Automated Analysis of Renewable Energy Datasets (“EE/RE Data Mining”)  
Brian Bush, Ryan Elmore, Dan Getman, Danny Inman, Eric Kalendra

Goals, Plans, Impacts

Goals: To dramatically improve the understanding of EERE energy efficiency and renewable energy data sets and the depth and efficiency of their analysis through the application of statistical learning methods (“data mining”), in the intelligent processing of these large and messy information sources.

Focus Areas:
- anomaly detection
- data cleansing
- forecasting
- pattern mining
- relational models
- automated reasoning

Application Areas:
- smart-meter data
- data-mining data
- financial-risk models
- predictive data mining

Impact:
- Results from this project have the potential to routinely add value to a wide range of projects across NREL and not just NREL centers.
- The rapid and efficient data mining techniques can significantly lower the costs associated with analysis of data within in-house programs and become a standard feature of such projects.
- This project will provide increased leverage to both NREL and EERE data sets.
- The addition of this capability to NREL can be leveraged in marketing sophisticated, complex analysis and projects to multiple sponsors, putting NREL another step ahead of competitors.

Monte Carlo Study: Imputation and Clustering Techniques

Goals: We designed a large-scale Monte Carlo simulation in order to evaluate the effectiveness of each imputation/clustering base combination to estimate the missing values. In particular, we are interested in answering the following:
- Is it better to impute using multiple imputation or simple split-based methods?
- Should we learn the clustering on the imputation function or data stream on or several properties related to the data stream representation of the stream?
- How robust is the fuzzy clustering method?

Monte Carlo Study: Data with Additional Missing Values

Example Simulation Run:
- We selected one representative month of data (May 2011) as the basis of our simulation study.
- In each simulation run, we included additional missing values so that we could compare the methods of imputation and clustering (see Figure 2).
- The results presented in Tables 1 and 2 are based on applying the imputation/clustering combinations to the “test” data (right panel of Figure 2).
- We verified our results by applying these methods to the validation data set given in the left panel of Figure 2 (data not shown).

Figure 1: Analysis flow diagram illustrating these imputation and clustering approaches used in this study.

Figure 2: Missingness maps of the original validation data (left) and the same data set with more missing values, not set, added (right). Red and blue lines represent missing and observed values, respectively.

Insights from RSF (Research Support Facility) Building Data

Heat recovery controlled building systems can be identified using data-driven techniques.

Weather and CO2 measurements can (sometimes) indicate meeting-room occupancy.

Power data can be used to identify monitoring and missing data.

The hierarchy of power meters in the RSF can be used to identify missing and missing data.

Anomaly and Error Detection in Large Spatial Datasets

Goals: Develop a reliable methodology to approximate:
- Processing and performing statistical analysis of very large spatial-temporal datasets.
- Identification of spatial and temporal trends in raw data and in resulting statistical datasets.
- Visualizing these phenomena for use in analysis and in identification of anomalies.
- Application of data correction and noise reduction techniques.

Results:
- Developed a scalable methodology for anomaly detection in large geospatial datasets using a combination of spatial data and spatial-temporal statistical methods.

Monte Carlo Simulation Results and Summary

Table 1: Clustering results for a typical run in the simulation comparing autocorrelation function-based clustering with weight-based clustering. The weight-based clustering criteria using cosine scores needed to be more stable than any of the other combinations. This is evidenced by the fact that the weights of the power array and Dam Carter data streams are assigned to their own clusters.

Table 2: Clustering results across all imputation/clustering combinations.

Summary: Clustering based on weight properties along with the multiple imputation produces the most reliable results in terms of consistency of cluster membership when using clusters. That is, 99% of out of 1000 simulation trials produced the same cluster assignment (see Table 3). These results provide a strategy for imputing missing observations and clustering assignments of power demand meters.

This presentation does not contain any proprietary, confidential, or otherwise restricted information.

LDRD FY13 Annual Review and Poster Session
Golden, Colorado
June 13, 2013
NREL/PO-620-64976