Wind power predicting
Case study

Mines Saint-Etienne / Majeure *Science des données*
M. Lutz (Octo Technology) and O. Roustant (Mines Saint-Etienne)
October - November 2015

The aim of the case study is to predict / forecast the electric power produced by a wind turbine. The data, simulated but realistic, correspond to 10-minute lagged variables : the electric power (response), wind speed, speed direction, and other variables; For confidentiality reasons, the data have been rescaled and are given without unit, except for angles which are given in degrees. More details are given in the slides.

Both statistical and machine learning approaches will be considered. Statistical models will allow to quantify the uncertainty on predictions. We ask you to follow the guideline below, starting by standard methods and finishing by an open challenge.

Organization: Groups of 4.
Evaluation: Your work should be presented in a report in pdf format, limited to 15 pages, given at the end of the case study by e-mail (mlutz@octo.com ; roustant@emse.fr) ; The R code must also be provided.

1. **[Day 1]** Data visualization and preparation. Look at each variable as well as couples of variables. Conclusions? For this step, and maybe during the whole exploratory stage, you should first work on a sample of moderate size (say 10% of the data) drawn at random (use 'sample' function in R), in order to reduce the running time (in particular for plotting procedures).

2. Split your data into a train set (say the first 75% data) and a test set.
   From now on, use the train set for modelling.

3. Construction of a linear model, $Y_t = \beta_0 + \beta_1 X_{1,t} + \ldots + \beta_1 X_{k,t} + u_t$. Notice that the *model* variables X_i may not be equal to the *problem* variables, but may be constituted by some *features* built by combining or transforming the problem variables (*feature engineering*).
   Evaluate its performance in prediction on the test set.

4. **[Day 2]** Short-tem forecasting with the linear model. Consider that the last 3 hours data are unknown, and forecast at t_0-3h the electric power at horizon 10mn, 20mn, ..., 3h, where t_0 is the present time. As no wind forecast is given, and as a toy study, we will use the future values of the predictors. (In practice one has to use meteorological forecasts).

   (a) By block-bootstrapping u_t (the residuals of the linear model).
   Give 95% prediction intervals.

   (b) By first modelling u_t as an ARMA time series $\Phi(B)u_t = \Theta(B)\varepsilon_t$, and doing (simple) boostrap on ε_t (the residuals of the time series).
   Give 95% prediction intervals.

5. **[Day 3]** Improvement of the linear model for prediction. Use a random forest model on u_t.
   Evaluate the performance in prediction of the approach in the same way of Question 2. Compare with the performance of the linear model.

6. Challenge! Propose your own model in order to improve the prediction performance on the test set.