Untitled

May 17, 2021

1 Assignment 5

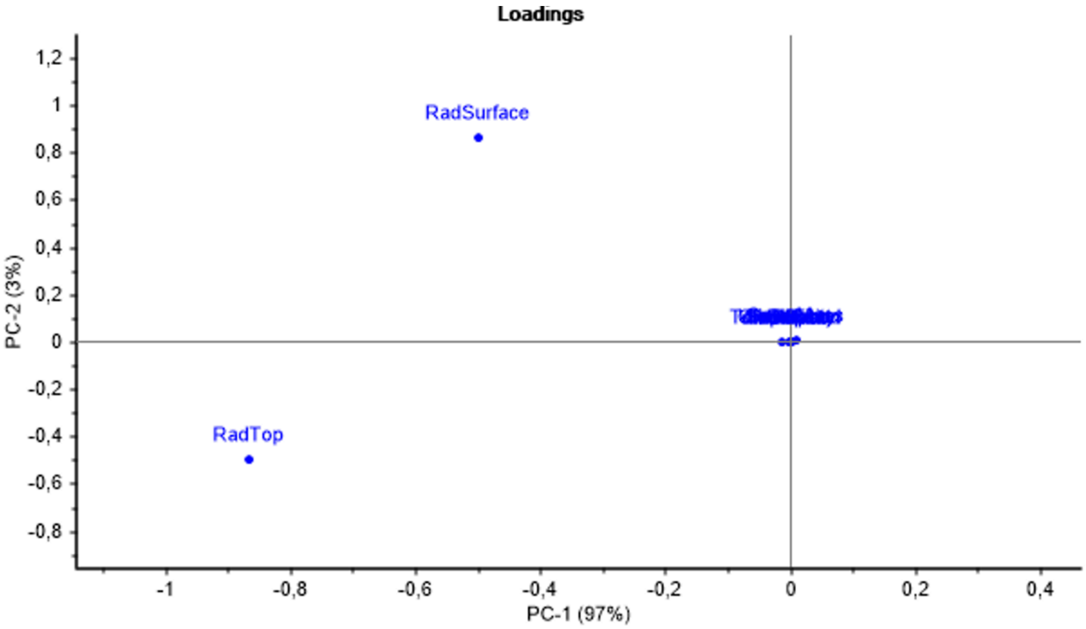
1.1 Part 1: Make PCA of the data with weighting = 1 and random cross validation.

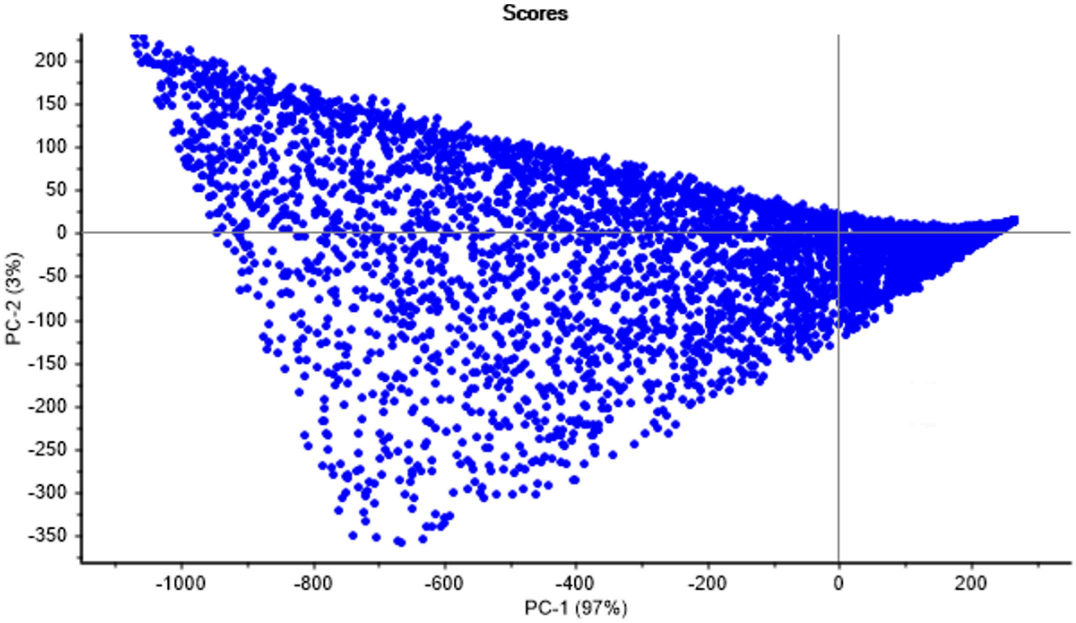
First I made the PCA according to what described in the task, with weighting = 1 and random cross validation.

1.1.1 Scores and Loadings:

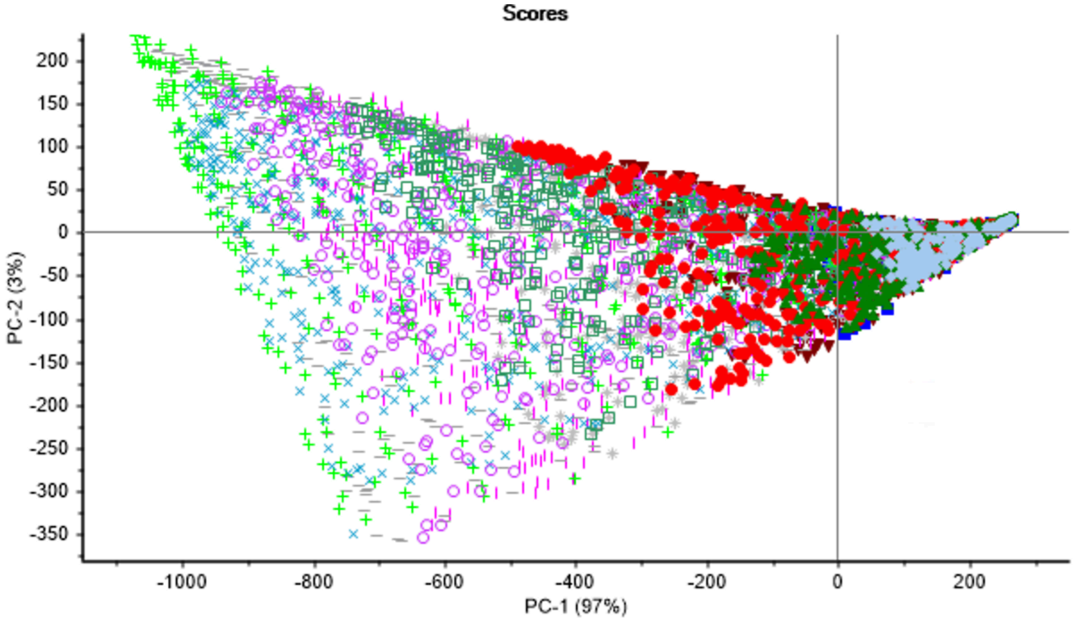
Below one can see the scores and loadings plots. The loading plot shows how much a varible is weighted when it comes to describing as much of the variance as possible. From this plot it is clear that most of the variance in the data set lays in the RadSurface and the RadTop variables. When we use 1 as weighting of the different variables the varibles with the biggest number will dominate the analysis. If one take a look at the columns of RadTop and RadSurface in the data set one can see that these values varies from 0 to several hundreds. Thus the variance in these variables will be big.

The scores plot shows the weighting of the different samples. One can see the scores and loadings plots in combination and the values up in the left hand corner will have a high value in the RadSurface variable. Likewise, the data in the lower left hand corner will have a high value in the RadTop variable.



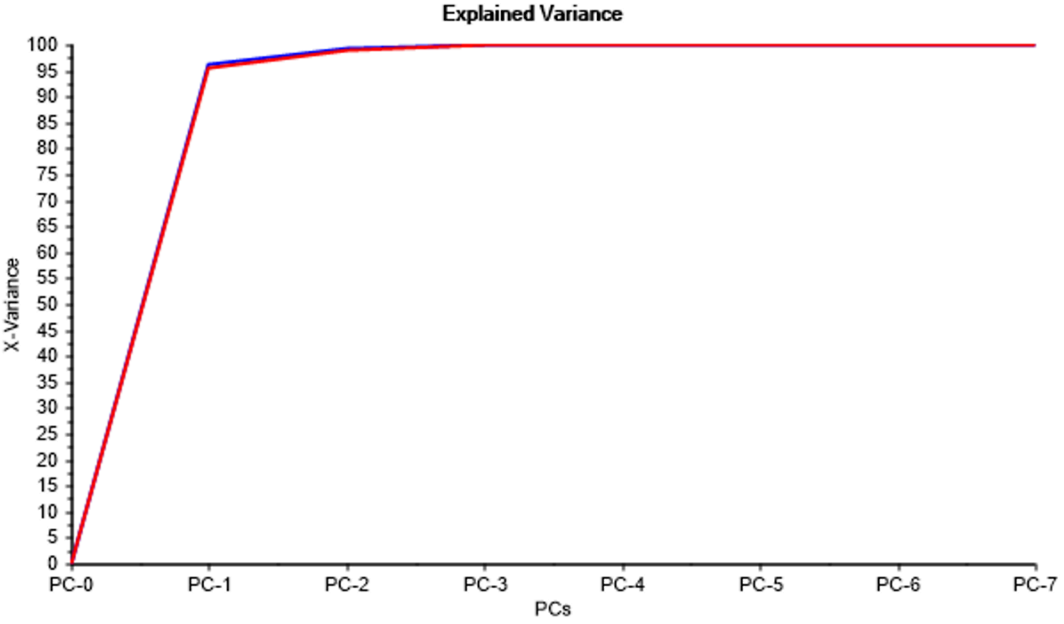


By adding grouping based on the different months to the scores plot one can see a clear pattern. As stated earlier, the scores in the lower and upper left of the plot has a higher value in the RadTop and RadSurface variables. And by inspection the the summer months with high radiation are much more represented on the left side of the plot, and vice versa, the winter months are represented on the right side.



1.1.2 Explained variance:

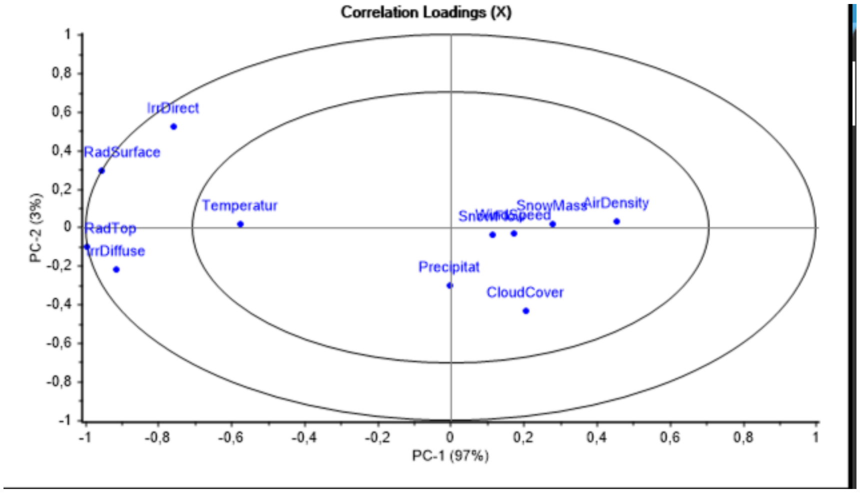
From the below plot showing the explained variance one can see that PC1 explanies 96% of the variance alone, and by also including PC2 one can describe 99% of the variance by only looking at two dimensions of the data.



1.2 Correlation loadings:

The correlation loadings shows the correlation between the score vectors and individual variables. It shows how much of the total variance of a variable that is explained by the PCs. If a variable laies on the outer circle in the correlation loadings plot 100% of its variance is explained by the PC. And if it lais in the inner circle, 50% is explained.

From the below plot one can see that for the RadSurface, RadTop, IrrDiffuse and IrrDirect variables almost 100% of the variance is explained. And for the rest of the variables under 50% of the variance is explanined by this model.



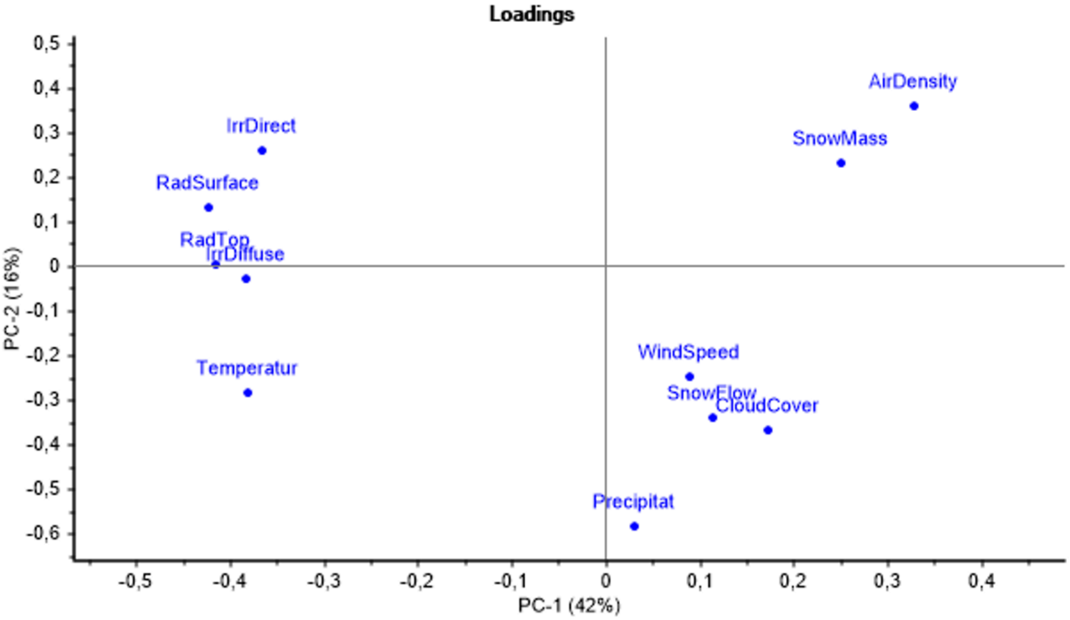
1.3 Part 2: Make PCA of the data with weighting = 1/Std.Dev and random cross validation.

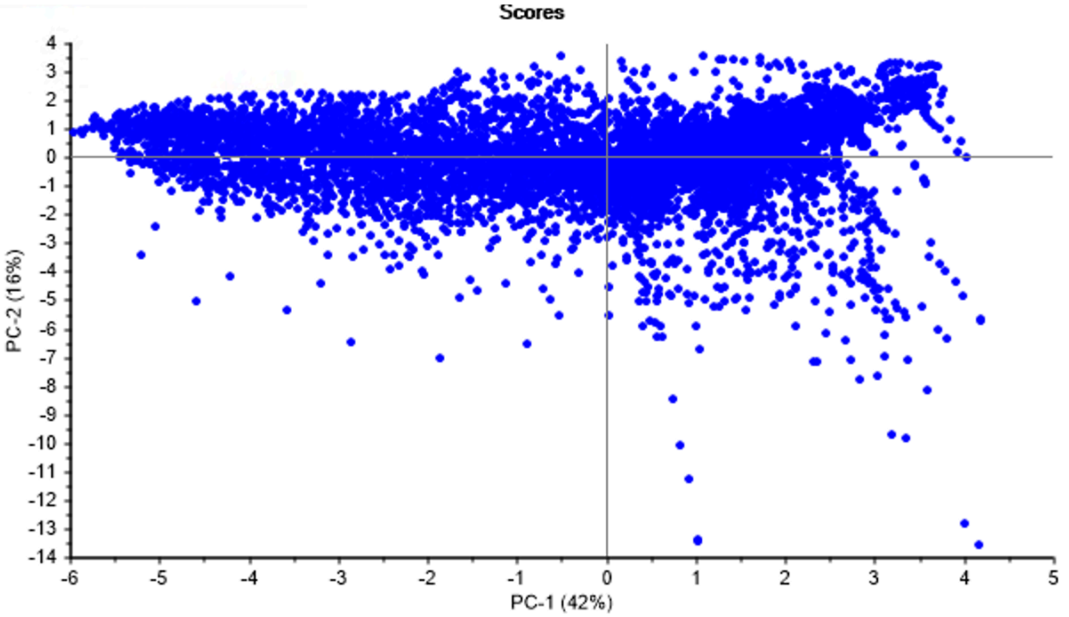
Now the PCA is computed with weighting equal to 1/Std.Dev of each variable.

1.3.1 Scores and Loadings:

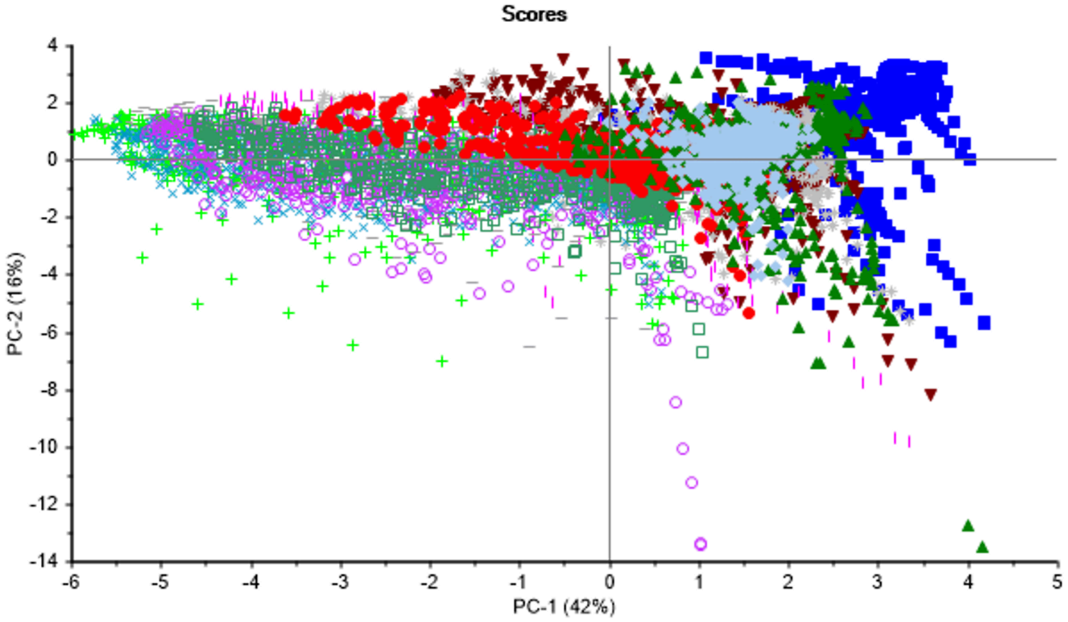
Using 1/Std.Dev as the weighting of the PCA results in a PCA that takes all of the vairbales into account. In the last part the PCA weighted all differnt variables the same and did not take the unit of the variable into account. This resulted in the variables with the highest values and biggest variance being represented much more. In this case all variances are divided by their standard deviation and thereby scaled.

From the loading plots it can clearly be seen that the variance of all different variables are taken into account. One can also see both positive and negative correlations between the different variables. For example the IrrDirect, RadSurface, RadTop and Irrdiffuse are naturally positively correlated. And the temperature and snow mass are negative correlated.



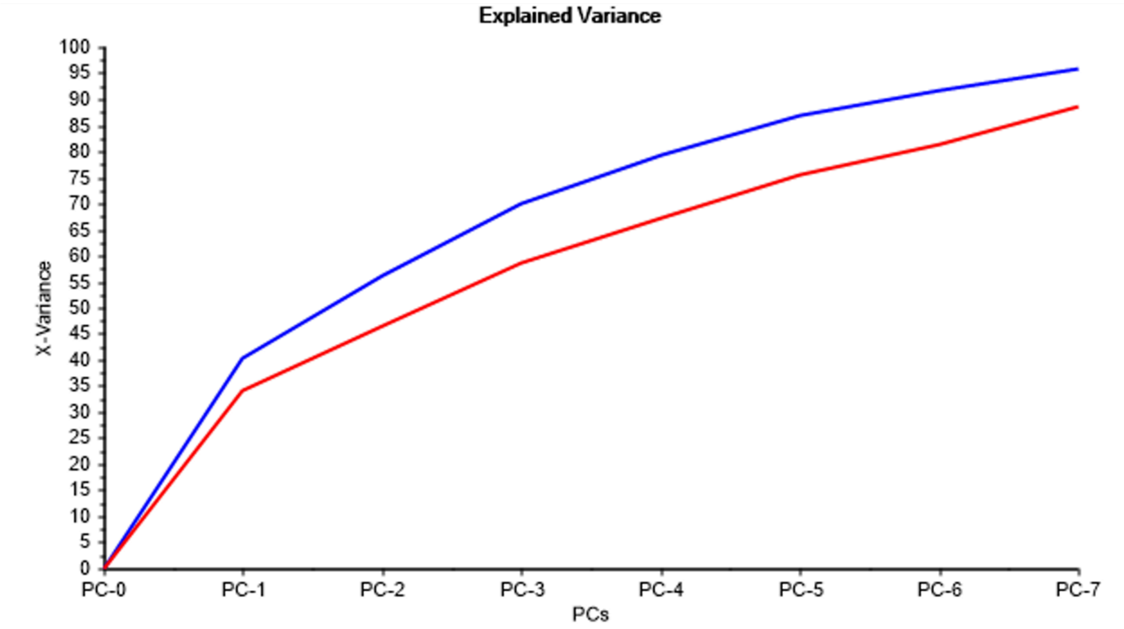


Again if we apply gruping on the different months on the score plot we can see clear connections with the loadings. All the samples on the leftmost side of the plot are from the summer months, something that as natural as the summer most has the most sun. Also, all the samples in the upper right are from the winter months, as these months has the most snow and lowest temperature.



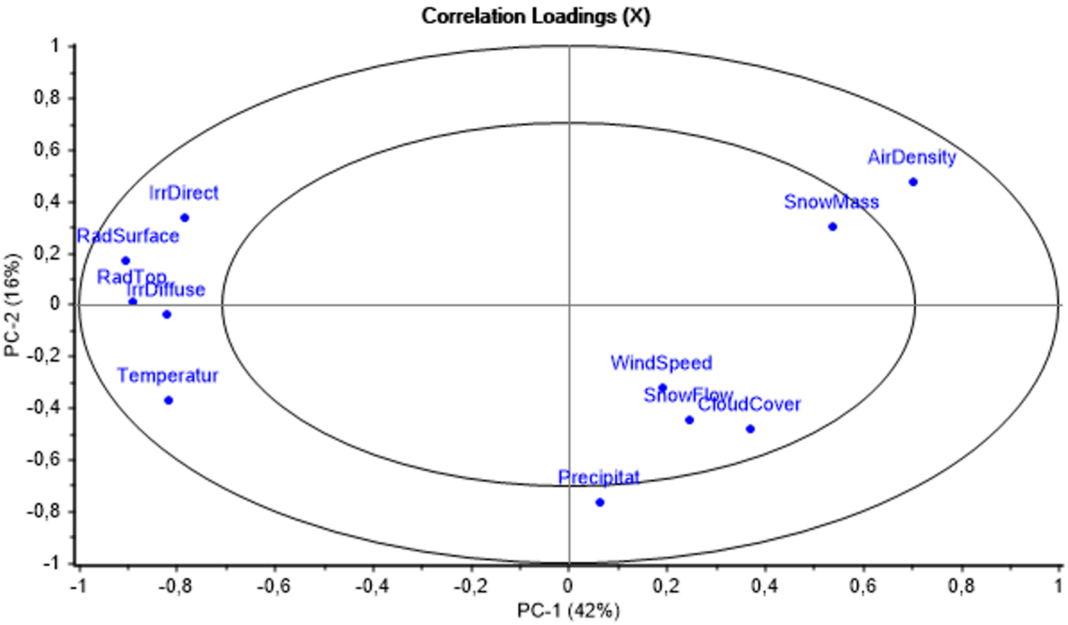
1.3.2 Explained variance:

For this PCA the PC1 and PC2 explaines much less of the total variance of the data set. This is natural, as we now are weighting all the differnt variances the same way and thus the variance are distributed into many more dimensions than in the last part. For this PCA



1.3.3 Correlation loadings:

By taking a look at the correlation loadings in this case it is clear that the PCA model explains a greater part of the variance of the different variabels than the one in part 1. Ecspecially the variables that was around origo in the last part are now more represented with almost 50% of their variance explained. This is a result of the new weighting of the data, as this results in the variance of the variables being more representative.

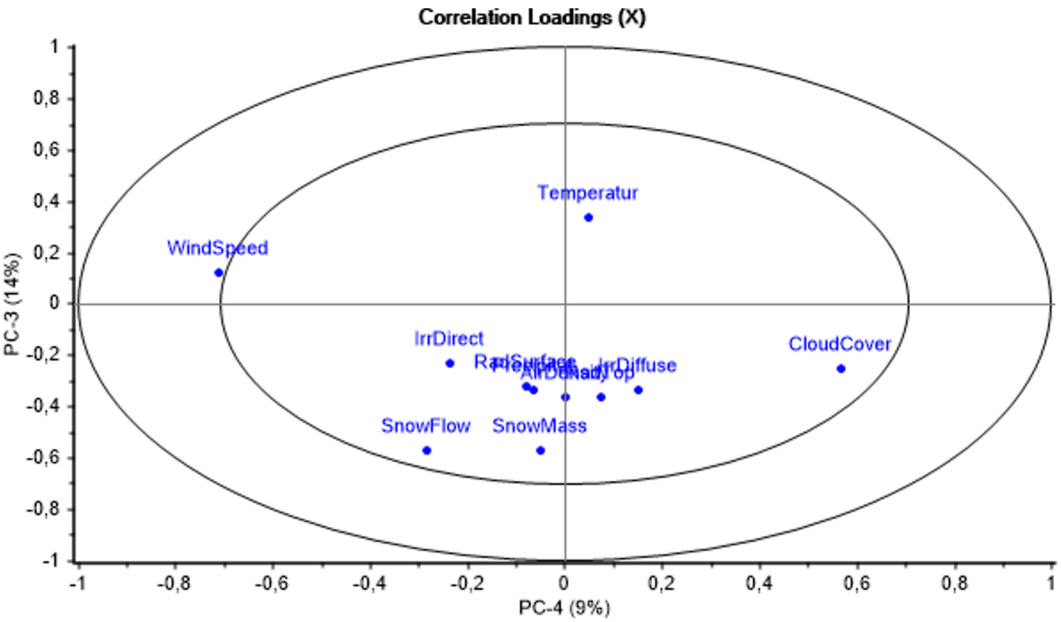


1.3.4 Optimal number of PCAs:

For this model PC1 and PC2 does not explain more than 56% of the total variance. Because of this one would include more of the PCs in order to get an acceptable amount of explained varinace. From the above plot showing the explained variance one can see that one need to include 5 or 6 PCs to get the amount of explained variance we had in the last part.

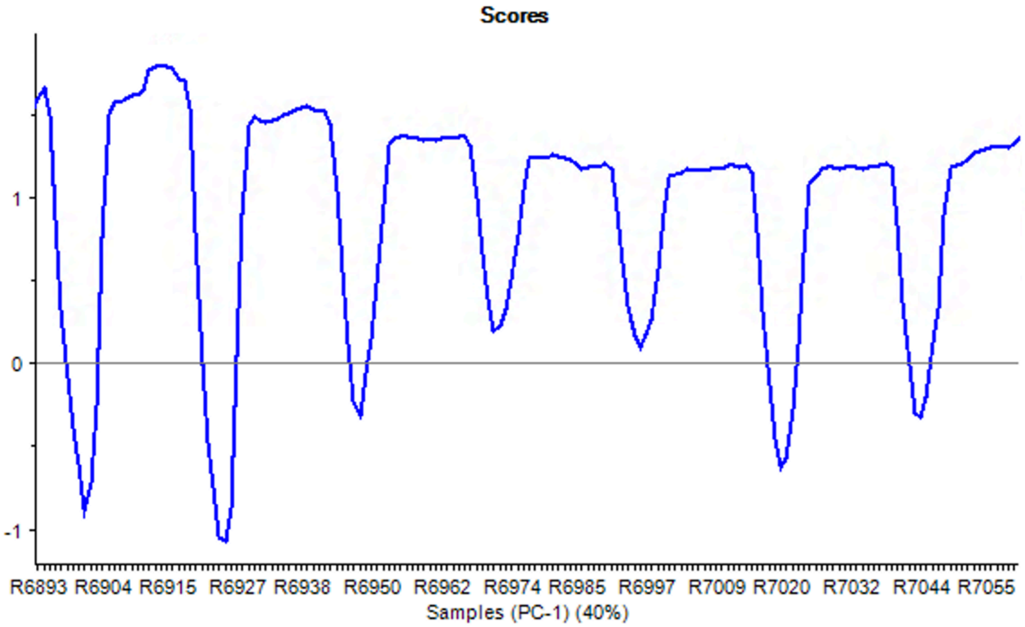
By inspecting the different PCs scores and loadings plots it is possible to extract what varibles the different PC catches. For examples will PC1 explain much of the variance correspondig to the radiation and irratdiaton. PC2 explains much of the variance connected to precipitation, snow mass and flow and air density. PC3 is just a combination of many of the variables, while PC4 explians tbe wind speed and cloud cover.

From the above one can conclude that the number of PC to include depends on what you are trying to analyze, as the different PCs represents different features of the data. But including 3 or 4 is probably enough in this case.



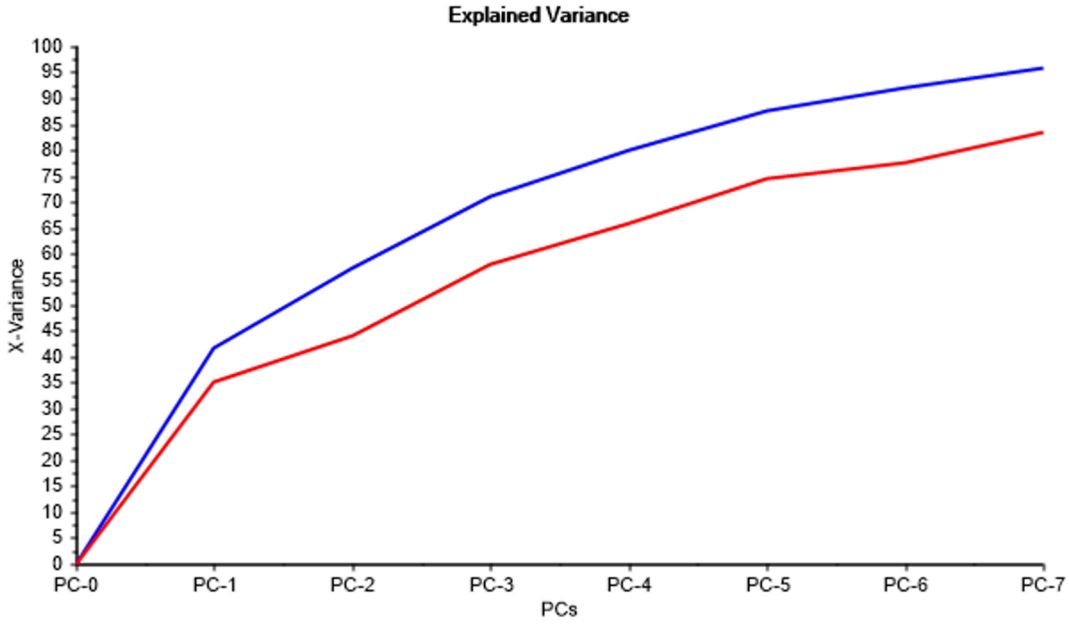
1.3.5 Line Plot of Scores:

Below you can see line a zoomed in line plot showing some of the daily variations. Each day consists of 24 samples. You can clearly see patterns.

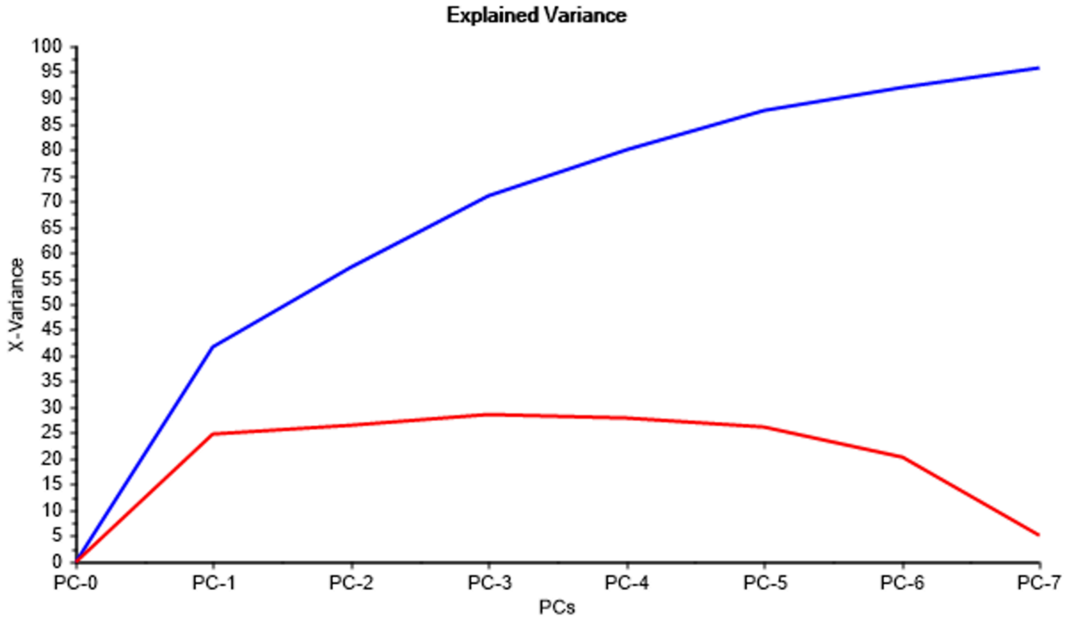


1.4 Part 3: Systematic validation

When using systematic validation I get exactly the same results as in the last part, exept for the validation in the explained variance is lower. This is due to the fact that the model no longer is validated randomly and thus performs worse on the validation. The explained varinace can be seen below



When changing to validation on day and night category pairs the data validation gets even worse.



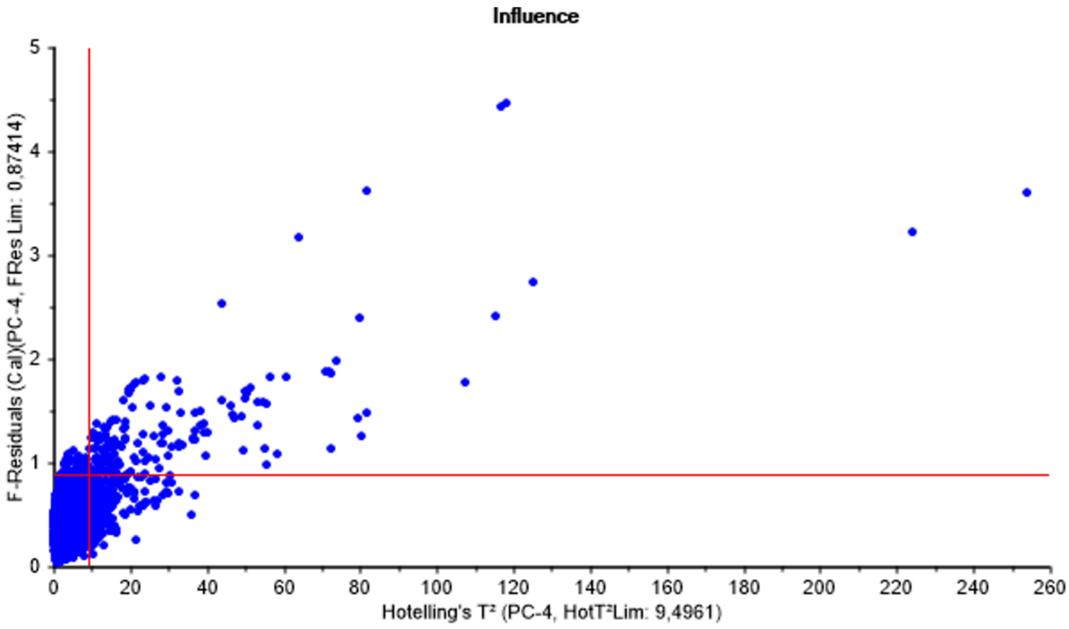
For these models it is only the validation part of the PCA that differs, and the PCA itsself is actually the same. THus, although it looks like it performs worse, the two latter PCA will do a just as good of a job and the same amount of PCs will be needed in order to achive the same performance.

1.5 Part 4: Look into the Hotelling´s T^2 and F-residuals plots and check for any outliers.

I decide to go for 4 PCs.

According to the below plot there may be many ouliers. An outlier will often lay in the big squar in the middle of the plot. Thus the points up in the right corner of the plot are probably outliers. By further inspection this point may be an outlier due to the unusually high snow flow. This sample is from the month of april where the snow flow uasually is 0. The same is the case for the sample second most to the right. This sample had a much larger snow flow then usual. The two samples up on the left had a higher wind speed than usual in that period, and thus is marked as an outlier.

After inspecting sveral of these possible outlier samples it looks like all of them comes from some variable with a higher or lower value than usual in that period.



1.6 Part 5: is it conceptually viable to include year, month and/or day/night if the purpose is to project new samples onto a training model for detecting changes in the x-variables?

As you can see from the above task there is a difference between the samples taken in the different months and day. This can be seen from the explained varience when the validation is done on day/night and how bad the model then performs. Thus in order to project new samples onto the model it is important to know if a sample is a day or a night sample and what month it is taken from.

1.7 Part 6: Project the 2017 samples on the 2017 PCA.

[ ]: