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

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

Company: Beedata Analytics SL

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
DSO	Distribution System Operators
ENTSO-E	European Network of Transmission System Operators for Electricity
ES	Spain
FR	France
GMM	Gaussian Mixture Model
kNN	K-Nearest Neighbor
LOF	Local Outlier Factor
MAD	Median Absolute Deviation
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

Short Description

Context

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 in the project we expect to improve the quality of the data managed by the data provider.

The AI data quality toolbox will have two main tasks: Identification of erroneous data and imputation of erroneous data. In order to implement these modules it's mandatory to use the data from the provider to identify use cases and properly train the core models.. In this deliverable we provide evaluation results of the outlier detection and imputation methods (big data).

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) , association of grid operators in Europe, and complementary on smart grids data from other use cases.

Challenges

Our proposal addresses the "Scenario No 6.: Advanced Data Quality Analysis of Data Exchange Platforms", by implementing an actively monitoring and performing outlier detection to flag errors in time series data to act as an early indicator for abnormality in the databases of the Transparency Platform, in the power system analysis, the effective collaboration between actors, and in the implementation of flexibility services.

The main challenge to be addressed is to design and implement a tool for data quality measurements divided in two main steps; i) Identification of missing or erroneous readings, ii) Imputation of missing or erroneous readings.

Proposed solutions

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

Outlier detection methods

Two different data scenarios areconsidered for the outlier detection methods and different approach is proposed per each scenario:

A) Big data scenario: Big number (>5K) of single high variable time series. That would be the case of the load or generation from distribution grid through SCADA or energy management systems, and smart meters measurements of consumers and prosumers. The method is based on using Daily Patern based method to detect the abnormal patterns in data that are considered outliers or anomalies or errors or noise or faults or defects. In this case the method is based on using a daily pattern based approach in order to identify abnormal dates in the time series. Clustering is the core of the pattern based approach. The core of the clustering approach is based on the Gaussian Mixture Model (GMM). The result is the ability to work in a context where the size of training sample grows as time went on, leading to more training time, more computation resources, failing to detect outliers on time.

B) Small data scenario: Small number of time series data. That would be the case of national load, generation, and the other time series data from the Transparency Platform, as well as or small business cases from the other scenarios (<5K). The method is based on obtaining a baseline model based on LSTM autoencoders, which is a self-supervised method based on neural networks. Our neural network anomaly analysis is able to flag the upcoming bearing malfunction well in advance of the actual physical bearing failure by detecting when the data readings begin to diverge from normal operational value.

Description of the imputation method

Each of the time series have specific domain properties and outliers. Although all the time series are energy domain related, each of them has different dynamics depending on different factors. The dynamics of time series can depend on economics, weather, logistics, etc. The used imputation 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 neighbors value. The weight is 1/d where d is the distance to the neighbor.

Each of the stages have different customization settings. Detection method customization is used to properly fit the method to the different kinds of time series provided by ENSTO-E.

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

Evaluation results

Outlier detection evaluation

Synthetic outliers are used as the data provided by ENTSO-E is not already classified and has no kind of label to be used as outlier identification. Synthetic data is an industry workaround to evaluate scenarios which are in the domain knowledge but with no available data. Synthetic outliers are created to evaluate the model under outlier scenarios not present in data. Outliers are domain specific but typical outlier patterns are:

- Spikes
- Plateaus
- Null values
- Anomalous patterns

The results of the evaluation method previously described is presented as benchmarking analysis. Recall will be the (from 0 to 1) main indicator and accuracy will be analyzed in each specific caseBenchmarking 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.
- Median Absolute Deviation (MAD). The Median Absolute Deviation is a robust measure of the variability of a univariate sample of quantitative data.

Table 2.4.1.1 - Specification of the outlier scenarios

Time Serie	Type of Outlier	LOF	MAD	PatternBased
ActualTotalLoad	global_spike_plateau	1	1	1
ActualTotalLoad	contextual	0.04	0	0.8
ActualTotalLoad	collective	0	0	0.95
AggregatedGenerationPerType_NUCLEAR	global_spike_plateau	1	1	0.89
AggregatedGenerationPerType_NUCLEAR	contextual	0.12	0	0.72
AggregatedGenerationPerType_NUCLEAR	collective	0	0	0.27
AggregatedGenerationPerType_SOLAR	global_spike_plateau	0.4	1	0.9
AggregatedGenerationPerType_SOLAR	contextual	0.04	0.6	0.99
AggregatedGenerationPerType_SOLAR	collective	0	0.44	0.85

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

ForecastedDayAheadTransferCapacities	global_spike_plateau	1	1	1
ForecastedDayAheadTransferCapacities	contextual	0.82	0	0.62
ForecastedDayAheadTransferCapacities	collective	0	0	0.52
TotalCapacityNominated	global_spike_plateau	1	1	0.51
TotalCapacityNominated	contextual	0.03	0	0
TotalCapacityNominated	collective	0.01	0	0.41

Imputation evaluation

See the imputations results in Table 2.4.2.1:

Table 2.4.2.1 - Specification of the outlier scenarios

TimeSerie	Type of gap	nRMSE
ActualTotalLoad	five_days	0.00787
ActualTotalLoad	partial_days	0.00207
ActualTotalLoad	single_days	0.00355
ActualTotalLoad	single_hours	0.00063
AggregatedGenerationPerType_NUCLEAR	five_days	0.00485
AggregatedGenerationPerType_NUCLEAR	partial_days	0.00170
AggregatedGenerationPerType_NUCLEAR	single_days	0.00242
AggregatedGenerationPerType_NUCLEAR	single_hours	0.00048
AggregatedGenerationPerType_SOLAR	five_days	0.04497
AggregatedGenerationPerType_SOLAR	partial_days	0.01425
AggregatedGenerationPerType_SOLAR	single_days	0.02627
AggregatedGenerationPerType_SOLAR	single_hours	0.00677
ForecastedDayAheadTransferCapacities	five_days	0.00422
ForecastedDayAheadTransferCapacities	partial_days	0.00036
ForecastedDayAheadTransferCapacities	single_days	0.00094
ForecastedDayAheadTransferCapacities	single_hours	0.00000
TotalCapacityNominated	five_days	0.06122

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

TotalCapacityNominated	partial_days	0.02585
TotalCapacityNominated	single_days	0.03851
TotalCapacityNominated	single_hours	0.00855

Conclusions

Conclusions regarding outlier detection:

- Manual curation of positives is required in order to calculate precision so F1-score
- Accuracy is highly related to the properties of the time series. As it was expected, high stochastic time series have worse results.
- Accuracy in global spike and plateau outliers is quite good in all methods
- Detection in contextual and collective outliers is only done by the pattern based algorithm. LOF could work under some specific time series properties
- Pattern based algorithms detects potential (false/true) positives which should be manually reviewed
- Automatic hyperparameter tuning must be used to improve the results once labeled data is available
- Hyperparameter tuning can be used to generate different kinds of outlier signals. Soft and strict hyperparameter tuning can be used to create warning and severe outliers
- Additional extra-information like holidays or special days could be added to pattern based method in order to improve clustering quality

Conclusions regarding imputation:

- Deviation is highly related to the properties of the time series. Best results are obtained in low stochastic time series and worse results are obtained in high stochastic time series
- Partial day gaps and single hour gaps are best predicted as they provide daily pattern contextual information to the imputation method
- Automatic hyperparameter tuning must be used to improve the results once outlier labeled data or extra information on gaps is available
- Additional extra-information like holiday, specials days or specific domain specific data (weather, market info, time serie correlation, …) could be added to improve prediction

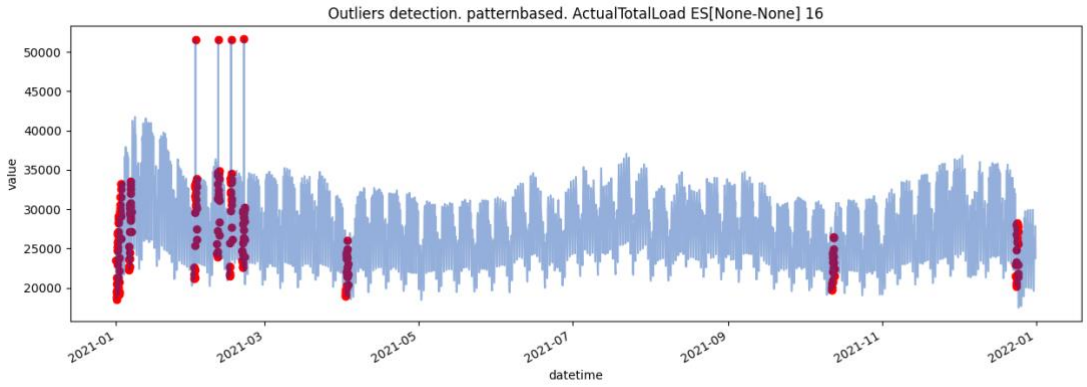
Figures

The results per each kind of time series introduced in the evaluation description are displayed below. The red dots are the predicted outliers detected in the time series by the outlier detection.

ActualTotalLoad.global_spike_plateau

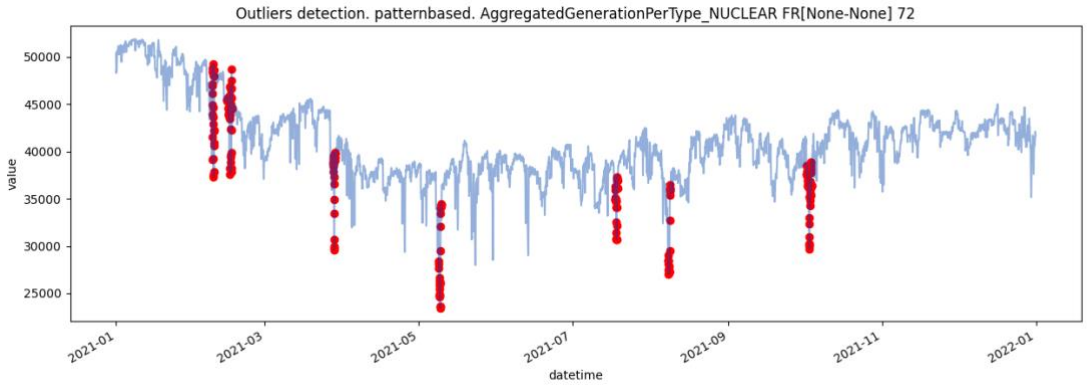
Daily-pattern based

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



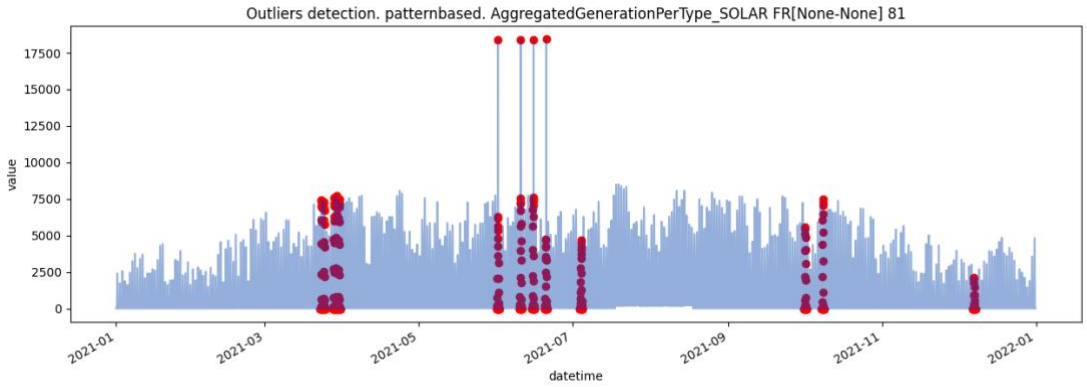
AggregatedGenerationPerType_NUCLEAR.contextual

Daily-pattern based



AggregatedGenerationPerType_SOLAR.global_spike_plateau

Daily-pattern based

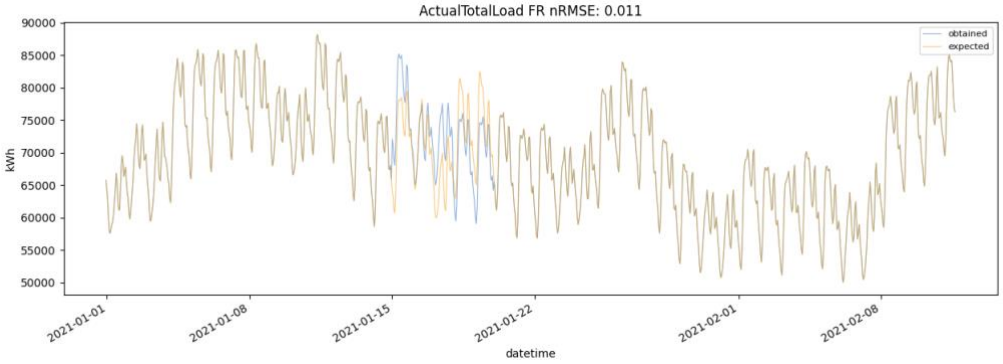


Imputation evaluation results

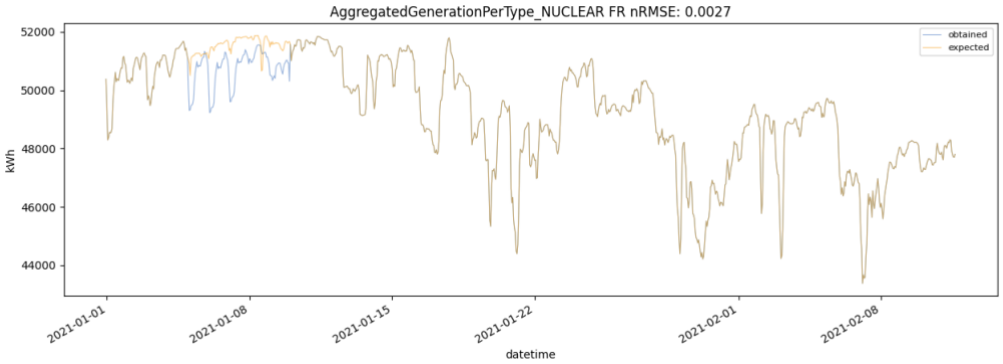
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 957739

The results per each kind of time series introduced in the evaluation description are displayed below. The blue time series corresponds to obtained imputation results and the orange time series corresponds to expected imputation results, so the original time series.

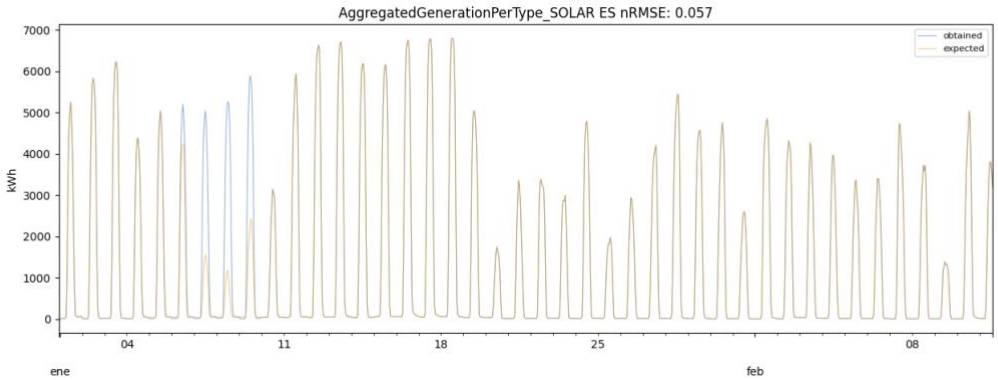
ActualTotalLoad.five_days



AggregatedGenerationPerType_NUCLEAR.five_days



AggregatedGenerationPerType_SOLAR.five_days



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