Power Source Detection

1. Introduction

nLine's Gridwatch sensor reports data every two minutes. Each of these sensors reports data from health faclity connected to either solar, grid, generator, or some combination of these. The data reported is a timeseries dataset by each sensor indicates frequency and voltage, and whether power is on or off (and other metadata fields) for every two minutes. For the time slots in the timeseries where the power is off (i.e. outage), there is a ***null*** frequency value. The goal of the power source detection algorithm is to determine which power source is connected at any point in time which would effectively enable computation of solar, grid, and generator usage time in each site (i.e. health facility in SL). Understanding how much each power source gets used is important for economic planning of health facilities as well as monitoring the impacts of solarization projects. By understanding generator usage time, the fuel consumption can be computed hence the impact of solarization on the fuel consumption can be determined as an economic output of such development projects . We therefore to aim classify each sensor's data point as either ***solar***, ***grid***, or ***generator.***

2. Power Source Detection Algorithms

Previously, we have explored more automated approaches to detect the active power sources, including unsupervised ML algorithms and Kalman filtering. The ML model would be trained on the available sensor data features like frequency, voltage, and time, and augment them with derived variables through feature engineering. Additional inputs fed into the model could include time of day, known power source equipment at each location, and more. By training on the unlabeled timeseries data, the ML algorithm clusters the data points to find inherent patterns that map to solar, grid or generator states.
While these methods would be generalizable, due to some complexities that exist in the data, these methods would perform less robustly by underfitting the data. To overcome these limitations, we developed a lightweight algorithm based

primarily on expert knowledge and exploratory data analysis. In this post, we explain the methodology and some of the advantages of our approach.

3. Statistical Rule-Based Power Source Detection Algorithm

The core methodology focuses on analyzing frequency and feature-engineered variables from the sensor-reported frequency data. Based on exploratory data analysis, domain knowledge of the local grid, understanding of how different power source technologies work, and statistical methods, we are able to characterize the frequency and its features hence label the data points as solar, grid, or generator.

In the sections below, we describe the variables used for classification — frequency and other features derived from frequency — and explain the knowledge of local grid connection that help in the classification process.

3.1. Defining Variables for Classification

3.1.1. Frequency

This is the frequency value as reported by the sensor. The expected nominal frequency for a Sierra Leone grid is 50Hz. From our exporatory data analysis of the data collected so far from SL as well as knowledge of the working of inverter-based systems:

- Solar systems exhibit frequency with a tight distribution around 50Hz.
- Grid frequencies hover near 50Hz but have more variance.
- Generator frequencies vary widely, have a wider deviation from 50Hz nomina, with some staying as low as ~43Hz and some as high as ~57Hz.

With this in mind, we create frequency thresholding rules to distinguish between the power sources.

3.1.2. Frequency deviation from the electrically closest grid-only-connected facility

Frequency alone is not always enough, as the ranges can overlap at times, rendering the thresholding rule alone inadequate. To tighten up the logic around grid classification for locations where grid is one of the power sources, we use the frequency deviation from the electrically nearest grid-only connected facility. We determine the electrical proximity by eyeballing frequency profiles and analyzing the frequency patterns. Electrically proximal locations usually have the same frequency profile.

We compare the sensor-reported frequency to that of the current grid-frequency. The closer the frequency is to the current grid frequency, the more likely it is that it is a grid frequency (For a synchronized grid, frequency measured anywhere on the grid at any point in time should be the same value). We set a maximum deviation threshold of 0.11Hz between the sensor-reported frequency and the current grid frequency to classify the frequency as “grid”. We determine the current grid frequency by computing the median frequency from sensors in rooms that are grid-only connected, and connected to the same grid.

3.1.3. Frequency spread/variability of the sensor data within a time window

We use the mean absolute deviation (MAD) of the sensor frequency timeseries within a sliding time window. The time window used is experimental based on the exploratory data analysis we’ve conducted on the SL data and is currently ~15 minutes (3 data points before and 3 data points after the current data point). Solar has the lowest variability while generators have the highest variability. We use a threshold of 0.1 neighborhood MAD for solar.

3.1.4. Label of the combination of power sources in the room

3.2. Other Considerations for Classification - Ground Truth Knowledge

For the case of SL, there is some ground-truth knowledge that helps in both the classification steps and the fine-tuning steps to ensure correctness. The qualitative data collected during deployment labels each room where a sensor is deployed based on the power source it is connected to.

- ***Knowledge of grid-only connected rooms in health facilities*** - as described above, we used grid-only connected rooms in the facilities as grid-references. Some of the rooms are connected to grid only hence we know any data coming from these are from the grid and are used as grid reference.
- ***Knowledge of generator-only or solar-only connected rooms -*** While the use of the variables/features described above help in classification, some of the rooms are only connected to one power source so they are automatically labelled as such.

The use of variables defined above for classification is therefore most useful for rooms where there are multiple power sources in use.

3.3. Classification Steps

1. Label as outage if*:*
    a. the power is out
2. Label data point as grid if**:**
    a. not outage
    b. the power source is grid only **OR**
    c. one of the power sources is grid and the frequency deviation from the geographically closest grid median is ≤ 0.11
3. Label data point as solar if**:**
    a. not outage or grid
    b. frequency value is between 49.8 to 50.2
    c. MAD within the time window is ≤ 0.1
    d. not one of the rooms we know solar isn't installed.
4. Label all others as generator
5. Perform smoothing

3.4. Smoothing

Some smoothing steps are applied based on some combination of: post classification exploratory analysis, knowledge of the working of solar/grid/generator, ground truth knowledge based on survey data as well as interaction with health facility (HF) officials through our project manager, contextual understanding of the grid network in SL, and observation of historical data in SL.

1. At PCMH, only allow for generator label when grid is off.
2. For Ola During, when the grid is off, set to solar if the frequency is within solar threshold.

We continue to incorporate some smoothing steps (not included here) as we keep learning about the data, SL context, HF-specific knowledge, and making intuitive observations.

4. Limitations

While the methodology leverages domain knowledge and exploratory analysis to correctly classify power sources, additional rigor could be applied to improve replicability and precision. Supplementing the frequency data with additional sensor data may also remove reliance on informal intuitions. The main limitations are:

- The method is specific to current Sierra Leone hospitals where nLine sensor deployments are, making the method less transferable without adjustment.
- It relies heavily on the sensor frequency value, whereas additional data sources could improve accuracy.
- Smoothing steps are based on informal observations rather than statistically robust methodology.

Additionally, it does not discuss handling of uncertainty in classification or confidence scores. Assessing the overall accuracy is also not covered.

5. Possible Improvements

The current power source detection algorithm is effective to a large extent, but suffers the limitations stated previously. There are ways we can continue improving the methodology to make it even more effective:

- We can use this rule-based approach to reliably label new data over longer time periods. This ground truth historical dataset can then be used to train more advanced ML models in a supervised manner, rather than relying purely on clustering unlabeled data.
- Feature engineering steps used in our methodology also provide a solid starting point for predictive ML algorithms. Variables like frequency distribution statistics, variability, and geospatial correlations/grid distances would be useful features for a supervised ML method.
- We can run these new ML approaches in parallel to our current system and continuously check performance against our labeling to ensure consistency and correctness. We would therefore only rely on ML if the accuracy of the ML model is acceptable.

By combining our methodology with ML in this gradual fashion over time, we can evolve model generalization without risking incorrectnesss. The best practices learned creating this initial algorithm will guide our evaluation of ML improvements while still accounting for local knowledge. This will lead to robust and replicable power source detection for health infrastructure.