AMERICAN UNIVERSITY OF ARMENIA

CAPSTONE PROJECT

Application of data-driven approaches for photovoltaic system performance measurement

---

Author:
Arin JOKAKLIAN
Ashot JANIBEKYAN
Hovsep AVAGYAN

Supervisor:
Arpine KOZMANYAN

A project submitted in fulfillment of the requirements for the degree of BS in Computer Science

in the

Zaven & Sonia Akian College of Science and Engineering

May 30, 2021

Contents

Abstract

Monitoring photovoltaic system performance can have multiple applications, from improving the systems efficiency to early detection of any potential and dangerous faults. The dependence of the system performance from a large variety of factors, such as weather and seasonal conditions, variability between different sections of the same plant and the limitations in the collection of monitoring data make this task particularly complex. The use of automated data analytics systems, on the other hand, can help alleviate and automate the management of the complex relationships and accelerate the data processing and analysis times.

In this work we investigate the possibility of development of data-driven performance control methods that can be easily scaled between multiple systems. For this purpose we developed two main anomaly detection approaches: i) a baseline approach that is based on simple principles and exploits our knowledge of the data; ii) a set of machine learning approaches that can potentially take into account more complex patterns present in the data. We compare the outcomes of the two approaches and give an interpretation to any differences observed between their outcomes. We conclude that despite the ability to better account for inter-variable dependencies the machine learning approaches are not able to provide the required level of scalability, but potential improvements can still be done to correct this. The baseline approach developed by us, on the other hand, despite ignoring some more complex relationships between the variables, provides reliable results and can be easily applied to new data.

As possible future extension of the work we recommend a number of improvements for the machine learning approaches and suggest to extend the analysis to larger datasets including different plants operating in varying conditions. Additionally, the availability of additional details about the power plant scale and configuration as well as close collaboration with field experts would be crucial for a fuller interpretation of the detected anomalies and their possible causes.

1. Introduction

The monitoring of photovoltaic system performance has multiple scopes. It allows on one hand to ensure that the yield of the system is optimal and in case of necessity take actions to improve the performance. On the other hand it can help identify early any malfunctions that could indicate or trigger more serious failures.

The complex dependency of the system performance on a range of conditions such as weather and seasonal variations, the operating conditions, plant configuration and scales as well as differences between different plant configurations makes this task particularly difficult.

Solar panels are formed by joining a number of photovoltaic cells, that are in charge of conversion of sunlight to electrical power through the photovoltaic effect [3, 1]. Arrays of solar panels are in their turn connected to inverters that are responsible for converting the direct current (DC) generated by the panels to alternating current (AC) - more suitable for distribution over the power gird. In this work we explore open source data [1] that includes information on power generation and conversion steps in two different power plants. We use this data to study any possible anomalies in the first step of power generation and create a number of simple checks to intercept these types of anomalies automatically. These types of checks could help identify early on possible performance falls or failures and give warnings to the field experts.

Due to the limited information available about the plants a full interpretation of the observed behaviour and its physical causes is difficult, we thus concentrate on two main aspects:

1. detection of behaviours and performance issues that we can classify as anomalous with high confidence
2. scalability of the method, its easy application to any new system with different configurations

We also pay particular attention to having a good understanding of the data and its behaviours and base the development of the data-driven approach on those insights.

The analysis and the implementation of the approaches has been performed in Python language in Jupyter Notebook environment. After the finalization of some of the approaches we also completed an easy-to-launch python script that runs the entire pipeline of anomaly derivation for a selected dataset and saves the final outcome in output files. This step simulates a conversation of a developed method into a code that can be integrated in other systems at production level. In the footnotes of this page we provide the link to access the repository with the complete code[2].

[1] https://www.kaggle.com/anikannal/solar-power-generation-data

[2] https://github.com/ArpineKoz/ds_starter

In the rest of the work we give a detailed description of how we proceeded and what we found. We start by giving a short description of the available data in Section 2. In the next section (Section 2.1) we detail the analysis done and the approaches developed for anomaly detection. In Section 3 we present the outcome of the approaches and their interpretation. We finally conclude with a discussion and conclusions on the results in Section 4.

2. Data

The data used in the paper has been gathered at two solar power plants in India over a 34 day period. Two sets of information are available for each plant: i) power generation dataset with around 68000 observations; ii) sensor readings dataset. The power generation datasets are gathered at the inverter level and each inverter has multiple lines of solar panels attached to it[3]. The sensor data is gathered at a plant level by a number of sensors optimally placed at the plant. The power generation dataset has the following variables: *DATE_TIME* - recorded at 15 minute intervals; *PLANT_ID* - common for the entire file; *SOURCE_KEY* - stands for the inverter id; *DC_POWER* - amount of DC power (in kW) generated by the solar panels attached to the inverter in the 15 minute interval; *AC_POWER* - amount of AC power (in kW) generated by the inverter in the 15 minute interval; *DAILY_YIELD* - daily yield is a cumulative sum of power generated on that day, till that point in time; and *TOTAL_YIELD* - total yield for the inverter till that point in time. The sensor readings dataset contains the following variables: *DATE_TIME* - recorded at 15 minute intervals; *PLANT_ID* - common for the entire file; *SOURCE_KEY* - stands for the sensor panel id (this will be common for the entire file because there is only one sensor panel for the plant); *AMBIENT_TEMPERATURE* - ambient temperature at the plant; *MODULE_TEMPERATURE* - temperature reading for the module (solar panel) attached to the sensor panel; and *IRRADIATION* - irradiation [4] for the 15 minute interval.

A number of important details are missing from the data, such as the scale and the capacity of the plants as well as any particulars on the panel, inverter setup and their connections. This limits somewhat our ability to give a full interpretation of the effects found in this work and their causes.

2.1. Method

In this section we describe our approach to identifying anomalies (potential losses in the total generated power) in the power generation data in the plants. The development of the approach can be divided in three main stages:

- data exploration,

[3] Throughout the work when we refer to panel sets or panel groups we will intend panel arrays connected to individual inverters.

[4] No units were provided for this variable in the data, but judging from the variable amplitude it is likely to be in kW/m^2.

- detection of anomalies using an ad-hoc data analytics approach (baseline approach),
- detection of anomalies using machine learning models.

The first step of in depth data exploration allowed to have a clear idea of the data and the relationships and patterns present among the variables. With a clear idea of these aspects we were able to define the main directions of work and the types of anomalies that could potentially be derived from the data. We were also able to outline an analytic approach to derive those anomalies using our knowledge of the data. This approach is simple and easily interpretable, hence we will use it as a baseline model in this work. The scope of the work, however is the creation of a method that is easily scalable across plants and can be applied with little effort to new data. For this reason we also evaluated approaches using machine learning models to derive the same types of anomalies. After the derivation of anomalies using the two approaches we compared their results and gave an interpretation to the agreements and differences between the outcomes.

Since only one months worth of data is available from as few as two plants, our strategy has been throughout the work to develop all the anomaly detection methods and pipelines on the data deriving from only one of the plants (Plant 1) and subsequently apply the developed methods on the second (Plant 2). This way we were able to imitate the development of the approach on limited data and its application on new and unseen data in order to evaluate the generalizability of our approaches.

2.2. Data exploration

The three quantities in the weather dataset (*AMBIENT_TEMPERATURE*, *MODULE_TEMPERATURE* and *IRRADIATION*) are of particular interests to us, since they are the direct or indirect drivers of the power generation. As expected, *AMBIENT_TEMPERATURE* and *MODULE_TEMPERATURE* have similar values, although *MODULE_TEMPERATURE* is usually higher (see Fig. 1). This is due to the fact that the module temperatures vary more easily due to outside irradiation and thus grow to higher values during the hours with highest irradiation in the day. We note, however that the module temperature, although highly correlated with the irradiation levels, has also a weak dependence on the ambient temperature. This is easily observed in Fig. 2, where the high correlation of the module temperature and the irradiation is evident. However the scatter in the correlation can be additionally explained by the ambient temperature. As is physically expected, high ambient temperature contributes to heating the module even in conditions of low irradiation. Similarly low ambient temperature contributes to lowering the module temperature when irradiation is at maximum.

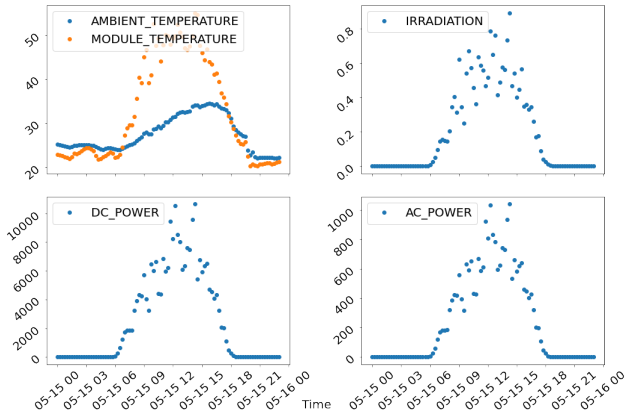

Figure 1. The behaviour of selected variables during the day. In the top-left corner we show the behaviour of ambient and module temperatures that are both affected by irradiation (top-right corner). Module temperature is strongly dependent on irradiation. Ambient temperature instead is more inert and has a lag in its maximum value with respect to irradiation maximum. In the bottom-left plot we show the overall DC power generated by a single panel group (a set of panels attached to a single inverter) on the same day. In the bottom-right corner we show the AC power generated by the inverter in question converting to AC power the DC power received from the panels.[6]

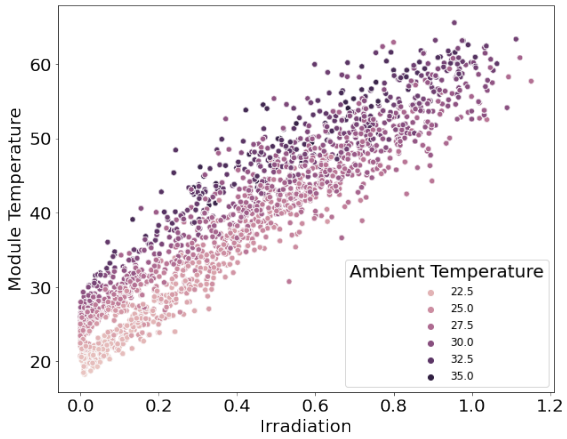

Figure 2. The complex dependence of module temperature from irradiation and ambient temperatures. The irradiation is the strongest factor driving the module temperature values, however the ambient temperature can enhance or mitigate the effect of irradiation.

[6]Typical DC to AC power conversion efficiencies are in the order of 10%. That is the AC power amplitude is around 10th of the DC power amplitude. In this plant the AC power seems to be at around 1% of the DC power (The data for the second plant instead behaves as expected). We think this could be due to some error in the data or the measurement, however our future treatment does not directly depend from the DC power amplitude values thus we leave the data as it is without correcting or modifying it.

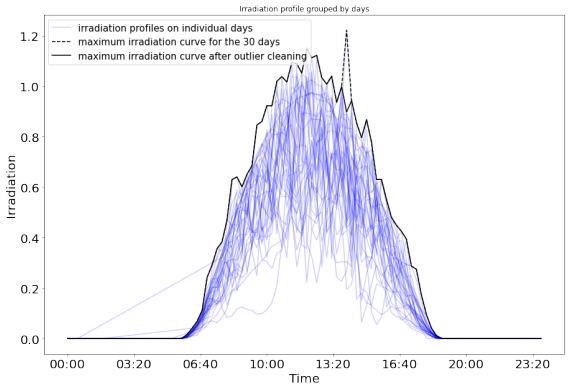

Figure 3. The irradiation profiles observed over the 30 days for Plant 1 (blue). The black dashed curve represents the maximum irradiation profile derived by taking the hourly maximum over the set of days. The black line represents the maximum irradiation profile after applying some outlier cleaning.

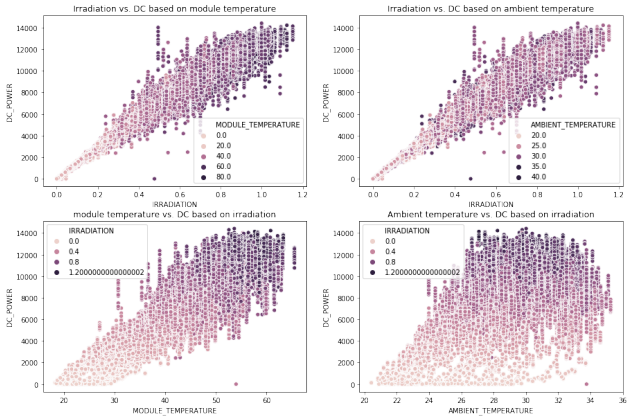

Figure 4. Top left: Dependence of DC power from irradiation and module temperatures. Top right: Dependence of DC power from irradiation and ambient temperatures. Bottom left: Dependence of DC power from module temperature. A cutoff power value is observed above certain temperature and irradiation values. Bottom right: Dependence of DC power and ambient temperature with a first cutoff observed for high values of irradiation and a second steep cutoff at higher values of ambient temperature.

The observation of the behaviour of the two ambient quantities (irradiation and ambient temperature) in the 30 day period available in the data helped us conclude that, as expected, any weather variations are most strongly and quickly observed in the irradiation data. The ambient temperature changes more slowly and is also more stable to small variations of weather conditions (such as clouds, etc). The irradiation instead can be highly variable during the day and this can be mainly due to the sun occlusion with clouds and their variability (Fig. 3). We can use this strong variability to define a possible measure of cloudiness using the fact that the daily profile of the maximal possible irradiation can be considered stable in a short period (one month in this case) where changes due to seasonal variability have moderate effect. We can potentially derive a 'perfect' irradiation profile per season. This possibility will be further detailed in Section 2.3.

In Fig. 1 we show example profiles of the DC power generated by a group of panels (connected to a single inverter) and AC power generated from the input DC power by the inverter itself. As expected, both signals are highly correlated between each other and with the irradiation. The DC power, however is the direct outcome of irradiation levels, thus any anomalies found in the DC power behaviour with respect to outside conditions will be due to irregularities in the solar panel condition or the system configuration up to the input point of the inverter. Instead any anomalies in the AC power behaviour with respect to the DC power would be indicators of irregularities in the inverter function itself as DC to AC power conversion is done by the inverter. We will construct our anomaly detection approach with this in mind. Moreover, in order to concentrate more on the panel behaviour and anomalies due to solar panels, we will mainly concentrate on anomaly detection in the DC power signal. Although some treatment of the AC power anomalies is also done.

In Fig. 4 we show in detail the inter dependence of power generation with the ambient variables that we discussed previously. As expected, the strong correlation of the DC power with irradiation is evident. There is still some scatter in this relationship, but as we will see in Section 2.3 it is due to varying performance of different inverters. An interesting behaviour can be noticed observing the tip of the DC power scatterplot in the three out of four images present in Fig. 4. It is evident (and is expected) that the DC power generation flattens out at high module temperatures (top-left and bottom-left images of Fig. 4). This is due to the fact that solar panels loose efficiency at high module temperature values. The high module temperatures, however are not only due to high irradiation, as we saw earlier, but can be reached with the additional effect of the ambient temperatures. In fact such high module temperatures are reached both due tot high irradiation and ambient temperature values. This is visible in the bottom-right image of Fig. 4 where two cutoffs can be observed in the relationship of DC power and ambient temperature. The top horizontal one is independent of the ambient temperature and is due to high irradiation values (possibly to the limits of irradiation levels in the period in question). The second one is visible beyond this section, at even higher ranges of ambient temperature, where the higher is the value of temperature the lower is the possible DC power generated..

This complex relationship between the generated power and weather variables means that a full treatment of the

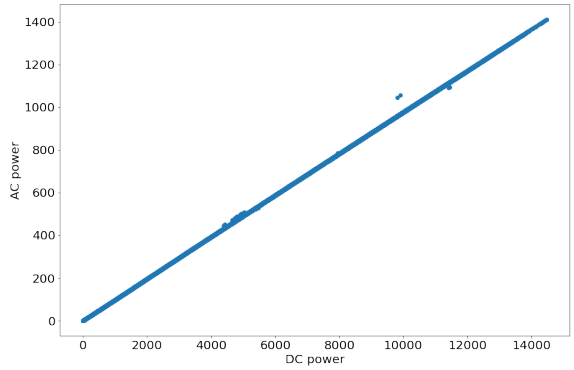

Figure 5. The relationship between the DC power generated by the panels and the respective AC power generated by the inverters. The strong linear relationship with a number of clear anomalies is evident.

power generation behaviour has to take into account all of the weather variables. Indeed some more information such as wind speed (that is not available in this data) would also add useful insight to the analysis.

We finally treat the conversion of DC power to AC power due to the inverters. As already mentioned the two signals are strongly correlated (see Fig. 5). This almost constant conversion efficiency is a result of the high stability of the inverters. A number of outliers are observed in the image, but could be due to some noise or instantaneous variations since they seem to be isolated points rather than some behaviour extended in time. We will treat them better in Section 2.3.3.

2.3. Development of the baseline approach for anomaly detection

In the following sections we provide a detailed description of the method used for the anomaly derivation through a data-analytics approach that we developed. This treatment helped us to better understand the data. This understanding was then crucial for the correct evaluation of machine learning approaches (described in Section 2.4).

2.3.1 Cloudiness measure

In Section 2.2 we showed and discussed how strongly is the generated power affected by the irradiation. This means that the yield and the output of the panels can be strongly affected by clouds and will be particularly low on cloudy days [2]. With this in mind, we wanted to introduce a measure that would allow to have a quantitative idea of how cloudy each day has been and to be able to check the correlation of any effects we observe with the weather conditions. For example it could be used to control whether power generation performance measures derived by us carry

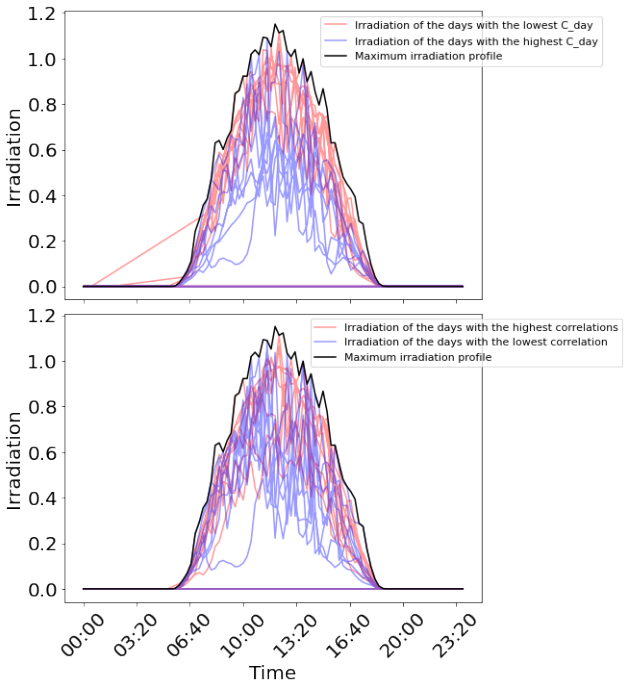

Figure 6. Top: Irradiation profiles of the least (8 profiles) and most (9 profiles) cloudy days found using the C_{day} formula. Bottom: Irradiation profiles of the least (8 profiles) and most (9 profiles) cloudy days found using the correlation as cloudiness measure.

any biases due to weather conditions (ideally they should reflect only the performance due to panel conditions).

With no variable that explicitly shows the cloudiness of the day, we decided to derive the value using the irradiation information. As introduced in Section 2.2 and in Fig. 3 the irradiation behaviour over a set of days can give us an idea of the maximum possible irradiation values for the given season (a 'perfect' irradiation profile). We thus derive a maximum irradiation curve by taking the maximum value of the irradiation over the set of days for each particular timestamp (see the dashed black line in Fig. 3). To derive a value representing the cloudiness of the day, we took the sum of the difference squared of the irradiation profile of each day and the maximum irradiation curve (see Eq. 1). [7]

$$C_{day} = \sum_i (I_{max} - I_i)^2, \quad (1)$$

where the sum runs over 24 hours and i represents a particular timestamp during the day.

[7] The maximum irradiation curve resulted to have an anomalous peak, probably due to some error in the data. We cleaned the curve for this type of anomalies by defining acceptable irradiation ranges through the mean and the standard deviation of irradiation values in each timestamp for a set of days. Any points outside the limit of mean ± 3std were removed as outliers. The maximum irradiation curve (solid black line in Fig. 3) and the cloudiness were calculated only after this step.

To make sure that this was the right approach, we introduced a second possible measure defined as the correlation value of the irradiation profiles and the maximum irradiation curve. We then compared how well do the two measures quantify the distance of a single irradiation curve from the maximum irradiation profile. For this comparison we plot the most and least cloudy days according to both measures and compare their distance from the maximum irradiation profile. The result is shown in Fig. 6. The blue curves represent the irradiation profile for the days that according to the given measure should be the least cloudy. The red curves represent the irradiation profile for the days that according to the given measure should be the most cloudy. We can see that in top image of the figure the red curves (least cloudy days) are more separated and easily distinguishable from the blue curves (most cloudy days), whereas this is less so in the bottom image. For the least cloudy days, the curves are closer to the maximum irradiation curve (low values of C_{day}), as opposed to the curves that are further from the maximum irradiation curve (high C_{day}). This is what was expected, because the cloudier is the day, the lower is the irradiation. The lower separation in case of correlation as measure of cloudiness is also expected, as it measures the correlation between the maximum and the individual curves, and not the difference/distance between those two. Thus, we select C_{day} as a measure of cloudiness between the two measures.

2.3.2 DC power generation anomalies and inefficiencies

In Section 2.2 we started investigating the relationship between the irradiation and the DC power generated by panel sets connected to single inverter. Fig. 4 brings to light a number of behaviours that are worth investigation. Firstly, it is worth understanding the origin of the scatter of the DC power at high values of irradiation. At the same time a number of points are visible that are clearly outside of this relationship and could be themselves some kind of failures or shutdowns of some plant components.

In order to investigate better these two issues we conducted a deeper analysis by looking at this dependence for individual panel sets on individual days. We notice that different panel sets can show different inclination with respect to the irradiation axis (see Fig. 7). This is particularly so near high irradiation value. By measuring this inclination we can define a way to quantify the performance of single panel sets and compare them against each other. Before we can do so, however it is important to identify the points that are outside this relationship, both because they could themselves represent anomalies in the panels (and corresponding warnings should be given when they occur) and because

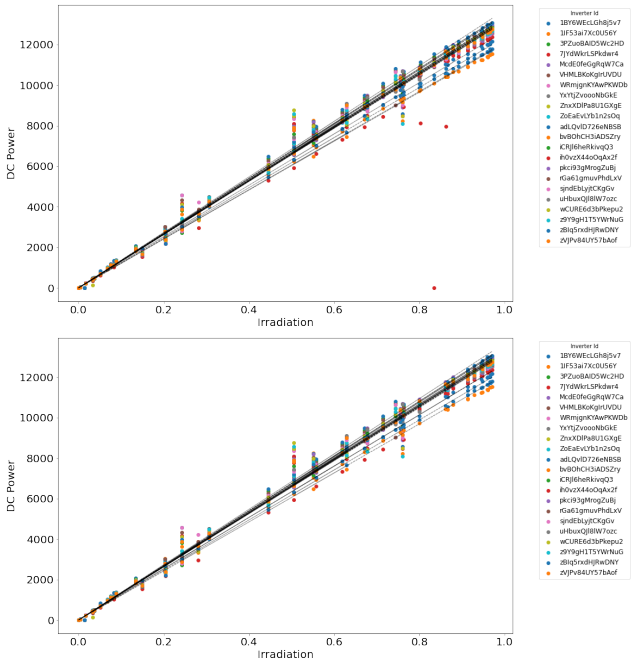

Figure 7. Top: Raw data showing the dependence of the generated DC power from irradiation for different panel sets on a single day (25 June 2020). The black dashed lines represent the linear fits used to describe these dependencies. The few outlying points are clearly visible for the panel set connected to the inverter "ih0vzX44oOqAx2f". They can potentially affect the fitted line for this inverter, so need to be removed before applying the fit. They could also represent panel performance anomalies and could necessitate a warning. Bottom: The data after cleaning for the described outliers. A slight difference in the fitted lines can be observed after outlier cleaning. In the plots there is a clear difference between some panel sets when comparing the amount of the DC power generated for fixed values of irradiation at high values of irradiation.

they represent noise that could mask the real performance of the panels (see for example the top image in Fig. 7).

Outliers as isolated anomalies. In order to identify and remove those points we first derive an approximate efficiency measure defined by the fraction of DC power and irradiation. We then use the distribution of this fraction for each panel set in order to identify the most anomalous values of the quantity with respect to its recent behaviour. More specifically we use a rolling time window for each panel set and tag as outliers the points that are 4 standard deviations away from the mean value of the distribution in the given time window. This identifies very well the most extreme outlying points (see the difference between top and bottom images in Fig. 7) allowing us to move on to the next step of the anomaly calculation, that of derivation of anomalies in the daily efficiencies of the panel sets.

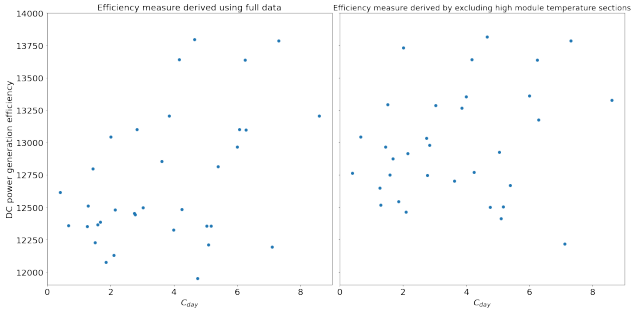

Figure 8. The correlation of the daily cloudiness the DC power generation efficiency calculated using the full data (left), or the data that only includes module temperatures (right) efficiency measure excluding the values where the module temperatures below 50°C (right). The use of data with low values of module temperatures moderates slightly the levels of correlation removing partially any weather biases present in the efficiency coefficients.

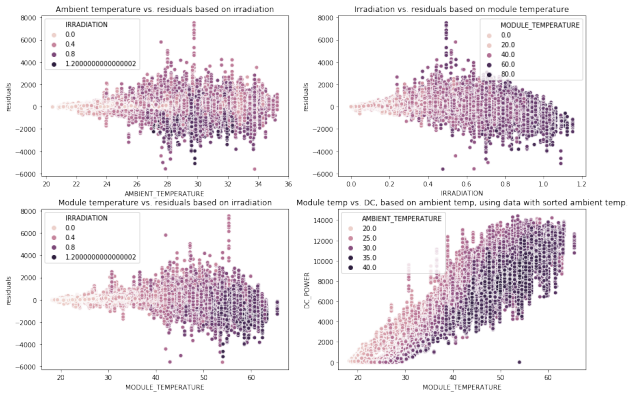

Figure 9. The residuals derived after applying a linear fit to the irradiation-DC power relationship and their dependence from weather variables. The clear bias (growth) of the residuals at high values of irradiation and module temperatures is evident in the top right and bottom left images.

Panel set daily efficiency. In order to quantify the performance a panel set during the DC power generation from the solar irradiation we make a linear fit (black dashed lines in Fig. 7 represent the fit done for every individual panel set) to describe their relationship. We use the coefficient of the fit as a measure of efficiency and would like to ensure that it is indeed measuring inverter performance and not weather related effects.

We compare the derived daily efficiency values with the cloudiness measure derived earlier. Fig. 8 shows this dependence for a single panel set. A certain level of correlation between the derived efficiency values and the daily cloudiness values can be observed. Moreover the correlation has the opposite behaviour to what one would expect. Higher cloudiness values seem to correspond to higher DC generation efficiency. We explain this effect in the following way. The relationship between irradiation and the DC power observed in Fig. 7 is not linear. It is more curved towards high irradiation values. This is due to the efficiency drop of power generation mentioned in the Section 2.2. It adds a secondary effect on the DC power generated for the highest values of irradiation making the dependence curved. The linear fit thus gets shifted more towards lower angles of inclination for the panel sets where this effect is the stronger (that have had the most irradiation input).

The described curving effect can be observed in Fig. 9. It shows the dependence of the residuals ($r = DC_{power} - Irradiation * Efficiency$) of the linear fit from the weather variables. No particular relationship is observed between ambient temperature and the residuals, since the residuals are all centered at 0. On the other hand, it can be clearly observed in the upper right and lower left plots of the figure that when the irradiation and module temperature rise, the residuals are shifted towards more negative values. This is a sign that the DC to irradiation relation described earlier is actually curved (not linear) for high values of irradiation and a large number of points in this relation fall below the fitted line (this is even visible for the fits demonstrated in Fig. 7). The reason for the curved behaviour is the loss of efficiency of the panel sets at high module temperatures discussed in Section 2.2 and visible as a cutoff of the DC power at high module temperature values in the lower right image of Fig. 9. We thus conclude that the linear fit carries biases due to the weather effects and does not satisfy our requirements at least in this simple form.

The complete solution to the problem would assume the use of a non-linear relation to describe the dependence of the DC power from the irradiation, however this would lead to a more complex treatment of the problem, which we plan to implement using the machine learning approaches in the following steps. We choose a simpler approach in order to correct the efficiency measure for biases by deciding to do the linear fit only for the lower range of module temperature values and ignoring the more complex sections of the data. It is fair to say that the relation between DC power and Irradiation is linear for the lower range of module temperature values. We choose the values lower than 50°C as such by observing the behaviour in the lower right image of Fig. 9.

This correction removes partially the correlation between the derived efficiency values and the cloudiness as shown in the right part of Fig. 8. Some level of correlation is still present. Nevertheless, we accept the reached level of weather bias correction and intend to implement additional checks in the future steps in order to ensure the conclusions we make about the panel performance do not carry any of these biases. To conclude, so far we have derived a daily efficiency measure for each individual panel set by using

linear regression on the data with module temperatures below 50°C. The distribution of the derived efficiency values is shown in Fig. 10.

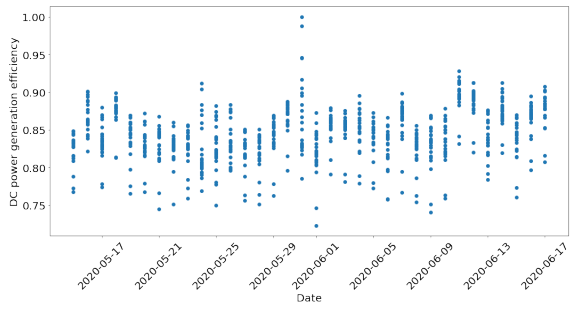

Figure 10. The final values of the DC power generation efficiencies. The values have been derived using only the sections of the data where the module temperature is below 50°C.

Having derived daily efficiency measures for the panel sets we can use those efficiency values to detect changes of panel set behaviour in time or compare panel set behaviour between each other. More specifically we will try to derive three types of behaviours:

- Panel set daily efficiency drops with respect to its recent behaviour
- Falling trends of panel set efficiency
- Panel set showing lower performance compared to the performance of other sets in the same conditions.

The first two of the mentioned behaviours, if detected, could be strongly affected by the biases that we expect to have in the daily efficiency value. The only behaviour that is not prone to weather biases is the third type that ensures independence from weather conditions, since the panel behaviour is always compared in similar conditions. To ensure the maximum possible independence from weather biases we will make additional arguments and checks for the first two types of anomalous behaviours.

Efficiency day-to-day drops. By looking for panel set efficiency drops we intend to look for any faults that occur in the panel system that could bring to sudden changes of their overall performance. Like mentioned before, we have to make sure the observed drops are not due to weather conditions.

To find the efficiency drops, we use as reference the behaviour of the panel set in a previous time window. We use the mean and standard deviation of the panel set efficiency in the given window to decide whether the efficiencies right after the period are anomalously low. If they are lower than a set limit (mean-2std) we consider that the panel set has shown particularly low efficiency on that day compared to its behaviour in preceding days.

Decreasing trend in efficiency. To look for any trends in the efficiency values of the panel sets we check the correlation of the efficiency values for every individual panel set with the number of the day in the year. The number of the day being a monotonically increasing quantity, should show negative correlation with the efficiency if the efficiency is showing a negative trend. Since the trends can well be a result of weather conditions we select only trends with very strong anti-correlation (i.e. Pearson correlation coefficient value close to -1). We additionally check that the observed trends are not common among panels to ensure again that the observed trends are not due to weather effect. Like before, also for this function we use a rolling window to check for the existence of the trend considering the recent behaviour of the panel set.

Panel sets with low efficiency. We finally check for panel sets that show a lower efficiency compared to the other panel sets in the same day. Given that the weather conditions are common for all the panel sets, any differences present between panel set efficiencies in a given day should be due to panel specifics and not due to external conditions. In order to select the panel sets that show different behaviour with respect to the rest of the groups we calculate the mean and the standard deviation of the efficiency values in a given day and qualify those that are 2 standard deviations below the mean value as anomalous.

We further check whether this anomalous behaviour has been continuous and give a stronger warning in case the low performance persists for more than a set amount of days (specifically, we set 7 days as a limit).

2.3.3 AC power conversion (inverter) anomalies and inefficiencies

In order to detect the AC power generation anomalies we use the tight relationship that exists between the DC and AC powers (also observed in SubSection 2.2). We fit a regression line to this relationship and use the residuals of single points with respect to their predicted AC power values to identify the points that are not following the observed tight behaviour. Specifically we calculate the mean and the standard deviation of the residual values and select the points that are away from the mean values by 3 standard deviations or more. Given the tightness of the AC power and DC power relationship the margin of the choice of this value is quite large. The outliers (observed easily in Fig. 11) are multiple standard deviations away from the mean value.

The outlier calculation described so far, however is based on the whole data. This means that when making a decision

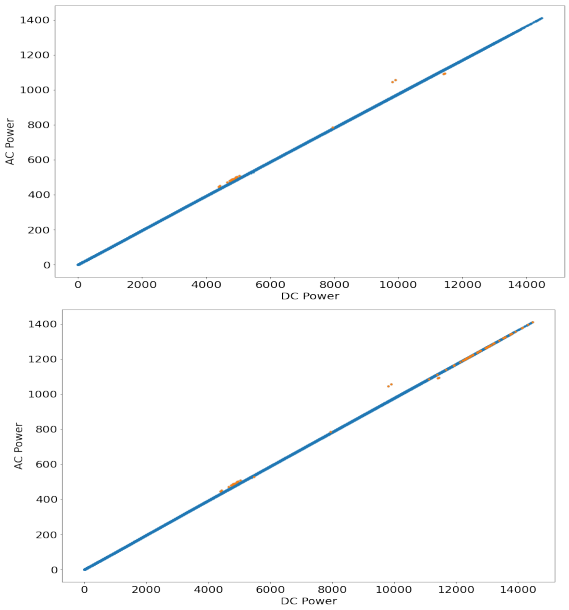

Figure 11. The linear relationship between AC power and DC power. In the top image the red points indicate the outlier points derived using the entire dataset. In the bottom image the red points indicate the outlier points derived using only the information available to the past from a given instance.

about outlier points on a specific date and at a specific instant we were also using values from the following days. In order to avoid this we need to make sure that we are doing both the regression fit and the outlier calculation only on the data that is available as of the specific day. Specifically, we calculate the conversion coefficient of DC to AC power, for each inverter for each day, by making a linear fit to the data of the current day. Then we calculate the residuals according to this fit and derive the outliers based on the mean and standard deviation of the residual values of the last 4 days. The outliers selected through this calculation are shown in the bottom image of Fig. 11. They are generally in agreement with the ones calculated using the entire data, with the exception of a stripe of points at high DC power values that result to have a slightly shifted behaviour. We will see in the following what are the reasons for this behaviour.

In order to better investigate the behaviour of the points at high DC power values we look at the residual values in more detail after removing the most extreme outliers identified so far. We note that the residuals of the remaining sets of points exhibit two different behaviours. We can observe this behaviour in Fig. 12 where a group of points with relatively high residuals is observed at high DC power values. Moreover there seems to be a relationship between the residual values and DC power generated at those points. This relationship can be explained considering the specifics of inverter functionality. Inverters are designed to work for

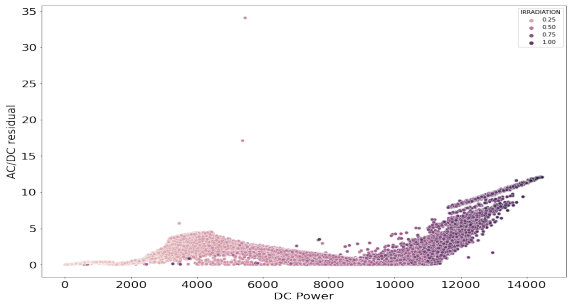

Figure 12. The observed distribution of AC to DC fit residuals and their dependence from DC power and irradiation. We see a correlation between the residual values and the DC power at high DC power values. When DC power is high (this naturally corresponds to high irradiation values) the values of residuals are also high. Moreover, there is a cutoff in the relationship showing that given our treatment of DC to AC power relationship the residual values never have low values in case of high DC power. This is explained by the falling power conversion efficiency of the inverters when the DC power is above the inverter design values.

specific power ranges, and they loose efficiency as the incoming power exceeds the design optimal (power clipping). Due to this reason the relationship between the AC and DC powers is actually slightly curved towards higher DC power values in Fig. 11 even if it is not visible by naked eye. Considering all of the DC power range is used for the regression line fitting, these points result slightly shifted from the fitted line and thus have higher residual values. The behaviour we observe then is the expected behaviour of the inverter, and should not be considered an anomaly. We thus leave this points in the analysis for the following steps.

We can use the daily conversion coefficients derived for outlier calculation above as measures of inverter efficiency and repeat the search for the different type of daily inefficiencies that we implemented in Section 2.3.2 for DC power to adopt for similar effificnes of AC power conversion. A daily measure of efficiency (conversion coefficients) can allow to detect:

- the inverters for which the conversion coefficient is paritcularly low with respect to other inverters,
- days on which the efficiency of inverter exhibits a sudden drop of efficiency
- inverters showing a degradation (falling trends of efficiency).

The implementation for deriving those types of behaviours is equivalent to the process described in Section 2.3.2.

We present the results of this analysis in Section 3.3 although the AC conversion anomalies are not the main focus of this work and in the rest of the work we will not concentrate on the aspects regarding the inverter functionality.

2.4. Anomaly detection with machine learning models

The approach developed in Section 2.3 has the advantage of being designed ad-hoc for the data at hand, hence uses the particularities of the data to develop a simple treatment of the problem. Like discussed in the same section, however it ignores some interesting dependencies and sections of data that could have important effect on the performance calculation.

In order to address this point and try to apply a more complete treatment of the data we make use of multiple machine learning approaches. As before, we concentrate here only on the direct current generation anomalies and ignore any effects that arise in the conversion from DC to AC current. This means that the main variables that we use in this section are the weather variables and the DC power, since the AC power inevitably contains effects of the power conversion deriving from the inverter itself.

Our methodology for deriving inefficiencies with machine learning is based on two main steps (see Fig. 13):

1. fit a model to the data in order to be able to describe the full dependence of the generated power from weather variables,

2. use the reconstruction error calculated between the models predicted value and the truth as a measure of inefficiency. Finally, derive the various types of anomalies discussed in the previous section from this inefficiency measure.

The basis of our methodology is the assumption that when used on data coming from all inverters the models should learn to represent the most common behaviour among various panel sets, making this way the less common behaviours easily identifiable through their difference from the model expectations.

The models that we used for this approach are the following:

- Linear Regression
- Decision Tree Regression
- Autoencoders

For the first two regression models we model the generated power based on the weather variables and use the predicted power value to calculate the error as an inefficiency measure. For the autoencoder we use all the 4 variables (3 weather variables and the DC power variable) as input and output of the autoencoder. We tested a number of architectures for the autoencoder from simple to high complexity and selected the more stable architecture (Fig. 15) that behaved according to our expectations.

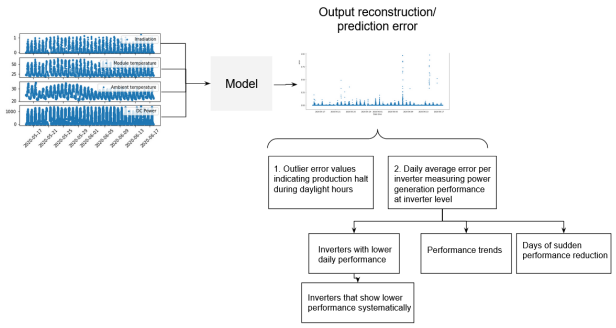

Figure 13. A diagram showing the main steps used for the anomaly derivation with machine learning models. The data is first used to train the model. After this the model output is used to construct a reconstruction error, which serves as a measure of inefficiency. Aggregated inefficiency values can be derived for single days and single panel sets and are used for detection of anomalies in the daily performances.

2.4.1 Regression models

We trained the regression models to recognise the dependence of the DC power from weather variables and expect the model to learn the most common behaviour among the panel sets. Against our expectations, however the models did not always necessarily learn to represent the statistically most common behaviour. In Fig. 14 we show an example region of predicted and actual generated power values. We expect the predicted value to pass in the most densely populated areas in the plot. This is generally true, however is not always the case. The reason for this is the high scatter of the data in the central regions of the daily power generation profile. The modelled behavior follows our expectations in the tails of the profile where the generated power values are more concentrated and better follow our expected behaviour.

On the other hand, the high scatter of the data and the behaviour learnt by the models imply that the consideration of an absolute error value, as a measure of inefficiency is not accurate (the absolute error can be high for both the most efficient and most inefficient panel sets). To ensure the high error value represents only inefficiencies we need to limit its calculation to the points that fall below the predicted DC power curve. We thus modify the prediction error calculation in a way as to include only the points below the predicted curve.

$$E_i(t) = \begin{cases} (P_{i,pred}(t) - P_i(t))^2, & \text{if } P_{i,pred}(t) > P_i(t) \\ 0, & \text{otherwise,} \end{cases}$$

where i is the index of a single panel set and t is the time. We will use $E_i(t)$ as a measure of inefficiency of the panel sets and derive the anomalies from it.

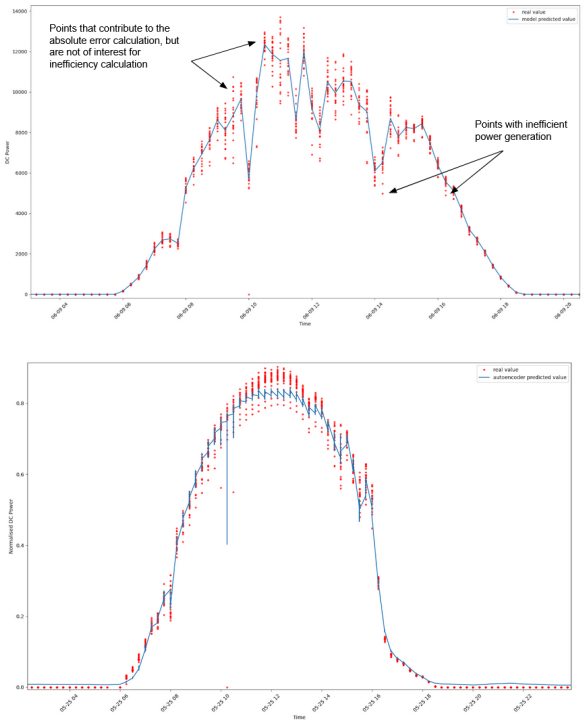

Figure 14. Comparison of real and predicted values of the generated DC power predicted using ambient variables (irradiation, ambient temperature, module temperature) for two sections of data. Top: Comparison of the prediction made by the decision tree model in a section of data with high scatter between different panel sets. It can be seen in the image that due to the high scatter the mean squared error will weight equally the points that are above the predicted curve (panel groups performing well)and those that are below this curve (panel groups performing badly). This limits the use of the mean squared error as an efficiency measure. In order to give a higher importance to the points that have relatively low power generation values the error value is set to 0 for the points that are above the predicted curve. Bottom: Prediction made by the autoencoder model for a section of data with high agreement (low scatter) between different panel sets. We observe that even in a section where due to low scatter the replication of the data behaviour should be relatively simple the autoencoder has not been able to learn to fully replicate this behaviour. Additionally the error value calculated for the points at the peak of the curve will be higher compared to the panel groups that are more efficient. This adds additional noise to the error value of the autoencoder models and makes its use as an inefficiency measure less reliable.

2.4.2 Autoencoders

We used the same 4 variables to train autoencoders to represent the relations between the variables. We tested a variety of architectures for the autoencoder, starting from very simple to highly complex architectures (shown in Fig 15). We found that the high complexity cases were unstable making it impossible to use them for our purposes. The reason for

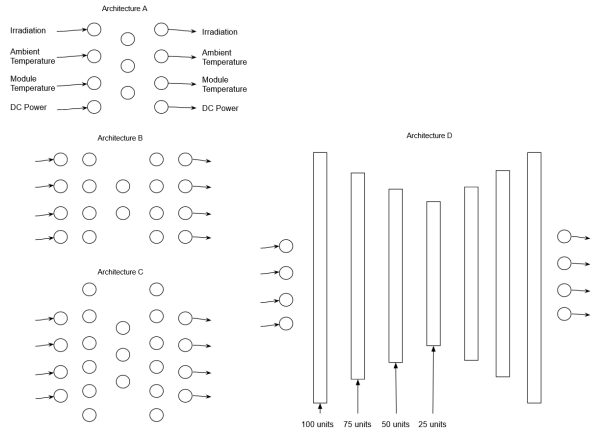

Figure 15. Autoencoder architectures tested in this work. Only four signals are used as input and output to specialize the autoencoder in anomaly detection in the DC power. Simple (Architecture A) medium (Architectuers B and C) and more complex (Architecture D) architectures were tested. The outcome of simple and medium architectures was almost equivalent. The more complex architecture exhibited high instability and was discarded.

the instability could be the small amount of the data that is not enough to help constrain the large number of parameters present in complex networks. We thus select only one of the more simple architectures for further analysis since the results of the stablec architectures are generally in agreement with each other.

It is important to note that the problem with the scatter in the data described for the regression models is also present for autoencoders. Moreover the autoencoder often does not even fit the data fully (see Fig. 14). At the same time, the error adjustment made for the regression models is not applicable to autoencoders, since for autoencoders the reconstruction error is calculated based on all of the variables used for the fit and cannot be limited to just one. Thus we leave the absolute error calculation as an inefficiency measure for the autoencoder and will use it as a measure of inefficiency for panel sets.

$$E_i(t) = (P_{i,pred}(t) - P_i(t))^2.$$

To summarise, the observation of the quality of the fit of the regression models and the autoencoder allowed to conclude that the error calculation should be modified in order to actually measure panel set inefficiency. Moreover, such an adjustment is impossible to make for the autoencoders and this in addition to the inadequacy of the fit of the autoencoder could make the results derived from it considerably less reliable. In any case we will present them in Section 3 in order to compare with the results of the other methods.

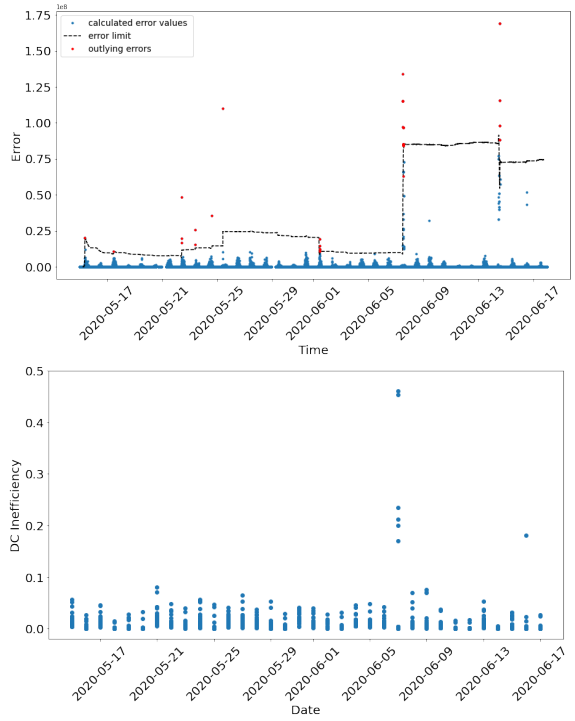

Figure 16. Top: Dervation of points with particularly high reconstruction error value. These points should be equivalent to the points where power generation suddenly halts during daylight hours. The blue dots represent the reconstruction error value calculated for each data point. The black dashed line shows the outlier limits calculated by using the standard deviation of the error values in the past week (10 standard deviations from 0). The red points are the outliers selected in this way. Bottom: The daily mean error (modified error definition for the regression models) value calculated for each panel set. This measures are used to compare panel set performances in the same day or in time.

2.4.3 Derivation of anomalies

Having defined an error value (inefficiency value) for each timestamp and each panel set we can proceed to identifying the 4 types of anomalies we defined in the last section.

Outliers as isolated anomalies. We use the mean and the standard deviation of the error value in a rolling window to derive outlier limits for the entire period. The points with error values that are above those limits are considered to be anomalies.

Panel set daily inefficiencies. We calculate the daily mean error for each panel set $E_i = \overline{(E(t)_i)}$, as a measure of overall daily inefficiency level. Our expectation is that the panel sets that have systematic inefficiencies will on average have a higher daily error value.

After calculating a daily inefficiency measure using the error value, we applied the same logic described in Section 2.3.2 to derive

- the panel sets that have relatively high inefficiencies compares to the rest in the same day
- the panel sets that show this behaviour consistently over a set of consequitive days
- panel sets that show trends of decreasing efficiency over a few days
- panel sets that show sudden falls of efficiency in the passage from one day to another

We adjusted the described functions in order to select the highest (or growing) values instead of lowest (and decreasing) values that were used in Section 2.3.2.

This concludes the description of the approaches we used to derive inefficiency signature in the power generation data. In the following we will present the insights we got from the analysis both from the baseline and model-based approaches.

3. Results

We have developed throughout the work a few approaches for finding anomalies in photovoltaic plant data. Having available only data from two different plants, we developed the approaches using exclusively the Plant 1 data and then tested them on Plant 2 in order to check how generalizable were the developed methods.

When thinking of the use of the final algorithms in a real-life situation two approaches can be used in order to extend the models to any new dataset:

1. Both model-based and baseline approaches used in this work are constructed in real-time logic[8] and make use only of the data to the past from a given point. This means that all of these methods can be trained and launched directly on any new plant data to derive anomalies on a daily, hourly or more frequent basis. In this way the history of the plant itself will be used to derive the reference 'normal' behaviour of the plant and there is no need to use an independent pre-trained model (see schematic in top image of Fig. 17).

2. Another possibility is to create a pre-trained model on a single dataset and use a small amount of new plant data to fine-tune the pre-trained model for the new setting. After this the fine-tuned model can be used on the new plant for the rest of the data (see schematic in bottom image of Fig. 17).

[8] This is only partially true for the model-based approaches, since due to lack of data we were forced to train the models on the entire months worth data. The subsequent anomaly derivation logic however has followed a real-time logic.

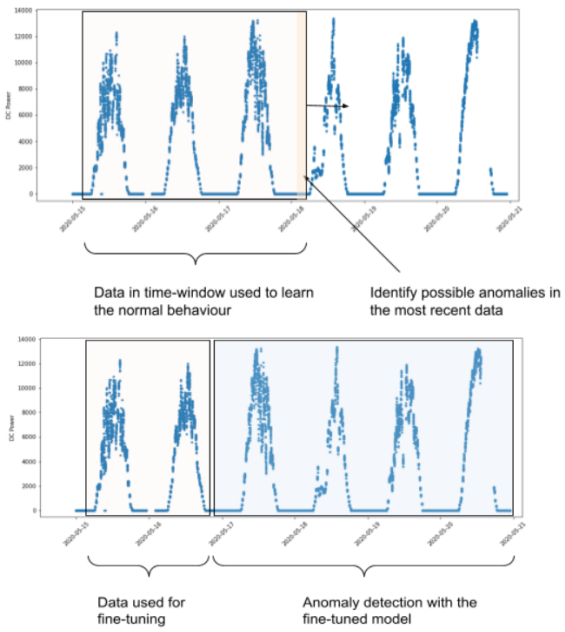

Figure 17. Graphical representation of the possible ways to apply the approaches developed on Plant 1 on Plant 2 or any future new dataset. The approach represented in the top part describes a process in which the model is re-fitted on the past data on every new run and derives the performances and inefficiencies based on the newly learnt behaviours. The process described in the bottom plot assumes the direct use of the model fitted on Plant 1 after a fine-tuning process done on a small amount of data from the new dataset. For anomaly detection purposes the first approach is preferred, as it is simpler both for implementation and deployment. It also ensures a better learning of the new datasets behaviour thus a better performance of the anomaly detection model.

Given that for an anomaly detection task a strictly predictive algorithm is not necessary, the second option only adds complexity to the final implementation and might not necessarily offer any potential improvement. The first approach is recommended, since it is lighter to implement and run, is better specialized for each individual plant data and can handle behaviour that changes in time, such as variation of 'standard' behaviour during the year.

Thus our approaches have been developed and applied with this deployment logic in mind.

We present the results of our analysis in the following sequence:

- Insights from data exploration
- Results of the baseline anomaly detection approach
- Results of anomaly detection approaches using a set of regression models or autoencoders
- Comparison of the results for the two plants and the scalability of all the methods studied in this work

3.1. Insights from data exploration

A detailed discussion of the exploration process and its results have been given in Section 2.2. We would however like to underline a number of points that are important for the evaluation of the approaches. Specifically, we observed in Section 2.3 that the effect of the weather and ambient variables on the generated DC power is not merely linear. The DC power generated is dependent both on solar irradiation and the panel temperature (here approximated with module temperature), with the latter having a negative effect on the generation efficiency as it increases. The panel temperature in its turn depends on both the irradiation levels and the outside temperatures. Thus, generally speaking, the model-based approaches that take into account more variables are more suited to fully exploiting all the patterns present in the data for a more accurate evaluation of the performance of panel sets.

The baseline approach developed by us cannot account for these complex dependencies (it actually ignores the sections of data with complex relations) and could potentially evaluate the panel group behaviour inaccurately in specific circumstances.

On the other hand the exploration of power generation curves in both plants shows an overall efficiency difference between the two plants (Plant 2 generates power at lower amplitude compared to Plant 1). This means that a model trained on one of the plants cannot be applied directly on another plant without an initial fine-tuning (scenario 2 discussed in the beginning of this section). The baseline model developed by us is invariant to DC power amplitude differences, although generally might also necessitate some fine-tuning of the limits used in the method. We have not been able to confidently evaluate the invariance of these limits since having only two datasets from different plants limits our ability to test this aspect. In any case, in the sense of the invariance to the amplitude of DC power values the baseline approach has an advantage with respect to the models.

Our analysis also showed a serious scalability problem for the model-based approaches. A very large number of outliers present in Plant 2 made the application of those approaches impossible. The main assumption for using machine learning models for anomaly detection is that most of points in the data behave normally. When this assumption is true, a small percentage of the point with anomalous behaviour can be identified by looking at the differences of expected behaviour learned by the model from most of the points and the actual behaviour of the anomalous points. In case of Plant 2 (see figures 21, 22) these points are too many and become important for the fit of the machine learning model. This puts a very strong limitation on the use of

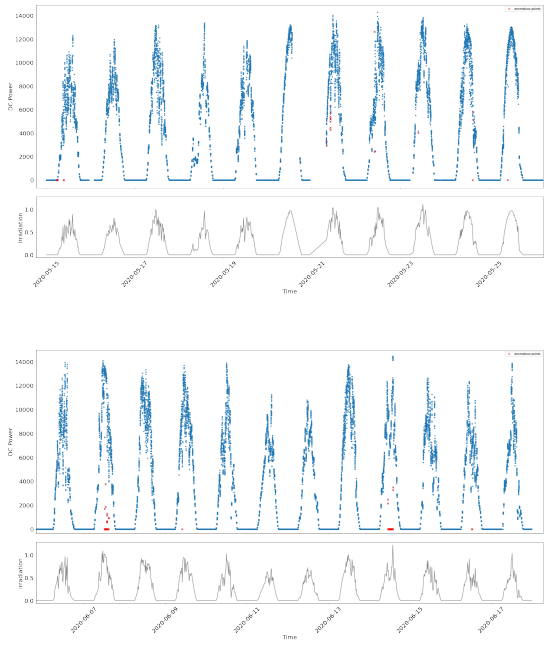

Figure 18. Two sections of DC power generation data showing the derived outlying points for Plant 1. They represent single points with anomalous values of efficiency and often include the situations in which the power generation halts suddenly during the daylight hours.

model-based models. A possible workaround could be removing those points before the use of the model. This has not been addressed in this work due to time limitations.

3.2. Inefficiency detection using the baseline model

We discuss the results obtained through the baseline approach developed by us both for Plant 1 and Plant 2. We will use them as baseline outcome or as 'truth values' when evaluating the performance of the model-based approaches. While doing this, however, it is important to remember that we potentially expect the model-based approaches to be better suited to describe the data and in some cases their disagreement with the baseline data could be justified. These cases would have to be manually evaluated and confirmed as such.

Plant 1

Outliers as isolated anomalies These are sudden changes/drops in the power generation that are not caused by weather variations. There are mainly 5 sections in which one or more panel sets have anomalously high DC power value (see Figure 18). These do not represent any physical interest for us. The anomalies we are interested in and that would eventually lead to alarms are those with particularly low DC power value. There are an overall 10 sections where one or more panel sets show this behaviour. We show these sections in the Figure 18.

Finally, almost all of the methods are prone to boundary effects, where the growth of the DC power curve on the very first day is considered anomalous by the approach, since this is the first time this behaviour is observed in that time range. We ignore those points as it is very easy to differentiate and remove these type of effects.

Panel set daily efficiencies

After deriving the panel set efficiencies as described in Section 2.3.2 we selected panel sets and periods in which the sets efficiencies had the following patterns:

- Panel sets that show lower efficiency on a selected day relative to other panel sets
- Panel sets that show low efficiency behaviour systematically (in a given time window)
- Panel sets that show sudden drops of efficiency
- Panel sets that show decreasing efficiency trends

Panel sets with low efficiency

In Fig. 10 we already showed the final distribution of panel set efficiencies over the period of study. Two panel sets that show low efficiency almost every day are clearly distinguishable in the plot. Indeed according to our analysis and the functions derived in Section 2.3.2 these panel sets are often selected as showing low power generation efficiency with respect to the other panel sets in equivalent weather conditions. These can be better observed in Fig. 19 where the two anomalous panel sets are coloured differently with respect to the rest of the panel sets. The latter only rarely exhibit relatively low efficiency values.

In some cases the power generation difference for these panel sets is so evident that it can be observed directly on the DC power generation curves (see an example section in Fig. 19). The difference between the generated power values between panel sets is particularly clear in the tails of the daily DC power profile (where the scatter among the different panel set values is minimal). This is possibly a factor that makes our method particularly sensitive at detecting these inefficiencies in these sections of the data.

The days on which these panel sets are not identified as the least efficient of the day are the ones with the higher overall variability among the panel set efficiencies, where the behaviour of those panels sets does not stand out as much.

Decreasing trend in efficiency

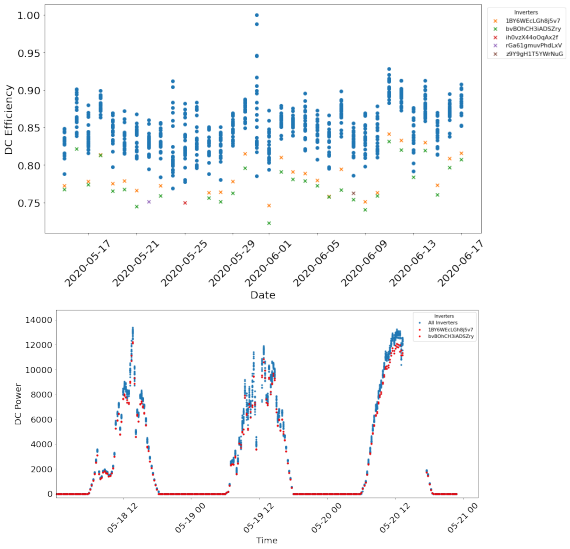

Figure 19. Panel groups selected as low performing in separate days for Plant 1. Top: The daily efficiency (blue dots) of panel groups connected to individual inverters in the period analysed. The coloured crosses mark the panel groups with low performance in separate days. Two of the panel groups show low performance on multiple days. Bottom: The DC power generated by the two panel groups that perform badly for multiple days. The lower value of generated DC power for these panel groups is clearly visible in the raw data for some days.

From our analysis three panel sets show a falling trend in the efficiency (Fig. 20). It is, however, noteworthy to observe that the three panel sets exhibit this decreasing trend in a correlated manner. This could be an indication that this trend is somehow related to weather effects, since for patterns unrelated to weather effects we do not expect to observe correlation between panel sets. On the other hand, patterns created due to weather effects are expected to be found in the majority of panel sets not just few of them, so the observed behaviour could also be a result of connections present between various panel systems.

Efficiency day-to-day drops

Finally we observe six occasions on which the panel set efficiency suddenly drops. We ensure those drops happen only to single panel sets as observing such behaviour for groups of panel sets could also be an indication that the behaviour could be weather related. We note that this type of anomaly can generally include also cases of the previous type, if the decreasing efficiency trend happened at very high speed. This actually the case for one of the panel sets appearing in both top and bottom images of Fig. 23.

Like in the previous case, we have not devised ways to be able to check with confidence that this type of behaviours are not biased due to weather. In this light, among

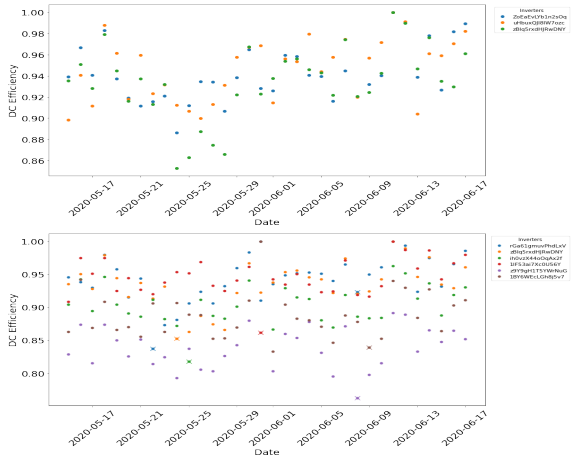

Figure 20. Top: The power generation efficiency for the three panel groups of Plant 1 that were selected as showing falling trend in performance. The behaviour is clearly visible in the period from 17 to 25 May. Bottom: Panel groups from Plant 1 that showed sudden performance drops. The days on which the anomalous drop was observed are marked by coloured crosses. The colours indicate the specific panel group.

the anomaly types that we derived, the most robust against the weather related biases are the sudden falls of power generation (outlier points) and the detection of panel sets with relatively low efficiency in a day, since the latter is derived by comparing different panel sets between themselves while they are in common weather conditions. We will thus consider mainly those inefficiencies when comparing results between approaches.

Plant 2

Before presenting the anomalies derived through the baseline method for Plant 2 we remind that this method was developed by considering only the data available from Plant 1. After the development, the prepared pipeline was applied directly on Plant 2 data proving that our method can be easily applied to completely new and unseen data.

Outliers as isolated anomalies

The number of outlying points in Plant 2 is very high. It is clear that there is a systematic problem in Plant 2 leading to frequent power conversion cuts during the day (see sample of data in Fig. 21).

Panel set daily efficiency

There is no systematic behaviour of low efficiency observed among panel sets of Plant 2. Only one panel set seems to fall often behind in efficiency. Similar to the case of Plant 1, also for Plant 2 the power generation gap can be easily observed by looking at power generation profiles of the panel sets (Fig. 22).

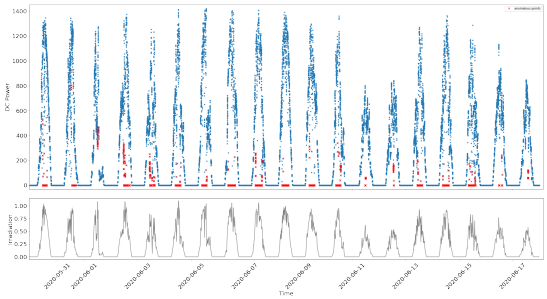

Figure 21. A section of DC power generation data showing the derived outlying points for Plant 2. They represent single points with anomalous values of efficiency and often include the situations in which the power generation halts suddenly during the daylight hours. The presence of some systematically occurring panel group shut-downs is clearly visible. This behaviour could also be due to some maintenance works done on the plant, during which the panel groups are shut down for checks.

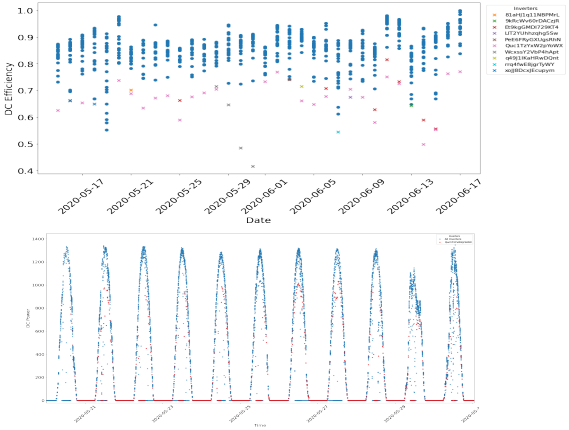

Figure 22. Panel groups selected as low performing in separate days for Plant 2. Top: The daily efficiency (blue dots) of panel groups connected to single inverters in the period analysed. The coloured crosses mark the panel groups with low performance in separate days. One panel group shows low performance behaviour on multiple days. Bottom: The DC power generated by the selected panel group. The lower value of generated DC power is clearly visible in the raw data for some days.

Decreasing trend in efficiency

We observe one of the panel sets follow a clear behaviour with falling efficiency in the course of six days in the time period in from 25 to 29 May (Fig. 23). This is the only panel set showing such behaviour and its very high negative inclination gives some confidence that this behaviour is not due to any weather effects. It would then be interesting to clarify what could potentially be the cause of such efficiency falls as well as the reason for the eventual recovery

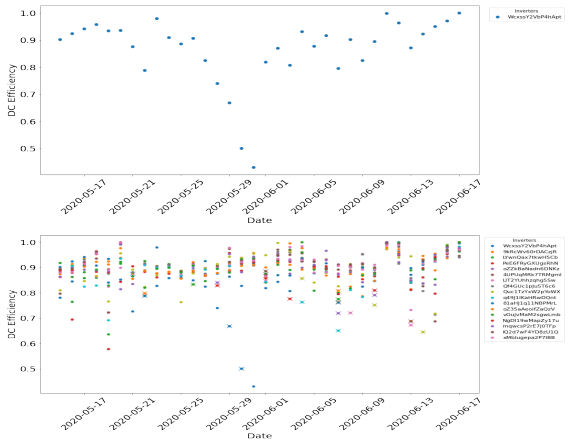

Figure 23. Top: The power generation efficiency for one panel group of Plant 2 that was selected as showing falling trend in performance. The behaviour is clearly visible in the period from 25 to 29 May. Bottom: Panel groups from Plant 2 that show sudden performance drops. The days on which the anomalous drop was observed are marked by coloured crosses. The colours indicate the specific panel groups.

of power generation efficiency.

Efficiency day-to-day drops

We finally present the panel sets that exhibit sudden day-to-day efficiency falls for the period observed in Plant 2 in Fig. 23. Like in the case of Plant 1, the panel set showing the falling trend is also included among the cases with day-to-day efficiency falls.

The results presented in this section show that the baseline approach that was developed based on simple principles but with consideration of the particularities and the expected behaviour of the data has proven to be a promising algorithm that can eventually be used in practice to quickly identify problems in solar energy production plants. Moreover, the successful application of the approach to Plant 2 data shows that the logic and the arguments used for the method development and hence the method itself is easily generalizable to new data.

The main weak side of the baseline model is its incomplete treatment of weather variables and so the danger that some of its results might be biased due to weather effects. For some of our conclusions we are able to ensure high confidence by making weather invariant comparisons. For others a deeper analysis (non-linear models to describe DC to Irradiation dependence for efficiency derivation) and perhaps a dataset covering a wider range of variables would have been necessary to confirm the invariance of some types of anomalies with respect to the weather conditions.

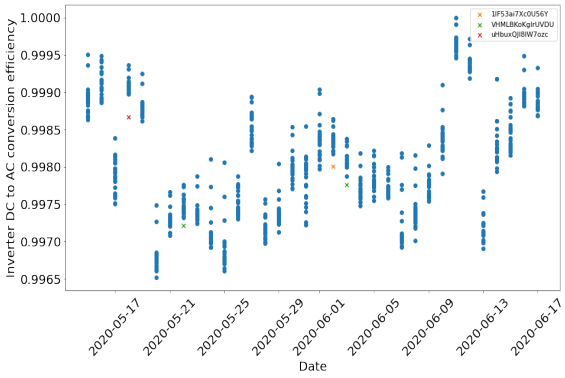

Figure 24. The DC to AC power conversion efficiency for inverters of Plant 1. The inverters selected as low performing on specific dates are marked with coloured crosses.

3.3. AC power conversion inefficiencies

In this section we briefly summarise the anomalies derived for the inverter performance by studying the conversion efficiency of DC power to AC power in Section 2.3.3. We do not concentrate on the power conversion behaviour in this work, but analyse it briefly to ensure to have a complete view of the anomalies we select and the possible effects they might have.

Like detailed in Section 2.3.3 after removing some clear problematic power conversion values we derive daily power conversion efficiency measures for every inverter. We then use those efficiency values to select any inverters that show low performance, performance trends or performance falls. In Fig. 24 we show the inverters that are selected as low performing on particular days. The fact that no anomalies have been detected in the DC power generation behaviour of those panel groups means that the detected anomalies are due to the inverter performance itself. On the other hand the fact that the panel groups that were selected as low performing in the DC power generation do not appear among these inverters reassures that the anomalous behaviour detected in the DC power generation behaviour was really due to panel performance (with no possible cause from the inverter itself).

3.4. Anomaly detection using Machine Learning methods

Having now discussed the outcome of the baseline approach we can compare the results of all the other methods to it. For the comparison we rely mainly on comparing the following types of anomalies:

- The single anomalous points selected as outliers (sudden falls of energy generation curve)
- Low efficiency days selected for individual panel sets

Outliers as isolated anomalies

In Fig. 25 we present the distribution of the single outlier points according to the model-based approaches and the baseline model.

In general, the baseline model seems to be more generous in the selection of the outliers, even if this is very dependent on the choice of the limits for outlier selection. In the sections where the most clear power generation falls have taken place all the models seem to have captured the problem sooner or later. There is, however, also a large amount of points that have been selected as outlying by the models and that do not necessarily coincide with any of the cases selected by the baseline model.

At this point it is difficult to judge about the reliability of the results without any input from field-experts. For the selection of the right approach for alarms and warnings given in case of a single outlier points a decisive factor for the selection of the approach are the preferences of the user and the feedback of the user regarding the frequency and accuracy of the alarms that are set off. Thus we consider that the final decision on the selection of the model and the best fitting limits is impossible at this level and that all of the approaches used are able to single out the most important points, those of complete fall of power generation during the day.

Daily inefficiency of individual plant groups

When looking instead on the daily selection of the least efficient panel sets (Fig. 26) the simpler models (regression and decision tree) seem to be in agreement with the baseline model. The reason for this is the correction of the error calculation that we made in Section 2.4 for the regression models. This adjustment was impossible for the autoencoder, thus as confirmed by the figure the decision on the most inefficient panel set during the day becomes very unstable due to additional noises present in the autoencoder error value.

4. Conclusions and Discussion

In this work we used power generation data from two different power plants to study the possibility to define a data-driven approach that could be used in practice on multiple power plants to identify anomalies in real time. The eventual solutions that could be based on the proposed approaches would provide a support system for the plant experts raising warnings whenever necessary. These criteria defined at the beginning of the project traced the main range of requirements that a potential approach would need to satisfy. They were adjusted accordingly as we gained new insights from the detailed analysis of the data.

(a) (Original requirement) In order to continuously monitor the plants condition during the day the algorithm

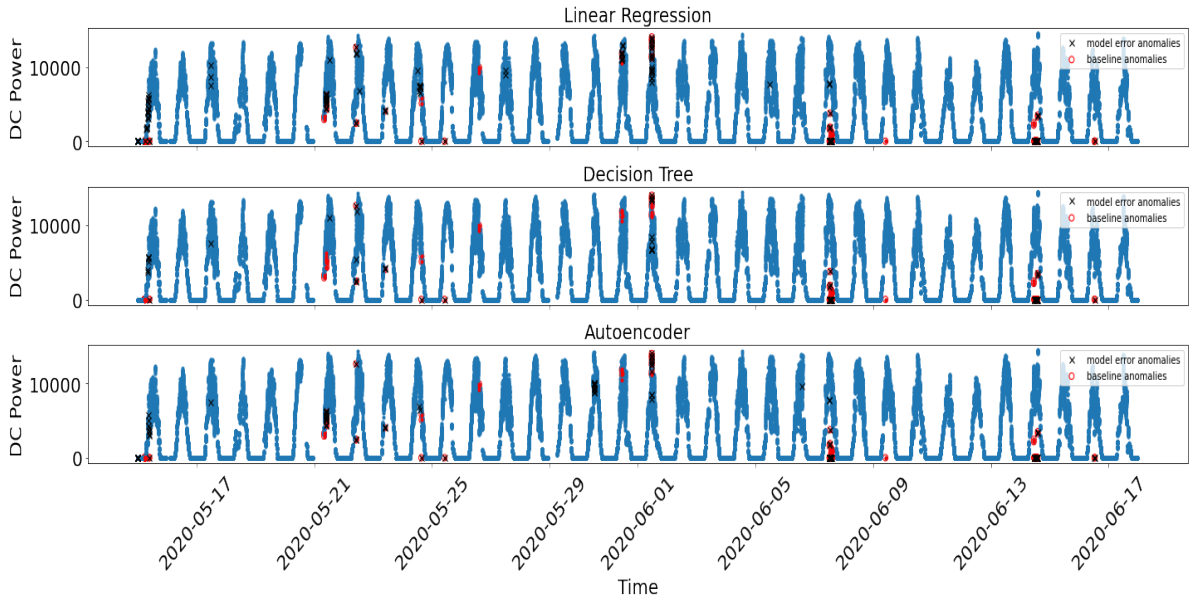

Figure 25. Comparison of selected anomalous performance sections for Plant 1 using our baseline approach (in red) and machine learning approaches (in black). It is clear that the methods perform equally in the detection of sudden energy production halts during the daylight hours. The differences between the three methods could be due to the differences of the quality of the fit of the three methods to the data and the ability to represent more accurately the data or the ability to account for the effect of all the ambient variables on power generation. Specifically the approach using a decision tree often does not agree with the selection of the baseline method for outlying points located near the peaks of the daily profile. This could be due to the fact that the decision tree is better able to account for the complex behaviour in the highest irradiation and temperature conditions where the power generation has non-linear dependence from the other variables and where the baseline model has limited accuracy. The group of anomalous points selected in the leftmost part of the data by all four methods (baseline method and three machine learning approaches) is a result of boundary effects and should not be considered.

would need to run periodically during the day (e.g. evaluation of the situation every 5 minutes) and thus the timescales necessary to complete a run of the algorithm pipeline (including any re-training, if necessary) should necessarily be below these values.

(Modified form after the data analysis) We found that the more relevant anomalies for the plants performance and health monitoring are anomalies that are observed on timescales well above the order of minutes (over 1 or more days). Thus, in general, the margins for the completion time of a single pipeline cycle can be larger. Short time scales would be of higher importance in cases when the users need specifically to be warned of any sudden changes in panel behaviour (for example sudden halts in panel power generation found in the data).

(b) An eventual solution based on the proposed approach (hence the approach itself) should be easily applicable to any new plant, thus should be easily extended and functional on any new dataset derived from power plants of various scales, configurations and working in a variety of conditions.

(c) Ideally user defined accuracy measures would also be included among the main requirements for the approach, however not having access to any feedback from potential users we defined some general criteria for the frequency and accuracy of the warnings that the eventual system would need to have. The amount of alarms that the user receives from the system should be manageable and useful. If some anomalies are observed frequently in the studied data, they should either not be considered for warnings or should be only assigned low priority warnings. Less frequent anomalies that we evaluate to be relevant (high confidence of anomaly, scale of the effect on the performance levels and eventual costs associated with the performance losses) are instead anomalies that should receive the most attention from the experts and are the best candidates for the eventual warnings.

We evaluated multiple approaches in terms of these criteria along with the technical evaluation of the approaches applicability and results. The approaches included:

- a relatively simple data analytics method (our baseline approach),

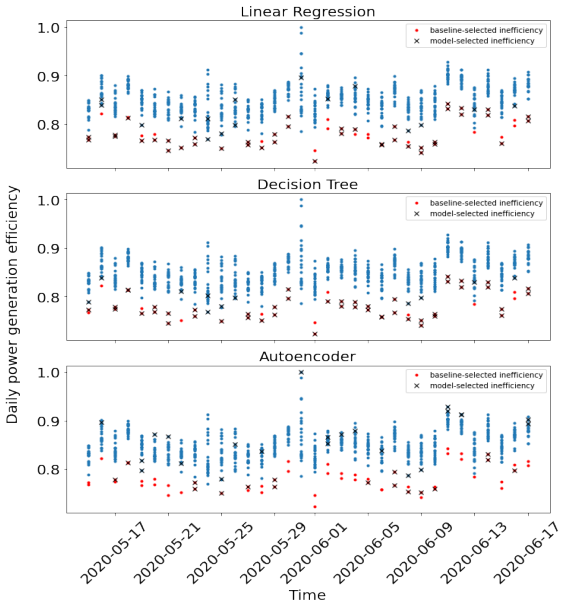

Figure 26. The agreement between the baseline method and the machine learning approaches regarding the selection of panel groups with low efficiency. The blue dots represent the efficiency value as calculated following the baseline approach[10]. The red dots are the daily performance values selected by the baseline methods that result to be particularly low with respect to the values of the rest of panel groups in the same day. The black crosses show the performance values selected by the machine learning approaches using error-based inefficiency values. The inefficiencies were calculated based on a corrected version of error (see the equations at the end of Section 2.4) for the cases of Linear Regression and Decision Tree. Due to this reasons these methods show better agreement with the baseline model. The error correction was not possible for the autoencoder model and this explains the difference between the results of the autoencoder and the other regression models.

- and more advanced approaches using machine learning methods (regression with machine learning and autoencoders).

Since the panel power generation efficiency can be very different between different plants, the machine learning models will need to be trained on the individual plant in order to be able to evaluate the anomalies effectively. This does not introduce additional complications, since with the data available to us, a retraining over the history of 1 month for a single plant took an order of few minutes at most. This means the machine learning approach can in practice be used to even make real-time monitoring of the plants performance.

On the other hand, a number of technical issues were encountered in the approaches using machine learning models. For example the complexity of the power generation behaviour and the variability between the single panel sets in the same plant would not allow the models to define and reproduce a 'typical' efficient behaviour. This is a fundamental step before being able to select the anomalies based on their differences from the 'typical' behaviour.

The baseline approach did not suffer from this effect, since we had greatly simplified the treatment of the data during its development. Thus, the baseline approach was able to reproduce well the more common behaviour in the less complex (less scattered) sections of the data. Understandably this also limited on some level the reliability of the baseline approach since it was not considering the sections of the data with the most complex behaviour. Nevertheless the behaviours found with the baseline approach were studied in detail and qualified as likely candidates for alarming anomalies. The baseline approach was also extended to the new unseen dataset with ease and without any need to fine-tune parameter choices [11].

Using the baseline method we were able to identify a large number of cases where the individual panel sets' power generation efficiency drops dramatically or falls to 0 in short sections of time (see examples in Fig. 25). These were particularly numerous for Plant 2 where these shutdowns happened for multiple panel groups in a single day. From these type of anomalies we suggest to raise alarms only for those that show a dramatic drop, but do not fall completely to 0, since the complete shutdown could also be a result of maintenance works (see for example the cases identified on 22 and 23 of June by all of the approaches in Fig. 25). This however is less likely to be the case for Plant 2 where the power shutdowns seem to have regular nature but do not show any patterns resembling maintenance works.

Apart from instantaneous anomalies we also defined a way to measure the daily efficiency of individual panel sets and derived anomalies among these values. In both plants we observe one or more panel set that often shows lower performance compared to the other sets. This is particularly evident for Plant 1, where two panel sets (see Figures 19 and 26) seem to underperform most of the days in the 30 day period. A similar, but less frequent underperformance is observed on only one panel set for Plant 2 (Fig. 22). If the observed lower performance behaviours are constant (but for some reasons this was not caught by our algorithms) then they could be a result of configuration differences between panel sets (such as positioning, inclination angle, or

[10] Note that the representation of the DC efficiency in the plot is according to the baseline model. It thus does not represent the values of inefficiencies derived using the model-based methods. The image should be used only for comparison of panel sets selected as low efficiency according to various methods.

[11] This was only observed using the two plants we had at hand and a fine-tuning could still be necessary in case of use on plants with very different behaviour in the future.

outside factors that for some reasons are fixed for the specific panels and panel sets). In this case additional checks would need to be implemented in order to silence some type of alarms for those panel sets. A constant underperformance could also be caused due to some degradation or malfunction that had triggered a performance drop at earlier stages. If the performance differences are are not stable, as suggested by our method, then they can be caused by the differences in panel properties between the panel sets (e.g. the panel sets with lower performance could be composed of panels that get heated easier or whose power generation is affected easier by module temperatures due to panel degradation). In the last two scenarios the warnings would be justified and would help notify about the necessity of maintenance checks to find any faulty behaviour or degradation in the interested panel sets.

We calculate that due to the lower efficiency of the two panel sets of Plant 1 the overall energy produced by each of them over the studied 30 day period is around $9 - 10\%$ less compared to the average produced by the remaining panel sets. This means each of the panel sets induces a potential $9 - 10\%$ production loss with respect to the production levels of panel sets behaving normally.

Another anomaly worth noting is the steep decrease of the efficiency observed for one of the panel sets in Plant 2 (top plot in Fig. 23). We do not know the reasons for the performance fall and its eventual recovery (could have been a control and maintenance made after noting the anomalous behaviour), however the alarms designed by our approach would have been raised on the last day before the reset.

To further develop the baseline model we suggest the use of more complex non-linear approximations to the DC to irradiation dependence, in order to be able to account for the sections of data that are not described with simple linear behaviour.

The simpler machine learning approaches although mostly in agreement with the above results, however were hard to apply to the new unseen datasets due to differences present in the data quality between the sets (a fact that did not pose an issue for the baseline approach). We thus suggest a number of improvements (e.g. better cleaning of the data before the application of the models) that could be made in the future in order to make the machine learning approach more scalable.

We think that with the suggested improvements the two approaches could eventually merge into one that uses simple linear regression with two or three features on a data that passes through a thorough data cleaning step beforehand (like was done in the baseline approach).

Overall we think that given the restricted dataset and times for the analysis the studied approach and the results suggest that with the application of the above recommendations and after testing on a larger set of plants the analysis can give beneficial results. A field expert evaluation would be necessary in order to judge and compare the advantages that this method could offer from the point of view of the potential user.

Acknowledgements

The authors would like to thank Prof. Artak Hambaryan for the discussion and valuable insights provided to help explain and assess the results of this project.

References

[1] How do solar panels work? the science of solar explained. URL https://solect.com/the-science-of-solar-how-solar-panels-work/.

[2] R. A. Abdoulatif Bonkaney, Saïdou Madougou. Impacts of cloud cover and dust on the performance of photovoltaic module in niamey. *Journal of Renewable Energy*, 2017(1), 9 2017. URL https://www.hindawi.com/journals/jre/2017/9107502/.

[3] G. Knier. How do photovoltaics work?, 8 2008. URL https://science.nasa.gov/science-news/science-at-nasa/2002/solarcells.