FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Cellular Time Activation Networks, a novel approach applied to photovoltaic anomaly detection

David da Silva Moreira Freire

WORKING VERSION

Mestrado Integrado em Engenharia Eletrotécnica e de Computadores

Orientador: Cláudio Domingos Martins Monteiro

June 30, 2023

Resumo

A proliferação de centrais fotovoltaicas de dimensão industrial levou à necessidade de métodos para detetar e classificar falhas nos seus componentes, sendo que estas que podem ter impactos económicos significativos. Neste trabalho, explora-se o estado da arte das ferramentas de deteção de falhas e estimação do estado aplicadas ao campo dos sistemas PV, com foco na compreensão do seu funcionamento, identificando-se pontos fortes e possíveis limitações. Relevar-se-á quais os métodos baseados em aprendizagem computacional mais utilizados. Ainda assim, reconhece-se o contributo dos diversos domínios para colmatar este tipo de problema, desde a teoria dos grafos a processamento de sinal, aprendizagem profunda e aprendizagem quântica. Efetuam-se comparações e propostas de melhoria aos algoritmos existentes, e desenvolvida uma nova abordagem para abordar o tema de deteção de falhas. Com a retrospeção das ferramentas contemporâneas de maior sucesso, e pela oferta de uma nova abordagem, o objetivo deste trabalho é fornecer aos operadores de instalações fotovoltaicas o aumento na fiabilidade e eficiência dos seus sistemas. Além disso, há a possibilidade de que a ferramenta desenvolvida seja aplicável para outros problemas de coesão de dados, impactando positivamente os diversos tipos de domínios de sistemas orientados a dados.

Abstract

The increase in utility-scale photovoltaic power plants has led to the need for effective methods for detecting and classifying component faults, which can have significant economic impacts. This work assesses the current state of fault detection and state estimation tools in the field of PV systems, focusing on understanding how these tools work and identifying their strengths and limitations. It is seen that machine learning makes up the majority of state-of-the-art fault detection and classification algorithms. Still, many fields have contributed to this problem, from graph theory to signal processing, deep learning, and quantum machine learning. Consequently, this work compares and proposes improvements to existing approaches or a novel technique developed to address this issue. By examining the most successful tools to date and offering new solutions, the intention is to help PV plant operators improve the reliability and efficiency of their systems. The developed methodology is also expected to become a generalistic data cohesion algorithm, positively impacting other data-driven fields.

Agradecimentos

Aliquam id dui. Nulla facilisi. Nullam ligula nunc, viverra a, iaculis at, faucibus quis, sapien. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Curabitur magna ligula, ornare luctus, aliquam non, aliquet at, tortor. Donec iaculis nulla sed eros. Sed felis. Nam lobortis libero. Pellentesque odio. Suspendisse potenti. Morbi imperdiet rhoncus magna. Morbi vestibulum interdum turpis. Pellentesque varius. Morbi nulla urna, euismod in, molestie ac, placerat in, orci.

Ut convallis. Suspendisse luctus pharetra sem. Sed sit amet mi in diam luctus suscipit. Nulla facilisi. Integer commodo, turpis et semper auctor, nisl ligula vestibulum erat, sed tempor lacus nibh at turpis. Quisque vestibulum pulvinar justo. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Nam sed tellus vel tortor hendrerit pulvinar. Phasellus eleifend, augue at mattis tincidunt, lorem lorem sodales arcu, id volutpat risus est id neque. Phasellus egestas ante. Nam porttitor justo sit amet urna. Suspendisse ligula nunc, mollis ac, elementum non, venenatis ut, mauris. Mauris augue risus, tempus scelerisque, rutrum quis, hendrerit at, nunc. Nulla posuere porta orci. Nulla dui.

Fusce gravida placerat sem. Aenean ipsum diam, pharetra vitae, ornare et, semper sit amet, nibh. Nam id tellus. Etiam ultrices. Praesent gravida. Aliquam nec sapien. Morbi sagittis vulputate dolor. Donec sapien lorem, laoreet egestas, pellentesque euismod, porta at, sapien. Integer vitae lacus id dui convallis blandit. Mauris non sem. Integer in velit eget lorem scelerisque vehicula. Etiam tincidunt turpis ac nunc. Pellentesque a justo. Mauris faucibus quam id eros. Cras pharetra. Fusce rutrum vulputate lorem. Cras pretium magna in nisl. Integer ornare dui non pede.

David Freire

"You should be glad that bridge fell down.
I was planning to build thirteen more to that same design"

Isambard Kingdom Brunel

Contents

1 CellTAN Application **1**
1.1 Case study . . . 1
1.1.1 Data analysis . . . 2
1.1.2 Data Cleaning . . . 6
1.1.3 Photovoltaic Plugin . . . 8
1.1.4 CellTAN Configuration . . . 11
1.1.5 Simulation and Results . . . 11
1.1.6 Inverters-Satellite mismatch . . . 11
1.1.7 Inverter underperformance . . . 11
1.1.8 Scaling up . . . 11

A CellTAN Development **13**
A.1 Statistical tests for measure of association . . . 13
A.1.1 Pearson's chi squared test . . . 13
A.1.2 Fischer's exact test . . . 13
A.1.3 Odds ratio . . . 13
A.1.4 Phi coefficient . . . 14
A.1.5 Contingency coefficient C . . . 14
A.1.6 Theil's U . . . 14
A.2 Technology stack . . . 15
A.3 Cell configuration . . . 15

B CellTAN Application **17**
B.1 MPPT Curve . . . 17
B.2 Data analysis . . . 17

List of Figures

1.1 Inverter AC side power from 2020 to 2022, used for the knowledge base. 2
1.2 Inverter AC side power from 2023-01-01 to 2023-01-05, used for testing. 3
1.3 Pair plot of AC power from both inverters (2020 to 2022), using scatter (left) and KDE (Kernel Density Estimation) (right). 3
1.4 Pair plot of AC power from both inverters (2023), using scatter (left) and KDE (Kernel Density Estimation) (right). 4
1.5 Pair plot of DC side voltage and current from inverter one (2020-2022), using scatter (left) and KDE (Kernel Density Estimation) (right). 4
1.6 Pair plot of DC side voltage and current from inverter two (2020-2022), using scatter (left) and KDE (Kernel Density Estimation) (right). 5
1.7 Scatter pair plot of the AC power, tilted and horizontal global irradiance for both inverters (2020 to 2022). 6
1.8 Inliers (orange), outliers (blue), and decision boundaries (black) using three different anomaly detection algorithms on the AC Power from both inverters. 7
1.9 Inliers (orange), outliers (blue), and decision boundaries (black) using three different anomaly detection algorithms on the tilted irradiance and AC Power from both inverters. 8
1.10 Inliers (orange), outliers (blue), and decision boundaries (black) using three different anomaly detection algorithms on the horizontal irradiance and AC Power from both inverters. 9
1.11 Inliers (orange), outliers (blue), and decision boundaries (black) using three different anomaly detection algorithms and their combination on the DC side voltage and current from both inverters. 10
1.12 Inverter underperformance region based on the lowest boundary of current-voltage inliers. 11

A.1 Technology stack of the Cell and Hub of CellTAN. 15

B.1 "PV panel power characteristics as a function of the DC voltage and solar irradiance." 17
B.2 Pair plot of DC side voltage and current from inverter one (2023), using scatter (left) and KDE (Kernel Density Estimation) (right). 18
B.3 Pair plot of DC side voltage and current from inverter two (2023), using scatter (left) and KDE (Kernel Density Estimation) (right). 18
B.4 Scatter pair-plot of AC power from the two inverters with cloud coverage and temperature (from satellite). 19

List of Tables

1.1 Available variables from two inverters and a satellite. 1

Abreviaturas e Símbolos

AC	Alternating Current
ANN	Artificial Neural Network
CNN	Convolutional Neural Network
CXN	Cell Complex Neural Network
DC	Direct current
DL	Deep Learning
DNN	Deep Neural Network
LSTM	Long short-term memory
MCD	Minimum Covariance Determinant
ML	Machine Learning
PV	Photovoltaic
RBFNN	Radial basis function neural network
RMM	Recurrent Neural Network
SC	Short Circuit
SRC	Sparse Representation Classifier
STC	Standard Test Conditions
SVM	Support Vector Machine

Chapter 1

CellTAN Application

1.1 Case study

The experiments validating CellTAN's behavior incorporate two neighboring grid-tied string inverters from the same PV farm with common satellite data. Their only known characteristics are:

- Inverter one: 12.5kW nominal power; 14.4kW peak power; fixed tilt and azimuth; installed January 1st, 2013.
- Inverter two: 15kW nominal power; 15.84kW peak power; fixed tilt and azimuth; installed January 1st, 2013.

Variable	Source	Unit	Label
AC side power	Inverter (1 & 2)	W	ac_power
AC side current	Inverter (1 & 2)	A	ac_current
AC side voltage	Inverter (1 & 2)	V	ac_voltage
DC side power	Inverter (1 & 2)	W	dc_power
DC side current	Inverter (1 & 2)	A	dc_current
DC side voltage	Inverter (1 & 2)	V	dc_voltage
Global tilted irradiance	Satellite	W/m^2	global_tilted_irradiance
Global horizontal irradiance	Satellite	W/m^2	global_horizontal_irradiance
Cloud coverage	Satellite	%	cloud_coverage
Air temperature	Satellite	ºC	temperature

Table 1.1: Available variables from two inverters and a satellite.

Table 1.1 represents the available variables and corresponding labels used to identify them in the cell's inputs and graphs. These variables are sampled every 10 minutes from May 31, 2020, at 5:00 am to April 30, 2023, at 7:30 pm (with gaps). We utilized data from 2020 until the end of 2022 for the cells' knowledge base, and any information from 2023 onwards is considered new and used for testing. Since there is no production at night, the data's original database does not

store values for this period. Not accounting for the night as missing samples, we have around 98% of data availability.

Analyzing and cleaning raw inverter and satellite data is essential to take full benefit of CellTAN's capabilities. As seen in its formulation stage, having a clean knowledge base contributes to correctly identifying anomalous situations. Therefore, the following sections focus on these two steps, contributing to understanding the anomalies' domain and frequency of occurrence.

1.1.1 Data analysis

Before data visualization, and regarding the variables in table 1.1, we eliminate those that will not benefit the CellTAN. We determined that AC side voltage is insignificant since the grid mandates it in a grid-tied inverter. We have decided to only use the measure of power instead of using the AC side current measure since, in conjunction with voltage, it provides the same information. To simplify things further, we do not need to consider the power on the DC if considering both the DC side current and voltage measures.

We examine all variables related to satellite data to determine which ones could be useful, not making any premature assumptions.

1.1.1.1 Power

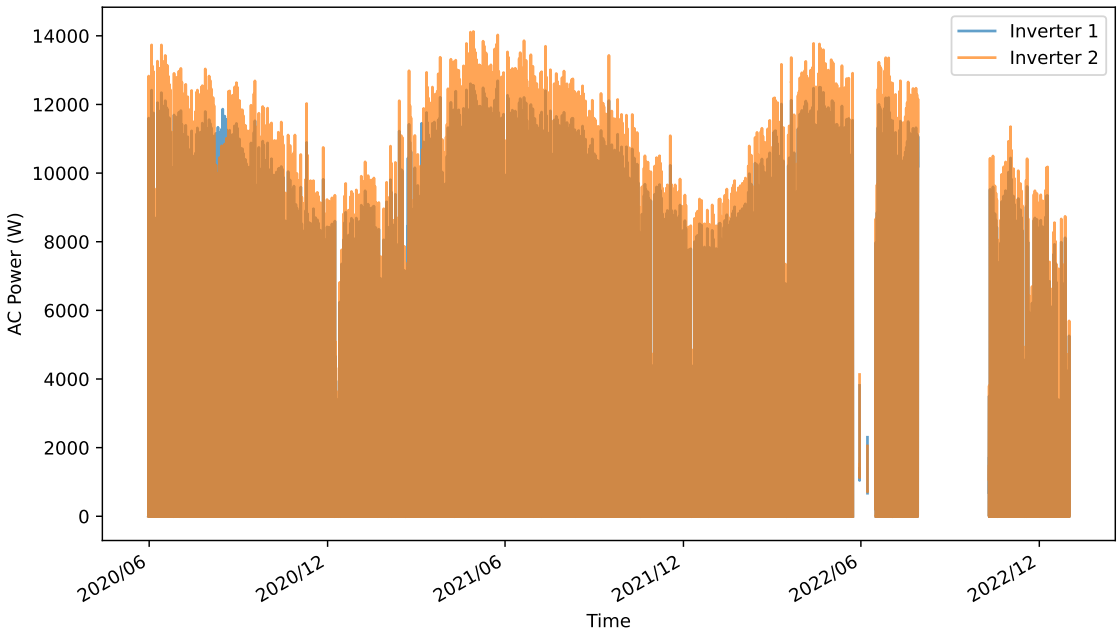

Figure 1.1: Inverter AC side power from 2020 to 2022, used for the knowledge base.

Figure 1.1 shows the power profile of the two studied inverters. Right away, we notice that the power of inverter two caps at around 14kW, while inverter one usually maxes at 12kW. This information is coherent with their ratings. We can notice two relatively large chunks of missing data, with the gaps occurring in mid to late 2022.

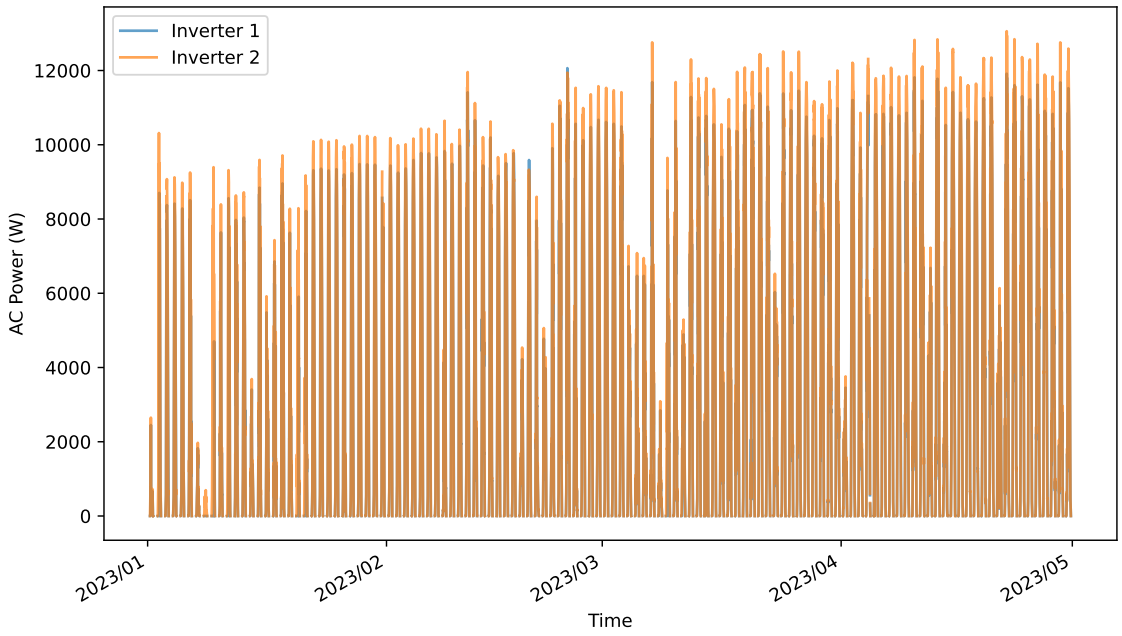

Figure 1.2: Inverter AC side power from 2023-01-01 to 2023-01-05, used for testing.

Figure 1.2 represents the power profile on the portion of data used for testing. When performing a closer inspection (with more zoom), we could hand-pick some fault occurrences in both datasets, with the majority being one inverter off while the other continues regular operation. However, these will be more noticeable during different types of data analysis, such as pair plotting. Regardless, the cases that will matter are in the test data since these scenarios will not exist in the knowledge after cleaning. In Section 1.1.5, you will find a selection of carefully chosen scenarios.

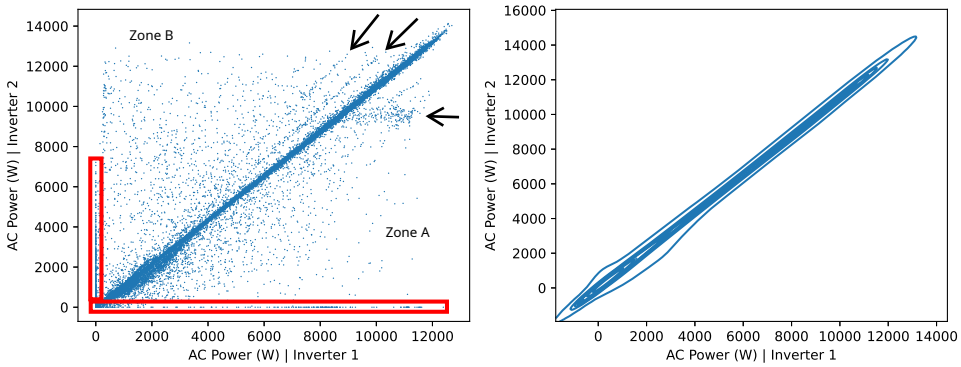

Figure 1.3: Pair plot of AC power from both inverters (2020 to 2022), using scatter (left) and KDE (Kernel Density Estimation) (right).

Figure 1.3 lets us better understand the relationship between the two inverters. As expected, since they are neighboring, they have a strong trend line, leading them to a high Pearson coefficient: 0.97753. However, the noise from outliers is noticeable in the scatter. We define Zone A as the zone where inverter two underperforms compared to one and Zone B as the opposite. The

black arrows in the graph display secondary trend lines in Zone B, indicating scenarios of inverter one performing consistently less than expected. Another arrow also points out a cloud in Zone A (right under the trend line) of the opposite scenario. Furthermore, the red rectangles highlight instances where one inverter was functioning while the other was not. CellTAN must flag these situations, so we should remove them from the knowledge base. The KDE visualization confirms that most samples lie close to the primary trend.

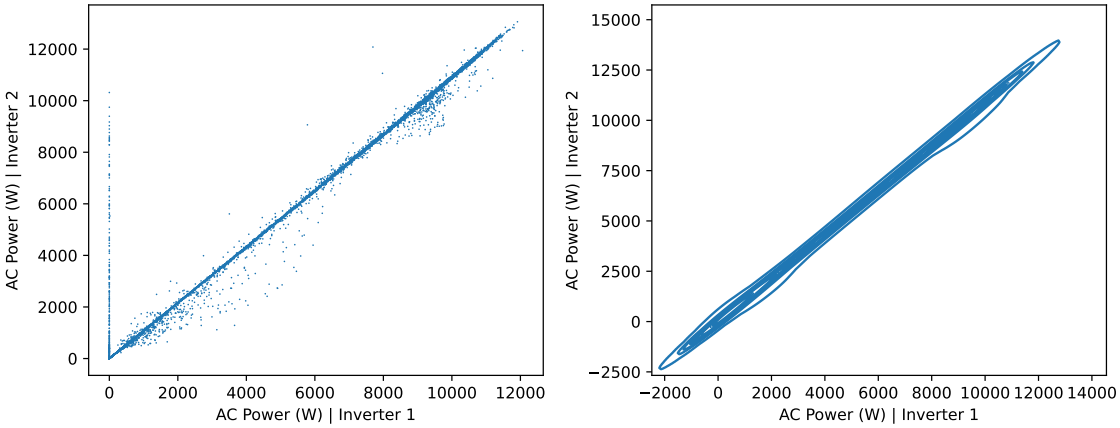

Figure 1.4: Pair plot of AC power from both inverters (2023), using scatter (left) and KDE (Kernel Density Estimation) (right).

From 1.4, it is clear that test data has fewer outliers than the previous. Nonetheless, there are many occurrences of inverter one being inoperational. Besides, there are also a considerable amount of samples below the trend line, meaning the underperformance of inverter two.

1.1.1.2 Voltage and Current

Both inverters are equipped with MPPTs to maximize the power output from their strings. As a result of this power converter, the current and voltage readings should fall within the optimal range of the I-V curve.

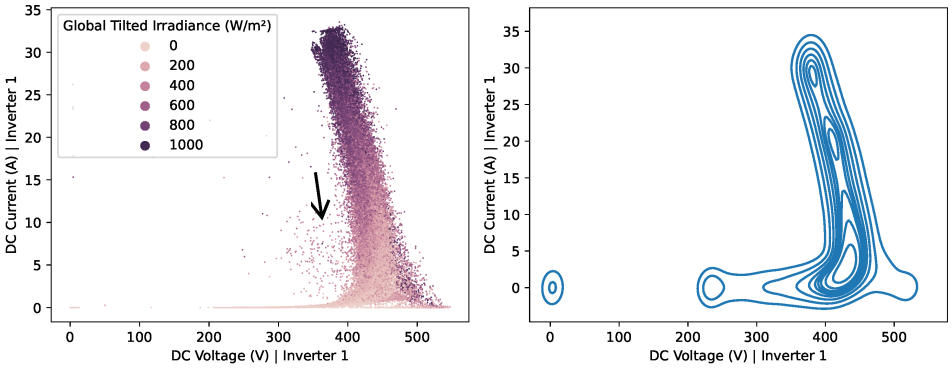

Figure 1.5: Pair plot of DC side voltage and current from inverter one (2020-2022), using scatter (left) and KDE (Kernel Density Estimation) (right).

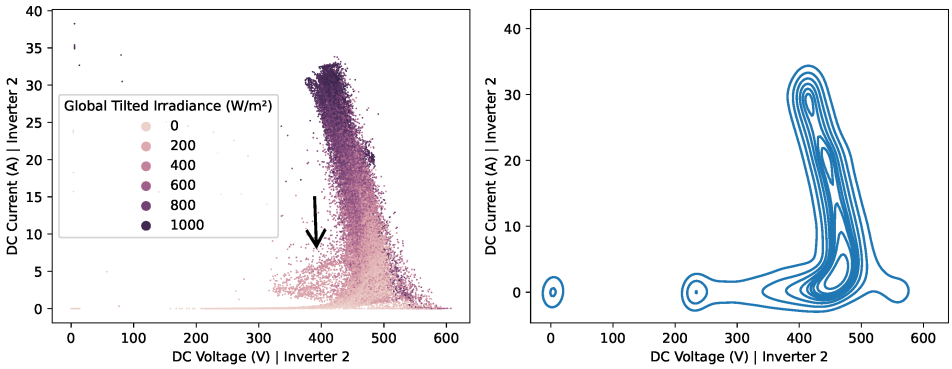

Figure 1.6: Pair plot of DC side voltage and current from inverter two (2020-2022), using scatter (left) and KDE (Kernel Density Estimation) (right).

Regarding DC side voltage and current, figures 1.5 and 1.6 demonstrate the operating range of the inverter's MPPT. Both kickstart production at around 400 V and operate until close to 600 V. Between this range, the central column of samples represents voltage-current points relative to the knee of the strings' I-V curve (see figure B.1), with irradiance generally increasing along its height. Some instances are outside this normal operation range (outliers), especially in the zone marked by the black arrow. This zone has particular interest given that it represents scenarios of underperformance, which could mean the occurrence of faults. It is denser in figure 1.6, meaning that inverter two has more underperforming situations, which was not completely clear from the previous analysis (figure 1.3). Appendix B.2 presents the same charts, but for test data (figures B.2 and B.3).

1.1.1.3 Satellite

Irradiance should be the satellite variable most related to inverters' power. Therefore, we scatter both tilted and horizontal irradiance against AC power from inverters one and two.

Figure 1.7 shows the relationship between irradiance and AC power. We expected a positive correlation, and it exhibits such. However, the large radius around the central trend line demonstrates cycles around it that resemble some kind of hysteresis (especially with horizontal irradiance). By adding a color that displays the hour in each sample, we can affirm that these paths occur due to the fixed nature of the installed PV panels, having a characteristic curve from low to high irradiance in the sunrise and another from high to low during the sunset. Because they produce more with less sunlight in the morning, we can infer that they are oriented slightly towards the east.

Although most instances appear inside the sunrise-sunset paths, there are some outliers. The most notable are the ones of non-zero irradiance with zero production. Either error in satellite data or some anomaly in the inverters causes these odd scenarios, so we target them for cleaning the knowledge base.

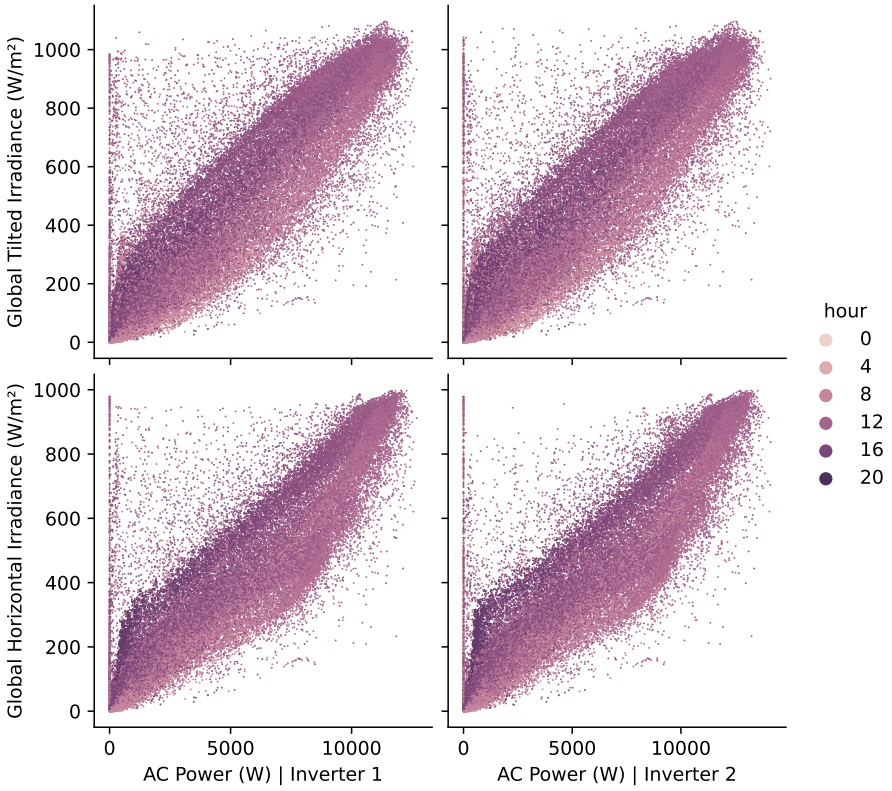

Figure 1.7: Scatter pair plot of the AC power, tilted and horizontal global irradiance for both inverters (2020 to 2022).

Regarding the rest of the meteorological variables (cloud coverage and temperature), we deemed them unnecessary since they do not demonstrate a direct relationship with inverter behavior (see figure B.4). We added KDE visualizations and plots for the test period in appendix B.2.

1.1.2 Data Cleaning

Now, we take our previous analysis and start the data-cleaning process. We used the Scikit-Learn Python Library and tried out some anomaly detection algorithms for outlier identification on our dataset: Robust Covariance [1], Isolation Forest [2] [3], and Local Outlier Factor [4]. The code used for our plotting derived from an example made by Alexandre Gramfort, available on the Scikit-Learn website [5].

1.1.2.1 Power

Based on the geographical proximity of the inverters, we have determined that the area around the trendline shown in figure 1.1 indicates standard operating points. Therefore, we must remove any samples deviating from this reference to clean the knowledge base.

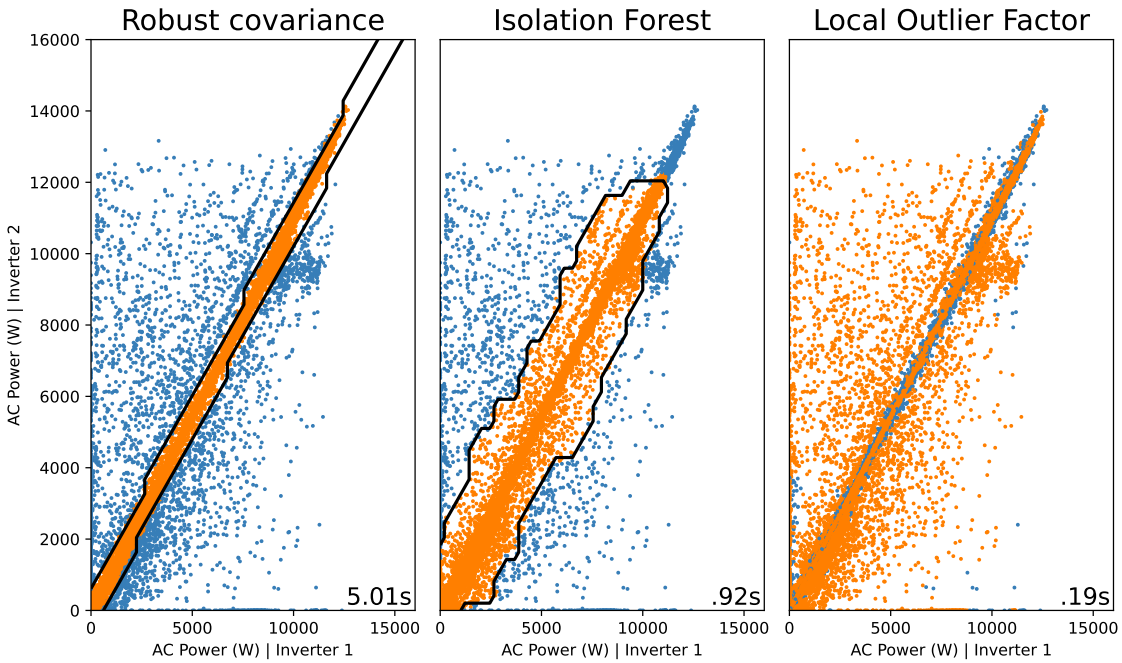

Figure 1.8: Inliers (orange), outliers (blue), and decision boundaries (black) using three different anomaly detection algorithms on the AC Power from both inverters.

In Figure 1.8, we can see a visual representation of various anomaly detection algorithms' inliers, outliers, decision boundaries, and their computation time (bottom right). These algorithms are fine-tuned for approximately 14% of outlier contamination. Based on the results, we determined that the Robust Covariance algorithm was the most effective since it extracted the region around the trend line without issues. We selected it to clean our knowledge base.

1.1.2.2 Satellite

In section 1.1.1.3, we discovered that the inverters follow a standard production pattern based on the irradiance and hour of the day. The specific area within the paths is considered the central region of standard operation. To clean up the data, we will remove all other instances.

Once again, figures 1.9 and 1.10 demonstrate the effectiveness of the first algorithm, now for irradiance and power data. It is, once again, the chosen one for cleaning the knowledge base.

1.1.2.3 Voltage and Current

Cleaning voltage and current data were the most challenging thus far. The peculiar data distribution was an issue for all of the tested algorithms, and we combined them all to obtain the best inlier-outlier separation.

The charts in figure 1.11 illustrates how the algorithms have difficulty keeping up with the data's distribution and preserving the densely populated areas. Neither algorithm effectively identifies the outliers separately, but their combination proved much more effective.

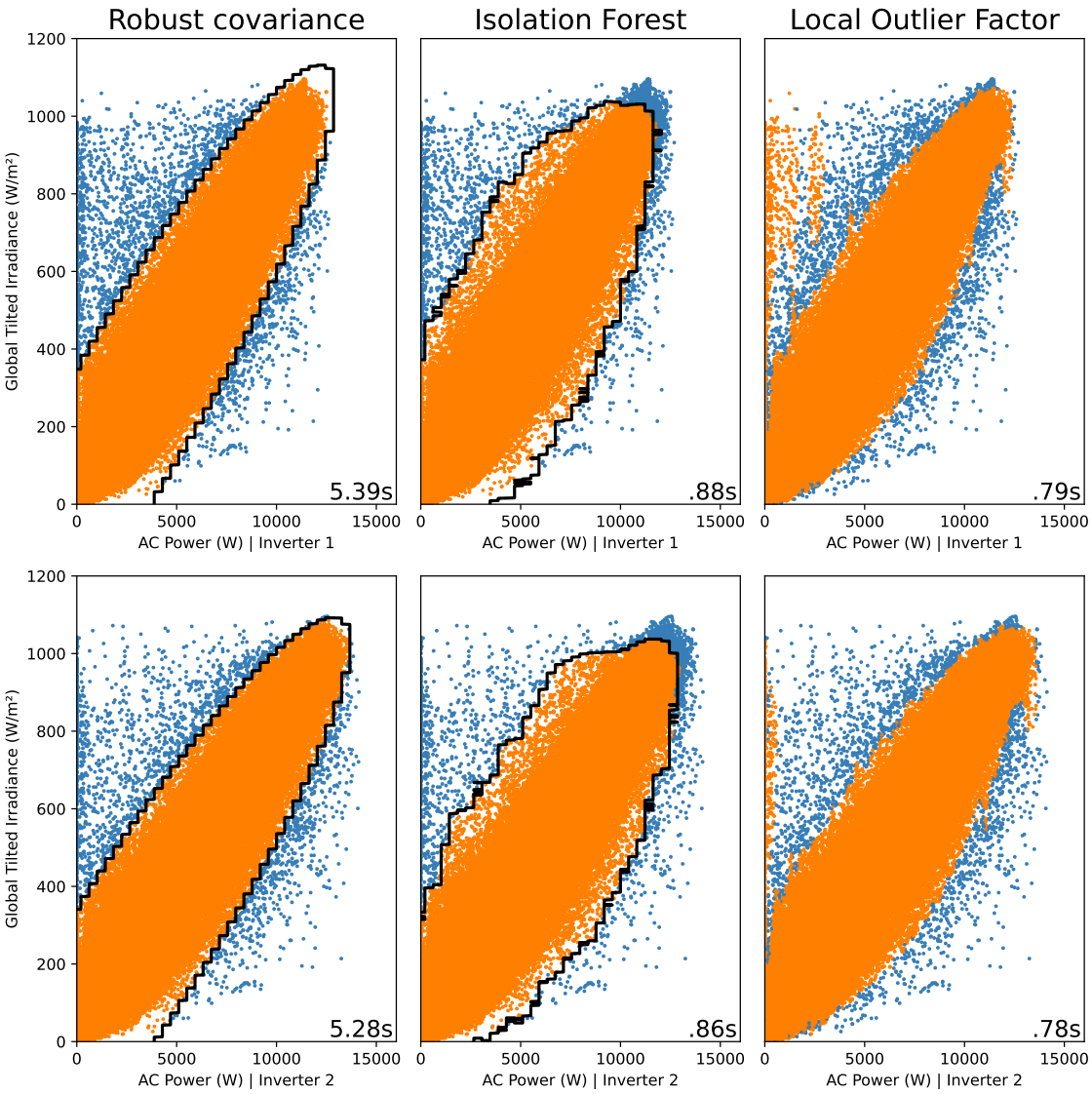

Figure 1.9: Inliers (orange), outliers (blue), and decision boundaries (black) using three different anomaly detection algorithms on the tilted irradiance and AC Power from both inverters.

1.1.3 Photovoltaic Plugin

According to the CellTAN formulation, cells cannot share their variables' data directly to maintain privacy. As a result, this limits plugins to work with only the information available in the cell they are running in. Therefore, inverter underperformance is the only situation considered for the PV plugin, based on the analysis from section 1.1.1.2. Access to both the inverter's data and satellite data allows the owner to employ other algorithms that leverage this knowledge for better situational and fault discernment, but this falls back to the "classical" centralized algorithms, which is not the scope of this work.

To define the minimum operational boundary, we use the current and voltage samples resulting from the cleaning process to define inverter underperformance. Figure ref shows the proposed

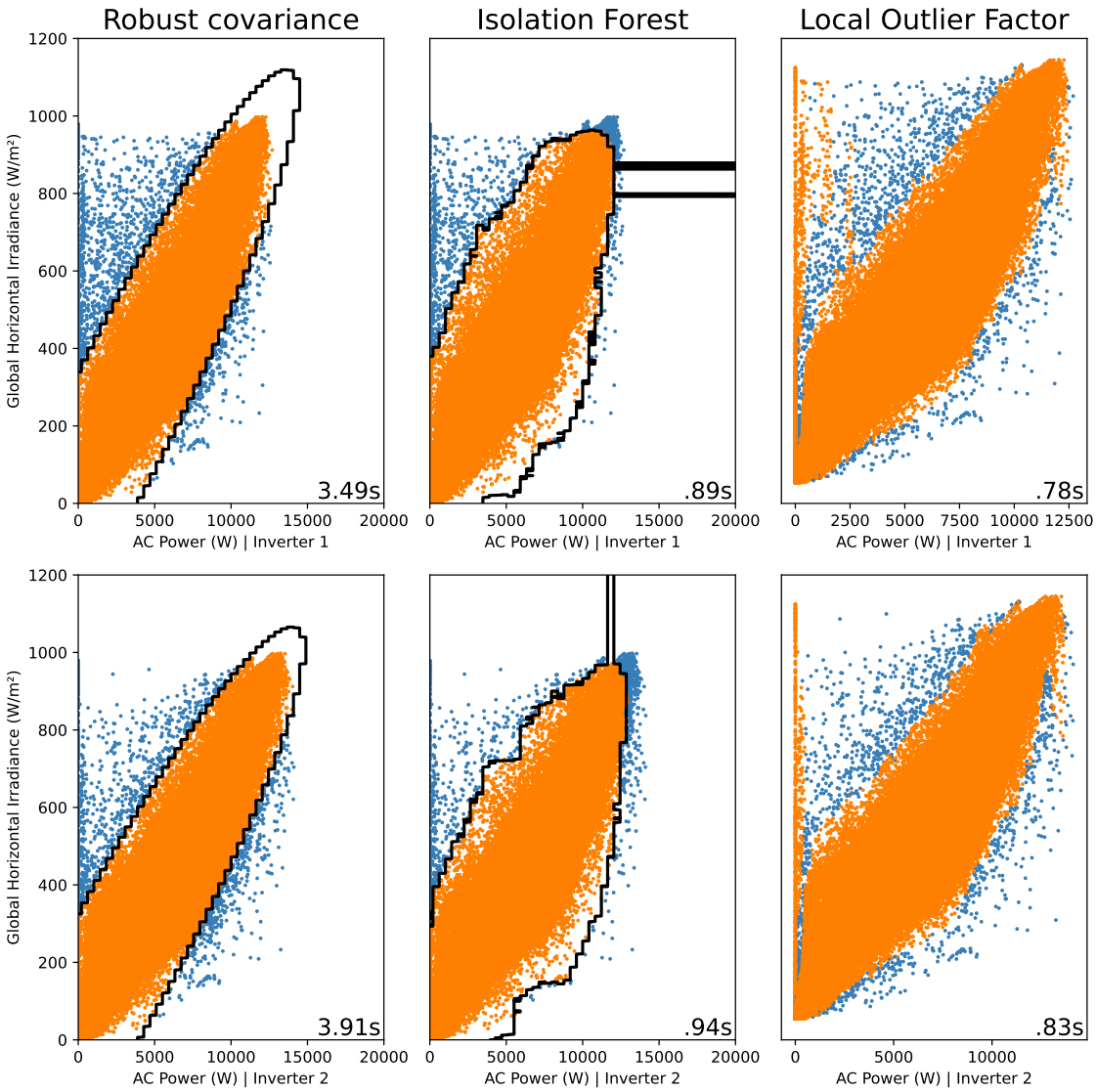

Figure 1.10: Inliers (orange), outliers (blue), and decision boundaries (black) using three different anomaly detection algorithms on the horizontal irradiance and AC Power from both inverters.

algorithm, which consists of the following steps:

- Clean I-V data;
- Find the boundary points of lowest voltage;
- Fit a logarithmic function (1.1) to these points;
- Consider all points below the curve as situations of inverter underperformance.

Since data cleaning is already covered, we describe ways of determining the boundary points. To begin, we establish bins for the current samples with a size of 1 A. Next, we examine each bin and determine the minimum voltage for that particular region via either the absolute minimum or

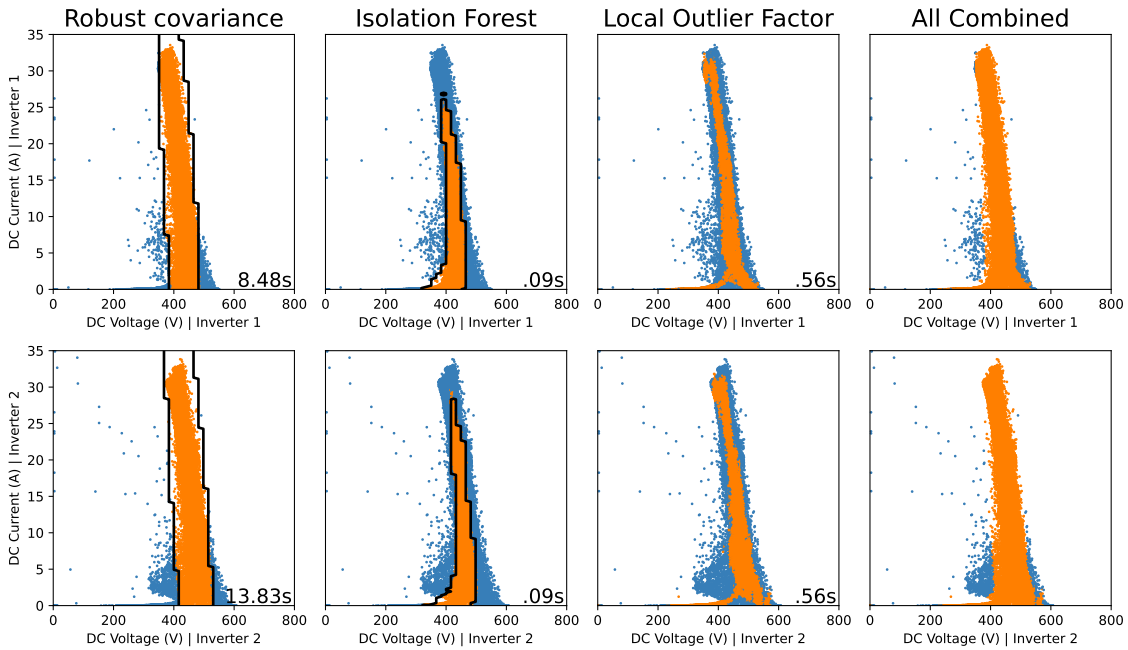

Figure 1.11: Inliers (orange), outliers (blue), and decision boundaries (black) using three different anomaly detection algorithms and their combination on the DC side voltage and current from both inverters.

a quantile. Once we have selected the reference minimum voltage, we record the bin's median current and minimum voltage as a boundary point. After registering all boundary points, we use a curve-fitting algorithm to apply a decaying logarithmic function to them (equation 1.1).

$$V = a \times log(b \times I + c) + d \times I + e \tag{1.1}$$

where:

$$V : \text{Voltage (V)}$$
$$I : \text{Current (A)}$$
$$a, b, c, d, e : \text{Unknown curve parameters}$$

Figure 1.12 illustrates the outcome of using the described algorithm on our case study inverters. With the fitted curve's parameters, the PV plugin can analyze a new sample and determine if it falls under the underperformance category—whenever it does, the plugin flags that situation and reports back to the cell. The fitted logarithmic curves are:

$$V_{inverter1} = 19.2 \times log(170.7 \times I_{inverter1} - 37.0) - 2.5 \times I_{inverter1} + 269.5$$

$$V_{inverter2} = 20.8 \times log(169.8 \times I_{inverter2} - 41.8) - 2.6 \times I_{inverter2} + 288.2$$

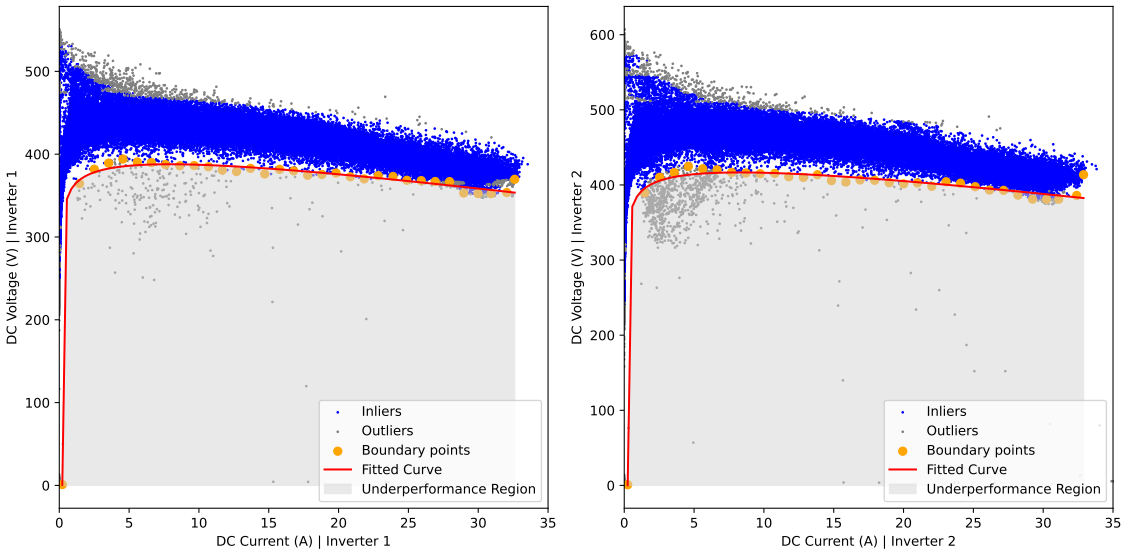

Figure 1.12: Inverter underperformance region based on the lowest boundary of current-voltage inliers.

1.1.4 CellTAN Configuration

...

1.1.5 Simulation and Results

1.1.5.1 Inverter-Inverter mismatch

...

1.1.6 Inverters-Satellite mismatch

...

1.1.7 Inverter underperformance

...

1.1.8 Scaling up

...

Appendix A

CellTAN Development

A.1 Statistical tests for measure of association

A.1.1 Pearson's chi squared test

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{A.1}$$

where:

χ^2 : Chi-squared statistic

O_i : Observed frequency for category i

E_i : Expected frequency for category i

k : Number of categories or cells in the data

A.1.2 Fischer's exact test

$$p = \frac{\binom{a}{x}\binom{b}{y}}{\binom{N}{n}} \tag{A.2}$$

where:

p : p-value of the test

a : Number of successes in group A

b : Number of successes in group B

x : Number of successes of interest in group A

y : Number of successes of interest in group B

N : Total number of observations

n : Number of observations in group A

A.1.3 Odds ratio

$$OR = \frac{a \cdot d}{b \cdot c} \tag{A.3}$$

where:

OR : Odds ratio
a : Number of successes in group A
b : Number of failures in group A
c : Number of successes in group B
d : Number of failures in group B

A.1.4 Phi coefficient

$$\phi = \sqrt{\frac{\chi^2}{N}} \tag{A.4}$$

where:

ϕ : Phi coefficient
χ^2 : Chi-squared statistic
N : Total number of observations

A.1.5 Contingency coefficient C

$$C = \sqrt{\frac{\chi^2}{N+\chi^2}} \tag{A.5}$$

where:

C : Contigency coefficient
χ^2 : Chi-squared statistic
N : Total number of observations

A.1.6 Theil's U

$$U(x|y) = \frac{H(x) - H(x|y)}{H(x)} \tag{A.6}$$

Entropy of variable x:

$$H(x) = -\sum_{i=1}^{n} p(x_i) \log(p(x_i)) \tag{A.7}$$

Conditional entropy of variable x given variable y:

$$H(x|y) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log\left(\frac{p(x_i, y_j)}{p(y_j)}\right) \tag{A.8}$$

A.2 Technology stack

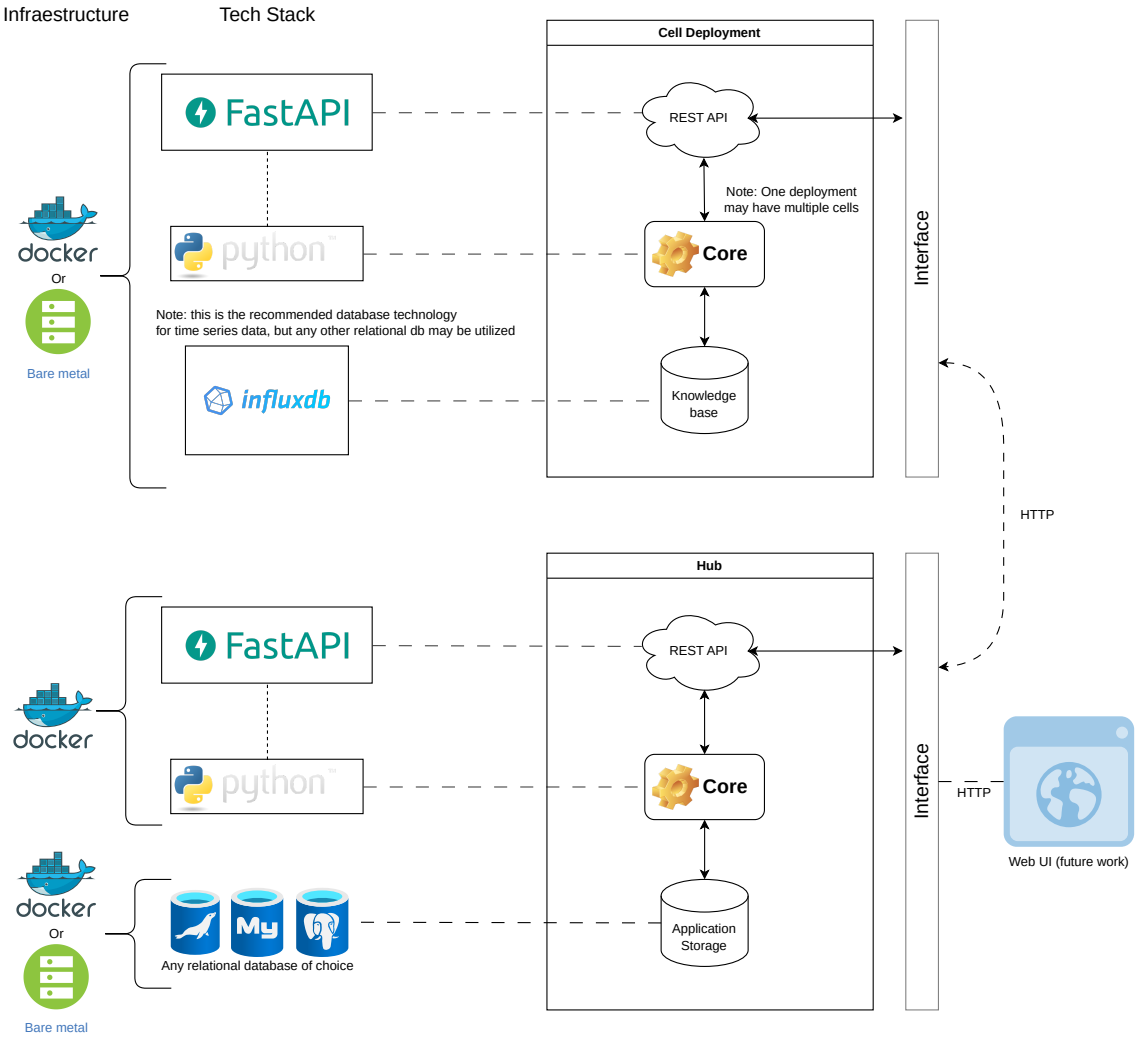

Figure A.1: Technology stack of the Cell and Hub of CellTAN.

A.3 Cell configuration

Appendix B

CellTAN Application

B.1 MPPT Curve

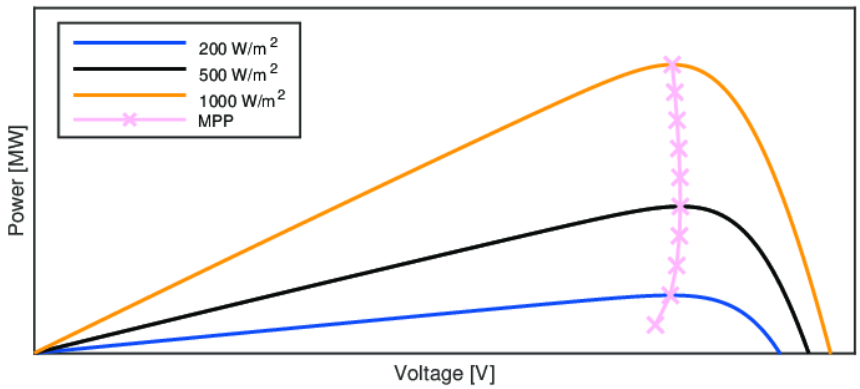

Image source and copyright: [6].

Figure B.1: "PV panel power characteristics as a function of the DC voltage and solar irradiance."

B.2 Data analysis

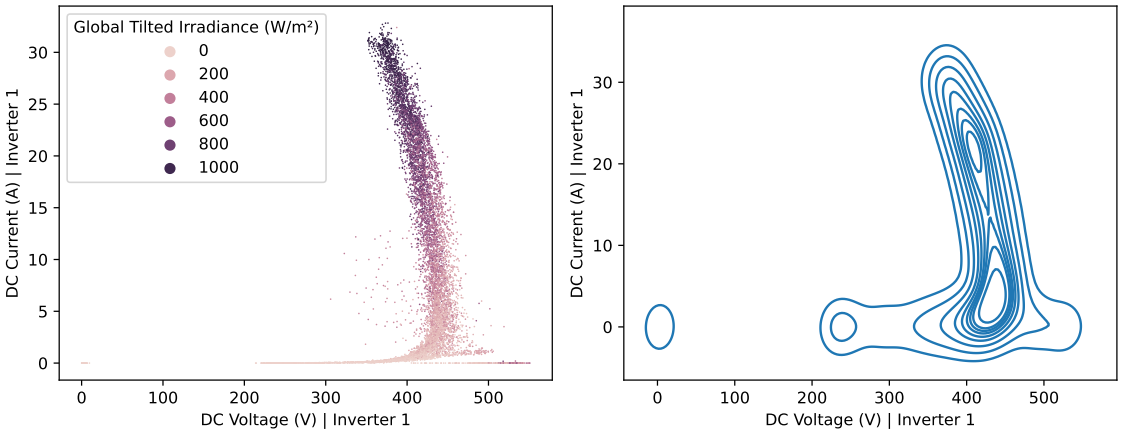

Figure B.2: Pair plot of DC side voltage and current from inverter one (2023), using scatter (left) and KDE (Kernel Density Estimation) (right).

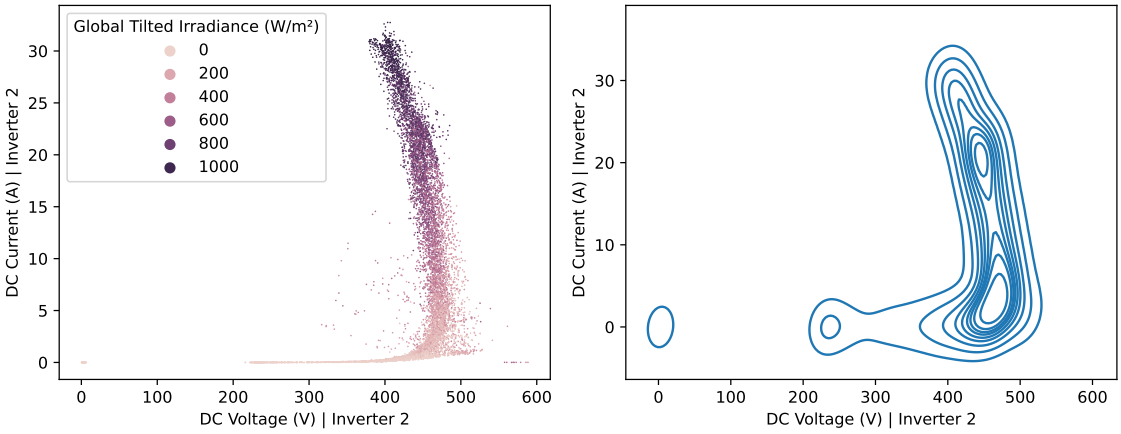

Figure B.3: Pair plot of DC side voltage and current from inverter two (2023), using scatter (left) and KDE (Kernel Density Estimation) (right).

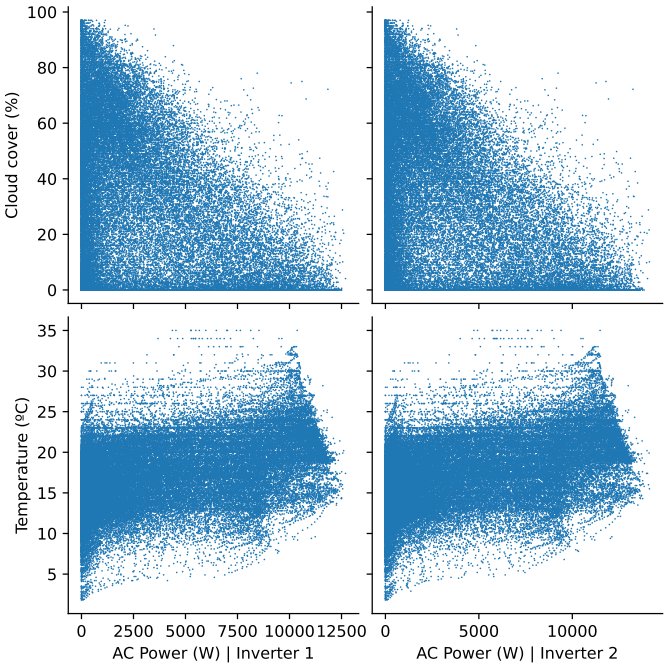

Figure B.4: Scatter pair-plot of AC power from the two inverters with cloud coverage and temperature (from satellite).

References

[1] P. J. Rousseeuw and K. V. Driessen, "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, vol. 41, pp. 212–223, 1999.

[2] F. T. Liu, K. M. Ting, and Z. H. Zhou, "Isolation forest," pp. 413–422, 2008.

[3] F. T. Liu, K. M. Ting, and Z. H. Zhou, "Isolation-based anomaly detection," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 6, p. 1, 3 2012.

[4] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander, "Lof: Identifying density-based local outliers," *SIGMOD 2000 - Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pp. 93–104, 2000.

[5] "Comparing anomaly detection algorithms for outlier detection on toy datasets — scikit-learn 1.2.2 documentation." `https://scikit-learn.org/stable/auto_examples/miscellaneous/plot_anomaly_comparison.html#sphx-glr-auto-examples-miscellaneous-plot-anomaly-comparison-py`. Accessed: 2023-06-06.

[6] A. Lunardi, L. F. Normandia Lourenço, E. Munkhchuluun, L. Meegahapola, and A. Sguarezi, "Grid-connected power converters: An overview of control strategies for renewable energy," *Energies*, vol. 15, p. 4151, 06 2022.