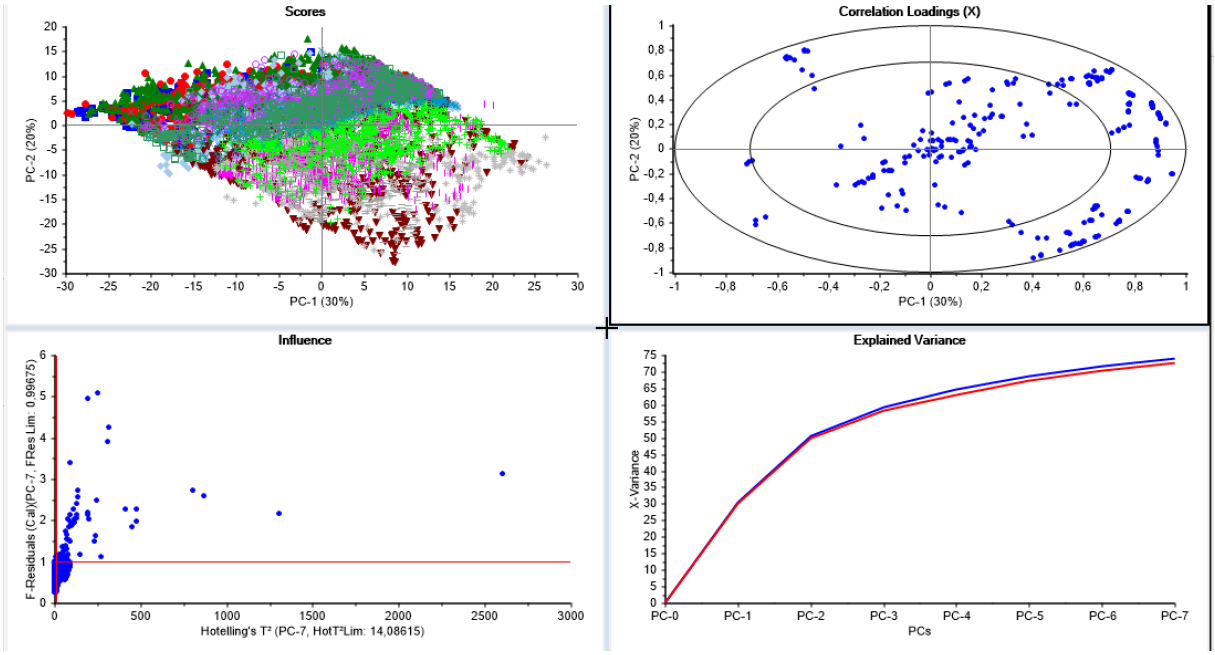
Assignment 8 – Regression modelling and validation

The objectives with the analysis is to train and test various models for the ActuralPower.

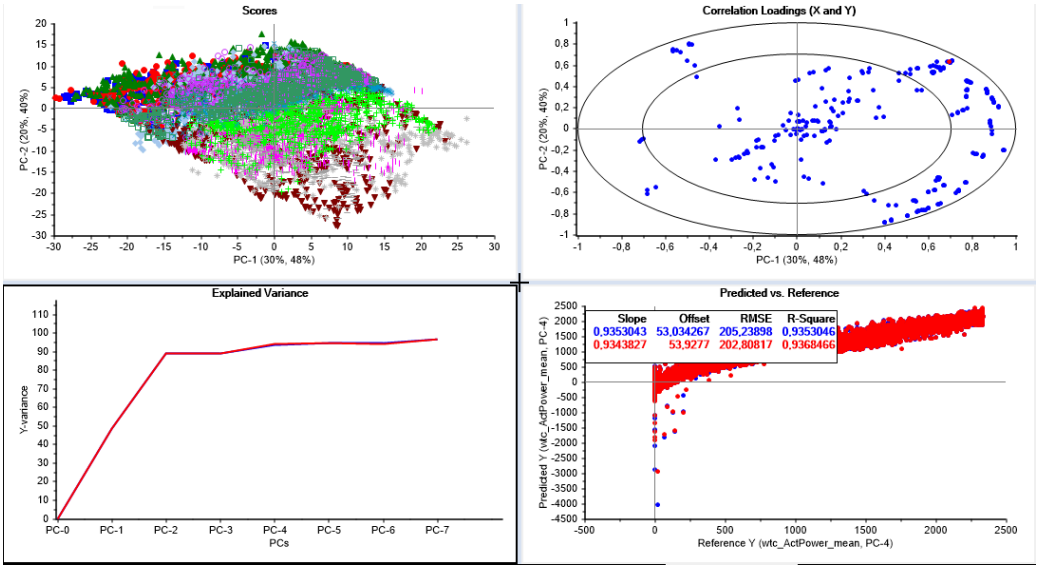
Start with PCA and see if there are any time dependencies (hint: show Month as category variable in Sample Grouping)



In the scores plot there is some clustering according to the different months. Thus it seems to be time dependency.

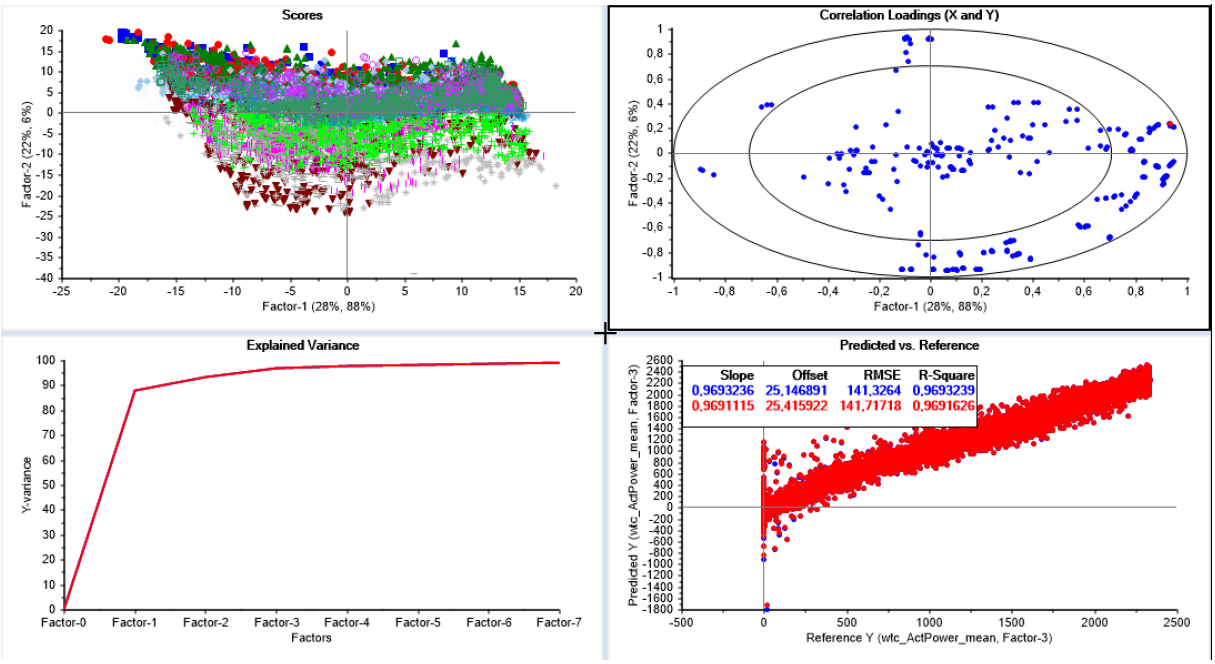
Try both with PCR and PLSR. Are there differences correlation structure in the correlation loadings for the two methods?

PCR



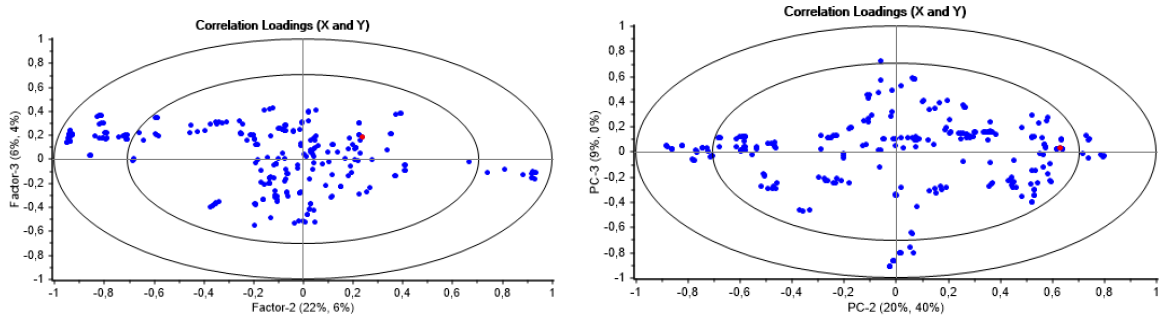
This is the result of the PCR. It shows the same tendencies to clustering as the PCA plot.

PLSR

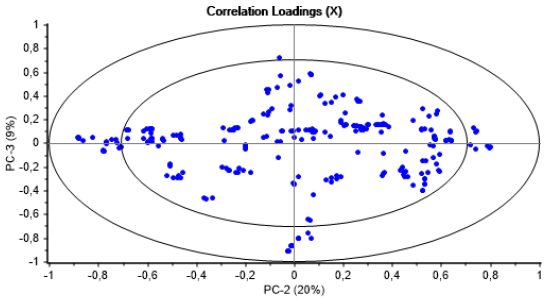


For all three plots the correlation loadings are the same, except the correlation loadings plot for PLS is rotated.

Looking at the correlation loadings for higher PCs:



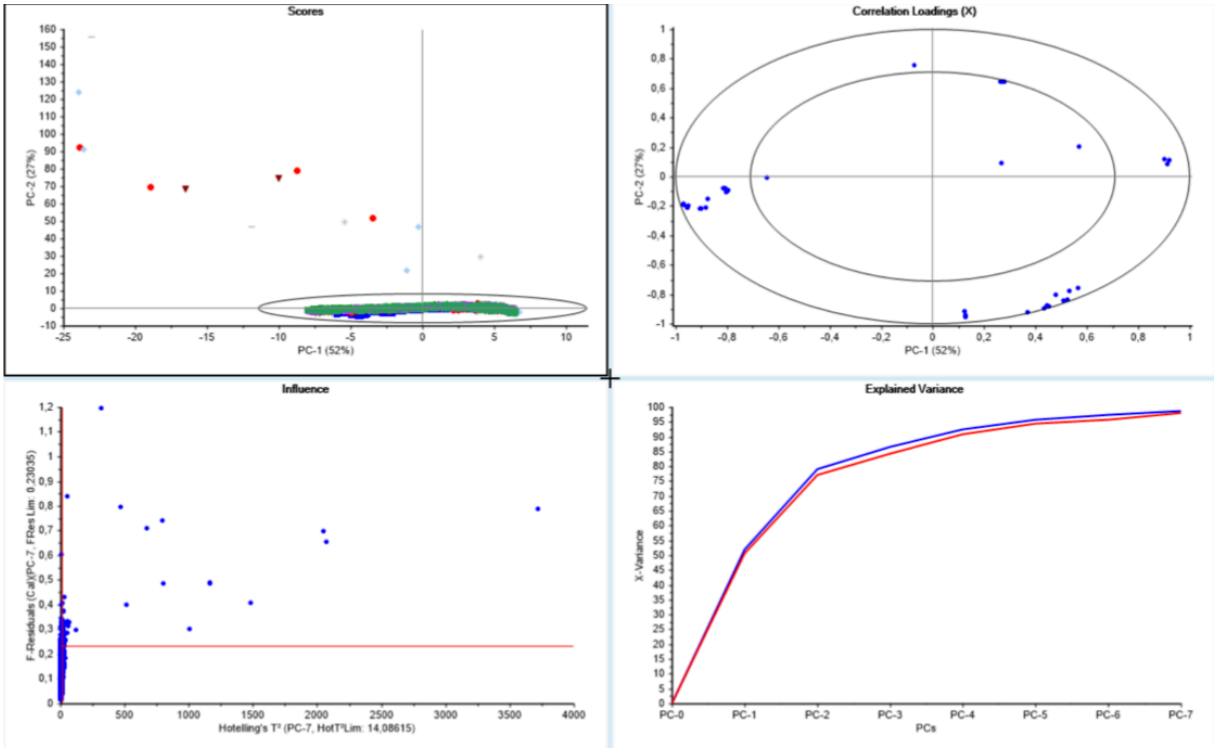
Correlation loadings for PC-2 and PC-3 (Factor-2 and Factor-3) for PLSR to the left and PCR to the right. PCA correlation loadings for PC-2 and PC-3 under, but the same as PCR.



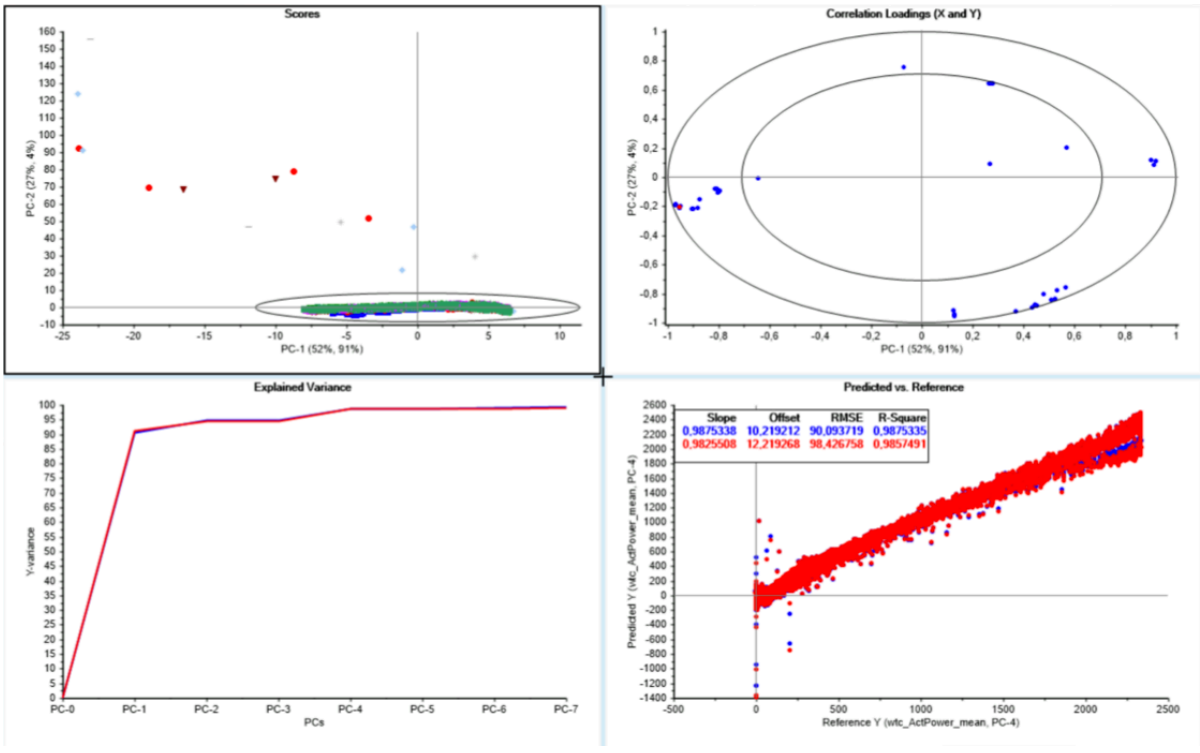
Here we can observe that there is more difference than in the lower PCs.

Make models with selected column subsets and the column set "AllX"; compare results
Find the optimal number of PCs/Factors.

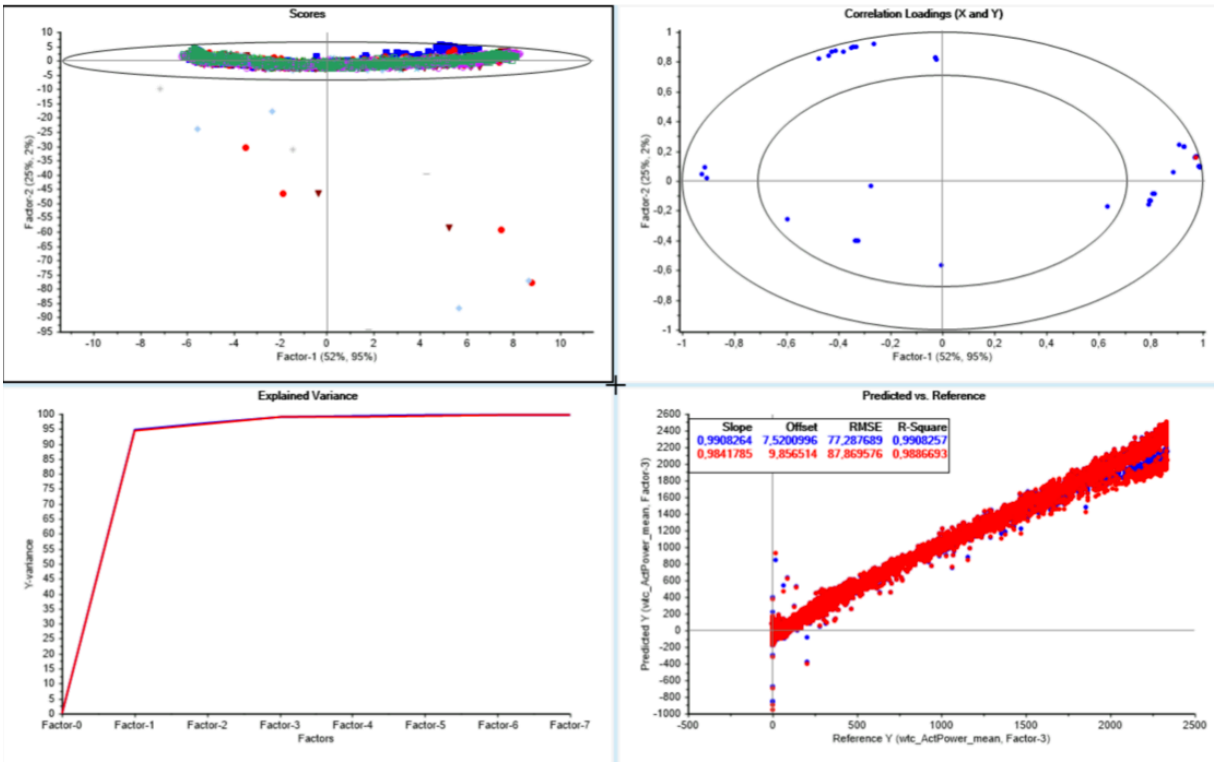
PCA - Grid



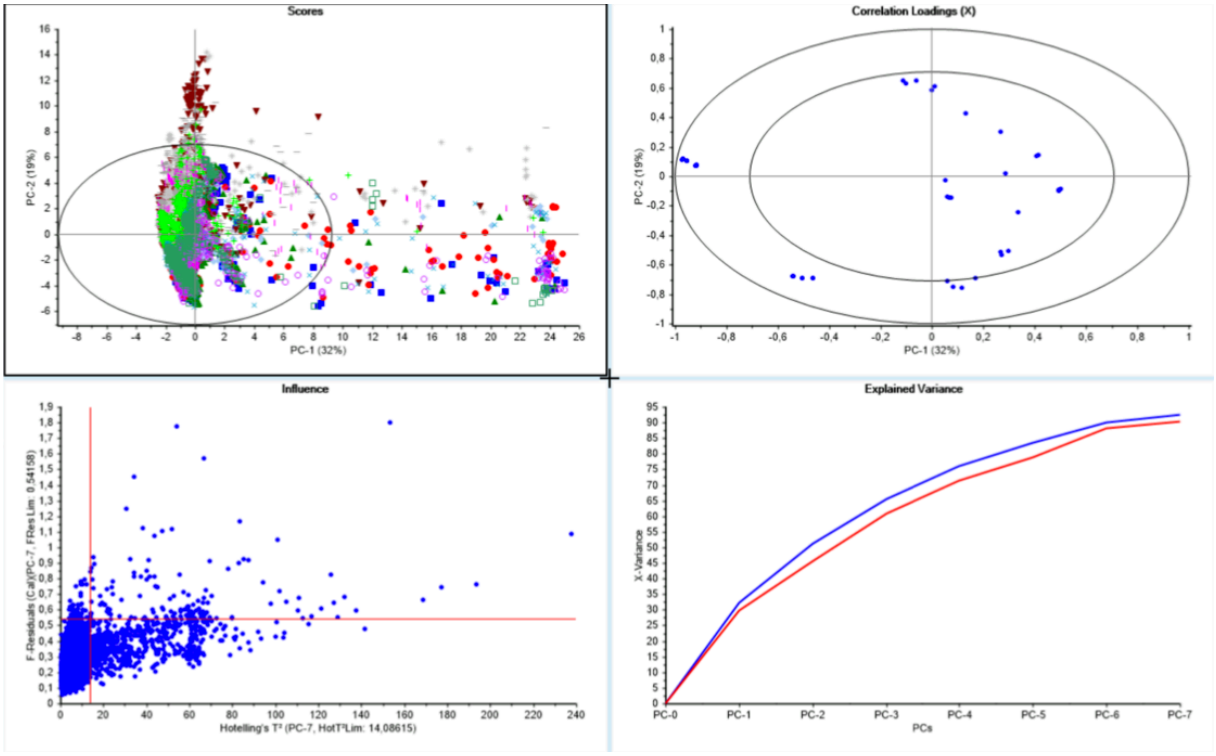
PCR – Grid



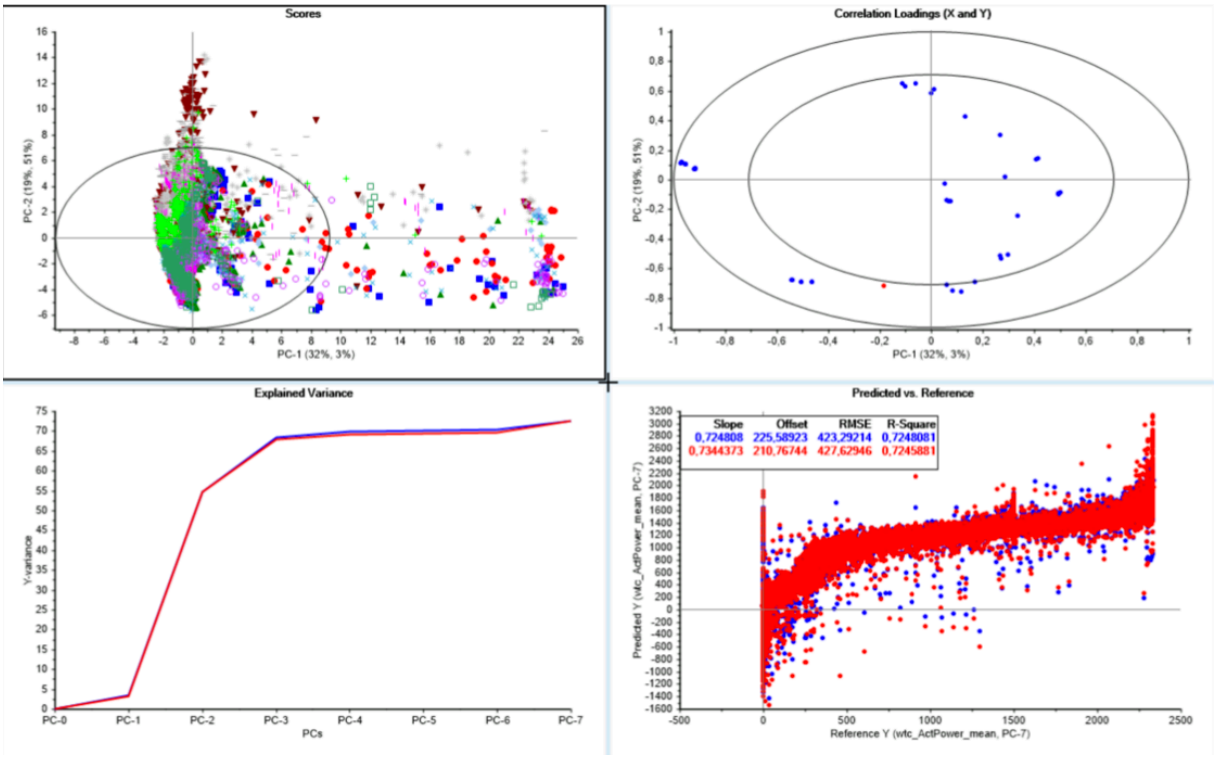
PLS – Grid



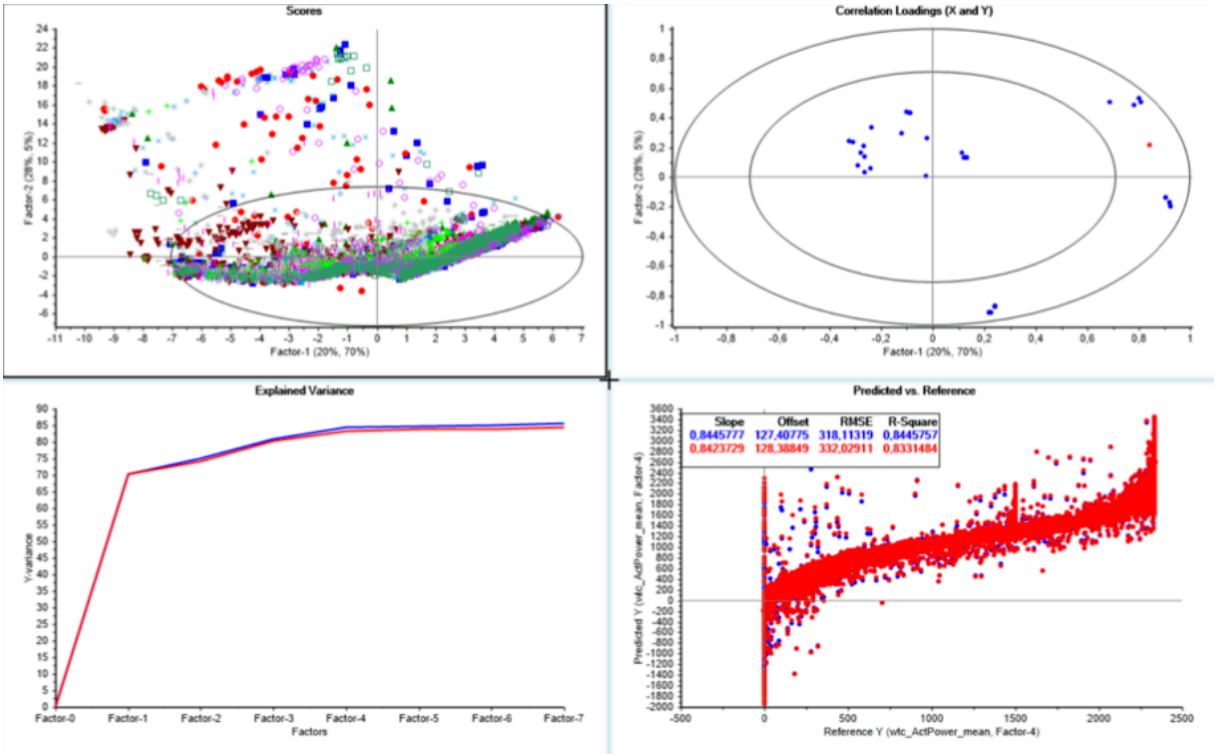
PCA – Turbine



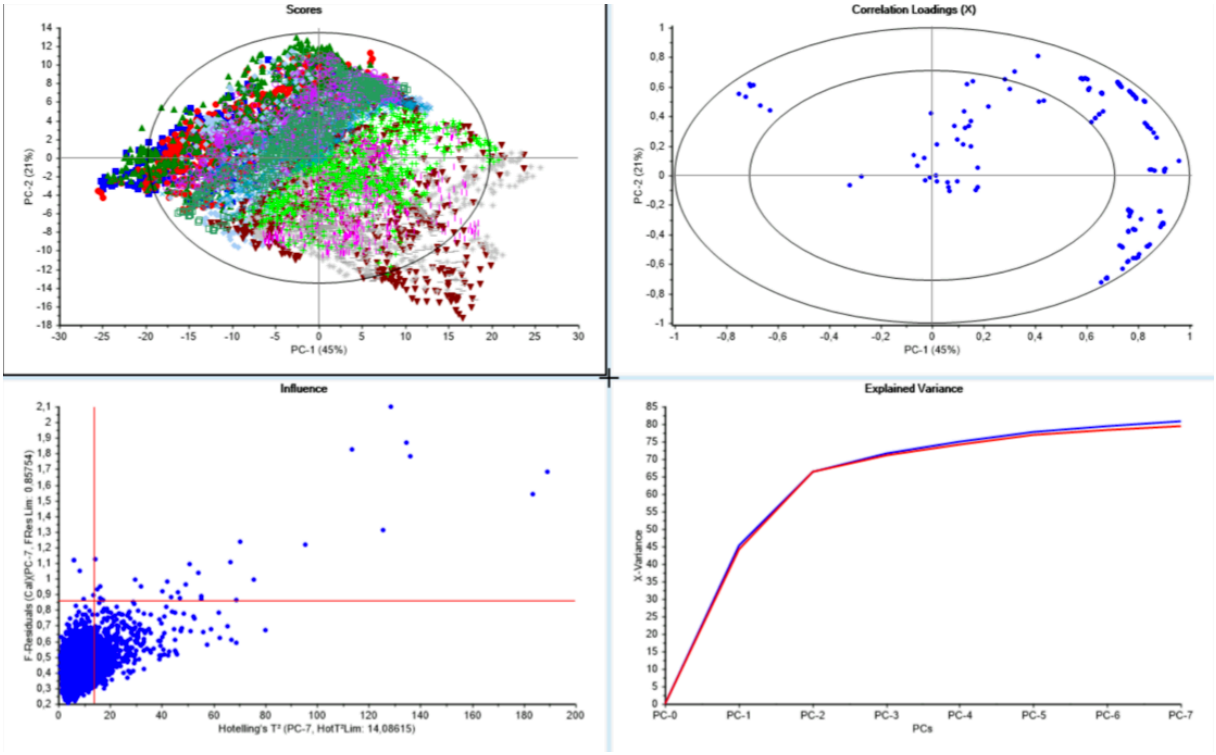
PCR – Turbine



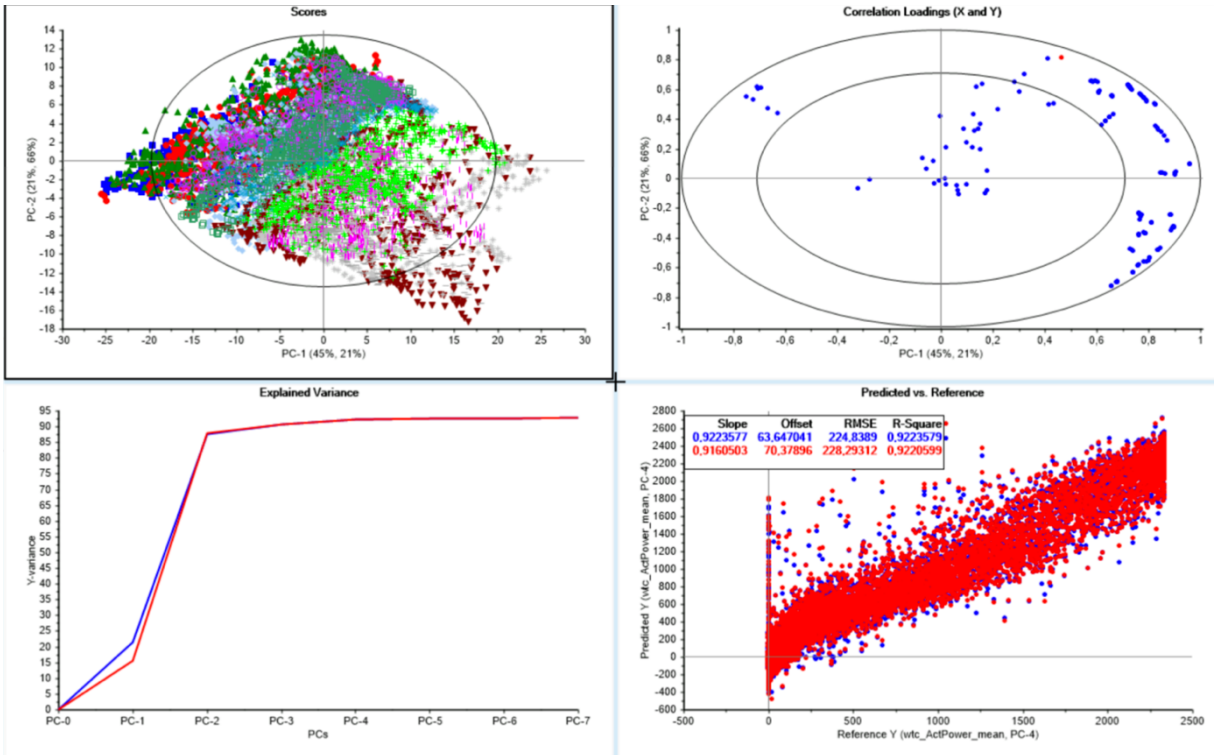
PLS – Turbine



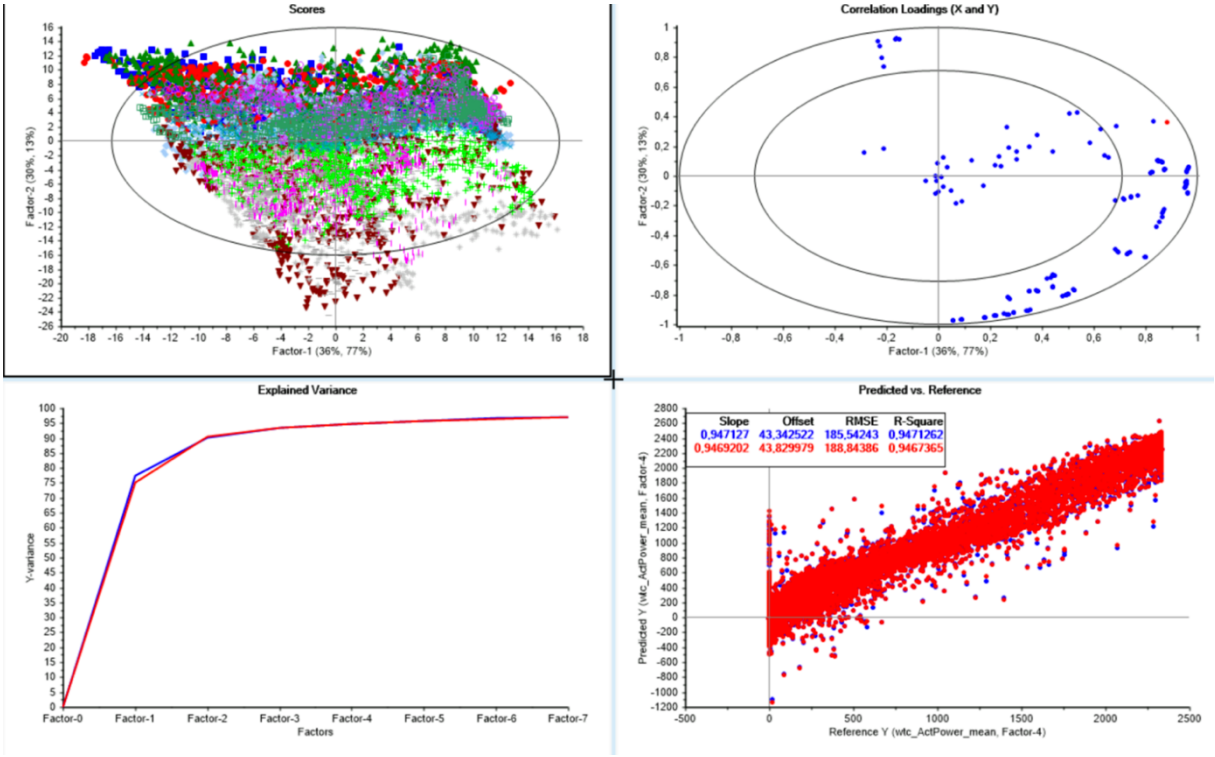
PCA – temperature



PCR – Temperature



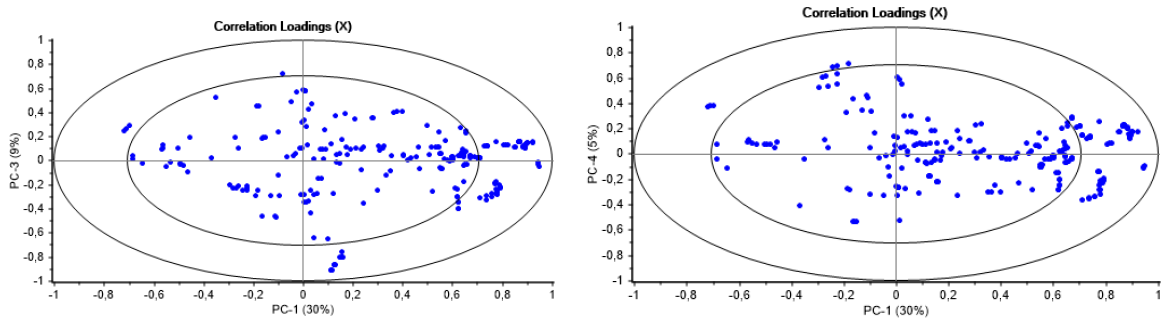
PLSR – temperature



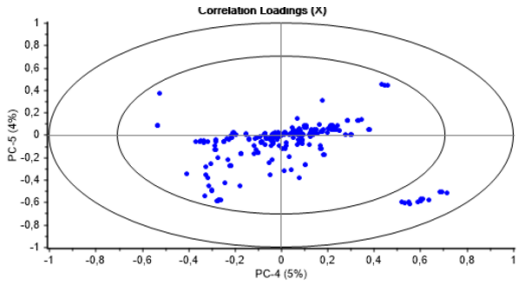
Both Grid and Turbine seem to have a lot of outliers for all plots, but not Temperature. The R^2 values for Turbine is lower than the ones for Temperature and Grid, they are quite high at over 0.9.

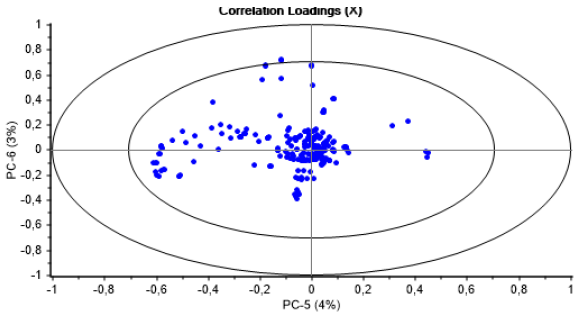
Find the optimal number of PCs/Factors

PCA: PC-1 and PC-2 explain a lot of the variance (50% in total), but the graph does seem to grow after PC-2 as well, so maybe 5 PCs.



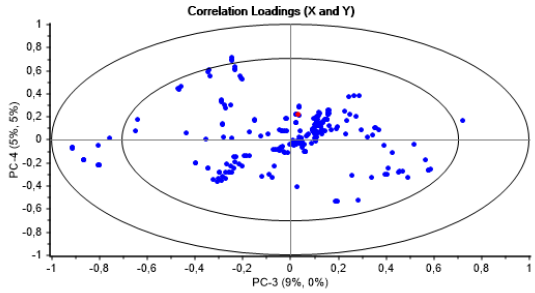
Looking at the correlation loadings plot for PC-3 and PC-4 they contribute 9% and 5%, but they might be important still as the variables contributing to them might not be of the same importance for the other, lower PCs.

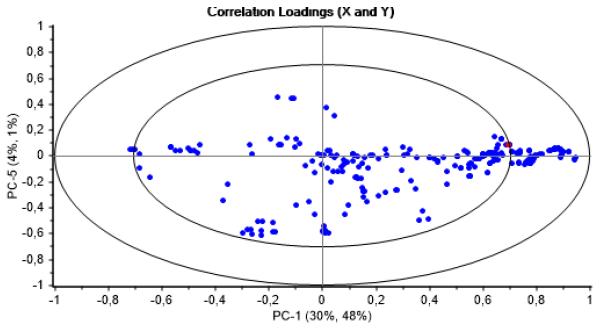




Looking at the correlationg loadings for PC-4, PC-5 and P_6 it is clear that the loadings are gathering close to the center. For PC-4 and PC-5 there are some variables lying outside the inner ellipse, towards the outer, but for PC-5 and PC-6 there are only one variable outside the inner ellipse. Thus the number of PCs should be 5 for PCA.

PCR: Looking at the explained variance in the PCR graph it is clear that most of the explained variance is covered by PC-1 and PC-2. PC-3 and PC-4 could be included, but after this the graph flattens out.

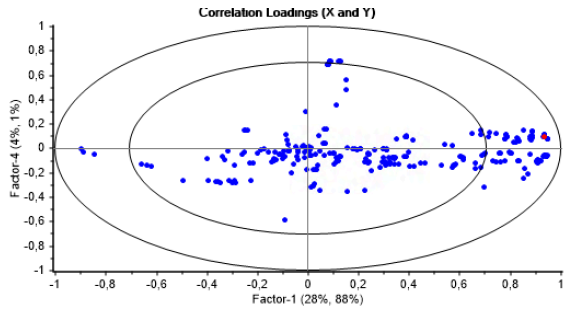


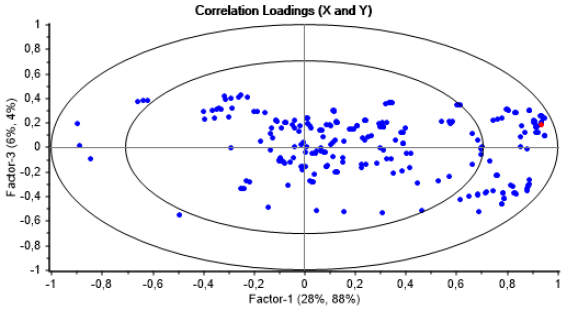


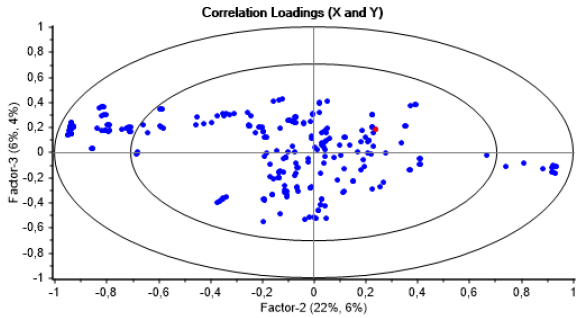
The above plots show that PC-3 covers 9%, and the correlation loadings for PC-3 and PC-4 does have variables outside the inner ellipse, some close to the outer. The second figure on the other hand shows that PC-5 is redundant. Thus four is the optimal number of PCs.

PLS: Looking at the explained variance plot for PLS it looks like PC-1 covers most of the explained variance. The graph is flat after PC-3, my guess is therefore that 3 PCs is optimal.

Taking a closer look at the correlation loadings to not guess, but be certain of the optimal number of PCs:

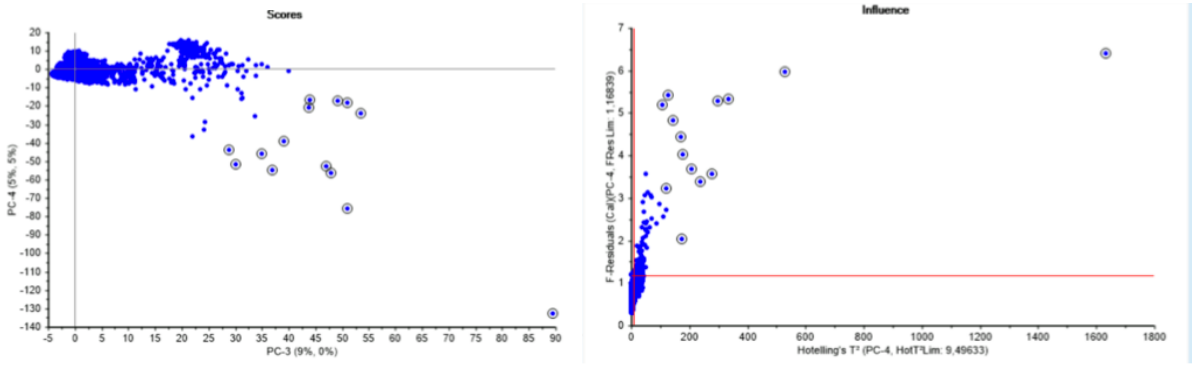




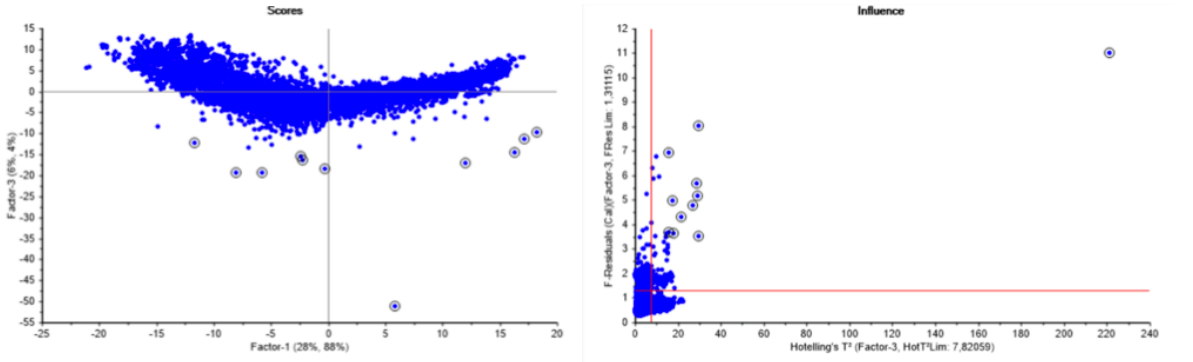


The plots confirmed my guess. I would say that the optimal number of PCs is 3 for PLS.

Identify outliers and remove them if it can be justified. Does the RMSE improve after removing them?



By removing some of the outliers in the PCR model the RMSE value is improved from 205.2/208.1 to 192.1/198.7.



By removing outliers from the PLS model the RMSE has a small improvement from 141.3/147.0 to 140.8/146.5.

Test different validation schemes, compare RMSE

By using a cross validation systematic 112233 PLS the RMSE is 140.3/147.1, and using the same validation scheme for PCR we get a RMSE value of 192.1/198.5. By using systematic cross validation the results are more similar to using cross validation with categorical values.

Select a subset of the variables by marking in the correlation loadings plot and/or regression coefficients plot. What is the least number of variables to keep and still have the same RMSE as with all variables (or the best subset of X)?

By selecting subplots and recalculating the results show that the RMSE-value stays the same for the same number of variables decided by PCs and factors. This can be due to choosing variables strongly correlated to Y.