CARNEGIE MELLON UNIVERSITY
DATA ANALYTICS (COURSE 18-899)
ASSIGNMENT 2

You should submit, using Canvas, a report in the form of a PDF document (Student_ID-Name-DAassignment2.pdf). Include a cover-sheet on the assignment with your name and the required details. Number the pages, graphs, tables and answers carefully to correspond with the questions. Each answer should be supported by Matlab or R or Python code, graphs and calculations. The submission deadline is 24:00 Rwandan Time (CAT) on **Monday 10 February 2020**. If you prefer to use R and Python for this assignment, the report should provide a list of the non-built-in libraries you used in your code.

1. Intraday on-shore wind power generation measured every hour for one year is available from the csv file WindGeneration.csv. Load the data into your computer and produce a graphic showing the time series of the wind generation over time. Is there evidence of annual seasonality?

2. Plot the change in wind generation over time as a percentage of the maximum generation. Is there evidence of annual seasonality?

3. Consider positive and negative ramps in wind power generation, x(t), as a percentage of the maximum, over the hourly timescale. An hourly ramp is therefore defined as r(t,d) =100*[x(t+d)-x(t)]/max(x) where d=1 for an hourly sampling period. Construct empirical cumulative distribution functions (CDF) for both the positive and negative ramps and plot these with the probability on a vertical logarithmic axis. Plot the CDF for a normal distribution with mean-zero and standard deviation from the observations. Is the normal distribution a good model for wind power extremes?

4. National power system operators are tasked with the challenge of balancing supply and demand. They need to understand the variability in wind generation over different timescales. Investigate variability over timescales from one hour to one day by plotting the 1%, 5%, 95% and 99% percentiles. This can be achieved using distributions of the ramps r(t,d) with d =1,2,…,24.

5. Calculate and plot the autocorrelation of wind generation for lags over 10 days. Comment on the structure of the autocorrelation.

6. Calculate and plot the autocorrelation of change in wind generation for lags over 10 days. Include horizontal lines to detect statistically significance values ($p<0.05$). Is there any evidence of diurnal seasonality? Might it be more appropriate to model the change in wind generation than the wind generation?

7. Use a variance ratio test to investigate the structure of the wind generation time series. Can the null hypothesis of a random walk be rejected? Is there evidence of either mean-reversion or mean-aversion?

8. Estimate the optimal window for simple moving average. Is there a simple benchmark that improves on the persistence benchmark?

9. Evaluate the mean-Absolute-error (MAE) performance of the persistence benchmark forecast over forecast horizons from one hour to one day. Plot MAE as a percentage of the maximum generation for the persistence benchmark.

10. Loop over the number of parameters to use in an ARIMA model for describing wind generation and use information criteria (AIC and BIC) to find the optimal ARIMA model.