Data Quality Issues in Solar Panels Installations: A Case Study*

Dumitru Roman
Antoine Pultier
SINTEF Digital
Oslo, Norway

Xiang Ma
SINTEF Industry
Oslo, Norway

Ahmet Soylu
Oslo Metropolitan University
Oslo, Norway

Alexander G. Ulyashin
SINTEF Industry
Oslo, Norway

ABSTRACT

Solar photovoltaics (PV) is becoming an important source of global electricity generation. Modern PV installations come with a variety of sensors attached to them for monitoring purposes (e.g., maintenance, prediction of electricity generation, etc.). Data collection (and implicitly the quality of data) from PV systems is becoming essential in this context. In this position paper, we introduce a modern PV mini power plant demo site setup for research purposes and discuss the data quality issues we encountered in operating the power plant.

CCS CONCEPTS

• **Information systems → Information systems applications**.

KEYWORDS

Solar panels, monitoring, data quality, data pipeline

ACM Reference Format:
Dumitru Roman, Antoine Pultier, Xiang Ma, Ahmet Soylu, and Alexander G. Ulyashin. 2022. Data Quality Issues in Solar Panels Installations: A Case Study. In *Proceedings of the 2nd International Workshop on Software Engineering and AI for Data Quality in Cyber-Physical Systems/Internet of Things (SEA4DQ '22), November 17, 2022, Singapore, Singapore.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3549037.3564120

1 SOLAR PANEL INSTALLATION AND MONITORING

As a major renewable energy source, solar photovoltaics (PV) [9] nowadays provide 3.1% of global electricity generation. PV monitoring is an essential part in any PV plant. Monitoring sensors and their working principles, controllers used in data acquisition systems, data transmission methods, and data storage and analysis technologies are very important in a monitoring system [3]. PV system monitoring may be the best way to maximize the performance of PV systems. However, each monitoring system affects in a different way the PV system performance [4]. PV monitoring systems have been proposed in the literature, e.g., based on open-source solutions with wireless and low-cost systems [5]. Others focus on the design and implementation of microcontroller based wireless PV modules [1]. Diagnostic techniques and algorithms were proposed to monitor photovoltaic plants, to predict failures and to enhance PV system performance [8]. Recognition Technologies (RT), Artificial Intelligence (AI), and Machine Learning (ML) enable drones and make the monitoring of large-scale solar power plants easier [2]. Data collection (and implicitly the quality of data) from PV systems is essential in this context, not only for better maintenance but also for better prediction of electricity generation.

A modern PV mini power plant demo site with 58 solar panels was installed on the roof of SINTEF building at Forskningsveien 1, Oslo, Norway, for research purposes, amongst others to collect and analyze the data from the PV plant and its associated sensors. Fig. 1.a shows a picture of the installation. To monitor the PV plant performance, various sensors are required. These include environmental sensors, data loggers, infrared cameras, etc. A CMP6 pyranometer (Fig. 1.b), a DustIQ soiling monitoring system (Fig. 1.c), and a ClimaVUE50 mini weather station (Fig. 1.d) are selected and installed for monitoring purposes. The Tigo system[1] is deployed to monitor the current, voltage and electricity output from each panel group through module optimizers and invertors. Thus, the information about radiation, dust related parameter, wind speed, environmental temperature, panel temperature, and power generation values can be monitored in real-time.

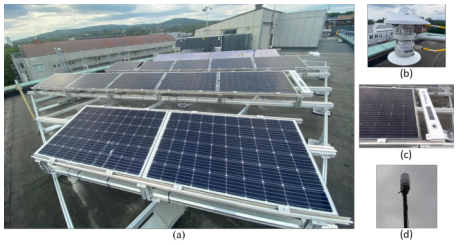

Figure 1: (a) PV demosite at SINTEF in Oslo, Norway; (b) CMP6 Pyranometer; (c) DustIQ Soiling Monitoring System; (d) Campbell Scientific CimaVUE 50 weather station.

Based on the large data that is collected, this demosite provides a unique opportunity to evaluate the power generation performance and explore the relation between different environmental variables which influences the energy output. For this purpose, a data pipeline was designed and implemented to collect and store data, and make it available for analysis. Fig. 2 depicts the data pipeline: data is collected from the sensors installed on the solar panels, as well as related sensors (e.g., from invertors, weather station), but also from external data providers (e.g., weather forecasting). Data that comes from proprietary systems (e.g., Tigo, SMA[2]) is firstly transmitted to corresponding proprietary cloud systems, after which it is downloaded, integrated/merged with the other data in the form of time

*This work received partial funding from the projects DataCloud (H2020 101016835), Super PV (H2020 792245), BigDataMine (NFR 309691), and SINTEF SEP-DataPipes.

SEA4DQ '22, November 17, 2022, Singapore, Singapore

ACM ISBN 978-1-4503-9459-8/22/11.
https://doi.org/10.1145/3549037.3564120

[1] https://www.tigoenergy.com
[2] https://www.sma.de

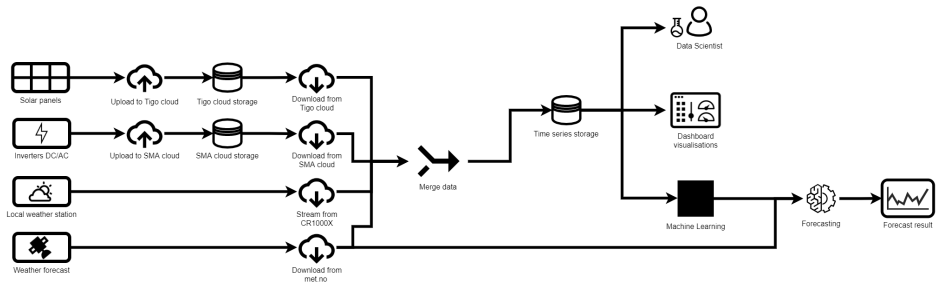

Figure 2: Data pipeline for collecting data from the PV installation.

series data, and stored for further analytics (ranging from basic visualization on dashboards to advanced AI/ML analytics). In the process of designing and implementing this pipeline we identified and experienced various data quality issues which we describe in the following section.

2 DATA QUALITY ISSUES

Missing Data. Missing data is a common problem that many reasons can cause. It can be hardware or software faults and last from seconds to weeks. We, for example, experienced the loss of electrical power for our data logger, which took a few days to fix. Our data is also streamed to the cloud using an IoT gateway and a broadband internet connection, which can be unreliable from time to time. The software could be missconfigured. For example, we kept only the last three months of data in the data logger at the beginning of the experiment, while we thought we kept everything forever. We also identified the risk of major hardware failures, such as broken cables or sensors. When data is missing for a very short period of a few sensors, interpolation can be considered. However, when the data is missing for longer periods, the data should not be used for analysis, especially for ML analytics. Removing the whole period is better than using default or dummy data.

Inconsistent Timing. The data sources in our systems have different sampling rates, from a few seconds to one hour, with clocks that are not necessarily synchronized tightly. We also experienced issues with time zones. Some sources also have delays in their availability. For example, a sensor uploads its data at a one-minute sampling rate, but only every half hour. We can use sensor fusion software methods to address such problems. We need to synchronize the timestamps, though it can be tricky to figure out the minor differences between the clocks. We also sometimes need to have a single sampling rate for all sensors, and we need to decide how we interpolate or sample the data. Average and linear interpolation are the primary solutions we use, but more advanced methods can also be used. For stream processing, we need to buffer the data for long periods, waiting for all our sensors to upload their data.

Unknown Condition. While an utterly broken sensor can be easy to spot, a sensor that produces inaccurate data can be challenging. Perhaps one temperature sensor fells from its solar panel to the floor while still reporting temperatures. One sensor may not be calibrated correctly or be replaced by another model with different characteristics. One classic solution in the big data domain is to use more data, so these issues are merely noise and could be ignored. In our case, we need to detect the quality changes using fault detection algorithms and careful data analysis.

Changes in the Experiment Environment. When doing an experiment outdoor for an extended period, years, for example, we should be prepared to observe significant changes in the environment. For example, we observed a new construction that obstructed the sun for a major part of the time for some of our solar panels. The panels could also be relocated or have their position adjusted. In our case, we decided to simulate the differences in sun exposure for the solar panels that are now mainly in the shade. But sometimes, the data should be dropped from the datasets.

Not Large Enough Experiment. We would like to have many years of data with many weather conditions in many locations, which would significantly increase the value of the data. One solution would be to share the data. People already share the energy production with little weather information on websites such as PVOutput[3]. Having a few more sensors in such community-sourced datasets would be valuable.

3 SUMMARY AND OUTLOOK

We introduced a modern PV mini power plant demo site and discussed the data quality issues we encountered in operating the plant. In future work we plan to identify and implement specific data pipeline solutions and strategies addressing the identified data quality issues [6, 7].

REFERENCES

[1] M Reyasudin Basir Khan et al. 2012. Wireless PV Module performance monitoring system. In *Proceedings National Graduate Conference 2012*. 1–4.
[2] Nallapaneni Manoj Kumar et al. 2018. On the technologies empowering drones for intelligent monitoring of solar photovoltaic power plants. *Procedia computer science* 133 (2018), 585–593.
[3] Siva Ramakrishna Madeti and SN Singh. 2017. Monitoring system for photovoltaic plants: A review. *Renewable and Sustainable Energy Reviews* 67 (2017), 1180–1207.
[4] Eneko Ortega et al. 2017. Study of photovoltaic systems monitoring methods. In *2017 IEEE 44th Photovoltaic Specialist Conference (PVSC)*. IEEE, 643–647.
[5] José Miguel Paredes-Parra et al. 2018. PV module monitoring system based on low-cost solutions: Wireless raspberry application and assessment. *Energies* 11, 11 (2018), 3051.
[6] Dumitru Roman et al. 2021. Big Data Pipelines on the Computing Continuum: Ecosystem and Use Cases Overview. In *Proceedings of the Symposium on Computers and Communications, 2021*. IEEE, 1–4.
[7] Ahmet Soylu et al. 2022. Data Quality Barriers for Transparency in Public Procurement. *Inf.* 13, 2 (2022), 99.
[8] Asma Triki-Lahiani et al. 2018. Fault detection and monitoring systems for photovoltaic installations: A review. *Renewable and Sustainable Energy Reviews* 82 (2018), 2680–2692.
[9] Marta Victoria et al. 2021. Solar photovoltaics is ready to power a sustainable future. *Joule* 5, 5 (2021), 1041–1056.

[3] https://pvoutput.org