Group
Alba Ordoñez
Charles Thérond
Ioan Catana
Karine Petrus
Stéphane Mulard

A dashboard to explore solar farm performances

Final Report - 28 juin 2019

1. INTRODUCTION

In coherence with our previous Project Milestones, we have decided to create a dashboard to explore and analyze the performance of the solar farm that we have studied for Engie during our Fil Rouge project. The gitHub with the code developed for this project can be found in: https://github.com/igr204engie/DataViz_Engie.

2. REMINDER : PRESENTATION OF THE DATA

We chose to work with a dataset from Engie, the French energy company. This data is made of multivariate time series related to the production of electricity from a French solar farm located in the South of France (Blond, Haute-Vienne). Four of us have already been working with this dataset as part of our “Fil Rouge” project with Engie.

The main goal of our “Fil Rouge” project, and what makes this data interesting, is to **understand the performance of this solar farm and in particular identify, detect, analyze and categorize the different forms of anomalies** occurring throughout the time period.

The work we carried out for the “Fil Rouge” enabled us to gain many interesting insights about this data and get a better understanding of the factors involved in performance issues. We were able to create a predictive model for the normal or “expected” behaviour of a given zone of the farm, based on a set of environmental and inner characteristics. We then used this model to compare the actual values observed with the expected values, and from there identify the abnormal values and in particular the under-performances. Last we worked on categorizing the different kinds of abnormal behaviour, with a view to automating their detection.

The solar farm of Blond is made of approximately 25,000 photovoltaic solar panels, the production of which is aggregated into 8 independent inverters. Each inverter actually corresponds to a specific zone of the farm. The electricity produced at the inverter level is measured every minute. Practically what we got from Engie and what we have been working on is the following :

- **8 independent signals coming from the 8 zones (inverters)**, representing the electrical production every minute of the day, during a period of 1.5 year, from the 31st of May 2017 to the 4th of November 2018. Together with the production, we also have a number of attributes for these inverters.
- **The corresponding meteorological data**, which is crucial to understand and explain the production of electricity : the solar irradiance, which is the main feature involved in the physical phenomenon, the temperature, the hydrometry, etc.

The original data was supplied as a list of ".csv" files that we significantly processed for errors, missing values, etc. and turned into "DataFrame" format (tabular format from the Python Pandas package).

One challenging aspect with this multivariate series of data is to display them so as to reveal the key information they carry about the performances of the production and the different types of anomalies. But they are many ways to approach the visualization.

The final dataset that we have created and that we are working on has the following characteristics:

- 104,416 samples, corresponding to measures every 5 minutes between 4am and 8pm during for about 500 days
- 10 main characteristics among about 40 characteristics or attributes available:
    - Production : electrical measures, surface temperatures (quantitative)
    - Meteo : irradiances (horizontal, tilt and corrected), temperatures, azimuth (quantitative attributes)
    - Anomalies / maintenance and events : days with labelled anomalies, type of anomaly. Note that there are mainly 4 types of anomalies that will be described further on in this report: short loss, long loss, late start and temporary failure (anomalies type are nominal attributes).
    - Other : day of the year, hour (quantitative attributes)
- 8 inverters corresponding to 8 different zones of the farm.

3. OVERVIEW OF THE DASHBOARD

3.1. Target users

During our assessments we have considered two kinds of users:

- **Business users** : this would be users without a technical background in Data Science, more concerned with an immediate operational usefulness of the dashboard. For these users the primarily aim is to visualize the data. That's why we provide an interface that shows the production of electricity and the existing anomalies over a given period.
- **Data scientists** : these users already have the knowledge and skills to understand complex visualizations as well as specific knowledge of photovoltaic power production. Typically this would be a member of the data science team at Engie, like the people supervising our project. For them, the primarily aim would be to gain a better understanding of the performance and especially the anomalies of the farm.

When we designed our dashboard, we kept the two options in mind, and although the visualizations remain relatively simple and easy to use for a business user (tab 1 and 2), we think the dashboard is better suited for data scientists. Following Tufte's design principles, we have tried to give to the user the greatest number of ideas in the shortest time with the least ink. Even if the user does not have extensive experience with visualization tools, the dashboard is relatively easy to understand as we have included some explanations when needed. With the developed interactions, the dashboard allows investigating many aspects of the data so that users can explore and share their findings with their colleagues.

3.2. Primary functionalities

The developed dashboard includes three levels of visualization for accomplishing mainly the following visualization tasks:

- **Characterize the days of productions** of the solar farm over a period of 2 months and for one of the 8 zones (also referred to as 'inverters') : we have added interactions that allow exploring this data and identifying possible days with anomalies. This design also helps understanding the daily production in a 3D

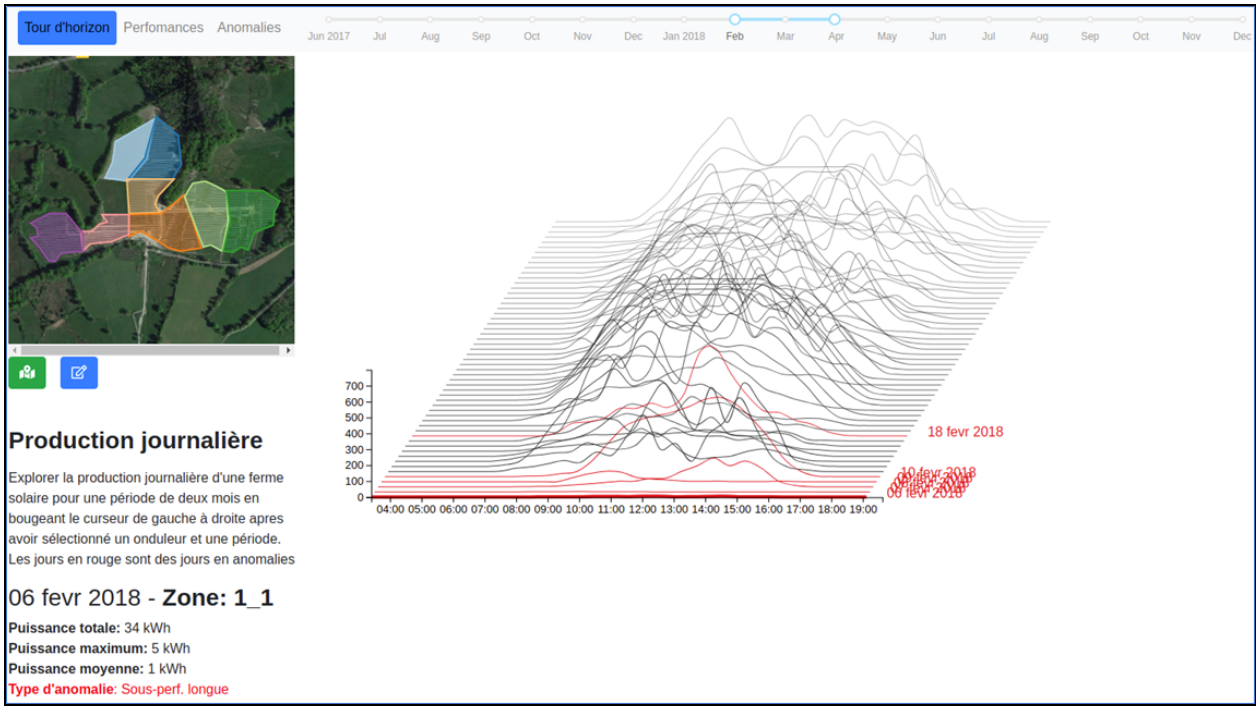
manner, with the third dimension representing several consecutive days. This allows some daily productions to be compared with others.

- **Characterize the performance** of the different zones of the solar farm over a chosen period of time**:** we have added interactions that allow selecting specific zones of the farm to explore if possible under performances have occurred. The user has actually the possibility of analyzing detailed weekly and daily views of the production in case some interesting behaviors are noticed.
- **Determine days with anomalies and evaluate if there were properly tagged as such by different algorithms**: we have added interactions that allow selecting different types of algorithms and letting the user evaluate whether they performed well (taking into account the F1 score metric).

3.3. Overview of the three tabs

The three primary functionalities presented above are covered in our dashboard which is decomposed into three different tabs: 'Tour d'horizon', 'Performances' and 'Anomalies'. The user can click to switch from one view to the other.

The screenshot below shows the first tab, when the application starts:



Note that the three tabs all share a *colored-coded clickable map* of the 8 zones of the farm to be able to select a specific zone. These zones have been shaped using the aerial photograph of the solar farm and its different zones). We have chosen on purpose color as perceptual feature in the visualization. Indeed, this feature can be discriminated preattentively to identify which zones of the farms are being analyzed.

3.4. Technical overview

For designing the dashboard, we have decided to use two programming languages:

- **Python**: It has the advantage of offering many packages for visualization. More specifically, our dashboard exploits the Dash package which is a layer of the Flask (web-app) and Plotly (graphic) packages. Dash gives the possibility to rapidly implement interactive graphs with a wide range of visualization interactions. It also facilitates the interaction with the javascript language.
- **Javascript**: The Javascript D3 library has been used to create features and interactions in our dashboards that were difficult to implement with Python.

The implemented dashboard is actually a web application and to accelerate its development, we have used the bootstrap framework (https://getbootstrap.com/) for the formatting and design of the different tabs.

For the readability and the maintenance of the code related to the web application, we have created a main program file that structures all the components of the dashboard (graphs and selectors) and imports the necessary libraries and other program files. Each graph is implemented in a separate program file that is either in Python or Javascript.

4. TAB 1 : Overview of the production during a period of time

4.1. Overview

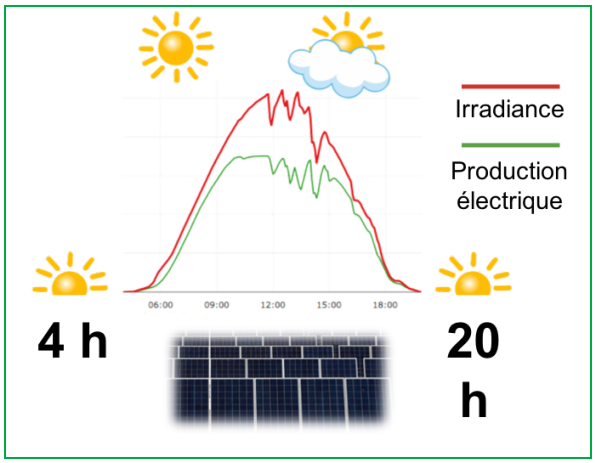
The main idea of this tab is to allow the user to navigate through the days of production of the solar farm, and more specifically to view the production of a single inverter over a period of time.

The production corresponds to the amount of electricity produced by the inverter with regards to the surrounding irradiance.

The best scale to use to understand this process is the daily production as the physical phenomenon is closely related to the rising and setting of the sun and the daily weather.

The image below shows a typical plot that can be observed during the day :

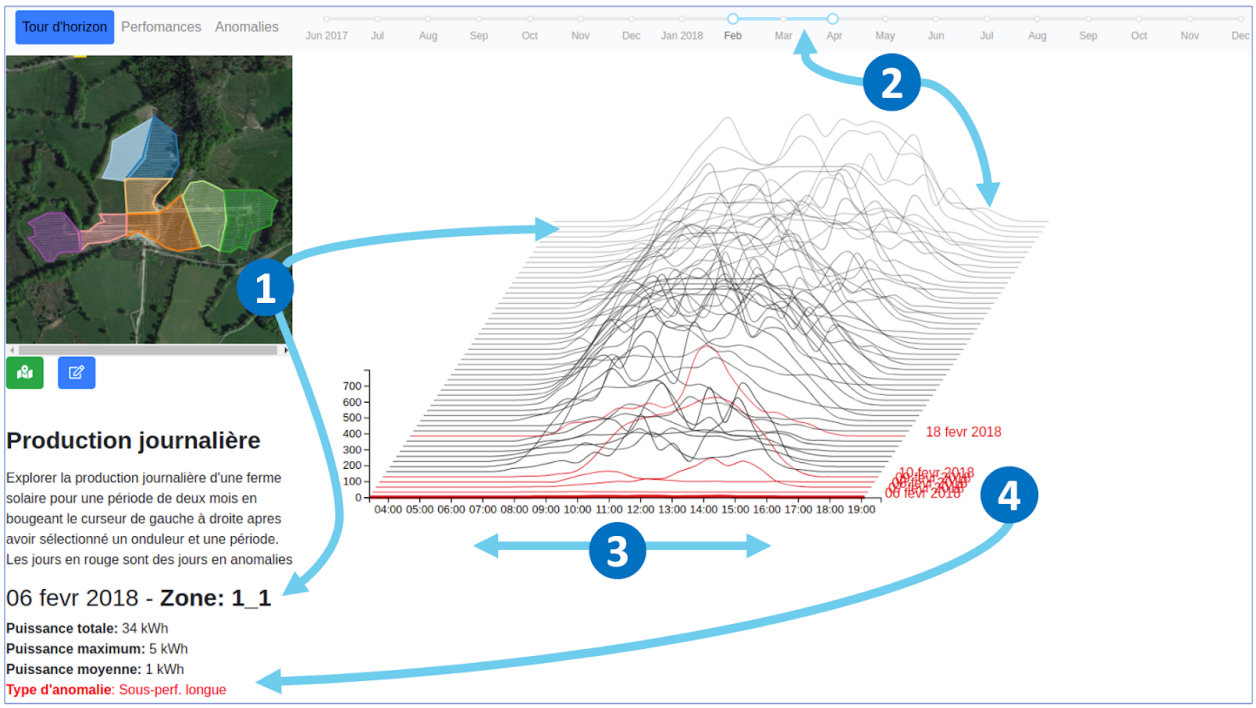
- As the sun rises in the morning, showed by the irradiance curve in red, the solar panels start producing electricity, shown by the green curve.
- At a certain amount of irradiance, the production reaches a plateau, which corresponds to the physical limit of a set of panels.
- During the day clouds and rain can occur, causing the irradiance to fluctuate, which in turn causes the production to fluctuate.
- Last, as the sun goes down, the amount of electricity produced decreases to zero.



4.2. Designed tab

Looking at different options to represent our production data, we came across an original way to visualize the history of production through a given period, demoed here : https://charts.animateddata.co.uk/uktemperaturelines/. This display concerns the UK temperatures. It is not easy to explain how it works without trying it, but basically when the user scrolls up or down, either using the trackpad of the PC, the wheel of the mouse or the side scroll bar of the browser, the production curves are displayed in front of each other with a slight offset. Thus going up and down is a very efficient way to navigate through the days. We got strong inspiration from the code used to display the UK temperatures which was written in D3-version 3. However, we re-wrote it from scratch using D3-version 4 and included interactions that were more adapted to our needs.

The result of the developed tab is shown below together with some numbered interactions between the graphs:



The interactions we included on this tab are the following:

- **Interaction 1**: The map selector lets the user see where the inverters are located and which inverter is already selected (by the opacity rate). When the pointer of a mouse is over an inverter, it automatically updates the history of production and the description associated with the selected zone.
- **Interaction 2**: The time slider allows selecting a period of 2 months going from June 2017 to December 2018 (last day of data is November, the 4th). The selection of the period automatically updates the history of the production on the selected period.
- **Interaction 3**: Moving the mouse cursor from left to right allows navigating through the days of production across the selected time period.
- **Interaction 4**: By placing the mouse cursor on a given day there is an update of the characteristics description for this day. Useful information such as the total, the maximum and the average productions are displayed. In addition, if the selected day is an anomaly, the type of anomaly is displayed.

4.3. Interesting insights

By analyzing the history of production during the winter season (more specifically the months of February and March 2018), we figured out that there were 5 consecutive days that were tagged as anomalies over the entire solar farm. Over this time period, the production was almost null for the different zones. Further investigation of the solar irradiance showed that the sun activity was relatively strong during those days. Thus, how could we explain the absence of production of the solar panels? Actually, we checked the weather forecast in the region some days before that period of time. It turned out that there was some strong snow falls that possibly covered the solar panels explaining therefore the lack of production despite of very shiny days. Moreover, the analysis of the history of production during the winter season also allowed to detect that it was from the end of February that the production started increasing again to reach interesting exploitation levels for the energy company Engie.

4.4. Possible improvements

The interesting insights that were found from the history of production required checking the solar irradiance level and the weather forecast. Including this information in the description of the characteristics of the days is something to be considered in the future.

5. TAB 2 : Comparing performances between inverters

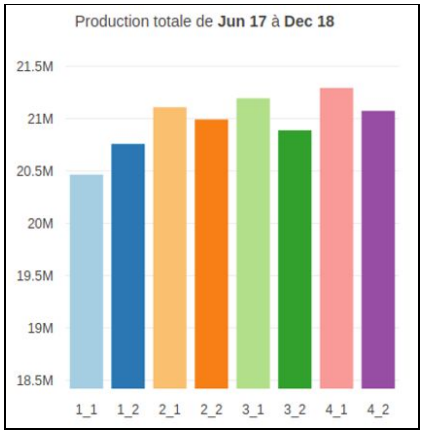
5.1. Overview

The main objective of this tab is to compare the performances of all or a selection of inverters over a given period of time. In particular, users should be able to compare the overall performance of a set of inverters over the entire period but also at a weekly and daily level.

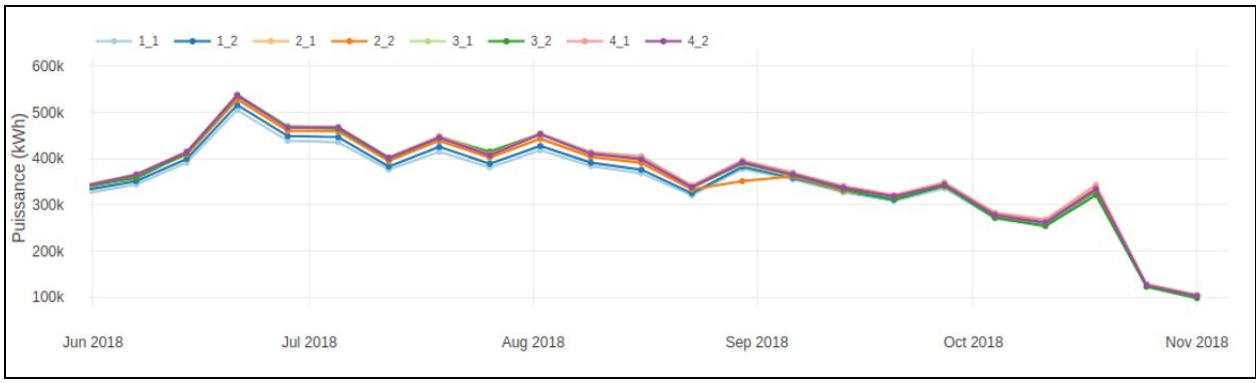
5.2. Designed tab

We have developed a tab that combines different graphs in addition to the map of the solar farm. These graphs are illustrated and described below:

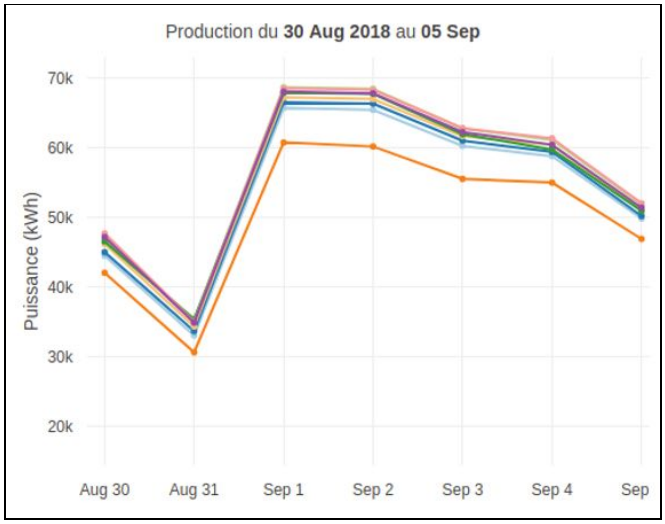
- **Bar plot to compare the total production:** This bar plot summarizes the total production over the current selected period for all the selected inverters of the solar farm. On the horizontal axis we indicate each inverter ID. The bars will have different colors specific to each inverter.



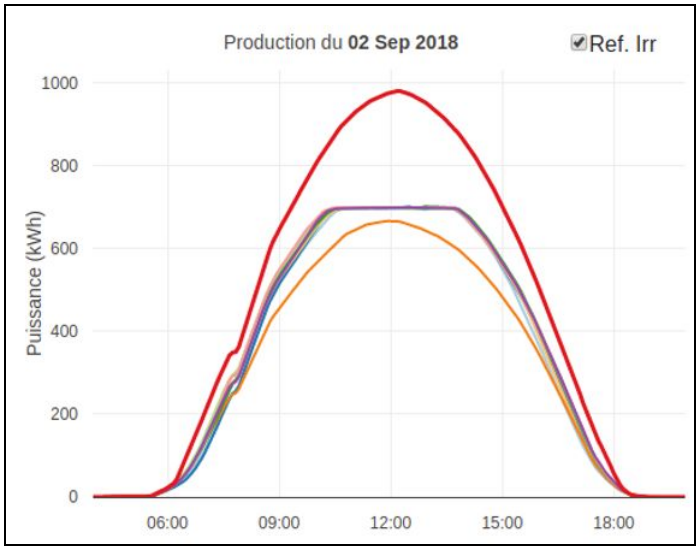
- **Weekly production evolution**: This graphical component is intended to show weekly production evolution over a specific period for all 8 inverters. Each inverter is represented by a different color indicated in the plot legend at the top left corner. These colors match the color of the zones on the map.



- **Performances by day over a week:** This graphical component is intended to show daily production evolution over a week for all 8 inverters. The colors of the curves match the color of the zones on the map.



- **Detailed performance of a given day:** This graph represents the production of the 8 inverters together with the irradiance progress over a given day. The colors of the curves also match the color of the zones on the map. Day hours are represented on the horizontal axis, production and irradiance are represented on the vertical axis.



In the end, by combining all the graphical components presented above we created the tab displayed below:

The tab includes some interactions between the graphs:

- **Interaction 1**: The map selector lets the user select one inverter or a set of inverters. This automatically updates the bar plot to compare the total production, the weekly production evolution, the performances by day over a week and the detailed performance of a given day. By default, the first zone is selected, causing all the graphics to show something.
- **Interaction 2**: The time slider allows selecting the desired period of time. By default, the entire period going from June 2017 to December 2018 is selected but the user can select a shorter period which should be at least 3 months long. The selection of the period automatically updates the bar plot to compare the total production, the weekly production evolution, the performances by day over a week, and the detailed performance of a given day.

- **Interaction 3**: The selection of a *week point* induces an automatic update of the weekly production evolution (this represents 7 values of production, one per day in the week) and the detailed performance of the first day of the week.
- **Interaction 4**: The selection of a *day point* induces an automatic update of the detailed performance of the selected day .

5.3. Interesting insights

By analyzing the different graphs of the tab, we detected a period of significant under-performace for the zone 2_2 and for the week of August 30, 2018. As it can be seen in the detail of the production of September 2nd, this area shows an under-performance throughout the day compared to the other zones. This behavior cannot be explained by a meteorological phenomenon.

5.4. Possible improvements

This tab was originally conceived for comparing the performances between the different inverters. However, it does not clearly appear from the developed tab which inverter outperforms the others in terms of energy production for a given period. To fix this, in the area where we displayed the weekly production evolution (top right corner), we could consider adding a tick box allowing to select different types of plots. We can keep the option showing the weekly production evolution which is useful to detect that the production is lower during the winter period, compared to the summer. In addition, we could include another plot displaying the difference between the weekly production evolution of an inverter and the mean production computed over all inverters for the week. This plot would allow identifying which inverters have production above average (i.e. positive values).

6. TAB 3: Identify and analyze performance losses and “outliers days”

6.1. Overview

The main objective of TAB3 is to explore in a detailed manner the outliers of production, which in our context are the days during which something unusual happens, resulting in a temporary loss of production compared to an expected production. During our project we identified four different types of loss for a given day :

- **Short loss** : abnormal loss of production during up to an hour
- **Long loss** : abnormal loss of production for several hours
- **Late start** : absence of production in the morning until a point of normal production
- **Temporary failure** : no production at all during the failure.

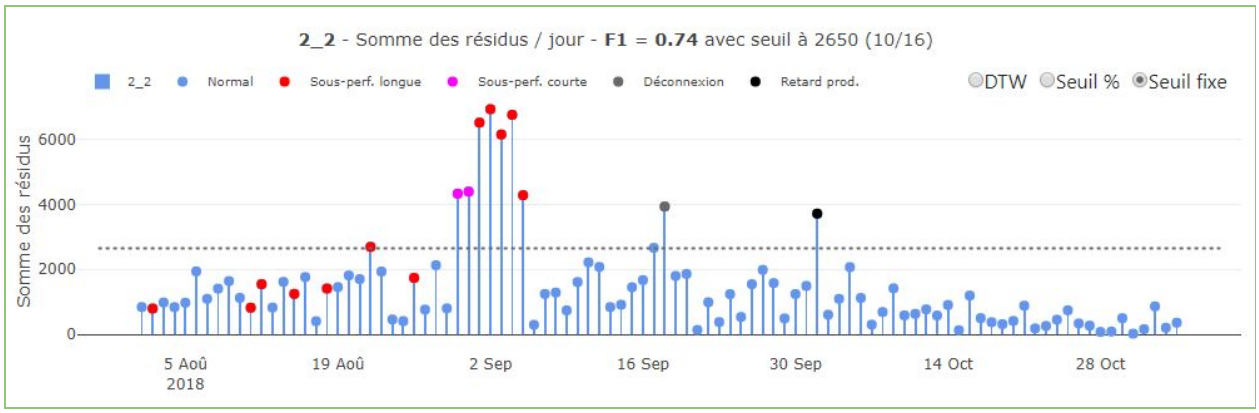
6.2. Designed tab

In addition to the map of the solar farm we have added the following graphs:

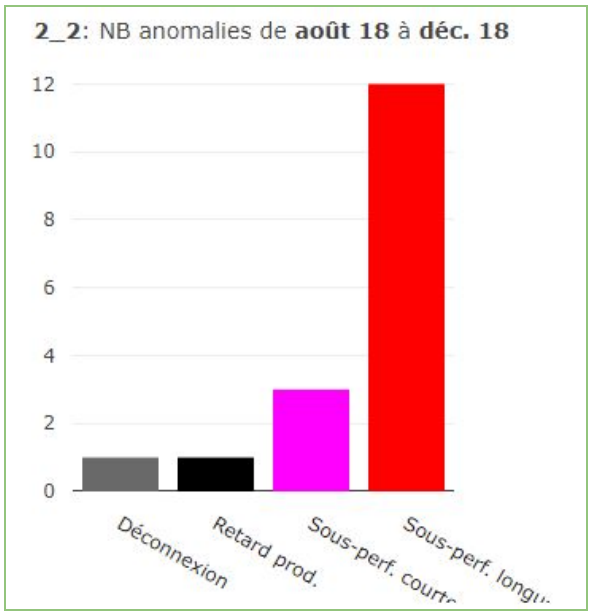
- **Stem graph with daily loss and outliers identified with thresholds**: The daily loss or residue is the difference between what has been produced and the expected production for a day.

  With the visualization below, we want to provide an analysis of the residues to investigate if there is a correlation or not with a spotted anomaly for a given period of time. Coloured dots (not blue) show the ground truth anomalies and the dashed line shows the decision threshold.

  We can also choose between a fixed threshold, a percentage threshold and a threshold based on the DTW (Dynamic Time Warping) similarity. The following plot uses a fixed value threshold to detect the outliers. To evaluate the performance of the chosen method, the F1 score has been indicated. The higher the value, the more accurate is the method with between parentheses the number of anomalies detected among the actual number of anomalies that exist.

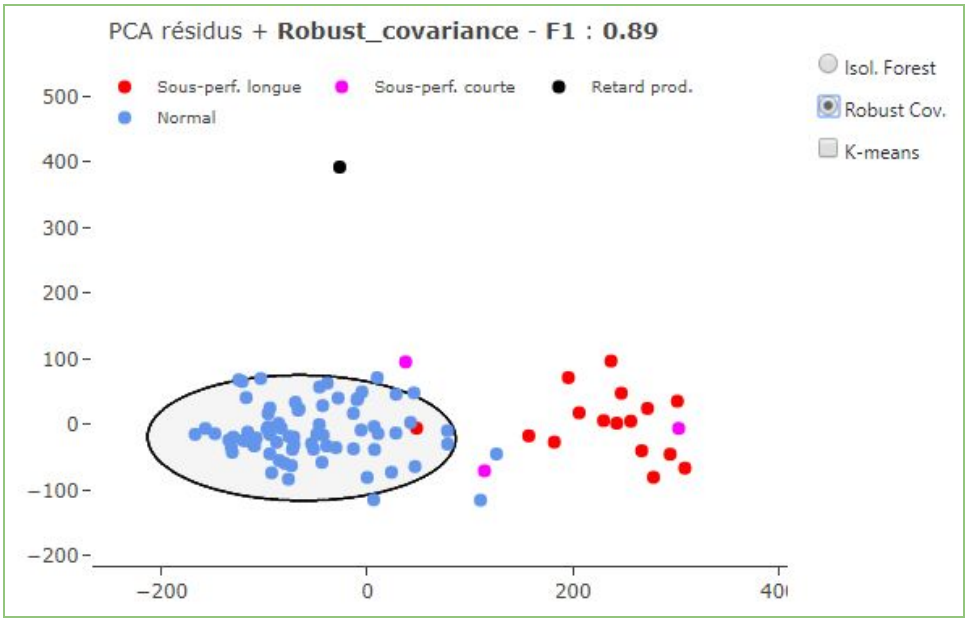


- **Bar plot with the Distribution of outliers per type by inverter**: The number of anomalies during the selected period of time is shown by type in the below bar plot for a specific inverter. We have the same color code for the three plots for the anomalies types. This graph allows to easily determine which anomaly type is most represented during the period of time.

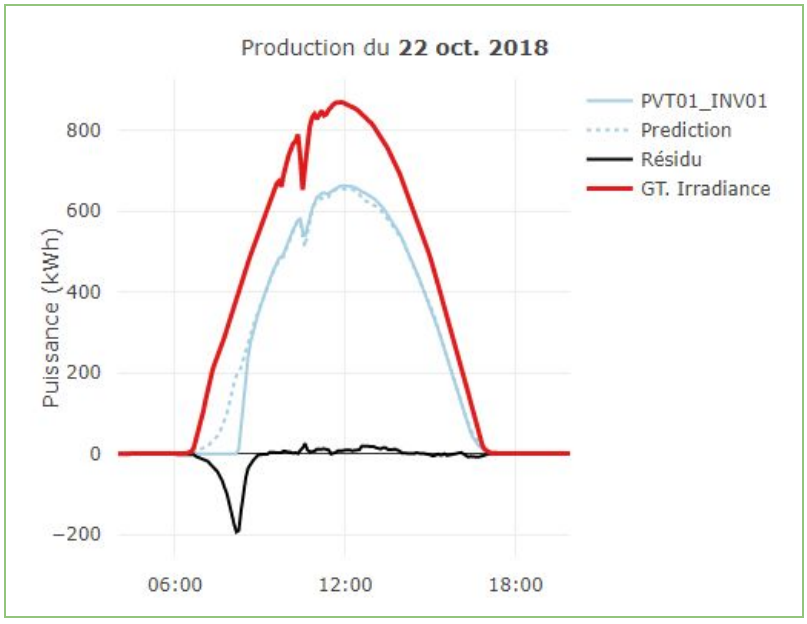


- **Scatter plot to identify outliers from clustering**: Depending on the chosen method, one of the graphs shown below will be displayed. Three methods are proposed for clustering, each of them corresponds to a specific clustering algorithm:
    - Isolation forest
    - Robust covariance
    - K-means

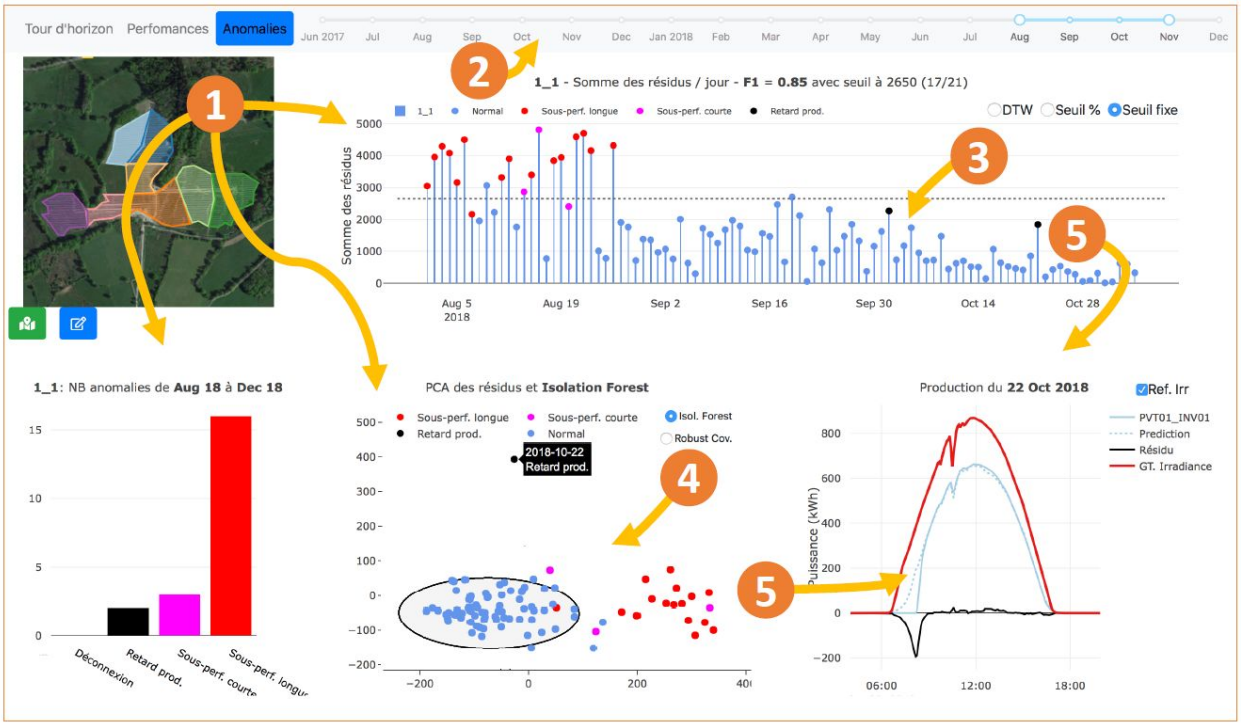
  The F1 score is calculated and allows to compare between methods, but also with the above stem graph and the threshold methods. Each point represents one day and the anomalies detected are the points which are situated outside the boundary. The corresponding anomaly colors have the same color code as the previous two graphs.



- **Expected production vs actual production for a day**: The graph below represents the irradiance in red, the production as well as the prediction of the current inverter as the day progresses. The color of the production curve also matches the color of the zones on the farm map. The prediction is represented by a dashed curve and has the same color as the selected inverter. Day hours are represented on the horizontal axis.



In the end, by combining all the graphical components presented above we created the tab displayed below:



Similar to the first and second tabs, this tab also includes some interactions between the graphs:

- **Interaction 1**: The map selector lets the user select one inverter in the solar farm. This automatically updates the stem plot with daily loss and outliers identified with thresholds, the bar plot with the distribution of outliers per type by inverter and the scatter plot to identify outliers from clustering. By default, the first zone is selected, causing all the graphics to show something.
- **Interaction 2**: The time slider allows selecting the desired period of time. The selection of the period automatically updates the stem plot with daily loss and outliers identified with thresholds, the bar plot with the distribution of outliers per type by inverter and the scatter plot to identify outliers from clustering.
- **Interaction 3**: Selection of the thresholding method to detect anomalies. The days above the threshold are the anomalies detected, the days with coloured circles (not blue) represent the real anomalies.

- **Interaction 4**: Selection of the clustering algorithm and update of the anomaly detection contour. Each of the clustering approaches may identify different types of outliers more clearly and we can compare the F1 score between them.
- **Interaction 5**: The circles from the stem plot could be clicked to display the details of the day of production in the right-hand bottom corner. The circles from the scatter clustering plot could also be clicked to display the details of the day of production in the right-hand bottom corner.

6.3. Interesting insights

By analyzing the graphical components associated to zone 1_1, we noticed that the two production late start anomalies corresponding to the 2nd and the 22nd of October 2018 were not detected by either the fixed and the percentage threshold methods. This is because the residual sum is not as great as the threshold value. However these are well detected by the DTW distance based method which compares the similarity between two time series. This example also provides a very good intuition behind the DWT distance algorithm.

6.4. Possible improvements

When investigating the stem graph with daily loss and outliers identified with thresholds, it could be nice to know upfront which of the methods performs best between the DTW, the fixed and the percentage thresholds. This additional information could be added in the title instead of needing to click and remembering the values of the F1 score associated to the different methods. The same applies to the scatter plot used for identifying outliers from clustering methods. Moreover, the investigation of the daily expected and actual productions indicated that some days that appear as normal should be in principle tagged as anomalies. An option could be added to allow the data scientist changing the label of the day and the type of anomaly according to his/her interpretation. Last it would be very useful to include a control allowing changing the threshold value and see how the F1 score would change.

7. CONCLUSION

We have decided to create a dashboard primarily aimed data scientists, with three tabs to explore many aspects of our data. This tool should help these users exploring and sharing their findings with their colleagues. It is built upon the extensive work we carried out on our data during our Fil Rouge project.

The first tab is designed for a quick exploration of the data. The second one allows an in depth analysis of the data to gain a better understanding of the performances of an inverter. The last tab requires a slightly higher technical level and an understanding of the Machine Learning work that has been carried out on the data to better interpret the results and the anomalies displayed over the given period. This includes understanding which metrics have been used to decide what days are outliers, what days are normal and why. Should we have decided to aim our tool at business users, we would have chosen a different set of designs, like calendar views directly displaying anomalies.

We also understand that some of these tabs may appear very dense when all the graphs are displayed together. However, it should be noted that the path of interactions has been thought to navigate through the graphs one after the other, going from a general view to a more detailed view.

Strengths

- The main strengths of the developed visualizations is the ability for the user to cross different pieces of information in order to have a complete view of the problem: from a quick exploration of the daily production to the manual checking of the information concerning the detected anomalies (e.g. date of anomaly, type, severity).
- We also think that our dashboard can be very useful for the data scientists at Engie, particularly when it comes to investigating the behavior of specific days and find out if they present anomalies.
- Visually, we decided to produce a dashboard that did not require the user to scroll up or down the page. We think it improves usability by limiting the work with the mouse or trackpad. Indeed, all the information would fit in the same window and would directly be available for users. Users need less than three clicks to access all the data which makes their navigation experience easier and more intuitive.

Further improvements

Some improvements have already been suggested for each of the tabs. Below are some other suggestions:

- At the moment, the choice of the range for the time period is a tricky subject. First it has to be tuned differently for each tab, as some designs cannot accommodate too long a period. Second, we have to set a global “maximum” and “minimum” for the time period, currently set at approximately 1.5 years for the complete data and 3 months for the test data. What if we get more data ? How would we accommodate a longer period in our design to keep the graphics legible ? This is something that would require extra thoughts.
- As it stands the design might frustrate the more advanced data scientists as we have not included yet extra controls to change all the parameters and hyper-parameters in the design. Now it is important to be able to “fiddle” with the parameters to push the boundaries of exploration. This could be for a next version.

Last word

We are pleased to say that our work has been received very positively by the Data Science team at Engie to whom we have presented the application today (27th June 2019) :-)