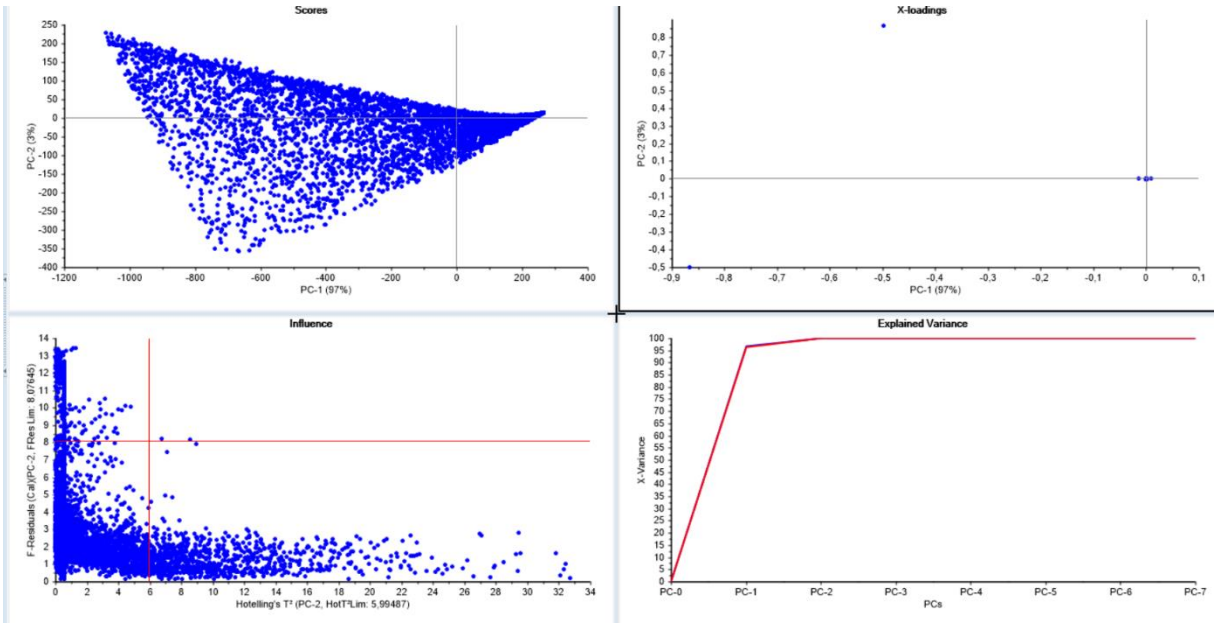
Assignment 5

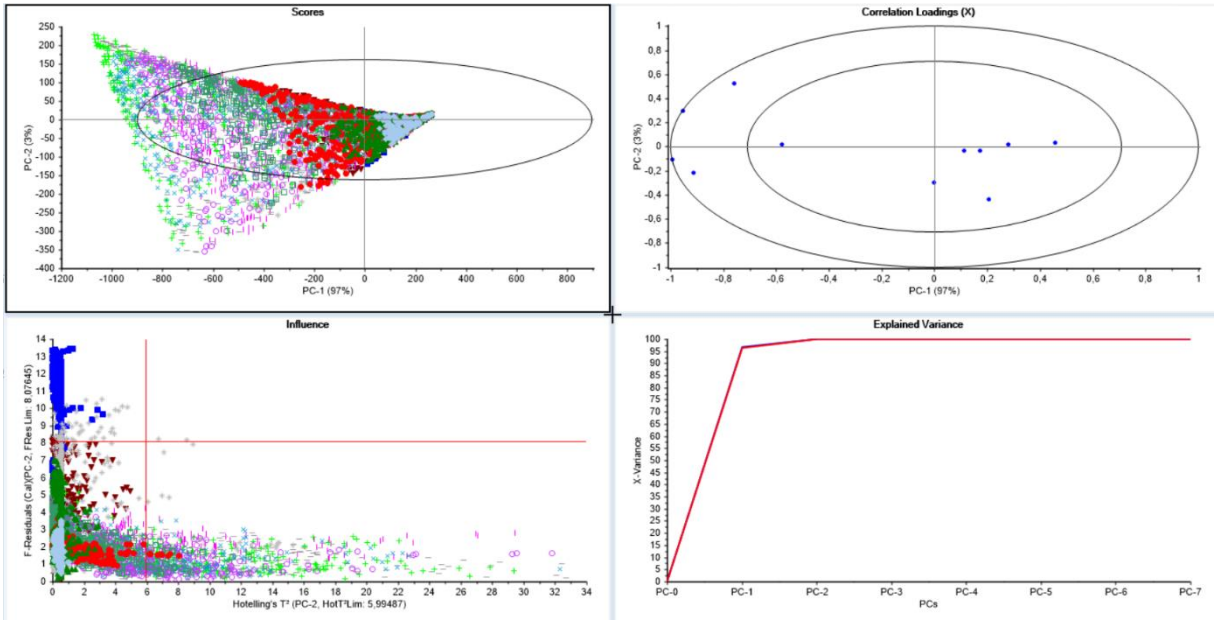
Part 1: PCA with weights = 1



1. Interpret scores loadings and explained variance

Answer: For Scores we see that PC1 represents more of the data than PC2. From the loadings plot we see that variable 9 and 10 are explained very well in PC1 and PC2. From the explained variance most of the variance is explained in PC1 (96%). The addition of PC2 only gives us an additional 3% explained variance.

Part 2: PCA with weights = 1, correlation loadings and Hotelling's T 2 ellipse

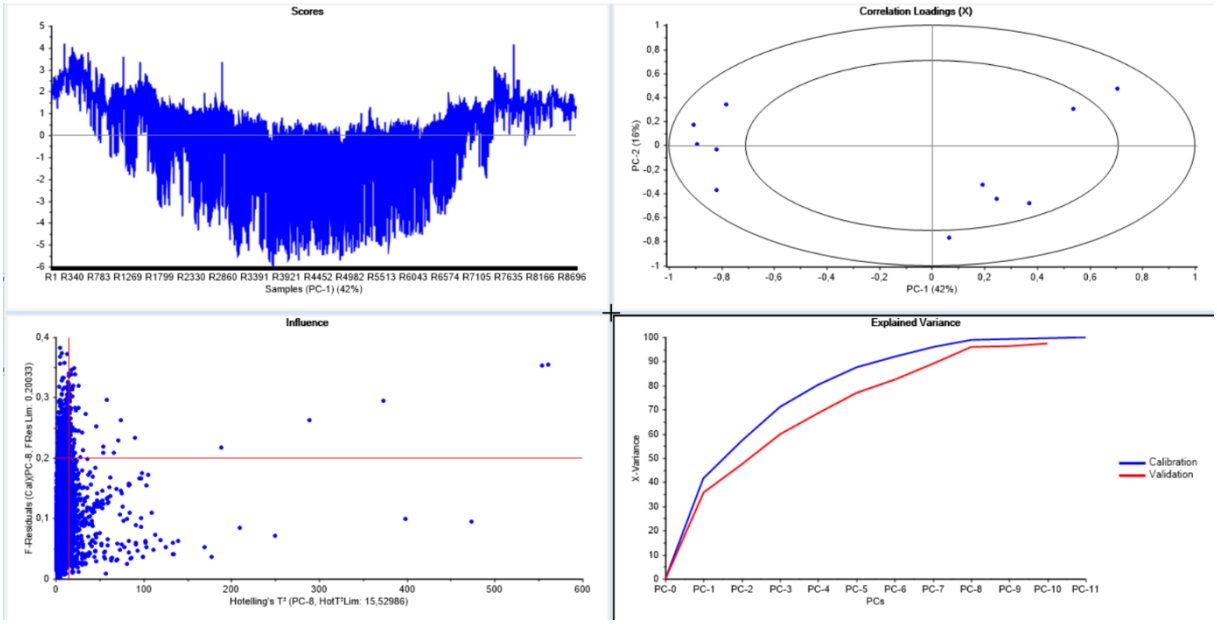


2. Explain why the plot has changed.

Answer: Loadings: Using Correlation loadings we can see the data structure more clearly, and the importance of each variable. The correlation loadings shows the correlation between variables

and the explained variance along different components. Also groupings of variables shows how different variables are correlated to each other.

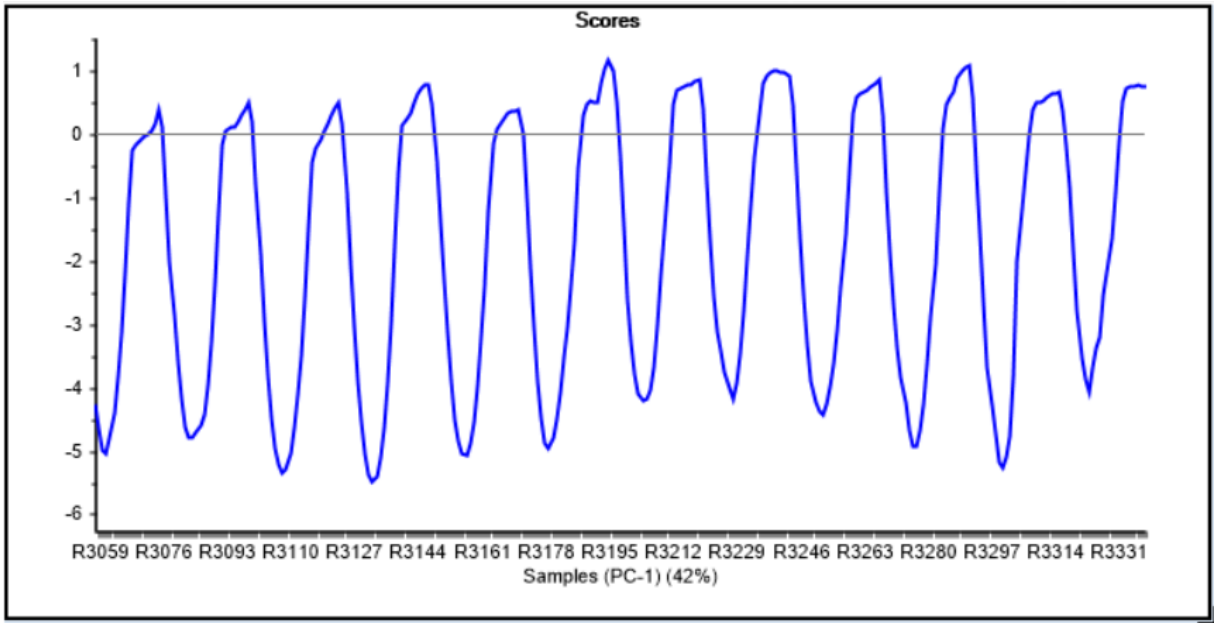
Part 3: PCA with weights = 1/Std.Dev



3. <u>Interpret the model again. Decide on the optimal number of PCs</u>

Answer: After 8 components there is very little increase in explained variance. Therefore the optimal number of PCs is 8.

Part 4: Scores as line plots

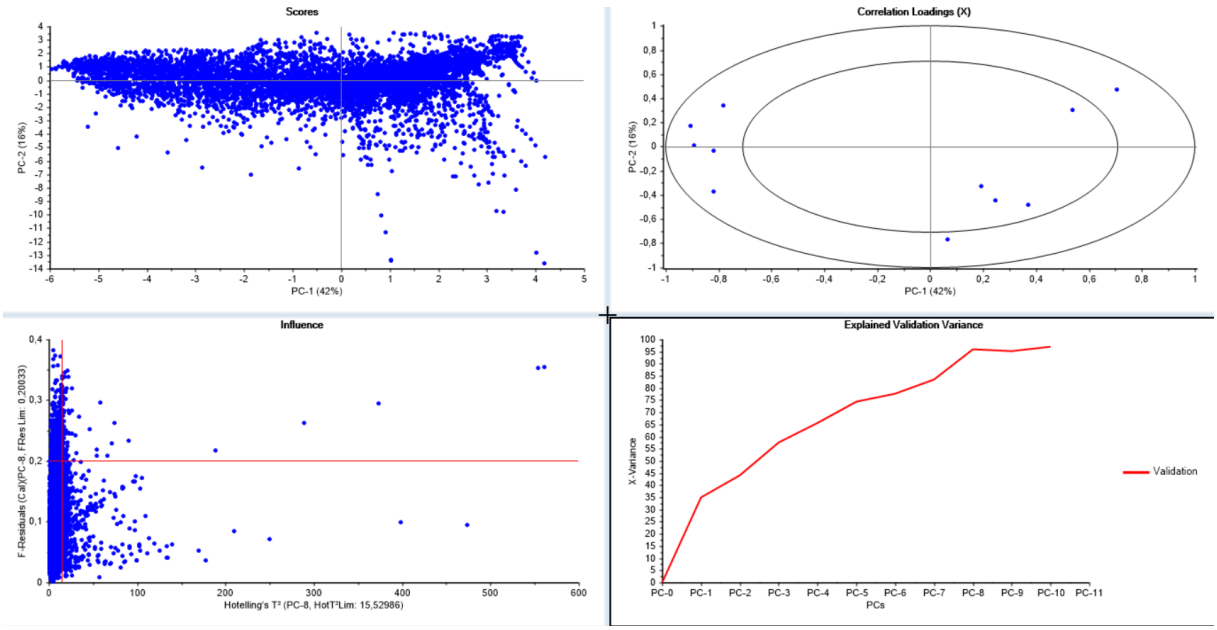


4. <u>Zoom in to show only scores for some days (use the Frame scale). Do you see the daily systematic pattern?</u>

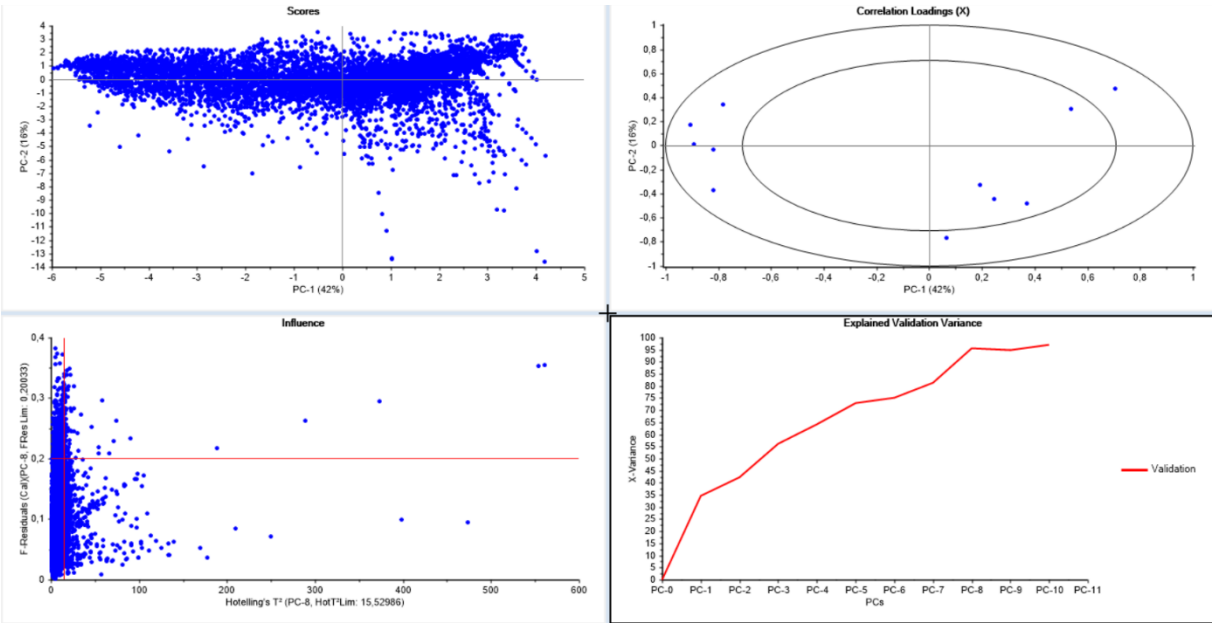
Answer: We see that the scores oscillate in waves. The pattern is clear as the amount of sunlight varies systematically throughout the day.

Part 5: Try other cross validation set-ups:

<u>Systematic (111 222 333)</u>



<u>Category variables (day/night, month)</u>



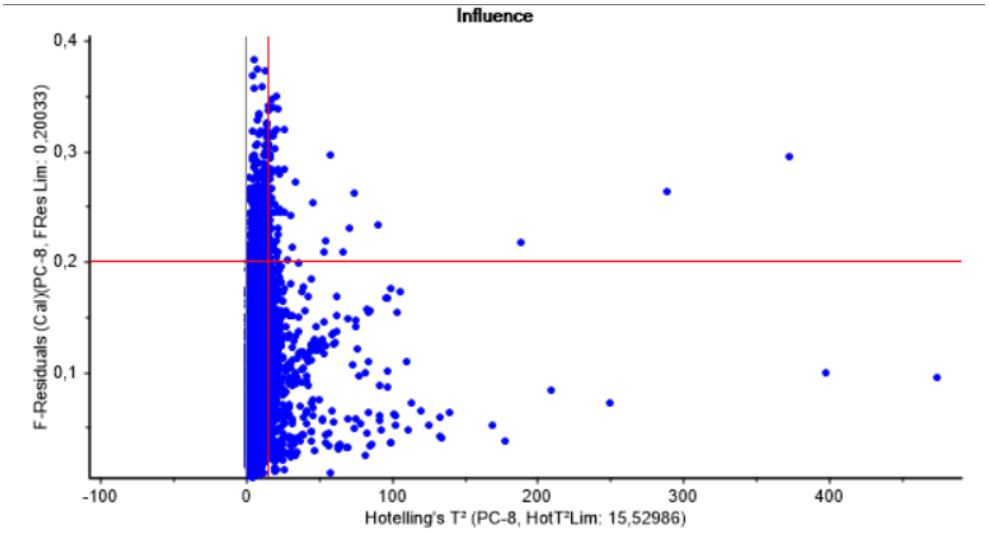
5. Compare the explained validation variance for these model

Answer: The explained validation variance is almost the same for Systematic (112233) and Category variables (month) cross validation set-ups. But it is slightly higher for Sytematic.

6. Decide on the optimal number of components (compare to the model with random CV above)

Answer: For both Systematic (112233) and Category variables (month) cross validation set-up, the optimal number of components is 8, as seen from the low increase in explained variance. This is the same number of optimal components as with random CV.

7. Look into the Hoteling's T 2 and F-residual plots if there are any outliers (NB! Decide on the optimal number of PCs first)

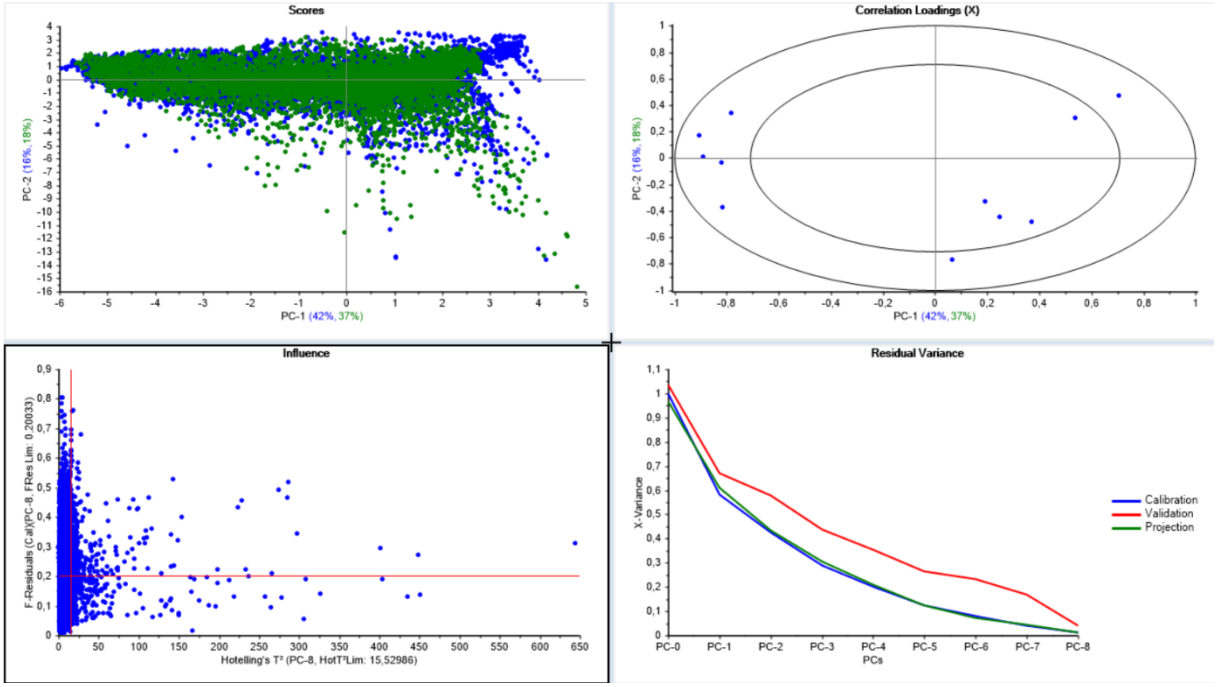


We see that we have data outside of the Hotellings and F-residual limit. These datas would be considered outliers.

8. Discuss if it is 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

Answer: It is not conceptually viable to include all variables (day/night, month, year) as day and night are dependant on each other and year is constant.

9. Project 2017 dataset onto model based on 2016 data set (symstematic (112233) with 8 components)



dd

Answer: