Data mining for automatic fault detection and diagnosis from photovoltaic monitoring data

Asset management | The timely and skilful interpretation of performance data from the monitoring of operational PV power plants is vital to improving the management and thus profitability of those plants over their lifetime. Achim Woyte shows how data mining and artificial intelligence can serve the management of solar assets

Professional photovoltaic plants today are virtually always monitored. Asset managers collect operational data from heterogeneous portfolios of plants. The data originates from on-site sensors and inverters. They are recorded by local dataloggers and sent to a central monitoring platform. Such platforms include a database, dashboards for supervision, operations and maintenance (O&M) and reporting purposes, analytical tools and data export functions.

3E operates the hardware-independent performance monitoring and reporting platform SynaptiQ. From the monitoring data of our customers we see that, on average, their PV plants perform as well as expected. SynaptiQ lets them continuously improve the plant availability and performance while streamlining their business processes. At the same time, performance ratios (PRs) are widely spread, even for plants from the same portfolio. Although monitored, many plants perform far below expectation (Figure 1). Obviously, these plants are not managed as well as they could be.

Monitoring is more than collecting data and aggregating them into contractual and financial key performance indicators (KPIs). Probably, the bottom tier plants in Figure 1 are followed by an operator and their KPIs are reported regularly. To the asset manager, their overall performance must look weak but not yet alarming.

Performance monitoring allows O&M contractors to increase their business efficiency through fast fault detection and focus on actual faults and solutions. It serves asset managers to see what's going on at the plant and device level and how fast their O&M partners intervene. In case of component failure or module degradation, they can identify and prove causes for warranty claims.

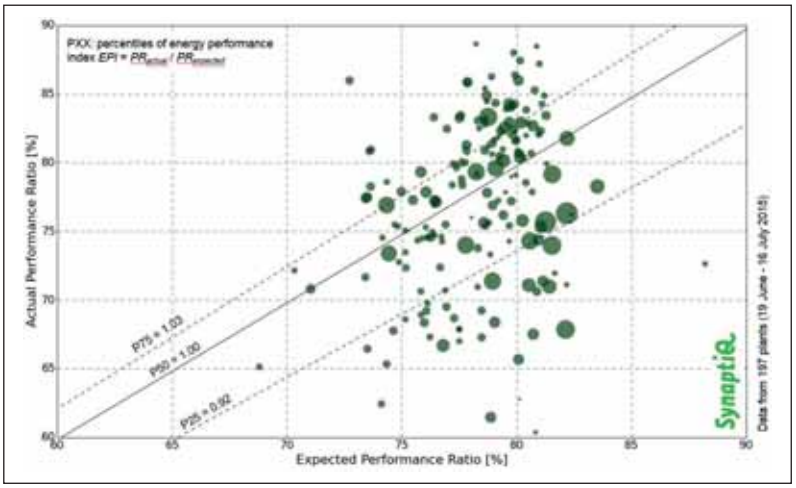

Figure 1: Actual performance ratio (PR) from monitoring versus expected PR based on plant specification and real weather data for 197 PV plants in Europe; the worst 25% perform more than 8% below expectation; the dot size represents the size of the PV plant

Back in 2012, 3E launched an extensive programme to introduce automatic fault detection and diagnosis into its monitoring tools. Today 3E offers the PV Health Guard as a monthly or quarterly fault report to their SynaptiQ customers as well as a one-time Historical PV Health Scan, e.g., before the end of the warranty period or for plants changing ownership.

Approach

Asset managers can easily implement automatic fault detection and diagnosis functions themselves as a step of data post-processing. The data is exported from the monitoring database and can then be mined with spreadsheets, scripting tools or dedicated data mining packages. However, particularly for large portfolios, plants with many inverters or with string monitoring, this task becomes quite challenging due to the complexity of the underlying data structures as well as the sheer size of the data sets.

3E uses the Python programming language. The application programming interface (API) of SynaptiQ allows the user to directly query SynaptiQ's monitoring and plant configuration databases with Python. The API can be made available to customers.

Automatic fault detection and diagnosis are common activities in the condition monitoring of industrial processes and machines. The energy conversion process we want to monitor is illustrated in Figure 2. The different measurements as indicated may be interpreted as external process variables. The PV Health Scan for

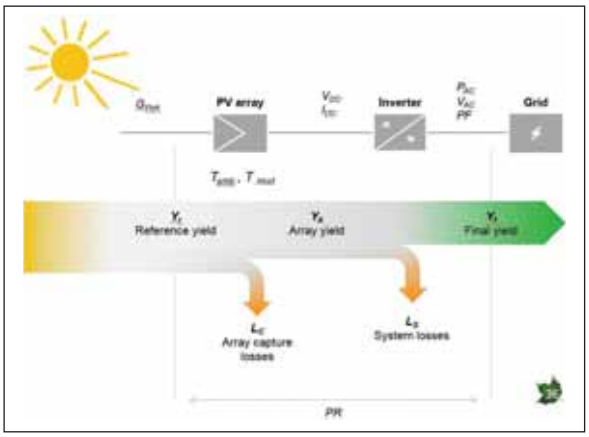

Figure 2: Energy flow in a grid-connected PV system; with measurements of plane-of-array irradiance (GPoA), DC voltage and current (VDC, IDC), AC power, voltage and power factor (PAC, VAC, PF), ambient and module temperature (Tamb, Tmod); yields (Y), losses (L) and performance ratio (PR)

this process consists of the steps analysis, fault detection and diagnosis. Analysis allows the plant operation to be reviewed in detail but it does not include any evaluation or decision step. Fault detection tells us whether anything is wrong, what is wrong and where it is wrong. And diagnosis tells us why it is wrong and how we can solve it.

Analysis

We start the analysis step with a data integrity check, removing outliers and identifying periods of missing data. Local irradiance sensors are validated by comparing their measurements to satellite-based irradiance. We then compute the high-level KPIs, performance ratio and energy-based availability for the PV plant and its individual arrays/inverters. In a second step, we compute the losses over the energy conversion chain for the plant and its arrays/inverters. We show how these loss components behave over time and whether they differ for the different arrays/inverters. If the plant contains string monitoring, the different loss components are also computed for each string.

Moreover, we review the correlation of measurements for the individual samples recorded. Finally, if the dataset contains several years of data, we also review the structural degradation over time.

In short, the analysis step largely relies on the detailed allocation of losses over the energy conversion chain, over the different instances of each component type (e.g., strings, arrays, inverters) and over time. It creates value for the user through the quantitative details and their visual presentation.

For the analysis, we build further on the conventions, guidelines and recommended practices from IEC 61724 [1], the European Joint Research Centre in Ispra [2], the International Energy Agency's Photovoltaic Power System Programme (IEA-PVPS) [3] and SolarPower Europe's O&M Best Practice Guidelines [4].

Fault detection

Our approach to fault detection is model-based: we compare the process variables as measured in the field to their expected reference values based on a model of the process. A simple and frequently used method to do this is limit checking of the measured variable. The measured value is compared to the reference value. If a certain range around the reference value is exceeded, this indicates a fault.

For PV monitoring, evaluating the process variables directly as they have been measured over time is not very effective due to the often high noise. Instead, it is more promising to work with derived indicators that correspond to different parts of the process. These so-called features should be representative for the underlying process and uncorrelated with each other. When a feature is evaluated positive, i.e., a threshold is exceeded, we call this a symptom.

For the PV Health Scan we have developed several feature sets for different parts of the work flow. Features for the 'Data Integrity Check' are the daytime recording fraction, i.e., the fraction of the monitoring period during daytime for which measurements have been recorded, and the fraction of outliers over the total number of measurements. For the 'Solar Sensor Check', we use features to check the clock setting, the sensor orientation and its calibration (see example in Table 1). For the 'Performance and Loss Analysis' on plant and component level, we have defined features in line with the different loss components. For the 'Degradation Analysis' we use annual degradation rates.

The threshold values for the different features can initially be set based on expert knowledge. A better way is to compute the feature sets for a sufficiently large sample of healthy plants and then chose, e.g., the P5 and P95 percentiles for each feature as thresholds. Finally, by applying fuzzy logics for decision making or machine learning based on classification, it is possible to evaluate the features more gradually in line with their severity. This should contribute to improving the overall selectivity of the fault detection algorithm.

Diagnosis

While automatic analysis and fault detection are relatively straightforward, the diagnosis step is the most challenging. We look for a conclusion on the underlying root cause through a comprehensive analysis of the different symptoms. In practice, we see that the symptoms from different parts of the workflow, e.g., Data Integrity Check, Solar Sensor Check, Performance and Loss Analysis and Degradation Analysis, complement each other. A human domain expert can synthesise these symptoms and draw a conclusion based on human experience. The challenge is to make the machine evaluate and synthesise from the symptoms and return a few probable suggestions on the root cause.

Notably, the obvious analogy to the medical world is not solely semantic but also practical. A blood test returns a feature set consisting of concentrations of lipids, glucose, hormones, etc. If the reference range for any of these features is exceeded, this is flagged as being 'abnormal'. Medical imaging devices come with post-processing tools that check features like minimum thickness of a layer of tissue or optical density and raise a flag as well if the reference range is exceeded. Both tools perform a fault detection; however, the diagnosis is left to the physician.

We can apply different approaches to move from fault detection to diagnosis. 'Inference-based methods' are suited if the link between root causes and symptoms can be expressed through a known set of logical rules, referred to as knowledge base. By means of the knowledge base, a so-called inference engine can then compute the most probable root causes. In artificial intelligence, this kind of systems is commonly called 'expert systems'. Setting them up requires a realistic translation of domain expertise into the knowledge base, which easily becomes quite tedious.

'Classification methods' are suited if a sufficiently large empirical dataset is available for training. They are a family of machine learning methods. For the case of PV fault diagnosis, this would ideally be a set of monitoring data from many plants over several years along with detailed maintenance logs. Classification

methods can be applied without explicit knowledge of the underlying causalities. However, the preparation of a meaningful training set can be quite tedious as well. For a practical overview of the advantages and drawbacks of different machine learning algorithms, we recommend the documentation of the Python machine learning package scikit-learn [5].

Currently, 3E is exploring inference-based as well as classification methods. Inference-based methods work well for simple causalities. For example, in Use Case 1 below, the slope of the sensor calibration is significantly too low. In this situation, the sensor should be cleaned or calibrated. Other common faults like near shading or a wrong orientation can be excluded since they would cause a different set of symptoms. This reasoning can easily be formulated in logical rules. For the example of Use Case 2, deciding whether a reduction in array current is due to a disconnected string, module degradation or inefficient maximum power point tracking is less straightforward. The symptoms for these faults are quite distinctive and a PV expert should be able to read them. Nevertheless, it looks much more promising to implement this intelligence through machine learning than through an explicit knowledge base.

Both approaches are potentially very useful for fault diagnosis in PV. At the same time, their limitations become clear from the medical analogy. Artificial intelligence and data mining can point towards possible root causes; however, asset managers and O&M contractor will still rely on their domain experience and personal judgement for a long time. Automatic fault detection and diagnosis can simplify this work and help them to manage large and heterogeneous portfolios much more efficiently.

Use case 1: radiation sensor calibration

Case description

A 860kW rooftop installation in Belgium returned an annual PR of 82% which appears to be a normal performance. A Solar Sensor Check was run for the on-site irradiance sensor over the annual data set from June 2015 to May 2016.

Fault detection and diagnosis

The Solar Sensor Check evaluates the fault illustrators (features) as listed in Table 1. Where the fault illustrators are situated in the normal range, the observation is labelled 'OK'. Where this is not the case, a symptom is triggered. Finally, a human expert verifies the textual description.

	Fault illustrator [unit]	Fault illustrator value	Observation
Data recording: plausibility of maximum irradiance	maximum Irradiance [W/m^2]	1037	OK
Data recording: plausibility of minimum irradiance	minimum Irradiance [W/m^2]	0	OK
Data recording: completeness	daytime recording fraction [%]	99.98	OK
Total irradiation	Mean bias error [%]	-9.4	Measured irradiation too low
Clock setting	approximated time shift [min]	-15	Deviation in clock setting
Sensor orientation	estimated azimuth & tilt [degree]	4°, 13° -> 6°, 13°	OK
Sensor calibration: linearity	non-linear term	1.037	OK
Sensor calibration: offset	sensor offset [W/m^2]	-3.88	OK
Sensor calibration: slope	sensor gain	0.922	Slope of calibration is too low

Table 1. Example of fault illustrators (features) and diagnosis for a solar radiation sensor installed in Belgium, data from 1 June 2015 to 31 May 2016

Of the three symptoms indicated in Table 1, the low value of 'Sensor calibration: slope' is the decisive one for the diagnosis. The sensor systematically shows 92.2% of the satellite-based irradiance only. Accordingly, the linear regression line in Figure 3 approximates the real calibration of the sensor. Consequently, the measured irradiation over the period is 9.4% too low. This deviation may be due to severe soiling or bad calibration.

Moreover, a deviation in clock setting is observed. However, the clock setting error of -15 minutes is in the order of magnitude of the sampling period and does not disturb the measurement as such.

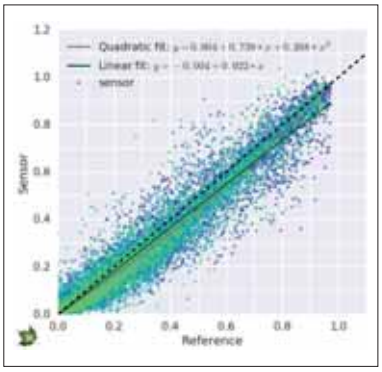

Figure 3. Sensor irradiance versus satellite-based reference irradiance with linear and quadratic regression; the slope is 8% too low; 860 kW site in Belgium, data from 1 June 2015 (blue) to 31 May 2016 (green), orientation 30° south

Economic impact

In line with the mean bias error as listed in Table 1, the sensor in this use case recorded 9.4% too little irradiation over the year. Accordingly, the performance ratio as computed with this reference yield is 10.6% too high. While the real PR for this year was a low 74%, the O&M contractor could report 82%. For the 860kW plant built in 2011, this over-optimistic performance evaluation hides a loss of €40 000 per year. With recurrent monthly sensor checks, this faulty calibration would have been detected after one month.

Use case 2: disconnected strings

Case description

After five years of operation, a 250kW rooftop plant in France was checked on behalf of a third-party investor. The plant contains three central inverters and no string monitoring. Amongst other things, a PV Health Scan revealed that, for one inverter, several strings had been disconnected for more than a year.

Fault detection and diagnosis

Figure 4 shows the high-level KPIs and losses per array/inverter. The overall PR of this plant for the study period of one year is low at 69%. The current-based losses are too high for all arrays/inverters, and especially for Inverter 1 with current-based losses of 17.8%. Accordingly, threshold checking of the overall losses triggered a fault for this array; however,

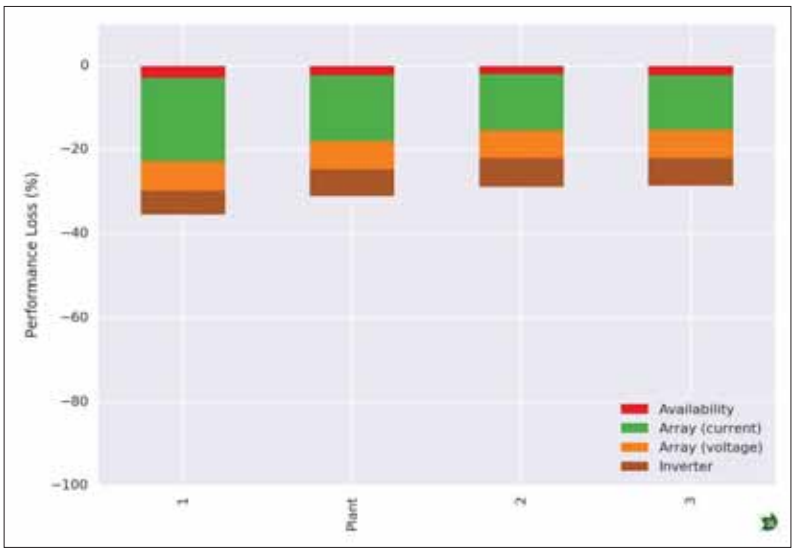

Figure 4. Performance losses split into loss types and stacked for each array/inverter and the entire plant, sorted in ascending order; particularly the current-based losses (green) are generally too high and worst of all for array/inverter 1; the losses are normalised to the reference yield. Data from 1 July 2015 to 30 June 2016

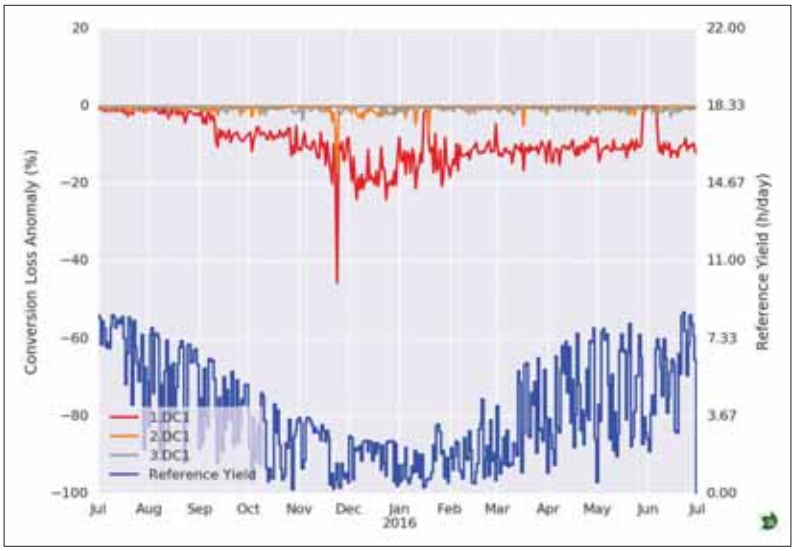

Figure 5. Conversion loss anomaly for current-based array losses per day, comparing the conversion losses of all arrays/inverters to the best in class of each day. The blue line shows the reference yield and hence the available irradiation for each day. Starting on 11 September 2015, the losses for inverter 1 (red) are much higher than for the others

it is insufficient to conclude on a root cause.

Figure 5 reveals that the conversion loss anomaly for current-based losses is almost constant over time. We can identify two distinctive events: on 11 September 2015, the current losses for array/inverter 1 dropped by -5.5% below the others. On 27 October 2015, the losses dropped further to -11%. This situation persisted until the end of the study period.

Array/inverter 1 counts 18 strings of modules. Accordingly, a disconnection of one or two strings would lead to a systematic power loss of 5.5 and 11%, respectively. These distinct conversion loss anomaly values, together with their relatively sudden changes, point towards one or two disconnected strings at this array/inverter.

Economic impact

Obviously, in the given case the O&M contractor did not see these string faults. Their effect on the overall performance is quite small, namely 11% on an array/inverter level and 3.7% on a plant level. Moreover, the O&M contractor did not act on the overall low performance of the plants.

For the owner, only the string faults caused a loss of approximately €6,000 per year. With recurrent monthly Health Scans, the string faults would have been detected and repaired after one month.

Outlook

Data mining and machine learning can boost the revenues from PV plant operation by up to 10% simply by making O&M more agile. Already today, readily available solutions for automatic analysis and fault detection can be plugged into the PV performance monitoring platforms.

The step from fault detection to automatic diagnosis is still challenging. The machine can suggest probable root causes for common faults and formulate recommendations. However, for the final interpretation and formulation of remediation actions, we still rely on human experts for now. ■

Acknowledgement

3E's work on automatic fault detection and diagnosis has received funding from the European Union under the Horizon 2020 programme, contract 662189 – MANTIS – ECSEL-2014-1 as well as from the Brussels Capital Region – Innoviris under the MANTIS project.

Author

Achim Woyte manages product innovation at 3E. He is a recognised international expert in photovoltaic system technology and power systems. In both these fields, he has contributed to 3E's international track record for strategic consultancy and innovation. Achim Woyte is an electrical engineer from the University of Hannover (Germany) and a PhD in engineering from the Katholieke Universiteit Leuven (Belgium).

References

[1] IEC 61724 ed1.0, *Photovoltaic System Performance Monitoring-Guidelines for Measurement, Data Exchange and Analysis*. International Electrical Commission, 1998.
[2] G. Blaesser and D. Munro, 'Guidelines for the Assessment of Photovoltaic Plants Document A Photovoltaic System Monitoring', Commission of the European Communities, Joint Research Centre, Ispra, Italy, EUR 16338 EN, Issue 4.2 (June 1993), 1995.
[3] A. Woyte *et al.*, 'Analytical Monitoring of Grid-connected Photovoltaic Systems - Good Practice for Monitoring and Performance Analysis', IEA PVPS, Report IEA-PVPS T13-03: 2014, Mar. 2014.
[4] S. Degener and J. Watson, 'O&M Best Practices Guidelines', SolarPower Europe, Brussels, Belgium, Jun. 2016.
[5] scikit-learn, 'Machine Learning in Python'. [Online]. Available: http://scikit-learn.org. [Accessed: 20-Nov-2016].