

A Framework for Comparison of Spatiotemporal and Time Series Datasets

David Biagioni, Brian Bush, Ryan Elmore, Dan Getman, Danny Inman

Goals, Plans, Impacts

Goal: To develop statistical methods for the cross-comparison and relative-quality evaluation of datasets focused on the interpretation of analytical results and the validation of modeled data through comparison to known or source datasets.

Impact: Results from this project have the potential to routinely add value to a wide range of projects across most, if not all, NREL centers.

- Developing a methodology to routinely apply techniques for the cross-comparison and relative-quality evaluation of large spatiotemporal datasets constitutes a significant and novel contribution to energy data science.
- Increased visibility in the data science community will foster the development of high-value collaborations with academic institutions, enlarge the energy-data territory that NREL.
- The addition of such a capability to NREL can be harnessed in marketing sophisticated, complex analysis projects to multiple sponsors, putting NREL another step ahead of competitors.

Next Steps: This research is being undertaken in a series of phases that will ensure its availability and applicability to researchers at NREL. Currently, we are completing the initial assessment of potential statistical and computational methods and moving into the second phase of research in which those methods are applied.

This will involve several in-depth analyses focused on

- Comparing multiple spatiotemporal datasets with overlapping coverage in space and time
- Comparing time series between different spatial locations within the same spatiotemporal datasets
- Comparing spatial data between multiple time slices within the same spatiotemporal dataset
- Comparing multiple time windows within one or more temporal datasets
- Comparing irregularly spaced point datasets to gridded data representing similar parameters.

Objective

- Develop a non-parametric approach for making inter- and intra-dataset comparisons of large resource datasets.
- Time series were assessed for normality.
- Box-Cox was used to transform the data to quasi-normal.
- Temporal dependency was removed from the Box-Cox transformed data using autoregression.
- Residuals were compared using Canonical Redundancy Analysis (RDA).
- Non-parametric significance tests were performed using bootstrapping.

Illustrative Example

- National Solar Resource Database (NSRDB).
- Are stations within a region the same over a given time period?
- Are stations within a region the same over a given time period?
- Results of the RDA analysis suggest that three of the four stations have strong similarities in their underlying data structure.
- This approach may be applied across a range of data sets, including comparing multiple parameters across disparate datasets.

Comparing Datasets using Non-Parametric Methods

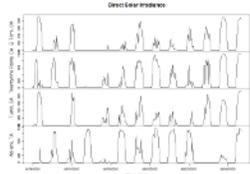


Figure 2. Direct solar irradiance for four NSRDB stations in the Western US. An autoregressive model was used to remove the effects of autocorrelation on the data. From this plot, it is evident that a strong underlying structure exists in the data.

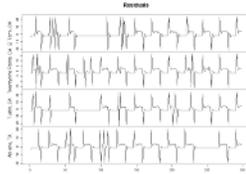


Figure 3. Residual plots of four NSRDB stations in the Western US. An autoregressive model was used to remove the effects of autocorrelation on the data. From this plot, it is evident that a strong underlying structure exists in the data.

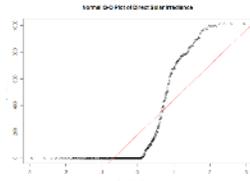


Figure 4. Quantile-quantile (Q-Q) plot for the Twenty Nine Palms, CA station. Comparing the sample quantiles (open circles) to the theoretical normal quantiles (red line) the data do not follow a normal distribution.

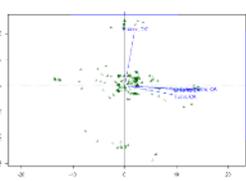


Figure 5. Biplot of the redundancy analysis (RDA). The green triangles represent the fitted site scores of the dissimilarity measures and the blue lines represent the station data. The two axes are the first (x) and second (y) principal coordinates. The cosine of the angle between any two stations (blue lines) represents the correlation.

Exploration of Simultaneous Variability using Principal Component Analysis

Objective

To develop a method of deriving where, when, and at what aggregation level we can detect and express correlation between the variability of two parameters within a large spatiotemporal dataset.

Approach

- Using NSRDB Stations data, calculate variability for wind speed and solar irradiance at daily, weekly, monthly, and annual aggregation levels.
- Calculate a difference in the simultaneous variability at each aggregation level.
- Perform PCA on the difference value at each aggregation level, including monthly PCA calculated using the daily values.
- Calculate the variability of the difference values across each aggregation level.
- Classify results of the PCA based on the variability and maximum values of the difference calculations.

Results

Stations with low contributions to the first 10 eigenectors demonstrate lower variability, range, and maximum values of the difference between variability of wind speed and solar irradiance (DIR). Stations with higher contributions demonstrate higher values in these parameters.

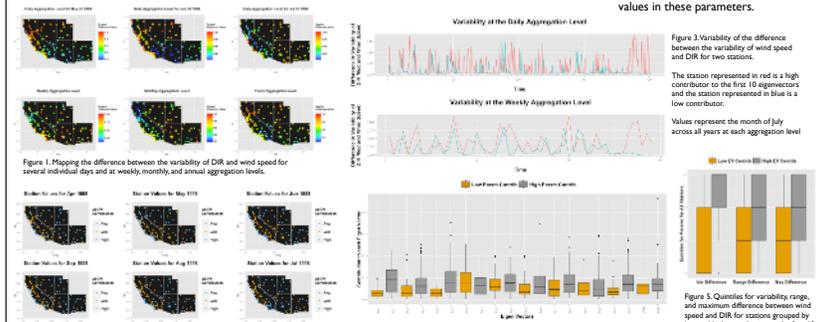


Figure 1. Mapping the difference between the variability of DIR and wind speed for several individual days and at weekly, monthly, and annual aggregation levels.

Figure 3. Variability of the difference between the variability of wind speed and DIR for two stations. The station represented in red is a high contributor to the first 10 eigenectors and the station represented in blue is a low contributor. Values represent the month of July across all years at each aggregation level.

Figure 2. Mapping the contribution of each station to the first 10 Eigenectors.

Figure 4. Contribution of all stations to the first 10 eigenectors grouped by high and low values representing the variability, range, and maximum difference between wind speed and DIR.

Non-Parametric Detection of Bias in Diffuse Solar Irradiance

Motivation

- Non-parametric statistical methods combine power with robustness.
 - Accessible and safe for non-statisticians
 - Apply broadly to many types of EERE data
 - Don't require strong assumptions about data
 - Avoid false positives
 - Perform nearly as well as parametric methods
 - Well suited for rapid calculation in distributed computing environments

Approach

- Our research focuses on multidimensional applications of non-parametric methods, particularly those with spatio-temporal extensions.
- Datasets comparison often fits into a common data-processing pattern.
- Such data-processing patterns can be implemented in HPC environments as a map-reduce operation.

Example (see diagram to the right)

- Study Question:** How do NSRDB station's irradiance "measurements" (dataset #1) compare with those measurements from the same hour of the same day five years previously (dataset #2)?
- Results of Sign Test:** There is a statistically significant bias detectable between these datasets in many geographical regions.
- Diagnosis:** There is a several W/m² bias of the dataset #1 relative to dataset #2.



Evaluation of Data Formats for Statistical Analysis of Large Matrices on HPC

Objective

To determine the best storage and analysis data format for statistical analysis of large matrices on high performance computing resources

Approach

Part of our preliminary experimentation involved running benchmarks to see how standard analyses would scale on Peregrine. The following results show the average time to complete principal component analyses on data sets of varying size (7.6 MB, 76MB, 763 MB, and 7.45 GB), differing numbers of cores (96, 144, 192, 240, and 288) and cores per node (16 vs 24). The left panel shows the timings for all combinations and the right panel shows the speedups gained in using 144, 192, 240, and 288 cores relative to using 96 cores.

