Automated Shift Detection in Sensor-Based PV Power and Irradiance Time Series: Preprint

Preprint

Kirsten Perry and Matthew Muller

National Renewable Energy Laboratory

Presented at the 49th IEEE Photovoltaic Specialists Conference (PVSC 49)
Philadelphia, Pennsylvania
June 5-10, 2022

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

Conference Paper
NREL/CP-5K00-83062
November 2022

Automated Shift Detection in Sensor-Based PV Power and Irradiance Time Series: Preprint

Preprint

Kirsten Perry and Matthew Muller

National Renewable Energy Laboratory

Suggested Citation
Perry, Kirsten and Matthew Muller. 2022. *Automated Shift Detection in Sensor-Based PV Power and Irradiance Time Series: Preprint*. Golden, CO: National Renewable Energy Laboratory. NREL/CP-5K00-83062. https://www.nrel.gov/docs/fy23osti/83062.pdf.

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

Conference Paper
NREL/CP-5K00-83062
November 2022

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

NOTICE

This work was authored by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Solar Energy Technologies Office. The views expressed herein do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

Cover Photos by Dennis Schroeder: (clockwise, left to right) NREL 51934, NREL 45897, NREL 42160, NREL 45891, NREL 48097, NREL 46526.

Automated Shift Detection in Sensor-Based PV Power and Irradiance Time Series

Kirsten Perry[1],
Matthew Muller[1]
[1]NREL, Golden, CO, 80401, USA

***Abstract*—PV power and irradiance sensor-based measurements are prone to error, resulting in issues such as abrupt time series data shifts. These shifts, which are usually unintentional, may be caused by software or hardware configuration changes on a PV system, and do not reflect an actual change in overall system performance. Locating these shifts and segmenting the associated time series aids in more accurate future PV analysis. In this research, an offline changepoint detection (CPD) algorithm that automatically detects these abrupt data shifts in sensor-based time series is introduced. Data shift periods in 101 daily PV power and irradiance time series were labeled manually by two solar experts. These data streams represent sensor-based measurements, and display a variety of data shift behaviors. A changepoint detection algorithm was tuned using the 101 labeled data streams, with each model configuration's ability to detect labeled changepoints benchmarked using metrics such as F1-score, recall, and Rand Index. Best performing models on seasonality-corrected data streams include the Pruned Exact Linear (PELT) method, the Binary Segmentation method, and the Bottom-Up method, all scoring an average F1-score of 0.76 or greater at detecting labeled changepoints within a 30-day window for the labeled data sets. To promote further research in this space, we are releasing the labeled data shift sets on U.S. Department of Energy's (DOE) DuraMAT Data Hub, and the associated algorithm in the Python PVAnalytics package.**

***Index Terms*—data shift, changepoint detection, solar, irradiance, power, data quality**

I. Introduction

The use of high-quality PV data is paramount for effective monitoring of photovoltaic (PV) systems, including running advanced analytics routines to estimate system performance and degradation rate [1]. When poor underlying data is fed into PV models, results may be inaccurate and lead to uninformed business decisions that further impact system health and longevity. Consequently, it is paramount to use high-quality and valid data when performing these analytics routines.

Invalid solar time series data, which includes missing data periods and outliers, may occur as result of power outages, equipment failures, and communication issues [1]. One particular issue in power and irradiance data streams is abrupt data shifts. Some example data shifts, taken from sensor-based time series data, are shown in Figure 1. Data shifts such as these are frequently introduced unintentionally, as a result of replacing hardware or by software configuration changes [2]. It is important to note that these shifts do not reflect an actual change in system performance, but are generally a result of data acquisition issues; for example, by changing the scale factor for a particular data output, or by converting an AC energy data stream to an AC power data stream and not adequately documenting the change.

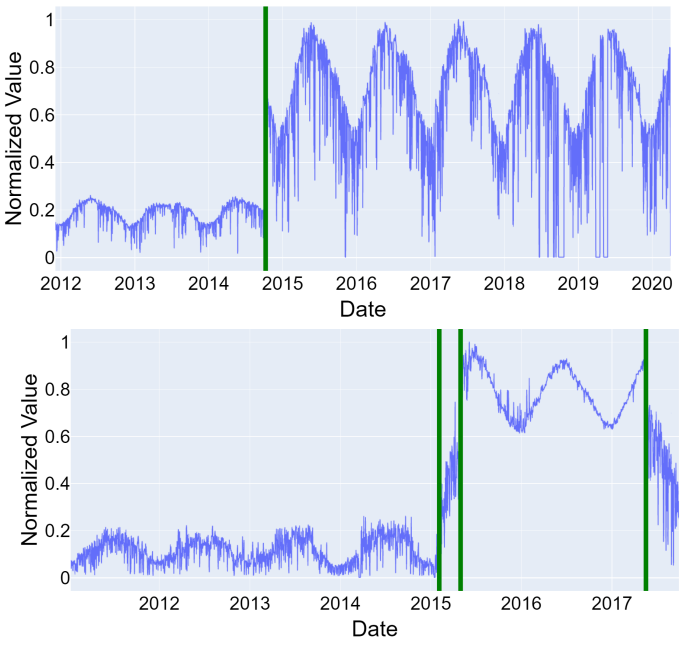

Fig. 1. Two example daily time series with data shifts, taken from data streams in the NREL PV Fleets project. Daily data points are represented in purple, and vertical green lines represent manually labeled changepoints.

Some limited research has been performed exploring the consequences of performing PV analysis using time series with data shifts. In particular, Jordan et al. [2], [3] have examined the effects of abrupt data shifts on degradation estimates in PV time series data. [3] corrects artificially introduced data shifts in a sensor-based temperature time series via a scaling factor, to obtain accurate degradation estimates for a system. Similarly, [2] uses multiple techniques such as standard least squares regression (SLS) and a year-on-year (YoY) approach to correct data shifts to accurately estimate degradation rates.

Lindig et al. [4] also addresses data shifts in the context of data quality filtering for accurate solar performance loss and useful-lifetime calculations. This research specifically recommends filtering out data shift periods in time series where the cause of the data shift is unknown. However, [4] does not provide any process for detection of data shifts in PV time

series data, instead relying on manual inspection by a trained analyst to identify shift periods.

The methods previosuly described require manual identification of data shift periods, which may not be feasible if hundreds or even thousands of data streams must be analyzed. This research aims to automate the process of detecting abrupt data shifts in PV data automatically via offline changepoint detection (CPD), without making any data assumptions. The resulting time series segments can then be analyzed separately, or the shortest segment can be removed during later analysis, similar to the process recommended by [4].

CPD is the process of detecting changes in an underlying signal, in particular a time series [5]. The concept of CPD dates back to the 1950's [6], and has various applications in speech processing [7], climatology [8], and financial analysis [9], among others. Offline CPD is a subcategory of CPD, where changepoints are detected in a signal after all data points have been collected [5]. Offline CPD can be formulated as a model selection problem, where we want the best possible segmentation of a signal with a specific quantitative criterion minimized [5]. CPD can be described as a combination of three elements: a cost function, which is a measure of the homogeneity between separate time series segments; a search method, which is the particular procedure employed to solve the optimization problem in question; and a constraint, which is either the number of changepoints in a sequence (if known) or penalty value associated with the goodness-of-fit term [5].

CPD has been applied to PV time series data previously, for different applications. Specifically, Theristis et al. [10], [11] applied the Facebook Prophet CPD algorithm to performance ratio (PR) time series with a non-linear degradation rate, with the intent of detecting degradation rate changes. This research differs from ours in that we are not looking to detect changes in degradation rates across a time series; rather, we are attempting to detect issues with the raw data itself, which occurs as a result of data acquisition problems. To our knowledge, this is the first attempt to automatically identify this particular issue in sensor-based PV data.

II. Methods

A. Data Sets

To build and validate the shift detection algorithm, 101 data sets representing unique sensor-based irradiance and power data streams were collected and labeled. Time series were collected from multiple PV solar installations, available via the NREL PV Fleets Initiative [12]. The PV Fleets Initiative is a US Department of Energy-funded project, where operational PV plant data is aggregated into a centralized cloud repository for the purpose of large-scale degradation analysis across the US. This database contains sensor-based time series data for over 1700 sites across the United States.

Data streams representing a variety of data shift behaviors, including scaling issues, were selected for labeling. Each time series was summed over a daily basis.

Two experts manually labeled data shifts in each of the 101 daily summed time series. A binary labeling strategy was used. Each individual data point in the time series was either labeled as a "data shift" point, where a major data shift occurs in the time series, or as a "regular" point where no change occurs. This labeling strategy was used to facilitate finding the specific point in a time series where a shift occurs, so data can be partitioned into individual issue-free segments.

B. Data Pre-Processing

Before applying the shift detection algorithm, each daily summed time series data set was cleaned, with the intent of removing egregious single-point outliers and anomalies. Specifically, the following steps were performed:

- All negative data days were removed.
- Stale data readings, i.e. consecutive repeat daily readings, were identified and removed from the time series. A consecutive repeat window of 6 readings or more was used to identify stale reading periods.
- All values less than the 1st percentile of data and greater than the 99th percentile of data were removed.
- Each daily time series was min-max normalized.

Irradiance and PV power time series can show extreme seasonality year-over-year. Removing seasonality helps to make the time series more stationary, and aids in detecting data shifts more accurately. Seasonality was removed from each time series via the following logic:

- The median value of each day of the year was calculated, resulting in a 365 day-long time series, with a median value for each day in the year. So, for example, in a three-year long time series, the three values occurring at January 1st in the time series are used to calculate the median value of January 1st.
- At each day in the time series, the median day value calculated in the previous step is subtracted from the normalized time series value.

An example time series, pre- and post-seasonality removal, is shown in Figure 2. It is important to note that all time series must be at least two years in length for this strategy to work, or seasonality cannot be calculated and removed.

C. Shift Detection Algorithm

The Python Ruptures changepoint detection package was used to develop and tune the shift detection algorithm [13]. An offline, unsupervised CPD algorithm was tuned using the manually labeled data sets. The following changepoint algorithm parameters were varied during grid search, to find the best-performing algorithm combination on the labeled data:

- Search method: Binary Segmentation, Window-based, Pruned Exact Linear Time (PELT), and Bottom-Up methods
- Cost function: radial basis function (rbf), L1, and L2
- Penalty: value between 10 and 100 inclusive, at intervals of 10. Higher penalty values cause heavier filtering of changepoints, resulting in fewer total detected changepoints. Because we are attempting to automatically detect

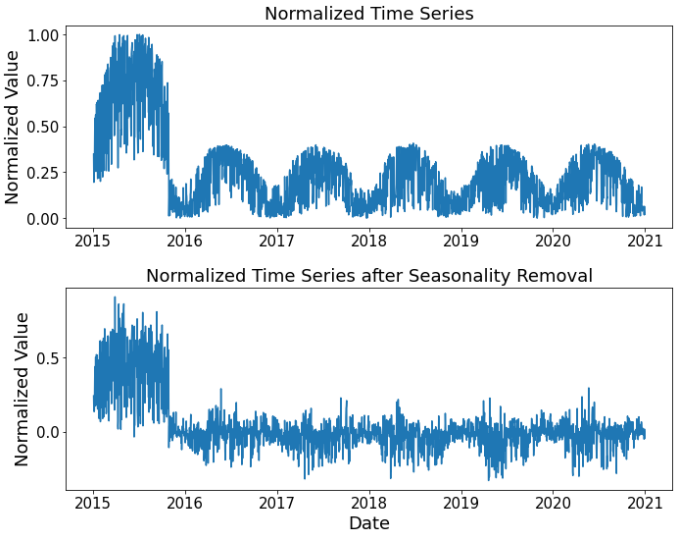

Fig. 2. An example time series before seasonality removal, and after seasonality removal.

data shifts with no prior knowledge of the time series, we do not pass a specific number of changepoints to find, and instead rely on a penalty threshold.

- Width: value between 10 and 110 inclusive, at intervals of 20. Width only applies to the window-based method, and acts as the length of the sliding window.

In addition to running seasonality-corrected data through the CPD algorithm, normalized time series data (with no seasonality correction) was also run. This was to identify the best parameter combination for situations where seasonality cannot be removed, i.e. time series shorter than 2 years in length.

D. Benchmarking Algorithm Performance

The ability of each model to successfully detect change points in the labeled time series was assessed, using CPD-specific F1-score and precision metrics developed by the Alan Turing Institute [14].

F1-score and recall are defined via the following equations, respectively [14]:

$$\text{F-score} = \frac{2 \cdot \text{recall} \cdot \text{precision}}{\text{recall} + \text{precision}} \tag{1}$$

$$\text{recall} = \frac{TP}{TP + FN} \tag{2}$$

where TP represents the number of true positives, and FN represents the number of false negatives. A true positive is defined as a detected changepoint within 30 days of a labeled changepoint. This 30-day period acts as a margin-of-error period, to allow for small discrepancies where the detected changepoint may be a few days off from the actual labeled changepoint. Recall is the fraction of relevant changepoints detected out of all correctly-labeled changepoints. A recall of 1 indicates that all labeled changepoints were found. An F1-score of 1 indicates perfect precision and recall, where the only changepoints detected by the algorithm are true positives.

In addition to measuring F1-score and recall, the Rand Index was measured for each test case. The Ruptures Python package implementation for the Rand Index was used [13]. The Rand Index measures the similarity between data clusters; in the case of changepoint detection, it analyzes the similarity between changepoint-separated time series segments [13]. For this research, the Rand Index is a valuable metric because we are most concerned with having consistent time series segments as a final output, not necessarily the changepoints themselves. The Rand Index is defined via the following equation [13]:

$$\text{Rand} = \frac{N_0 + N_1}{T(T+1)/2} \tag{3}$$

where N_0 represents the number of pairs of samples that belong to the same segment in a sequence T that has been split into segments T_1 and T_2, and N_1 is the number of pairs of samples that belong to different segments according to T_1 and T_2. The Rand Index is normalized between 0 and 1, where 0 indicates complete disagreement and 1 indicates complete agreement.

III. Results

Tables I and II show the five best CPD model configurations based on average F1 score for seasonality-corrected and normalized data streams, respectively. Table III shows Rand Index scores for these particular models. Generally, better overall metric scores were achieved when seasonality-corrected data was used.

Using seasonality-corrected data, the best-performing CPD model, a PELT model, achieved an F1-score of .767 and a Rand index value of 0.871. For normalized-only data, the best performing model, a window-based model, achieved an F1-score of .745 and a Rand Index value of .848. It is noteworthy that the best performing model overall (PELT model with seasonality-removed data) had one of the slowest overall run times, with an average run time of 50.81 seconds per a data stream, with the average data stream length of approximately 2300 data points. Several models achieved slightly lower average F1-scores on the data set, but have far faster average run times. When looking at model performance in terms of both time efficiency and accuracy, we recommend using the Bottom-Up model for seasonality-corrected data (average F1-score of 0.76 and Rand Index of .878, with an average run time of 0.26 seconds). For normalized-only data, we recommend using the Window-based model with the highest average F1-score (F1-score of .745 and Rand Index of .848, with a run time of .2 seconds).

The Rand Index scores for the highest-scoring models are particularly promising (values greater than .8), as they indicate that the final time series outputs are well-segmented. This is particularly important, as we want to perform analysis on data periods that are consistent and free of massive data shifts.

TABLE I
TOP 5 PERFORMING CPD CONFIGURATIONS ON LABELED, SEASONALITY-CORRECTED DATA

Model	Cost	Penalty	Recall	F1	Run Time (s)
PELT	rbf	40	.734	.767	50.81
Binary Seg	rbf	50	.705	.763	2.24
Binary Seg	rbf	40	.726	.762	2.37
PELT	rbf	50	.708	.761	54.99
Bottom-Up	rbf	40	.729	.760	.26

TABLE II
TOP 5 PERFORMING CPD CONFIGURATIONS ON LABELED, NORMALIZED DATA (NO SEASONALITY CORRECTION)

Model	Cost	Penalty	Width	Recall	F1	Run Time (s)
Window	rbf	30	50	.698	.745	.200
Window	rbf	40	50	.671	.741	.199
Window	rbf	20	30	.736	.739	.191
Window	rbf	20	50	.747	.737	.199
Window	rbf	50	50	.654	.736	.206

IV. PVANALYTICS INTEGRATION & FURTHER RESEARCH

The models developed in this research are available for public use via the Python PVAnalytics package [15]. In addition to automated data shift detection, the package includes functionality for identifying the longest continuous time series segment that is free of data shifts and isolating it for further analysis. This detection-and-segmentation process is illustrated in Figure 3.

Our logic for selecting the longest time series segment for further analysis is currently rudimentary, and we plan to develop more advanced processes for analysing and selecting the "best" time series segment for further analysis, based on each segment's overall data quality and availability. We also plan to further investigate whether data shifts caused by scaling issues or similar can be identified and corrected (rather than eliminated), without compromising the overall quality of the time series and biasing future analyses.

ACKNOWLEDGMENT

This work was authored in part by Alliance for Sustainable Energy, LLC, the manager and operator of the National Renewable Energy Laboratory for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under Solar Energy Technologies Office (SETO) Agreement Numbers 38258.

TABLE III
AVERAGE RAND INDEX FOR THE TOP PERFORMING MODELS

Data Type	Model	Cost	Penalty	Width	Rand
Season-Removed	PELT	rbf	40	NA	.871
Season-Removed	Binary Seg	rbf	50	NA	.864
Season-Removed	Binary Seg	rbf	40	NA	.867
Season-Removed	PELT	rbf	50	NA	.859
Season-Removed	Bottom-Up	rbf	40	NA	.878
Normalized	Window	rbf	30	50	.848
Normalized	Window	rbf	40	50	.840
Normalized	Window	rbf	20	30	.857
Normalized	Window	rbf	20	50	.838
Normalized	Window	rbf	50	50	.828

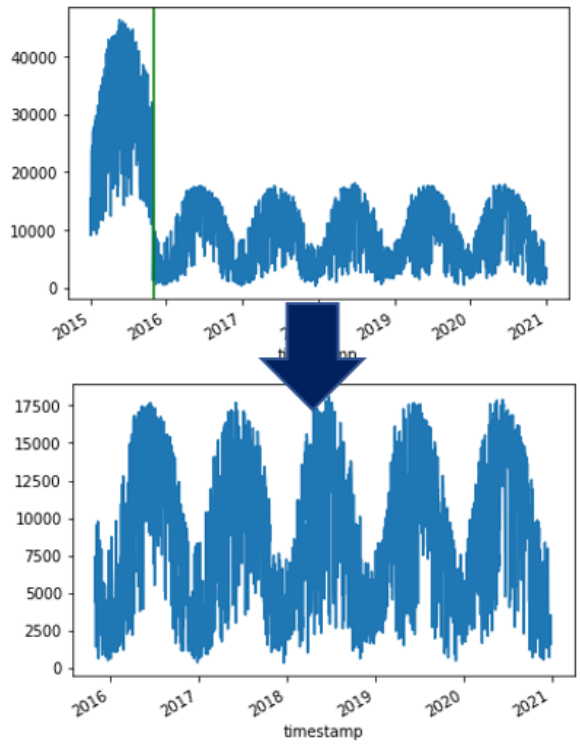

Fig. 3. PVAnalytics pipeline for automated detection of data shifts in time series, and segmentation of the longest sequence free of data shifts. In the upper image, data shifts are automatically detected (see green line) via the PVAnalytics detect_data_shifts() function. In the lower image, the longest time series sequence free of data shifts is segmented, using the PVAnalytics get_longest_shift_segment_dates() function.

REFERENCES

[1] A. Livera, M. Theristis, E. Koumpli, S. Theocharides, G. Makrides, J. Sutterlueti, J. S. Stein, and G. E. Georghiou, "Data processing and quality verification for improved photovoltaic performance and reliability analytics," *Progress in Photovoltaics: Research and Applications*, vol. 29, no. 2, pp. 143–158, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/pip.3349

[2] D. C. Jordan, C. Deline, S. R. Kurtz, G. M. Kimball, and M. Anderson, "Robust pv degradation methodology and application," *IEEE Journal of Photovoltaics*, vol. 8, no. 2, pp. 525–531, 2018.

[3] D. C. Jordan and S. R. Kurtz, "Analytical improvements in pv degradation rate determination," *2010 35th IEEE Photovoltaic Specialists Conference*, pp. 002 688–002 693, 2010.

[4] S. Lindig, A. Louwen, D. Moser, and M. Topic, "Outdoor pv system monitoring—input data quality, data imputation and filtering approaches," *Energies*, vol. 13, no. 19, 2020. [Online]. Available: https://www.mdpi.com/1996-1073/13/19/5099

[5] C. Truong, L. Oudre, and N. Vayatis, "Selective review of offline change point detection methods," *Signal Processing*, vol. 167, p. 107299, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165168419303494

[6] E. S. Page, "A test for a change in a parameter occurring at an unknown point," *Biometrika*, vol. 42, no. 3/4, pp. 523–527, 1955. [Online]. Available: http://www.jstor.org/stable/2333401

[7] Z. Harchaoui, F. Vallet, A. Lung-Yut-Fong, and O. Cappe, "A regularized kernel-based approach to unsupervised audio segmentation," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 1665–1668.

[8] J. Reeves, J. Chen, X. L. Wang, R. Lund, and Q. Q. Lu, "A

review and comparison of changepoint detection techniques for climate data," *Journal of Applied Meteorology and Climatology*, vol. 46, no. 6, pp. 900 – 915, 2007. [Online]. Available: https://journals.ametsoc.org/view/journals/apme/46/6/jam2493.1.xml

[9] M. Lavielle and G. Teyssière, *Adaptive Detection of Multiple Change-Points in Asset Price Volatility*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 129–156. [Online]. Available: https://doi.org/10.1007/978-3-540-34625-8_5

[10] M. Theristis, A. Livera, L. Micheli, C. B. Jones, G. Makrides, G. E. Georghiou, and J. S. Stein, "Modeling nonlinear photovoltaic degradation rates," in *2020 47th IEEE Photovoltaic Specialists Conference (PVSC)*, 2020, pp. 0208–0212.

[11] M. Theristis, A. Livera, C. B. Jones, G. Makrides, G. E. Georghiou, and J. S. Stein, "Nonlinear photovoltaic degradation rates: Modeling and comparison against conventional methods," *IEEE Journal of Photovoltaics*, vol. 10, no. 4, pp. 1112–1118, 2020.

[12] D. C. Jordan, K. Anderson, K. Perry, M. Muller, M. Deceglie, R. White, and C. Deline, "Photovoltaic fleet degradation insights," *Progress in Photovoltaics: Research and Applications*, vol. n/a, no. n/a. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/pip.3566

[13] C. Truong, "ruptures: change point detection in Python," https://github.com/deepcharles/ruptures, 2018.

[14] G. J. J. van den Burg and C. K. I. Williams, "An evaluation of change point detection algorithms," 2020.

[15] PVLib, "PVAnalytics," https://github.com/pvlib/pvanalytics, 2020.