Weather Data Cleansing for Electrical Load Forecasting

Aubrey Condor

University of Central Florida

INTRODUCTION/ ABSTRACT

Energy forecasting has become widely applicable in the utility industry for system planning and operations. One of the main drivers of electricity demand is weather. Programs such as SAS Energy Forecasting use hourly temperature readings as an input to their forecasting models. A lack of quality temperature data would contribute to less accurate forecasts. Despite observable daily as well as yearly seasonal patterns in the data, temperature can be very volatile, making it difficult to accurately fill in missing data.

In order to monitor and improve the quality of temperature data, we sought to create an algorithm in SAS Enterprise Guide to automate the process of identifying missing and bad data, as well as imputing cleansed data into a temperature series. We examined different methods for filling in short missing series and long missing series using 5 years of historical data from Virginia weather stations. For long missing series, we could fill in using data from nearby stations (in distance and elevation) or from previous time periods of the same temperature series. We also looked at what could qualify as “bad” temperature data.

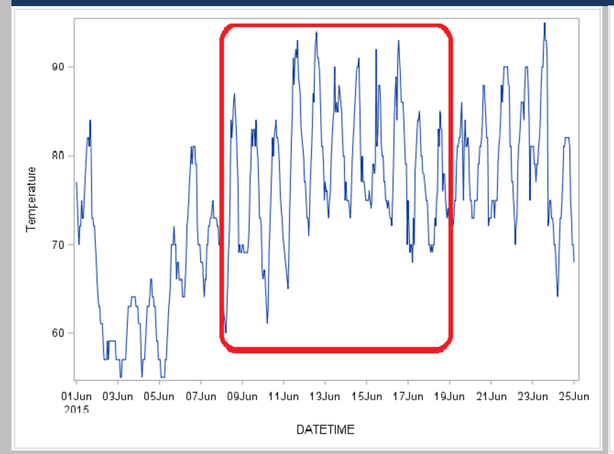
The weather data cleansing algorithm runs relatively quickly with little input from the user. The process is friendly to analysts and business users alike because it is easy to implement regardless of skill level. We think that the forecasting process can be simplified and improved with the use of the algorithm.

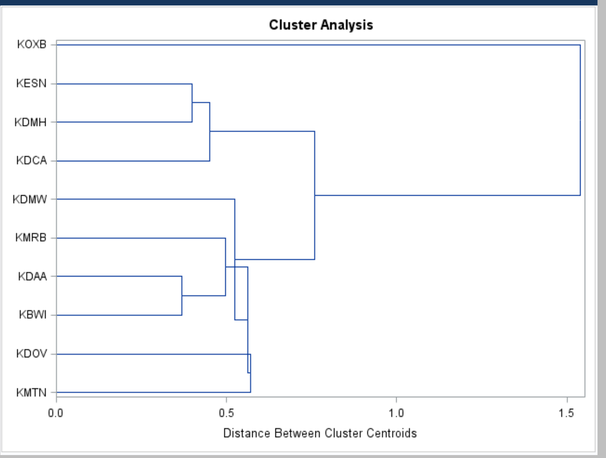
METHODS

Different methods for filling in missing temperature data were examined. For a short series, a linear interpolation was calculated. For a long series, we used data from a nearby station, where non-missing data were found to be most similar. The mean squared difference of the temperature series was used as a similarity measure. Clustering methods were examined as well.

After successfully filling in missing data, we sought to identify ”bad” data. We came up with a method to flag potentially misleading data by looking at the temperature change over an hour, and comparing the change to nearby stations. This process only identifies possible issues with the data. Such issues should then be reviewed by an industry professional for reasonability.

SAMPLE ANALYSIS





Sample Code

```
/* macro to partition distance table, call other macro to
cycles through each pair of relevant weather stations*/
%macro stndist(stn= );
proc sql outobs=5; *only grab top 5 distances for the
relevant station;
create table stndist as
select * from dist where station1 = "&stn";
quit;
data _null_; *run through macro combine for relevant
station combinations;
set stndist;
call execute('%combine(stn1='||strip(station1)||',
stn2='||strip(station2)||');');
run;
%mend stndist;
```

Sample Output

Missing Temperature Data

id	Station	firstdate	lastdate	count
1	KDMH	08JUL15:03:00	08JUL15:04:00	2
2	KJYO	19MAR13:03:00	20MAR13:04:00	26
3	KMTN	03JAN14:03:00	03JAN14:07:00	5
4	KMTN	08JUN15:01:00	19JUN15:00:00	264
5	KOXB	16OCT15:01:00	13JAN16:00:00	2136