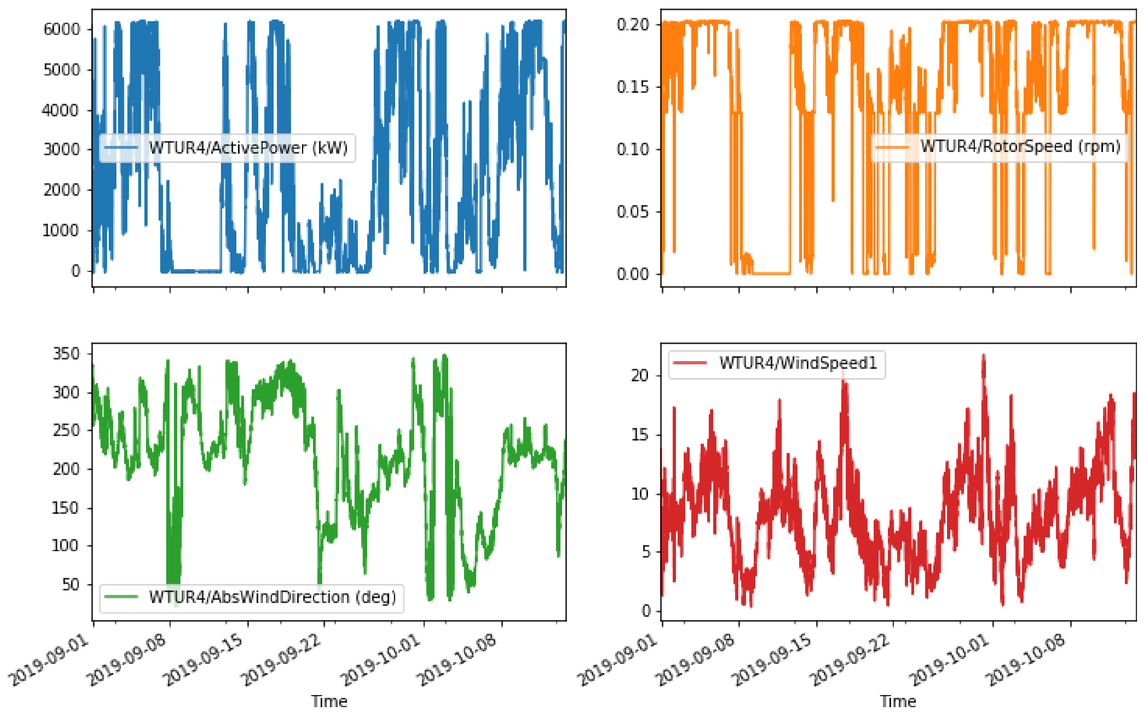
Data Exploration and Conclusion

Yohann Jacob Sandvik
Institute of Electronic Systems - NTNU

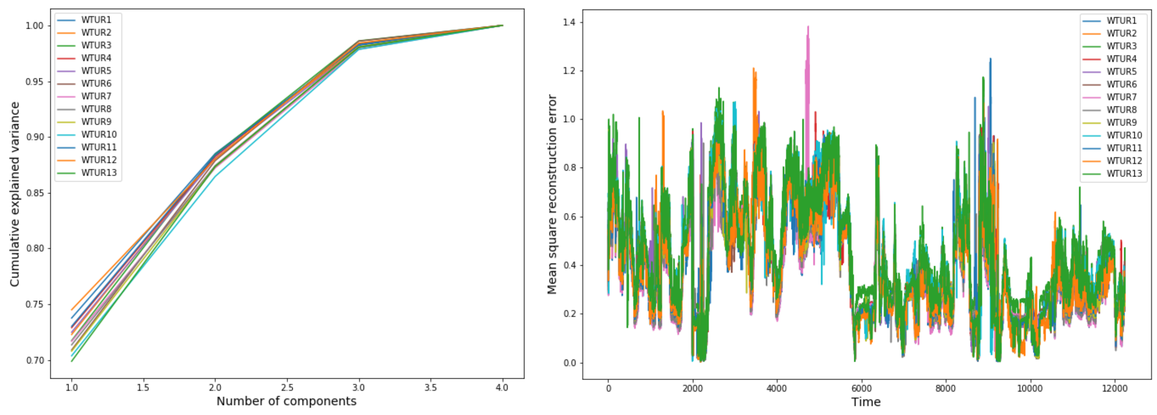
All Values Associated with A Turbine



Dataset Description

- This dataset is taken from a wind farm in Norway with 13 wind turbines.
- Each turbine can be viewed as a multivariate time series, with four dimensions: active power produced by a turbine in kilowatts, wind speed measured in front of the blades in meters per second, the rotational speed of the rotor in rpm and the absolute direction of the wind speed in degrees.
- The values are sampled every five minutes from the 31. of August 2019 00:00 until the 13. of September 2019 21:55, which totals 12239 samples per univariate time series and 48956 samples per wind turbine.
- The dataset chosen does not have any missing values, so there was no requirement for estimating missing values.

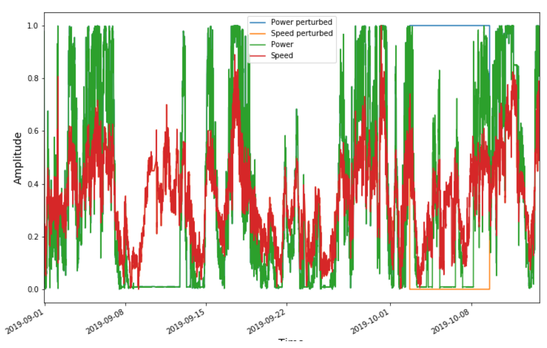
Cumulative Explained Variance and Reconstruction Error



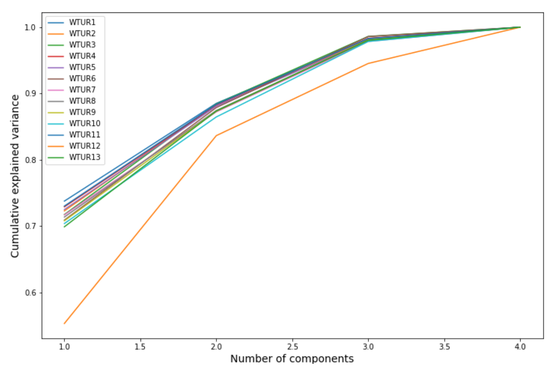
What Can PCA Do?

- The total variance is the sum of the variences of the individual principal components.
- The explained variance of a principal component is the ratio of the variance of said component to the total variance of all the principal components, and the cumulative explained variance of n principal components is the sum of the explained variance of the n first principal components.
- The rightmost plot shows the cumulative explained variance of each turbine for a given number of principle components.
- The leftmost plot shows the total MSE of the reconstructed time series associated with the different wind turbines.
- From the figure one can see that all the cumulative explained variance curves, and reconstruction error curves follow more or less the same shape, with the exception of turbine seven that has a slight spike at round about 5000 samples.

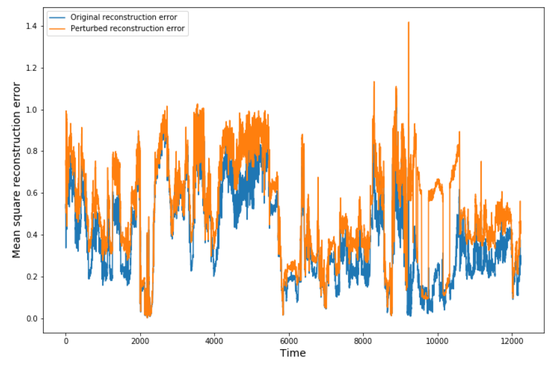
Perturbation



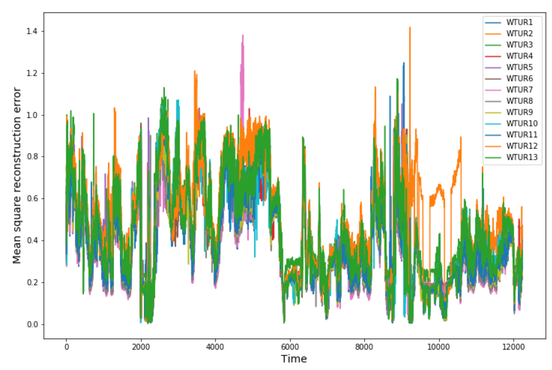
CEV



Pert vs. Unpert



Reconstruction Error



Discussion

- From the cumulative explained variance plot, it is possible to see that artifical perturbation is quite evident.
- The reconstruction error also shows the artificial perturbation.
- The most interesting observation might be that the reconstruction error caused by artificial perturbation could yield a ”threshold” as to what can be considered anomalous behaviour.
- As for example the peak in reconstruction error of turbine seven at around 5000 samples.

Conclusion

The methods chosen to be most relevant to be explored in a master thesis are:

- Feature extraction using PCA, and clustering in reduced feature space.
- Model-based time series representation using a univariate ARMA model, and possibly a multivariate vector AR model.
- Since there are many open-source python implementations of different clustering algorithms, the only restriction set on which clustering algorithms to test is ease of implementation. So algorithms such as Self-organizing maps are excluded.