Probabilistic Wind Power Forecasting

Hongdi Fu, Steven Xu, Yishen Wang and Prof. Daniel Kirschen

Department of Electrical Engineering, University of Washington

INTRODUCTION

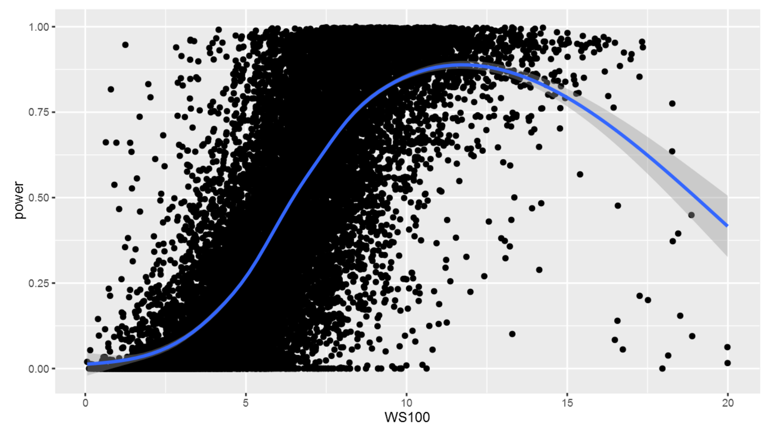
- Machine learning has become an effective tool for interpretation and prediction in real-life tasks involving large data sets.
- Renewable energy industries are interested in estimating the energies consumed and produced in a future time period based on historical trend.
- Compared to point forecasting, probabilistic forecasting gives more comprehensive information about future events by assigning probabilities to each outcome.
- Data are provided by CrowdANALYTIX GEFCOM 2014 with raw inputs of zonal(U) and meridional(V) components of the wind speed, timestamp as well as the observed wind power output.

OBJECTIVES

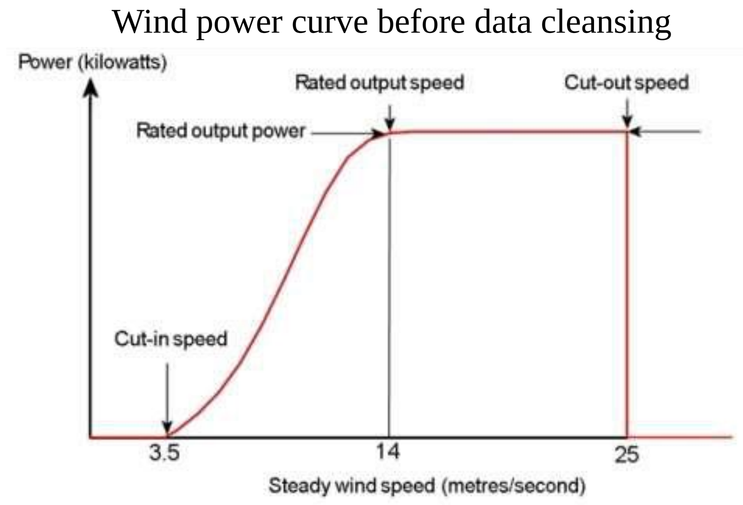
- ***Forecast*** the probabilistic distribution(in quantiles) of the wind power generation based on history input data.
- ***Replicate*** using some of the famous methods from published work.
- ***Explore*** and apply some of the newest machine learning algorithm(R packages).
- ***Compare*** with published results to seek improvements and deficiencies.

DATA CLEANSING

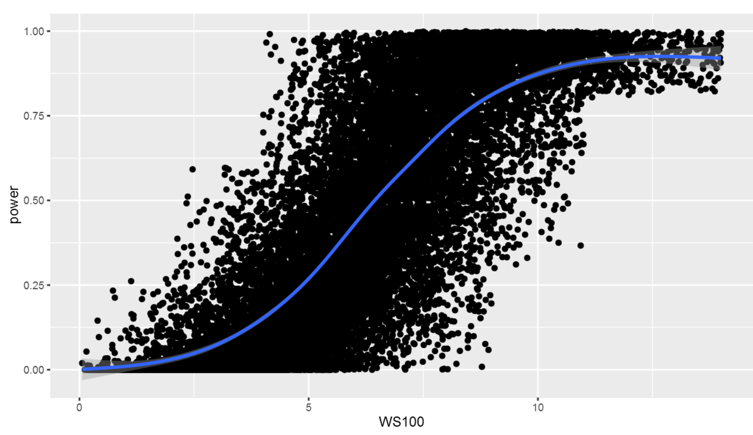
- ***Exclude*** time periods which have power equals to 0 continuously for more than 48 hours.
- ***Omit*** time spots where responding power is not defined(NA).
- ***Remove*** outliers according to the theoretical relation.



Wind power curve before data cleansing



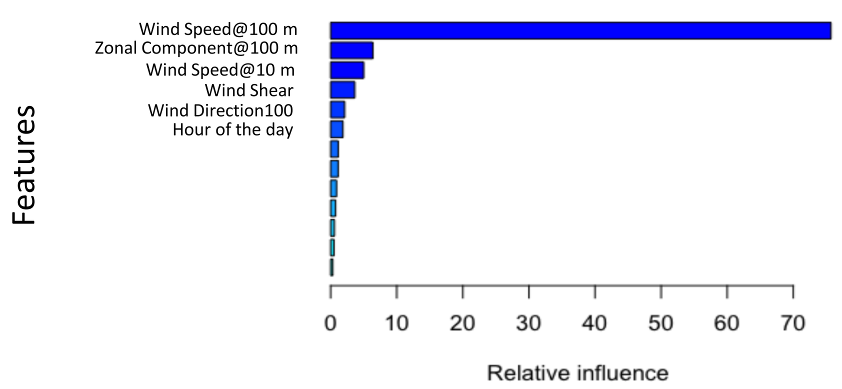
Theoretical relation between wind power and speed



Wind power curve after data cleansing

FEATURE SELECTION

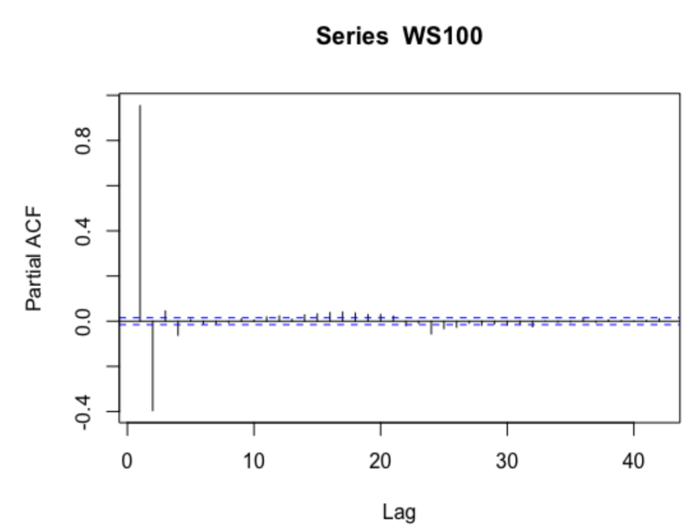
- Tool: Build-in feature importance function in Gradient Boosting Machine(GBM)



- Wind speed: $\sqrt{U^2+V^2}$
- Wind direction: $\frac{180}{\pi}\tan^{-1}(U,V)$ or $\frac{180}{\pi}\tan^{-1}(U,V)+360$
- Wind shear: $|WS100-WS10|$
- Hour: $\cos(\frac{2\pi * Hour}{24})$

LAGGED VARIABLES

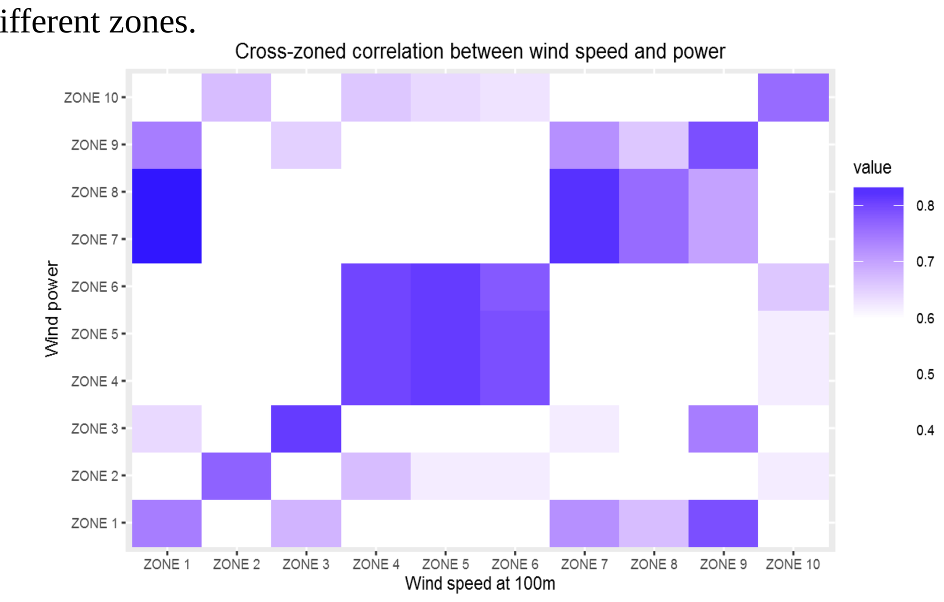
- Lagged variables are almost always considered in time series analysis for better prediction due to the time-dependent nature of the data. Original inputs are shifted by selected time units to form lagged variables.
- Tool: Auto correlation function(acf) in R.



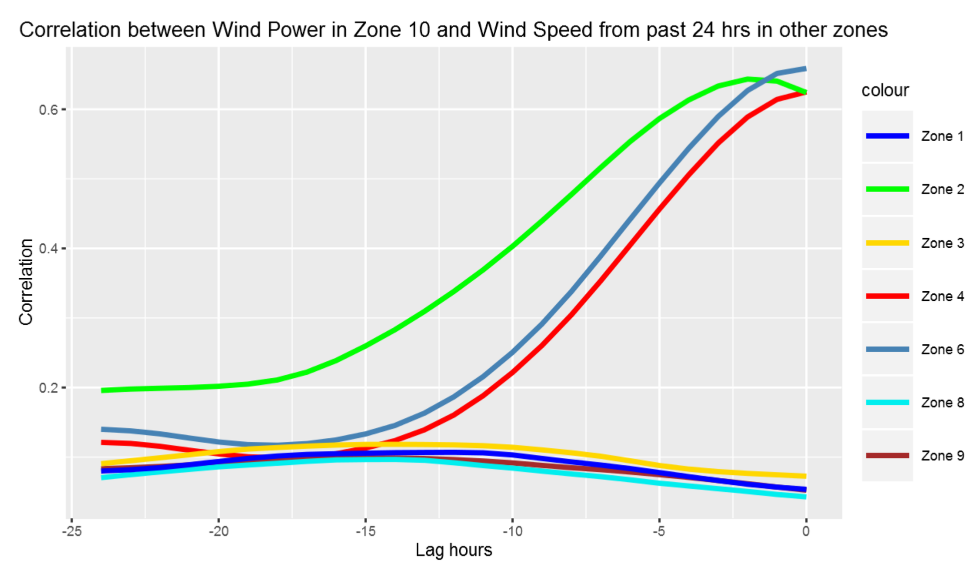
- Correlation is peaked between WS_t and WS_{t-1}.

Zonal Correlation

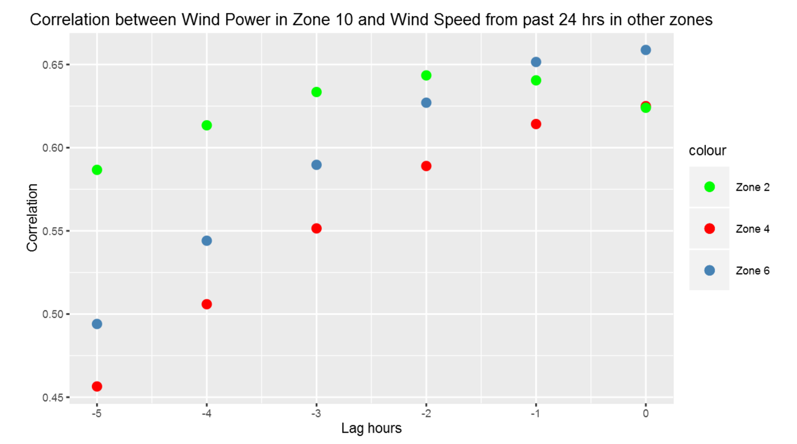
- Wind farms in this task are located in the same region of global, therefore it is intuitive to think that there might be cross correlation between different zones.



- Darker region infers higher cross-correlation



- Zone 2, 4, 6 have the highest correlation with zone 10. In general, high correlation is reached starting from t-3.

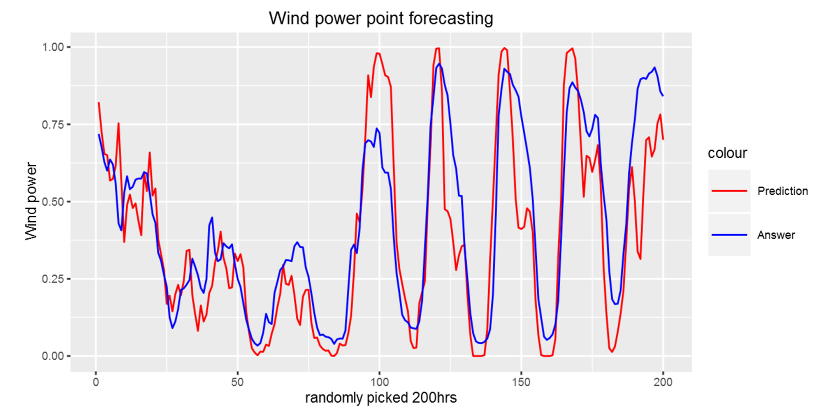


POINT FORECASTING

- Tool: eXtreme Gradient Boosting(XGBoost), a variant of GBM but much faster.
  1. A loss function to be optimized
  2. A weak learner to make predictions(Decision trees)
  3. An additive model to add weak learners to minimize the loss function
  4. Learn a regression predictor -> Compute residuals -> Learn to predict the residuals -> Adds up to the original model -> Repeat

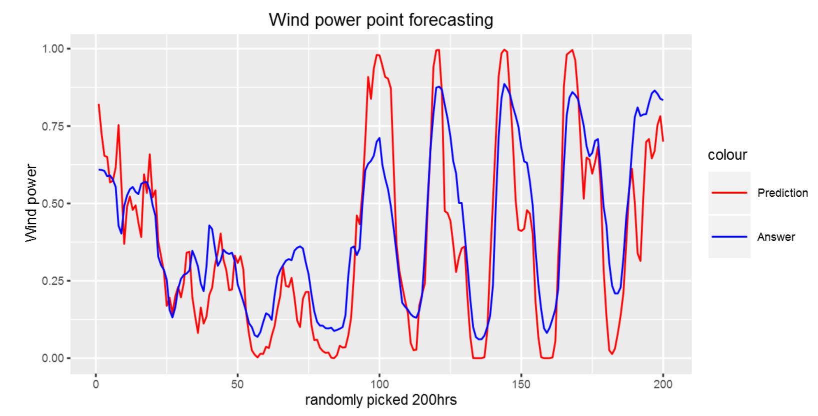
GBM

MSE: 0.02588



XGBoost

MSE: 0.02352

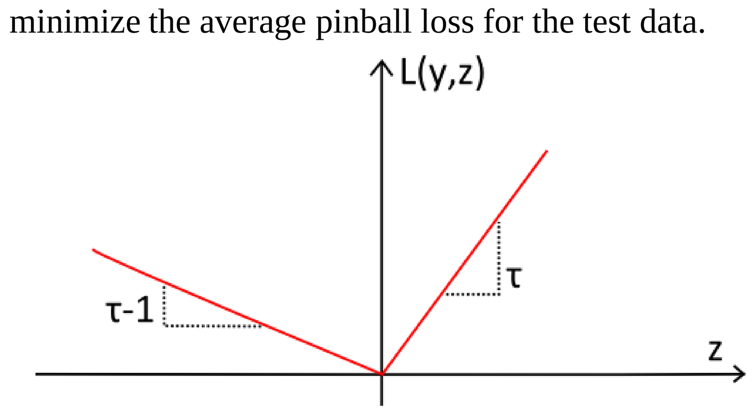


PROBABILISTIC FORECASTING

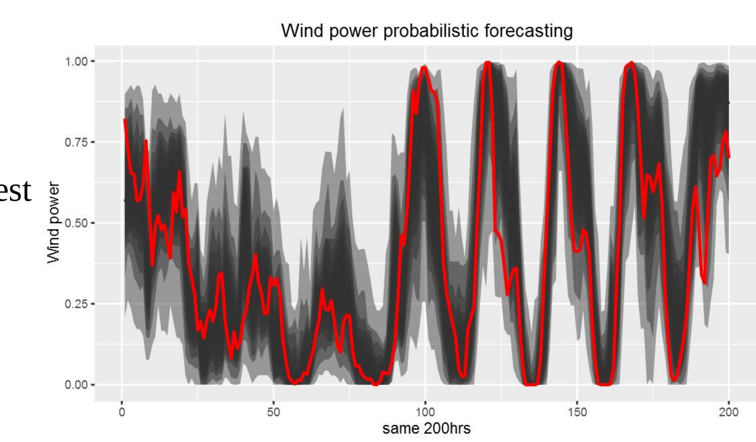
- Tool: GBM with built-in quantile regression; quantregForest
- Similar to point forecasting, but use quantile loss(pinball loss) instead of squared loss.
- Pinball loss function
  1. Let τ be the target quantile, y the real value and z the quantile forecast. Then $L\tau$, the pinball loss function, can be written as:

$$L_\tau(y,z) = \begin{cases} (y-z)\tau & \text{if } y \geq z \\ (z-y)(1-r) & \text{if } z > y \end{cases}$$

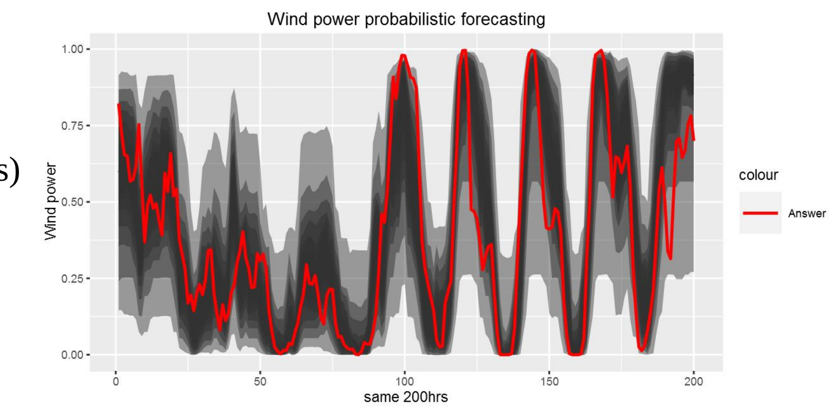
- We want to minimize the average pinball loss for the test data.



quantregForest



GBM(quantile loss)

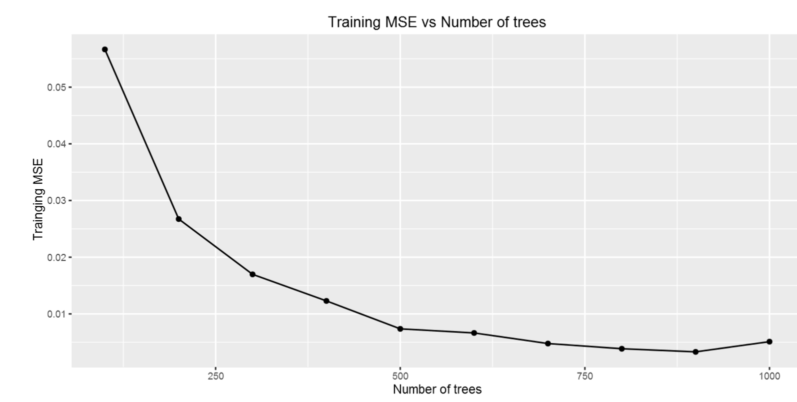


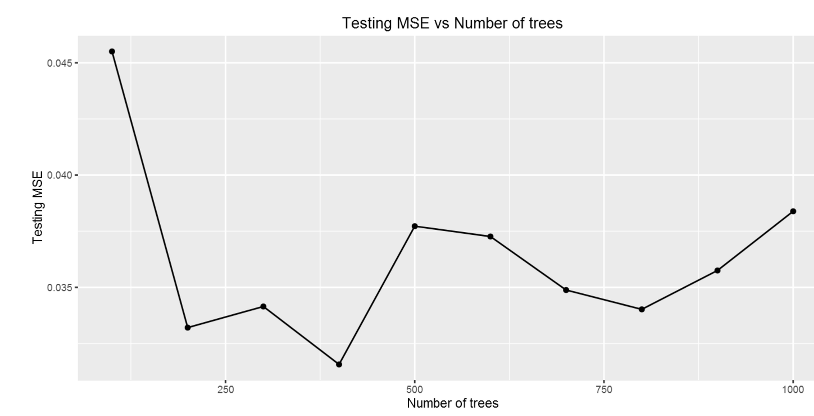
Models	Bench mark	quantregForest	GBM (quantile)
Pinball loss	0.07584	0.06373	0.03452

- Bench mark was given by GEF2014 by assigning constants to each quantile

PARAMETER TUNING

- GBM and XGBoost have lots of tunable parameters
- Regular cross validation and bootstrapping: problematic
- Possible mistakes: use future data to predict history data, inconsecutive sample data
- Choice: self-designed validation process





- We seek value of parameter that gives both low training error(bias) and testing error(variance).

CONCLUSION AND EXPANSION

- The combination of XGBoost as point forecasting model and GBM (quantile loss) as probabilistic forecasting model gives a very impressive forecasting score.
- Parameter tuning can be done more sophisticatedly
- More interesting packages to be explored: MXNet, QRNN
- Possible reason for zones with same input but different outputs might be the difference in wind turbine
- Periodicity of wind speed, wind direction etc. can be taken in to consideration as some published work suggested.

RESOURCES AND REFERENCES

1. Landry, Mark, Thomas P. Erlinger, David Patschke, and Craig Varricio. "Probabilistic Gradient Boosting Machines for GEFCom2014 Wind Forecasting." *International Journal of Forecasting* 32.3 (2016): 1061-066. *Probabilistic Gradient Boosting Machines for GEFCom2014 Wind Forecasting*. Elsevier, 26 Mar. 2016. Web. 16 Dec. 2016.

2. Zhang, Yao, Jianxue Wang, and Xifan Wang. "Review on Probabilistic Forecasting of Wind Power Generation." *Renewable and Sustainable Energy Reviews* 32 (2014): 255-70. *Review on Probabilistic Forecasting of Wind Power Generation*. Elsevier, 31 Jan. 2014. Web. 16 Dec. 2016.

3. Probabilistic Forecast Application in Power Systems – Yi Luan, Bolun Xu University of Washington