Yu Ding

Barnes Professor of Industrial & Systems Engineering

Feb 13, 2015
ISEN Big Data related areas

- Advanced Manufacturing
- Human & Organizational Systems
- Operations Research
- System Informatics

Yu Ding (ISEN)
Dr. Andy Banerjee, system-level simulation visualization for healthcare, security and manufacturing applications.

Dr. Yu Ding, applied statistics and quality engineering for energy, manufacturing, and security applications.

Dr. Andy Johnson, production economics and nonparametric estimation for energy, manufacturing, and healthcare applications.

Dr. Justin Yates, spatial and geo-data analysis and social media networks for security and logistics applications.

Dr. Li Zeng, applied Bayesian statistics and statistical process control for bio-manufacturing and nano-manufacturing.

• **Data**: Manufacturing & production data, material characterization data, economics data and logistic data.

• **Engineering Applications**: Manufacturing, healthcare, energy, and security and logistics.
**Nano informatics**

- **Data**: Static and dynamic (video) nano imaging data, in the amount exceeding 1.2 GB.
- **Data Providers**: TAMU Nanomaterials Processing and Atomic Imaging Lab; Chinese National Nano Science Center; KAUST; Georgia Tech Manufacturing Institute; FSU’s High Performance Material Institute.
- **Collaborators**: Huang (STAT), Mallick (STAT), Liang (MEEN), Ji (ECEN), Bukkapatnam (ISEN).
Anatomy of a wind turbine: gearbox is particularly prone to failure.

Construction of wind turbines at a remote site.

**Data**: Real data in the amount of 120 GB and simulated data (by NWTC) of 1.2 TB.

**Data Providers**: Risø National Lab, Nationa Wind Technology Center (NWTC), commercial wind farms, and ABB (pending).

**Collaborators**: Huang (STAT), Mallick (STAT), Genton (STAT), Xie (ECEN), Kumar (ECEN), Kezunovic (ECEN), Johnson (ISEN), Moreno-Centeno (ISEN), Ntaimo (ISEN).
System informatics promotes an effective and seamless integration of physical knowledge and first principle models with data driven methodologies.
A case study: multiple-dependency model for wind power curve


• Technical objective: build a probability predictive model $p(y|x)$, in which
  
  ▶ $y$ is the power output of a turbine;
  
  ▶ $x$ comprises the vector of, **at least**:
    1. wind speed $v$;
    2. wind direction $d$;
    3. air density $\rho$ (incorporating temperature and air pressure);
    4. turbulence intensity $t_b$;
    5. humidity $h_m$;
    6. above-hub wind shear $w_a$;
    7. below-hub wind shear $w_b$. 
Current practice

- Industrial standard (recommended by IEC): Binning

- Other data driven methods:
  - Multivariate product kernels;
  - Smoothing spline method (SSANOVA);
  - Bayesian Additive Regression Trees (BART): a Bayesian version of the classification and regression tree method.

- **Challenge**: balance capability to capture system complexity and scalability.
Our approach

• Physical law provides us some clue.

Physics behind:

\[
\text{Power} = \frac{1}{2} \rho \cdot A \cdot C_p \cdot v^3
\]

• At least three important factors affect wind power generation;

• Functional relationship nonlinear with function form unknown;

• Interactions exist among the factors.

• This insights motivate a hybrid structure that we name it an Additive multivariate kernel (AMK) model:

\[
p(y|x) \leftarrow \text{kernel}(v, d, \rho) + \text{kernel}(v, d, h_m) + \text{kernel}(v, d, t_m) + \ldots
\]

- A three-component multivariate kernel captures the critical factor interactions (with wind speed and wind direction);
- An additive structure ensures the scalability of the final model.
What difference does this make?

- **Six sets of turbine data for comparison.** WT 1 - WT4: inland turbine datasets; WT 5 - WT6: offshore turbine datasets.

- **Noise level:** WT1-WT3’s noise relatively low, WT4 data sees a 50% increase, and WT5 and WT6’s noise level triples that of WT1-WT3’s.

- **Against Binning (industry practice):** AMK reduces predictive error by as much as 45%.

- **Against BART:** Comparably on WT1-WT3 data; but considerably better on WT4-WT6 data. AMK beats BART by 18% on WT4 data and over 30% on WT5-WT6 data.

- **Implication:** AMK is competitive because of the special kernel model structure advised by the physical insights. BART tries to learn the intrinsic structure through the data. BART does so well enough when data are of good quality but its capability becomes remarkably less effective when data are noisy.