[EDA] Exploratory Data Analysis

Note: this report summarizes the output that can be found on the Jupyter Notebook (`EDA/EDA_ExploratoryDataAnalysis.ipynb`) which also includes the source code in-line (html can be consulted). This report has been created with the sole purpose of meeting the requirement of 5 pages for the visual report.

The report is an analysis of the file `dataset_processed.csv` which has been created by performing an ETL process (Extract Transform Load) of the original CSV file, running 10 minutes averages, on the cloud.

Missing Data Analysis

First things first, ETL (Extract Transform Load) shows that the dataset contains gaps (identified by NaN values when pivoting on the output format). The following gaps are identified. The analysis also reveals that the gaps are per asset (all the variables for that asset are missing for those gaps).

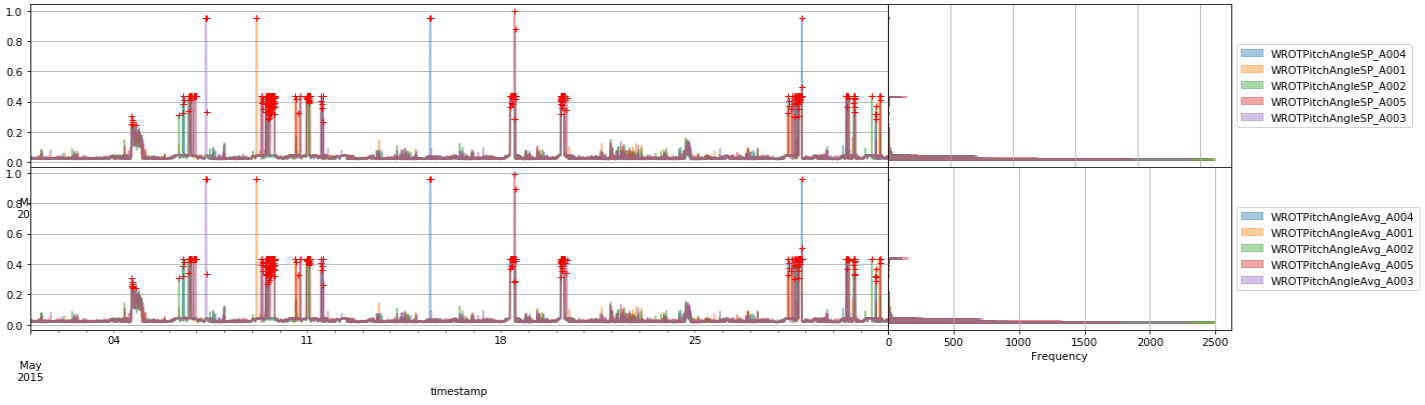
start	stop	duration	asset
2015-05-06 06:20:00	2015-05-06 13:30:00	0 days 07:10:00	A005
2015-05-07 06:20:00	2015-05-07 13:20:00	0 days 07:00:00	A005
2015-05-18 09:40:00	2015-05-18 12:20:00	0 days 02:40:00	A004
2015-05-21 11:10:00	2015-05-21 11:40:00	0 days 00:30:00	A002
2015-05-21 11:40:00	2015-05-21 12:10:00	0 days 00:30:00	A004
2015-05-25 09:50:00	2015-05-25 13:20:00	0 days 03:30:00	A003
2015-05-25 11:40:00	2015-05-25 12:00:00	0 days 00:20:00	A005
2015-05-26 09:00:00	2015-05-26 11:10:00	0 days 02:10:00	A001
2015-05-27 11:10:00	2015-05-27 11:30:00	0 days 00:20:00	A003

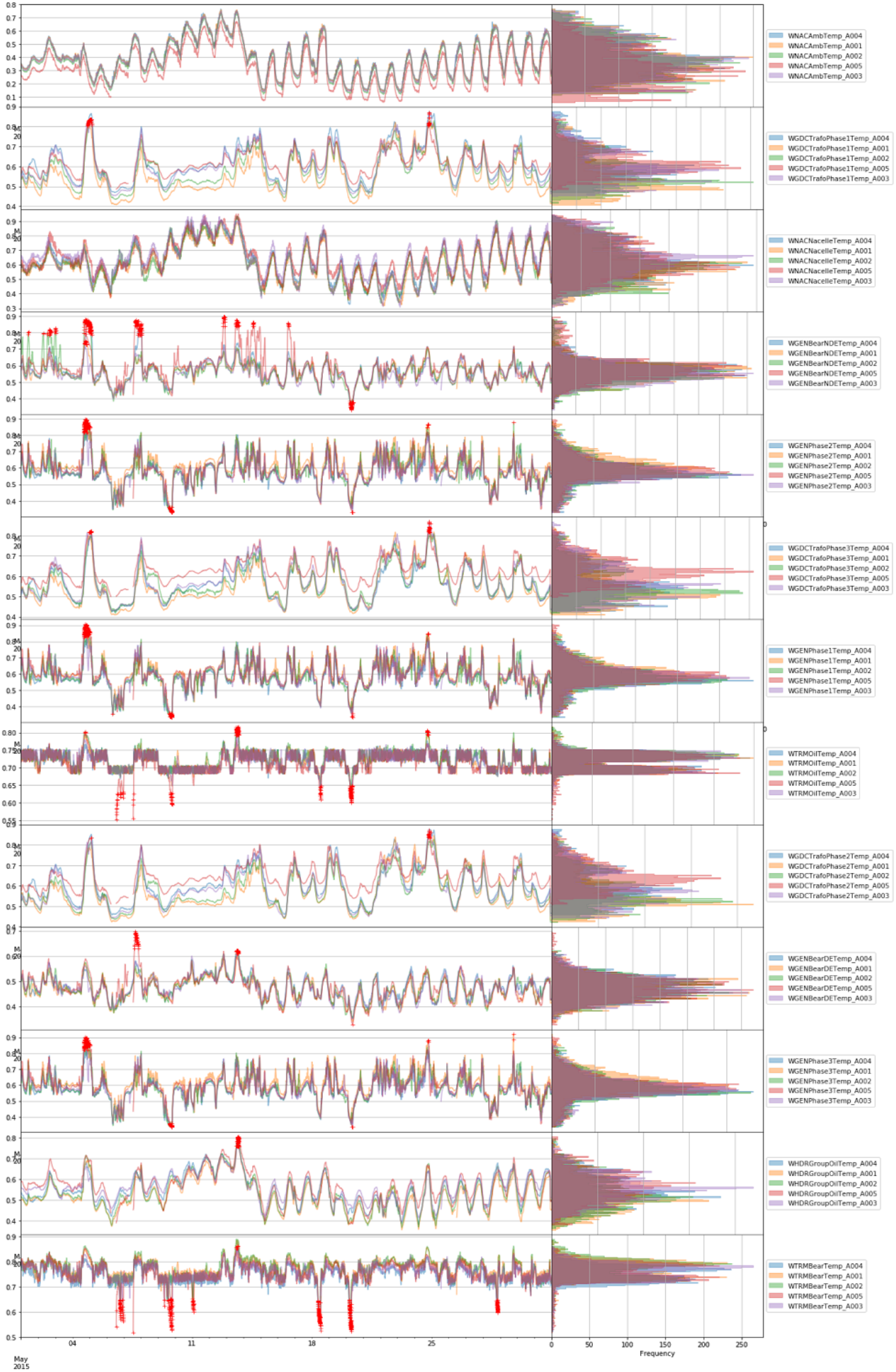
Time series analysis

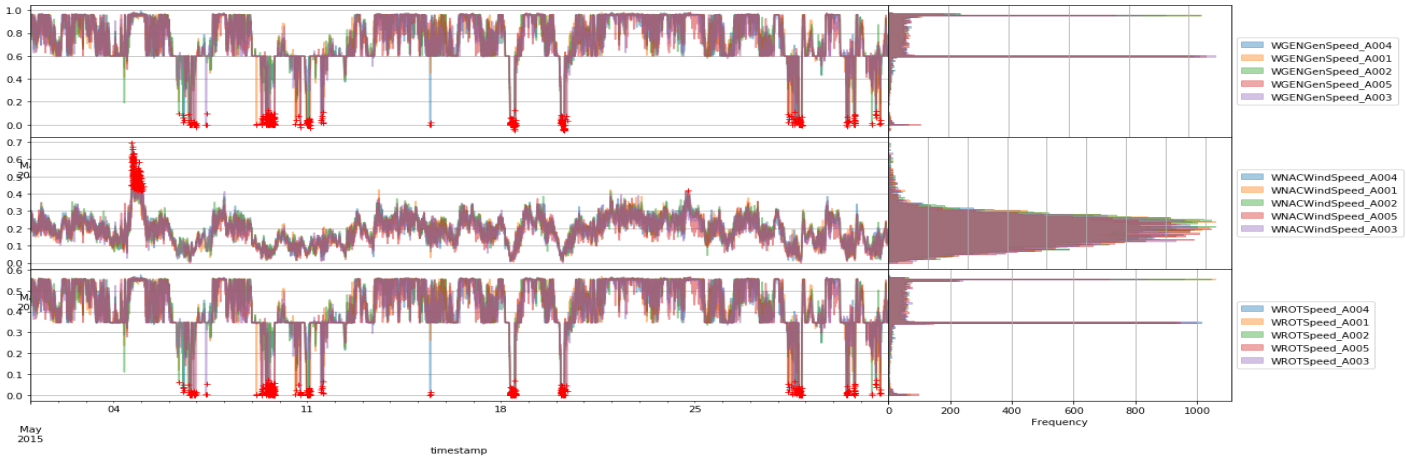
The data contains signals which come from sensors and therefore represent a physical quantity (temperature, speed, pressure). The statistical nature of these signals is to be analyzed, but a very good first approximation is to find outliers defined in a statistical sense as those points that are more than 3-sigmas away from the mean. The outliers are highlighted in red in the signal.

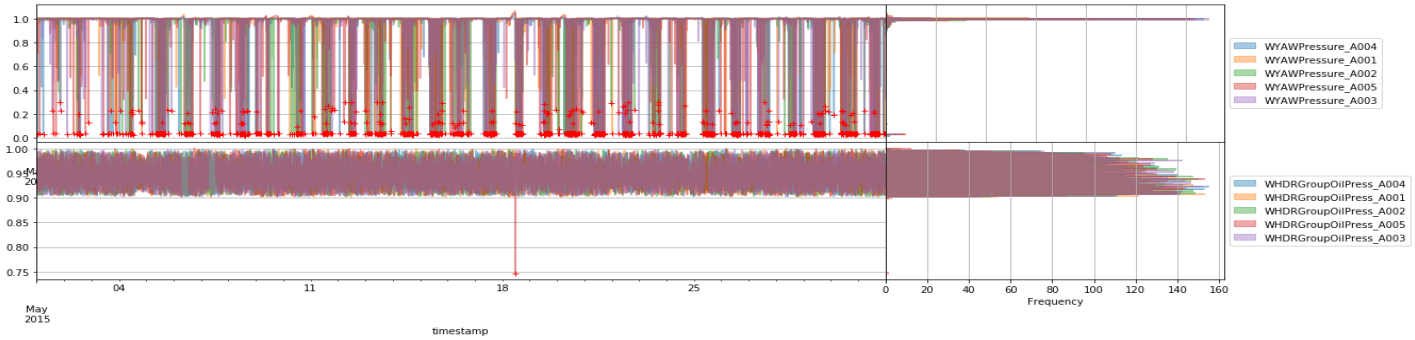
$$o_x(n) = 1 \leftarrow (x(n) > \mu_x + 3\sigma_x) \cup (x(n) < \mu_x - 3\sigma_x)$$

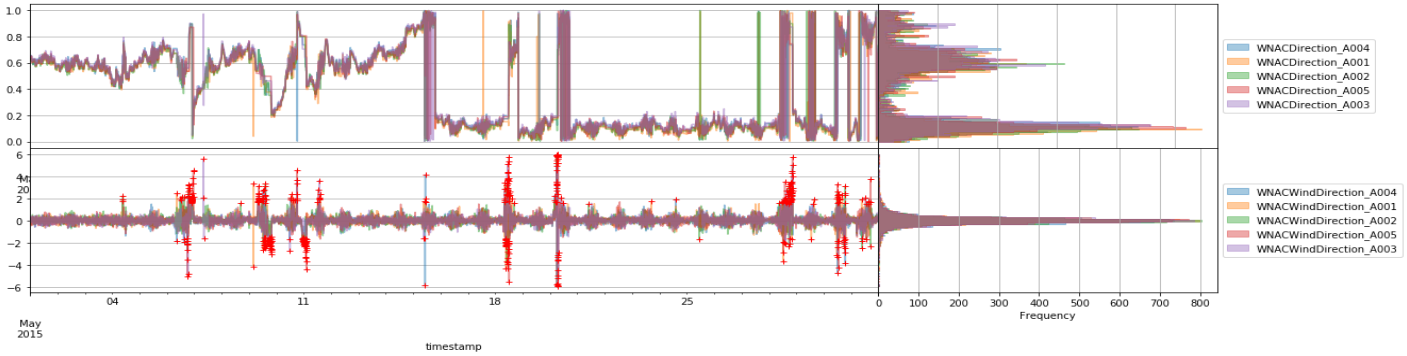
Given the high level of correlation among the same variable of different assets the following plot shows the 28 variables, where each color is an asset. The histogram is plotted on the side and the outliers are highlighted in red. The plots are grouped per variable nature (e.g. temperature, speed, angle, power)

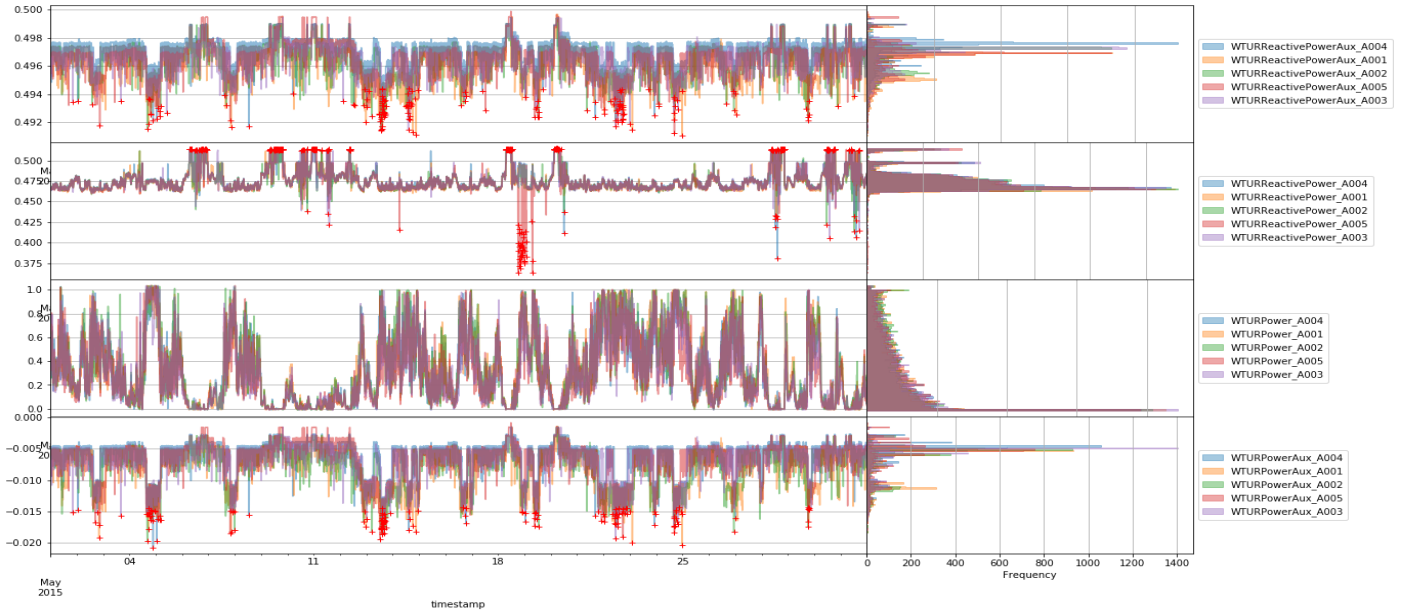


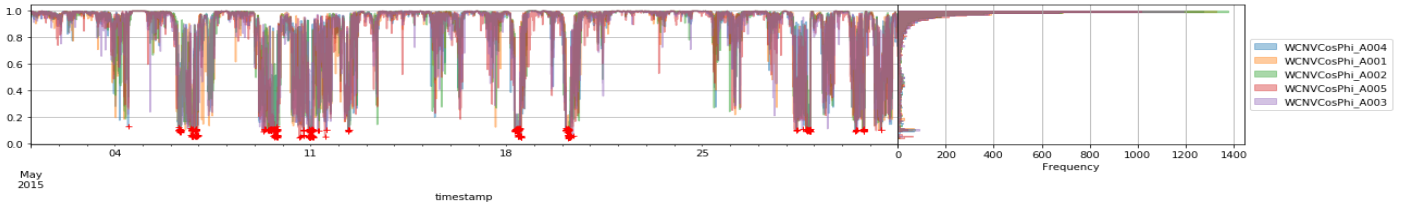

WNACAmbTemp_A004
WNACAmbTemp_A001
WNACAmbTemp_A002
WNACAmbTemp_A005
WNACAmbTemp_A003
WGDCTrafoPhase1Temp_A004
WGDCTrafoPhase1Temp_A001
WGDCTrafoPhase1Temp_A002
WGDCTrafoPhase1Temp_A005
WGDCTrafoPhase1Temp_A003
WNACNacelleTemp_A004
WNACNacelleTemp_A001
WNACNacelleTemp_A002
WNACNacelleTemp_A005
WNACNacelleTemp_A003
WGENBearNDETemp_A004
WGENBearNDETemp_A001
WGENBearNDETemp_A002
WGENBearNDETemp_A005
WGENBearNDETemp_A003
WGENPhase2Temp_A004
WGENPhase2Temp_A001
WGENPhase2Temp_A002
WGENPhase2Temp_A005
WGENPhase2Temp_A003
WGDCTrafoPhase3Temp_A004
WGDCTrafoPhase3Temp_A001
WGDCTrafoPhase3Temp_A002
WGDCTrafoPhase3Temp_A005
WGDCTrafoPhase3Temp_A003
WGENPhase1Temp_A004
WGENPhase1Temp_A001
WGENPhase1Temp_A002
WGENPhase1Temp_A005
WGENPhase1Temp_A003
WTRMOilTemp_A004
WTRMOilTemp_A001
WTRMOilTemp_A002
WTRMOilTemp_A005
WTRMOilTemp_A003
WGDCTrafoPhase2Temp_A004
WGDCTrafoPhase2Temp_A001
WGDCTrafoPhase2Temp_A002
WGDCTrafoPhase2Temp_A005
WGDCTrafoPhase2Temp_A003
WGENBearDETemp_A004
WGENBearDETemp_A001
WGENBearDETemp_A002
WGENBearDETemp_A005
WGENBearDETemp_A003
WGENPhase3Temp_A004
WGENPhase3Temp_A001
WGENPhase3Temp_A002
WGENPhase3Temp_A005
WGENPhase3Temp_A003
WHDRGroupOilTemp_A004
WHDRGroupOilTemp_A001
WHDRGroupOilTemp_A002
WHDRGroupOilTemp_A005
WHDRGroupOilTemp_A003
WTRMBearTemp_A004
WTRMBearTemp_A001
WTRMBearTemp_A002
WTRMBearTemp_A005
WTRMBearTemp_A003
04
11
18
25
May
2015
timestamp
0
50
100
150
200
250
Frequency


WGENGenSpeed_A004
WGENGenSpeed_A001
WGENGenSpeed_A002
WGENGenSpeed_A005
WGENGenSpeed_A003
WNACWindSpeed_A004
WNACWindSpeed_A001
WNACWindSpeed_A002
WNACWindSpeed_A005
WNACWindSpeed_A003
WROTSpeed_A004
WROTSpeed_A001
WROTSpeed_A002
WROTSpeed_A005
WROTSpeed_A003
May
2015
timestamp
Frequency


WYAWPressure_A004
WYAWPressure_A001
WYAWPressure_A002
WYAWPressure_A005
WYAWPressure_A003
WHDRGroupOilPress_A004
WHDRGroupOilPress_A001
WHDRGroupOilPress_A002
WHDRGroupOilPress_A005
WHDRGroupOilPress_A003
May
2015
timestamp
Frequency

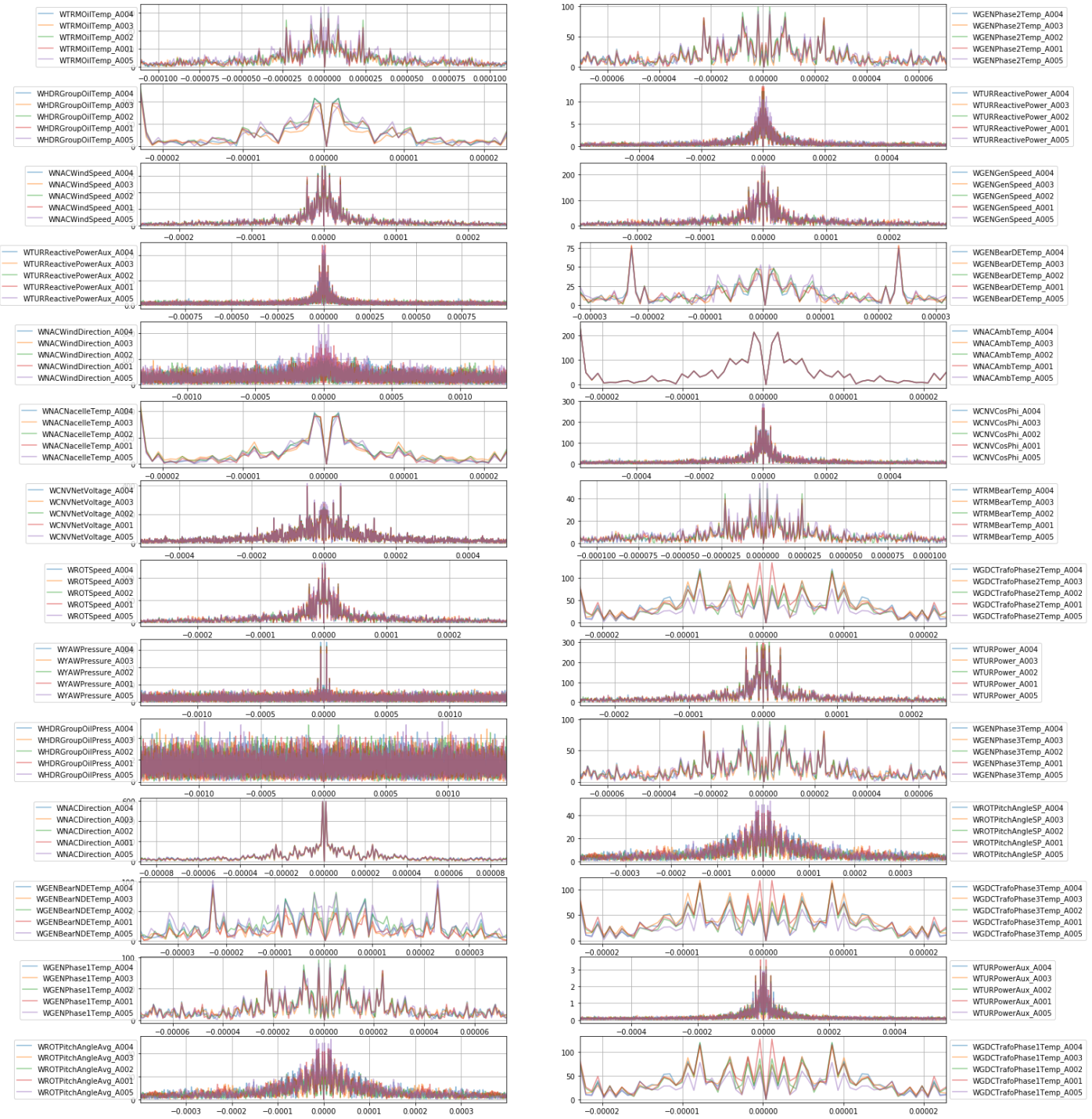

WNACDirection_A004
WNACDirection_A001
WNACDirection_A002
WNACDirection_A005
WNACDirection_A003
WNACWindDirection_A004
WNACWindDirection_A001
WNACWindDirection_A002
WNACWindDirection_A005
WNACWindDirection_A003
May
2015
timestamp
Frequency


WTURReactivePowerAux_A004
WTURReactivePowerAux_A001
WTURReactivePowerAux_A002
WTURReactivePowerAux_A005
WTURReactivePowerAux_A003
WTURReactivePower_A004
WTURReactivePower_A001
WTURReactivePower_A002
WTURReactivePower_A005
WTURReactivePower_A003
WTURPower_A004
WTURPower_A001
WTURPower_A002
WTURPower_A005
WTURPower_A003
WTURPowerAux_A004
WTURPowerAux_A001
WTURPowerAux_A002
WTURPowerAux_A005
WTURPowerAux_A003
May
2015
timestamp
Frequency


WCNVCosPhi_A004
WCNVCosPhi_A001
WCNVCosPhi_A002
WCNVCosPhi_A005
WCNVCosPhi_A003
May
2015
timestamp
Frequency

Spectral analysis

These time signals are continuous and can also be inspected from the spectral point of view. Since almost all signals contain a bias, the best visualization is given by removing the DC component and adjusting the shown bandwidth so 95% of the energy of the signal is contained (otherwise most plots would be seen as a dirac delta at DC).

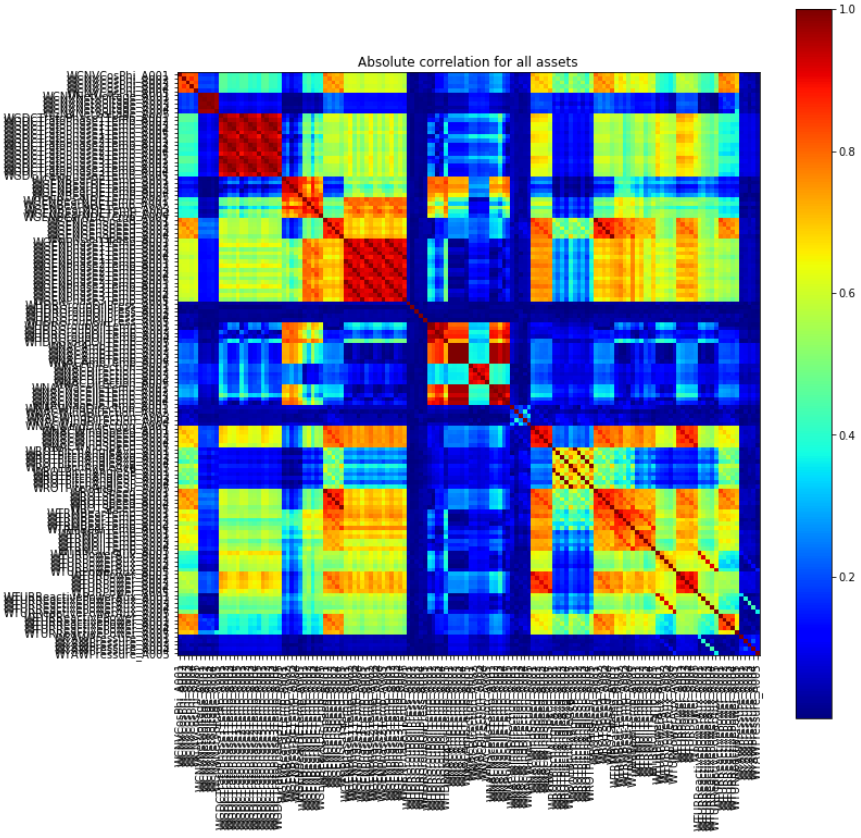


Correlation analysis

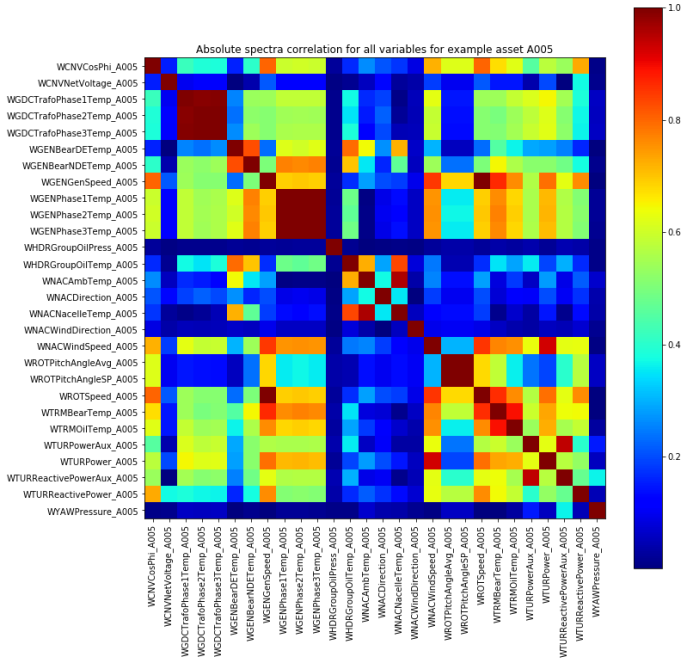
There is a high level of correlation among the signals. There are three types of correlations observed: correlation among assets, correlation among variables within an asset in the time domain, correlation among variables within an asset in the spectral domain.

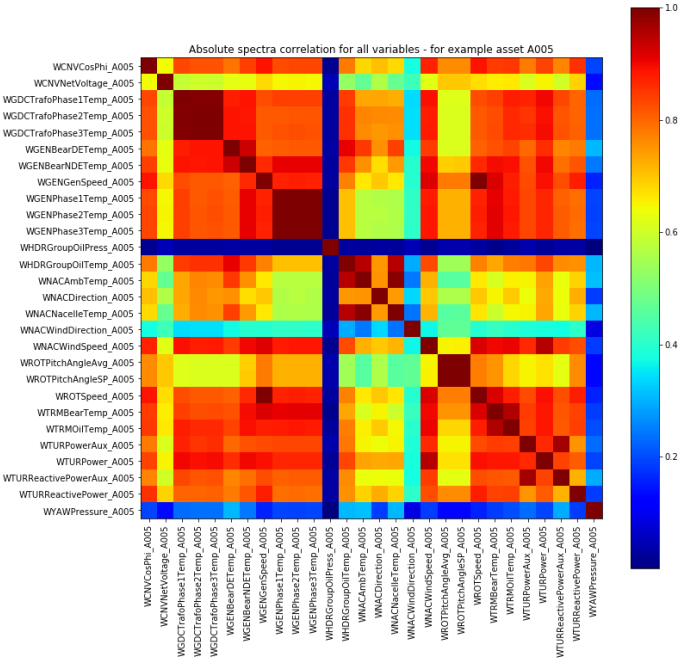
Correlation among assets

The following plot represents the correlation matrix of all variables for all assets (28*5=140). The hypothesis to confirm is that the data is highly correlated (i.e. correlation shows a clustering of 5 elements around the same variable), which does.



Correlation matrix among variables on the same asset on time and spectral domain





One can observe high correlation among temperatures, angles, power and reactive power, wind speed and power, speed and temperature, phase1 temp and speed. Others like pressure seem to be totally decorrelated to the rest.

AOB:
An analysis on signal clustering has also been carried out on the notebook with inconclusive results.