Data Science Approach to Time Series Analysis of Real-World PV modules

Yang Hu[1], Mohammad A. Hossain[1], Yifan Xu[2], Timothy Peshek[1], Jiayang Sun[2], Roger H. French[1]

[1]Solar Durability & Lifetime Extension Center, Material Science and Engineering, Case Western Reserve University, Cleveland OH, USA

[2]Center for Statistical Research, Computing and Collaboration (SR2c),Department of Epidemiology and Biostatistics, Case Western Reserve University OH, USA 44106

GREAT LAKES ENERGY INSTITUTE

SDLE — SOLAR DURABILITY AND LIFETIME EXTENSION CENTER

Motivation

Photovoltaics capacity exceed 100 GW in 2012

In the US, a solar project will be installed, on average ,every 4 mins.

Lifetime & Degradation Science for PV Modules

- Convened by U. S. DOE, Basic Energy Sciences

Current warranty 25 year

- Degradation rate R_d less than 1.0%/year

Qualification testing not sufficient for reliability

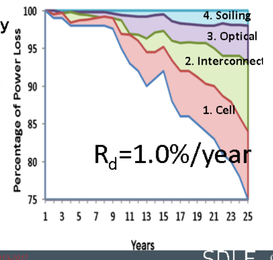
- Does not guarantee lifetime

Quantify influence of each stress in real-world

- Requires better understanding of degradation mechanisms
- Lead to more sensible testing protocol

Outdoor testing is vital !

- It is the typical environment PV systems work with
- It is the only way of correlating real-world degradation mechanisms to indoor accelerated test



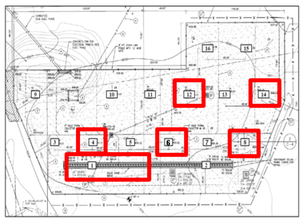
Case study of 60 modules on SDLE SunFarm

60 modules of 20 brands

- Distributed on 6 sites
- Three samples from each brand
- Same brand on the same site
- Brand name A- T

Observation period:

- November 25 ,2012
- To May 31, 2013



Purpose:

- Interpreting the information
- Develop a data cleaning and data munging procedure
- Integrate analytic procedure with Energy CRADLE
- Improve experimental design
- Evoke interests of further analysis

Analytical methods

- Raw data validation
- Exploratory data analysis
- Data assembly for performance metrics
- Subsampling of data
- Clustering

Snow covered PV module

Snowy days

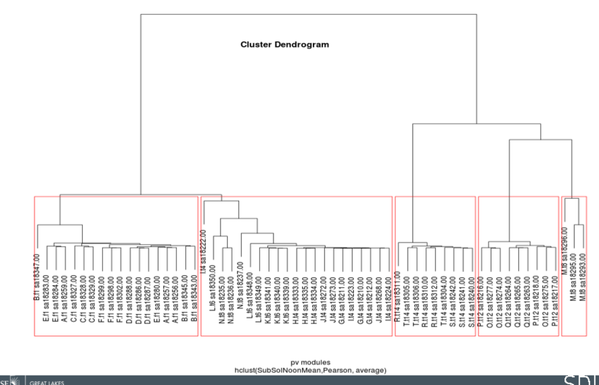
- Plot PR
- As a function of time
- Colored Snowy days data
- PR was very low or zero
- Snow cover mostly appear on fixed rack
- Also seen on stopped trackers

Snow cover

- Partially uncover help melting snow
- Arise thermal stress in PV module
- Potentially one of the degradation causes

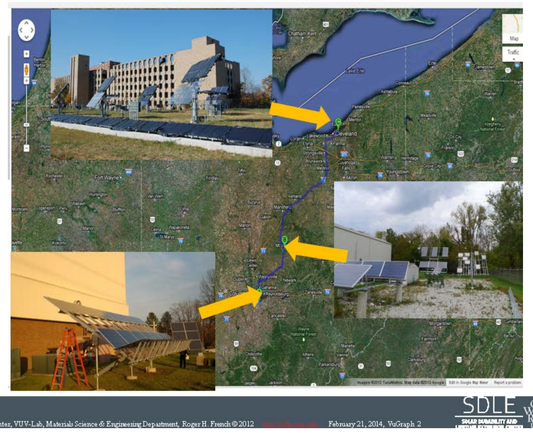
Hierarchical clustering of 60 modules

SubSolNoonMean , Pearson, Average



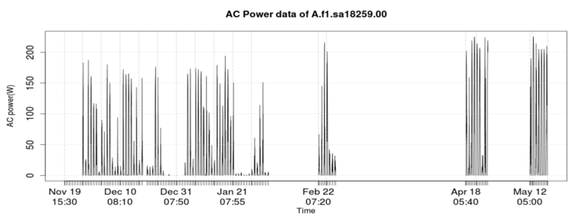
SunFarm Network

S-DLE Center SunFarm
Cleveland, OH
Lakeview 1MW SunFarm
Cleveland, OH
Under construction
Replex SunFarm
Mt. Vernon, OH
AEP Dolan Center SunFarm
Columbus, OH
Q-Lab (Existing)
Arizona
Bwh
Florida
Aw
Taiwan - UL
Taitung
Lujhu
India – Gujarat
IIT gandhinagar



Raw data validation

1. **Baseline I-V characteristics**
   - SPIRE solar simulator 4600
2. **Power data**
   - Enlighten micro-inverter user interface
   - Three "gaps" in data collecting due to unexpected grid disconnection and human error
   - 99 days' data out of 180 days power data was collected.
3. **Weather data**
   - Global Horizontal Irrdiance (GHI) data from Nov. 2012 to May. 2013 Sampling rate 1 min
   - Covert to plane of array (POA) irradiance
4. **Data alignment**
   - Alien power data and weather data



Snow covered irradiance sensor

Irradiance sensor cross check

Sensor 1:

- Horizontal
- Sampling rate 1min
- Black curve

Sensor 2:

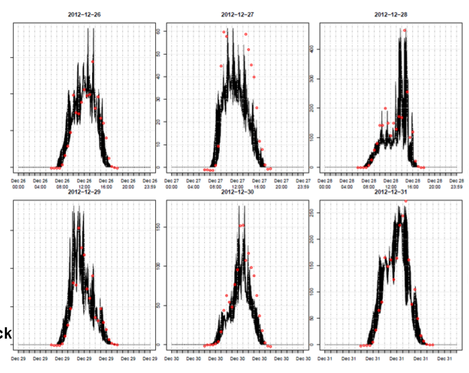
- Horizontal
- Sampling rate 30mins
- Red dots

Data alignment

Visually determine

- "2012-12-27"
- "2012-12-28"
- "2012-12-30"
- "2013-01-04"
- "2013-01-25"

Redundant sensors and sensor cross check are needed !

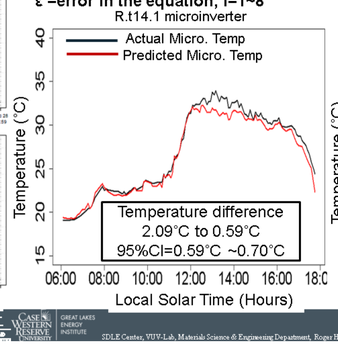


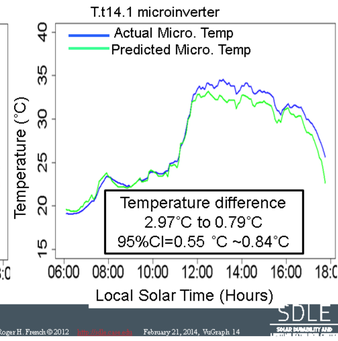
Predictive Model based on Linear regression

A linear regression model was used to develop an equation to predict the microinverter temperature

$$Micro.T_i = \sum_{j=1}^{8}\beta_{0j}x_{ij} + (\sum_{j=1}^{8}\beta_{1j}x_{ij})Ambient.T_i + (\sum_{j=1}^{8}\beta_{2j}x_{ij})Module.T_i + (\sum_{j=1}^{8}\beta_{3j}x_{ij})Irradiance_i + (\sum_{j=1}^{8}\beta_{4j}x_{ij})Power_i + \varepsilon$$

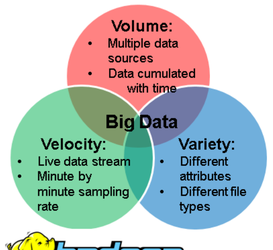
ε =error in the equation; i=1~8





Data Science Approach

Big Data



hadoop

- Open-source software framework
- Storage and large scale processing of data
- On clusters of commodity hardware

Energy-CRADLE

- Data Sources
  - SunFarm Network
- Data Acquisition
  - Local pre-storage
- Data Processing
  - Hadoop Distributed File System (HDFS)
  - HBase non-relational, distributed database
- Data Representation
  - Energy-MiMi, web interface

R language

- Free software programming language
- Statistical computing and graphics
- Well developed packages for statistical analysis

Exploratory data analysis on integrated performance

Baseline

Evaluation of each module

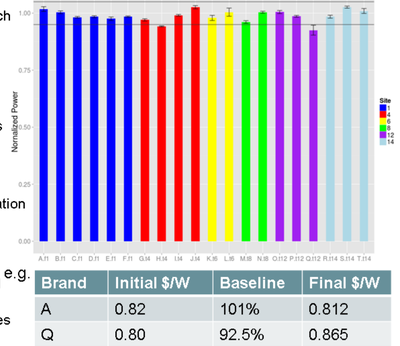
- 16 I-V curve were taken for each module
- Maximum power output(P_{max})
- Corrected to Standard Test condition
- According to IEC 60891
- Deviations of 16 measurements between 0.04%-0.9%

Evaluation of 20 brands

- Means of three samples
- +/- 5%common market expectation
- Brand H and Q didn't reach

Help decision making

- Power plant owner
- Module price based on nominal power
- Deviation also influence modules performance in strings



e.g.

Brand	Initial $/W	Baseline	Final $/W
A	0.82	101%	0.812
Q	0.80	92.5%	0.865

Hierarchical clustering of 60 modules

Data set:

Subseted SolarNoon Mean

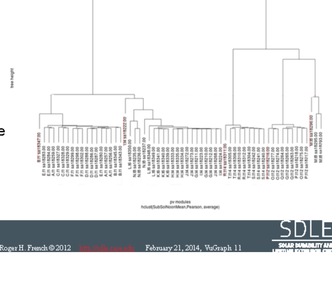
- PR time series data of 60 modules
- Subtract snow covered days
- There are 85 days in common

Distance metric:

- Euclidean, 'ordinary' distance in 85 dimensions
- **Pearson, Pearson Correlation measures the similarity in shape between two profiles.**
- Dynamic time warping, measuring similarity between two temporal sequences which may vary in time or speed. Widely used in Time series analysis
- Data was aligned, Euclidean, Pearson are applicable

Linkage criteria:

- Distance between sets of observations as a function of the pairwise distances
- Single : shortest distance
- Complete : longest distance
- **Average: average distance**



Conclusions

Enabled remotely diagnostic of PV system performance in the field

- Will be applied to Energy CRADLE

Data analytics procedure was built

- Including raw data validation
- data alignment
- sensor data cross check
- data assembly
- data sub-sampling

Data clustering

- Observed potential groups in data, conformed with clustering analysis.
- Solar noon time PR found malfunctioning brands and trackers
- Distinguished fixed rack and trackers

Future Work

Improve SunFarm metrology and study design

- Irradiance metrology and redundant sensors

Predictive model

- Identifying major contribution stressors to the over all system performance lost
- Feature selection

Time Series Analysis (TSA)

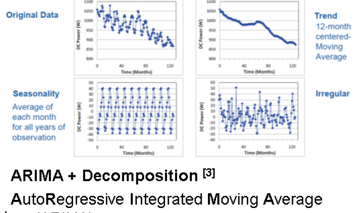
Continuous data monitoring:

- **Photovoltaic for Utility Scale Application(PVUSA)**
  - Power data, ambient T, wind speed
  - Estimated module temperature
  - Correct to PVUSA test condition (PTC)
- **Performance Ratio (PR)**
  - Compare real-world performance to performance under standard test condition(STC)
  - Quantitatively compare the system performance of systems with different configurations at different locations
- **DC/POA Temperature corrected**
  - "Total uncertainty fluctuates somewhat from dataset to dataset –DC/POA Temp-corr best performing"[1]

Point in time measurements:

- **Indoor I-V curve**
  - With indoor solar simulator
- **Outdoor I-V curve**
  - With portable solar simulator
  - Both methods need to disconnect module from array

Classical Decomposition [2]



ARIMA + Decomposition [3]

AutoRegressive Integrated Moving Average (ARIMA)

References:

[1] Jordan, Dirk C., and Sarah R. Kurtz. "Photovoltaic degradation rates—an analytical review." *Progress in Photovoltaics: Research and Applications* 21.1 (2013): 12-29

[2] S.G. Makridakis et al., "Forecasting", New York, John Wiley & Sons 1997.

[3] Jordan, Dick. *Methods of Analysis of Outdoor Performance Data.* National Renewable Energy Laboratory, 2011

Data assembly

Performance metrics

- In order to quantitatively compare the system performance of systems with different configurations at different locations,

Normalized instantaneous quantities

$Y_I = POA/G_0, G_0 = 1kW/m^2$

$Y_{DC} = P_{DC}/P_0$

$L_C = Y_I - Y_{DC}$

$Y_{AC} = P_{AC}/P_0$

$L_S = Y_{DC} - Y_{AC}$

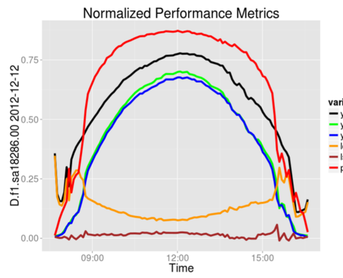
$PR = Y_{AC}/Y_I = (P_{AC}/POA)/(P_0/G_0)$

Solar time

- Convert performance metrics
- From local time to solar time

Data filtering

- PR of +/- 15 min solar noon time
- Seven data points for each

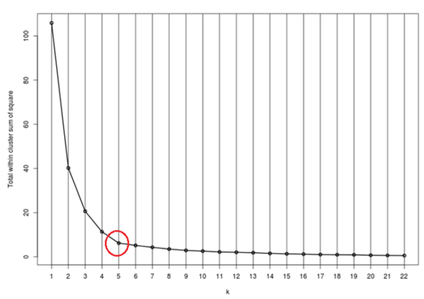


Kmeans Clustering

Number of clusters k.

Within Cluster Sum of Squares (WCSS)

- Assign each observation to k clusters such that the total WCSS is minimized.
- Plot WCSS as a function of K
- WCSS always decrease
- "Elbow point"
- Adding more cluster doesn't give a much better modeling of data



K-mean clustering indicate that the number of clusters is 5

Partition the hierarchical clustering result into 5 groups

Acknowledgement

SDLE center was established through funding through the Ohio Third Frontier, Wright Project Program Award tech 12-004

Bay Area Phocovoltaic Consortium Prime Award No. DE-EE0004946, Subaward Agreement No. 60220829-51077-T.

Department of Material Science and Engineering:

SDLE center :

Mohammad A. Hossain

Daniel Dryden

Nicholas Wheeler

Dr. Timothy Peshek

Prof. Roger French

Department of Epidemiology and Biostatistics

SR2$_c$ Center:

Dr. Yifan Xu , Prof. Jiayang Sun

Department of Electrical Eng. & Computer Sci.

Division of Biomedical Informatics:

Yashwanth R. Gunapati, Yi Hou,

Prof. G.Q. Zhang

this presentation contains no confidential information