Distributed Statistical Computation for Comparing Massive Spatiotemporal Datasets

D. Biagioni, B. Bush, R. Elmore, D. Getman, R. Inman

National Renewable Energy Laboratory, September 2015

NREL/PO-6A20-65091

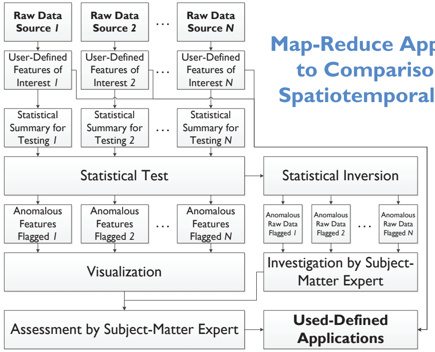
Motivation

- Non-parametric statistical methods combine power with robustness, especially regarding Type I errors.
- Accessible and safe for non-statisticians
- Apply broadly to many types of renewable energy data
- Do not 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.
- Dataset comparison often fits into a common data-processing pattern.
- Such data-processing patterns can be implemented in high performance computing environments as a map-reduce operation.

Map-Reduce Approach to Comparison of Spatiotemporal Data



Raw Data Source *1* · **Raw Data Source 2** · … · **Raw Data Source *N***

User-Defined Features of Interest *1* · User-Defined Features of Interest 2 · … · User-Defined Features of Interest *N*

Statistical Summary for Testing *1* · Statistical Summary for Testing 2 · … · Statistical Summary for Testing *N*

Statistical Test

Statistical Inversion

Anomalous Features Flagged *1* · Anomalous Features Flagged 2 · … · Anomalous Features Flagged *N*

Anomalous Raw Data Flagged *1* · Anomalous Raw Data Flagged 2 · … · Anomalous Raw Data Flagged *N*

Visualization

Investigation by Subject-Matter Expert

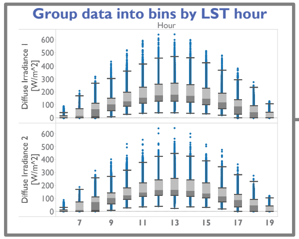
Assessment by Subject-Matter Expert

Used-Defined Applications

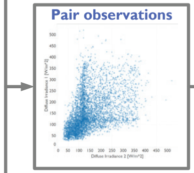
- Analysts are typically faced with deciding which of *N* datasets is most appropriate for their application.
- Analysts rarely use datasets in their raw form, but typically aggregate, transform, or summarize them into a set of features for their application.
- One finds that rigorous statistical comparison of raw datasets usually indicates that they are statistically different, which is not particularly informative to analysts. Only by applying statistical tests to the user-defined features of interest, however, can one determine if the datasets differ for the analyst's application.
- Statistical tests for comparing datasets typically rely on transforming, binning, or summarizing the data before applying the test. The result of the test is the identification of the anomalous features, which can then be visualized and used in assessing the consequences of using each dataset.
- Statistical inversion techniques can allow one to trace back the anomalies identified in the features of interest back to the characteristics of the raw datasets. Subject-matter experts can then focus on determining the fundamental cause of the anomalies and assess the severity of their impact on applications.

Example Application: Non-Parametric Detection of Bias in Diffuse Solar Irradiance

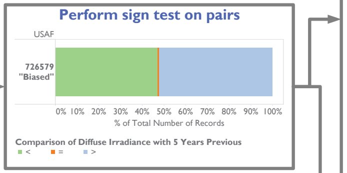
- *Study Question:* How do measurement stations' irradiance reports (dataset #1) compare with the reports from the same hour of the same day five years previously (dataset #2)?
- *Results of Sign Test:* There is a statistically significant bias detectable between these datasets in many geographical regions.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



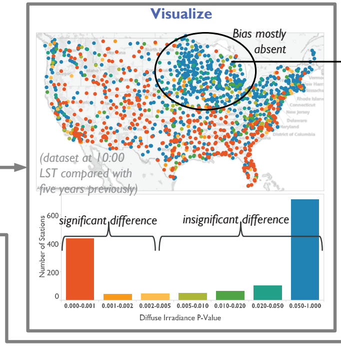
Group data into bins by LST hour



Pair observations



Perform sign test on pairs

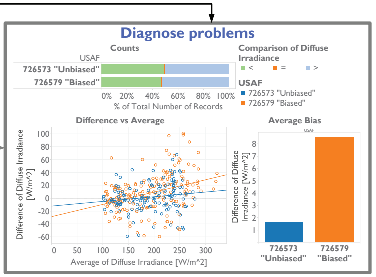


Visualize

Bias mostly absent

(dataset at 10:00 LST compared with five years previously)

examination of USAF 726573 and 726579



Diagnose problems

Example Application: Non-Parametric Detection of Anomalies in a Solar Resource Dataset

Goals

- Robustly identify differences in resource datasets.
  - Applicable to any kind of resource data.
  - Usable by engineers and analysts generally, not just by statistical specialists.
- Apply non-parametric statistical tests to two or more spatiotemporal datasets at high resolution in time and geography.
  - Numerous spatial bins (e.g., one degree by one degree, or smaller).
  - Numerous temporal bins (e.g., month of year with hour of day).
  - Avoid assumptions (hypotheses) about the underlying probably distribution functions.

Example application

- Compare the DHI for the years 2006 and 2007 in the preliminary version of a new solar radiation gridded dataset.
- Computational requirements
  - 20 nodes
  - 320 cores
  - 400 GB memory
- One degree by one degree spatial grid
- Temporal grid of month of year with hour of day

Summary of results

Test	*Wallclock*	*Node*	*Core*	*Result*
Mean-variance	1.5 hrs	30 hrs	610 hrs	No anomalies
Kruskal-Wallace	2.0 hrs	45 hrs	740 hrs	**Anomalies**
Anderson-Darling	11.5 hrs	235 hrs	3740 hrs	**Strong anomalies**

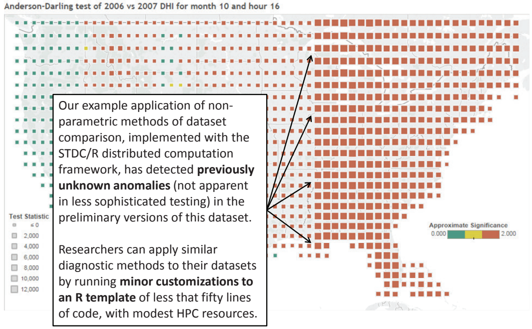
General statistical approach

- Bin the two or more datasets into the same set of spatial and temporal bins.
  - These bins are slices and groupings in space and time.
  - The use of binning reconciles differences in spatial and temporal resolution between the datasets, so different grids, point vs grid data, etc., can be compared.
- Use non-parametric statistical tests to identify differences, biases, or anomalies between the datasets.
  - The tests rely on comparing the empirical distributions of the datasets in corresponding bins.
  - The tests naturally include diagnostics to handle cases where one dataset has far fewer points than another.
  - Where justifiable, parametric statistical tests could also be used.
  - Instead of hypothesis testing, descriptive statistics could just be collected for each bin.
- Collect and analyze the test statistics from each of the spatiotemporal bin.
  - Apply corrections to account for the testing of (the numerous) multiple hypotheses.
  - Rank and highlight differences in order of significance, for visualizations.

Pseudo-code for example application

```
# Initialized distributed computation.
init.grid()

# Use one-degree bins for all of CONUS.
lons <- c(-125.0, -66.0) ; nlon <- 59
lats <- c(  24.0,  50.0) ; nlat <- 26

# Use two-dimensional bins by month and hour of day.
time_keys <- c("month", "hour")

# Read HDF5 data files.
dsa <- get_data_list_by_lonlat('year2007.h5', 'dhi', lons, lats, nlon,
                               nlat, prefix='dsa')
dsb <- get_data_list_by_lonlat('year2008.h5', 'dhi', lons, lats, nlon,
                               nlat, prefix='dsb')

# Apply Anderson-Darling test to each spatiotemporal bin.
results.list <- test_bins(dsa, dsb, ad_test, cols=time_keys)

# Organize results into a distributed data structure.
results.dtables <- concat_bins_safe(results.list)

# Gather results onto the rank 0 node.
results.data <- gather(results.dtables, rank.dest=0, unlist=FALSE)
if (comm.rank() == 0) {

  # Concatenate distributed results into a single reformatted table.
  results.dtable <- unbin_result(concat_bins(results.data, rank0=TRUE))

  # Save results.
  save(file='anderson-darling.RData', list=c('results.dtable'))
}

# End distributed computation.
```

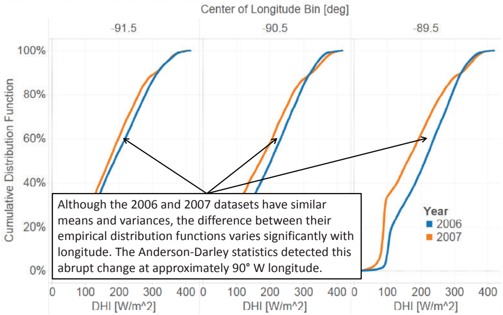
Anderson-Darling test of 2006 vs 2007 DHI for month 10 and hour 16



Our example application of non-parametric methods of dataset comparison, implemented with the STDC/R distributed computation framework, has detected **previously unknown anomalies** (not apparent in less sophisticated testing) in the preliminary versions of this dataset.

Researchers can apply similar diagnostic methods to their datasets by running **minor customizations to an R template** of less that fifty lines of code, with modest HPC resources.

Empirical distribution of DHI for month 10 and hour 16



Although the 2006 and 2007 datasets have similar means and variances, the difference between their empirical distribution functions varies significantly with longitude. The Anderson-Darley statistics detected this abrupt change at approximately 90° W longitude.