Forecasting Daily Solar Energy Production Using Robust Regression Techniques

Gilles Louppe (@glouppe)
Université de Liège, Belgium

Peter Prettenhofer (@pprett)
Graz University of Technology, Austria

Problem statement

Goal

Short-term forecasting of daily solar energy production based on weather forecasts from numerical weather prediction (NWP) models.

Challenges

- High volatility

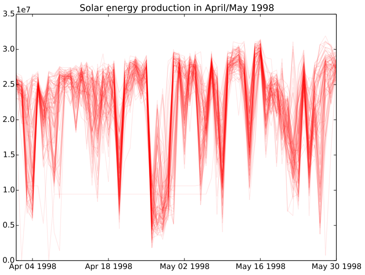
  rapidly changing weather conditions

- Noisy response

  hardware failure

- Noisy inputs

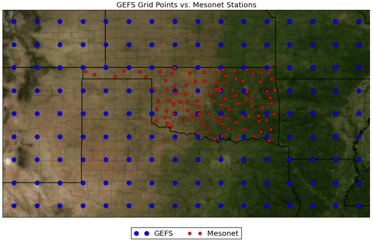
  inaccuracy of NWP model



Data

Solar energy production

- 98 Oklahoma Mesonet sites
- Total incoming solar energy in Jm^{-2}
- Time period : 1994 - 2007

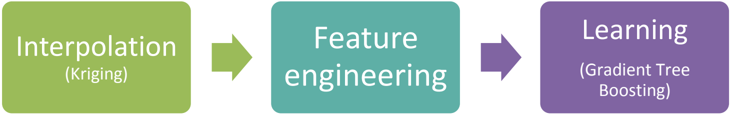


Courtesy : Dr. Amy McGovern

Numerical weather prediction

- NOAA/NCEP GEFS Reforecast, 5 forecasts per day
- Ensemble comprises 11 members (one control)
- 15 measurements (temp, humidity, upward radiative flux, ...)

Overview of our approach



1. **Interpolation** of meteorological measurements from GEFS grid points onto Mesonet sites ;
2. Construction of **new variables** from the measurement estimates ;
3. **Forecasting** of daily energy production using Gradient Boosted Regression Trees, on the basis of the local measurement estimates.

Kriging

Goal : Estimate meteorological variables (temperature, humidity, ...) locally at all Mesonet sites.

For each day d, period h and type f of meteorological measurement :

1. Build a local learning set

$$\mathcal{L}_{dhf} = \{(\mathbf{x}_i = (\text{lat}_i, \text{lon}_i, \text{elevation}_i), y_i = \overline{m_{idhf}})\},$$

   where $\overline{m_{idhf}}$ is the average value (over the ensemble) of measurements m_{idhf} of type f, at GEFS location i, day d and period h ;
2. Learn a Gaussian Process from $\mathcal{L}_{dhf}$, for predicting measurements from coordinates ;

   (Fitting is perfomed using *nuggets* to account for noise in the measurements.)
3. Predict measurement estimates $\widehat{m_{jdhf}}$ at Mesonet stations j from their coordinates.

Feature engineering

Goal : Build a learning set $\mathcal{L}$ from the measurement estimates.

1. Concatenate the estimates at all periods h and for all types f, for each Mesonet station j and day d :

$$\mathcal{L} = \{(\mathbf{x}_{jd} = (\widehat{m_{jdh_1f_1}}, \widehat{m_{jdh_1f_2}}, ...), y_{jd} = p_{jd})\}$$

   where p_{jd} is the energy production at Mesonet station j and day d.

2. Extend inputs $\mathbf{x}_{jd}$ with engineered features :
   - Solar features (delta between sunrise and sunset)
   - Temporal features (day of year, month)
   - Spatial features (latitude, longitude, elevation)
   - Non-linear combinations of measurement estimates
   - Daily mean estimates
   - Variance of the measurement estimates, as produced by the Gaussian Processes

Predicting energy production

Goal : Predict daily energy production at Mesonet sites.

1. Learn a model using Gradient Boosted Regression Trees (`sklearn.ensemble.GradientBoostingRegressor`), predicting output y from inputs $\mathbf{x}$;
   - Use the *Least Absolute Deviation* loss for robustness ;
   - Optimize hyper-parameters on an internal validation set ;
2. For further robustness, repeat Step 1 several times (using different random seeds) and aggregate the predictions of all models.

Results

Evaluation

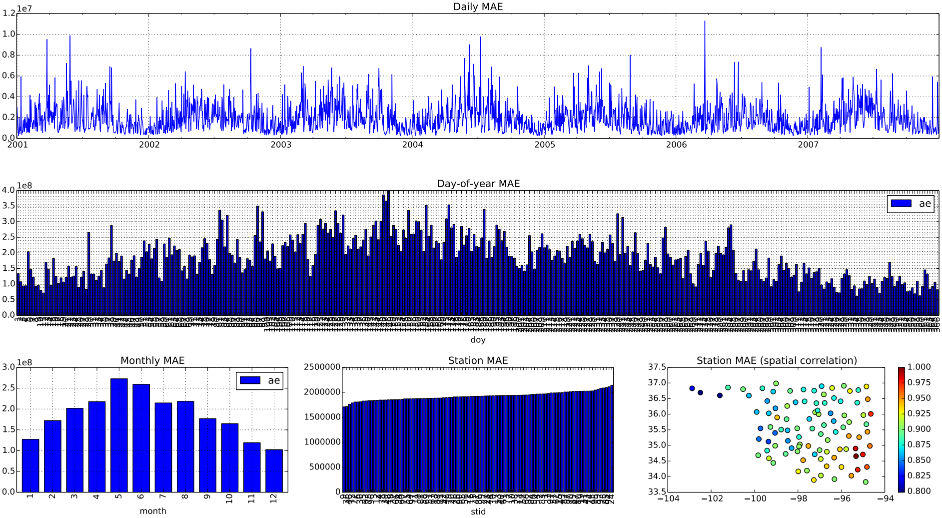
- Held-out data from 2008 - 2012.
- Mean Absolute Error (MAE) as metric :

$$MAE = \frac{1}{JD} \sum_{j=1}^{J} \sum_{d=1}^{D} |p_{jd} - \hat{p}_{jd}|$$

Results

Method	Heldout-Score [MAE]	Δ [%]
GMM	4019469.94	46.19%
Spline Interp.	2611293.30	17.17%
Kriging + GBRT	2162799.74	-
Best	2107588.17	-2.62%

Error analysis



Conclusions

✓ **Competitive** results (4th position) ;

✓ **Robust** approach at all steps of the pipeline ;

✗ Including additional data from nearest GEFS grid points might have further improved our results.

Questions ? g.louppe|peter.prettenhofer@gmail.com

Kriging illustration