Data-Driven Estimation of Soiling Loss and Optimal Cleaning Schedule for a Utility-Scale PV Plant

by

Abhinav Bora

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2023

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Soiling of panels in solar power plants can reduce production levels. In this thesis, we estimate the effect of soiling on power production and efficiency, as well as the gains from cleaning. Power data from a plant in southwest India was recorded every 5 minutes spanning 6 months. We analyzed this data to estimate efficiency degradation rates resulting from accumulation of soil and dust. The major challenge was filtering dataset noise/anomalies due to variations in micro-weather conditions. The key contribution of the thesis is a data-driven cleaning schedule algorithm. The algorithm detects cleaning events and produces a segmentation of the timeline into cleaning and soiling intervals. From the cleaning intervals we estimate the gains from panel cleaning, and from the soiling intervals we calculate the rate of power/efficiency loss. We apply these results to solve optimization problems regarding the cleaning schedule of a solar power plant. For example, by comparing the cost of cleaning against the potential gains in power production, we answer the questions "Which panel should I clean first/on this day?" and "Which day should I clean all panels?". We hope that the contributions of this research will provide important insights for any party working with solar power data.

Acknowledgements

I would like to acknowledge Professor Keshav for his patient guidance throughout my graduate studies and for inspiring a passion for practical research with real-world impact. Starting from the initial URA in Fall 2020, it has been my great privilege working with you these past 3 years, learning not only what it takes to succeed in academia but life in general. Thank you for your consistent advice and direction in our weekly meetings. I left every conversation feeling confident in my research with plenty of interesting new approaches to consider. For the opportunities you have given me, I will forever be grateful. Needless to say, without your support and encouragement this thesis would not be possible.

I would also like to thank Professor Golab for agreeing to co-supervise on short notice, keeping updated with thesis progress, and providing feedback every step of the way. As well, thank you to Professor Wong and Professor Al-Kiswany for reading the thesis and providing excellent comments for improvement. Finally, thank you to Sharat and Hugh from Quadrical for providing access to the PV data. Your insights and knowledge into the solar field have been invaluable throughout this research.

Dedication

To my parents, grandparents, and little sister, for their wholehearted support, belief, and inspiration throughout my life.

Table of Contents

Author's Declaration ii

Abstract iii

Acknowledgements iv

Dedication v

List of Figures x

List of Tables xii

1 Introduction 1

1.1 Soiling . . . 1

1.1.1 Quantifying Soiling Losses . . . 3

1.1.2 Mitigating Soiling Losses . . . 4

1.1.3 Optimal Cleaning Schedule . . . 5

1.2 Research Problem . . . 6

1.3 Research Approach . . . 7

1.3.1 Overview . . . 7

1.3.2 Real-World Issues and Solutions . . . 7

1.4 Thesis Structure . . . 8

2 Related Work **10**

2.1 Taxonomy . . . 10

2.2 Research Goals . . . 11

2.3 Literature Survey . . . 12

2.4 Discussion . . . 12

2.4.1 Lab Studies Papers . . . 12

2.4.2 Operational Studies . . . 14

2.4.3 Soiling Loss . . . 15

2.4.4 Cleaning Optimization . . . 17

3 Dataset Overview **19**

3.1 Data . . . 19

3.2 Step 1 - Tracking Soiling Losses . . . 21

3.3 Efficiency Calculation . . . 22

3.4 Inconsistencies in Data . . . 23

4 Methodology **28**

4.1 Data-Cleansing Methodology . . . 28

4.1.1 Data Filters . . . 28

4.1.2 Cumulative Efficiency Ratios . . . 29

4.1.3 Cloudy Day Detection . . . 30

4.1.4 Best-Fit Pyranometer Measurements . . . 33

4.1.5 Summary . . . 36

4.2 Human-Recorded Cleaning Logs . . . 36

4.2.1 Step 2 - Segmenting the Efficiency Timelines . . . 36

4.2.2 Unreliable Recording of Cleaning Events . . . 36

4.3 Data-Driven Segmentation Algorithm . . . 37

4.3.1 Motivation . . . 37

4.3.2 Initial Attempts . . . 38
4.3.3 Final Version . . . 39
4.4 Step 3 - Cleaning Schedule Optimization . . . 43
4.4.1 Extraction of Key Information . . . 44
4.4.2 Cleaning Benefit Model . . . 44
4.4.3 Cleaning Profit Calculation . . . 46
4.4.4 Simulations to Understand D and G . . . 49
4.4.5 Optimal Cleaning Dates . . . 53
4.4.6 Summary . . . 55

5 Results **56**
5.1 Original Data . . . 56
5.2 Timeline after Data-Cleansing . . . 57
5.3 Segmented Timeline . . . 59
5.4 Extracting Key Information . . . 60
5.5 Current Cleaning Profit Calculation . . . 61
5.6 Future Profit Projections . . . 62
5.7 Optimization Results . . . 63
5.8 Evaluation . . . 64
5.8.1 Periodic Cleaning . . . 66
5.8.2 Optimal Cleaning . . . 68
5.8.3 Profit Comparison . . . 69

6 Discussion **71**
6.1 Contributions . . . 71
6.2 Limitations . . . 72
6.3 Future Work . . . 74

References **77**

APPENDICES 81

A Heatmap Plots of PV Plant 82

B Cleaning Log Excerpts 86

C Segmentation Algorithm Code 87

D Gradual/Back-to-Back Cleanings 89

E K-Means Clustering for Union Cleaning Records 90

List of Figures

1.1 Badhla Solar Park - 2245 MW PV Plant in Rajasthan, India [23] 2

1.2 2023 Energy Market Forecast, United States [16] 3

1.3 Clean vs Soiled PV Module. Image presented in [10] 4

1.4 Impact on Incident Sunlight due to Soiling. Diagram by Al Hicks (NREL, USA) as presented in [38] 5

1.5 Thesis approach, including issues encountered at each step 9

2.1 Relative soiling ratios showing spatial variation in plant performance, as presented in [19] 15

3.1 Simple PV diagram 19

3.2 Contradicting values of relative power and irradiance 24

3.3 Noisy irradiance and power readings for Nov 14, 2020 25

3.4 Micro-weather conditions relating to cloud cover causing uneven shading of PV plant 26

4.1 Clear day with good correlation coefficient = 0.998 31

4.2 Cloudy day with poor correlation coefficient = 0.764 31

4.3 Partially cloudy day with good correlation coefficient = 0.996 32

4.4 SCB 454 Power Output with `Hz_Irr_1_pyrano` Irradiance Readings 34

4.5 SCB 454 Power Output with `Hz_Irr_2_pyrano` Irradiance Readings 34

4.6 Algorithm behaviour for a typical efficiency pattern indicating cleaning interval 41

4.7 Erroneous spike in efficiency timeline, safely ignored by our algorithm . . . 42
4.8 Segmented Efficiency Timeline Example . . . 43
4.9 Cleaning Benefit Model . . . 45
4.10 Cleaning Benefit Area Calculation . . . 47
4.11 Progression of GD as we approach rainfall . . . 50
4.12 Cleaning Profit Timeline for 15 Days Till Rainfall . . . 51
4.13 Cleaning Profit Timeline for 5 Days Till Rainfall . . . 52
4.14 Cleaning Profit Timeline for SCB 246 . . . 54

5.1 Raw Efficiency Timeline for SCB 216 . . . 57
5.2 Clean Efficiency Timeline for SCB 216 . . . 58
5.3 Segmented Efficiency Timeline for SCB 216 . . . 59
5.4 Cleaning Profit Timeline for SCB 216 . . . 63
5.5 Baseline 6-Week Power Production Timeline . . . 65
5.6 Cleaning Profits . . . 70

A.1 Heatmap of Latest Efficiencies . . . 83
A.2 Heatmap of Latest Soiling Rates . . . 84
A.3 Heatmap of Average Power Production Gains from Cleaning . . . 85

B.1 Cleaning Records for SCB 454 . . . 86

C.1 Segmentation Algorithm Main Body . . . 87
C.2 Class Definition of a Segment . . . 88

D.1 Missed Detection of Cleaning Period . . . 89

E.1 Clustered Efficiency Timelines . . . 91
E.2 Clustering Results Projected on Site Map . . . 92
E.3 Example of a Union Cleaning Record . . . 93

List of Tables

2.1 Summarized analysis of related works 13

4.1 Profit Calculations 51

5.1 Extracted Information from each Segment 60

Chapter 1

Introduction

Photovoltaic (PV) modules produce electricity by absorbing energy from a virtually unlimited resource: sunlight, via a process known as the photovoltaic effect [17]. *Utility-scale PV plants* (Figure 1.1) aim to maximize these capabilities, generating solar energy using arrays of panels [36]. A typical plant setup, illustrated in Figure 3.1 (Section 3.1), consists of panels organized into *strings* feeding into *string combiner boxes* (SCBs).

As a renewable and economic energy source [40], solar power has seen accelerating growth in recent years with major PV plants being developed in many countries [36]. For example, the Energy Information Administration (EIA) forecasts 29.1 GW of new utility-scale solar installations added in the U.S. for 2023, more than doubling the previous record of 13.4 GW (Figure 1.2) [35]. Further, in 2021 solar PV accounted for 3.6% of world-wide electricity production and overall PV power generation increased by a record 179 TWh, up 22% from 2020 [21]. To sustain this growth, industry research concentrates on affordability and performance of solar technologies [32], aiming to improve production efficiency by minimizing losses in energy yield.

1.1 Soiling

Key areas of PV research include the development of high-efficiency materials, solar plant configuration to maximize energy generation, and improving the durability of PV modules [31]. However, the single most influential factor impacting energy yield is incoming solar irradiance. If the sunlight levels reaching the PV module are reduced, the system will suffer proportional losses in electricity generation. Therefore, *soiling*, the accumulation of dirt,

Figure 1.1: Badhla Solar Park - 2245 MW PV Plant in Rajasthan, India [23]

dust, and other particulate matter on the surface of PV modules, presents a considerable challenge for plant operators (Figure 1.3). Soiling particles disrupt normal optical processes by absorbing, reflecting, and scattering incident sunlight away from the PV module (Figure 1.4), diminishing performance [38].

A 2018 analysis of global solar power production estimates that soiling reduced energy yield by 3%-4%, amounting to an annual revenue loss of € 3-5 billion, which is expected to rise up to 4%-5% and more than € 4-7 billion by 2023 [22]. These losses are driven by growth in PV plant deployments to regions with arid, dusty climates which present great energy potential due to high levels of irradiance, but also exhibit harsh environmental conditions favouring soiling [22]. Moreover, soiling is expected to further intensify in coming years with climate change leading to rising temperatures and risk of droughts [1]. As a result, there is heightened interest in measurement and mitigation of soiling losses impacting utility-scale PV plants.

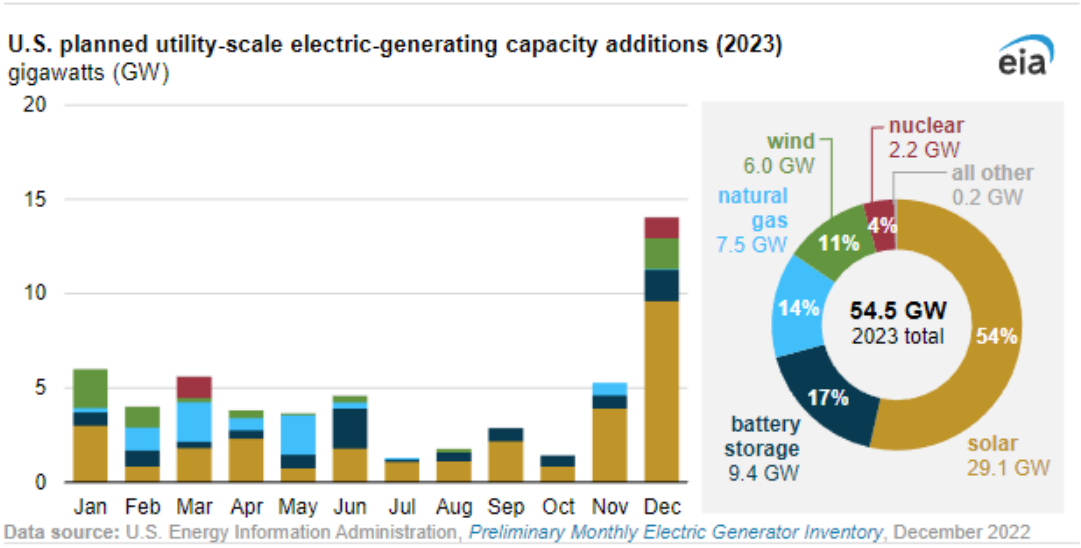


Figure 1.2: 2023 Energy Market Forecast, United States [16]

1.1.1 Quantifying Soiling Losses

The dominant soiling measurement techniques at this time involve the installation of soiling stations and sensors [2]. Soiling stations consist of two PV modules - one left exposed to soiling and one cleaned regularly to use as reference. The power outputs from both are compared to estimate soiling losses for the PV plant. Similarly, optical soiling measurement (OSM) sensors infer losses by comparing incoming vs transmitted energy through the PV glass, as affected by the optical characteristics of soiling materials deposited on top [3]. In addition, soiling image analysis (SIA) sensors operate on aerial photos of PV modules, using image-processing methods to estimate loss, based on soiling area coverage of the module surface [3]. While these techniques use different methods to measure soiling loss, they suffer from issues of cost, necessity for maintenance, and reliability [3].

Alternatively, some measurement techniques accurately detect soiling loss across a PV plant, without the need for any external instrumentation. These methods, known as *soiling extraction algorithms*, convert PV systems themselves into detectors and analyze power output data from modules to directly estimate soiling losses/rates [3]. While the use of raw system data allows for in-depth analysis, non-soiling factors such as noise and operational errors, with potential to invalidate results, must be identified. In this thesis, we develop data-cleansing methodology (Section 4.1) and a novel data-driven soiling extraction algorithm (Section 4.3) to quantify soiling loss for a utility-scale PV plant.

Figure 1.3: Clean vs Soiled PV Module. Image presented in [10]

1.1.2 Mitigating Soiling Losses

Given an accurate measurement of the ongoing production loss due to soiling, an optimal mitigation strategy can be developed. While this strategy differs from plant to plant depending on site-specific characteristics, such as PV system design/configuration, location-based environmental conditions (precipitation, soil aridity etc.), and soiling properties, in general, mitigation techniques fall under two categories: preventive and corrective measures [1].

Preventive measures aim to passively reduce soiling accumulation in the PV modules and involve efforts from the plant's engineering and construction teams [1]. One example is the use of anti-soiling coatings (ASC) which employ electrostatic forces to hinder particles from depositing [1]. Another interesting approach involves upside-down stowing of PV modules during the night, based on insights that soiling may become more pronounced at night [22][18]. An application of this strategy in India reported to reduce soiling loss by around 60% [5].

However, to our knowledge no preventive measures have been developed which completely eliminate the need for corrective measures such as cleaning. Cleaning is the most common mitigation technique and involves efforts from the plant's operation and maintenance (O&M) teams [22]. Cleaning methods can be divided into 3 categories:

- **Manual** - using dusting brooms and water mops/brushes [1]

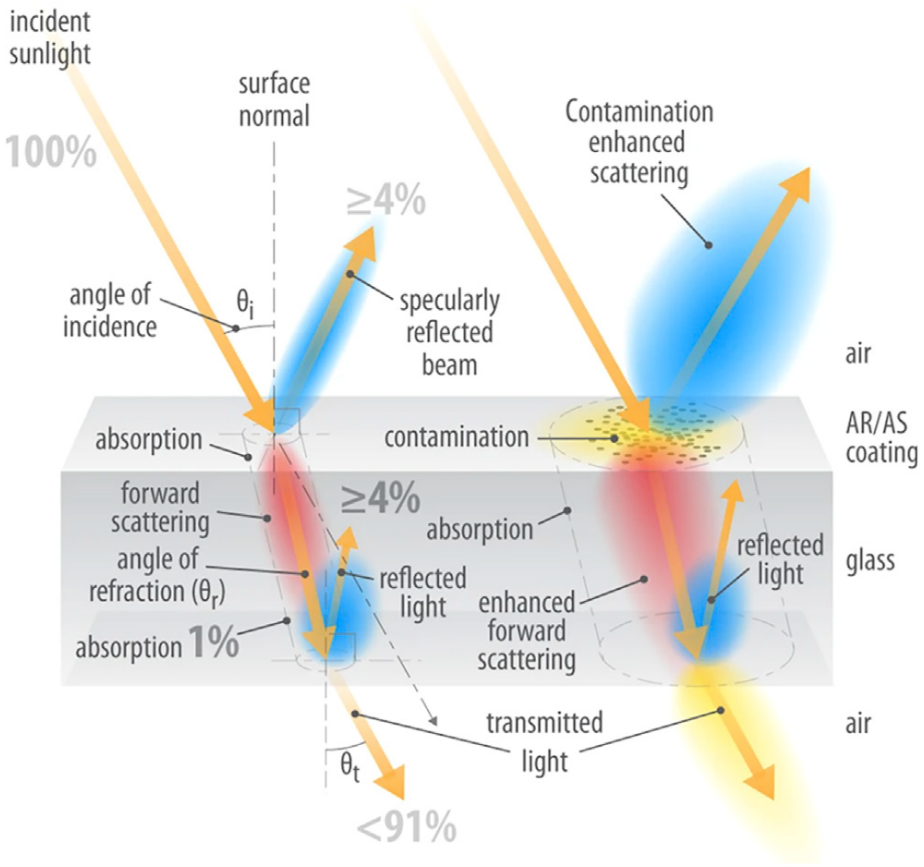


Figure 1.4: Impact on Incident Sunlight due to Soiling. Diagram by Al Hicks (NREL, USA) as presented in [38]

- **Semi-automatic** - via truck-mounted devices fitted with water tanks, or battery-powered portable robots and motorized brushes [1]
- **Fully-automatic** - devices designed to clean rows of modules at scheduled times without any human operation [1]

1.1.3 Optimal Cleaning Schedule

Regardless of the method, the cleaning of PV modules, especially in utility-scale PV plants, comes at a non-negligible price, consisting of both labour, time, transportation, and capital

costs [22]. In addition, scarcity of water in desert-like environments, the preferred locations for PV plants, can also drive up cleaning costs. Considering these costs, plant O&M teams may decide to distribute cleanings sparsely. However, losses in revenue begin accumulating each day the plant is not cleaned, due to decreases in power production from soiling. So while the cleaning costs can be minimized by cleaning as infrequently as possible, the soiling losses are minimized by cleaning as frequently as possible [25]. This presents an interesting dilemma for plant O&M teams, who must aim to balance both competing costs/losses and schedule cleanings to maximize plant production.

To further complicate matters, rainfall events, albeit rare and often insufficient in arid climates, also need to be considered as they can remove soiling at no cost [26]. Cleanings performed immediately before or after rainy days are particularly inefficient, as they provide limited production benefit relative to the cleaning cost.

Therefore, while the choice of cleaning method differs from site-to-site, cleanings at all plants need to be scheduled carefully, at times that minimize total cost/loss while maximizing total energy yield. For this, accurate measurements of current soiling loss and rates, as well as historical analysis of past cleanings is required. In this thesis, we develop an optimal cleaning schedule for a utility scale PV plant using knowledge on past cleaning quality/gains (Section 4.4).

1.2 Research Problem

As detailed in the previous sections, soiling is a major problem for PV plant operators. From their point-of-view, it is obvious that cleaning is necessary given the significant revenue losses due to soiling. However, what is not obvious is exactly when to clean for optimal profit, especially given rainfall events.

To determine an optimal cleaning schedule, a detailed cost-benefit analysis is crucial. This requires calculations of soiling losses and expected power production gains from cleaning, both of which are site-specific and involve careful examination of the site's PV data (consisting of time series of power outputs for the plant's PV modules).

Now, while plant operators have access to the PV data, they have trouble analyzing it due to the abundance of noise within the data and the lack of tools/algorithms to address it and extract the relevant values. Instead, cleanings are often based on rules of thumb such as performing cleanings once a week/month or cleaning specific modules which look sufficiently soiled. Of course, this leads to sub-optimal cleanings, meaning avoidable cleaning costs and soiling losses are incurred.

1.3 Research Approach

In this thesis, we approach soiling and cleaning from a purely **data-driven** perspective. Specifically, we develop data-driven algorithms and techniques to calculate soiling losses, estimate cleaning benefits, and ultimately optimize cleaning schedules. Plant operators can apply these to schedule cleanings which maximize profit.

1.3.1 Overview

Given raw PV data, we describe the following 3-step approach towards an optimal cleaning schedule:

1. Use PV data to calculate module-efficiency timelines (Chapter 3 and Section 4.1)
    - Tracking performance of modules over time
2. Segment timelines into cleaning and soiling intervals (Section 4.2 and 4.3)
    - Marking periods of soiling based on deteriorating module efficiency
    - Identifying cleaning events based on upswings in efficiency
3. Extract soiling losses and cleaning gains from segmented timelines, and use to optimize cleaning schedule (Section 4.4)
    - Calculating performance impact during soiling intervals and power production gains from cleanings
    - Comparing with cleaning costs and upcoming rainfall events to determine optimal cleaning dates for each PV module

Implementation details follow in the referenced sections.

1.3.2 Real-World Issues and Solutions

In processing PV data from any utility-scale plant, there are several problems that arise due to *real-world* conditions relating to the nature of the hardware, the environment, and day-to-day plant operations. Many such problems were encountered over the course of the thesis work, as we progressed through each step in the approach, and motivated the development of data-driven solutions, which are the key contributions of the thesis. These issues and solutions, along with brief comparisons to existing works, are detailed below:

- Problem # 1: Noise/anomalies in PV data such as operational errors, shading due to clouds, and micro-weather conditions making it difficult to calculate accurate efficiency timelines for Step 1 (Chapter 3)
  - Contribution # 1: Development of **data-cleansing methodology** (filters, cumulative efficiency ratios, cloudy day detection, best-fit pyranometer measurement) which provides insights/techniques to identify and address noise (Section 4.1)
    - Existing work lacks systematic methodology and often only consists of visual recognition and exclusion of outliers in the dataset
- Problem # 2: Unreliable recording of cleaning events by the cleaning crew, proving problematic when trying to segment efficiency timelines based on cleaning events for Step 2 (Section 4.2)
  - Contribution # 2: Development of a **data-driven timeline segmentation algorithm** which tracks changes in slope to detect cleanings and soiling intervals without needing human-recorded cleaning logs (Section 4.3)
    - Existing algorithms calculate day-to-day differences between performance metrics, marking cleanings whenever the difference is greater than a threshold value. This method is vulnerable to noise/outliers and lacks robustness compared to our approach.
- Problem # 3: Incomplete cleanings by the cleaning crew, complicating cleaning gain estimation in Step 3 as we cannot assume perfect cleanings
  - Contribution # 3: Determining expected cleaning gains based on gains of past cleaning intervals, rather than simply inverting the soiling loss (Section 4.4)
    - Existing works assume perfect cleanings

Note that all algorithms, techniques, and solutions are also **data-driven**, requiring only the PV data to address the real-world issues. This is to the benefit of plant operators and owners, as an optimal cleaning schedule, which maximizes profit, can be determined using already available PV data without installing costly equipment.

1.4 Thesis Structure

Figure 1.5 details the steps, issues, solutions, and contributions presented in the thesis.

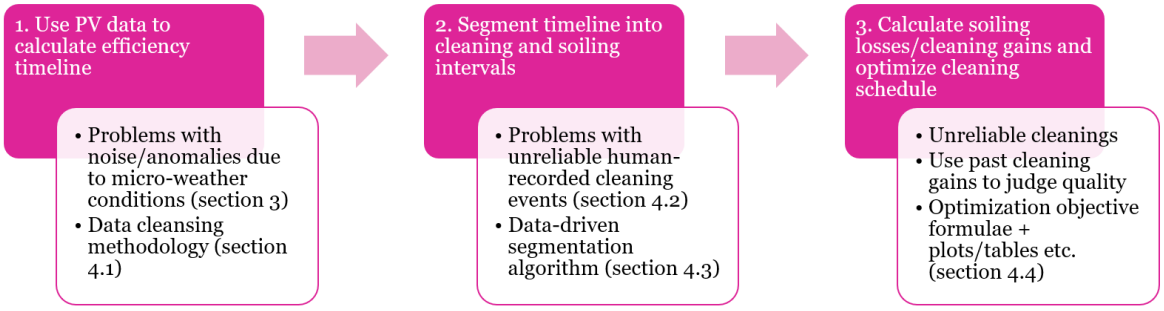


Figure 1.5: Thesis approach, including issues encountered at each step

The remainder of the thesis is structured as follows. A detailed examination of existing works is presented in Chapter 2. In Chapter 3, we outline the available PV data, establish formulae for efficiency calculations, and investigate anomalies in the dataset. Key contributions are detailed in Chapter 4. Finally, results are presented in Chapter 5 and discussed in Chapter 6.

Chapter 2

Related Work

2.1 Taxonomy

Soiling and its impact on PV performance has been a topic of research for nearly 8 decades [9]. With recent technological advances driving down PV cost, leading to increasing investments and market incentives, soiling-related research output has observed exponential growth in the last decade [9]. In general, research on the soiling PV performance relationship can be classified into two branches:

- **Operational-Level Research** - **Directly measuring soiling losses** for large-scale plants using **operational-level PV data**
- **Laboratory Studies** - Development of optical/physical/technical soiling models to **estimate expected soiling loss** using **small-scale experimental setups**

Operational-level research is relatively sparse, due in part to limited access to utility-scale PV data. Papers in this branch often lack detail, for example, calculating site-wide results rather than analyzing soiling losses across different areas of the site.

Lab studies are common. However, they estimate only projected soiling losses for PV plants based on experiments conducted using a small number of modules. Further, lab studies do not consider the real-world characteristics (Section 1.3.2) exhibited by large-scale solar plants where the operations team isn't rigorous in their duties. Instead, calculations are based on experiments performed under strict supervision of research personnel with

perfect conditions, perfect monitoring, perfect cleanings etc. As such, their methodology and the subsequent results are not applicable in a real-world setting.

In contrast, the research in this thesis is both extensive and large-scale, with data from individual PV modules used to calculate per-module soiling losses. Real-world characteristics leading to issues of data reliability are thoroughly investigated and resolved. In this way, we aim to address limitations of existing work and provide data-driven analysis of soiling loss and optimal cleanings for a utility-scale PV plant.

2.2 Research Goals

Here are our research goals:

- Data-driven - Calculating soiling/cleaning results based on analysis of PV data
- Large-scale - Working with operational-level PV data from a utility-scale plant, rather than using data from small-scale experimental setups
- SCB-level metrics - Calculating soiling losses/rates and optimal cleanings at an SCB level as opposed to site-wide results
  - Necessary due to the **spatial variability** of soiling. Soiling is a non-uniform phenomenon, especially across a large utility-scale PV plant, and different areas within a single site can see vastly different soiling conditions [1]. Thus, PV modules will exhibit varying soiling rates/losses and require different optimal cleaning schedules
  - If site-wide losses or optimal cleanings are required, the SCB-level calculations can easily be aggregated to produce site-wide results
- Robust to real-world conditions and noise - Not assuming perfect data/monitoring, perfect cleanings etc.
  - Developing data-driven solutions to systematically detect data anomalies/noise, determine the root causes, and address/work-around them
- Real-world results - Not projections and estimations based on experimental models
  - Calculations and methodology involving actual PV plant data, producing results which can be used and applied by the plant operators

- Robust to imperfect cleaning - Not assuming perfect cleanings which fully restore PV system efficiency
  - Instead, calculating expected production gains from cleaning based on quality and gains of past cleanings (or a similar method)
- Cleaning schedule optimization - Using soiling loss and cleaning gain results to determine the optimal cleaning dates
  - Accounting for cleaning costs and rainfall distributions

2.3 Literature Survey

We now present a survey of the PV soiling and cleaning field. Papers are organized into Table 2.1 to highlight their contributions and shortcomings with respect to our research goals.

Note that while all of the surveyed papers calculate soiling loss, none consider imperfect cleanings. Instead, they assume cleaning will negate all soiling loss (cleaning gain = soiling loss) and thus do not calculate cleaning gains separately for optimization. Moreover, no paper accounts for real-world conditions and no paper presents soiling results on a per-SCB level.

Further discussion on the papers' contributions and limitations in comparison to thesis work is presented in the following section.

2.4 Discussion

We now explore a subset of particularly intriguing and relevant papers in more detail.

2.4.1 Lab Studies Papers

Representative of most lab studies, Micheli et al. [29] analyze data from soiling stations to develop a model for predicting soiling loss, using parameters such as pollution indexes, land characteristics, and meteorological data. The purpose is to determine if soiling losses can be estimated without analyzing site-specific PV data, and instead using more widely-available parameters.

		Desired Characteristics						
		Data-driven	Large-scale	SCB-Level Metrics	Real-World Conditions	Real-World Results	Imperfect Cleanings	Cleaning Schedule Optimization
Lab Studies	Micheli et al., 2017 [29]	✓						
	Bessa et al. [3]							
	Deceglie et al., 2018 [11]	✓						
	Micheli et al., 2019 [28]	✓						✓
	Jones et al. [25]	✓						✓
	Urrejola et al. [39]	✓						✓
	Besson et al. [4]	✓			≈			
	Rodrigo et al. [34]	✓						✓
	Chiteka et al. [7]	✓						✓
	Yazdani et al. [41]	✓						✓
Operational-Level	Mejia et al. [26]	✓	✓			✓		
	Micheli et al., 2021 [27]	✓	✓			✓		✓
	Skomedal et al. [37]	✓	✓			✓		
	Deceglie et al., 2016 [12]	✓	✓			✓		
	Gostein et al. [19]	✓	✓	≈	≈	✓		
	Pavan et al. [33]	✓	✓			✓		
	Diouf et al. [13]	✓	✓			✓		✓
	Thesis	✓	✓	✓	✓	✓	✓	✓

Table 2.1: Summarized analysis of related works

The problem is the reliability and accuracy of such a model. Soiling is an extremely complex process, with hundreds of potential parameters, including precipitation patterns and concentrations of airborne particulate matter, contributing to its impact on PV systems [29]. Rather than modelling the influence of each parameter to project the expected soiling loss at a site, using the plant's PV data to directly calculate production losses is much simpler and more accurate [3]. Moreover, plant operators have direct access to their site's PV data. And so, it is considerably easier to follow data-driven algorithms to calculate **actual soiling loss**, as presented in this thesis, instead of labouring to acquire external data on all model parameters, only to determine an **estimate** of the soiling loss.

Continuing, the paper highlights additional problems, common to other lab studies such as those by Chiteka et al. [7] and Yazdani et al. [41]. Estimations of soiling loss for large-scale PV plants are based on data from small-scale experimental setups, including soiling stations consisting of only 2 modules. Further, since these setups are strictly monitored and perfectly maintained by the research team, the collected data is not subject to noise and anomalies stemming from real-world conditions. Due to these differences, any methodologies and results derived from lab studies cannot be representative of utility-scale PV plants in the real world.

2.4.2 Operational Studies

Operational-level research focuses on analysis of large-scale data. For example, Mejia et al. [26] quantified losses in efficiency of 186 residential and commercial PV sites due to soiling, with results indicating upwards of 0.1% soiling loss per day for some sites. On average, soiling loss was 0.00051% per day in efficiency, close to the 0.0007% daily loss calculated in this thesis (Chapter 6). However, the results are calculated for dry periods with respect to rainfall events observed at nearby weather stations. They do not consider manual/scheduled cleanings and do not make any attempt to apply the calculated rates, instead aiming to only quantify soiling loss at PV plants under natural conditions.

In addition, similar to all other operational-level papers, site-wide soiling losses are calculated rather than per-module/per-block soiling losses. As explained earlier, soiling varies across different areas of a single site, and an extensive analysis of area-specific soiling losses is necessary to accurately reflect soiling conditions and schedule cleanings in specific areas to save costs (as detailed in Section 4.4). This spatial variability of soiling is explored briefly by Gostein et al. [19] with block-level soiling losses being calculated across a PV plant (Figure 2.1), but further analysis and application is left to future studies. In this thesis, we examine spatial variability of soiling thoroughly, calculating SCB-level soiling losses (Appendix A Figure A.2) to schedule area-specific optimal cleanings.

Figure 2.1: Relative soiling ratios showing spatial variation in plant performance, as presented in [19]

Finally, although noise/anomalies with the large-scale PV data are identified, the data-cleansing methodology is very limited. Instead of robust systematic detection and resolution of noise, such as the techniques presented in this thesis (Section 4.1), large portions of the noisiest site/days data are manually identified and excluded using thresholds [26]. This methodology is prone to mistakes and cannot reliably address all sources of noise.

2.4.3 Soiling Loss

All surveyed papers calculate soiling loss, but using a variety of techniques. We now describe a soiling monitoring literature review by Bessa et al. [3] to discuss different monitoring methods and soiling loss calculation algorithms in comparison to the thesis work.

There are, in general, 3 classes of soiling monitoring approaches:

1. Soiling monitoring instrumentation - Involving installation of soiling stations and sensors as discussed in Section 1.1.1
    - This is expensive and requires regular maintenance from plant personnel [3]
2. Soiling estimation models - Studying the correlation between soiling and environmental parameters as discussed in Section 2.4.1

- Relatively lower accuracy than the other 2 methods [3]

3. Soiling extraction algorithms - Direct measurement of power output loss due to soiling by analyzing PV system data

    - Usually two steps: Detection of cleaning events and corresponding soiling intervals, followed by calculation of soiling loss using data within the soiling intervals

Our approach to soiling loss calculation, namely the efficiency timeline segmentation algorithm detailed in Section 4.3, can be classified as a soiling extraction algorithm. This method is preferred for two reasons. First, no expensive soiling detectors are required, as all values needed for calculations are already measured by PV system themselves [3]. Second, cleaning logs for PV plants are often unreliable or missing. Detecting cleaning events directly from PV data eliminates the dependence on the plant's O&M team to maintain an accurate cleaning record. The caveat, however, is that thorough data-cleansing methodology (Section 4.1) is necessary to avoid non-soiling related issues affecting the data-driven calculations.

A well-known soiling extraction algorithm is the *Stochastic Rate and Recovery (SRR) model*, which serves as the basis for many other adaptations [3]. This algorithm detects cleanings based on positive changes in the performance metric. For instance, supposing we examine efficiency timelines of PV modules, SRR would calculate deltas between efficiencies of consecutive days and mark cleanings for outlier deltas ie. positive deltas larger than Q3 + 1.5*IQR, where Q3 and IQR are the third quartile and interquartile range of all deltas [11]. However, this method is prone to error with noise, such as day-to-day spikes in efficiencies resulting in incorrect cleanings being detected. In the development of our segmentation algorithm, we faced similar issues when initially trying to mark cleanings based on increases in efficiency above threshold values. However, transitioning from this value-based approach to the slope-based approach (tracking changes in efficiency slope rather than the values itself) detailed in Section 4.3 proved to be more robust.

Adaptations of the SRR model have also been proposed, such as the algorithm by Skomedal et al. [37], which tracks shifts in median values rather than day-to-day values. However, a threshold parameter is still required to be manually tuned and calibrated to the plant's data [37]. No such site-specific parameters/thresholds are needed for our segmentation algorithm, as instead of comparing changes in efficiency value, we simply compare changes in the direction/sign of the efficiency slope.

Once the cleaning events have been identified, to determine soiling loss between cleanings, the SRR model uses the Theil-Sen Estimator to calculate slope and rate of daily

soiling loss, as described by Deceglie et al. [12]. This is exactly the same soiling loss calculation implemented in our segmentation algorithm, due to the Theil-Sen Estimator's effectiveness in the presence of outliers.

2.4.4 Cleaning Optimization

While every surveyed paper calculates soiling loss, some also apply these results to optimize cleanings. Here, we explore different optimization approaches.

Micheli et al. [28] analyzed data from soiling stations in California and Arizona to investigate the seasonal variability of soiling rates. Similar to most lab studies, the data is small-scale and no data-cleansing methodology is provided. Once again, site-wide soiling rates are calculated, rather than investigating per-module rates. However, while previous approaches calculated annualized soiling rates, in this paper, seasonal dependence of soiling rates is suggested and data from each month is examined separately to calculate monthly soiling rates. In this thesis, we also respect time-of-year variability of soiling, and calculate soiling rates for each soiling interval individually. Doing so allows for the use of only the most recent soiling rates from the latest or current soiling interval. This results in a more accurate representation of current soiling conditions for cleaning optimization, when compared to using annualized rates.

A fairly naive optimization approach is proposed in the paper, simply determining the day in which cleaning would have the highest positive impact on energy yield [28]. Here, no economic analysis is conducted, and only a single site-wide cleaning scenario is considered, rather than a cleaning schedule with multiple cleanings. Most importantly, cleaning is again assumed to restore the PV system to full operational efficiency. In comparison, the optimization approach proposed in this thesis (detailed in Section 4.4) considers multiple area-specific optimal cleanings performed at different dates. These dates are determined via cost-benefit analysis of cleaning costs and expected energy gains, which are calculated by examining the quality of past cleanings rather than assuming 100% cleaning effectiveness.

In addition to the cost-benefit analysis used in this thesis, there are other, more complicated economic metrics used to evaluate and determine the optimal cleaning schedule. For example, many of the surveyed optimization papers ([27] [25] [39] [4] [34]) employ varying financial models such as scheduling cleanings to maximize the Net Present Value (NPV) [14] and the Levelized Cost Of Energy (LCOE) [6]. However, these metrics require the input of many complex parameters which may not be available to plant operators. In contrast, our optimization model is relatively straightforward and easily applicable.

The thesis contributions, namely the data-cleansing methodology, the efficiency timeline segmentation algorithm, and the cleaning schedule optimization via smart cleaning-benefit calculations, address limitations found in existing works. Together, they constitute a novel approach for utility-scale PV plant operators to effectively monitor, and mitigate, soiling impact on plant production.

Chapter 3

Dataset Overview

3.1 Data

We analyze data from a utility-scale PV power plant in southwest India. The plant comprises of 207,015 solar panels organized into *strings*, which consist of PV modules connected in series. Power outputs from multiple strings are combined into one output by a *string combiner box* (SCB) and outputs from multiple SCBs are then fed into *inverters* on route to the utility grid. Figure 3.1 shows a typical PV plant unit.

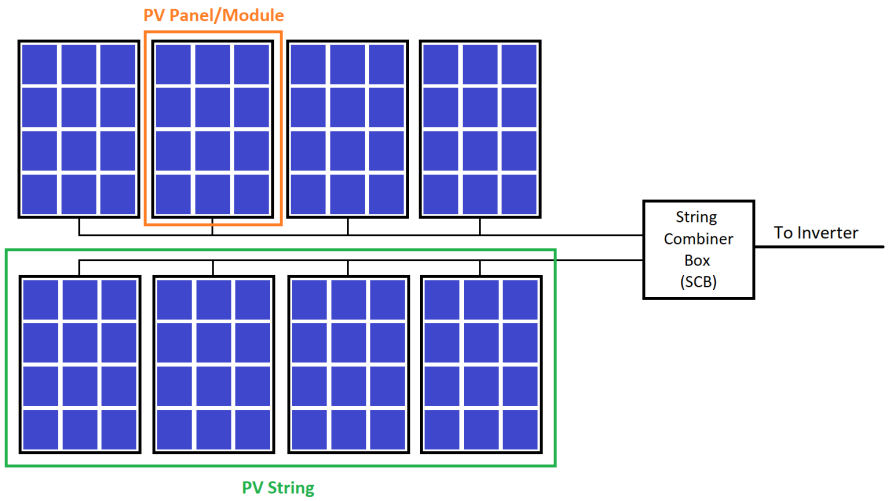


Figure 3.1: Simple PV diagram

There are 2 types of SCBs in the PV plant: **seasonal** and **tracker**. The two exhibit different axes of movement. For seasonal SCBs, the tilt angle of the panels can be adjusted freely, while for tracker SCBs, both the tilt and the orientation (rotation) of the panels can be calibrated.

To collect PV data, a monitoring component is added to the SCBs which measures performance-critical metrics such as power output, current, voltage etc. in real-time. For our plant, this data is recorded from every SCB in 5-minute intervals starting at 00:00 on Nov 1st 2020 and ending at 23:55 on April 30th 2021. The relevant fields from this dataset utilized by the thesis work are:

1. **meter_id** - ID of the SCB
2. **time_only** and **date_only** - time and date of the reading
3. **Pwr_DC_scb** - power output generated at the SCB
4. **capacity_scb** - total capacity (size) of all modules connected to SCB
5. **Hz_Irr_1_pyrano / Tilt_Irr_1_pyrano** - horizontal and tilted irradiance incident on pyranometer #1 located on-site
6. **Hz_Irr_2_pyrano / Tilt_Irr_2_pyrano** - horizontal and tilted irradiance incident on pyranometer #2 located on-site
7. **Module_Temp_pyrano** - temperature at pyranometer located on-site

In total, the dataset comprises of 5-minutely readings taken from 480 SCBs, with each SCB indicating performance of a different area/zone within the power plant. Figure A.1 in Appendix A visualizes the distribution of SCBs across the plant site. As discussed in Section 2.2, this level of granularity within the data, ie. SCB-level as opposed to site-wide measurements, is ideal as it allows us to analyze spatial variability in soiling losses across the power plant and schedule area-specific optimal cleanings.

Note also that the irradiance and temperature measurements are not taken at the SCBs, but rather by 2 different on-site pyranometers.

Along with the PV data, cleaning schedule information is also provided, indicating cleaning dates and methods for the SCBs (excerpt shared in Appendix B). The relevant fields from this dataset are:

1. **date** - date of cleaning
2. **meter_tag** - identifying the SCB which was cleaned
3. **cleaning_type** - one of three values:
   (a) 'water cleaning' = manual cleaning with water
   (b) 'dry cleaning' = manual cleaning without water
   (c) 'rain cleaning' = heavy enough rain occurred that manual cleaning was not required
4. **rainy_day** - one of three values:
   (a) -1 = some rain occurred, but not enough to clean the panel
   (b) 0 = no rain
   (c) +1 = enough rain occurred to clean panel (may not be as clean as a manual cleaning)

3.2 Step 1 - Tracking Soiling Losses

Following the research approach from Section 1.3.1 (Figure 1.5), the first step towards an optimal cleaning schedule is to track soiling losses. However, simply measuring losses in the power output is not sufficient. This is because the power output of an SCB can vary depending on non-soiling related factors, particularly the irradiance input on the panels. For example, a soiled module on a sunny day will generate more power than a clean module on a dark day. A better indicator of PV performance is the SCB *efficiency*, ie. the percentage of incoming irradiance that is converted into power. For a clean panel, we expect the conversion efficiency to be at a maximum, and as soiling begins to deposit, we expect the efficiency to gradually decline.

Therefore, using the available readings from SCBs, we first calculate day-to-day efficiency values, with the end goal for this step being to develop an **efficiency timeline** for each SCB. Then, by examining the timeline, we can extract key pieces of information such as how fast the SCB is losing efficiency during the soiling intervals (ie. the soiling losses/rate), when cleaning events occur and how much production gain they provide (ie. the cleaning benefits), and which SCBs are being impacted the most at the moment (optimal targets for cleaning). Eventually, this information will be used as input to optimize cleaning schedules of the SCBs.

3.3 Efficiency Calculation

The basic efficiency calculation involves determining the ratio of irradiance converted to power output. However, this ratio alone is not a complete measure of performance. There are other non-soiling related factors, such as temperature and capacity, which also affect the power output and thereby the efficiency. If we are to use changes in efficiency as an indicator of soiling loss and cleaning gains, we must ensure that the only parameter affecting efficiency is soiling. Thus, the extra factors should be considered and controlled.

Temperature

It is well known that temperature affects the performance of PV systems [24]. The PV power temperature coefficient of the module, provided by the manufacturers, indicates how strongly the power output depends on the module temperature. It is a negative number since the power production decreases with increasing temperature [15]. For the PV modules in our solar plant, the standard operating temperature is $24°C$ and the temperature coefficient $\alpha = -0.25\%°C^{-1}$, indicating that for every $°C$ above $24°C$ the power output of the modules will decrease by 0.25%. Since temperature will vary hour-to-hour and day-to-day, each power reading will need to be temperature corrected before being compared with the irradiance, in order for an accurate calculation of efficiency.

This correction is shown below, with $P_{measured}$ = power reading, T = temperature reading, and $P_{T-corrected}$ = temperature-corrected power output.

$$P_{T-corrected} = \frac{P_{measured}}{1 - \frac{\alpha}{100}(T - 24)} \tag{3.1}$$

Capacity

Another factor affecting the power output is the capacity of the SCB, ie. the size of the strings and modules connected to the SCB. For larger SCBs connected to more modules, the power output will also be higher. With the power plant receiving similar amounts of irradiance, a naive efficiency calculation will incorrectly record higher efficiencies for higher capacity SCBs. Since we will be comparing efficiency loss across different SCBs, this problem will directly impact results. Therefore, each power reading will be capacity corrected, before being compared with the irradiance, in order for a fair and accurate calculation of efficiency.

This correction is shown below, with $P_{measured}$ = power reading, C = SCB capacity, and $P_{C-corrected}$ = capacity-corrected power output.

$$P_{C-corrected} = \frac{P_{measured}}{C} \tag{3.2}$$

Overall Calculation

The overall efficiency calculation, with both temperature and capacity controlled, is shown below, with E = calculated efficiency and I = irradiance reading.

$$E = \frac{P_{measured}}{I \cdot [1 - \frac{\alpha}{100}(T - 24)] \cdot C} \tag{3.3}$$

We now have a formula to convert each 5-minutely reading, using the recorded power, irradiance, temp, and capacity, into a performance metric - ie. efficiency. Since we have corrected for all possible factors affecting the power output, except for soiling, this efficiency should be a true indicator of PV performance and any shifts in its value can be safely attributed to soiling loss alone.

What remains now is to aggregate these 5-minutely efficiencies into daily efficiencies, creating a PV performance timeline. This is done in Section 4.1.2.

3.4 Inconsistencies in Data

The efficiency calculation formulated in the previous section is still missing one important factor/parameter: noise! Despite correcting for all theoretical factors affecting power output, anomalies in the data will still pose issues for accurate efficiency calculation. Any noise in the measured signals can cause power output to be recorded incorrectly, skewing the calculated efficiency and possibly resulting in incorrect soiling losses or cleaning events being detected. Fortunately, noise encountered in PV datasets tends to follow recognizable patterns correlating to real-world conditions. In this section, we aim to identify inconsistencies and determine their root causes, before presenting a complete data-cleansing methodology in Section 4.1 to address them.

Identification

While working with the efficiency calculations and verifying the results, we noticed several inconsistencies in the recorded PV data such as large power readings for low irradiance measurements and vice versa. One instance is shown in Figure 3.2. Notice that when comparing readings from times 13-40 to 13-45, we see a decrease in measured irradiance (Hz_Irr_1_pyrano) as well as an increase in measured temperature (Module_Temp_pyrano).

Theoretically, this should result in a decrease in power production, but instead we observed an increase in recorded power output (Pwr_DC_scb). This resulted in an abnormally large efficiency value and a noticeable positive shift in the aggregated efficiency for the day compared to surrounding days, which will skew soiling loss calculations and hinder cleaning event detection.

	Hz_Irr_1_pyrano	Module_Temp_pyrano	meter_id	Pwr_DC_scb	capacity_scb	time_only	date_only
4063897	815.900	46.335	454	132.380	166.4	13-00	2020-12-09
4064294	804.418	46.374	454	132.766	166.4	13-05	2020-12-09
4064691	704.763	45.281	454	123.564	166.4	13-10	2020-12-09
4065088	661.768	45.677	454	112.237	166.4	13-15	2020-12-09
4065485	736.390	47.516	454	129.122	166.4	13-20	2020-12-09
4065882	679.747	47.270	454	129.426	166.4	13-25	2020-12-09
4066279	630.571	48.904	454	133.508	166.4	13-30	2020-12-09
4066676	692.085	47.690	454	124.330	166.4	13-35	2020-12-09
4067073	657.754	45.591	454	128.070	166.4	13-40	2020-12-09
4067470	472.214	46.038	454	135.712	166.4	13-45	2020-12-09
4067867	392.855	43.958	454	85.629	166.4	13-50	2020-12-09
4068264	477.701	43.287	454	103.273	166.4	13-55	2020-12-09
4068661	656.666	42.660	454	125.934	166.4	14-00	2020-12-09

Figure 3.2: Contradicting values of relative power and irradiance

Attributing this to some sort of glitch in the data measurement, we decided to investigate further by plotting daily irradiance and power readings together to examine their correlation. While not all days' plots reflected these inconsistencies, one particularly noisy plot is shown in Figure 3.3.

Notice that once again, there are increases and decreases in power output (spikes and dips in the plot) without corresponding shifts in the irradiance measurement. However, with the plot it's obvious that the major issue isn't the lack of the corresponding shifts, since they do seem to be present, but rather their position in relation to the power output updates. While ideally, power and irradiance should mirror one another very closely, here, it appears that the spikes and dips in one value appear slightly before or after the other's.

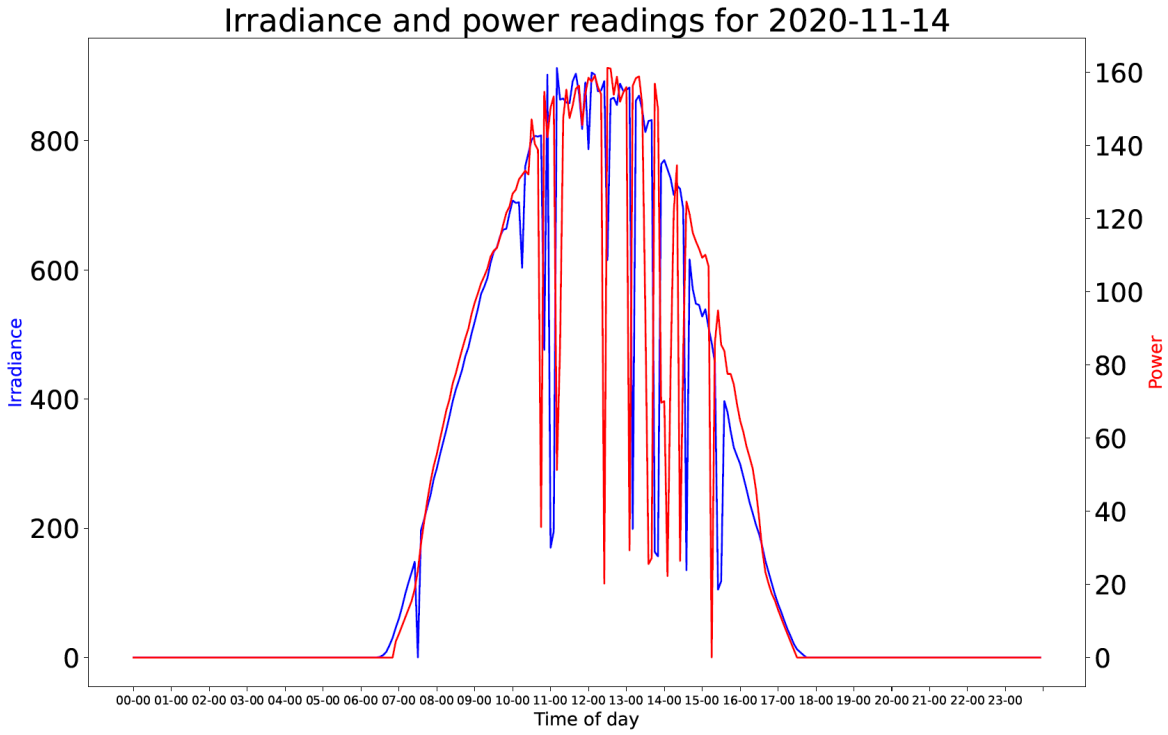


Figure 3.3: Noisy irradiance and power readings for Nov 14, 2020

This suggested a possible time-shift synchronization issue between the pyranometer's irradiance readings and the SCB's power readings.

Root-Cause

While a calibration issue between the pyranometer and SCB seems likely, let's revisit the observations from the data and plots:

- An increase/decrease in either irradiance or power, without a corresponding update in the other
- An increase/decrease in either irradiance or power with a delayed update in the other
- Only some days exhibited these inconsistencies while others seemed unaffected

With these observations in mind, we determined the root cause of the noise to be micro-weather conditions ie. uneven cloud cover causing uneven shadowing across areas of

the PV plant. Specifically, some parts of the plant may be overcast by clouds while other parts see clear skies. Remember that the irradiance measured is not incident on the SCB modules, instead it is measured by a pyranometer in a different location of the plant.

As such, the pyranometer may be shadowed, lowering measured irradiance, while the SCB is not, leaving measured power unaffected, and vice versa (Figure 3.4). This would lead to the conflicting recordings of irradiance and power output we observed. This also explains why some days (being cloudy) showcased these inconsistencies while other days (being clear) did not. Further, cloud cover could pass over the pyranometer and SCB at different times, shading one before the other and vice versa, which explains the misaligned/delayed patterns in the irradiance and power plots.

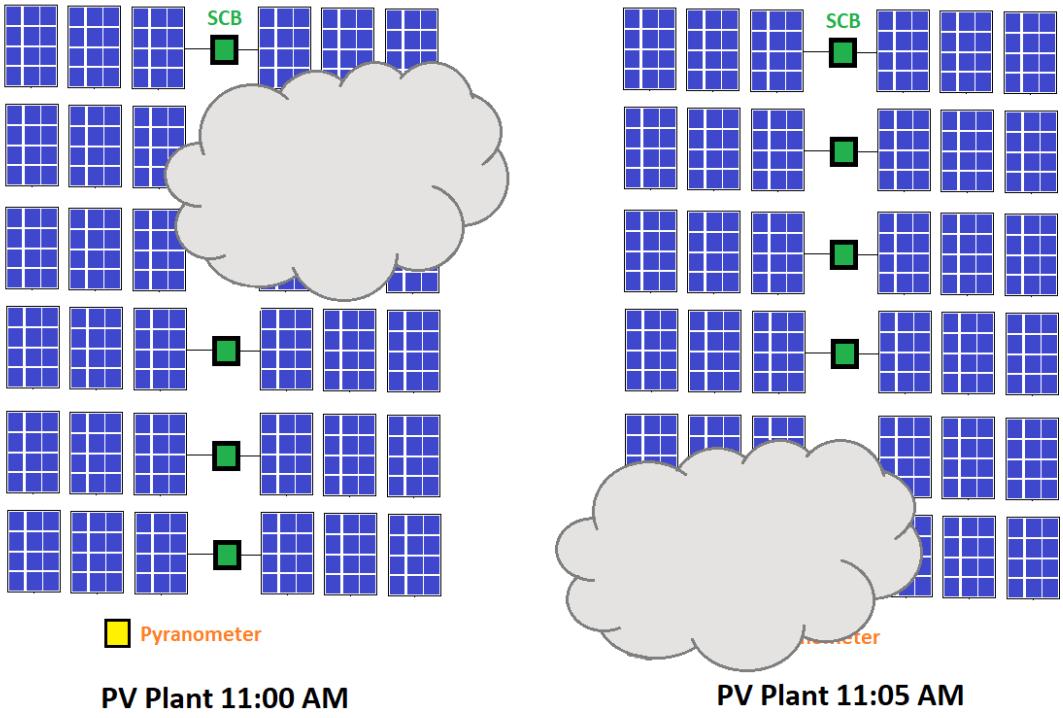


Figure 3.4: Micro-weather conditions relating to cloud cover causing uneven shading of PV plant

In comparison, a possible time-shift synchronization issue between the pyranometer and the SCB recordings cannot explain the lack of noise on some days vs others, since a

time-shift in the values should consistently affect all readings. It also cannot explain the changes in irradiance/power without a corresponding change in the other, as a time-shift should just delay the corresponding update rather than remove it completely.

How to Address This

A clean efficiency timeline, free of noise/anomalies due to micro-weather conditions, is necessary if we are to extract meaningful and accurate results. Data points affected by this issue must be systematically detected and either adjusted or excluded to mitigate their impact on efficiency calculations.

We present a complete data-cleansing methodology in Section 4.1, which addresses the major anomalies caused by uneven cloud cover as well as other more minor issues.

Chapter 4

Methodology

4.1 Data-Cleansing Methodology

Here, we present several data-cleansing techniques applied to the PV dataset to help address issues caused by noise/anomalies and generate clean efficiency timelines.

4.1.1 Data Filters

The first method we employed was to restrict the data used for efficiency calculations. Two different types of data filters, listed below, were applied to the dataset readings.

1. **Time of Day** - Restricted to readings taken between **10 am - 2 pm** to filter for peak power production, avoiding early morning and late evening conditions which are particularly foggy and cloudy, and instead focusing on hours receiving direct sunlight without diffraction/diffusion [8].
2. **Power $>$ 0 and Irradiance $>$ 600** - To avoid any glitches with readings where pyranometer was out-of-order to re-calibrate or SCB was not recording any power, ie. operational errors; also focusing on readings with peak irradiance levels

This ensures that the remaining data is free of any obvious operational errors, before we move on to addressing more complicated issues in the following sections.

4.1.2 Cumulative Efficiency Ratios

In Section 3.3, we calculated 5-minutely efficiency values. To plot a day-to-day efficiency timeline, we need to aggregate these values into daily efficiencies. However, with the micro-weather conditions skewing calculations, the aggregation must avoid the influence of erroneous efficiencies on the daily result.

Initial Attempts

- Mean - Calculating daily efficiency by averaging all 5-minutely efficiency values for the day
- Median - Calculating daily efficiency by taking the median of all 5-minutely efficiency values for the day
- Smoothing - Before calculating the daily mean or median, adjust/smooth each 5-minutely efficiency value using adjacent readings to account for sudden bursts of cloud cover only in effect for small periods of time

Final Version

While the initial attempts displayed varying degrees of effectiveness, the most successful attempt at generating clean efficiency timelines with daily aggregated values was the calculation of cumulative efficiency ratios.

Rather than calculate 5-minutely efficiency from individual 5-minutely power readings and then aggregate into a daily result, we instead calculated the daily efficiency directly from the cumulative power and irradiance readings for the day. This formulation of daily efficiency is shown below.

For each 5-minutely reading, we first calculate the temperature and capacity corrected power output ($P_{corrected}$) using Equations 3.1 and 3.2 as explained in Section 3.3. Then we calculate daily efficiency (E_{day}) by determining the cumulative ratio between corrected power and irradiance for the day.

$$E_{day} = \frac{\sum_{day} P_{corrected}}{\sum_{day} I} \tag{4.1}$$

In calculating daily efficiency using the total power output and total irradiance input during the day, we minimize the impact of outliers on the final result. With cumulative

ratios, outlier readings will be combined with all other readings for the day before being used in efficiency calculations, thereby limiting their influence on the final result.

This method also has the added advantage of accounting for the delayed updates in power vs irradiance due to cloud cover (Figure 3.4). For instance, even if there is a delay in corresponding spikes, both will be included in the day's summations for power and irradiance, and thus reflected in the final efficiency calculation.

4.1.3 Cloudy Day Detection

While cumulative efficiency ratios certainly minimize the impact of noisy readings, on some particularly cloudy days this impact cannot be avoided and the efficiency calculations are heavily distorted. We must look to exclude these days completely from our day-to-day efficiency timelines, else risk miscalculations of soiling loss and incorrect cleaning event detection.

Detecting cloudy days with high noise impact is not a trivial matter. For example, visually, the cloudy days are obvious by looking at plots of power and irradiance readings and identifying days with jagged curves indicating shading as opposed to smooth curves indicating clear skies. What is not so obvious is whether a cloudy day will produce noisy readings. It is possible, on a cloudy day, for the pyranometer and modules of an SCB to be equally shaded, meaning every dip in power output will be mirrored by a dip in irradiance. Readings such as these will not result in noisy calculations and must not be excluded even though they are produced on a cloudy day.

Correlation Coefficient

With this line of reasoning, we look for a metric not simply to detect cloudy days but rather to determine the **impact** of the cloudiness on the correlation between power and irradiance ie. the quality of the readings. This is exactly the purpose of the correlation coefficient.

The daily correlation coefficients between power and irradiance, bounded between -1 and 1, should give an accurate measure of the noise impact of clouds on distorting the power and irradiance readings. By using the correlation coefficient as our metric, we calculate the level of similarity and correspondence/alignment between the power and irradiance curves, which is directly affected by micro-weather conditions being different over the SCB's location vs over the pyranometer's location.

The correlation coefficients for a clear day with minimal noise vs a cloudy day with high noise impact are shown in Figures 4.1 and 4.2.

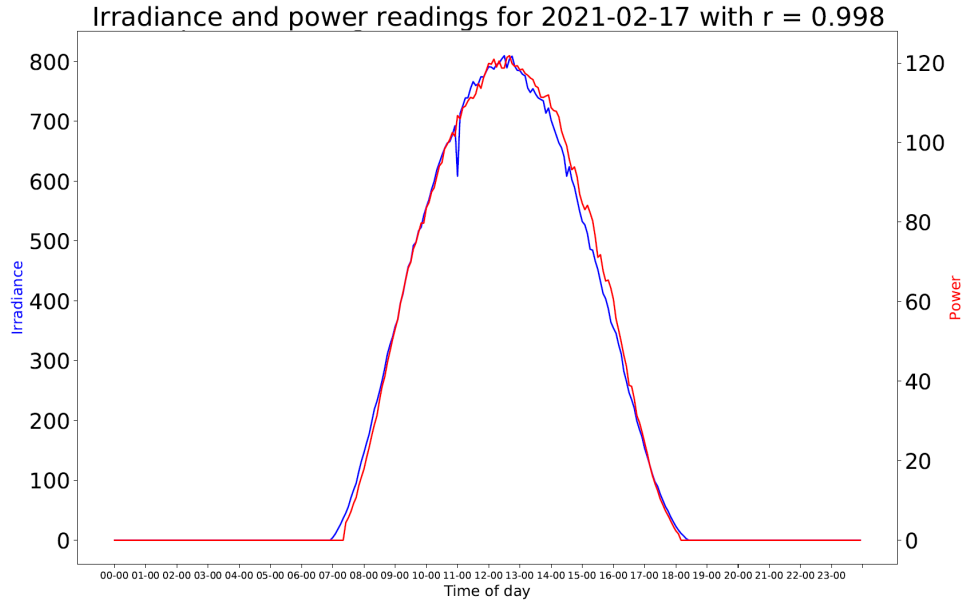


Figure 4.1: Clear day with good correlation coefficient = 0.998

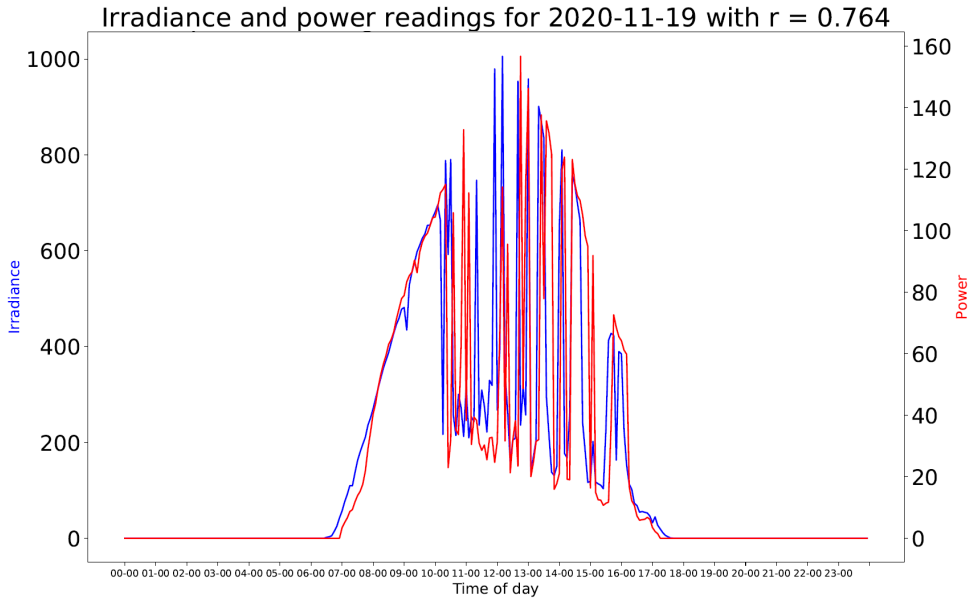


Figure 4.2: Cloudy day with poor correlation coefficient = 0.764

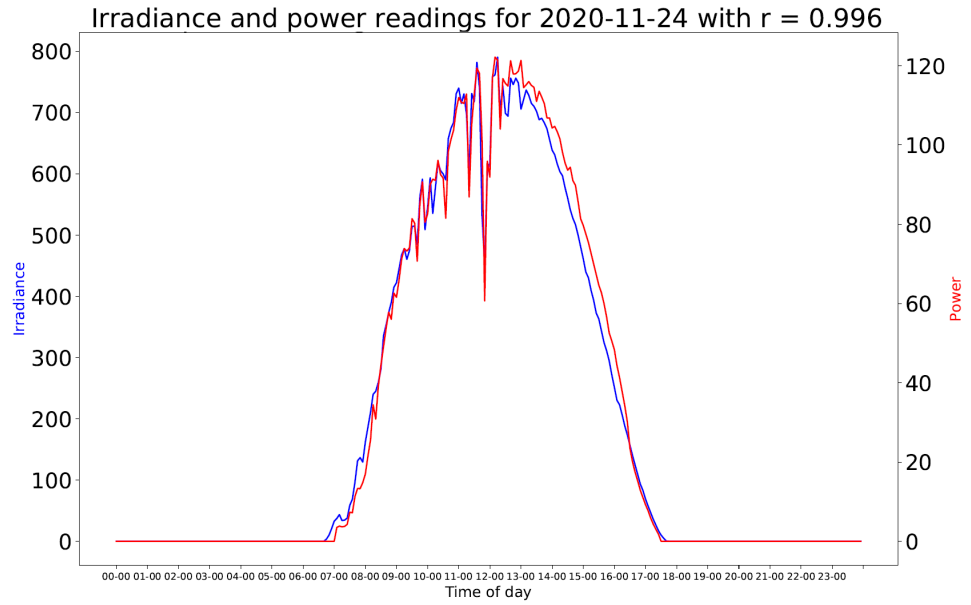


Figure 4.3: Partially cloudy day with good correlation coefficient = 0.996

In comparison, Figure 4.3 shows the correlation for a partially cloudy day. Note here that even if there is some jaggedness in the curves, indicating cloud cover, as long as the power values are aligned nicely with the irradiance changes (ie. the cloud cover affects both the SCB and pyranometer locations similarly), then the correlation coefficient is still high. This is exactly what we required. Since readings from this partially cloudy day will not distort the efficiency calculations despite the cloud cover, they should not be excluded and that is exactly what the correlation coefficient reflects.

Threshold

Having decided that the correlation coefficient is a suitable metric for determining noise impact, we must now decide on a threshold for excluding particularly noisy days. In principle, a correlation coefficient score above 0.7 or 0.8 typically indicates strong correlation. In practice, examining the scores from the dataset we found the majority of correlation coefficients falling between 0.85 - 1. Therefore, we decided on a cutoff of **0.8**, excluding data from days with correlation coefficients below this score. This proved to be effective in limiting the impact of noise.

Note that different sets of cloudy days are excluded for each SCB. This is because SCBs experience different levels of cloudiness and noise impact on power readings, depending on their location relative to the pyranometer.

4.1.4 Best-Fit Pyranometer Measurements

Motivation

Recall from Section 3.1 that we have irradiance measurements recorded from 2 pyranometers at different locations on the plant site, with a total of 4 irradiance readings listed below:

- **Hz_Irr_1_pyrano** - horizontal irradiance incident on pyranometer #1
- **Tilt_Irr_1_pyrano** - tilted irradiance incident on pyranometer #1
- **Hz_Irr_2_pyrano** - horizontal irradiance incident on pyranometer #2
- **Tilt_Irr_2_pyrano** - tilted irradiance incident on pyranometer #2

Given that the noisy readings result from varying levels of cloud cover over the SCB modules vs over the pyranometer, it may be possible to exploit the multiple irradiance readings.

Since the 2 pyranometers are in different locations, the irradiance reading from one pyranometer may be a better fit for some SCBs' power readings. For example, power readings from SCBs closer to pyranometer #1 will likely exhibit better correlation with its irradiance readings compared to the readings from the further away pyranometer #2. This is because cloud cover will affect both pyranometer #1 and its nearby SCBs similarly.

To verify this theory, we plotted power-irradiance curves for SCB 454 on Nov 13, 2020 using different pyranometers. The two plots are shown below in Figures 4.4 and 4.5. Clearly, SCB 454 exhibits better correlation (score of 0.866 vs 0.986) with pyranometer #2's readings.

In addition to the relative location between an SCB and the pyranometers, the SCB type in comparison to the type of the irradiance reading can also be a factor. Recall from Section 3.1 that there are 2 types of SCBs in the PV plant: seasonal and tracker, with different axes of movement regarding tilt and orientation. Depending on the SCB type, certain irradiance measurements (Tilted vs Horizontal) from different pyranometers could exhibit better correlation. Therefore, a data-driven analysis to determine the best-fitting measurement for each SCB is certainly warranted.

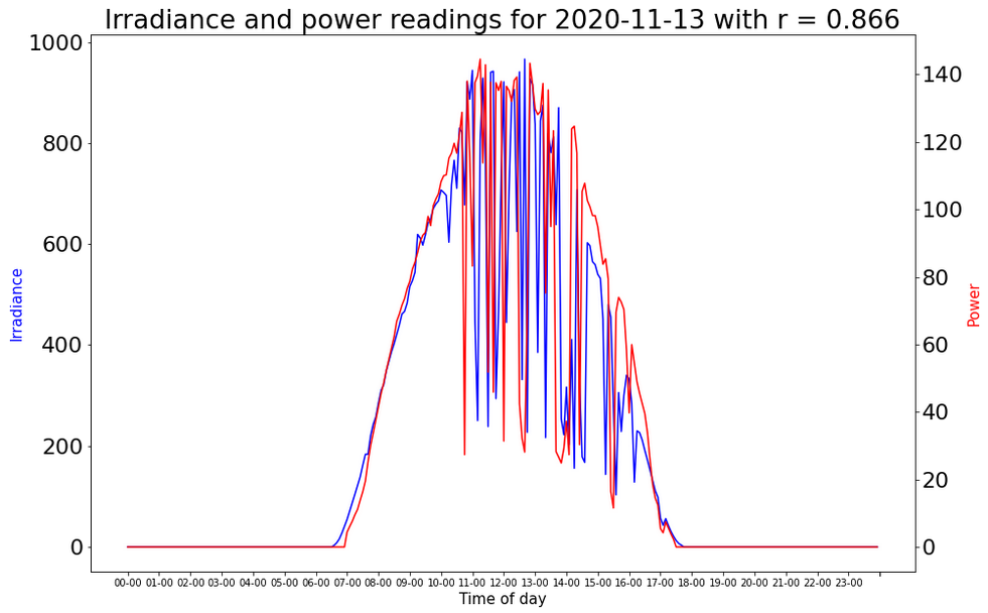


Figure 4.4: SCB 454 Power Output with `Hz_Irr_1_pyrano` Irradiance Readings

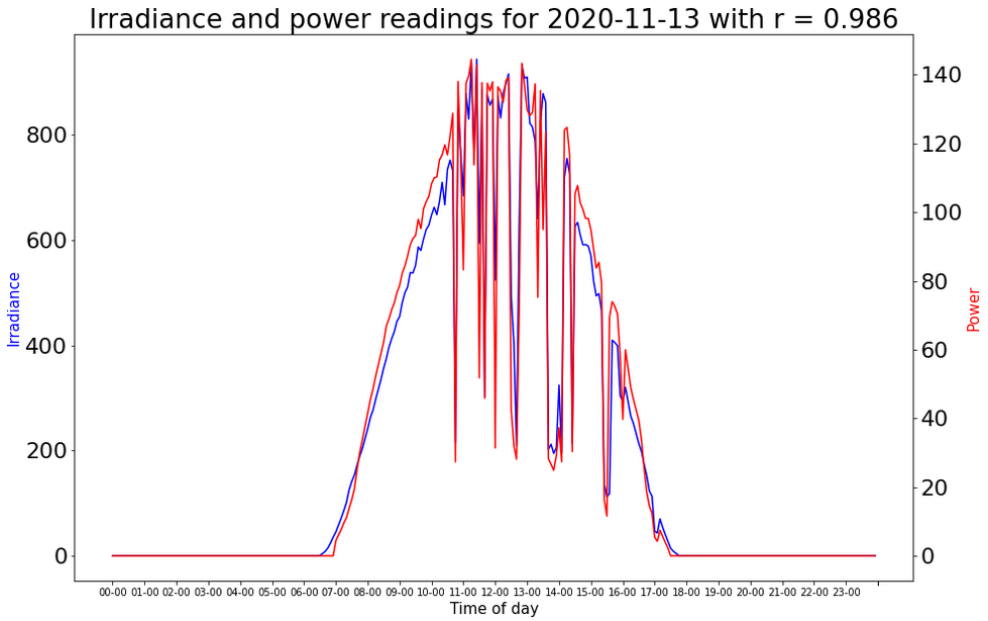


Figure 4.5: SCB 454 Power Output with `Hz_Irr_2_pyrano` Irradiance Readings

Approach

Ideally, we want to analyze the daily correlation between an SCB's power readings and each of the 4 irradiance readings, such that for each SCB, on each day, the best-suited irradiance reading is used for efficiency calculations. However, alternating between different irradiance measurements for different days in an SCB's efficiency timeline will introduce variance, modifying the efficiency calculations each time a different pyranometer is chosen.

Instead, we decided to evaluate each irradiance reading for best fit to an SCB's power data over the entire timeline. For each SCB, the single best-fitting irradiance reading is then used to calculate efficiency for all days.

Implementation (Metrics)

To evaluate the different irradiance measurements, we experimented with several metrics listed below.

- Root Mean Square Error (RMSE) between normalized SCB powers and irradiance readings
- Wasserstein (Earth Mover) Distance between normalized SCB powers and irradiance readings
- Correlation coefficient metrics - Calculating daily scores between SCB powers and irradiance readings, then:
  - Choosing the irradiance measurement which correlates best on majority of days
  - Choosing the irradiance measurement which has a score of at least 0.8 for the most number of days (ie. contributes the most clean data to efficiency timeline)
  - Choosing the irradiance measurement with the highest average daily score

Each of these metrics was tested and found to produce similar distributions of best-fitting irradiance measurements. Thus, which metric to use is not too important. Rather the key insight here is that we need to consider the different available irradiance measurements, instead of naively using one.

For this thesis, the 4 irradiance readings were evaluated using the correlation coefficient metric which chooses the measurement producing a score of at least 0.8 for the most number of days. Different SCBs fitted best with different irradiance readings, although `Hz_Irr_2_pyrano` and `Tilt_Irr_2_pyrano` were the best fit for the majority of SCBs.

Note that once the best-fitting irradiance reading for an SCB is determined, it is the one used for cloudy-day detection, as well as efficiency calculations, to produce a clean efficiency timeline.

4.1.5 Summary

Combining the efficiency calculations in Section 3.3 and the techniques discussed above, we form a **complete data-cleansing methodology**.

Applied to any PV dataset, it can systematically identify and resolve non-soiling related noise impacting data-driven calculations. Further, insights from the encountered micro-weather conditions, as well as the techniques developed to address them, will prove useful for plant operators aiming to analyze their own datasets.

In the thesis, this data-cleansing methodology was applied to generate clean efficiency timelines for each SCB (see Section 5.2 for an example), successfully completing the first step of our research approach (Section 1.3.1 and Figure 1.5).

4.2 Human-Recorded Cleaning Logs

4.2.1 Step 2 - Segmenting the Efficiency Timelines

The second step towards an optimal cleaning schedule is to segment the efficiency timelines into cleaning and soiling intervals by determining when cleaning events occurred. Once we have a list of cleaning dates for each SCB, their efficiency timelines can be divided into periods of soiling, where efficiency deteriorates, and periods of cleaning, where efficiency improves. This extracts both soiling losses and cleaning benefits from the timelines, which are used to optimize the cleaning schedules of each SCB.

Fortunately, along with the PV dataset for the solar plant, we are also provided with cleaning logs, indicating cleaning dates and methods for the SCBs, as detailed in Section 3.1. Unfortunately, any human-recorded input cannot be 100% reliable. In this section, we explore the cleaning log dataset and analyze its reliability. In the next section, we study an alternative method for cleaning event detection driven by the PV data itself.

4.2.2 Unreliable Recording of Cleaning Events

Using the provided cleaning logs, we divided the efficiency timelines of each SCB into multiple soiling intervals. Analyzing the efficiency over the course of these soiling intervals we noted various instances of unexpected behaviour, indicating the unreliability of the cleaning logs.

Findings

- Cleaning event recorded for an SCB on a given date, but no noticeable efficiency improvement in timeline
- Noticeable efficiency improvements in timeline, but no cleaning event recorded for the SCB during those time periods

Overall, the cleaning logs seemed riddled with **false positives** and **false negatives** as indicated by the above observations. All findings pointed towards the unreliable recording of cleaning events by the cleaning crew.

Statistical Analysis

In light of these observations, we decided to investigate further. Looking at all records, we performed a statistical analysis to evaluate the accuracy of the cleaning logs. Specifically, we tracked the fraction of recorded cleanings actually resulting in a corresponding increase in the efficiency timeline, as well as the total possible missed cleaning events.

Tallying the cleaning logs across all SCBs, we found that 75% of the recorded cleanings do correspond to efficiency increases in the timeline (true positives), while 25% do not (false positives). As well, there were numerous instances identified where efficiency improved without any record of a cleaning event (false negatives). In many of these cases, recorded cleanings were instead found in the logs of neighbouring SCBs. This suggests that the crew cleaned a group of SCBs, while only recording the cleaning in one SCB's log.

Regardless of the exact numbers, these findings and subsequent investigations confirm the poor quality of the cleaning logs. Whether due to a lack of attention to detail from the cleaning crew or an error in data recording, it is obvious that the human-recorded cleaning logs are unreliable, with incomplete, missed, or blatantly incorrect records.

4.3 Data-Driven Segmentation Algorithm

4.3.1 Motivation

There are two drawbacks of using human-recorded cleaning logs:

- Reliability - Mistakes and oversights from the cleaning crew will result in unreliable cleaning logs with incorrect and missed recordings.
- Availability - With some PV plants the logs may not even be maintained in the first place!

If we continue as is, using the cleaning logs to determine cleaning events, the timelines' division into soiling and cleaning intervals would be inaccurate. This inaccuracy would propagate to the downstream calculations of soiling loss and cleaning gains, resulting in a sub-optimal cleaning schedule and lost profit for the plant owners and operators. Therefore, an accurate segmentation of the efficiency timelines, without dependence on provided cleaning logs, is critical to our research problem.

Visually inspecting the patterns in the timelines, it's not too difficult to determine when cleanings must have occurred. As such, a data-driven algorithm should be able to analyze trends in efficiency and detect cleaning events.

Ultimately, the PV data is the only source of truth, revealing the complete cleaning schedule of each SCB as patterns in its efficiency timeline. So rather than using human-recorded cleaning logs as a proxy, we develop an algorithm capable of recognizing these patterns directly from the PV data itself.

4.3.2 Initial Attempts

We experimented with two initial data-driven approaches:

- Efficiency Thresholding - Identifying efficiency increases in the timeline greater than a threshold value and marking them as cleaning events
- *Change Point Detection* (CPD) Algorithm - Iterating through efficiency timeline and at each data point predicting the next data point's value using the slope (Theil-Sen Estimator) of the past points
  - If the difference between the predicted value and the actual value is in the upwards direction and greater than some threshold, mark as cleaning event (change point)

Limitations

While these techniques generated reasonable cleaning schedules for the most part, the cleanings detected weren't always accurate. For instance, both methods performed poorly in many edge cases such as **gradual or back-to-back cleanings** ie. cleanings performed for an SCB over multiple consecutive days (perhaps due to extended rainfall). In these cases, cleanings either were not detected, because the overall cleaning efficiency increase was broken into multiple smaller increases which avoided detection from the threshold (see

Figure D.1 in Appendix D for an example), or if they were detected, cleaning was only marked for the first day, rather than the entire cleaning period.

Further, each of these methods requires **threshold calibration** which is not practical. With the various cleaning methods for different PV plants, it is impossible for one threshold to apply correctly for all cleanings, even if the threshold is tuned for site-specific data. If the threshold value is too small, **outlier data or noise** may be identified incorrectly as cleanings. On the other hand, if the value is too large, some **cleanings will be missed**.

However, while these approaches alone did not work as needed, together their underlying ideas contributed towards development of the final data-driven segmentation algorithm.

4.3.3 Final Version

Description

The final version involves segmentation of the efficiency timeline into generally up-sloping (ie. cleaning) and generally down-sloping (ie. soiling) intervals. Specifically, the algorithm iterates through the timeline, monitoring Theil-Sen slopes of a moving window of points. Whenever there is a change in direction of the slope (between the previous windows and the current window), a change-point is marked, indicating the beginning of a cleaning interval (if the slope changes from downwards to upwards) or the beginning of a soiling interval (if the slope changes from upwards to downwards).

Compared to the initial value-based threshold approaches (Section 4.3.2), this slope-based approach is much more robust. For example, instead of comparing deltas between individual efficiency values, we now examine directional changes in slope of multiple efficiency values. This means that calibrating specific threshold values to fit all data points is no longer needed, and only the monitoring of the slope's sign between positive and negative is required. Moreover, problematic outliers in the timeline are safely ignored, because while the delta between values will signal incorrectly, the direction of the Theil-Sen slope will remain unaffected (Figure 4.7). Instead of being thwarted by a single anomaly, our slope-based algorithm will only signal when multiple points begin trending in the other direction. Finally, the gradual back-to-back cleanings plaguing our initial approaches are also addressed, being grouped together into a cleaning interval as a generally up-sloping period (Figure 4.6).

Implementation Details

The algorithm requires 2 parameters:

- Window Size - This is the number of points to use when calculating slopes. We set this to 5 (based on trial and error) which was adequate in identifying gradual cleanings while also detecting 1-day cleanings.
- Choice of Slope Calculation - Between the Theil-Sen Estimator and Linear Regression best-fit, Theil-Sen slopes were more effective in the presence of outliers and thus chosen as the method for slope calculation.

To segment an efficiency timeline, the algorithm takes the following steps (code shown in Appendix C):

1. Determine the initial slope direction for the 1st window of points beginning from the 1st data point, labelling it as the previous slope direction
2. Move to the 2nd data point
3. Determine the slope direction in the current window of points, beginning from the current data point
4. If the previous and current slopes have different directions:
   (a) Mark the minimum point in the current window (if the slope changes from downwards to upwards) or the maximum point in the current window (if the slope changes from upwards to downwards) as a change-point
   (b) Skip to the change-point and set the previous slope direction to the current slope direction

   Else continue to the next point
5. Repeat from Step 3

Figure 4.6 shows the general detection of change-points under a typical soiling-cleaning-soiling efficiency pattern observed in the timelines.

Note that from the 1st data point, the slope is calculated to have a downwards direction (soiling). This remains the case for the subsequent windows of points, until *Window 1* as highlighted in the diagram. For this window of points, labelled in red from 1-5, the Theil-Sen slope changes direction from downwards to upwards. The minimum point in this window, labelled in green, is marked as a change-point and the current slope direction becomes upwards (beginning of cleaning). The slope remains upwards until *Window 2* as

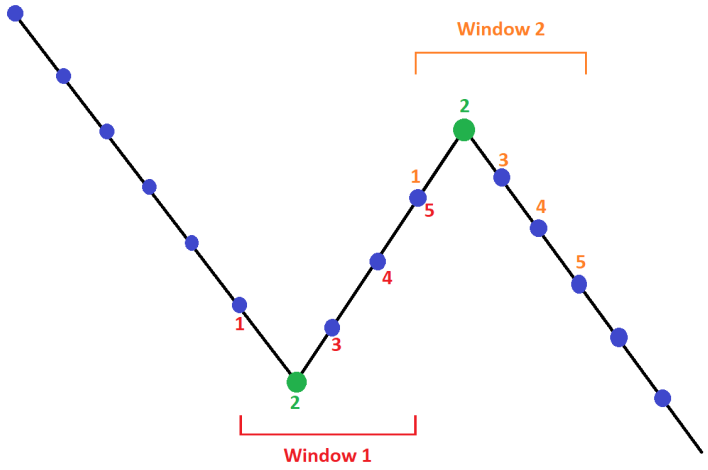


Figure 4.6: Algorithm behaviour for a typical efficiency pattern indicating cleaning interval

highlighted in the diagram. For this window of points, labelled in orange from 1-5, the Theil-Sen slope changes direction from upwards back to downwards. A change-point is marked at the maximum point in this window, labelled in green, and the current slope becomes downwards (back to soiling), remaining so for the rest of the timeline.

Notice also that marking change-points as the minimum or maximum points in the window allows us to track the exact end-points of the soiling and cleaning intervals in cases where a change in slope may be detected before the peak or trough points are reached.

Continuing, Figure 4.7 showcases a common outlier amongst an otherwise normal soiling timeline.

Naive value-based threshold methods, which compare differences between individual data points, are likely to mark this jump in efficiency as a cleaning event. Smarter value-based techniques, such as the existing soiling extraction algorithms discussed in Section 2.4.3, may ignore this outlier but mis-characterize other outliers where the value of the threshold is not perfect. In contrast, since the direction of the Theil-Sen slope remains downwards throughout the timeline, despite the spike in efficiency our algorithm will not prematurely mark a cleaning event. Of course, if the increase in efficiency is persistent with multiple high-efficiency points following the initial spike, the slope will change to upwards, and the algorithm will mark a cleaning event. As such, when compared to the change in value, the change in direction of slope is much more reliable and robust.

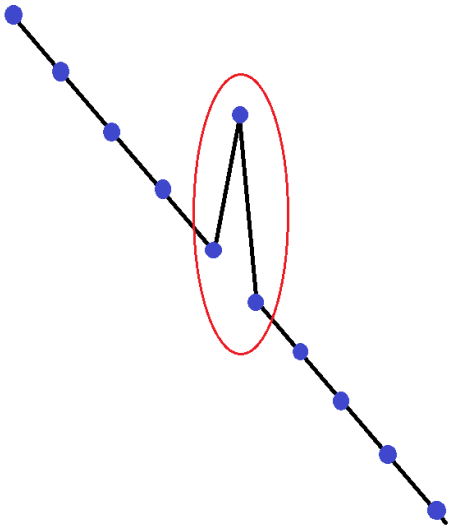


Figure 4.7: Erroneous spike in efficiency timeline, safely ignored by our algorithm

Summary

Figure 4.8 shows an example of a segmented efficiency timeline following our data-driven algorithm.

Notice that the identified cleaning (marked in green) and soiling (marked in red) intervals follow exactly from the visible efficiency patterns in the data. We also see that the algorithm is robust to minor ups-and-downs throughout the timeline, with generally up-sloping and generally down-sloping segments accurately detected.

We now have a completely data-driven method to detect cleaning events and segment the efficiency timeline generated in Step 1, without depending on the unreliable human-recorded cleaning logs. This successfully completes the second step of our research approach (Section 1.3.1 and Figure 1.5)

Figure 4.8: Segmented Efficiency Timeline Example

4.4 Step 3 - Cleaning Schedule Optimization

Given an accurate segmentation of each SCB's efficiency timelines, all that remains is extracting the soiling losses from the soiling intervals and the cleaning gains from the cleaning intervals. Once we have calculated these values for each SCB, we can then generate a profit function for optimization, which compares the cleaning cost with the cleaning benefit while accounting for any upcoming rainfall events.

In this section, we detail the smart-estimation of cleaning gains based on past cleaning intervals, develop a total cleaning benefit model considering rainfall events, and perform cost-benefit analysis to schedule cleanings which maximize profit.

4.4.1 Extraction of Key Information

For each SCB, by examining its segmented efficiency timeline, we extract information on soiling loss and cleaning gains.

- **Soiling Loss** - For each soiling interval in the timeline, we calculate the Theil-Sen slope of all the enclosed efficiency points. This slope is the soiling rate of the SCB for the time period defined by the soiling interval, representing the daily loss in efficiency due to soiling.
- **Cleaning Gains** - For each cleaning interval in the timeline, we calculate the increase in power production from the start-date to the end-date of the interval. This increase is the cleaning gain for the time period defined by the cleaning interval, representing the power production impact of cleaning.

Expected Cleaning Gain

The expected cleaning gain for an SCB is then defined as the average cleaning gain of all cleaning intervals in the past 60 days, accounting for any seasonal dependence. However, if the SCB is currently being cleaned, this latest cleaning interval is ignored to avoid smaller production gains due to unfinished cleaning.

Note that predicting future gains with 100% accuracy is not possible. Our method is still only an estimate, with limitations discussed in Chapter 6. However, since we consider the possibility of imperfect cleanings by the cleaning crew and judge the expected gain based on patterns in power production gains experienced from past cleanings, this is a much more knowledgeable estimate than existing methods, which assume cleaning gain = soiling loss.

Per-SCB Calculations

The soiling losses and expected cleaning gains are calculated for each SCB individually (Appendix A Figures A.2 and A.3), allowing for analysis into the spatial variability of soiling and area-specific optimal cleanings.

4.4.2 Cleaning Benefit Model

With the expected cleaning gains calculated for each SCB, there is still one more step before cost-benefit optimization - **total** cleaning benefit calculations.

Suppose we were to clean an SCB today. Not only would we observe power production gains today, but also for every day thereafter, when compared to the production of the same SCB had it not been cleaned. This benefit is visualized in Figure 4.9.

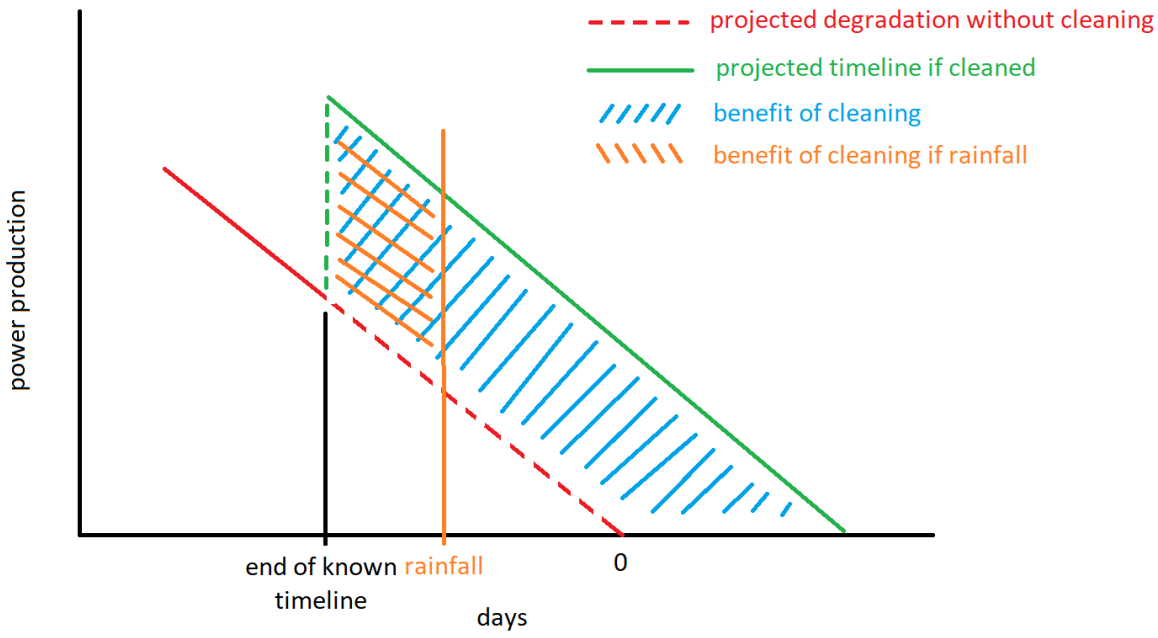


Figure 4.9: Cleaning Benefit Model

Without cleaning, the power production of an SCB is expected to degrade (shown in red) due to soiling. With cleaning, we expect to observe an increase in power production equal to the expected cleaning gain (dotted green line). After cleaning, the SCB will again be exposed to soiling and power production will degrade (solid green line) according to the soiling rate. However, the cleaned production timeline will maintain higher levels of production over the soiled timeline, with the difference being equal to the expected cleaning gain. As such, the **total benefit of cleaning** will be the **accumulated gains** in power production over the entire course of the timelines (shaded blue area), barring any other production-altering events.

Now, rainfall is one such production-altering event which will impact the total cleaning benefit. For simplicity, we assume a rainfall event will raise efficiency to the maximum power production, stopping the accumulation of cleaning gains. In this case, the total cleaning benefit will be given by the orange shaded area.

Of these two areas, we decided to use the area restricted by rainfall as the cleaning benefit, simply because accounting for rainfall is an important element in cleaning optimization. Practically, it is unreasonable to consider an *end* to these timelines with power production reaching 0 as shown in the diagram. Based on calculations from our dataset, we determined that with the current soiling rates, it would take over 200 days for the SCBs to stop producing power and so it is safe to assume a rainfall event will occur before then.

This cleaning benefit model does not directly incorporate the soiling loss (power production losses due to soiling), instead only using the cleaning gains (power production gains from cleaning). However, notice that the cleaning gain implicitly includes the soiling loss, with the size of the gain depending on how much production was lost originally due to soiling. The key insight here is that using the cleaning gain, rather than the soiling loss alone, allows us to also consider the impact of unreliable cleanings, while implicitly accounting for soiling loss.

4.4.3 Cleaning Profit Calculation

Profit Function

With a cleaning benefit model in place, we now formulate a profit function (Equation 4.2) defining the profit from cleaning today, based on a cost-benefit analysis.

$$\begin{aligned} \text{Profit} &= GDR - C \\ \text{where:}\quad G &= \text{expected kWh gain from cleaning} \\ D &= \text{number of days till rainfall} \\ R &= \text{PPA conversion rate from kWh to \$} \\ C &= \text{cleaning cost of any SCB} \end{aligned} \tag{4.2}$$

G represents the expected cleaning gain (in kWh) calculated by averaging power production gains from the past 60 days of cleaning intervals, as detailed in Section 4.4.1. The remaining parameters are all user input. D represents the expected number of days till rainfall, which is input by plant operators. This can be determined using meteorological data. R represents the conversion rate from kWh to \$ based on the *power purchase agreement* (PPA) agreed upon between the plant owners and the purchaser. This is value paid (ie. revenue earned) for every kWh of power produced. Finally, C represents the cleaning costs of an SCB. This can vary site-to-site dependent on cleaning methods, labour wages etc.

Details

While R and C are typically constant values, G and D can vary, taking on a range of values depending on SCB type, capacity, plant location etc.

The term of interest is GD, which represents the gain in power production summed across the days until rainfall ie. the total cleaning benefit in kWh when an SCB is cleaned vs not cleaned. With the other parameters being constant, this term dictates the profit from cleaning. The relationship between G and D, as well as their impact on profit is examined in Section 4.4.4. Figure 4.10 shows a visualization to help understand GD.

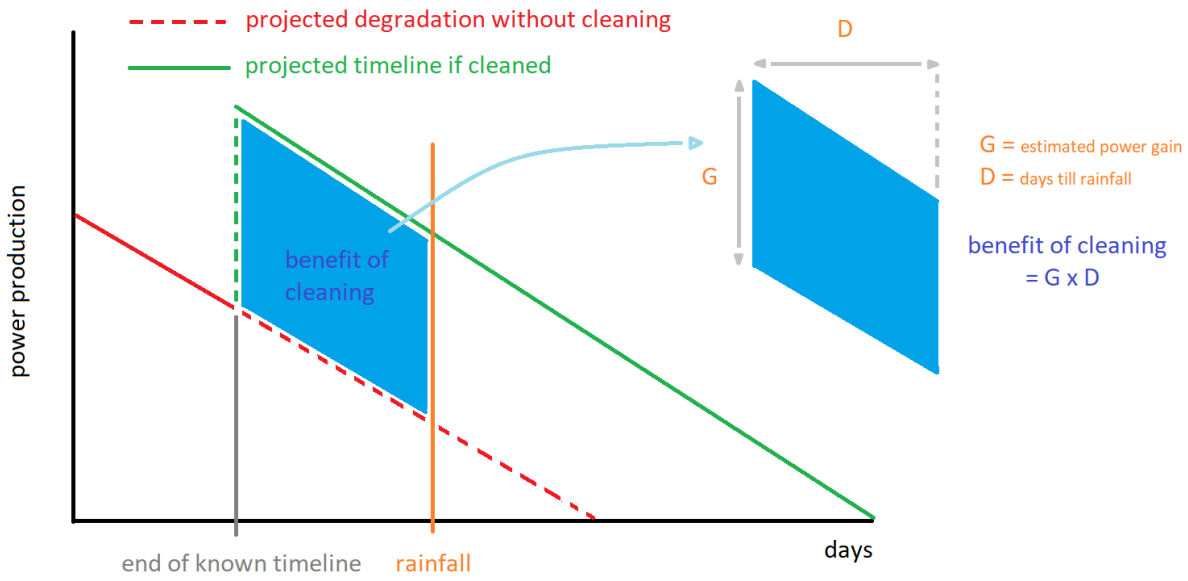


Figure 4.10: Cleaning Benefit Area Calculation

Note here that we assume cleaning due to a rainfall event will bring the SCB back to the maximum production level. With this in mind, we can simply sum the kWh gains until the rainfall event to estimate the total cleaning benefit (in kWh), GD.

Another important calculation is the estimation of the expected kWh gain G when the SCB is cleaned. As explained earlier, to determine its value we look back at previous cleanings and average their resulting gains. However, we cannot naiively apply this gain since there is a maximum limit to how much power an SCB can produce. To account for

this, we use past data to determine the maximum daily power production of the SCB and then calculate the expected kWh gain as detailed in Equation 4.3.

$$\begin{aligned} G &= \min\{\bar{g}, P_m - P_c\} \\ \text{where:} \quad G &= \text{expected kWh gain from cleaning} \\ \bar{g} &= \text{average kWh gain from past cleanings} \\ P_m &= \text{maximum power production} \\ P_c &= \text{current power production} \end{aligned} \tag{4.3}$$

This ensures G accounts for the maximum possible power production and does not exceed it.

Optimizing Today's Cleaning Decisions

With the profit function defined, optimizing cleaning decisions is relatively straightforward. Specifically, the following cases, regarding cleanings scheduled for today, are solved:

- For a given SCB, is it beneficial to clean today? - Yes, if the profit is positive, otherwise No
- If cleaning is to be performed today, which SCB should be prioritized? - Compare the profits from cleaning for all SCBs and choose the SCB with maximum profit
- Is it beneficial for all SCBs to be cleaned (site-wide cleaning) today? - Sum profits from cleaning for all SCBs. Some will lose money (for example if they've just been cleaned), while others are profitable. If the total profit is positive, then Yes else No

Further, since profit is calculated per-SCB, plant operators can gain more detailed insights on cleaning different areas of their plant. For instance, suppose a cleaning is scheduled for today, but the cleaning crew can only finish cleaning half of the plant - either the northern side or the southern side. In this case, plant operators can simply total profits from SCBs on each side and target the more profitable side. This level of insight would not be possible if we had calculated site-wide results.

Determining Optimal Cleaning Dates

As detailed, by using an SCB's current data we can calculate profit if the SCB were to be cleaned **today**, and perform **immediate** optimizations such as deciding whether to clean today, which SCB to target etc. However, looking ahead to the **future** and determining the

optimal cleaning date with maximum profit requires simulation of the SCB's performance and projection of profits from future cleanings.

We explore simulations of profit and its dependence on key parameters in Section 4.4.4. In Section 4.4.5 we apply the simulation to calculate optimal cleaning dates.

4.4.4 Simulations to Understand D and G

Description

Pending rainfall events complicate profit calculations, offering the possibility of cleaning at no cost. In these cases, it is unclear exactly when to clean or whether to clean at all. For example, suppose we've just cleaned an SCB and the next rainfall event is forecast to be 15 days away. Of course cleaning very soon thereafter would not be profitable as the SCB would not be sufficiently soiled. Similarly, cleaning too close to the date of rainfall would also be sub-optimal, as rain provides free cleaning. Perhaps exactly in the middle would work or perhaps it is optimal to not clean at all? It is intriguing to observe cleaning profit in these cases and determine the optimal action.

In this section, we simulate the above scenarios and examine how different parameters (D and G) affect profit. Specifically, we project the profit from cleaning on each day, from the day just after an SCB has been cleaned, till the day of rainfall. This allows us to observe the relationship between D and G and its impact on 1. the total cleaning benefit (GD) and 2. the profit.

The plot we generate should help plant operators visualize when there is no benefit from cleaning (sub-optimal), when there is some benefit from cleaning (fair), when that benefit is the largest (optimal), and how that benefit declines as we approach the rainfall event. In addition, by varying the days till rainfall, we can understand its effect on cleaning profit.

Calculations

Recalling Equation 4.2, the profit is directly proportional to D and G, with R and C being constants.

Now, consider an SCB which has just been cleaned, with a long period (D) until the next rainfall. Since D is at its maximum, we will be accumulating gains G over a long period of days and naturally we would expect the cleaning benefit area to be large. However, from Equation 4.3 we see that G will be very small since we do not expect a large power gain from cleaning with the SCB having just been cleaned and currently producing power at

near maximum levels (the minimum term in Equation 4.3 will be $P_m - P_c$). Thus, the total cleaning benefit, ie. the area GD, will be relatively small (Figure 4.11 Day 0), resulting in minimal, possibly negative profit with the cleaning costs outweighing the benefit.

As days pass and power production P_c decreases day-by-day due to soiling, the expected cleaning gains G increase ($P_m - P_c$ grows larger in Equation 4.3). However, D decreases by 1 every day. Overall, the area of GD (Figure 4.11 Day 15) will grow larger as the relatively large increase in height will outweigh the relatively small decrease in width. As a result, we expect the profit value will increase initially after cleaning. Note that we expect profit to increase quadratically due to a constant increase in G and a constant decrease in D leading to linear 1st differences. This is shown via sample calculations in Table 4.1 on the next page.

Eventually the effect of the decrease in D (nearing rainfall event) outweighs the effect of the increase in G and the cleaning benefit area starts to decrease (Figure 4.11 Day 30), leading the profit to decrease as well. At one point, $P_m - P_c$ will grow to be larger than $\bar{g}$ causing G to take on the constant value of $\bar{g}$ (see Equation 4.3). Thus, the expected gains from cleaning will just equal the average gains from the past. With G constant, and D still decreasing by 1 every day, we expect the profit to move from a quadratic shape to a linear one.

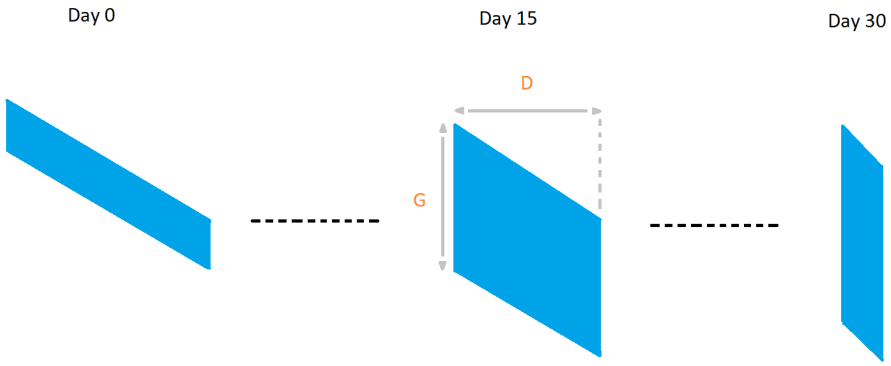


Figure 4.11: Progression of GD as we approach rainfall

Table and Plot Results

Key values from the simulation are shown in Table 4.1 and the corresponding profit curve is shown in Figure 4.12.

Days Since Cleaning	$\bar{g}$	P_c	$P_m - P_c$	G	D	GD	$GD_i - GD_{i-1}$
0	1500	13000	0	0	15	0	-
1	1500	12850	150	150	14	2100	+2100
2	1500	12700	300	300	13	3900	+1800
3	1500	12550	450	450	12	5400	+1500
4	1500	12400	600	600	11	6600	+1200
5	1500	12250	750	750	10	7500	+900
6	1500	12100	900	900	9	8100	+600
7	1500	11950	1050	1050	8	8400	+300
8	1500	11800	1200	1200	7	8400	0
9	1500	11650	1350	1350	6	8100	-300
10	1500	11500	1500	1500	5	7500	-600
11	1500	11350	1650	1500	4	6000	-1500
12	1500	11200	1800	1500	3	4500	-1500
13	1500	11050	1950	1500	2	3000	-1500
14	1500	10900	2100	1500	1	1500	-1500
15	1500	10750	2250	1500	0	0	-1500

Table 4.1: Profit Calculations

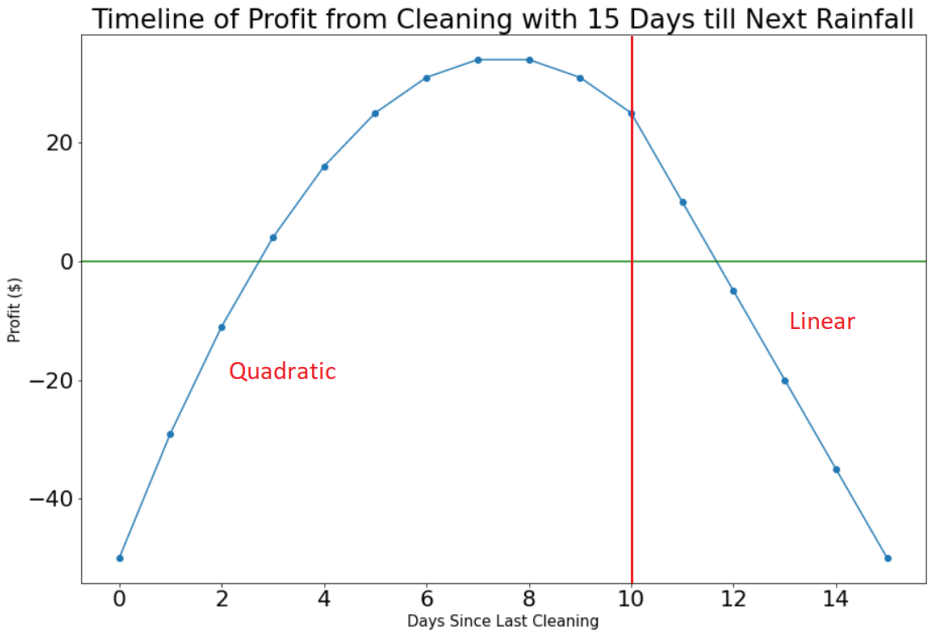


Figure 4.12: Cleaning Profit Timeline for 15 Days Till Rainfall

The $GD_i - GD_{i-1}$ column of Table 4.1 expresses the change in GD showing its quadratic and linear sections from Figure 4.12.

Overall, plant operators should be wary of cleanings too close to rainfall or too close to the recent cleaning date which result in losses. Cleanings scheduled in-between will result in profit, with the optimal cleaning dates indicated in the plots. Finally, the inversely proportional relationship between D and G, and its impact on profit, is also highlighted by the rise and fall of the curve.

Varying Days Till Rainfall

Continuing the simulation, we also examined the effect of a fast approaching rainfall on the cleaning profits. The profit curve for the exaggerated case of only 5 days till rainfall is shown in Figure 4.13.

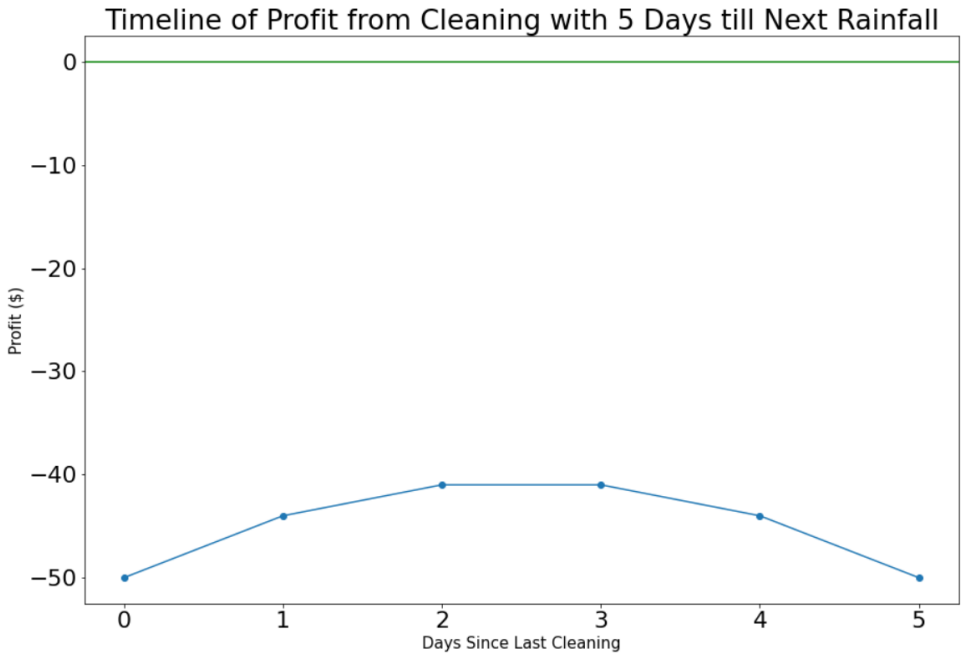


Figure 4.13: Cleaning Profit Timeline for 5 Days Till Rainfall

As expected, with rainfall just around the corner, cleaning is not profitable on any day. The optimal action in this case is to not to clean at all!

4.4.5 Optimal Cleaning Dates

Description

In the previous section we simulated a scenario where an SCB had just been cleaned. We can also apply those simulations to the real world with real data based on the current state of SCBs. The goal is to project future SCB performance based on current data and determine the optimal cleaning date which results in maximum profit.

Calculation

To determine profit from cleaning today, we follow the profit calculation detailed in Section 4.4.3 using the average cleaning gains ($\bar{g}$) extracted from the SCBs' timelines (Section 4.4.1) and the expected cleaning gains (G) calculated following Equation 4.3, which accounts for current and maximum production levels (P_c and P_m).

To determine profit from future cleanings, we follow the same profit calculations except while updating some of the parameters. Specifically, D will decrease by 1 every day as we get closer to rainfall. In addition, we will need to project future power production and update G according to P_c. This is because as the SCB degrades in performance due to soiling, power production will decrease and the expected cleaning gain may increase.

To project future power production, we calculate the daily production loss from soiling (L). This involves the current production (P_c) and the current efficiency (E_c), which come directly from the SCB data, as well as the current soiling rate (S_c), ie. daily efficiency loss as calculated in Section 4.4.1. Using these 3 values, we calculate L as follows:

$$L = \frac{S_c}{E_c} P_c \tag{4.4}$$

This is a simple proportion calculation to determine the change in power production from P_c if the efficiency E_c decreased by S_c daily.

Example

We now apply these calculations to project the profit curve for SCB 246 with 15 days till rainfall, as shown in Figure 4.14.

Note that unlike previous simulations where the SCB was recently cleaned, here SCB 246 is currently soiled resulting in an immediate positive profit from cleaning. The remainder of the curve behaves in the same way, with profit increasing initially and declining as we approach rainfall.

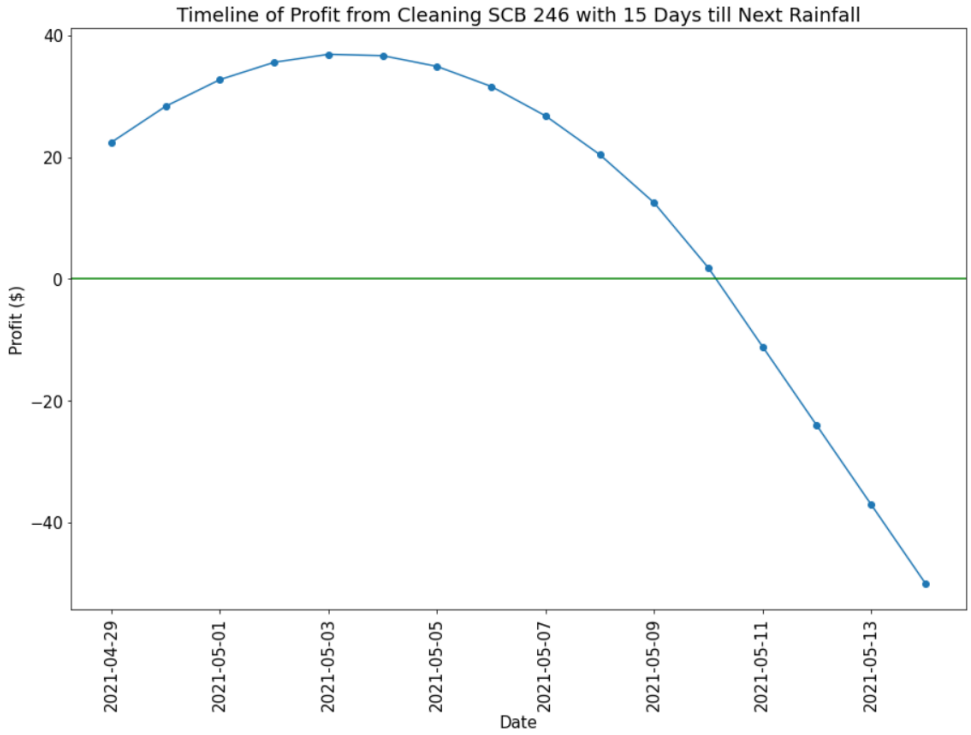


Figure 4.14: Cleaning Profit Timeline for SCB 246

Cleaning Schedule Optimizations

Using these projections, we can help answer the following questions regarding optimal cleaning dates:

- For a given SCB, when should cleaning be performed for maximum profit? - Examine the profit curve till the day of rainfall and choose the date with the largest profit
- When should site-wide cleaning be performed for maximum profit? - Sum profits from cleaning for all SCBs on all days until rainfall. Choose the date with the largest total profit.

Note that these are data-driven suggestions for the optimal cleaning dates conditional on the accuracy of rainfall forecasts. While for the most part they will hold true, predicting the future exactly is not possible and it may be the case that rain occurs earlier than forecasted, invalidating these projections.

4.4.6 Summary

Given an accurate segmentation of each SCB's efficiency timelines, we have now developed models and methodology to determine optimal cleanings and maximize profit.

Based on insights from past cleaning intervals, our smart cleaning-benefit calculations determine expected gains which account for unreliable cleaning quality. The cleaning-benefit model accumulates these gains until rainfall, to calculate total cleaning benefit. This is then compared with cleaning costs to formulate our profit function. Key characteristics and behaviour of the profit function, in a real-world scenario with varying parameters, is examined through simulations. Finally, we apply these simulations with the SCBs' current data to calculate optimal cleaning dates.

We have now successfully completed the final step of our research approach (Section 1.3.1 and Figure 1.5). Together with the data-cleansing methodology and the segmentation algorithm, this forms a comprehensive, data-driven approach for plant operators to optimize cleanings for their own PV plants.

Chapter 5

Results

To generate the results presented in this chapter, our 3-step research approach is applied to the PV data of all SCBs in the dataset. Starting with the raw data for each SCB, we first apply the data-cleansing methodology (Section 4.1) to generate clean efficiency timelines. The timelines are then divided into soiling and cleaning intervals through cleaning event detection using the segmentation algorithm (Section 4.3). Finally, daily cleaning profit is calculated and used to optimize cleanings for the SCBs and the PV plant as a whole (Section 4.4).

To showcase the results, we focus on data from SCB 216, a typical SCB, and examine its progression through each step of our approach.

5.1 Original Data

Using the raw PV data for SCB 216, we first examine its efficiency timeline as is, without performing any data-cleansing. The efficiencies are calculated for each 5-minute reading with temperature and capacity correction, as detailed in Section 3.3. To generate daily efficiency values, we simply average the 5-minutely efficiency for the day. These daily values are then plotted to form the raw efficiency timeline shown in Figure 5.1.

Note that without any data-cleansing in place, the efficiency calculations suffer from noise and operational errors. This is reflected in the timeline with high day-to-day variance in efficiency and no interpretable patterns in the data to define soiling and cleaning periods. Further, the efficiency values themselves are not in line with what we expect. From the plant specifications, the maximum operational efficiency of the PV modules is $\approx 15\%$, but

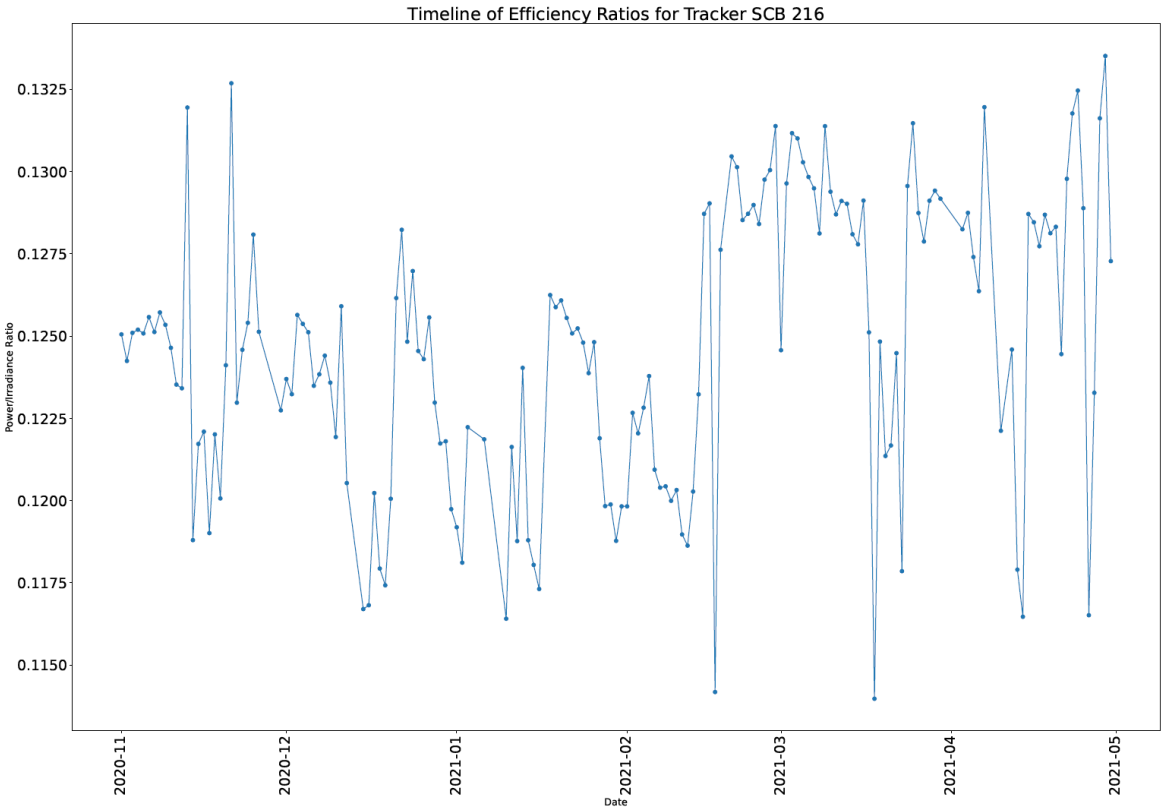


Figure 5.1: Raw Efficiency Timeline for SCB 216

here the calculated efficiencies only reach up to 13.25%. If the segmentation algorithm were to be applied with this timeline, the generated segments would be inaccurate and the results meaningless.

5.2 Timeline after Data-Cleansing

To address the noise and prepare the efficiency timeline for the segmentation algorithm, we now apply our data-cleansing methodology. Specifically this involves:

- Data Filters (Section 4.1.1) - Restricting data to periods of peak power production and avoiding any glitches or operational errors by using readings between **10 AM - 2 PM** with **Power** $>$ **0** and **Irradiance** $>$ **600**

- Cumulative Efficiency Ratios (Section 4.1.2) - Calculating daily efficiency using the **total corrected power output** and **total irradiance input** during the day
- Cloudy Day Detection (Section 4.1.3) - Excluding days with poor correlation (**coefficient score** $<$ **0.8**) between power and irradiance
- Best-Fit Pyranometer Measurements (Section 4.1.4) - Out of the 4 available irradiance measurements, using the one which has a correlation coefficient score of at least 0.8 with the SCB's power data for the most number of days. For SCB 216 this was `Tilt_Irr_2_pyrano`

The resulting clean efficiency timeline is shown in Figure 5.2.

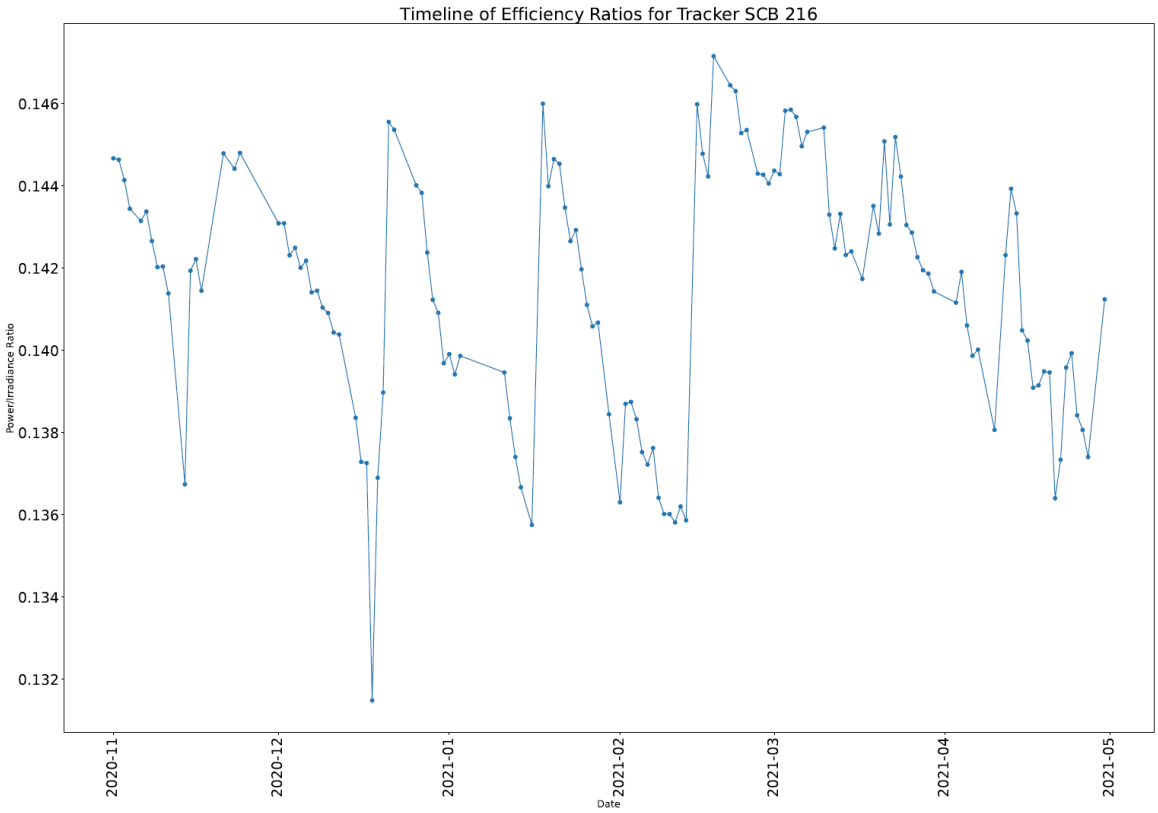


Figure 5.2: Clean Efficiency Timeline for SCB 216

Note that this timeline looks different than the raw timeline from Figure 5.1. The day-to-day variance in efficiency is noticeably lower with the impact of noise being limited

although not entirely eliminated. The timeline is sparser with some of the most cloudy days being removed, but enough data remains to observe clear patterns of soiling/cleaning periods for the segmentation algorithm to run effectively. Finally, the maximum calculated efficiencies have increased from 13.25% in Figure 5.1 to 14.6% in Figure 5.2, closer to the expected maximum of 15%. This improvement highlights the importance of data-cleansing, in particular the use of data filters and cumulative efficiency ratios, for accurate efficiency calculations.

5.3 Segmented Timeline

The segmentation algorithm detailed in Section 4.3 is now applied to the clean efficiency timeline. The resulting segmented timeline is shown in Figure 5.3.

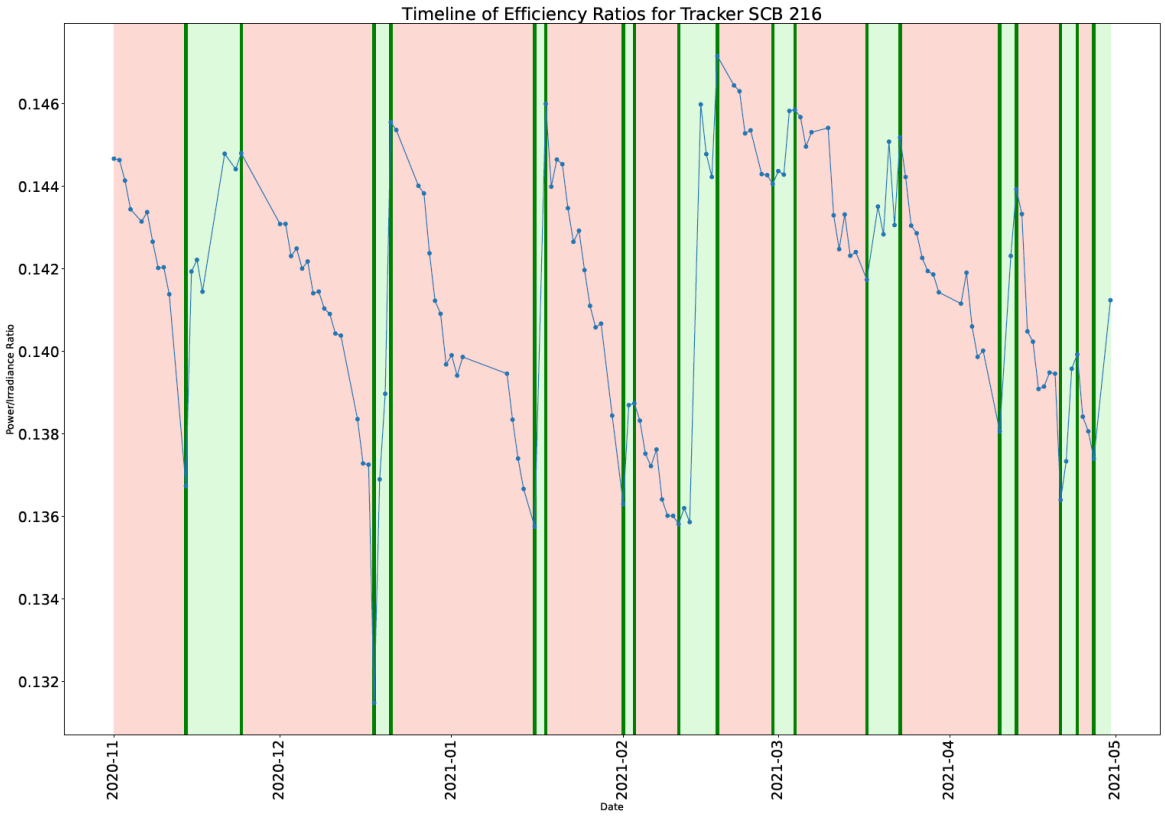


Figure 5.3: Segmented Efficiency Timeline for SCB 216

Notice that the marked cleaning (marked in green) and soiling (marked in red) intervals follow the visible efficiency patterns in the timeline. The timeline is divided into generally up-sloping and generally down-sloping segments with a high degree of accuracy, with the minor day-to-day variances in efficiency safely ignored.

5.4 Extracting Key Information

We now extract information on soiling loss and cleaning gains over the course of the time period defined by each interval in the segmented timeline. These results are organized into Table 5.1.

Segment Type	Start Date	End Date	Soiling Loss (%/day)	Cleaning Gain (kWh)
SOILING	2020-11-01	2020-11-14	-0.000345	-
CLEANING	2020-11-14	2020-11-24	-	+959.187
SOILING	2020-11-24	2020-12-18	-0.000329	-
CLEANING	2020-12-18	2020-12-21	-	+4325.158
SOILING	2020-12-21	2021-01-16	-0.000369	-
CLEANING	2021-01-16	2021-01-18	-	+889.055
SOILING	2021-01-18	2021-02-01	-0.000597	-
CLEANING	2021-02-01	2021-02-03	-	+3211.425
SOILING	2021-02-03	2021-02-11	-0.000380	-
CLEANING	2021-02-11	2021-02-18	-	+1947.294
SOILING	2021-02-18	2021-02-28	-0.000335	-
CLEANING	2021-02-28	2021-03-04	-	+210.042
SOILING	2021-03-04	2021-03-17	-0.000331	-
CLEANING	2021-03-17	2021-03-23	-	+160.745
SOILING	2021-03-23	2021-04-10	-0.000301	-
CLEANING	2021-04-10	2021-04-13	-	+883.171
SOILING	2021-04-13	2021-04-21	-0.000690	-
CLEANING	2021-04-21	2021-04-24	-	+1259.090
SOILING	2021-04-24	2021-04-27	-0.000751	-
CLEANING	2021-04-27	2021-04-30	-	-1684.189

Table 5.1: Extracted Information from each Segment

Note that while most of the cleaning gains indicate improvement in power production, the gain for the last cleaning interval is negative! This occurs because the incoming total irradiance was lower on the cleaning end date compared to the cleaning start date. So although efficiency improved from the start to the end of cleaning, the total power production actually decreased. This limitation and ways to address it for future work are discussed in more detail in Chapter 6.

In this specific case, the negative gain will not affect our calculations. Since the SCB is currently being cleaned, this last cleaning interval will be ignored so that an unfinished cleaning with low production gains does not impact the average cleaning gain.

5.5 Current Cleaning Profit Calculation

We calculate profit from cleaning today using Equation 4.2 (Section 4.4.3). The following parameters are calculated from the SCB's data:

- Average kWh gain from past cleanings ($\bar{g}$) - Examining cleaning gains for all cleaning intervals in the past 60 days from Table 5.1, we get:

$$\begin{aligned}\bar{g} &= \frac{1259.090 + 883.171 + 160.745}{3} \\ &= 767.67 \text{ kWh}\end{aligned}$$

- Maximum Daily Power Production (P_m) - Calculated directly from daily production data as: $P_m = 15140.128$ kWh
- Current Summed Power Production (P_c) - Calculated directly from latest day's production data as: $P_c = 12773.826$ kWh

Using these parameters, the expected kWh gain from cleaning G is calculated following Equation 4.3 as:

$$\begin{aligned}G &= \min\{\bar{g}, P_m - P_c\} \\ &= \min\{767.67, 15140.128 - 12773.826\} \\ &= \min\{767.67, 2366.302\} \\ &= 767.67 \text{ kWh}\end{aligned}$$

Additionally, the following user-input parameters, which are to be set by plant operators for their specific sites, are initialized to standard values:

- Number of days till rainfall (D) - Set to 15 following an examination of rainfall distribution for the PV plant site
- PPA conversion rate from kWh to \$ - Set to \$0.03 USD per kWh following standard rates offered for utility-scale PV plants in India [20]
- Cleaning Cost - Set to \$50 USD per SCB. In reality, this may differ from site to site dependent on cleaning methods, labour wages etc. and will be defined by the plant operator

With all parameters calculated, we now determine profit from cleaning SCB 216 today by following Equation 4.2:

$$\begin{aligned}\text{Profit} &= GDR - C \\ &= 767.67 \text{ kWh} \cdot 15 \cdot \$0.03/\text{kWh} - \$50 \\ &= \$295.45\end{aligned}$$

5.6 Future Profit Projections

To determine profit from future cleanings, we must first project future power production by calculating daily production loss (L) as detailed in Section 4.4.5. This calculation involves the following parameters:

- Current Summed Power Production (P_c) - Calculated directly from latest day's production data as: $P_c = 12773.826$ kWh
- Current Efficiency (E_c) - Retrieved directly from the efficiency timeline as: $E_c = 0.1412\%$
- Current Soiling Rate (S_c) - Retrieved from Table 5.1 as: $S_c = -0.000751\%/\text{day}$

The daily production loss is then given by Equation 4.4:

$$\begin{aligned}L &= \frac{S_c}{E_c} P_c \\ &= \frac{-0.000751\%/\text{day}}{0.1412\%} \cdot 12773.826 \text{ kWh} \\ &= -67.898 \text{ kWh/day}\end{aligned}$$

Using this daily loss, we can now project future production levels and calculate future values of G to determine profit. The projected profit curve for SCB 216 is shown in Figure 5.4.

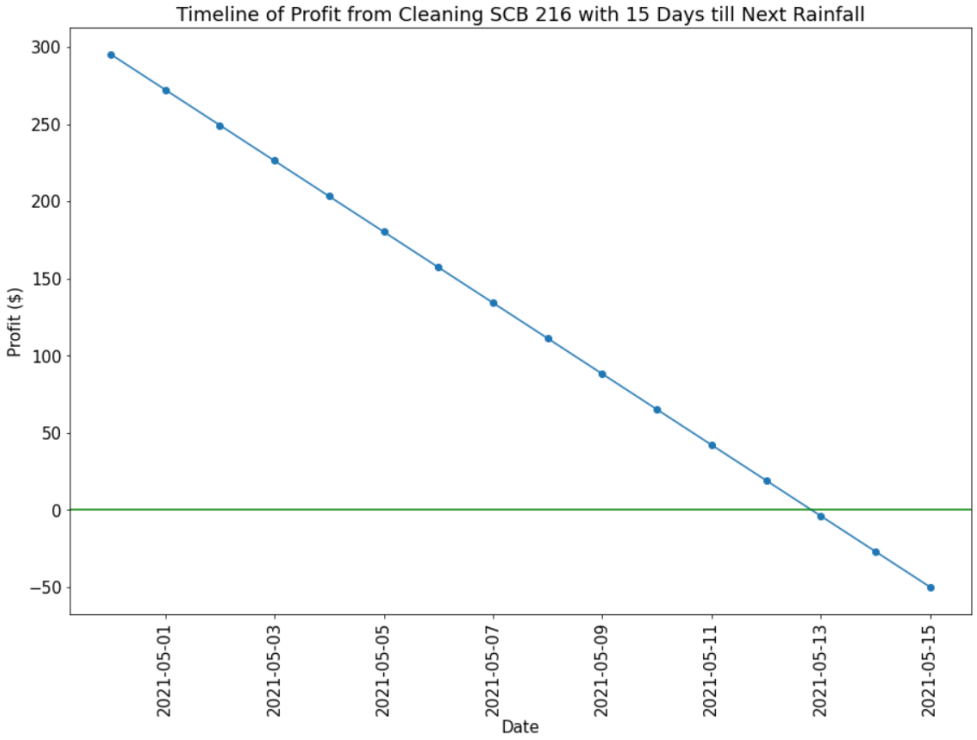


Figure 5.4: Cleaning Profit Timeline for SCB 216

Note that profit does not increase initially, and instead decreases linearly till the day of rainfall. This is because the level of soiling on the SCB is such that the optimal action is to clean immediately, and for every passing day in which the SCB is not cleaned, a loss in profit will be incurred.

5.7 Optimization Results

The calculations in the previous sections are performed for each SCB in the dataset to generate both profit from cleaning today and profit from future cleanings. With these results, the following optimization questions are answered:

- **Regarding cleaning decisions for today** (Section 4.4.3)
  - For a given SCB, is it beneficial to clean today?
    - For SCB 216 with a positive profit of **$295.45**, **Yes**
    - In total for this PV plant, cleaning is beneficial for **194** SCBs with positive profits but not for **57** SCBs with negative profits
  - If cleaning is to be performed today, which SCB should be prioritized?
    - Comparing profits from all SCBs for this PV plant, **SCB 444** should be prioritized for cleaning today with the maximum cleaning profit of **$1056.11**
  - Is it beneficial for all SCBs to be cleaned (site-wide cleaning) today?
    - Summing profits across all SCBs, we determine the total cleaning profit for today is **$59,417.08**. Since this is a positive profit, site-wide cleaning is indeed profitable today
- **Regarding optimal cleaning schedules** (Section 4.4.5)
  - For a given SCB, when should cleaning be performed for maximum profit?
    - For SCB 216, by examining the profit curve in Figure 5.4 we determine that the optimal cleaning date with maximum profit is **2021-04-30**
  - When should site-wide cleaning be performed for maximum profit?
    - Summing profits from cleaning for all SCBs on all days until rainfall, we determine that the optimal site-wide cleaning date is also **2021-04-30**

With the above questions answered by the thesis results, we provide PV plant operators with a complete guide to scheduling optimal cleanings. Whether dealing with cleanings per-SCB or site-wide, for today or in the future, all scenarios are explored in detail and solved by our optimization results.

5.8 Evaluation

In comparison to the periodic cleaning approaches applied by many solar plants, our optimal cleaning schedules should provide larger cleaning benefits while minimizing cleaning costs. To showcase the advantages, we evaluate the optimal cleaning schedule generated using our optimization approach by examining the cleaning profits over a 6-week segment of synthetic power production data. Figure 5.5 shows a plot of this timeline.

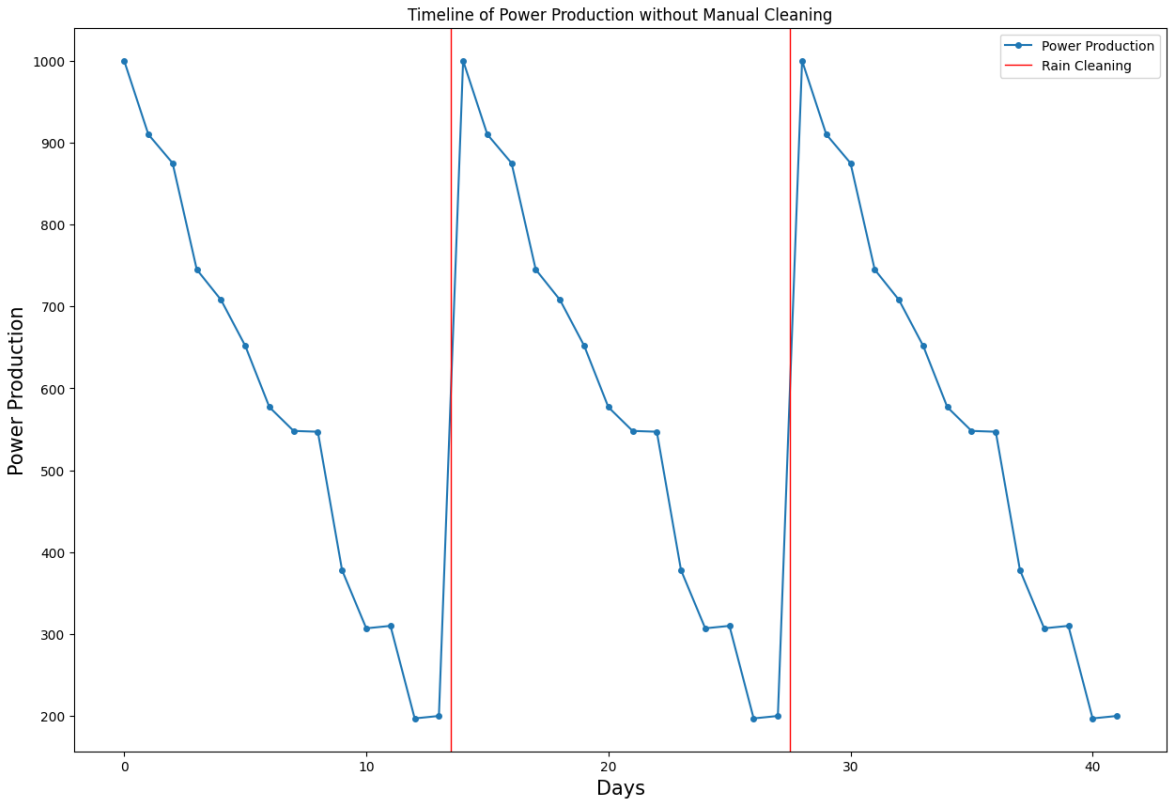


Figure 5.5: Baseline 6-Week Power Production Timeline

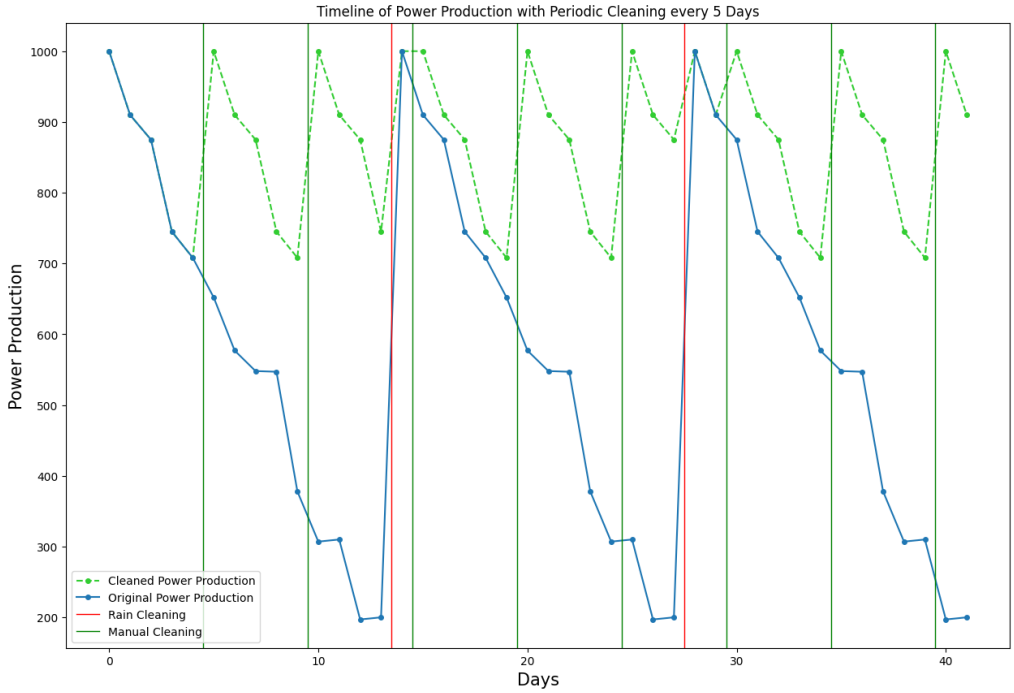
Unfortunately while working on this evaluation, access to real plant data was lost. However, even if we could get access, random rain cleanings would not allow for a clean, replicable trace for our evaluation. As a result, we decided to produce controlled synthetic power production data ourselves, aiming to get as close a representation as possible to real power production patterns. To do so, we chose a maximum power production value of 1000 and randomly generated linear daily production loss values drawn from a uniform distribution in the range [-200, 10] to form a 14-day timeline segment. This segment was then replicated 3 times with 2 rainfall cleanings restoring power production back to maximum, to form the 42-day power production timeline in Figure 5.5. Rather than a constant daily loss, the random loss is effective in replicating real-world patterns such as varying production losses and small jumps in production due to noise. Finally, we chose to work directly with power production instead of efficiency to help visualize the cost-benefit trade-offs of each cleaning approach and simplify the profit calculations. Since we

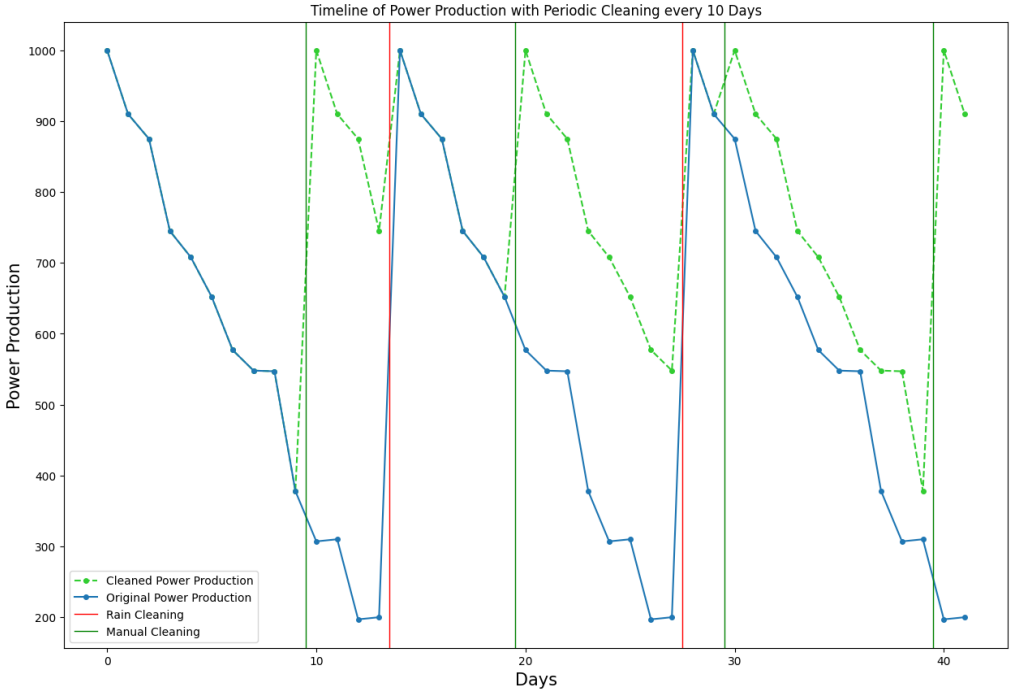
are only concerned with cleaning schedule evaluation, efficiency timelines, while useful for segmentation and soiling rate extraction, are not necessary.

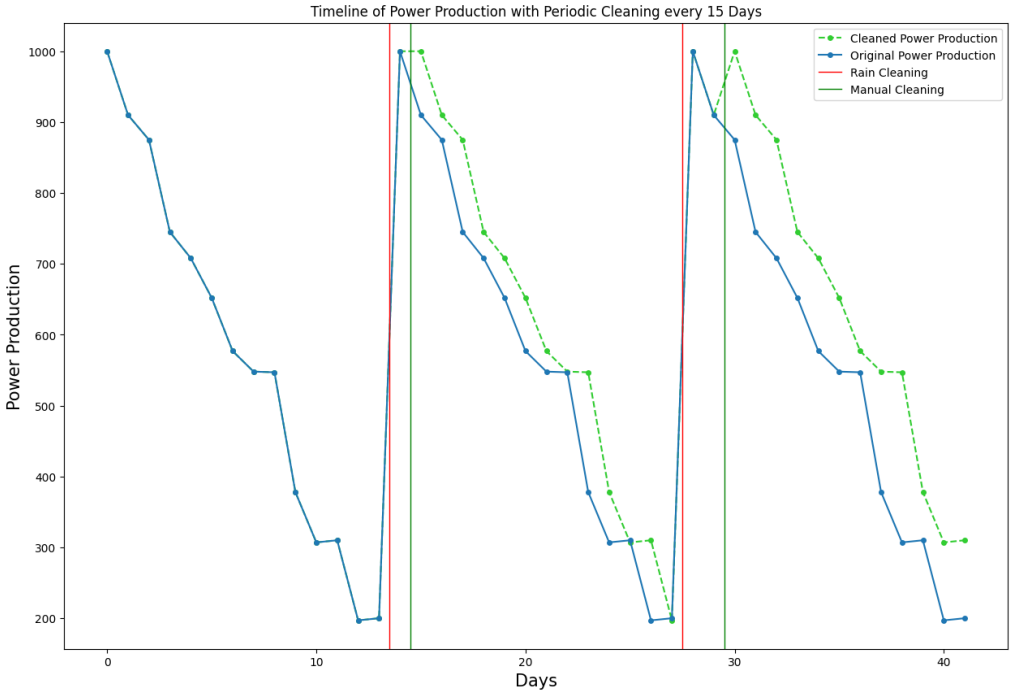
This data is used to gauge production benefits of both the periodic and optimal cleaning approaches. Note that the goal here is to replicate a typical power production timeline in the absence of any human intervention, which will serve as a baseline to compare the benefits of different manual cleaning approaches. For simplicity, we assume all cleanings, both rainfall and manual, restore power production to 100% capacity.

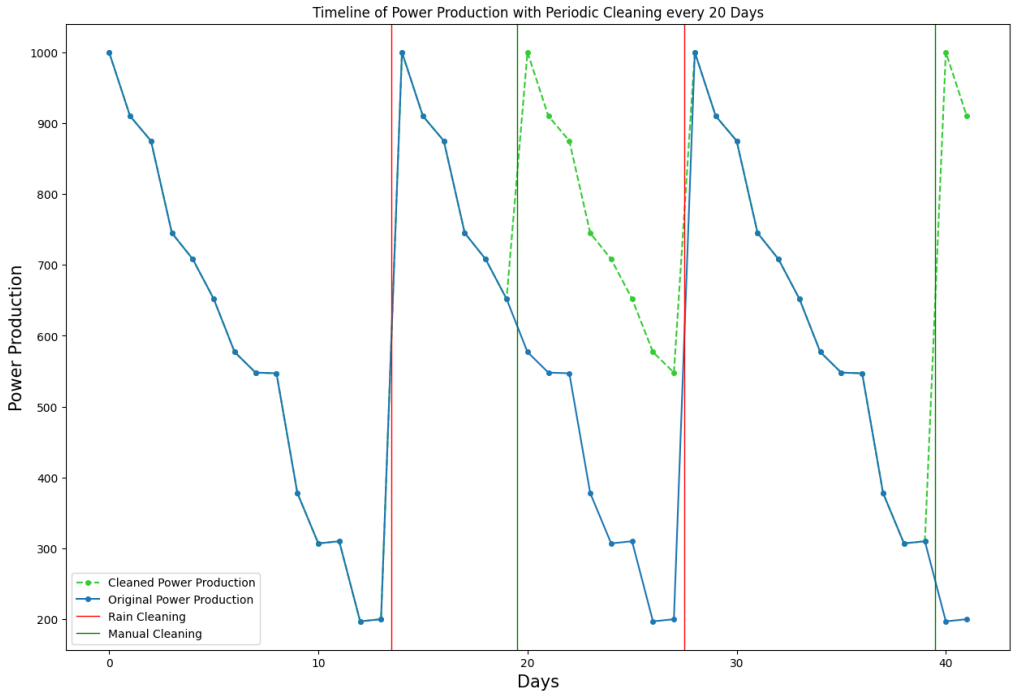
5.8.1 Periodic Cleaning

Using our baseline data, we now project the power production timeline when periodic perfect cleaning is performed. The resulting timelines for cleaning periods of 5, 10, 15, and 20 days are shown in the following plots.




Timeline of Power Production with Periodic Cleaning every 10 Days
Power Production
Days
Cleaned Power Production
Original Power Production
Rain Cleaning
Manual Cleaning

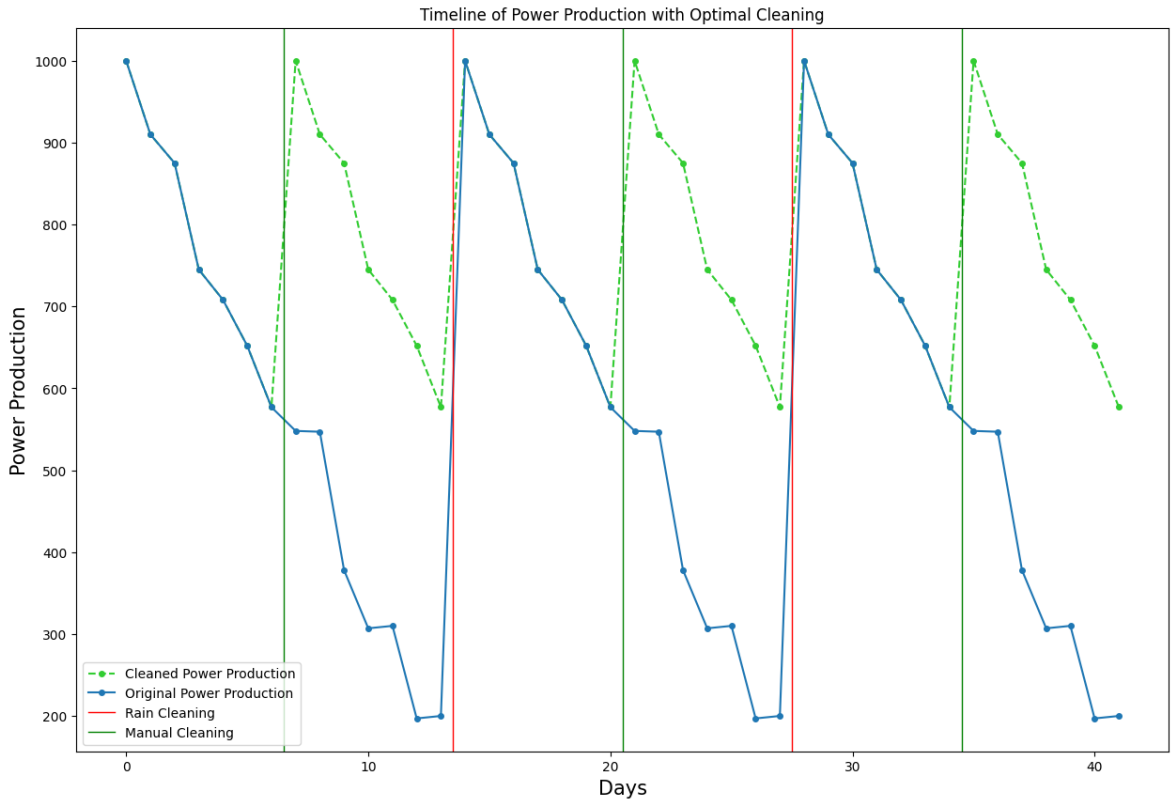

Timeline of Power Production with Periodic Cleaning every 15 Days
Power Production
Days
Cleaned Power Production
Original Power Production
Rain Cleaning
Manual Cleaning



With these plots, it is clear that periodic cleaning without any consideration for rainfall events will always be sub-optimal. Cleanings are either scheduled too frequently, too far apart, or too close to rainfall, leading to losses in production revenue and unnecessary cleaning costs being incurred.

5.8.2 Optimal Cleaning

We now apply our optimization approach to schedule cleanings which maximize profits while accounting for upcoming rainfall. This cleaning schedule and its power production timeline are shown in the following plot.



From this timeline, we see that the cleanings are evenly spaced between rainfall events, such that we are not cleaning too close to rainfall and optimizing our benefits. Further, with rainfall events being 14 days apart, only one cleaning per rainfall interval is sufficient to maximize production benefits relative to cleaning costs.

5.8.3 Profit Comparison

To calculate total cleaning profit from each timeline, we first calculate the total production benefit of each cleaning when compared to the baseline production levels. This is simply the difference between the cleaned power production and the original power production summed over the entire timeline. Then, using the PPA conversion rate ($0.03/kWh) and cleaning cost ($50) from Section 5.5 we convert this production benefit to a monetary value and subtract cleaning costs depending on the number of cleanings performed.

The total cleaning profits of each approach is shown in Figure 5.6.

Figure 5.6: Cleaning Profits

Comparing these profits with the scheduling of cleanings for each approach, we see that when cleanings are carefully spaced between rainfall events, profit is maximized. Moreover, frequent cleanings, such as when the cleaning period is 5-days, will lead to losses, while longer cleaning periods are generally more profitable. However, a 15-day cleaning period, which only schedules cleanings directly after rainfall events is the least profitable.

Overall, we see that the positioning of cleaning relative to rainfall is the most influential factor impacting profits. Therefore, it makes sense that our approach which calculates this exactly, produces the highest profit - more than double the highest profit for periodic cleaning. Note that these results are for a single synthetic trace meant to illustrate the importance of accounting for rainfall events. However, it is clear that the results generalise, with the cleaning benefits depending both on the trace and the cleaning interval.

Chapter 6

Discussion

6.1 Contributions

In this thesis, we approached the soiling and cleaning problem from a purely **data-driven** perspective. Working with measured PV data from a utility-scale plant, we encountered issues arising from real-world conditions relating to the nature of the hardware, the environment, and day-to-day plant operations. To address these problems, we developed data-driven solutions to extract key insights from the PV data and guide optimal cleanings.

Specifically, the thesis work makes 4 contributions to the PV soiling/cleaning research field. A brief overview of each contribution is provided below:

1. Data-Cleansing Methodology - Systematic techniques to a) **identify** non-soiling related noise/anomalies in PV data such as uneven shading due to micro-weather conditions, and b) **minimize** their impact on calculations to generate clean efficiency timelines
    - Using data filters, cumulative efficiency ratios, cloudy day detection via correlation coefficients, and best-fit irradiance measurements
    - Existing works rely on simple visual recognition and elimination of outliers in the dataset
2. Timeline Segmentation Algorithm - Monitors **changes in direction of slope** for a moving window of points to segment the efficiency timeline into generally up-sloping (cleaning) and generally down-sloping (soiling) intervals

- Rather than depending on the unreliable (sometimes unavailable) **human-recorded** cleaning logs, patterns in the efficiency timeline are analyzed for **data-driven** cleaning event detection
- Compared to value-based threshold methods in existing works, a slope-based approach is more reliable and robust against outliers
- Theil-Sen Estimator is used for slope calculations for added outlier resistance

3. Smart Cleaning-Gain Calculations - Accounting for **unreliable cleanings** and calculating expected cleaning gains based on the power production gains observed from past cleaning intervals
    - Existing works assume perfect cleaning
    - We consider the possibility of imperfect cleanings and calculate a **smarter data-driven estimate** based on the quality and power production patterns of past cleanings
4. Cleaning Schedule Optimization - Profit calculations (in light of rainfall events) via cost-benefit analysis to **optimize cleaning decisions for today** and **determine overall optimal cleaning dates**
    - Per-SCB results to account for **spatial variability** in soiling; this provides solutions to cleaning-optimization scenarios targeting different areas of a PV plant
    - Optimal cleanings are scheduled while taking into consideration impending **rainfall events** and their effect on cleaning profit

With this, we contribute a completely data-driven approach to monitor and mitigate soiling impact on power production, without the need to install expensive equipment. Plant operators can follow the thesis work and directly apply it to the data for their own sites at no added cost. As such, our approach not only presents key insights into PV soiling/cleaning but also provides immediate benefit in the real world.

6.2 Limitations

While the thesis work describes an effective approach to optimizing cleaning schedules, it is not beyond reproach. Limitations, specifically relating to the estimation of cleaning benefits and the calculation of future profits, are discussed below:

- Rainfall Cleaning Assumption - When defining our cleaning-benefit model in Section 4.4.2 (Figure 4.9) we assumed rainfall would completely clean the SCBs and level both cleaned and soiled timelines to the same power production
  - This simplified our calculations allowing us to stop accumulating cleaning gains at the rainfall date
  - In reality, depending on the rainfall intensity and duration as well as the current soiling state of SCBs, different cleaning effects may be observed, from no impact on soiling to complete cleanings
  - This would change our cleaning benefit calculations accordingly, and we might have to continue the accumulation of cleaning gains beyond the rainfall date
  - Estimating the level of cleaning from a rainfall event is complex and is left for future work
- Negative Cleaning Gains - Although rare, as highlighted in Section 5.4 (Figure 5.1) there is the possibility of negative power production gains being observed across some cleaning intervals
  - This occurs because the available sunlight for power production is lower on the cleaning end date compared to the cleaning start date. In this case, even though efficiency improves across the cleaning interval, power production will still decrease, simply because the higher efficiency cannot make up for the lack of available sunlight
  - While this may affect the cleaning gain of a specific interval, its overall impact on cleaning gain calculation is limited for 2 reasons:
    1. Such cases are scarce, with the overwhelming majority of cleaning intervals producing positive production gains. In arid regions with clear sunny conditions, an increase in efficiency due to cleaning will almost always result in a power production increase
    2. We average cleaning gains across the past 60 days of cleaning intervals and so a negative kWh gain will have little impact on our final estimate
- Future Power Production Projections - When projecting power production levels for future profit calculation in Section 4.4.5, we assumed a constant daily production loss proportional to the latest soiling rates and current production levels
  - In reality, increases in irradiance may actually result in production gains despite the efficiency decrease due to soiling

  - Since the profit calculations depend on accurate projections of power production, our cleaning schedule estimates may vary from the actual optimal cleaning dates, resulting in less profit being realized than the maximum possible in theory

- Dependence on Rainfall Forecasts - Similar to projecting future power production, our optimization results are also conditional on accurate rainfall forecasts
  - While we determine optimal cleaning dates based on the rainfall distributions of the plant site, it may be the case that rain occurs earlier than forecasted resulting in sub-optimal (but still profitable) cleanings

These limitations stem from uncertainty in future events, such as the quality and forecast of rainfall events, and projections of cleaning gain and power production. It is impossible to be 100% accurate and predict the future perfectly. Instead, we can only look back at past data patterns and develop smart data-driven estimates. This is what we do in this thesis and although our estimates might not produce true optimal results, for the most part we expect they will get as close as possible in practice.

6.3 Future Work

While our estimations can never be 100% perfect, there are some adjustments we can make to improve accuracy and address the aforementioned limitations. These extensions, along with other avenues of future work, are discussed below:

- Calculating Efficiency Gains - Rather than examine power production gains across cleaning intervals, we could calculate efficiency improvements instead
  - Using the current efficiency and the average efficiency gain from past cleaning intervals, the expected kWh gain from cleaning would be calculated proportional to the current power production
  - This may provide more accurate cleaning gain estimates and at the very least the limitation of negative kWh gains across cleaning intervals will no longer be a concern
  - Alternatively, future irradiance simulations from weather models could be used to project power production levels and kWh gains from cleaning; this would allow for a more complete estimate of cleaning gains and profits when compared to our approach of assuming a constant daily production loss

  - There is one challenge with this approach: since different irradiance measurements are used for different SCBs, different ranges of efficiencies will be calculated. This means the efficiency gain must first be normalized before being applied. A deeper investigation is necessary to avoid bias towards certain SCBs.

- Investigating Rain Cleanings - Examine the effect of rainfall, with varying intensity and duration, on soiling levels; the goal here is to generate insights into the expected production benefits of rain cleanings
  - Instead of assuming rain would completely clean SCBs in our cleaning-benefit model, we could input estimations of the rain-cleaning gains and adjust the cleaning-benefit calculation accordingly
  - This could also be applied to distinguish between rainfall vs manual human cleanings in the efficiency timelines; we could then average cleaning gains from only human cleaning intervals for more realistic calculations of cleaning gains

- Investigating Temporal Variability of Soiling - Soiling rates and losses for a utility-scale PV plant may vary dependent on the time of year, with soiling becoming more pronounced during summer months in particularly dry conditions
  - Although we don't examine it directly, we calculate soiling rates for every soiling interval across the dataset timeline, so a deeper investigation should be straightforward
  - Analyzing monthly or weekly rates could provide insights into soiling impact at different times of the year, which would help improve our cleaning schedule optimizations

- Applying the Approach to Different Datasets
  - If we had access to data from other plants, we could apply our approach to their PV datasets, evaluating the results and adapting our approach accordingly to avoid over-fitting to one specific PV plant
  - The ideal evaluation would include real-world implementation of the generated cleaning schedule and inspection of the resulting cleaning production gains and profits

- Analyzing Systematic Degradation of SCBs - All solar panels will naturally degrade in performance over time, due to wear and tear from exposure to harsh climate conditions such as high temperatures [30]

  - Statistically, we can expect a 0.5% decrease per year in power production due to panel degradation [30]
  - The same way we control for non-soiling factors when calculating efficiency in Section 3.3, panel degradation can also be accounted for to ensure the only factors affecting efficiency are soiling and cleaning events
  - However, we do not expect any noticeable impact in soiling rate calculations as the timescale for physical panel degradation is much longer (year-to-year) when compared to soiling which results in immediate efficiency losses

- Detailed Costs of Cleaning - Rather than a static cost of cleaning associated with every SCB, we can improve our cleaning cost estimates based on parameters such as the capacity
  - Of course, cleaning larger SCBs with more panels will require more water, more time, more workers etc. and so the cleaning costs should be defined as a function of SCB capacity
  - As well, an interesting extension to our current cleaning schedule optimization would be to consider route planning, where rather than only cleaning individual SCBs on their optimal dates, the neighbouring SCBs can also be targeted to save labour costs. Given our per-SCB calculations, we can aggregate results across SCBs to generate cleaning schedules for different areas of the plant.

The methodology presented in this thesis provides a sound framework to extract soiling and cleaning results from any PV dataset, serving as a solid foundation for future work and extensions.

References

[1] IEA-PVPS Task 13. Soiling losses impact on the performance of photovoltaic power plants. Technical report, International Energy Agency - Photovoltaic Power Systems Programme, December 2022.

[2] Alexandra Arntsen. Rethinking soiling. https://www.pv-magazine.com/2023/01/19/rethinking-soiling/, Jan 2023.

[3] João Gabriel Bessa, Leonardo Micheli, Florencia Almonacid, and Eduardo F Fernández. Monitoring photovoltaic soiling: assessment, challenges, and perspectives of current and potential strategies. *Iscience*, 24(3):102165, 2021.

[4] Pierre Besson, Constanza Munoz, Gonzalo Ramirez-Sagner, Marcelo Salgado, Rodrigo Escobar, and Werner Platzer. Long-term soiling analysis for three photovoltaic technologies in santiago region. *IEEE Journal of Photovoltaics*, 7(6):1755–1760, 2017.

[5] Sonali Bhaduri, Shashwata Chattopadhyay, Sachin Zachariah, Chetan Singh Solanki, and A Kottantharayil. Evaluation of increase in the energy yield of pv modules by inverting the panels during the non-sunshine hours. In *26th Int. Photovolt. Sci. Eng. Conf.*, 2016.

[6] K. Branker, M.J.M. Pathak, and J.M. Pearce. A review of solar photovoltaic levelized cost of electricity. *Renewable and Sustainable Energy Reviews*, 15(9):4470–4482, 2011.

[7] Kudzanayi Chiteka, Rajesh Arora, SN Sridhara, and CC Enweremadu. A novel approach to solar pv cleaning frequency optimization for soiling mitigation. *Scientific African*, 8:e00459, 2020.

[8] International Electrotechnical Commission. *Photovoltaic system performance - Part 1: Monitoring*. International Electrotechnical Commission, 2017.

[9] Suellen C.S. Costa, Antonia Sonia A.C. Diniz, and Lawrence L. Kazmerski. Solar energy dust and soiling r&d progress: Literature review update for 2016. *Renewable and Sustainable Energy Reviews*, 82:2504–2536, 2018.

[10] Oscar Darteh, Qi Liu, Collins Oduro, Xiaodong Liu, and Charity Adjei. *A Survey on an Artificial Intelligence Approach to Maintenance of Solar Photovoltaic Modules*, pages 507–517. 05 2021.

[11] Michael G Deceglie, Leonardo Micheli, and Matthew Muller. Quantifying soiling loss directly from pv yield. *IEEE Journal of Photovoltaics*, 8(2):547–551, 2018.

[12] Michael G Deceglie, Matthew Muller, Zoe Defreitas, and Sarah Kurtz. A scalable method for extracting soiling rates from pv production data. In *2016 IEEE 43rd Photovoltaic Specialists Conference (PVSC)*, pages 2061–2065. IEEE, 2016.

[13] Mame Cheikh Diouf, Mactar Faye, Ababacar Thiam, and Vincent Sambou. A framework of optimum cleaning schedule and its financial impact in a large-scale pv solar plant: a case study in senegal. *EPJ Photovoltaics*, 13:21, 2022.

[14] Easan Drury, Paul Denholm, and Robert Margolis. Impact of different economic performance metrics on the perceived value of solar photovoltaics. Technical report, National Renewable Energy Lab.(NREL), Golden, CO (United States), 2011.

[15] Swapnil Dubey, Jatin Narotam Sarvaiya, and Bharath Seshadri. Temperature dependent photovoltaic (pv) efficiency and its effect on pv production in the world – a review. *Energy Procedia*, 33:311–321, 2013. PV Asia Pacific Conference 2012.

[16] U.S. Energy Information Administration (EIA). Preliminary monthly electric generator inventory. https://www.eia.gov/electricity/data/eia860m/, Jan 2023.

[17] EnergyEducation. Photovoltaic effect. https://energyeducation.ca/encyclopedia/Photovoltaic_effect.

[18] Benjamin Figgis, Bing Guo, Wasim Javed, Klemens Ilse, Said Ahzi, and Yves Rémond. Time-of-day and exposure influences on pv soiling. In *2017 International Renewable and Sustainable Energy Conference (IRSEC)*, pages 1–4. IEEE, 2017.

[19] Michael Gostein, J Riley Caron, and Bodo Littmann. Measuring soiling losses at utility-scale pv power plants. In *2014 IEEE 40th Photovoltaic Specialist Conference (PVSC)*, pages 0885–0890. IEEE, 2014.

[20] Uma Gupta. Renew power signs ppas for 2 gw in india. https://www.pv-magazine.com/2022/05/03/renew-power-signs-ppas-for-2-gw-in-india/, May 2022.

[21] International Energy Agency (IEA). Solar pv. https://www.iea.org/reports/solar-pv, September 2022.

[22] Klemens Ilse, Leonardo Micheli, Benjamin W Figgis, Katja Lange, David Daßler, Hamed Hanifi, Fabian Wolfertstetter, Volker Naumann, Christian Hagendorf, Ralph Gottschalg, et al. Techno-economic assessment of soiling losses and mitigation strategies for solar power generation. *Joule*, 3(10):2303–2321, 2019.

[23] Mercom India. With 2,245 mw of commissioned solar projects, world's largest solar park is now at bhadla. https://mercomindia.com/world-largest-solar-park-bhadla/.

[24] AD Jones and CP Underwood. A thermal model for photovoltaic systems. *Solar energy*, 70(4):349–359, 2001.

[25] Russell K Jones, Abdulaziz Baras, Abdullah Al Saeeri, Ayman Al Qahtani, Ahmed O Al Amoudi, Yousef Al Shaya, Maher Alodan, and Shafi Ali Al-Hsaien. Optimized cleaning cost and schedule based on observed soiling conditions for photovoltaic plants in central saudi arabia. *IEEE journal of photovoltaics*, 6(3):730–738, 2016.

[26] Felipe A Mejia and Jan Kleissl. Soiling losses for solar photovoltaic systems in california. *Solar Energy*, 95:357–363, 2013.

[27] Leonardo Micheli, Eduardo F Fernandez, Jorge T Aguilera, and Florencia Almonacid. Economics of seasonal photovoltaic soiling and cleaning optimization scenarios. *Energy*, 215:119018, 2021.

[28] Leonardo Micheli, Eduardo F Fernández, Matthew Muller, and Florencia Almonacid. Extracting and generating pv soiling profiles for analysis, forecasting, and cleaning optimization. *IEEE Journal of Photovoltaics*, 10(1):197–205, 2019.

[29] Leonardo Micheli and Matthew Muller. An investigation of the key parameters for predicting pv soiling losses. *Progress in photovoltaics: research and applications*, 25(4):291–307, 2017.

[30] Benjamin Mow. Stat faqs part 2: Lifetime of pv panels. https://www.nrel.gov/state-local-tribal/blog/posts/stat-faqs-part2-lifetime-of-pv-panels.html, April 2018.

[31] Office of Energy Efficiency & Renewable Energy. Photovoltaics. https://www.energy.gov/eere/solar/photovoltaics.

[32] Office of Energy Efficiency & Renewable Energy. Solar energy research areas. https://www.energy.gov/eere/solar/solar-energy-research-areas.

[33] A Massi Pavan, Adel Mellit, and D De Pieri. The effect of soiling on energy production for large-scale photovoltaic plants. *Solar energy*, 85(5):1128–1136, 2011.

[34] Pedro M Rodrigo, Sebastián Gutiérrez, Leonardo Micheli, Eduardo F Fernández, and FM Almonacid. Optimum cleaning schedule of photovoltaic systems based on levelised cost of energy and case study in central mexico. *Solar Energy*, 209:11–20, 2020.

[35] Michael Schoeck. 2023 will see the most utility-scale solar added in a single year. https://pv-magazine-usa.com/2023/02/06/2023-will-see-the-most-utility-scale-solar-added-in-a-single-year/, Feb 2023.

[36] Solar Energy Industries Association (SEIA). Major solar projects list. https://www.seia.org/research-resources/major-solar-projects-list.

[37] Åsmund Skomedal, Halvard Haug, and Erik Stensrud Marstein. Endogenous soiling rate determination and detection of cleaning events in utility-scale pv plants. *IEEE Journal of Photovoltaics*, 9(3):858–863, 2019.

[38] Greg P Smestad, Thomas A Germer, Hameed Alrashidi, Eduardo F Fernández, Sumon Dey, Honey Brahma, Nabin Sarmah, Aritra Ghosh, Nazmi Sellami, Ibrahim AI Hassan, et al. Modelling photovoltaic soiling losses through optical characterization. *Scientific reports*, 10(1):1–13, 2020.

[39] Elias Urrejola, Javier Antonanzas, Paulo Ayala, Marcelo Salgado, Gonzalo Ramírez-Sagner, Cristian Cortés, Alan Pino, and Rodrigo Escobar. Effect of soiling and sunlight exposure on the performance ratio of photovoltaic technologies in santiago, chile. *Energy Conversion and Management*, 114:338–347, 2016.

[40] Rupert Way, Matthew C Ives, Penny Mealy, and J Doyne Farmer. Empirically grounded technology forecasts and the energy transition. *Joule*, 6(9):2057–2082, 2022.

[41] Hamed Yazdani and Mahmood Yaghoubi. Dust deposition effect on photovoltaic modules performance and optimization of cleaning period: A combined experimental–numerical study. *Sustainable Energy Technologies and Assessments*, 51:101946, 2022.

APPENDICES

Appendix A

Heatmap Plots of PV Plant

The data readings for the PV plant are recorded individually for each SCB, allowing for per-SCB and area-specific results. For comparison between SCBs and between different areas of the plant site, we plot heatmaps (Figures A.1, A.2, and A.3) of thesis results, namely the latest day's efficiencies, the most recent soiling rates, and the average power production gains from cleaning. Note that due to missing data, results for some SCBs are not shown.

These heatmaps help visualize the **spatial variability** of soiling. With different areas of the plant experiencing different levels of soiling, efficiencies and cleaning production gains also differ from one SCB to another. By calculating per-SCB results, we account for this variability and allow for detailed insights leading to area-specific optimal cleanings.

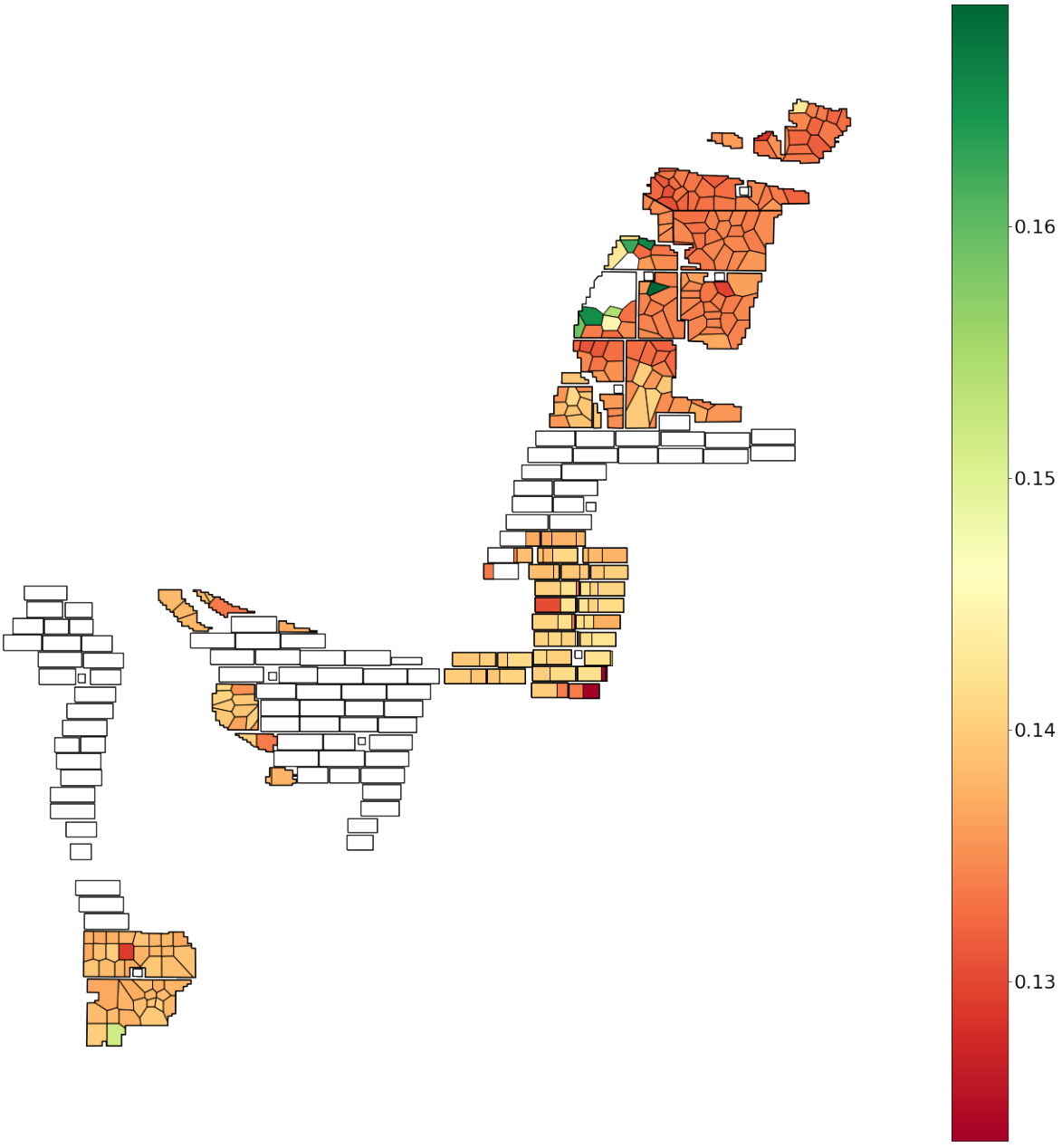


Figure A.1: Heatmap of Latest Efficiencies

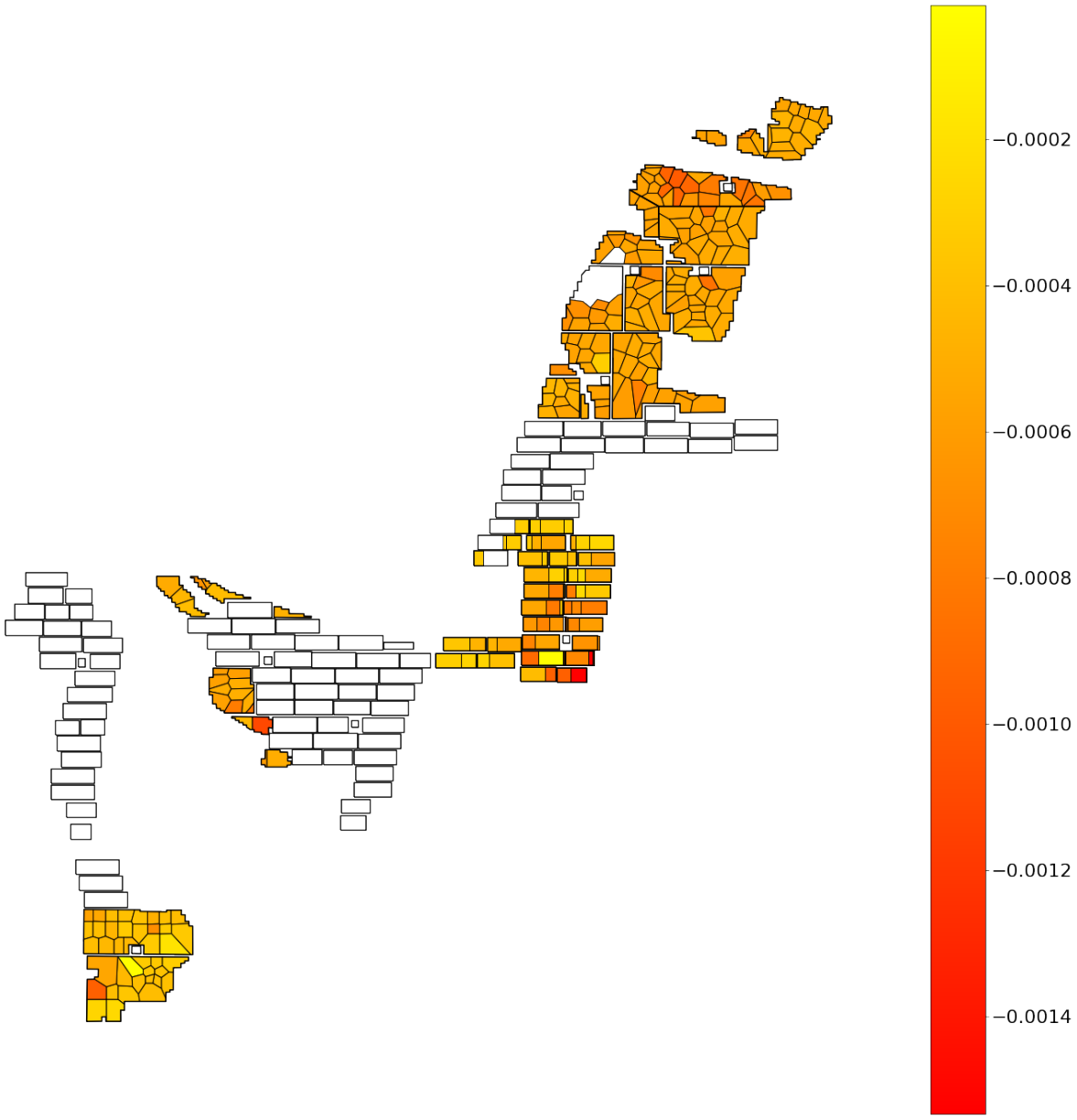


Figure A.2: Heatmap of Latest Soiling Rates

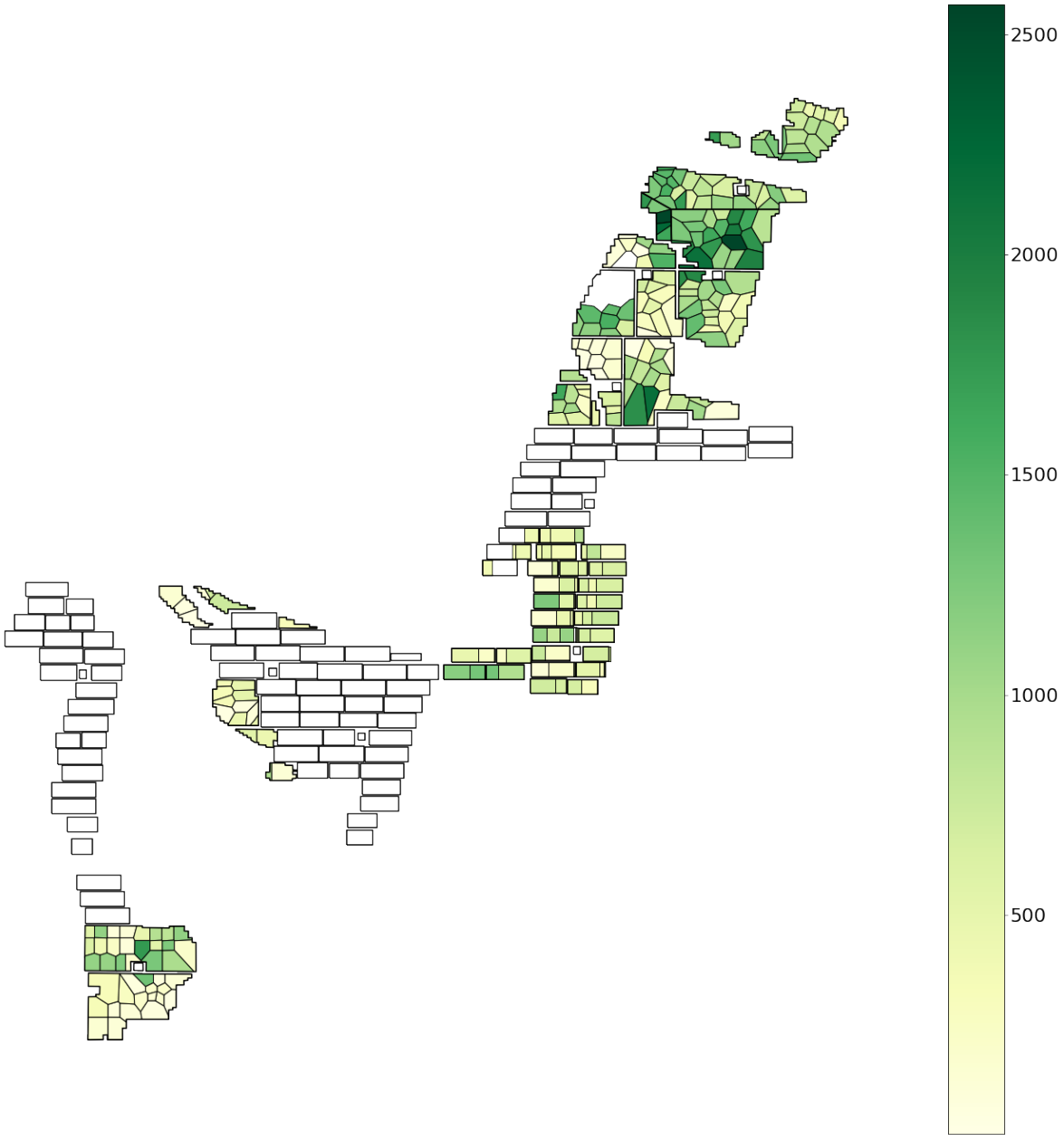


Figure A.3: Heatmap of Average Power Production Gains from Cleaning

Appendix B

Cleaning Log Excerpts

	date_only	meter_tag	rainy_day	cleaning_type	remarks
546	2020-11-28	BLOCK-8_INV-A_3_3	0	water cleaning	NaN
765	2020-12-08	BLOCK-8_INV-A_3_3	0	water cleaning	NaN
1274	2020-12-28	BLOCK-8_INV-A_3_3	0	water cleaning	NaN
1888	2021-01-16	BLOCK-8_INV-A_3_3	0	water cleaning	NaN
2033	2021-01-21	BLOCK-8_INV-A_3_3	0	water cleaning	NaN
2348	2021-02-03	BLOCK-8_INV-A_3_3	0	water cleaning	MCS Pipe Line Damaged due to Tractor Movement...
2478	2021-02-08	BLOCK-8_INV-A_3_3	0	water cleaning	MCS Pipe Line Damaged due to Tractor Movement...
3101	2021-02-18	BLOCK-8_INV-A_3_3	1	rain cleaning	Last Night Rain at Site
3503	2021-02-19	BLOCK-8_INV-A_3_3	1	rain cleaning	Rain at Site
3741	2021-03-06	BLOCK-8_INV-A_3_3	0	water cleaning	Cleaning effected due to borewell breakdown
4351	2021-03-23	BLOCK-8_INV-A_3_3	1	rain cleaning	Rain at Site
4523	2021-04-07	BLOCK-8_INV-A_3_3	0	water cleaning	Water Shortage in Borewell
4927	2021-04-11	BLOCK-8_INV-A_3_3	-1	rain cleaning	Rain at site
5330	2021-04-13	BLOCK-8_INV-A_3_3	-1	rain cleaning	Last Night Rain at Site

Figure B.1: Cleaning Records for SCB 454

Appendix C

Segmentation Algorithm Code

```
#segmenting timeline into upwards/downwards sections
timeline_dates = list(plot_df.date.values)
timeline_efficiencies = list(plot_df.efficiency.values)

#initial slope direction
current_window = [(timeline_dates[x],timeline_efficiencies[x]) for x in range(future_window_size)]
current_slope = slope_function(current_window)
current_direction = Segment.CLEANING if current_slope > 0 else Segment.SOILING

prev_date = timeline_dates[0] # tracks beginning of current segment (end of last segment)
i = 1 #index of current timeline point

#iterate through timeline and detect changepoints
while i <= len(timeline_dates) - future_window_size:
    #check slope of future window
    future_window = [(timeline_dates[i+x],timeline_efficiencies[i+x]) for x in range(future_window_size)]
    future_slope = slope_function(future_window)
    future_direction = Segment.CLEANING if future_slope > 0 else Segment.SOILING

    if not future_direction == current_direction:
        # take min/max of window as changepoint, because we may detect a change in slope before we reach the changepoint
        # so need to take min/max of the window where the change in slope occurs
        if future_direction == Segment.CLEANING and current_direction == Segment.SOILING:
            change_point = min(range(future_window_size), key=lambda x: future_window[x][1])
        elif future_direction == Segment.SOILING and current_direction == Segment.CLEANING:
            change_point = max(range(future_window_size), key=lambda x: future_window[x][1])
        final_segmentation_cs[meter_id].append(Segment(current_direction, prev_date, future_window[change_point][0]))
        prev_date = future_window[change_point][0]
        i += change_point
        current_direction = future_direction
        continue

    i += 1

final_segmentation_cs[meter_id].append(Segment(current_direction, prev_date, timeline_dates[-1]))
```

Figure C.1: Segmentation Algorithm Main Body

Full version of the code in a Jupyter Notebook can be found here.

```python
#Segment class to represent soiling or cleaning interval
class Segment:
    CLEANING = "CLEANING"
    SOILING = "SOILING"
    def __init__(self, trend, start_date, end_date):
        self.trend = trend
        self.start_date = start_date
        self.end_date = end_date
```

Figure C.2: Class Definition of a Segment

Appendix D

Gradual/Back-to-Back Cleanings

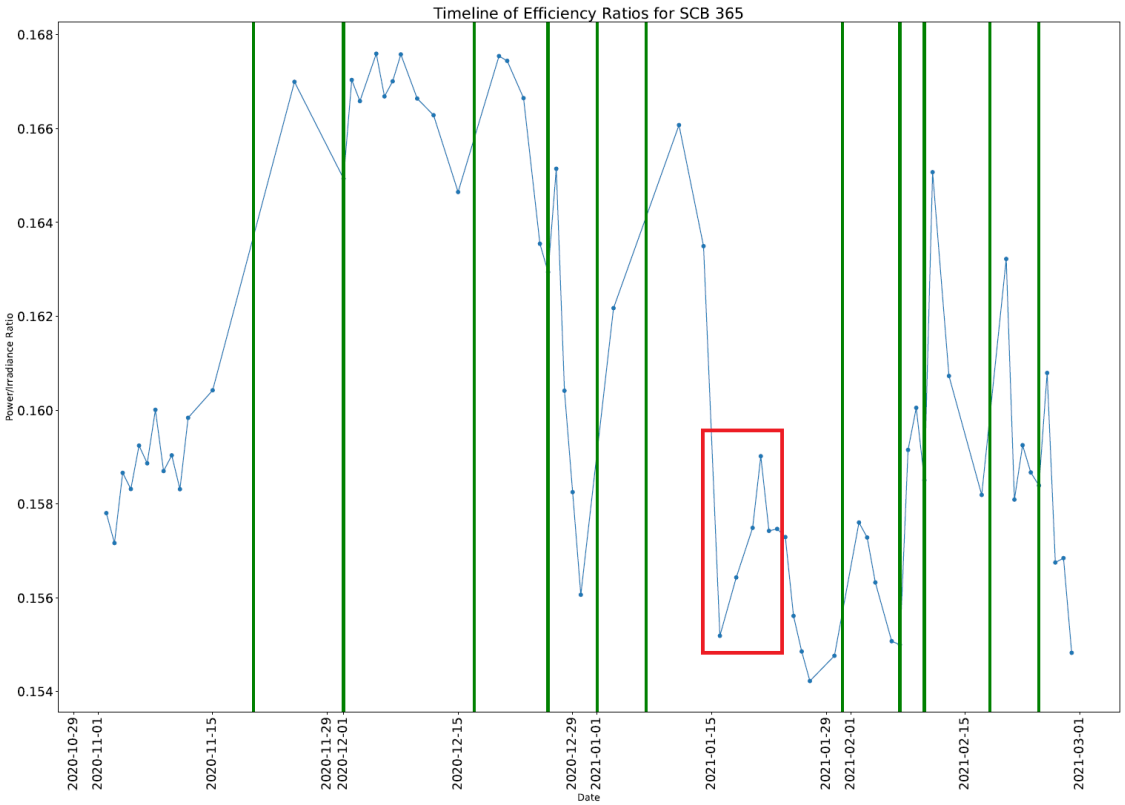


Figure D.1: Missed Detection of Cleaning Period

Appendix E

K-Means Clustering for Union Cleaning Records

Before developing the data-driven segmentation algorithm, we had made an attempt to fix the provided cleaning logs. While this attempt was ultimately unsuccessful, it is described here for reference.

During our analysis of the cleaning logs, we discovered many cases where the SCB efficiency improved with no record of a cleaning event. In some of these cases, we found records of cleanings corresponding to the efficiency improvement, but in the logs of neighbouring SCBs. This suggested that the crew may have cleaned a group of SCBs together, but only recorded the cleaning in one SCB's log. Now of course, this may not necessarily have been the case, but nevertheless it provided an avenue for further investigation. Specifically, if these cleaning groups could be identified, we could update the cleaning logs for those SCBs with missing records, adding the cleaning events which were only recorded for some SCBs in the group.

To determine potential cleaning groups, we analyzed the SCB efficiency timelines. The assumption here was that SCBs in the same cleaning group would have been cleaned at the same time and so their timelines would share patterns of efficiency improvement and decline. With this line of reasoning, we converted each SCB timeline into a vector of points (mean-normalizing the efficiency values) and applied **K-means clustering** to group SCBs with similar timelines. Some relevant clustering results are shown in Figures E.1 and E.2. Once the cleaning groups were determined, cleaning records from all SCBs in a group could be tallied to determine the **"union cleaning record"** for the group (Figure E.3).

While using clustering to leverage the cleaning records of neighbouring SCBs is intrigu-

ing, in practice there are several obvious issues. There is no guarantee that cleanings will be performed in groups and recorded for at least 1 SCB. Further, it may be the case that cleaning logs are not even maintained by the plant operators!

This is where a data-driven approach can come be utilized. Without requiring cleaning logs, data-driven techniques can detect cleaning events by analyzing trends in SCB efficiency. As such, this approach was left as an interesting exploration and focus was shifted to the segmentation algorithm described in Section 4.3.

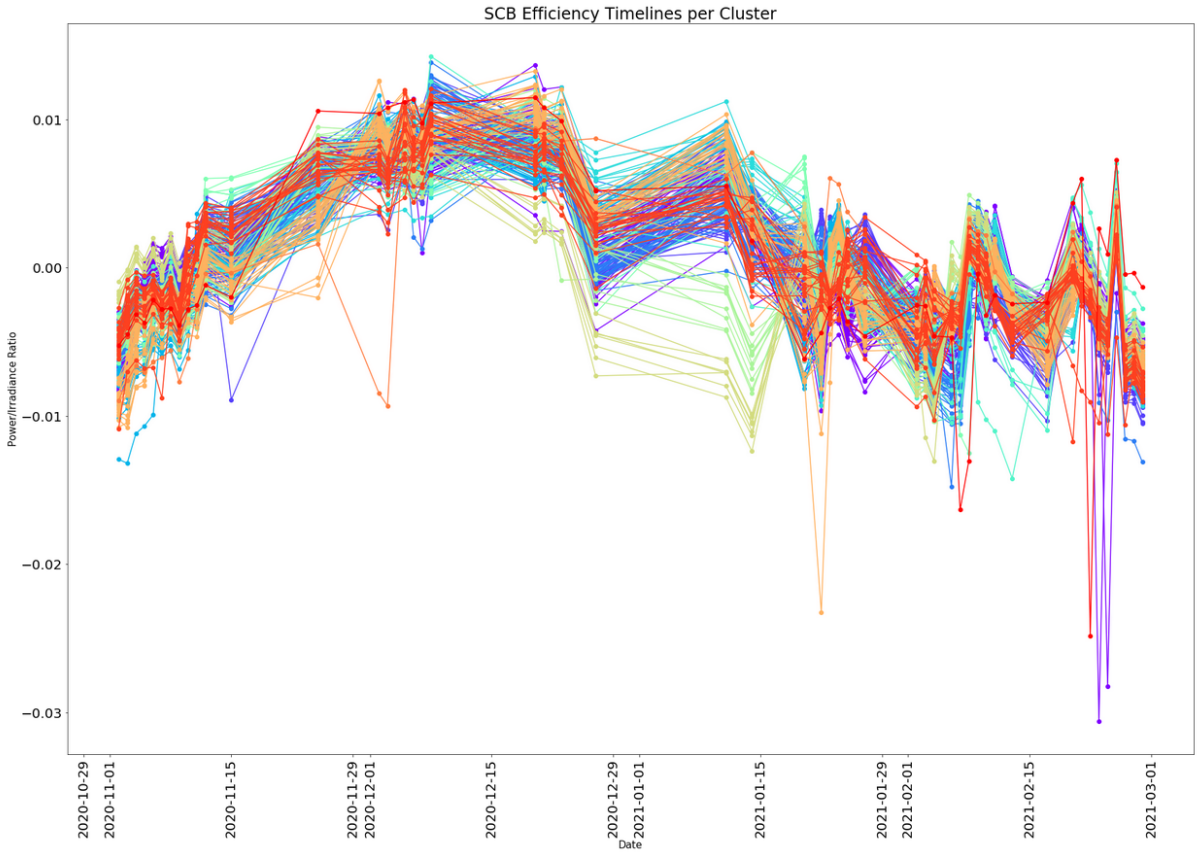


Figure E.1: Clustered Efficiency Timelines

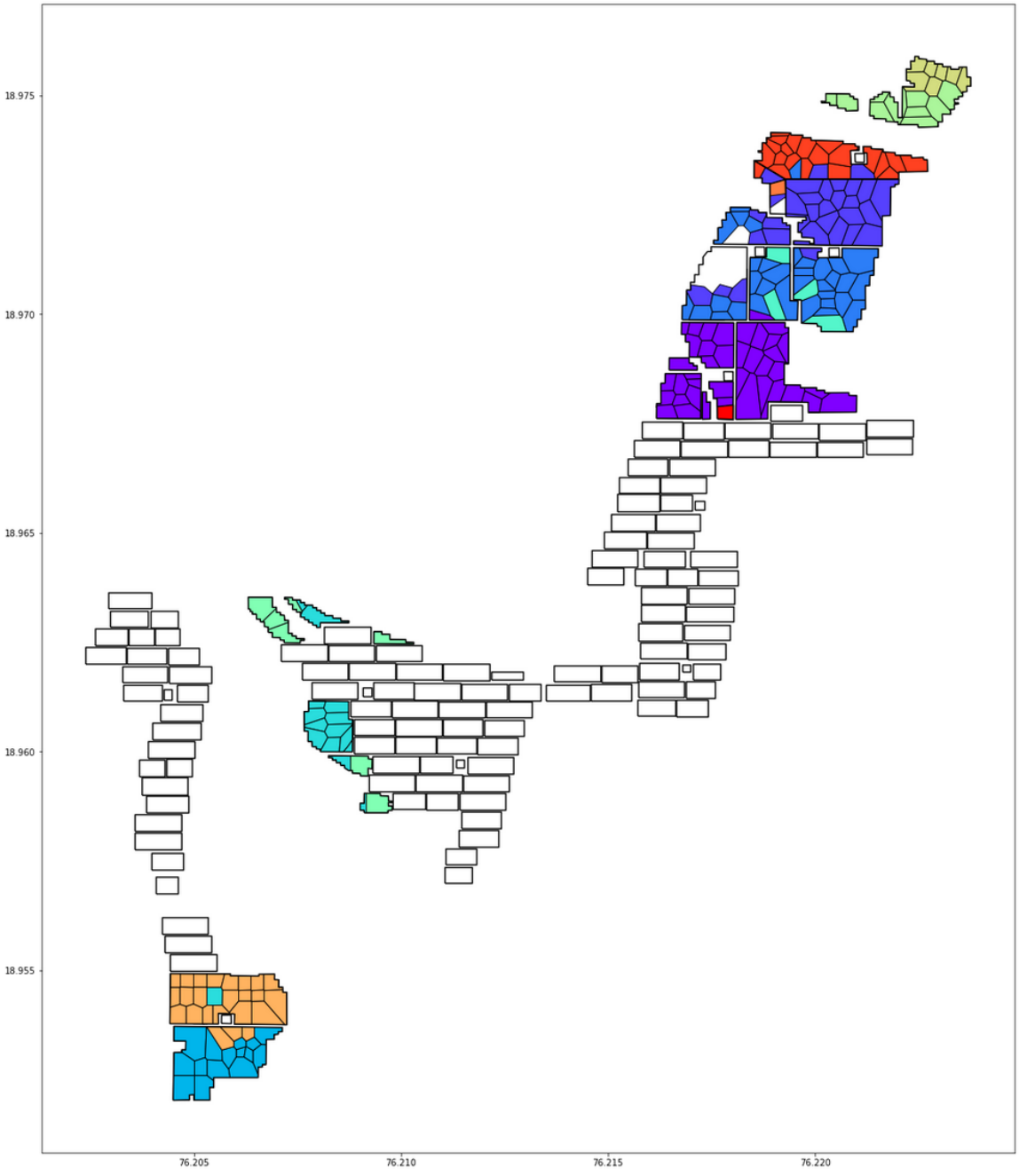


Figure E.2: Clustering Results Projected on Site Map

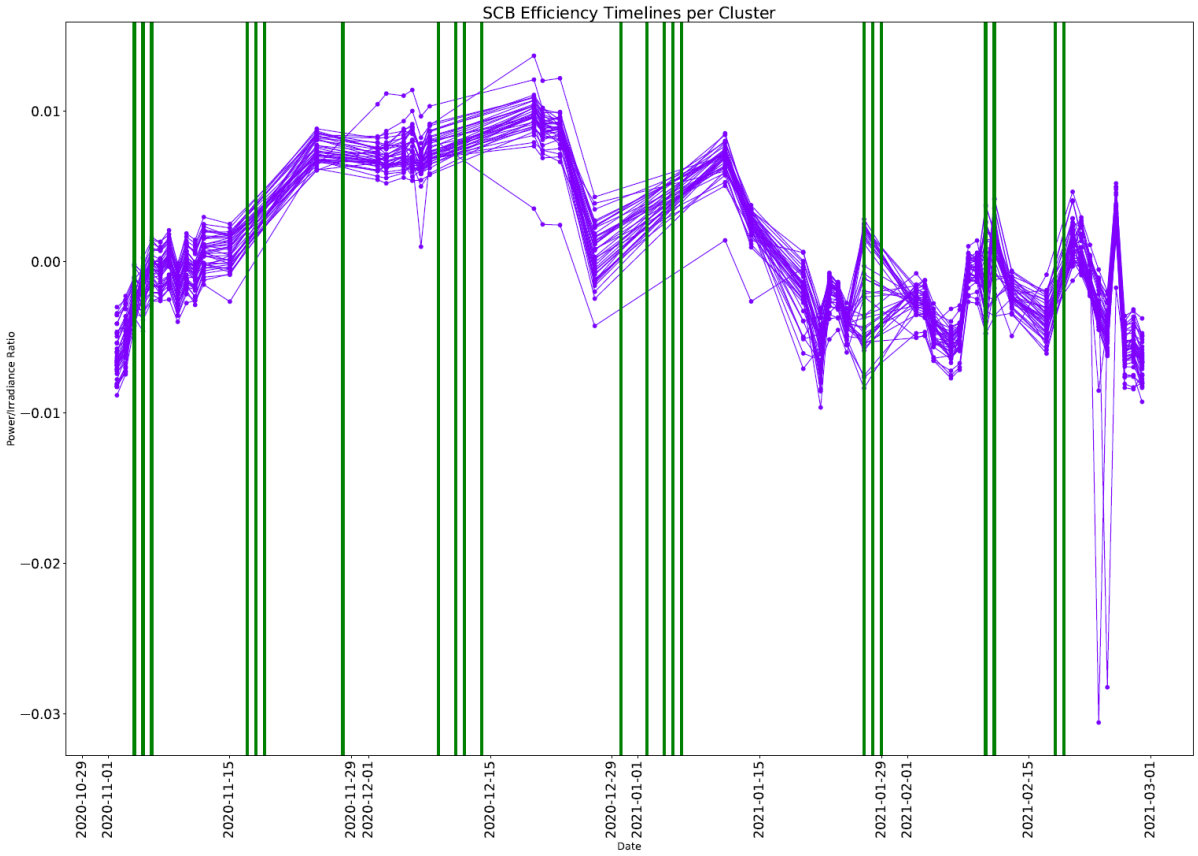


Figure E.3: Example of a Union Cleaning Record