FAIRification, Quality Assessment, and Missingness Pattern Discovery for Spatiotemporal Photovoltaic Data

William C. Oltjen[1], Yangxin Fan[1], Jiqi Liu[1], Liangyi Huang[1], Xuanji Yu[1], Mengjie Li[2], Hubert Seigneur[3], Xusheng Xiao[1], Kristopher O. Davis[2], Laura S. Bruckman[1], Yinghui Wu[1], Roger H. French[1]

[1]Solar Durability & Lifetime Extension Research Center, Case Western Reserve University, Cleveland Ohio, USA
[2]University of Central Florida (UCF), Orlando, Florida, 32816, USA
[3]Florida Solar Energy Center (FSEC), Cocoa, Florida, 32922, USA

Objective and Background

With access to a large number of time series data sets from many different power plants, it is important to focus on methods that can both speed up and improve our analysis

- **FAIRification** facilitates the ingestion and analysis of data
- **Data quality assessment** speeds the rate at which we can find high-quality data most suitable for modeling
- **Missingness Imputation** enables modeling on lower quality data sets

Data Source

The data used for this study is from the Sunsmart Schools set

- Mass deployment of solar with battery back -up
  - 15 minute interval time series data
  - Generally have a length of about two years
  - Can be as long as nine years

28 time series data sets from schools located throughout Florida

- Data sets include information about
  - Irradiance, Temperature, Power, and more

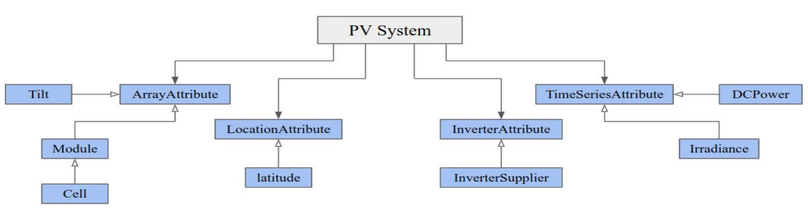
FAIRification and Ontology Development

Need to improve our ability to share data between groups

Do this by following the “FAIR” principles
- Findable
- Accessible
- Interoperable
- Reusable

Establish standards for our metadata through the creation of an ontology

Ontology is a formal dictionary of terms that also includes information on how those terms are related



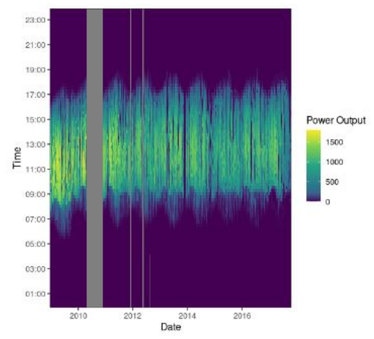
By including relationships between terms, we can navigate those relationships in order to gain new insights about data

Data Quality Assessment

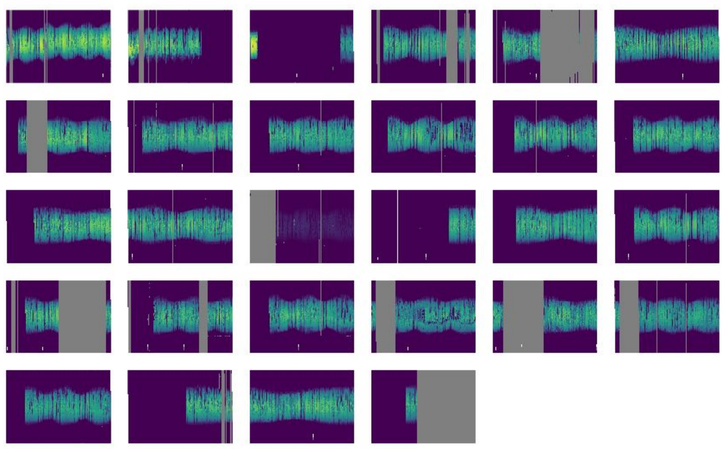
Missing data compromises statistical analysis and contributes to significant bias in the results

Need for rapid analysis of missing data

Heat maps provide a necessary visual
- Improve human interpretability
- Clear indication of missingness



Heatmaps can be easily applied at scale for a wide range of PV systems
- Missing data appears as gray, purple, or white blocks
- Data is colored in by the intensity of the power produced
- Seasonality in the data is apparent



Data Grading

Site	Outlier Percentage	Missingness Percentage	Longest Missing Gap	Length Requirement
1	B	A	C	P
2	B	A	D	P
3	B	A	A	F
4	B	B	D	P
5	A	C	C	P
6	B	A	A	F
7	B	B	D	P
8	C	A	A	P
9	B	A	A	P
10	B	A	A	P
11	B	A	A	P
12	C	A	A	P
13	C	A	A	F
14	B	A	A	F
15	B	B	D	P
16	B	A	A	F
17	B	A	A	F
18	B	A	A	P
19	A	D	D	P
20	B	A	C	P
21	B	A	A	P
22	B	B	D	P
23	B	C	D	F
24	B	B	D	P
25	B	A	A	P
26	B	A	A	F
27	C	A	A	F
28	A	D	D	F

Need for a quantitative measurement of data quality

- Assign letter grades based on:
  - Percentage of outliers
  - Missingess percentage
  - Longest missing gap
  - Length requirement

Can automatically sort data based on quality metrics

Letter grade	Outliers (%)	Missing percentage (%)	Longest gap (days)
A	Below 10	Below 10	Below 15
B	10 to 20	10 to 25	15 to 30
C	20 to 30	25 to 40	30 to 90
D	Above 30	Above 40	Above 90

Missingness Imputation

Linear Interpolation
- Linear fit using the two values before and after the missing block

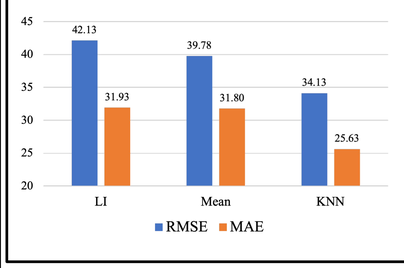
Mean Interpolation
- Fills the missing values with the mean of the overall data column

K-Nearest Neighbors
- Imputes data using an average of the K nearest neighbors

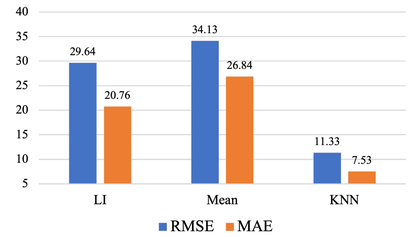
Spatiotemporal Graph Neural Network
- Utilizes spatial coherence of nearby PV systems to impute data
- Goal for future work

KNN provides the best imputations between these models
- Indicates the feasibility for the use of an st-GNN

Power



Module Temperature



Results

FAIRification
- The creation of a solar power plant ontology enhances both the standardization of metadata terms and the analysis capabilities of our data
- This process improves how data is shared both between organizations and within our own

Data Quality Assessment
- Data grading and quality heat maps provide the most efficient routes to understanding missingness in a data set

Missingness Imputation
- KNNs make use of neighboring power plants to provide the most accurate imputations, indicating the usefulness of an st-GNN

References
- M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, and A. Baak, “The FAIR Guiding Principles for scientific data management and stewardship,” Scientific Data, vol. 3, no. 1, pp. 1–9, Mar. 2016
- D. Moser, D. Bertani, A. J. Curran, R. H. French, M. Herz, and S. Lindig, “International Collaboration Framework for the Calculation of Performance Loss Rates: Data Quality, Benchmarks, and Trends,” in 36th European Photovoltaic Solar Energy Conference and Exhibition, Oct. 2019, pp. 1266–1271
- A. Curran, T. Burleyson, S. Lindig, D. Moser, R. French, and SDLE Research Center, “PVplr: Performance Loss Rate Analysis Pipeline,” Oct. 2020
- Willam C. Oltjen, Liangyi Huang, and Roger H. French, “FAIRmaterials: Make Materials Data FAIR,” Sep. 2021
- Roger H. French, Liangyi Huang, William C. Oltjen, Arafath Nihar, Jiqi Liu, Justin Glynn, and Kehley Coleman, “Fairmaterials,” Oct. 2021
- T. Kim, W. Ko, and J. Kim, “Analysis and Impact Evaluation of Missing Data Imputation in Day-ahead PV Generation Forecasting,” Applied Sciences, vol. 9, no. 1, p. 204, Jan. 2019
- A. R. T. Donders, G. J. Van Der Heijden, T. Stijnen, and K. G. Moons, “A gentle introduction to imputation of missing values,” Journal of clinical epidemiology, vol. 59, no. 10, pp. 1087–1091, 2006.
- A. Nihar, A. J. Curran, A. M. Karimi, J. L. Braid, L. S. Bruckman, M. Koyuturk, Y. Wu, and R. H. French, “Toward Findable, Accessible, ¨ Interoperable and Reusable (FAIR) Photovoltaic System Time Series Data,” in IEEE 48th PVSC Proceedings, Jun. 2021

Acknowledgements
This material is based upon work supported by the U.S. Department of Energy’s Office of Energy Efficiency and Renewable Energy (EERE) under Solar Energy Technologies Office (SETO) Agreement Number DE-EE0009347 and DE-EE0009353.
The views expressed herein do not necessarily represent the views of the U.S. Department of Energy or the United States Government.
This work made use of the Rider High Performance Computing Cluster in the Core Facility for Advanced Research Computing at Case Western Reserve University