast Transfer Capacity CH - FR

Preprocessing has been implemented after analyzing time series properties. 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. As the data provided is not labeled an unsupervised outlier method was proposed. The potential presence of non-labeled outliers in the training data can corrupt the results of the outlier detection methodology. The evaluation of the outlier detection algorithms using synthetic scenarios cannot be only based on the F1-score as the accuracy component would consider true positive outliers in the original as false positives. In the evaluation of the outlier detection method using the synthetic scenarios recall (see Figure 2.3.10) will be the main indicator and accuracy will be analyzed in each specific case.

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

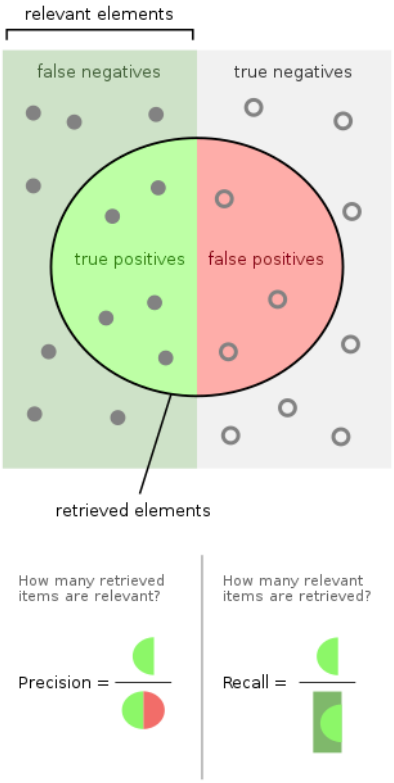


Figure 2.3.10. F1 description [6]

2.3.1.2 Task 3.5. Create synthetic outlier datasets. [M4]

Objectives of the task

Synthetic outliers are created to evaluate the model under outlier scenarios not present in data. Outliers are domain specific but typical outlier patterns are: i) Spikes, ii) Plateaus, iii) Null values. Different gains and time lengths are applied to these typical outlier patterns in order to evaluate in which cases the model is able to properly classify samples

Summary of progress towards objectives and description of the work performed

Synthetic outlier generator has been implemented to create the outlier scenarios agreed with partners and introduced in deliverable D1. See Table 2.3.2

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

Table 2.3.2 - Specification of the outlier scenarios

Id	Time serie	Type of outlier	Value (*)	Season	Duration	Frequency (*)	Pattern
1	Actual Total Load [6.1.A]	Global	x3 percentile %95 value in time serie	Summer	1 time step	1 time	Spike
2	Actual Total Load [6.1.A]	Global	x1.5 percentile %95 value in time serie	Summer	1 time step	1 time	Spike
3	Actual Total Load [6.1.A]	Global	x3 percentile %95 value in time serie	Winter	1 time step	1 time	Spike
4	Actual Total Load [6.1.A]	Global	x1.5 percentile %95 value in time serie	Winter	1 time step	1 time	Spike
5	Total Capacity Nominated [12.1.B]	Global	x3 percentile %95 value in time serie	Summer	1 time step	1 time	Spike
6	Total Capacity Nominated [12.1.B]	Global	x1.5 percentile %95 value in time serie	Summer	1 time step	1 time	Spike
7	Total Capacity Nominated [12.1.B]	Global	x3 percentile %95 value in time serie	Winter	1 time step	1 time	Spike
8	Total Capacity Nominated [12.1.B]	Global	x1.5 percentile %95 value in time serie	Winter	1 time step	1 time	Spike
9	Forecasted Day-ahead Transfer Capacities [11.1]	Global	x3 percentile %95 value in time serie	Summer	1 time step	1 time	Spike
10	Forecasted Day-ahead Transfer Capacities [11.1]	Global	x1.5 percentile %95 value in time serie	Summer	1 time step	1 time	Spike

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

11	Forecasted Day-ahead Transfer Capacities [11.1]	Global	x3 percentile %95 value in time serie	Winter	1 time step	1 time	Spike
12	Forecasted Day-ahead Transfer Capacities [11.1]	Global	x1.5 percentile %95 value in time serie	Winter	1 time step	1 time	Spike
13	Actual Total Load [6.1.A]	Global	x3 percentile %95 value in time serie	Summer	1 time step	4 times per month	Spike
14	Actual Total Load [6.1.A]	Global	x1.5 percentile %95 value in time serie	Summer	1 time step	4 times per month	Spike
15	Actual Total Load [6.1.A]	Global	x3 percentile %95 value in time serie	Winter	1 time step	4 times per month	Spike
16	Actual Total Load [6.1.A]	Global	x1.5 percentile %95 value in time serie	Winter	1 time step	4 times per month	Spike
17	Total Capacity Nominated [12.1.B]	Global	x3 percentile %95 value in time serie	Summer	1 time step	4 times per month	Spike
18	Total Capacity Nominated [12.1.B]	Global	x1.5 percentile %95 value in time serie	Summer	1 time step	4 times per month	Spike
19	Total Capacity Nominated [12.1.B]	Global	x3 percentile %95 value in time serie	Winter	1 time step	4 times per month	Spike
20	Total Capacity Nominated [12.1.B]	Global	x1.5 percentile %95 value in time serie	Winter	1 time step	4 times per month	Spike
21	Forecasted Day-ahead Transfer Capacities [11.1]	Global	x3 percentile %95 value in time serie	Summer	1 time step	4 times per month	Spike
22	Forecasted Day-ahead Transfer Capacities [11.1]	Global	x1.5 percentile %95 value in time serie	Summer	1 time step	4 times per month	Spike

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

23	Forecasted Day-ahead Transfer Capacities [11.1]	Global	x3 percentile %95 value in time serie	Winter	1 time step	4 times per month	Spike
24	Forecasted Day-ahead Transfer Capacities [11.1]	Global	x1.5 percentile %95 value in time serie	Winter	1 time step	4 times per month	Spike
25	Actual Total Load [6.1.A]	Global	x3 percentile %95 value in time serie	Summer	10 time step	2 times per month	Plateau
26	Actual Total Load [6.1.A]	Global	x1.5 percentile %95 value in time serie	Summer	10 time step	2 times per month	Plateau
27	Actual Total Load [6.1.A]	Global	x3 percentile %95 value in time serie	Winter	10 time step	2 times per month	Plateau
28	Actual Total Load [6.1.A]	Global	x1.5 percentile %95 value in time serie	Winter	10 time step	2 times per month	Plateau
29	Total Capacity Nominated [12.1.B]	Global	x3 percentile %95 value in time serie	Summer	10 time step	2 times per month	Plateau
30	Total Capacity Nominated [12.1.B]	Global	x1.5 percentile %95 value in time serie	Summer	10 time step	2 times per month	Plateau
31	Total Capacity Nominated [12.1.B]	Global	x3 percentile %95 value in time serie	Winter	10 time step	2 times per month	Plateau
32	Total Capacity Nominated [12.1.B]	Global	x1.5 percentile %95 value in time serie	Winter	10 time step	2 times per month	Plateau
33	Actual Total Load [6.1.A]	Contextual	Random values in min-max range (month)	Summer	1 day	1 time	Daily

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

34	Actual Total Load [6.1.A]	Contextual	Random value in min-max range (month)	Winter	1 day	1 time	Daily
35	Total Capacity Nominated [12.1.B]	Contextual	Random value in min-max range (month)	Summer	1 day	1 time	Daily
36	Total Capacity Nominated [12.1.B]	Contextual	Random value in min-max range (month)	Winter	1 day	1 time	Daily
37	Forecasted Day-ahead Transfer Capacities [11.1]	Contextual	Random value in min-max range (month)	Summer	1 day	1 time	Daily
38	Forecasted Day-ahead Transfer Capacities [11.1]	Contextual	Random value in min-max range (month)	Winter	1 day	1 time	Daily
39	Actual Total Load [6.1.A]	Contextual	Random value in min-max range (month)	Summer	1 day	2 times per month	Daily
40	Actual Total Load [6.1.A]	Contextual	Random value in min-max range (month)	Winter	1 day	2 times per month	Daily
41	Total Capacity Nominated [12.1.B]	Contextual	Random value in min-max range (month)	Summer	1 day	2 times per month	Daily

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

42	Total Capacity Nominated [12.1.B]	Contextual	Random value in min-max range (month)	Winter	1 day	2 times per month	Daily
43	Forecasted Day-ahead Transfer Capacities [11.1]	Contextual	Random value in min-max range (month)	Summer	1 day	2 times per month	Daily
44	Forecasted Day-ahead Transfer Capacities [11.1]	Contextual	Random value in min-max range (month)	Winter	1 day	2 times per month	Daily
45	Actual Total Load [6.1.A]	Collective	Random sort of daily values	Summer	1 day	1 time	Daily
46	Actual Total Load [6.1.A]	Collective	Random sort of daily values	Winter	1 day	1 time	Daily
47	Total Capacity Nominated [12.1.B]	Collective	Random sort of daily values	Summer	1 day	1 time	Daily
48	Total Capacity Nominated [12.1.B]	Collective	Random sort of daily values	Winter	1 day	1 time	Daily
49	Forecasted Day-ahead Transfer Capacities [11.1]	Collective	Random sort of weekly values	Summer	1 day	1 time	Daily
50	Forecasted Day-ahead Transfer Capacities [11.1]	Collective	Random sort of weekly values	Winter	1 day	1 time	Daily
51	Actual Total Load [6.1.A]	Collective	Random sort of daily values	Summer	1 day	2 times per month	Daily
52	Actual Total Load [6.1.A]	Collective	Random sort of daily values	Winter	1 day	2 times per month	Daily

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

53	Total Capacity Nominated [12.1.B]	Collective	Random sort of daily values	Summer	1 day	2 times per month	Daily
54	Total Capacity Nominated [12.1.B]	Collective	Random sort of daily values	Winter	1 day	2 times per month	Daily
55	Forecasted Day-ahead Transfer Capacities [11.1]	Collective	Random sort of daily values	Summer	1 day	2 times per month	Daily
56	Forecasted Day-ahead Transfer Capacities [11.1]	Collective	Random sort of daily values	Winter	1 day	2 times per month	Daily
57	Aggregated Generation per Type [16.1.B&C] Nuclear	Global	x3 percentile %95 value in time serie	Summer	1 time step	1 time	Spike
58	Aggregated Generation per Type [16.1.B&C] Nuclear	Global	x1.5 percentile %95 value in time serie	Summer	1 time step	1 time	Spike
59	Aggregated Generation per Type [16.1.B&C] Nuclear	Global	x3 percentile %95 value in time serie	Winter	1 time step	1 time	Spike
60	Aggregated Generation per Type [16.1.B&C] Nuclear	Global	x1.5 percentile %95 value in time serie	Winter	1 time step	1 time	Spike
61	Aggregated Generation per Type [16.1.B&C] Nuclear	Global	x3 percentile %95 value in time serie	Summer	1 time step	4 times per month	Spike

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

62	Aggregated Generation per Type [16.1.B&C] Nuclear	Global	x1.5 percentile %95 value in time serie	Summer	1 time step	4 times per month	Spike
63	Aggregated Generation per Type [16.1.B&C] Nuclear	Global	x3 percentile %95 value in time serie	Winter	1 time step	4 times per month	Spike
64	Aggregated Generation per Type [16.1.B&C] Nuclear	Global	x1.5 percentile %95 value in time serie	Winter	1 time step	4 times per month	Spike
65	Aggregated Generation per Type [16.1.B&C] Nuclear	Global	x3 percentile %95 value in time serie	Summer	10 time step	2 times per month	Plateau
66	Aggregated Generation per Type [16.1.B&C] Nuclear	Global	x1.5 percentile %95 value in time serie	Summer	10 time step	2 times per month	Plateau
67	Aggregated Generation per Type [16.1.B&C] Nuclear	Global	x3 percentile %95 value in time serie	Winter	10 time step	2 times per month	Plateau
68	Aggregated Generation per Type [16.1.B&C] Nuclear	Global	x1.5 percentile %95 value in time serie	Winter	10 time step	2 times per month	Plateau
69	Aggregated Generation per Type [16.1.B&C] Nuclear	Contextual	Random values in min-max range (month)	Summer	1 day	1 time	Daily

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

70	Aggregated Generation per Type [16.1.B&C] Nuclear	Contextual	Random value in min-max range (month)	Winter	1 day	1 time	Daily
71	Aggregated Generation per Type [16.1.B&C] Nuclear	Contextual	Random value in min-max range (month)	Summer	1 day	2 times per month	Daily
72	Aggregated Generation per Type [16.1.B&C] Nuclear	Contextual	Random value in min-max range (month)	Winter	1 day	2 times per month	Daily
73	Aggregated Generation per Type [16.1.B&C] Nuclear	Collective	Random sort of daily values	Summer	1 day	1 time	Daily
74	Aggregated Generation per Type [16.1.B&C] Nuclear	Collective	Random sort of daily values	Winter	1 day	1 time	Daily
75	Aggregated Generation per Type [16.1.B&C] Nuclear	Collective	Random sort of daily values	Summer	1 day	2 times per month	Daily
76	Aggregated Generation per Type [16.1.B&C] Nuclear	Collective	Random sort of daily values	Winter	1 day	2 times per month	Daily
77	Aggregated Generation per Type [16.1.B&C] Solar	Global	x3 percentile %95 value in time serie	Summer	1 time step	1 time	Spike

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

78	Aggregated Generation per Type [16.1.B&C] Solar	Global	x1.5 percentile %95 value in time serie	Summer	1 time step	1 time	Spike
79	Aggregated Generation per Type [16.1.B&C] Solar	Global	x3 percentile %95 value in time serie	Winter	1 time step	1 time	Spike
80	Aggregated Generation per Type [16.1.B&C] Solar	Global	x1.5 percentile %95 value in time serie	Winter	1 time step	1 time	Spike
81	Aggregated Generation per Type [16.1.B&C] Solar	Global	x3 percentile %95 value in time serie	Summer	1 time step	4 times per month	Spike
82	Aggregated Generation per Type [16.1.B&C] Solar	Global	x1.5 percentile %95 value in time serie	Summer	1 time step	4 times per month	Spike
83	Aggregated Generation per Type [16.1.B&C] Solar	Global	x3 percentile %95 value in time serie	Winter	1 time step	4 times per month	Spike
84	Aggregated Generation per Type [16.1.B&C] Solar	Global	x1.5 percentile %95 value in time serie	Winter	1 time step	4 times per month	Spike
85	Aggregated Generation per Type [16.1.B&C] Solar	Global	x3 percentile %95 value in time serie	Summer	10 time step	2 times per month	Plateau

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

86	Aggregated Generation per Type [16.1.B&C] Solar	Global	x1.5 percentile %95 value in time serie	Summer	10 time step	2 times per month	Plateau
87	Aggregated Generation per Type [16.1.B&C] Solar	Global	x3 percentile %95 value in time serie	Winter	10 time step	2 times per month	Plateau
88	Aggregated Generation per Type [16.1.B&C] Solar	Global	x1.5 percentile %95 value in time serie	Winter	10 time step	2 times per month	Plateau
89	Aggregated Generation per Type [16.1.B&C] Solar	Contextual	Random values in min-max range (month)	Summer	1 day	1 time	Daily
90	Aggregated Generation per Type [16.1.B&C] Solar	Contextual	Random value in min-max range (month)	Winter	1 day	1 time	Daily
91	Aggregated Generation per Type [16.1.B&C] Solar	Contextual	Random value in min-max range (month)	Summer	1 day	2 times per month	Daily
92	Aggregated Generation per Type [16.1.B&C] Solar	Contextual	Random value in min-max range (month)	Winter	1 day	2 times per month	Daily
93	Aggregated Generation per Type [16.1.B&C] Solar	Collective	Random sort of daily values	Summer	1 day	1 time	Daily

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

94	Aggregated Generation per Type [16.1.B&C] Solar	Collective	Random sort of daily values	Winter	1 day	1 time	Daily
95	Aggregated Generation per Type [16.1.B&C] Solar	Collective	Random sort of daily values	Summer	1 day	2 times per month	Daily
96	Aggregated Generation per Type [16.1.B&C] Solar	Collective	Random sort of daily values	Winter	1 day	2 times per month	Daily
97	Aggregated Generation per Type [16.1.B&C] WindOn	Global	x3 percentile %95 value in time serie	Summer	1 time step	1 time	Spike
98	Aggregated Generation per Type [16.1.B&C] WindOn	Global	x1.5 percentile %95 value in time serie	Summer	1 time step	1 time	Spike
99	Aggregated Generation per Type [16.1.B&C] WindOn	Global	x3 percentile %95 value in time serie	Winter	1 time step	1 time	Spike
100	Aggregated Generation per Type [16.1.B&C] WindOn	Global	x1.5 percentile %95 value in time serie	Winter	1 time step	1 time	Spike
101	Aggregated Generation per Type [16.1.B&C] WindOn	Global	x3 percentile %95 value in time serie	Summer	1 time step	4 times per month	Spike

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

102	Aggregated Generation per Type [16.1.B&C] WindOn	Global	x1.5 percentile %95 value in time serie	Summer	1 time step	4 times per month	Spike
103	Aggregated Generation per Type [16.1.B&C] WindOn	Global	x3 percentile %95 value in time serie	Winter	1 time step	4 times per month	Spike
104	Aggregated Generation per Type [16.1.B&C] WindOn	Global	x1.5 percentile %95 value in time serie	Winter	1 time step	4 times per month	Spike
105	Aggregated Generation per Type [16.1.B&C] WindOn	Global	x3 percentile %95 value in time serie	Summer	10 time step	2 times per month	Plateau
106	Aggregated Generation per Type [16.1.B&C] WindOn	Global	x1.5 percentile %95 value in time serie	Summer	10 time step	2 times per month	Plateau
107	Aggregated Generation per Type [16.1.B&C] WindOn	Global	x3 percentile %95 value in time serie	Winter	10 time step	2 times per month	Plateau
108	Aggregated Generation per Type [16.1.B&C] WindOn	Global	x1.5 percentile %95 value in time serie	Winter	10 time step	2 times per month	Plateau
109	Aggregated Generation per Type [16.1.B&C] WindOn	Contextual	Random values in min-max range (month)	Summer	1 day	1 time	Daily

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

110	Aggregated Generation per Type [16.1.B&C] WindOn	Contextual	Random value in min-max range (month)	Winter	1 day	1 time	Daily
111	Aggregated Generation per Type [16.1.B&C] WindOn	Contextual	Random value in min-max range (month)	Summer	1 day	2 times per month	Daily
112	Aggregated Generation per Type [16.1.B&C] WindOn	Contextual	Random value in min-max range (month)	Winter	1 day	2 times per month	Daily
113	Aggregated Generation per Type [16.1.B&C] WindOn	Collective	Random sort of daily values	Summer	1 day	1 time	Daily
114	Aggregated Generation per Type [16.1.B&C] WindOn	Collective	Random sort of daily values	Winter	1 day	1 time	Daily
115	Aggregated Generation per Type [16.1.B&C] WindOn	Collective	Random sort of daily values	Summer	1 day	2 times per month	Daily
116	Aggregated Generation per Type [16.1.B&C] WindOn	Collective	Random sort of daily values	Winter	1 day	2 times per month	Daily

(*) Syntax of the value is x times the percentile of the time series (ie x2 percentile %95 value, means that the outlier value will be 2 times the percentile %95 calculated over the specific time interval of the time series.)

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

Supported type of outliers are the most common outlier types identified in energy time series are:

- **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.

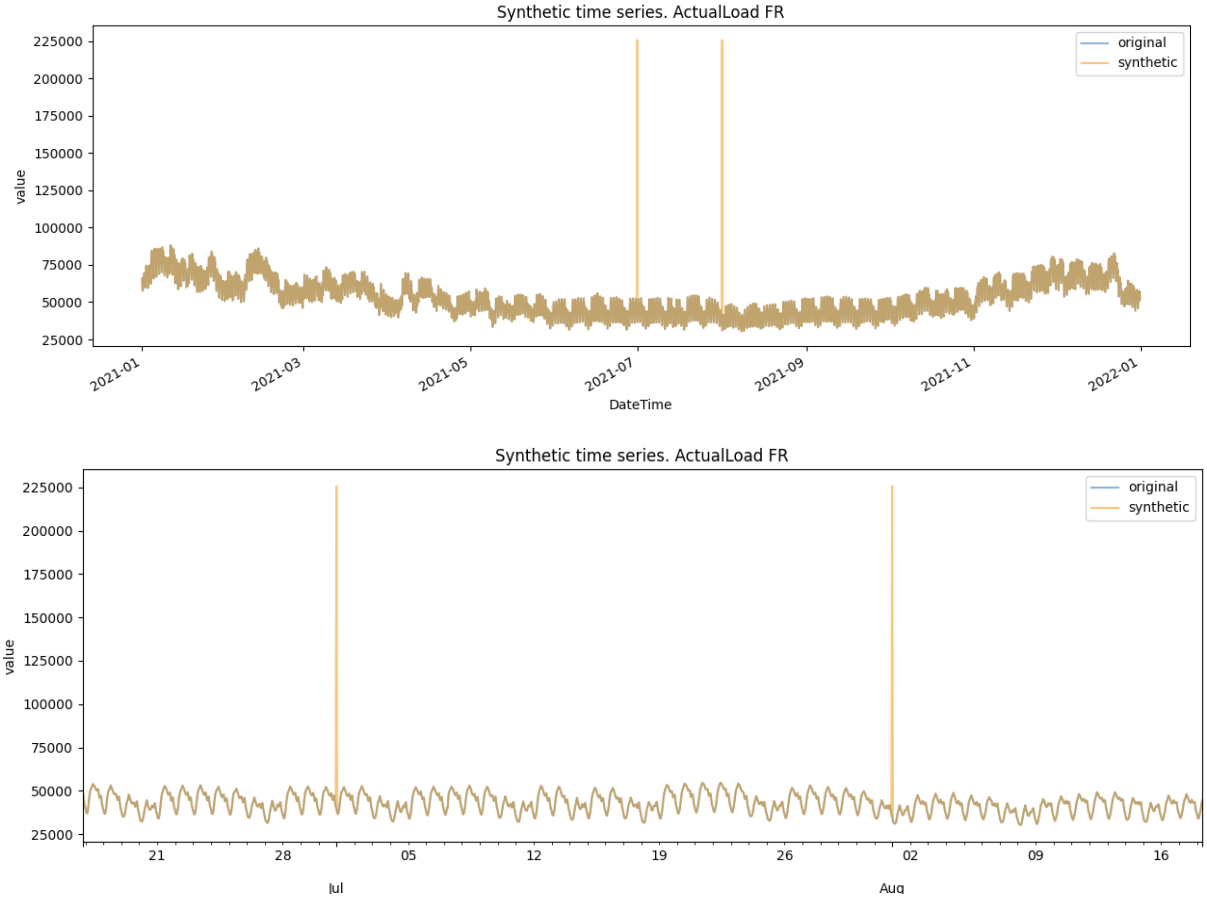
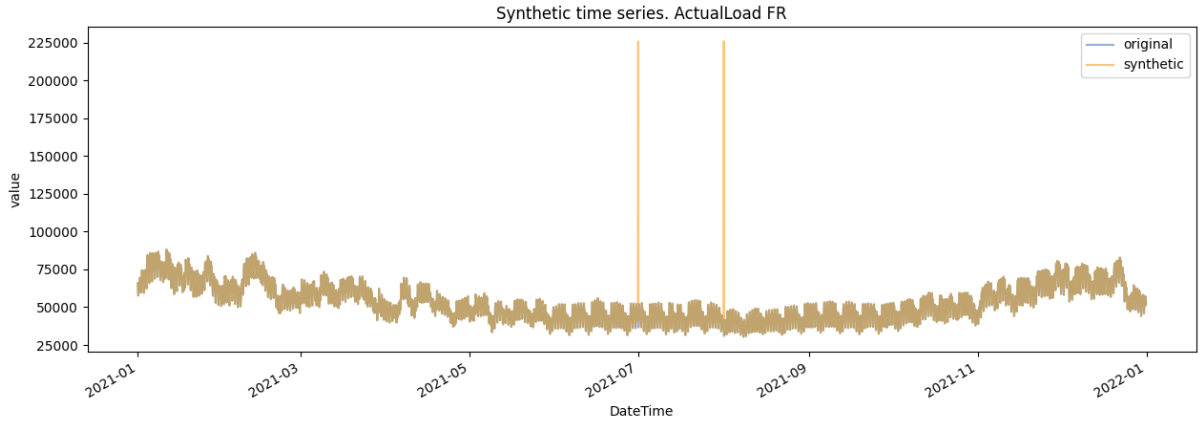


Figure 2.3.11 - Global outlier . Spike



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

Figure 2.3.12 - Global outlier . Plateau

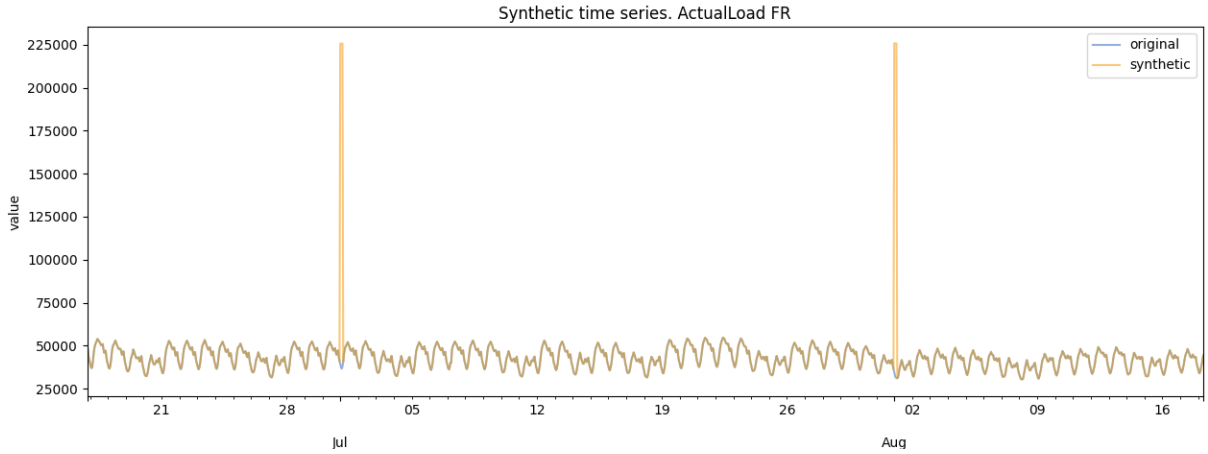


Figure 2.3.13 - Global outlier . Plateau

- **Contextual outliers.** Contextual outliers are data points whose value significantly deviates from other data within the same context. The "context" is almost always temporal in time-series data, such as records of a specific quantity over time. Values are not outside the normal global range, but are abnormal compared to the seasonal pattern. See example in Figure 2.3.14, 2.3.15, 2.3.16, 2.3.17

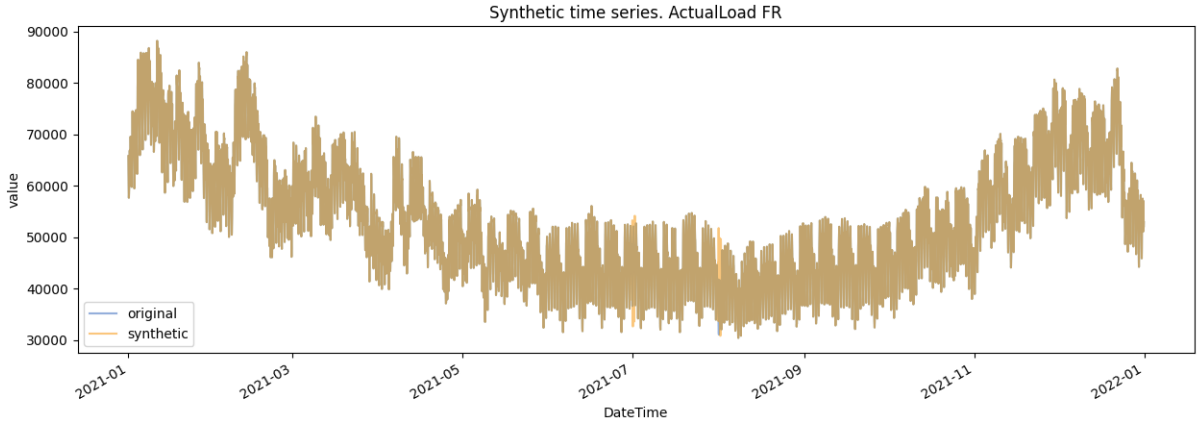


Figure 2.3.14 - Contextual outlier 2 day

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

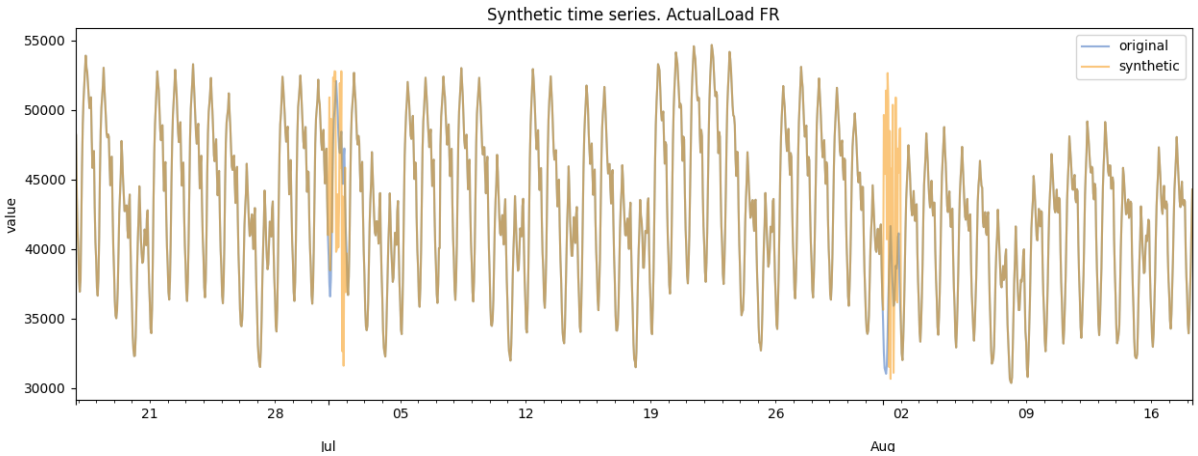


Figure 2.3.15 - Contextual outlier 2 day

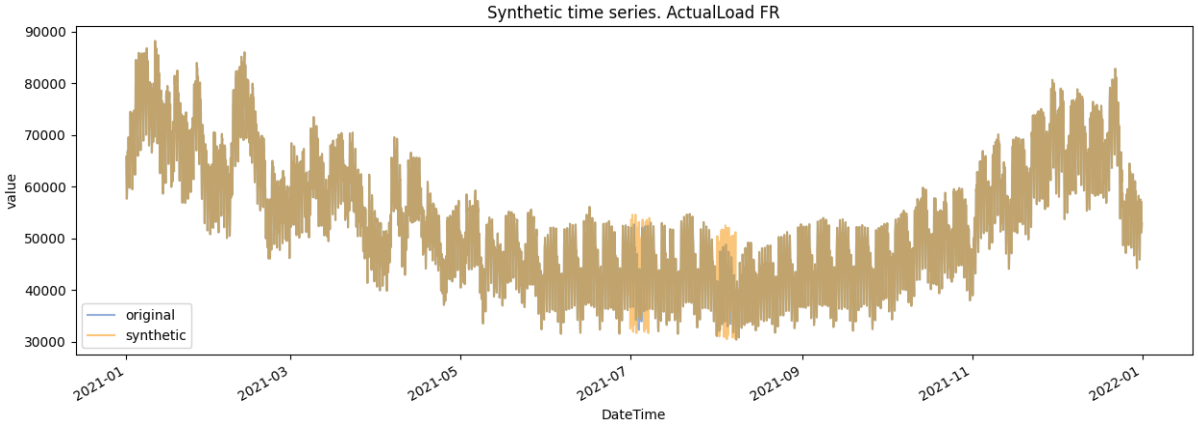


Figure 2.3.16- Contextual outlier 2 week

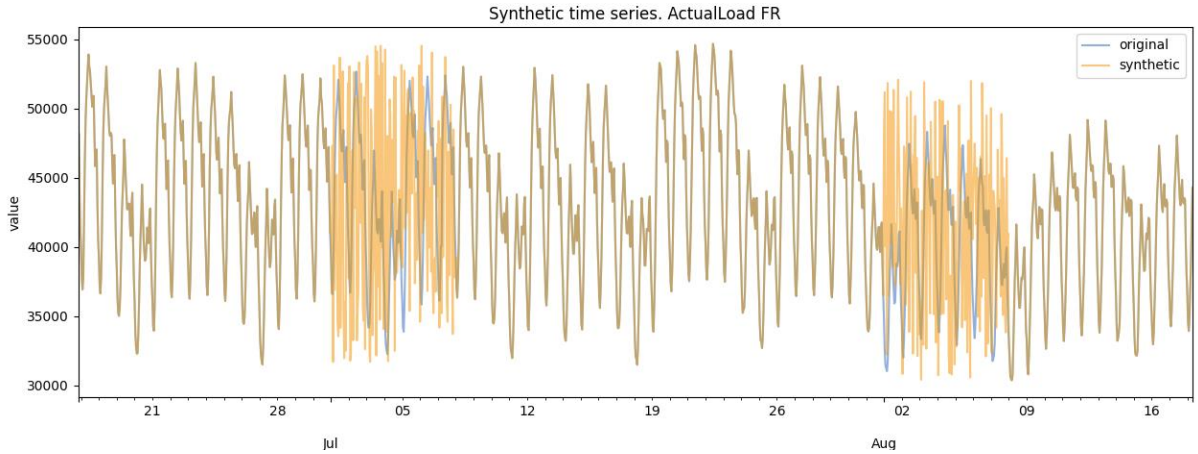


Figure 2.3.17 - Contextual outlier 2 week

- **Collective outliers**. A subset of data points within a data set is considered anomalous if those values

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

are considered as a collection which deviates 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. In time series data, one way this can manifest is as normal peaks and valleys occurring outside of a time frame when that seasonal sequence is normal or as a combination of time series that is in an outlier state as a group. See some examples in Figure 2.3.18, 2.3.19, 2.3.20, 2.3.21

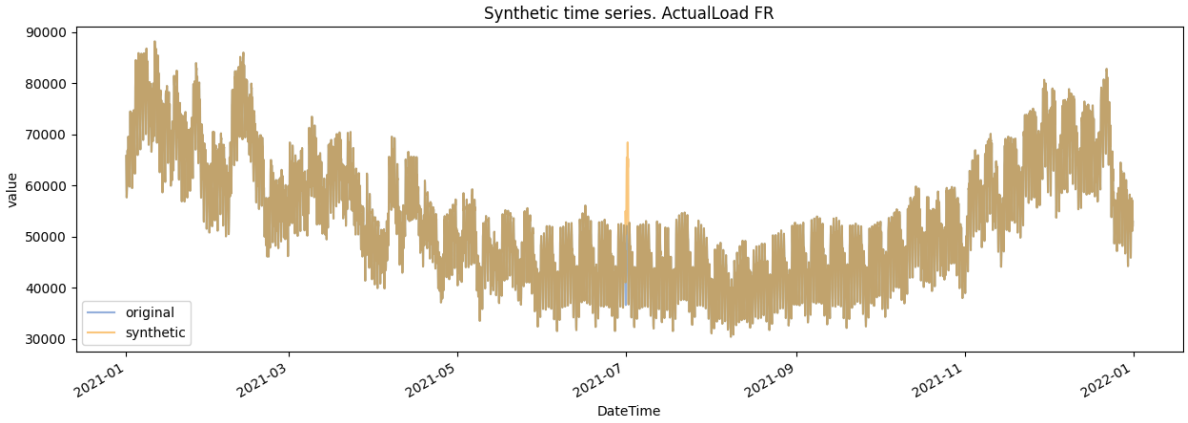


Figure 2.3.18- Collective outlier 1 day

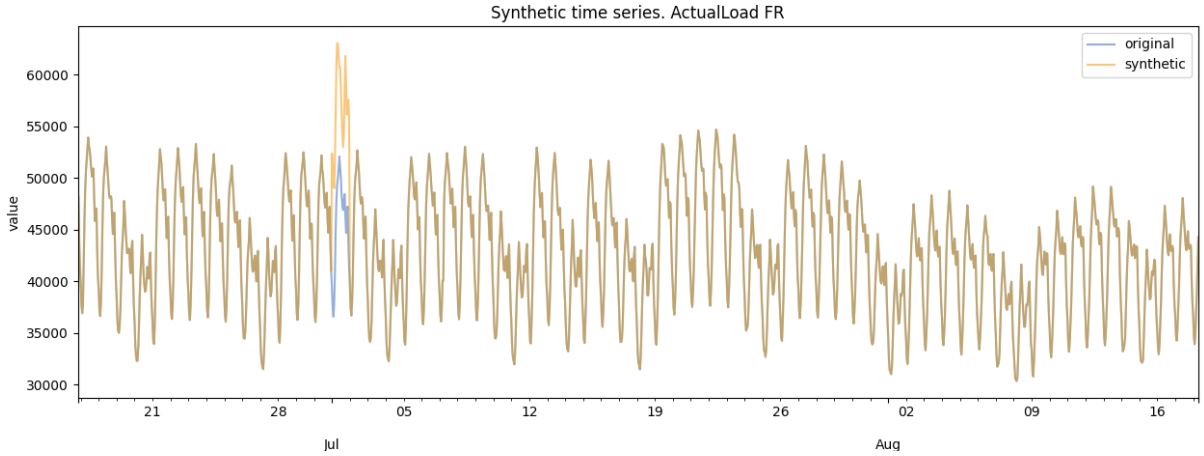


Figure 2.3.19- Collective outlier 1 day

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

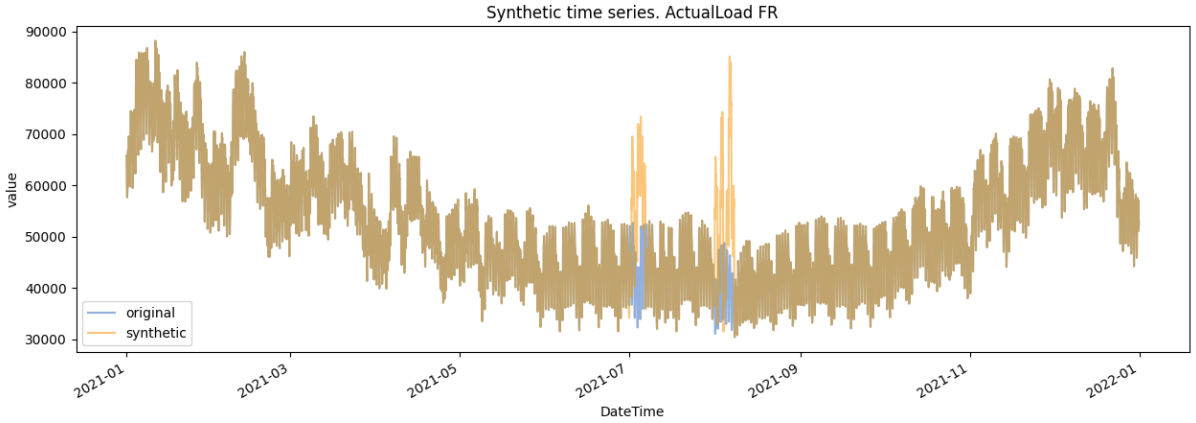


Figure 2.3.20- Collective outlier 2 week

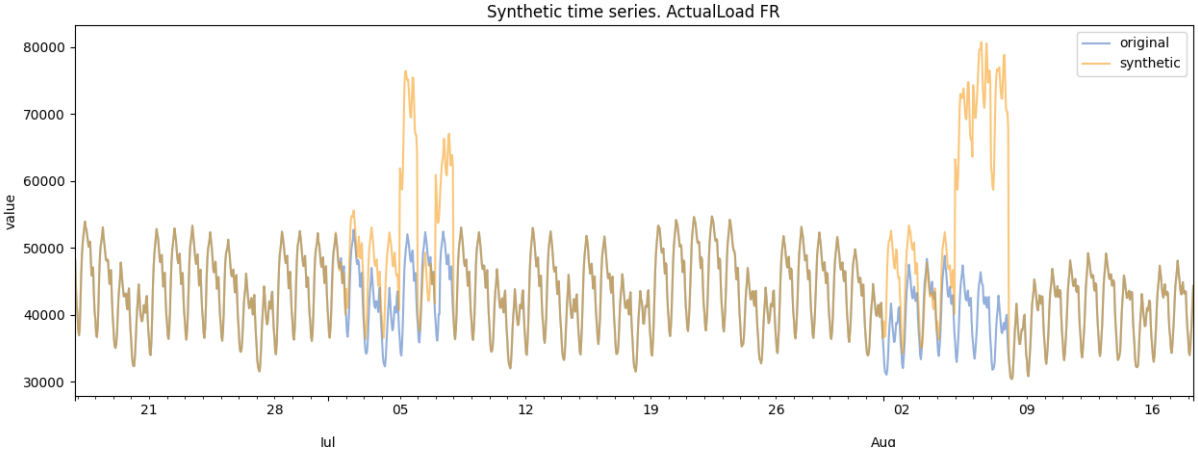


Figure 2.3.21- Collective outlier 2 week

2.3.1.3 Task 3.2a Define multiple subsampling criteria [M2]

Objectives of the task

Data used to train and test the model is subsampled using different window lengths. One single model is obtained per each of the window lengths. The criteria used to decide the number of models and window lengths depends on the specific time series properties and type of outliers to be detected. Smaller windows are used to detect short term outliers. Larger windows are used to detect mid-long term outliers. A weighted average of the multiple subsampling scores are used to obtain a single score. One specific issue to take care of is COVID-19. Different window lengths are going to be evaluated considering pre-COVID and COVID periods.

Summary of progress towards objectives and description of the work performed

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

Multiple window lengths have been analyzed in data exploration in order to identify some kind of pattern to be exploited by the model. As it was expected the load has typical daily and weekly seasonality, also annual patterns related to weather. At the same time, renewable generation also has a daily and annual pattern related to the weather and the availability of the renewable resource. Forecasted Day-ahead Transfer CapacitiesTotal Capacity Nominated series have a less clear pattern as the other series because they have multiple dependencies with high variation. The conclusions obtained from the analysis of the time series have been used in the model development.

2.3.1.4 Task 3.2b. Model training Big data. [M2-M3]

Objectives of the task

Train subset is subsampled considering previous criteria. A single model is trained per each of the window lengths. Different window lengths and contamination factors are evaluated in order to pick the best model so criteria. The evaluation methodology has been described at the concept section of the initial proposal.

Summary of progress towards objectives and description of the work performed

Two different approaches have been implemented to tackle different kind of outliers:

- Ensemble isolation forests. Outlier detection method focused in points and plateau outliers
- Pattern based. Outlier detection method focused on collective and contextual outliers

Ensemble isolation forests

The method is based on using ensemble isolation forests to calculate a weighted outlier score per sample. Isolation forest method builds a tree structure using randomly sub-sampled data. The number of splits required to isolate samples is equivalent to the path length from the root node to the end node. The averaged path length is the measure of normality. Samples with higher path length are less likely to be anomalies and samples with lower path length are more likely to be anomalies as they require less splits to be isolated. The ensemble isolation forests obtain multiple isolation forests for different sub-sampling windows. Different history windows are used to capture different dynamics in history. An outlier score is obtained per each of the criteria and weights are applied in order to obtain the final outlier score.

Pattern based

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

The method is based on using a pattern based approach in order to identify abnormal patterns in data. The abnormal patterns in data that are considered outliers or anomalies or errors or noise or faults or defects. Clustering is the core of the pattern based approach. In the first stage, data is preprocessed in order to characterize each of the days in the data in terms of absolute or relative daily profile. The process of clustering involves grouping of objects with more similarity. The inter similarity between the clusters is very less. The process of clustering is an unsupervised problem as the class labels for the data points are not known before. The core of the clustering approach is based on the Gaussian Mixture Model (GMM). A mixture model is a probabilistic model for representing the presence of subpopulations within an overall population, without requiring that an observed data set should identify the sub-population to which an individual observation belongs. Formally a mixture model corresponds to the mixture distribution that represents the probability distribution of observations in the overall population. However, while problems associated with "mixture distributions" relate to deriving the properties of the overall population from those of the sub-populations, "mixture models" are used to make statistical inferences about the properties of the sub-populations given only observations on the pooled population, without sub-population identity information. The GMM model is a probabilistic model that assumes all the data points are generated from a mix of gaussian distributions with unknown parameters. The final outlier score is obtained by identifying low density clusters which are supposed to represent abnormal patterns or daily patterns that are not similar in terms of distance to any of the identified clusters.

2.3.1.5 Task 3.3 Model training small data. Subsampling criteria [M3]

Objectives of the task

The same procedure for training used in task3.2 I also used for training of small data. Adam is used as a neural network optimizer. The default Tukey fence is implemented. Additional network architecture and Tukey fence tuning could be done in case of bad evaluation results.

Summary of progress towards objectives and description of the work performed

The method is based on obtaining a baseline model which represents the common dynamics of a time series and comparing its backcasting results against measured time series. Differences between backcasted and measured time series are analyzed in order to score dissimilarity. High dissimilar points are considered outliers. We're currently working on the analysis of current LSTM autoencoders implementation in order to be used as the core of the baseline models.

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

The proposed scoring is based on the differences between backcasted and measured time series via is done using Tukey fences 0. Outliers are values below Q_1-3(Q_3-Q_1) or above Q_3+3(Q_3-Q_1) or equivalently, values below Q_1-3 IQR or above Q_3+3 IQR.

2.3.1.6 Task 3.4 Model testing Big data and small data. [M4]

Objectives of the task

Model testing is done evaluating train model prediction on testing data subset. For the case of small data, model testing is done evaluating train model prediction on standardized/normalized data subset. Normalization/standardization is using training properties (mean, variance, etc.). The evaluation methodology is described at the Evaluation subsection of the Concept section.

Summary of progress towards objectives and description of the work performed

Benchmarking has been done to compare the results of the outlier detection algorithm against other methods used in the industry.

- Local Outlier Factor (LOF). The LOF algorithm is an unsupervised anomaly detection method which computes the local density deviation of a given data point with respect to its neighbors. It considers as outliers the samples that have a substantially lower density than their neighbors. The number of neighbors considered (parameter n_neighbors) is typically set 1) greater than the minimum number of samples a cluster has to contain, so that other samples can be local outliers relative to this cluster, and 2) smaller than the maximum number of close by samples that can potentially be local outliers [7]
- Median Absolute Deviation (MAD). The Median Absolute Deviation is a robust measure of the variability of a univariate sample of quantitative data. It can also refer to the population parameter that is estimated by the MAD calculated from a sample. For a univariate data set X1, X2, ..., Xn, the MAD is defined as the median of the absolute deviations from the data's [8]

$$\tilde{X} = \text{median}(X)$$

$$\text{MAD} = \text{median}(|X_i - \tilde{X}|)$$

- One-Class Super Vector Machine (SVM). The One-Class SVM algorithm is a variation of the SVM that can be used in an unsupervised setting for anomaly detection. A SVM constructs a hyper-plane or set of hyper-planes in a high or infinite dimensional space, which can be used for classification, regression or other tasks. Intuitively, a good separation is achieved by the hyper-plane that has the largest distance to the nearest training data points of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier [9]. The one-class SVM finds a hyper-

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

plane that separates the given dataset from the origin such that the hyperplane is as close to the data points as possible.

The results of the outlier detection are displayed below:

Global spike and plateau outliers:

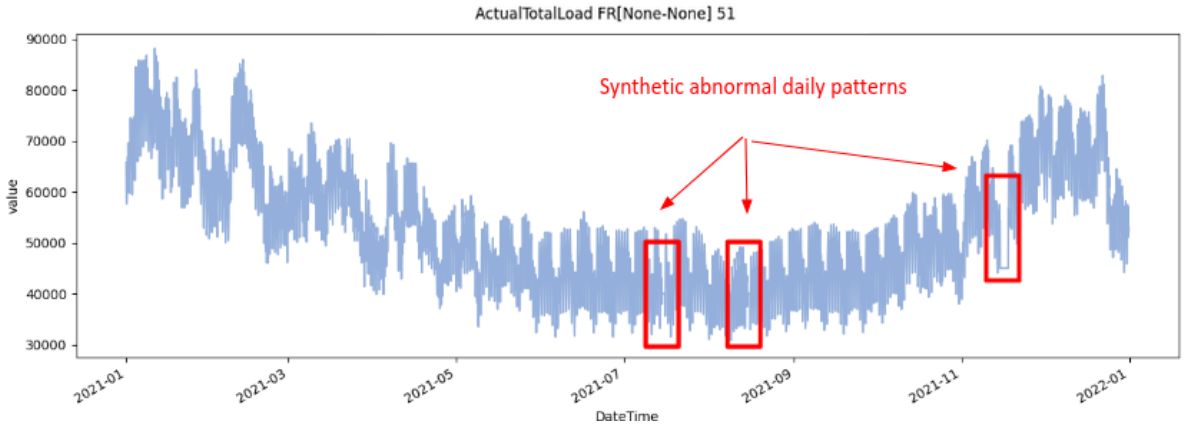


Figure 2.3.22- Synthetic abnormal daily patterns

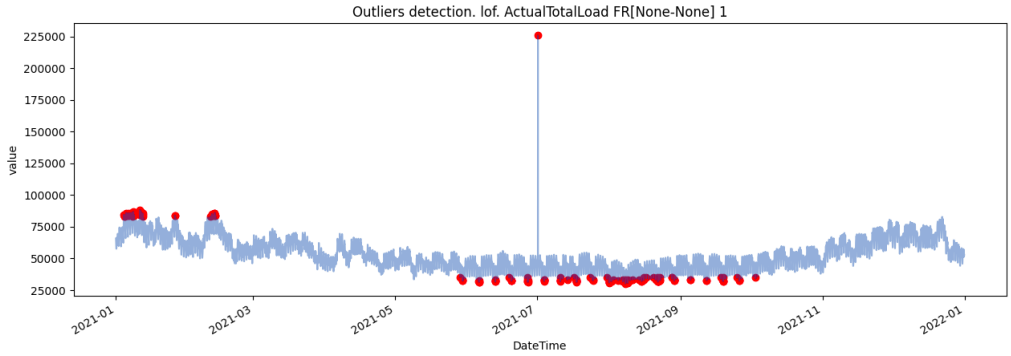
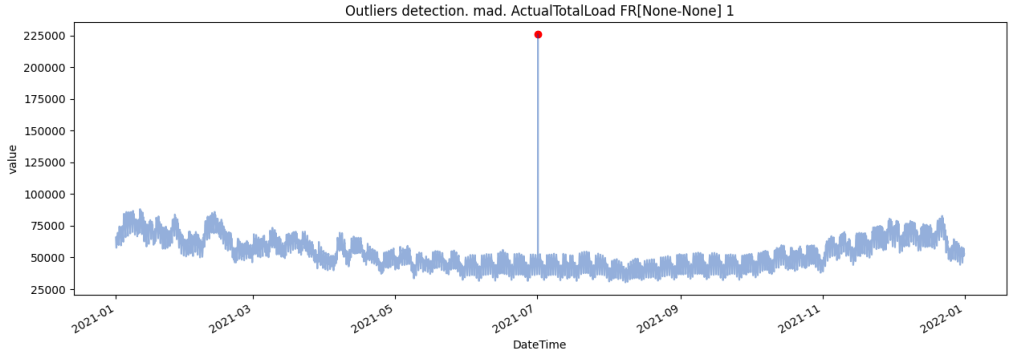


Figure 2.3.23- Outliers detected by LOF method



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

Figure 2.3.24- Outliers detected by MAD method

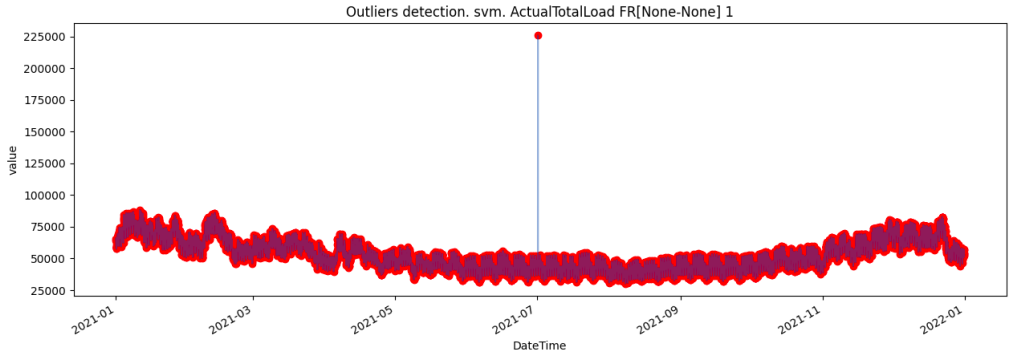


Figure 2.3.25- Outliers detected by One-Class SVM method

(implementation review pending)

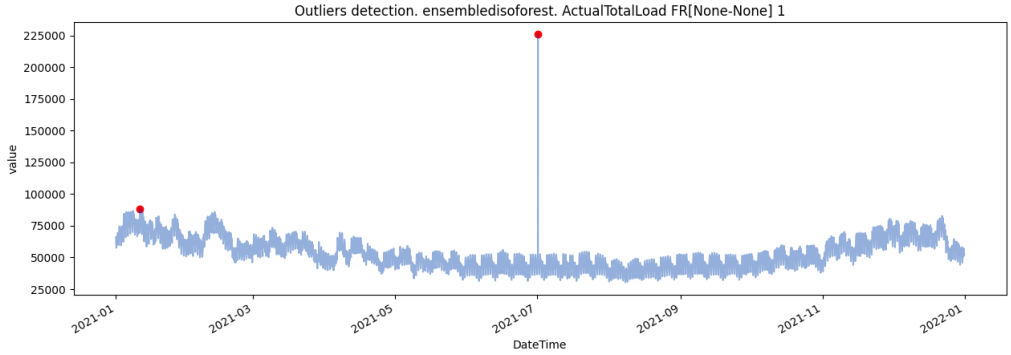


Figure 2.3.26- Outliers detected by EnsembleIsolationForest method

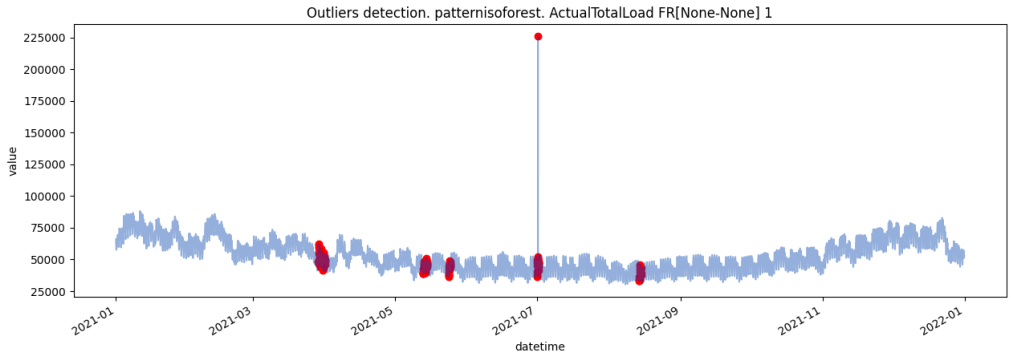


Figure 2.3.27- Outliers detected by PatternBased method (abnormal days)

False positives are being manually checked to confirm they're true false positives or they're true positives.

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

Contextual and collective outliers:

Actual Total Load FR

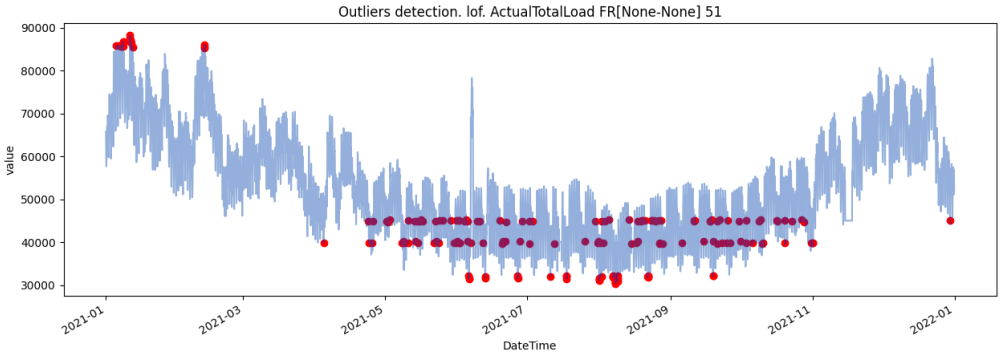


Figure 2.3.28- Outliers detected by LOF method

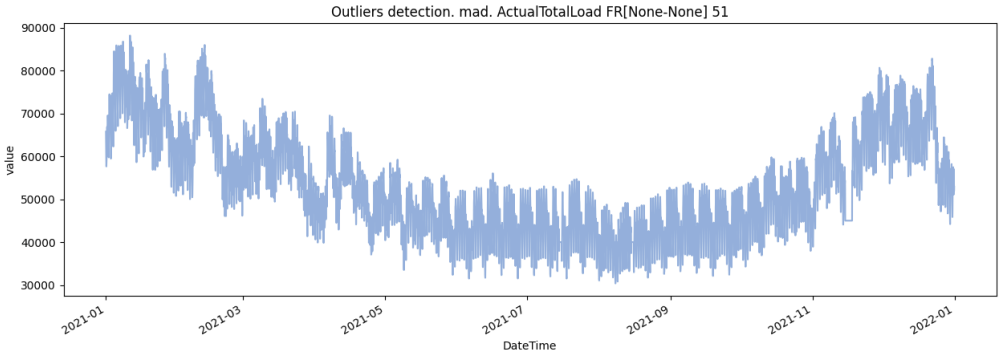


Figure 2.3.29- Outliers detected by MAD method

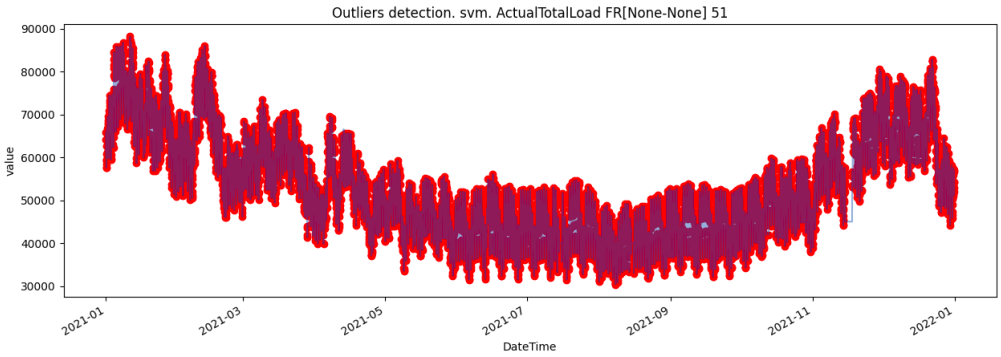


Figure 2.3.30- Outliers detected by One-Class SVM method

(implementation review pending)

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

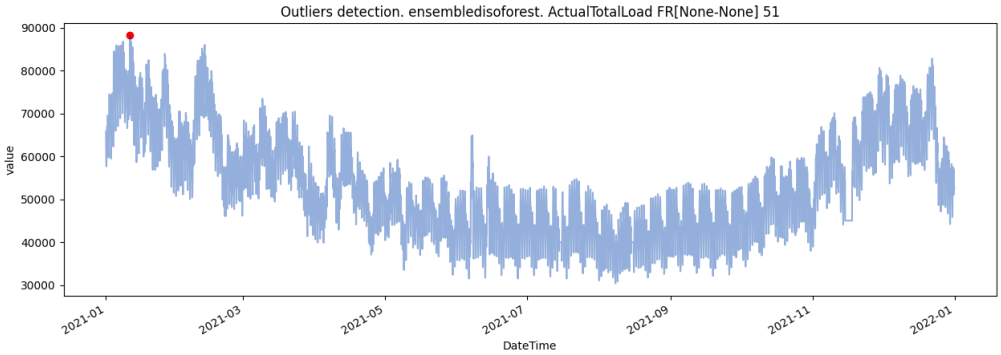


Figure 2.3.31- Outliers detected by EnsembleIsolationForest method

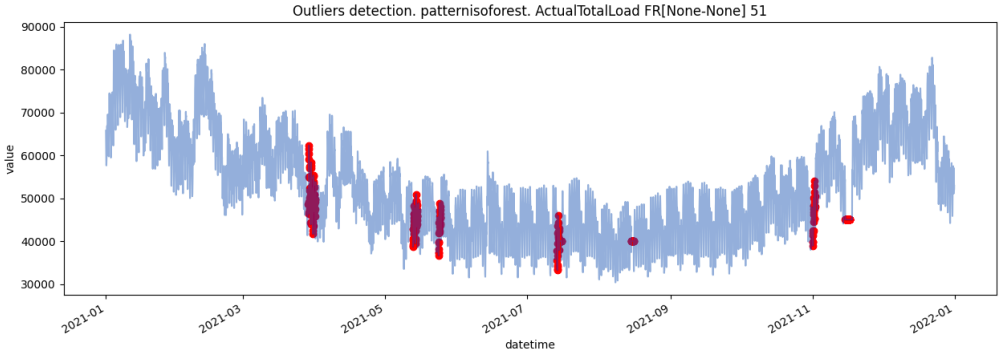


Figure 2.3.32- Outliers detected by PatternBased method (abnormal days)

TotalCapacityNomainated FR_CH

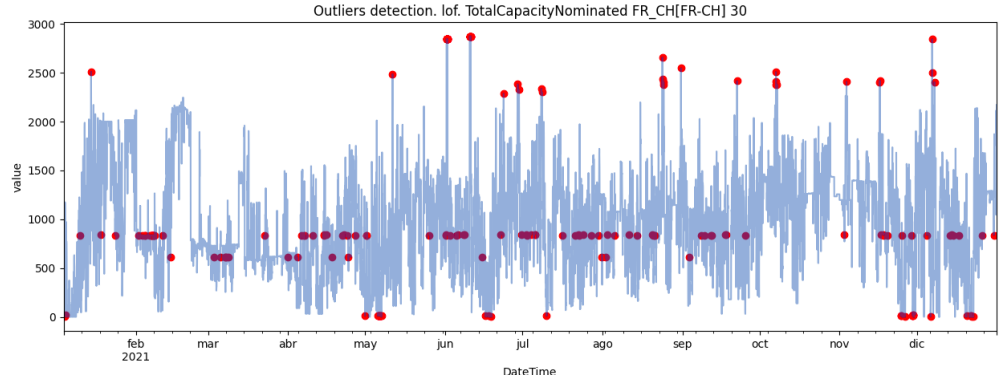


Figure 2.3.33- Outliers detected by LOF method

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

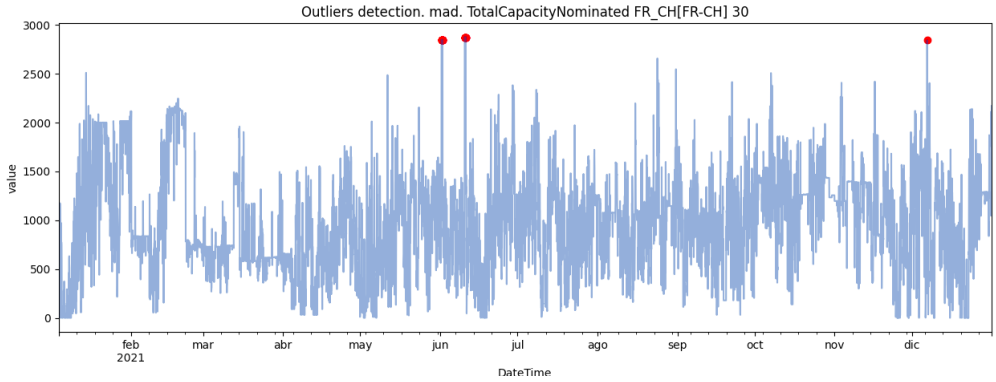


Figure 2.3.34- Outliers detected by MAD method

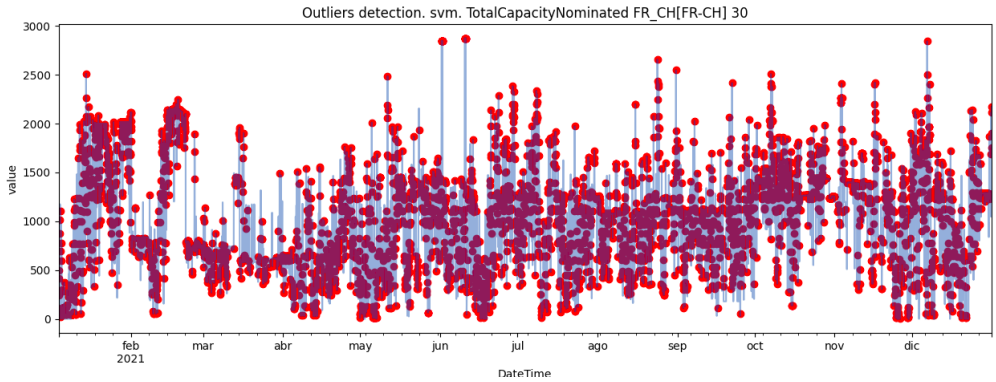


Figure 2.3.35- Outliers detected by One-Class SVM method

(implementation review pending)

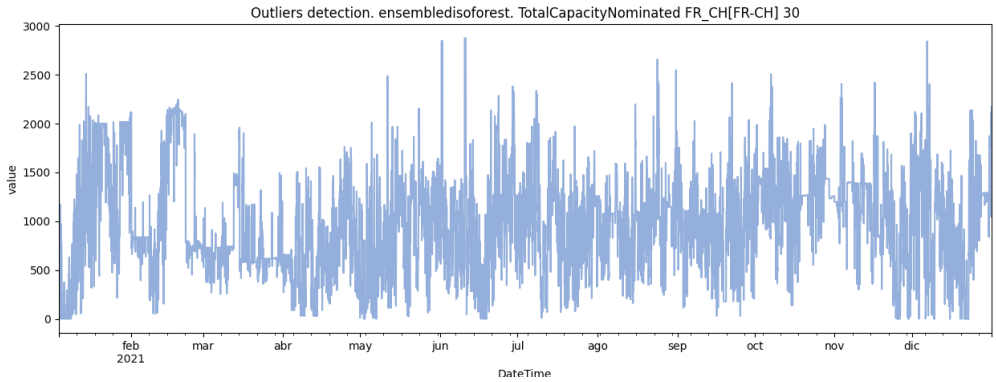


Figure 2.3.36- Outliers detected by EnsembleIsolationForest method

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

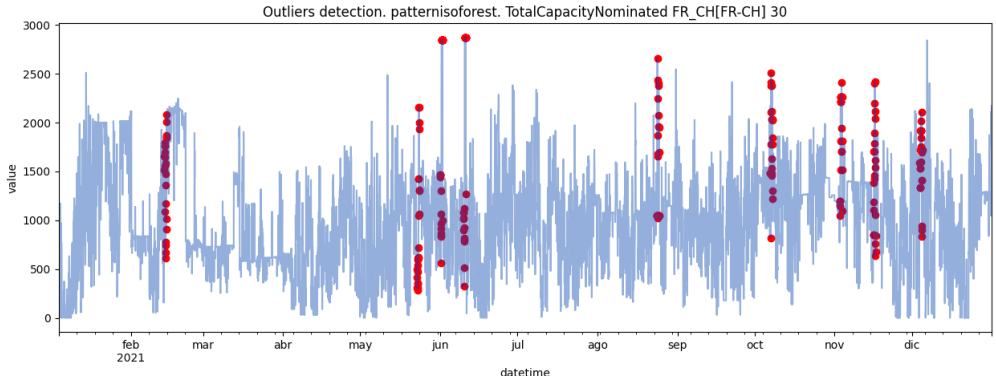


Figure 2.3.37- Outliers detected by PatternBased method (abnormal days)

First benchmarking has been done using 15 synthetic scenarios. See results below:

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 (*)

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

(*) Abnormal date is detected

Conclusions:

- 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

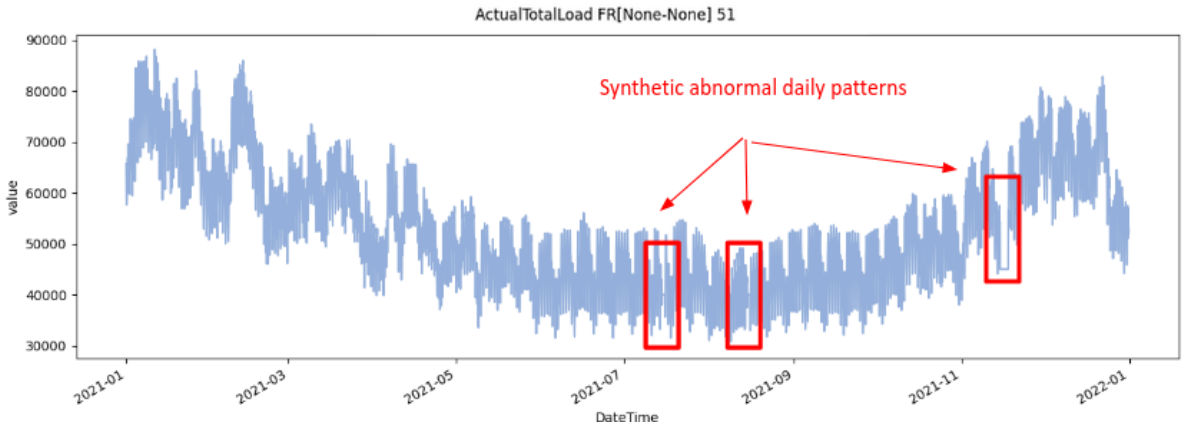


Figure 2.3.38- Outliers detected by PatternBased method (abnormal days)

2.4 Explanation of the work carried in T4. Build Imputation method - Big data and small data model

Objectives of the task

To build a big data module able to drop or fill with appropriate values, the missing or outlier values detected in previous tasks. The objective is to integrate two different tools in the process of outliers detection, both for big and small data samples.

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

2.4.1.1 Task 4.1. Data Pre processing [M3]

Objectives of the task

Same procedure described in #task 3.1 is implemented, followed by a next step where the dataset is splitted between train and test subsets. In both cases time series data and available labeled outliers are splitted. Training subset is splitted between training subset and validation. Already identified outliers, labeled outliers, are removed from the train time series to prevent model contamination.Data standardization/normalization is done to help kNN. Scaling all features to a common scale gives each feature an equal weight in euclidean distance calculations.

Summary of progress towards objectives and description of the work performed

Same data processing has been done as in Task 3.1

2.4.1.2 Task 4.2 . Model training and evaluation. [M3-M5]

Objectives of the task

Train subset is subsampled considering previously introduced tasks. A single model is trained per each of the window lengths. Different window lengths and k are evaluated in order to pick the best model so criteria. Model train is done using standardized/normalized train subset and model evaluation is done using standardized/normalized evaluation subset . The evaluation methodology is described at the Evaluation section.

Summary of progress towards objectives and description of the work performed

The big data imputation method has been implemented. The method is based on a variation of the kNN regression method over subsampled data. The subsampling criteria is mainly related to recent similar days in terms of calendar and load profile. Calendar similarity is based on labor/non-labour property or weekdays. Daily load profile similarity is calculated using euclidean distance between partial available load and partial load of all the other days. The ones with low euclidean distance between daily profiles are picked as similar profile days. The closest labor or non-labour days with a similar load profile are used by kNN method to evaluate neighbors' similarity.

Similarity is used as a weight of the contributions of the neighbors, The imputation result is the weighted average of the neighbor’s value. The weight is 1/d where d is the distance to the neighbor. Some of the methodology

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

settings like calendar criteria or number of neighbors can be tuned specifically per each of the 7 domains at the Transparency platform.

2.4.1.3 Task 4.3. Model testing [M4-M5]

Objectives of the task

Objective of testing for imputation is different than for detection, but the model is the same: baseline model. In one case it is used to identify the values "out of range" with respect to the baseline (outliers) and in the other case it is used for the prediction of the gaps (imputation). For testing and evaluation of imputation two steps are implemented: I) Evaluation data subset is used to evaluate the model, and tune it, and ii) Test data subset is used to blindly evaluate the model and is the one that ends up defining the accuracy of the model.I

Summary of progress towards objectives and description of the work performed

Imputation evaluation has been implemented in order to obtain imputation quality. First analysis has been done of the response of the imputation method in synthetic scenarios. See the imputation results under initial imputation testing scenarios:

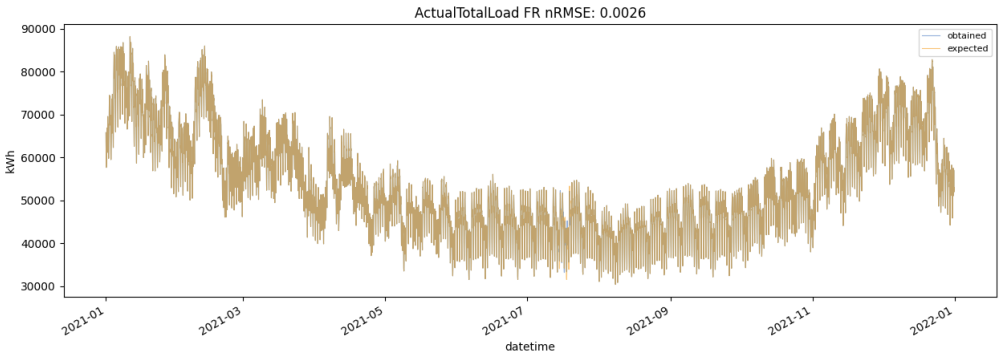


Figure 2.4.1- Imputation results in ActualLoad time serie (summer)

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

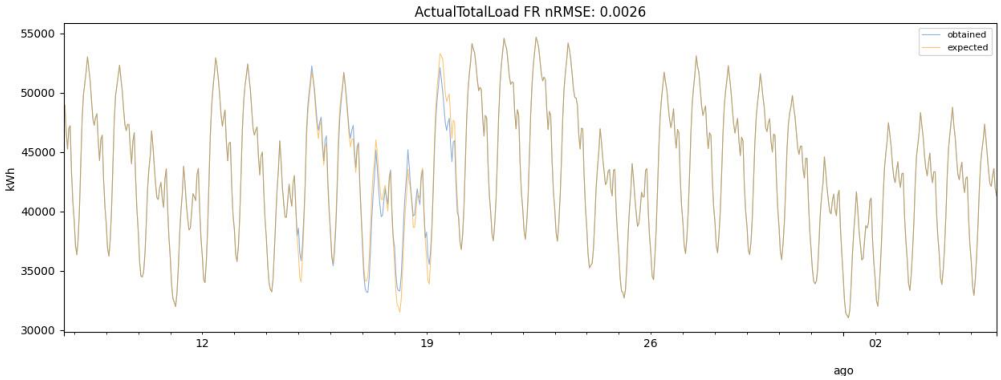


Figure 2.4.2- Imputation results in ActualLoad time serie zoom (summer)

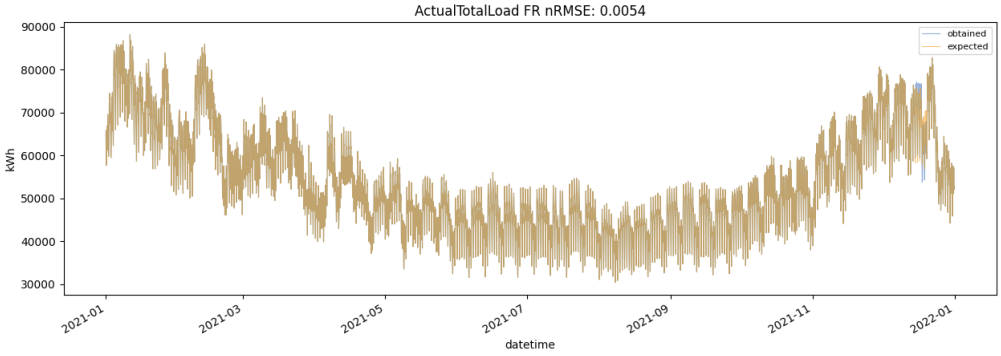


Figure 2.4.3- Imputation results in ActualLoad time serie (winter)

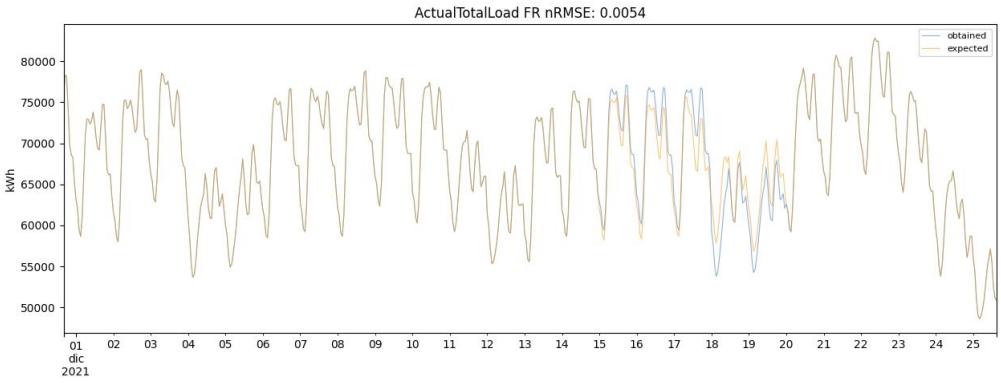


Figure 2.4.4- Imputation results in ActualLoad time serie zoom (winter)

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

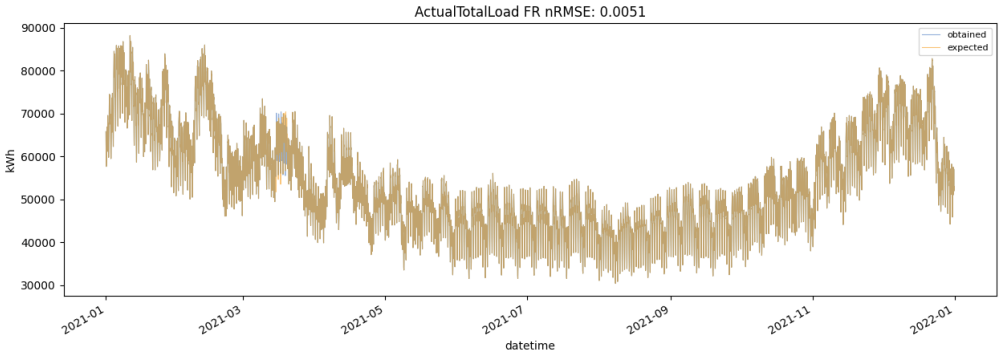


Figure 2.4.5- Imputation results in ActualLoad time serie (spring)

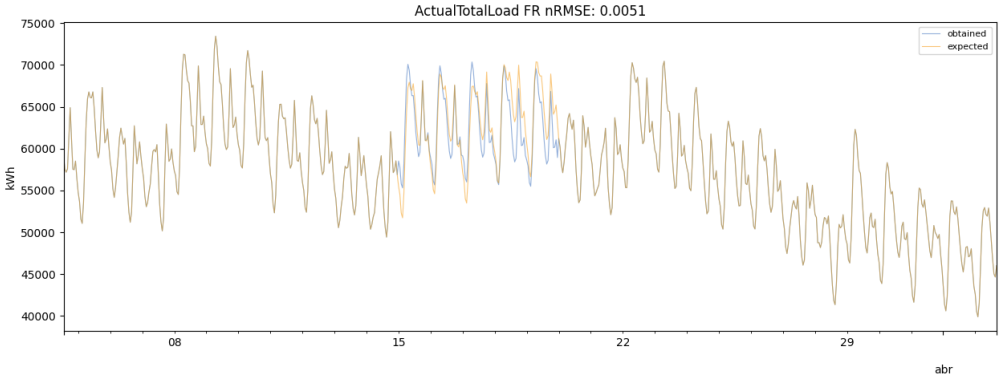


Figure 2.4.6- Imputation results in ActualLoad time serie zoom (spring)

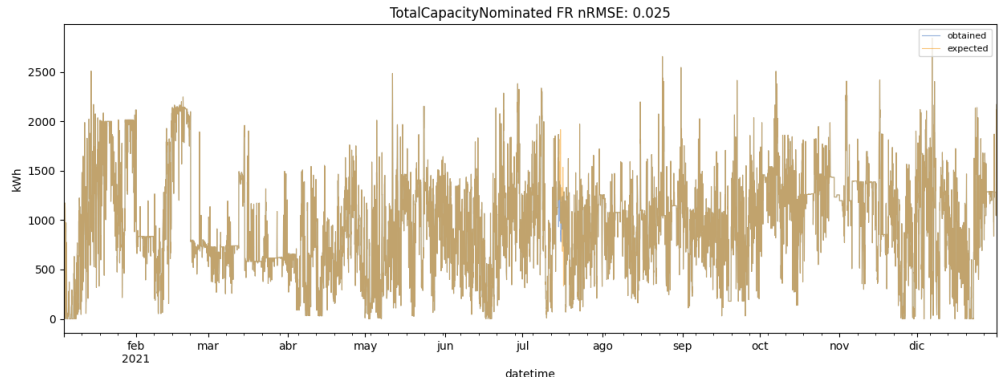


Figure 2.4.7- Imputation results in TotalCapacityNominated time serie (summer)

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

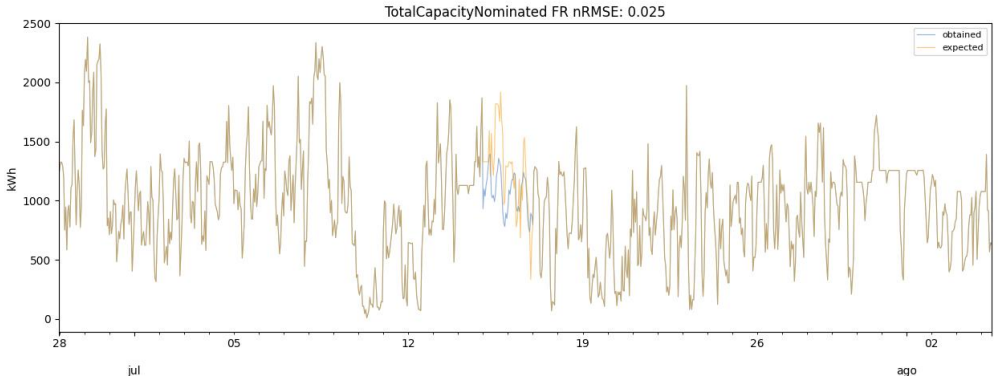


Figure 2.4.8- Imputation results in TotalCapacityNominated time serie zoom (summer)

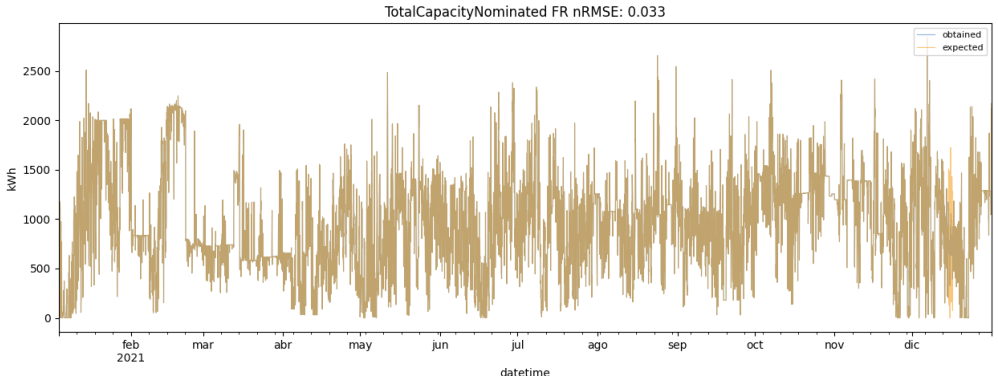


Figure 2.4.9- Imputation results in TotalCapacityNominated time serie (winter)

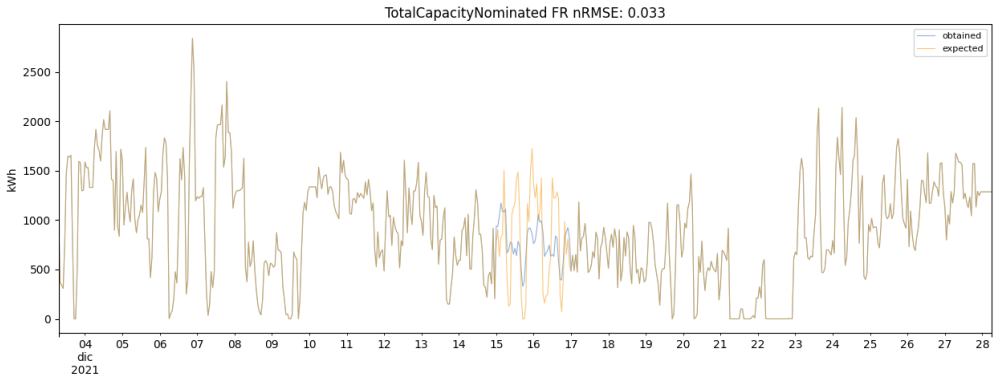


Figure 2.4.10- Imputation results in TotalCapacityNominated time serie zoom (winter)

Conclusions:

0 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 behavior 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 itself.

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

1 Imputation tuning (number of neighbors, lambda forgetting factor, etc) is required per each time series because of different behavior. Hyperparameter optimization could be used to tune imputation.

2 In cases when there are exogenous variables correlated with the time series it would be possible to consider them in the similarity analysis

3 In case of known daypart cyclic behavior in time series it would be possible to tune part of the day weighting in the similarity analysis

3 The impact of this subproject on OneNet and the general European Energy system

Data and analysis is increasingly becoming an integral part of everyday electricity system and more specific in data exchanges among TSOs, DSOs and consumers, which is the focus of the OneNet project, with a growing emphasis on data-led decision making across different organisations. Therefore, trust in the quality of data is vital. Outlier analysis plays an important role in maintaining this trust. Outliers can skew trends and forecasts modelled from data-sets, negatively impacting the quality and accuracy of decisions. If outliers are not identified and removed, models can become less accurate and effective. One of the impacts of our tool is minimizing risk in decisions and business processes, creating a more equitable system while boosting performance.

The increase of data quality, and the scalability of our proposal contributes to enabling interoperability and information exchange among the different actors of the energy ecosystem.

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

4 References

[1] OneNet. One network for Europe. https://onenet-project.eu/

[2] TSO – DSO – Consumer: Large-scale demonstrations of innovative grid services through demand response, storage and small-scale (RES) generation. H2020-LC-SC3-ES-5-2018-2020.

https://cordis.europa.eu/programme/id/H2020_LC-SC3-ES-5-2018-2020/es

[3] Building a Low-Carbon, Climate Resilient Future: Secure, Clean and Efficient Energy. H2020-LC-SC3-2018-2019-2020

https://clustercollaboration.eu/open-calls/building-low-carbon-climate-resilient-future-secure-clean-and-efficient-energy

[4] ENTSO-E. the European association for the cooperation of transmission system operators (TSOs) for electricity .https://www.entsoe.eu/

[5] European Distribution System Operators for Smart Grids (EDSO) . https://www.edsoforsmartgrids.eu/

[6] F1 score description https://en.wikipedia.org/wiki/F-score

[7] LOF introduction https://www.dbs.ifi.lmu.de/Publikationen/Papers/LOF.pdf

[8] MAD introduction https://en.wikipedia.org/wiki/Median_absolute_deviation

[9] SVM introduction https://scikit-learn.org/stable/modules/svm.html#mathematical-formulation

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

5 Glossary

Capacity. Capacity is the rated continuous load-carrying ability of generation, transmission, or other electrical equipment, expressed in megawatts (MW) for active power or megavolt-amperes (MVA) apparent power.

Decision Forest. A model created from multiple decision trees. A decision forest makes a prediction by aggregating the predictions of its decision trees. Popular types of decision forests include random forests and gradient boosted trees.

Decision Tree. A supervised learning model composed of a set of conditions and leaves organized hierarchically

Decoder. In general, any ML system that converts from a processed, dense, or internal representation to a more raw, sparse, or external representation. Decoders are often a component of a larger model, where they are frequently paired with an encoder.

Demand - Consumption. Demand is the rate at which electric power is delivered to or by a system or part of a system, generally expressed in kilowatts (kW) or megawatts (MW), at a given instant or averaged over any designated interval of time.

Encoder. In general, any ML system that converts from a raw, sparse, or external representation into a more processed, denser, or more internal representation. Encoders are often a component of a larger model, where they are frequently paired with a decoder.

Ensemble. A collection of models trained independently whose predictions are averaged or aggregated. In many cases, an ensemble produces better predictions than a single model. For example, a random forest is an ensemble built from multiple decision trees. Note that not all decision forests are ensembles.

Long Short-Term Memory. A type of cell in a recurrent neural network used to process sequences of data in applications such as handwriting recognition, machine translation, and image captioning. LSTMs address the vanishing gradient problem that occurs when training RNNs due to long data sequences by maintaining history in an internal memory state based on new input and context from previous cells in the RNN.

Precision. A metric for classification models. Precision identifies the frequency with which a model was correct when predicting the positive class.

Recall. A metric for classification models that described how many did the model correctly identify out of all the possible positive classes.

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