Locating Faults in Photovoltaic Systems Data

Alexander Kogler[1] and Patrick Traxler[1]

[1]Data Analysis Systems Group, Software Competence Center Hagenberg, Austria*
{alexander.kogler,patrick.traxler}@scch.at

Abstract. Faults of photovoltaic systems often result in an energy drop and therefore decrease the efficiency of the system. Detecting and analyzing faults is thus an important problem in the analysis of photovoltaic systems data. We consider the problem of estimating the starting time and end time of a fault, i.e. we want to locate the fault in time series data. We assume to know the power output, plane-of-array irradiance and optionally the module temperature. We demonstrate how to use our fault location algorithm to classify shading events. We present results on real data with simulated and real faults.

Keywords: fault location · fault detection · fault diagnosis · photovoltaics · shading · sustainable faults · fault classification · robust regression · pattern recognition

1 Introduction

Faults affect the performance of photovoltaic (PV) systems since most faults result in an energy drop, i.e. the PV system converts less solar energy into electrical energy than possible. Identifying faulty or inefficient PV systems is thus an important problem in data analysis with applications to maintenance of PV systems.

In addition to identifying faulty PV systems, we would like to analyze and explain detected faults. We consider the problem of estimating the starting time A and end time B of a fault and its application to discovering and classifying faults. We call this problem *fault location*. Fault location algorithms are of particular interest for analyzing sustainable faults, i.e. faults that occur frequently over some period of time. An example is shading. Shading usually occurs over the period of some weeks or months at some particular daytime (e.g. in the afternoon). This pattern allows us to classify a fault as a shading pattern. Classifying a fault as a shading pattern is an indication for real shading. Such algorithms are intended to support for example maintenance of PV systems.

We restrict to a basic sensor setting. We assume to know the power output P_t, the plane-of-array (POA) irradiance E_t, and optionally the module temperature

* The research reported in this paper has been supported by the Austrian Ministry for Transport, Innovation and Technology, the Federal Ministry of Science, Research and Economy, and the Province of Upper Austria in the frame of the COMET center SCCH.

T_t at time t. We do not make any assumptions on the PV system, i.e. the power inverter, the number of modules or anything else. In addition, we do not assume to have labeled data.

Results We describe an algorithm for locating faults in PV systems data. We present results for the accuracy and reliability of our algorithm, i.e. how well the algorithm estimates the starting time A and the end time B of some fault. Based on it, we describe a rule-based system for recognizing shading patterns. We present results on real data with real and simulated faults (energy drops). One of our data sets contains verified shading.

A crucial step in our method is robust linear regression. Robustness refers here to a high breakdown point [8]. We compare the effect of robust algorithms on fault location with the effect of non-robust algorithms on fault location. We demonstrate that robust algorithms improve fault location considerably.

Our approach is model- or rule-based. This is due to the lack of labels. Fault classification refers here to recognizing patterns in residuals. We describe these patterns by rules, which requires knowledge about the fault type. Fault detection works similar as in [4] and [9]. In [9], we detect faults by testing if a PV system behaves as a linear system, i.e. if the system observations fit a linear model. A fault is an energy drop in [4] and [9] similar as in this work.

Motivation Our knowledge about operating PV systems is usually very limited. Often we do not know the system design and the power output is the only available sensor information. This is the case for many residential PV systems, but also for solar power facilities. POA-irradiance sensors are also common although clearly not the standard, in particular for residential PV systems. This motivates our restrictive sensor setting. In the open problems section, we discuss the situation that no POA-irradiance sensor is installed.

The motivation for fault classification is that different faults may trigger different actions. Hardware defects such as module failures or sensor defects often require some form of maintenance. Faults such as snow or covering tree leaves are probably not critical and thus do not require any maintenance. Shading is a fault type somewhere in between. A chimney that covers a small fraction of solar modules of a roof-mounted PV system probably reduces the produced energy by a negligible amount of energy. A growing tree however may become or already is a crucial problem for power generation.

1.1 Related Work

Different approaches for detecting faults in PV systems are discussed in several publications like [2], [4], [3], [1] and [9]. However, only a few publications address the problem of fault diagnosis or fault classification. These problems require a more detailed analysis of detected faults.

In order to detect faults (energy drops) of PV systems and to determine their starting time and end time, we use system models similar to those presented in

[5]. The authors compare three different models in order to estimate the power output of PV modules.

These models abstract many details of PV systems and only require knowledge of quantities such as power output or POA-irradiance. In other words, it is not necessary to have knowledge of any system parameters of the PV systems.

One of our contributions is the recognition of sustainable, long-lasting faults. Our focus is on shading. Examples for other types of sustainable faults are panel coverage by snow or module defects.

These kinds of faults are also discussed in [4] but the authors pursue an entirely different method. The main difference to our approach is, that a sustainable event is indicated depending on the sun position at fault time. Furthermore, a considerable amount of historical data is required.

The authors of [2] establish a circuit-based PV module which has been developed using simulation software. Among other things, this model is used for generating test data and validating the proposed fault detection approach. This algorithm also determines different fault classes but they only give information about the severity in terms of energy loss.

In [1] fault detection is considered as a clustering task and is tackled by applying the minimum covariance determinant estimator. Like the method proposed in [9] also this approach does not provide for fault diagnosis (classification of faults).

In order to tackle the problem of fault detection and location estimation, we apply different linear regression methods. Regression methods like the Least Trimmed Squares (LTS) estimator [6] and the Least Median of Squares (LMS) estimator [7] are particularly applicable in fault detection because of their robustness against outliers, i.e. they have a high breakdown point [8]. Photovoltaic systems data is frequently contaminated with outliers (faults). In [9] ℓ_2-regression is used for residual analysis and fault detection due to its robustness property.

2 Method and Results

Our method consists of three steps: (1) robust estimation, (2) fault detection and location, and (3) pattern recognition (for shading patterns).

In the first step, we take data of a day as input and generate linear models for every PV system. In the second step, we take data of this day and the linear model computed in step (1) and compute the residuals. We check if during a day a fault happened. If so, we locate faults. Step one and two generate for every day and PV system information whether a fault happened and if so when. In step three, we take these results for up to several weeks and search for patterns in it. In particular, we check if the rules of a shading pattern are satisfied. In what follows, we describe these steps in detail and present experimental results.

2.1 Robust Estimation

In this section we describe PV system models and procedures for estimating unknown parameters. In addition, we present experimental results for the fitness

of the models. We consider a discrete daytime setting (5, 10 or 15 minutes intervals are realistic) and two different linear system models:

$$P_t = a \cdot E_t + b \cdot E_t^2 \tag{1}$$

and

$$P_t = a \cdot E_t + b \cdot E_t^2 + c \cdot T_t + d \cdot E_t \cdot T_t + e \cdot t + f \cdot t^2, \tag{2}$$

where P_t is the power output of the PV system, E_t is the POA-irradiance, T_t is the module temperature, and $t \in I$ is the discrete daytime. For example, if we have measurements for every hour, then $I = \{1, \ldots, 24\}$. The unknown parameters that we want to estimate are a, b, c, d, e, and f.

We consider the Least Trimmed Squares (LTS) estimator [2], which is a robust variant of the Ordinary Least Squares (OLS) estimator. LTS has a breakdown point of 0.5. This means that a bit less than 50% of data points can be arbitrarily corrupted with outliers (faults). Estimations via OLS and LTS are solutions to optimization problems. The objective function of the LTS estimator is the same as for OLS but considers only the 50% smallest squared residuals. We refer to [2] for the exact mathematical definition of LTS.

Table 1 shows experimental results for the LTS estimator and a single PV system. Power output estimates $\hat{P}_t$ are calculated based on energy data of the day that we want to check for potential faults. Model parameters are estimated based on data of the same day. We simulate energy drops by subtracting 10%, 30% and 50% during 2 hours of midday from the measured power output. We do this for every day. We calculate the R^2 value a.k.a. coefficient of determination and the model fitness as defined in Eq. 3 for each of 240 days. Table 1 shows the median of these 240 values for each setting.

Besides (ordinary) fitness values we calculate an adapted form of fitness. The adapted fitness is calculated on the 50% smallest squared residuals. It is a robust measure of fitness. We note the adapted R^2-value and fitness in Table 1 since it shows that e.g. the adapted R^2-value remains roughly constant in the presence of an energy drop whereas the R^2-value decreases with an increasing energy drop. This shows that LTS is robust against outliers in the data.

Table 1. Model fitness of LTS for a single PV system

	ordinary fitness				adapted fitness			
	model (1)		model (2)		model (1)		model (2)	
	R^2	fitness	R^2	fitness	R^2	fitness	R^2	fitness
drop = 0 %	0.9983	0.9836	0.9989	0.9883	0.9999	0.9947	1	0.9971
drop = 10 %	0.9824	0.9548	0.9849	0.9607	0.9999	0.9928	1	0.9961
drop = 30 %	0.8578	0.8929	0.8648	0.8997	0.9999	0.9929	1	0.9961
drop = 50 %	0.5965	0.8232	0.6096	0.8301	0.9999	0.9928	1	0.9961

We see in Table 1 that the consideration of module temperature T_t and time t has some effect. In what follows, we thus consider model of Eq. 2.

2.2 Fault Detection and Location

In this section, we treat fault location. The input is data for a PV system and day, i.e. (P_t, E_t, T_t) for $t \in I$, which we defined in Sec. 2.1. Note that t refers to daytime. In addition, we have a system model, parameters a, b, c, d, e and f in Eq. 2, that takes as input the POA-irradiance E_t, module temperature T_t, time t and outputs an estimate $\hat{P}_t$ for P_t. First, we calculate the residuals R_t as $R_t := P_t - \hat{P}_t$. Next, we detect faults by checking if the *fitness* [9]

$$F := 1 - \frac{\sum_{t \in I} |R_t|}{\sum_{t \in I} |P_t|} \tag{3}$$

is smaller than a threshold θ_{fit}. If $F < \theta_{\text{fit}}$, we say that a fault happened and continue with fault location. Otherwise, we say that no fault happened.

The output of our fault location algorithm Locate, Alg. 1, are pairs (A_j, B_j), where A_j is the starting time and B_j is the end time of the j-th fault. Alg. 1 outputs locations for all significant faults. Faults are energy drops. We control the significance by the parameter θ_{sig}, i.e. we output a fault location if the energy lost is significant in relation to the total energy. In our application, it suffices to work only with the three strongest energy drops since in most situations there are at most 3 significant energy drops per day.

The algorithm works by computing *deviation* and continuously checking if $R_i < \mathit{deviation}$ holds to find longest subsequences $I' \subseteq I$ of the residuals such that $R_i < \mathit{deviation}$ for every $i \in I'$. We compute *deviation* on a subset R' of the residuals R due to reasons of robustness.

Figure 1 shows results of the algorithm for a roof-mounted PV system with verified shading. The solid line depicts the measured power output and the dashed line the estimated one. Vertical lines indicate the calculated fault location. A chimney causes the shading that is particularly strong over summer. The energy drops are however relatively small. Detecting and locating such faults is difficult because of the small deviations.

Table 2 shows experimental results for fault location. We consider 240 days and a single PV system. We simulate energy drops, as above, by subtracting 10%, 30% and 50% during 2 hours of midday from the measured power output and consider a fault j as correctly located if the estimated starting time A_j and end time B_j differs by at most ± 60 minutes from the real starting and end time.

We conduct these experiments for another 8 PV systems. Table 3 shows the median of these 8 experiments. We see that the values show a similar behavior as the values in Table 2. Alg. 1 shows a consistent behavior over all 9 PV systems. The 9 PV systems are all from the same region in Austria.

Summarizing the results of Table 2 and 3, we clearly see that robust linear regression, LTS here, is superior to non-robust linear regression. Furthermore, Alg. 1 successfully locates most faults, in particular strong faults.

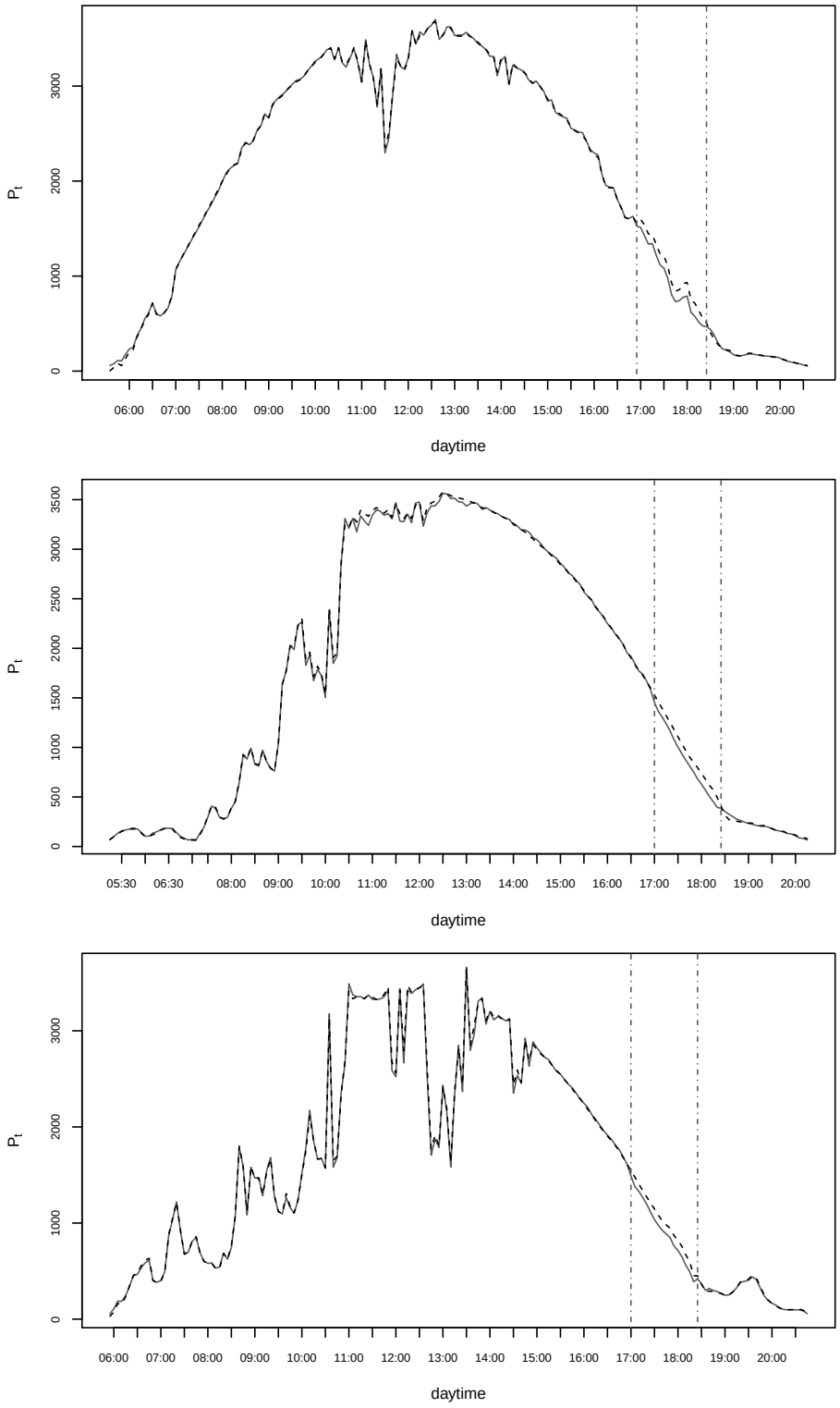

Fig. 1. Three days of a single PV system with verified shading

Algorithm 1 Algorithm Locate with parameter θ_{sig} (significance, typically 0.01). Its input is the generated power of the PV system P_i, $i \in I$, for a day and estimates $\hat{P}_i$ for it. Its outputs are the starting and end times of energy drops.

Calculate the total energy $total := \sum_{i \in I} P_i$
Calculate the residuals $R_i := \hat{P}_i - P_i$ for all $i \in I$
Let R' be the elements from R that are the $\lfloor \frac{|I|}{2} \rfloor$-smallest elements in $\{|R_i| : i \in I\}$
Set $deviation := \text{mean}(R') - \text{standard\_deviaton}(R') \cdot 3.0$
Set $j := 1$
for all $i \in I$ **do**
  while $X_i \geq deviation$ **do**
    Increment i
  end while
  $start := i$
  $energy := R_{start}$
  while $R_i < deviation$ **do**
    $energy = energy + R_i$
    Increment i
  end while
  Set $end := i - 1$
  if $energy > \theta_{\text{sig}} \cdot total$ and $i \geq start + 2$ **then**
    $A_j := start$, $B_j := end$, $E_j := energy$
    Increment j
  end if
end for
Output $(A_1, B_1, E_1), (A_2, B_2, E_2) \ldots$

2.3 Pattern Recognition (Shading Patterns)

In this section, we describe a sample application of Alg. 1. Given that we have detected a fault of a PV system during a day, we want to check if the fault has been possibly caused by shading. The idea is to look for a *shading pattern*. We model the shading pattern by three if-then rules. Let A be the starting time and B the end time of some detected fault.

1. Consider a time period of 7 days before the detected fault. If on at least 4 out of the 7 days some fault occurred with starting time in $A \pm 30$ minutes and end time in $B \pm 30$ minutes, then we say that the fault was possibly caused by shading.
2. Consider a time period of 21 days before the detected fault. If on at least 50% of the days in this period, some fault occurred with starting time in $A \pm 1$ hour and end time in $B \pm 1$ hour, then we say that the fault was possibly caused by shading.
3. Consider a time period of 60 days before the detected fault. If at least 50% of detected faults in this time period have starting time in $A \pm 2$ hours, end time $B \pm 2$ hours and if the number of such events is at least 10, then we say that the fault was possibly caused by shading.

Table 2. Fault location of a single PV system with simulated energy drops and parameters $\theta_{\text{fit}} = 0.99$ and $\theta_{\text{sig}} = 0.004$

	LTS regression		OLS regression	
	# detected f.	correct loc. (%)	# detected f.	correct loc. (%)
drop = 10 %	239	96.23	239	59.41
drop = 30 %	239	97.49	239	75.73
drop = 50 %	239	98.33	239	78.66

Table 3. Fault location of 8 PV systems with simulated energy drops and parameters $\theta_{\text{fit}} = 0.95$ and $\theta_{\text{sig}} = 0.004$

	LTS regression		OLS regression	
	# detected f.	correct loc. (%)	# detected f.	correct loc. (%)
drop = 10 %	197	48.54	159	12.69
drop = 30 %	242	79.34	242	51.03
drop = 50 %	244	90.43	244	59.02

We designed the conditions of the rules in such a way to indicate the confidence of the shading pattern, i.e. if rule (1) applies we are more confident than if rule (2) or (3) applies.

Table 4 shows experimental results concerning the recognition of shading patterns. We applied our method to our 9 PV systems and included the PV systems in the table only if rules fired at least 10 times. PV system (1) contains shading that has been verified on site. The other PV systems with recognized shading patterns have not been verified on site yet, but are likely to be caused by some form of shading. In the last column of Table 4 we noted the number of times at least one of the rules has fired.

Summarizing our results for recognizing shading patterns, we see that our rule-based approach, that is itself based on Alg. 1, successfully identifies PV systems with shading patterns. In particular, the faults of the PV system with verified shading have been correctly classified as shading patterns.

Table 4. Shading events

	# rule 1	# rule 2	# rule 3	≥ 1 rule
PV system (1)	10	23	33	36
PV system (2)	1	3	23	27
PV system (3)	0	10	30	39
PV system (4)	0	4	27	28
PV system (5)	0	1	25	25

3 Summary and Open Problems

We presented a method for analyzing PV systems data to locate and classify faults with a focus on shading. We identified fault location as an important step for fault classification. The starting time and end time of a fault are the basis for fault classification that works by identifying predefined patterns, e.g. a pattern indicating real shading. We verified our method on real data.

We considered a restrictive sensor setting. We assume to know the power output, the POA-irradiance and optionally the module temperature. An even more restrictive and common sensor setting includes only the power output. Studying this sensor setting remains as an open problem. One possible way to solve this problem is to compare the system behavior to nearby systems [10]. Another way is to integrate irradiance information from satellite images or weather stations. The problem with both ways or a combination of them is the low model accuracy that hinders reliable residual analysis.

We considered the case of recognizing shading patterns that indicate real shading events. Other important faults include covering snow and sustainable hardware defects. It remains as an open problem to classify them.

References

1. Braun, H., Buddha, S. T., Krishnan, V., Spanias, A., Tepedelenlioglu, C., Yeider, T., and Takehara, T. Signal processing for fault detection in photovoltaic arrays. In *37th IEEE International Conference on Acoustics, Speech and Signal Processing* (2012), pp. 1681–1684.
2. Chao, K. H., Ho, S. H., and Wang, M. H. Modeling and fault diagnosis of a photovoltaic system. *Electric Power Systems Research 78*, 1 (2008), 97–105.
3. Chouder, A., and Silvestre, S. Fault detection and automatic supervision methodology for PV systems. *Energy conversion and Management 51* (2010), 1929–1937.
4. Firth, S., Lomas, K., and Rees, S. A simple model of PV system performance and its use in fault detection. *Solar Energy 84*, 4 (2010), 624–635.
5. Marion, B. Comparison of predictive models for PV module performance. In *33rd IEEE Photovoltaic Specialist Conference* (2008), pp. 1–6.
6. Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., and Wu, A. Y. On the least trimmed squares estimator. *Algorithmica 69*, 1 (2014), 148–183.
7. Rousseeuw, P. J. Least median of squares regression. *Journal of American Statistical Association 79*, 388 (1984), 871–880.
8. Rousseeuw, P. J., and Leroy, A. M. *Robust regression and outlier detection.* John Wiley & Sons, 2005.
9. Traxler, P. Fault detection of large amounts of photovoltaic systems. In *Proc. of the ECML/PKDD 2013 Workshop on Data Analytics for Renewable Energy Integration* (2013).
10. Traxler, P., Gomez, P., and Grill, T. A robust alternative to correlation networks for identifying faulty systems. In *Proc. of the 26th International Workshop on Principles of Diagnosis* (2015), pp. 11–18.