Multivariate calibration

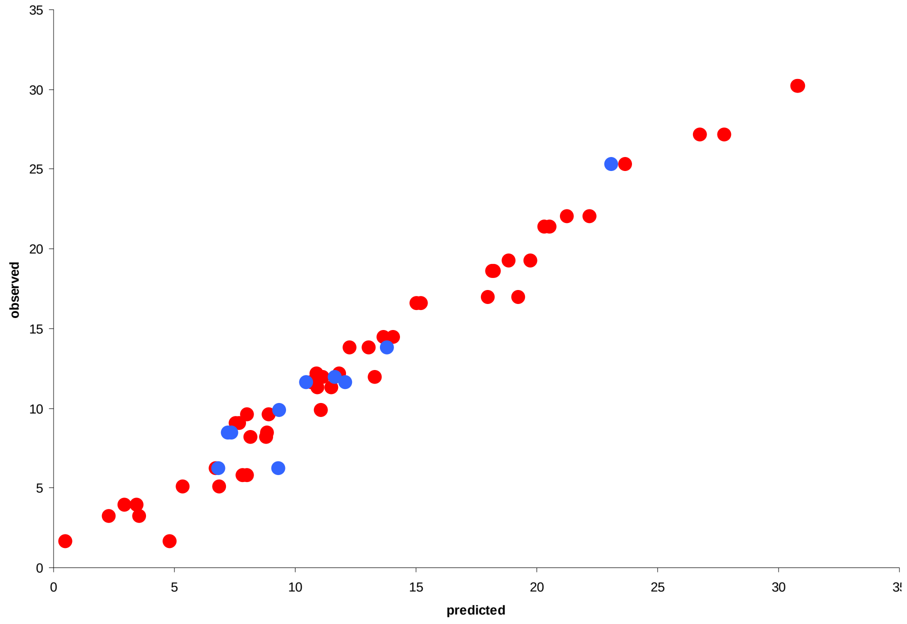
- What is calibration?

- Problems with traditional calibration
  - selectivity
  - precision
  - diagnosis

- Multivariate calibration
  - many signals
  - multivariate space

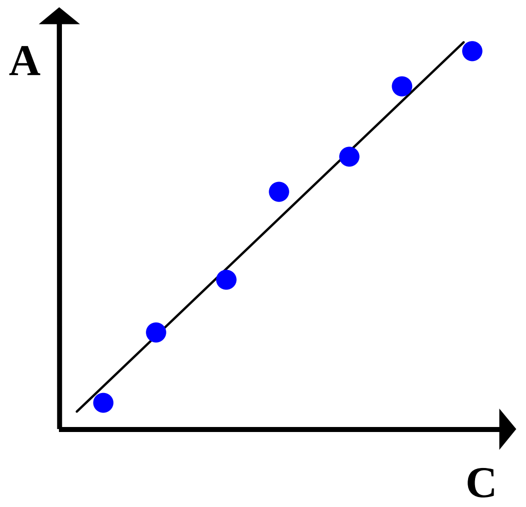
- How to do it?

- Example: Mix

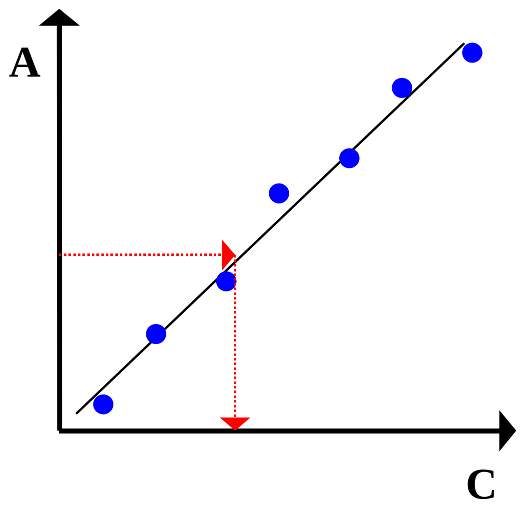


What is calibration?

- 1) Samples with known concentrations (c_i)
- Signal amplitudes (A_i) from measurement on samples
- Standard curve



- 2) New samples with **unknown** concentrations
- Measurements $\Rightarrow$ signal amplitudes, A_j
- $\Rightarrow$ predicted conc. values, c_j
  for new samples (from standard curve)

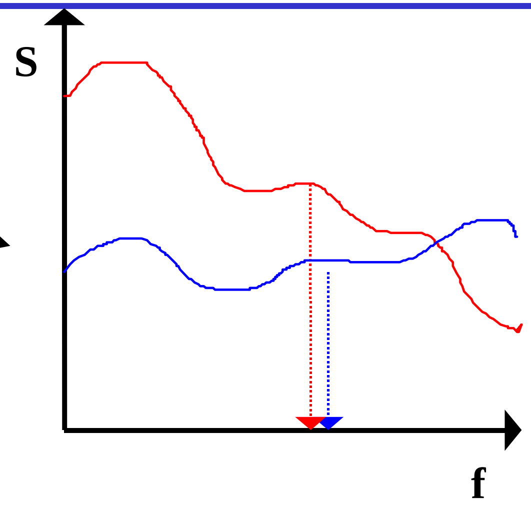


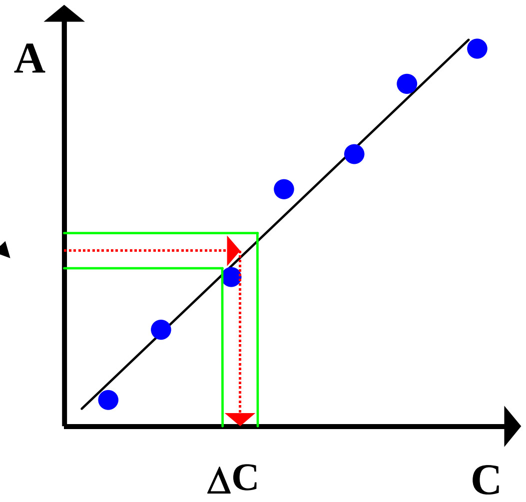
Problems with traditional calibration

- **Selectivity:** There is NO unique signal where ONLY the analyte absorbs.

- **Precision:** Noise in the signal amplitude is transferred to the predicted concentration for a new sample.

- **Diagnosis:** The standard curve is ONLY valid for samples similar to the ones in the calibration.





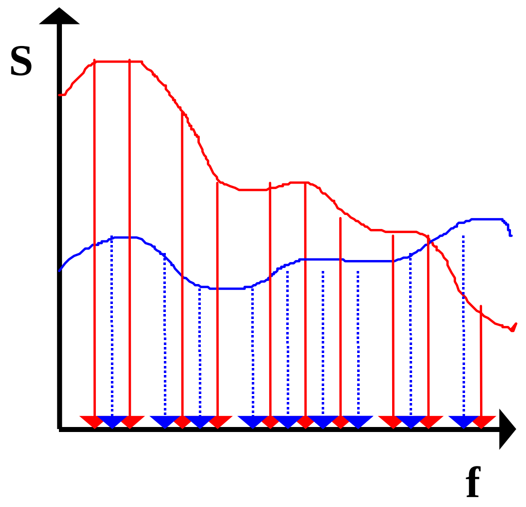
Multivariate calibration

- Many signals (spectrum digitised at K different wavelengths)

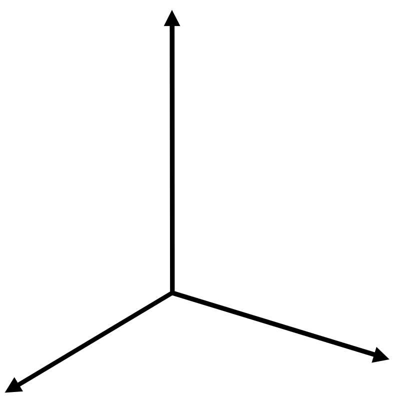
  $\Rightarrow$

  K variables

  K signals



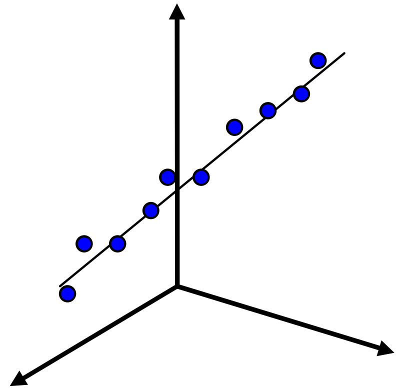
- Multivariate space
  - each variable defines a coordinate axis- Space with **K** coordinate axes.
  - Points, lines, distances, ..., have got the same properties in **K** as in **2** and **3** dimensions.



Multivariate calibration

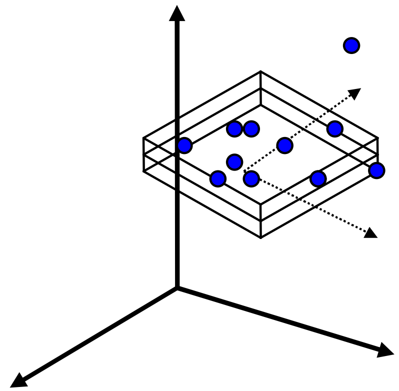
- One analyte (y-variable):

  - all points (digitised spectra) are describing a line ± noise in K-space.



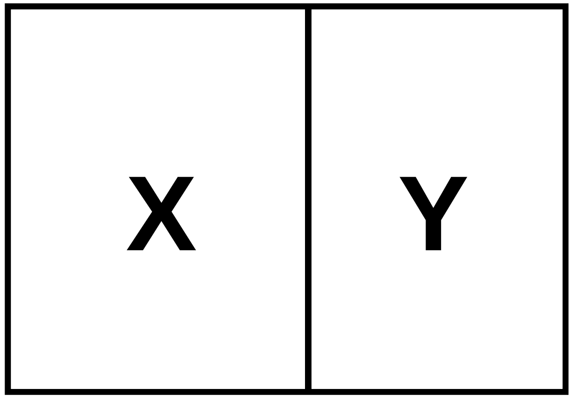
- One analyte + interacting compounds, or many analytes + interacting compounds :

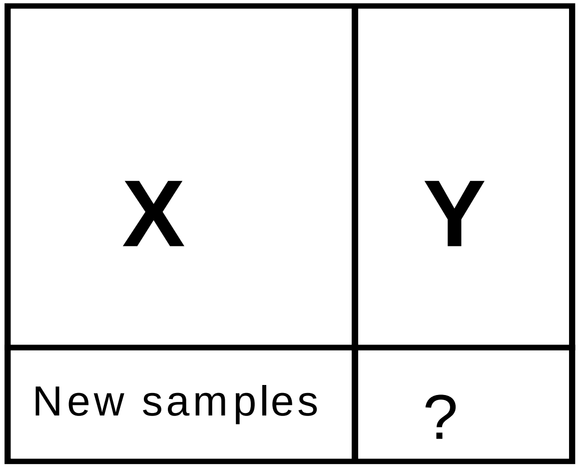
  - all points are describing a hyper-plane ± noise in K-space.



How to do it?

- Select samples representing the interesting variation. (Use design - FF, FrF, D-opt, Mixture)
- Measure Y-data for each sample using the "traditional" method i.e. the method we wish to replace.
- Use the “new method” (usually spectroscopy) to characterise the samples, these measurements are the X-data.
- Select calibration and test samples (PCA of X)
- Calculate a calibration model using PLS. Evaluate and interpret the model.
- **Test the model using external samples!**
- Use model for classification and prediction of new samples.





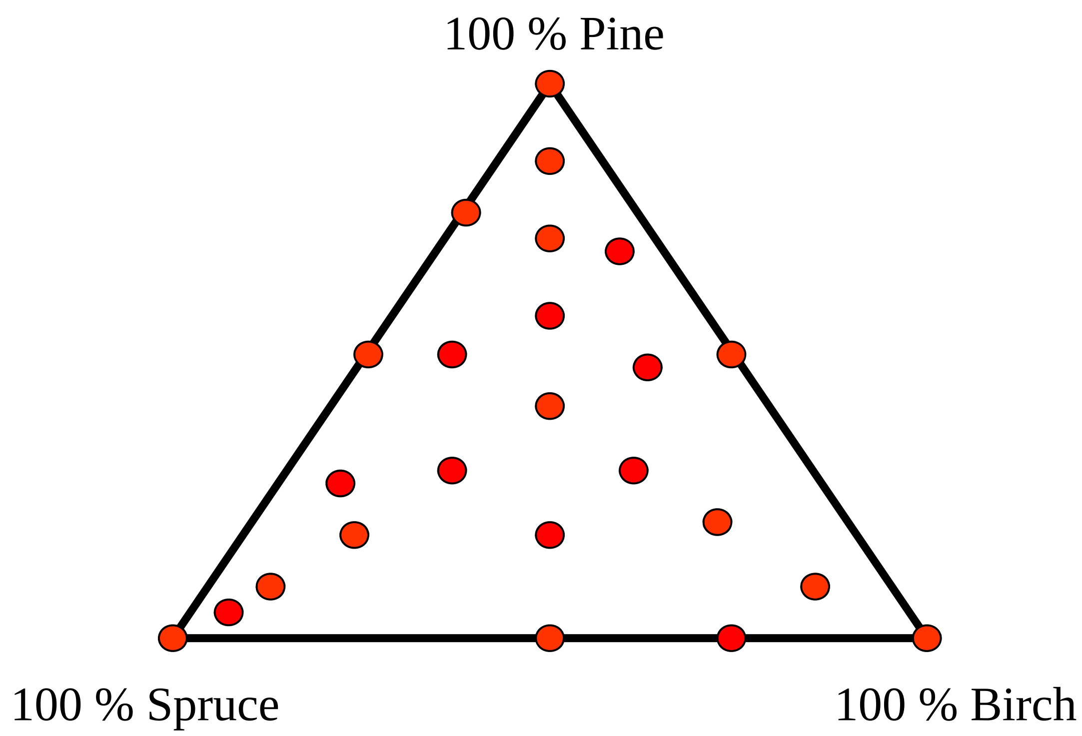
Application areas

• **Wheat, corn, ...**	Protein, water, fat	NIR
• **Peat, ….**	Water, energy, sugar, C, N, S	NIR
• **Lake water**	Humus acids, lignin sulfonates	UV
• **Beer, wine**	alcohol, protein, sugar, etc.	NIR, IR
• **Whisky, wine**	Taste, smell, “quality”	GC, HPLC
• ***Pulp, paper***	***Raw material***, lignin, products.	***NIR***, UV, IR, NMR
• **Pigs (living)**	Fat, meat, etc.	NIR
• ***Humans (living)***	Hair, blood, urine, skin, operations	NIR, FT-IR, NMR
• ***Plant material***	Screening for natural products	NIR
• ***Pharmaceuticals***	Compounds & metabolites	UV-Vis, FT-IR
• ***Process quality***	Sensors	NIR, IR, UV, GC
• **+ many more**		

Example - Mix (Prediction of wood mixes)

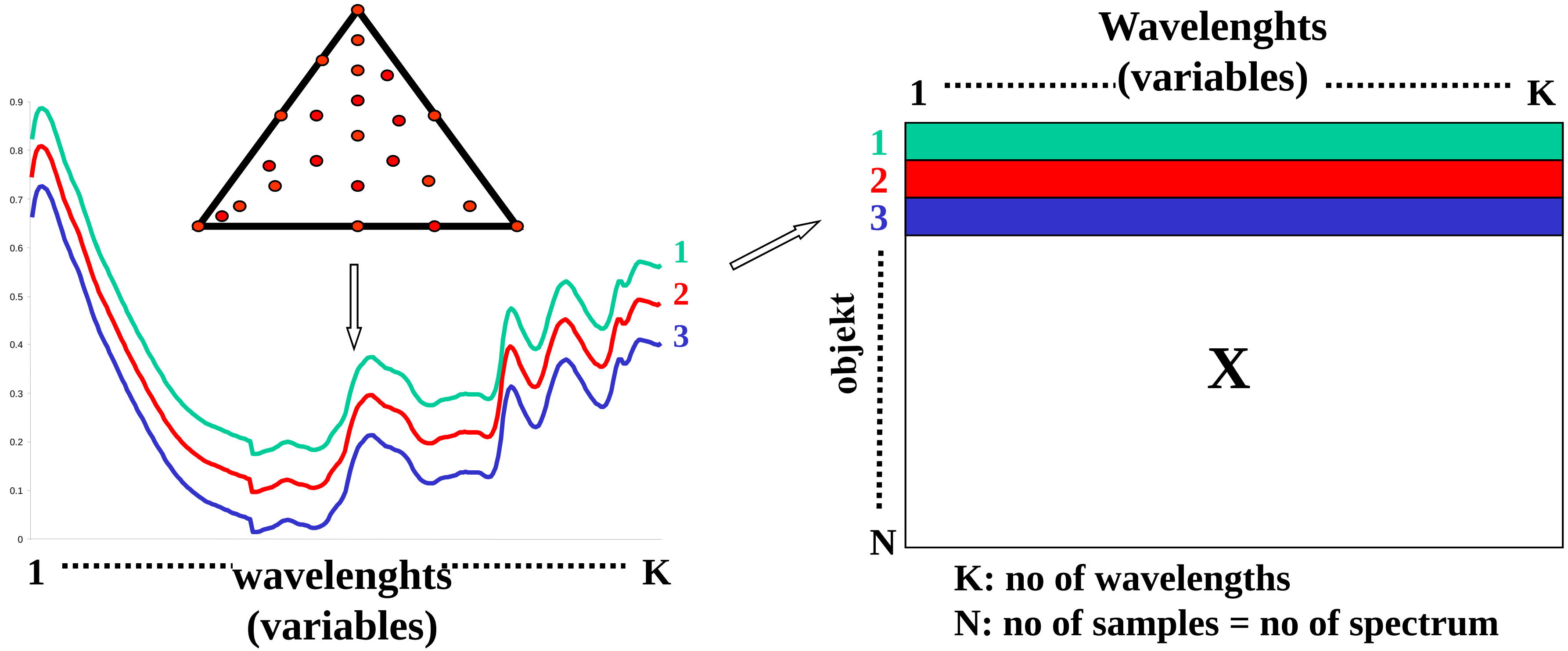
Pulp wood from three wood species (Pine, Spruce, Birch) was ground to a powder and mixed according to a mixture design. (30 samples in total)

- The sum of the three constituents in each mixture = 1 (100%) (Closure)



From spectra to data table

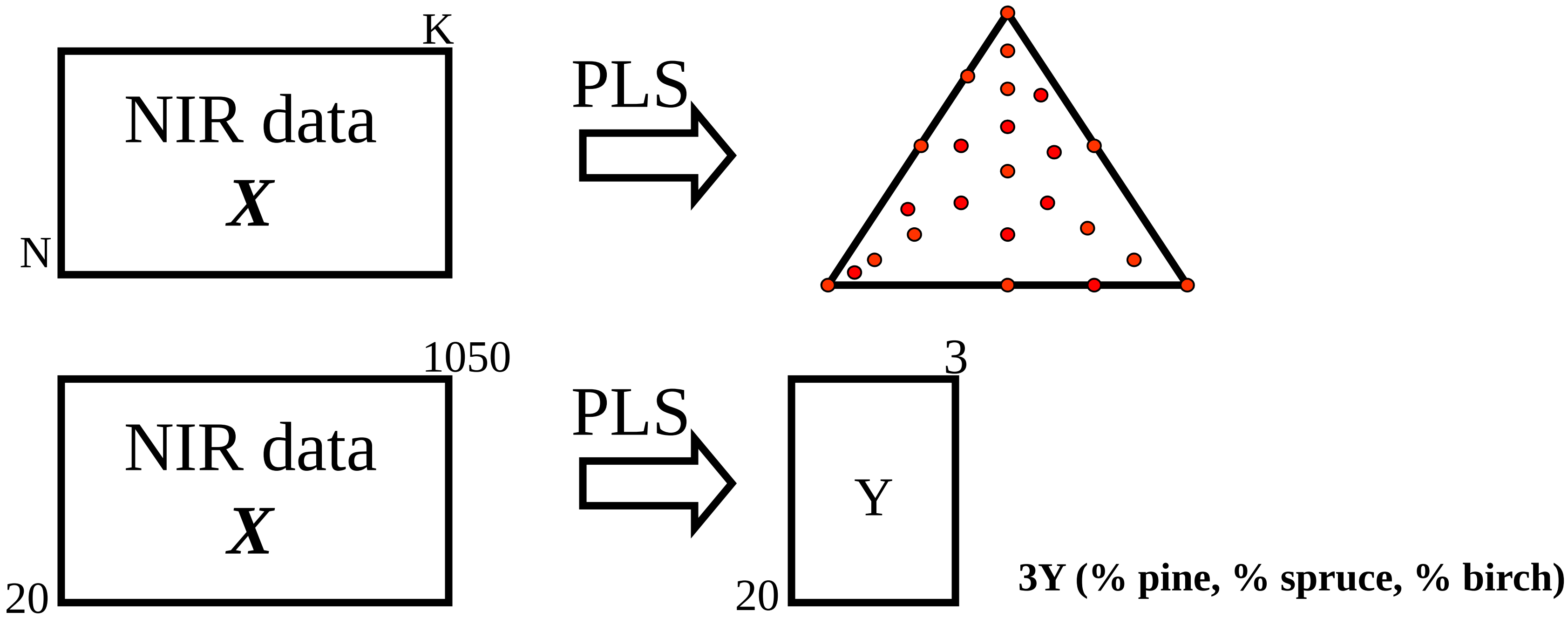
For each sample (mixture) a NIR spectrum was acquired. This generated 1050 wavelengths (variables) in the NIR region characterizing each sample.
Spectra were digitized giving the X data matrix below.



K: no of wavelengths
N: no of samples = no of spectrum

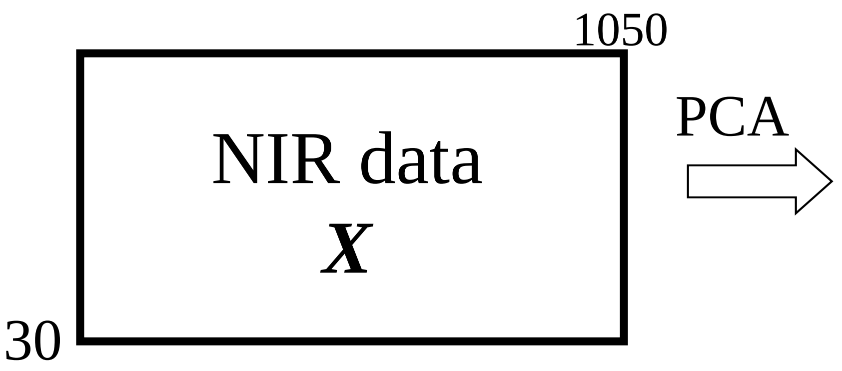
Exempel - Mix (Prediktion av vedblandningar)

The aim with the study was to use the NIR spectra of known mixtures of wood samples
To calculate a multivariate calibration model for prediction of sample mixtures (Y).
20 samples (mixtures) were used to calculate a calibration model (training set).
10 samples were selected for testing the models predictive ability (test set).



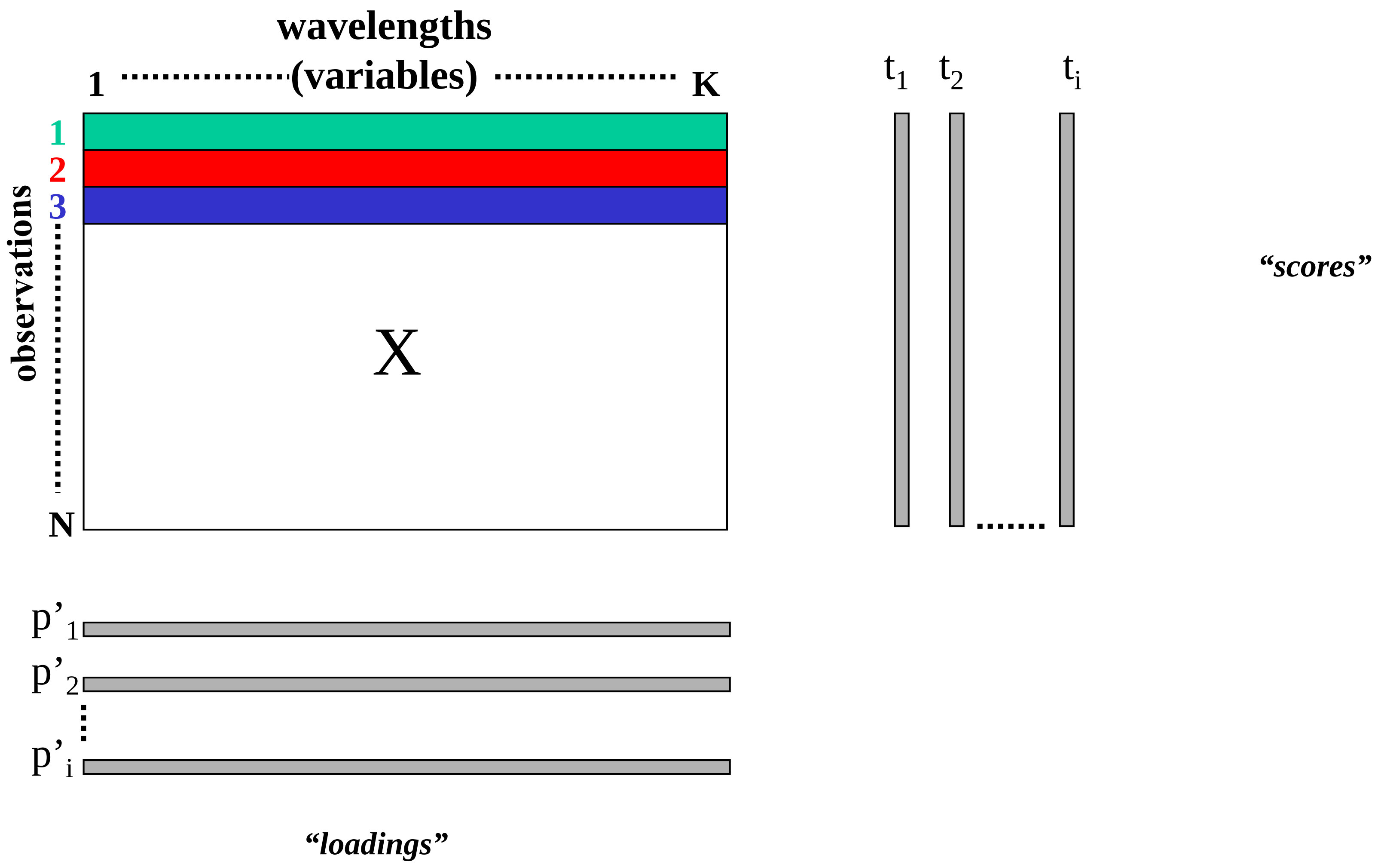
Selection of calibration & test set

Based on “scores” from the PCA of X (spectra) calibration (training) and test samples are selected. The calibration shall span the experimental space and give a good description of the entire space. The test samples shall be equally distributed over the entire space but not outside the limits set by the calibration (training) samples.

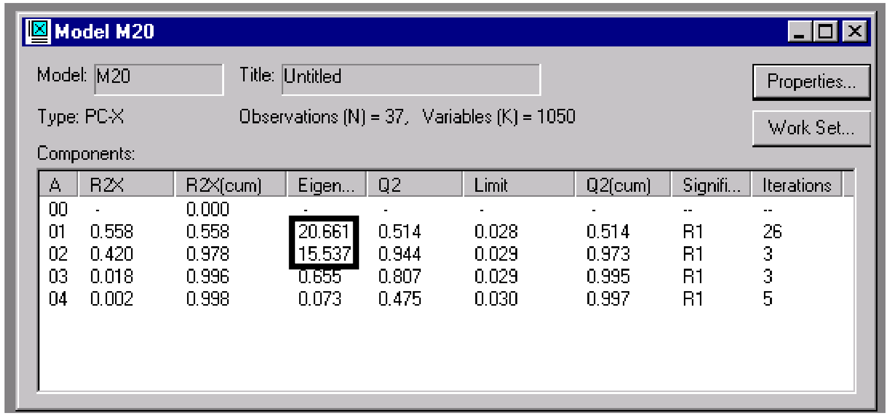


Selection of calibration and test samples.

PCA (Principal Component Analysis)



PCA of X (spectra)



Model M20

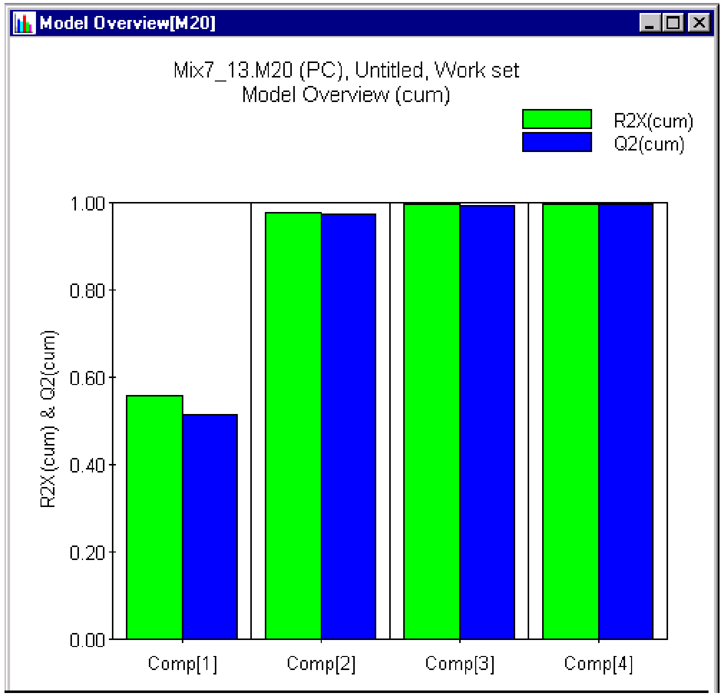
Model: M20 Title: Untitled

Properties... Work Set...

Type: PC-X Observations (N) = 37, Variables (K) = 1050

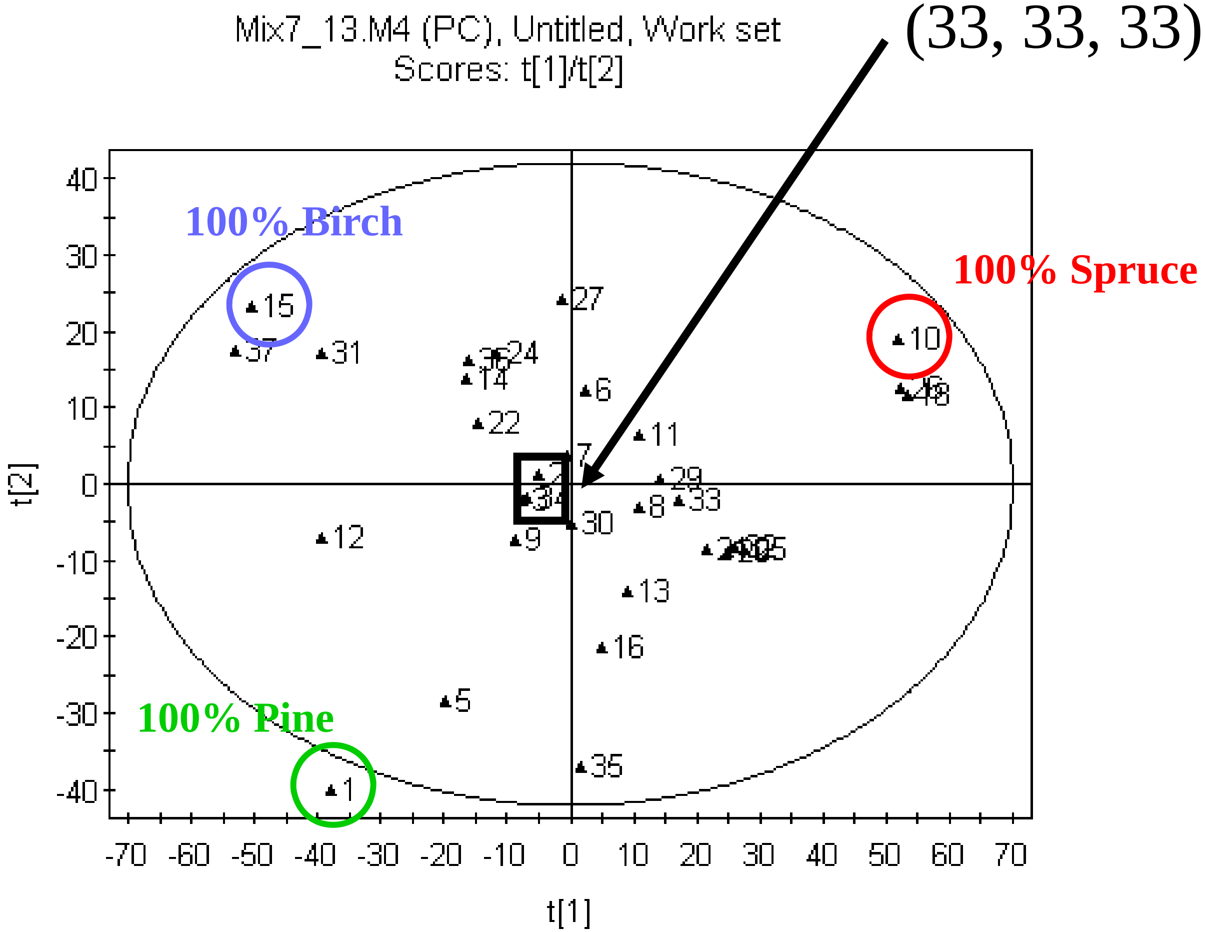
Components:

A	R2X	R2X(cum)	Eigen...	Q2	Limit	Q2(cum)	Signifi...	Iterations
00	-	0.000	-	-	-	-	--	--
01	0.558	0.558	20.661	0.514	0.028	0.514	R1	26
02	0.420	0.978	15.537	0.944	0.029	0.973	R1	3
03	0.018	0.996	0.655	0.807	0.029	0.995	R1	3
04	0.002	0.998	0.073	0.475	0.030	0.997	R1	5



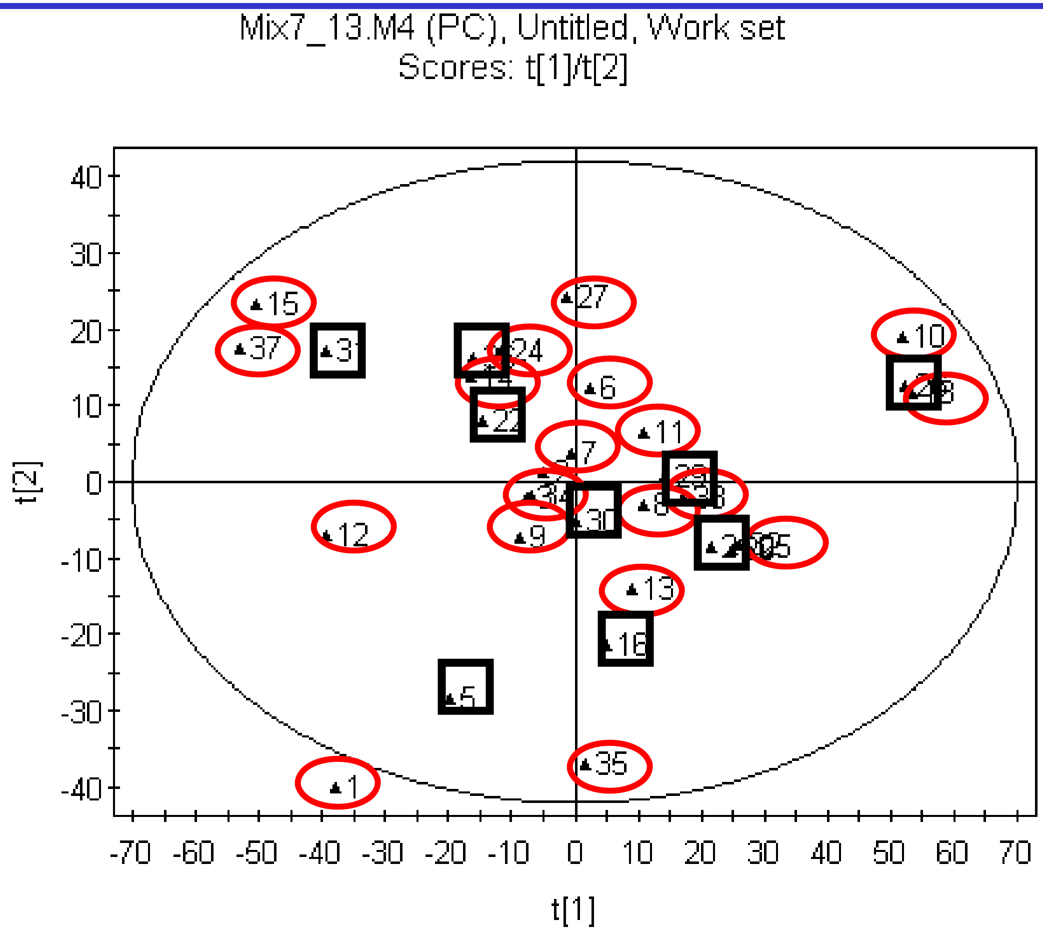
- 4 PCs significant according to cross validation
- 2 PCs significant according to eigenvalue (>2)
- After two PCs 97.8 % of the variation in X is described and 97.3 % of the variation in X can be predicted according to cross validation.
- Hence, we are describing the main part of the variation with two PCs and set for two PCs for interpretation of the systematic variation in X.

Interpretation of "scores" (t1/t2)



Scores contain discriminating information regarding wood mixtures.
I.e. spectra contain discriminating information regarding wood mixtures.

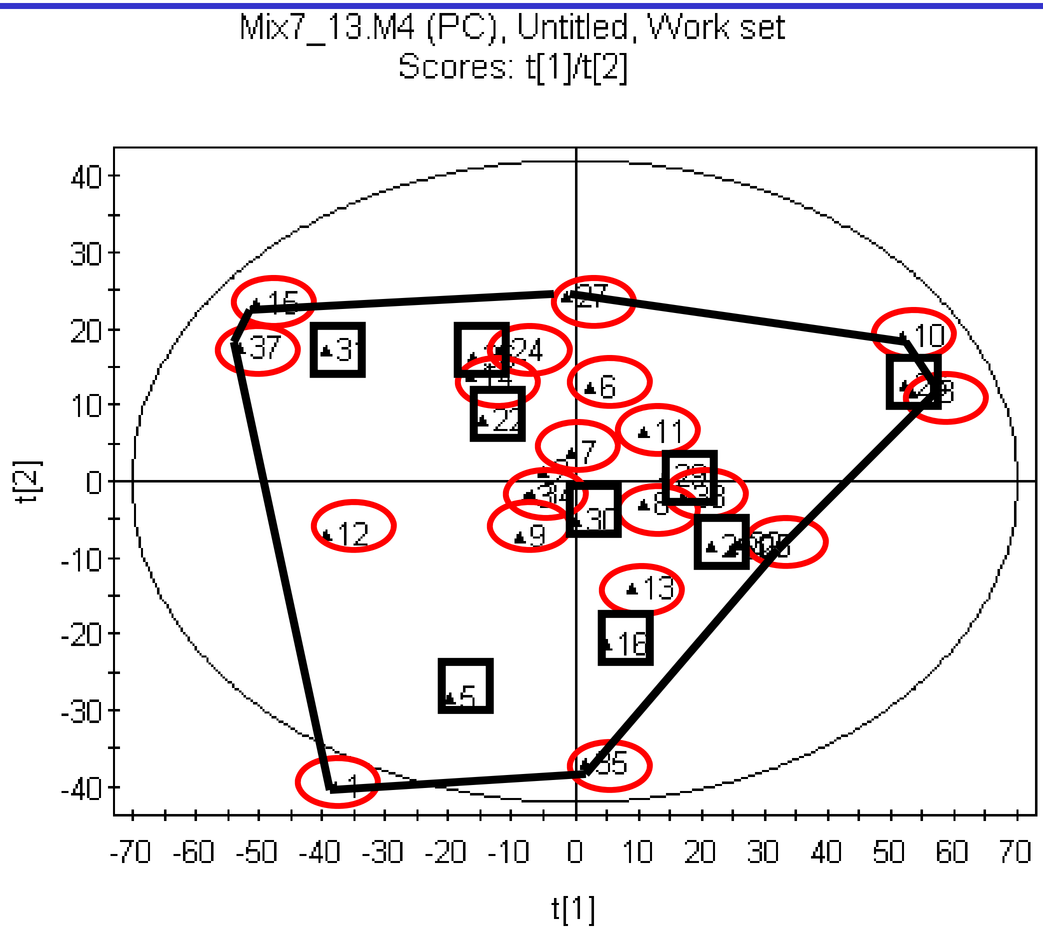
Selection of calibration & test set from "scores"



Calibration samples (circled) span the experimental space and are evenly distributed Over the whole surface.

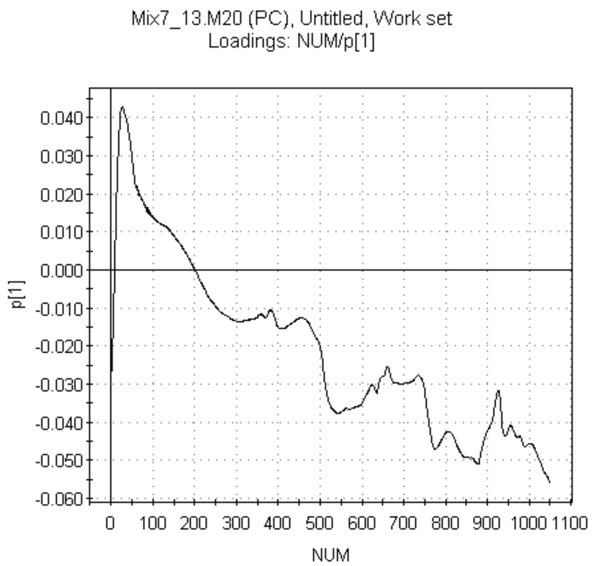
Test samples (in squares) are evenly distributed over the surface but not outside the model limits set by the calibration samples.

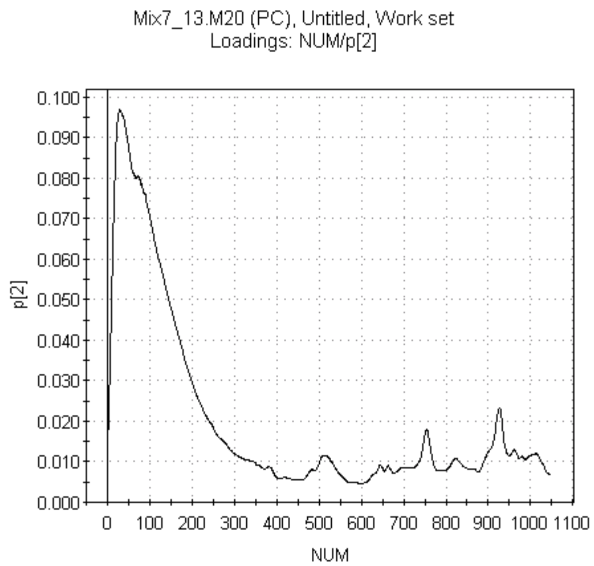
Model limits for calibration



The black lines in the score plot define the limits for the calibration model. Within these limits the model will be valid.

Loadings (p1 & p2) for PCA

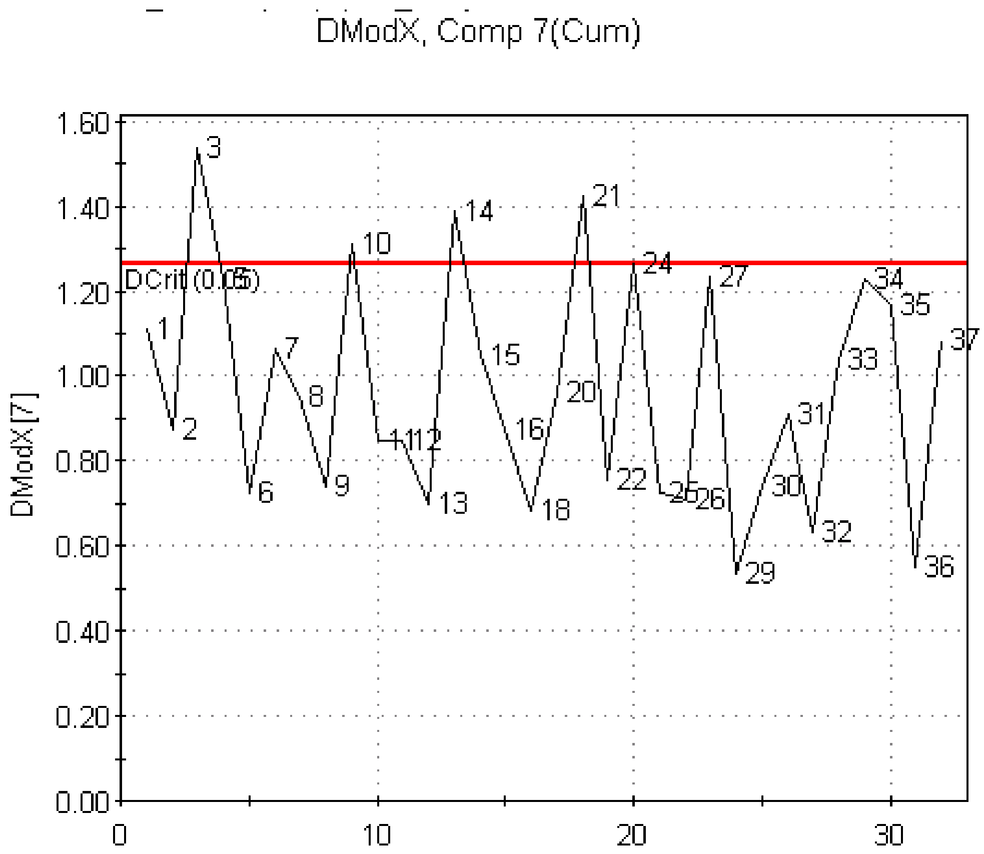




Loading (p1/wavelength) shows that the separation in the first PC depends on almost all wavelengths in the spectra.

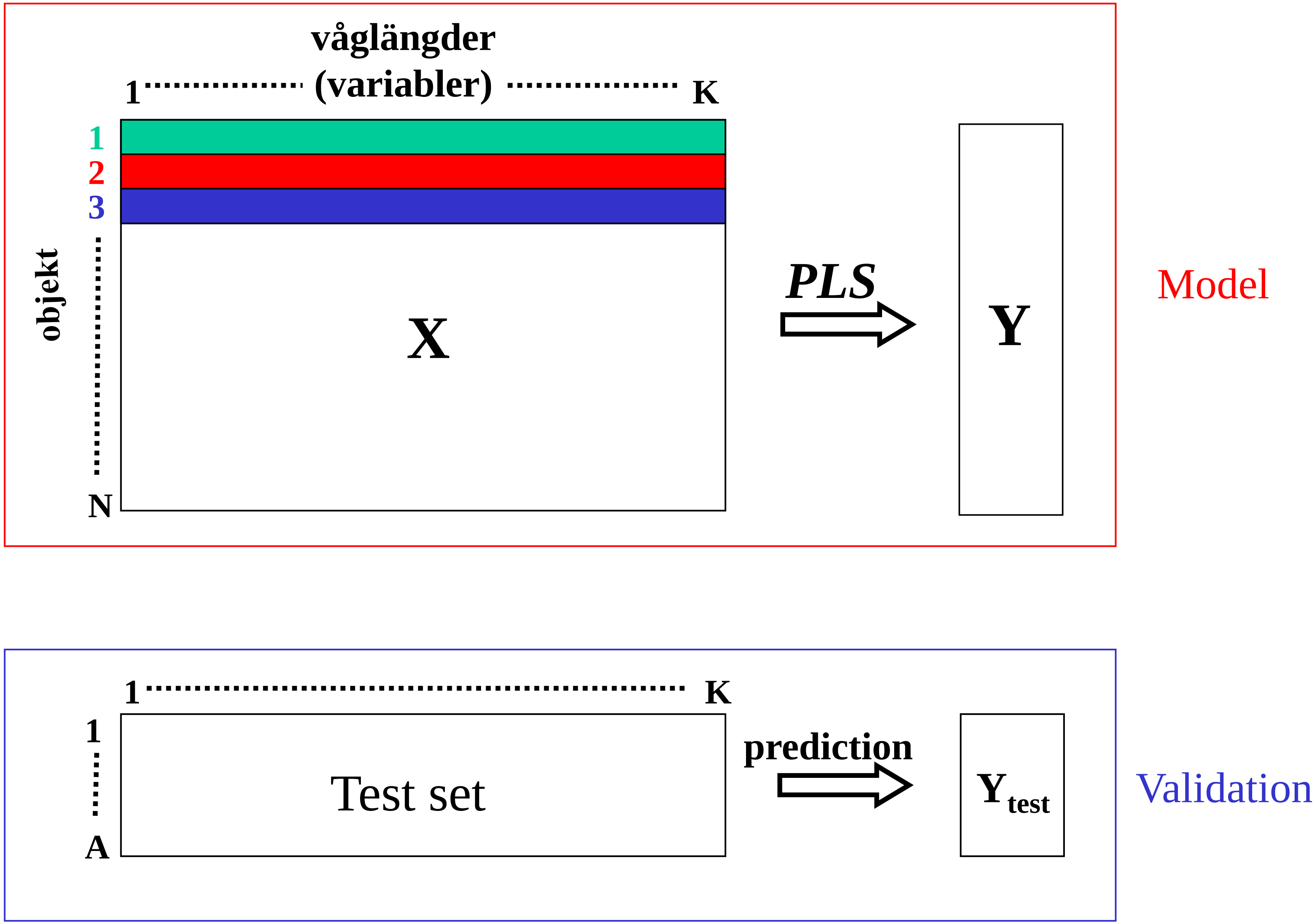
Loading (p2/wavelength) shows that the separation in the second PC depends on early wavelengths in the spectra.

DMod X for the 30 samples in X



DModX for the 30 samples don’t suggest any extreme outliers.

Calibration and Validation

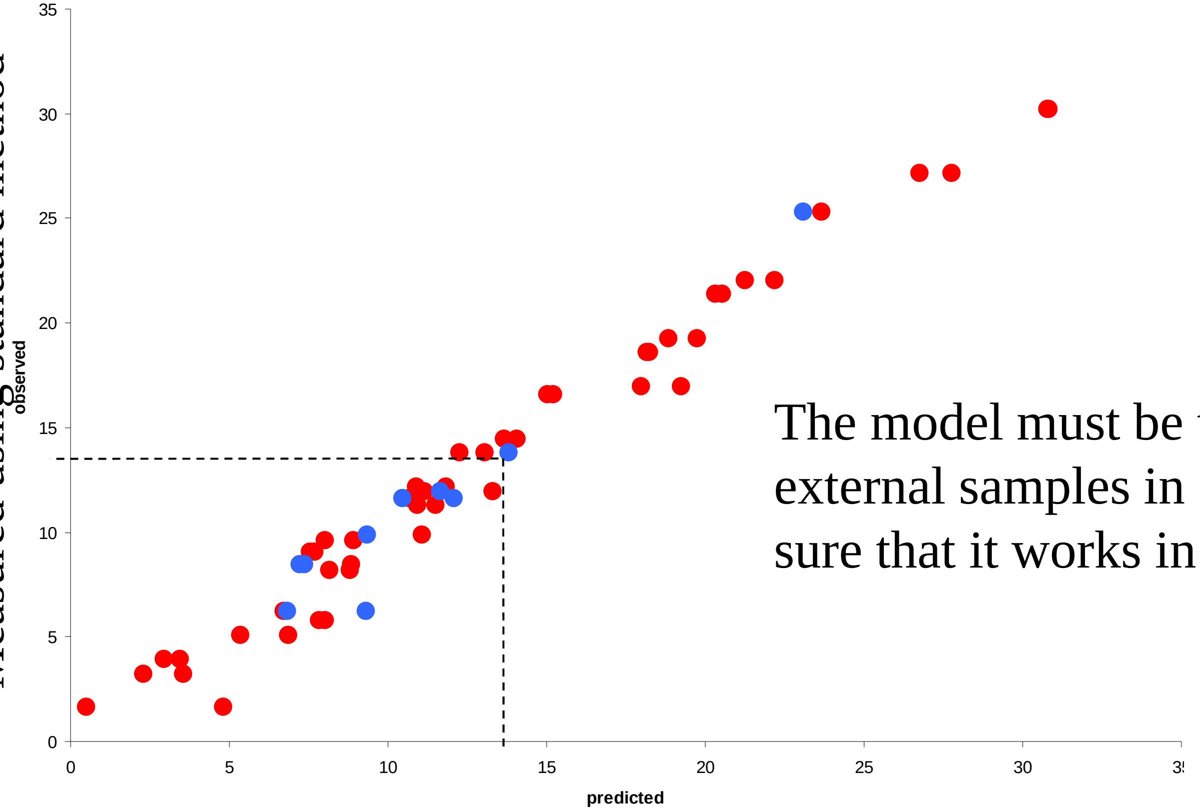


Estimate/Prediction

Estimate: Fit of model to calibration samples

Prediction: Prediction of test samples (not included in model)

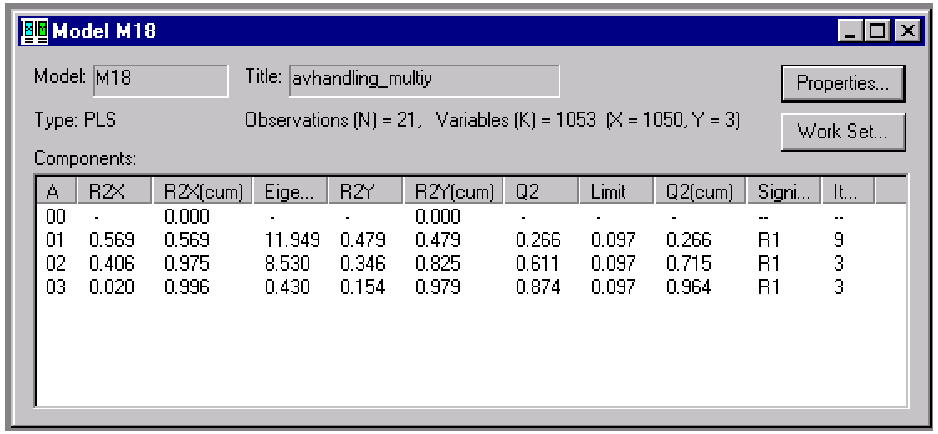
Measured using standard method



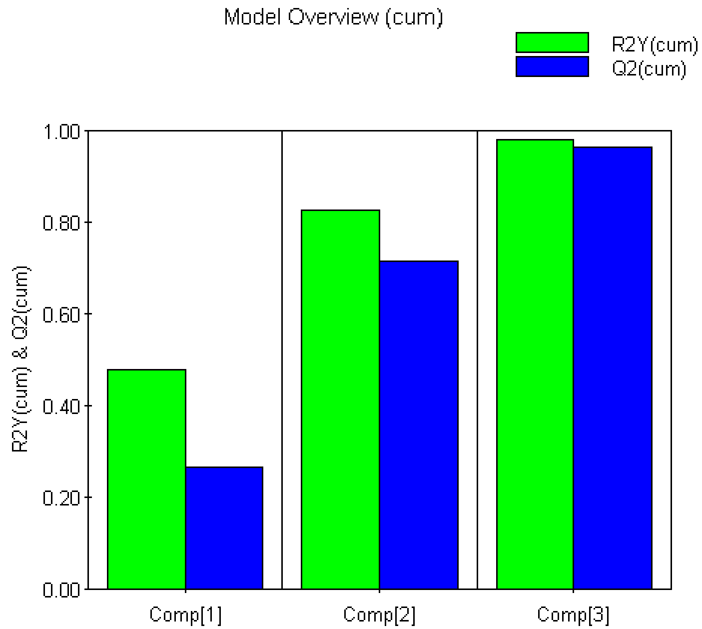
The model must be tested with external samples in order to make sure that it works in a real situation!

Predicted by PLS model

Calculation of calibration model (PLS)



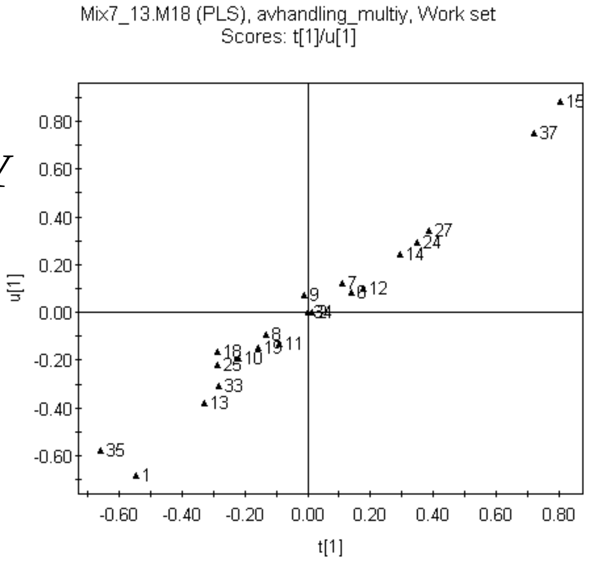
A	R2X	R2X(cum)	Eige...	R2Y	R2Y(cum)	Q2	Limit	Q2(cum)	Signi...	It...
00	-	0.000	-	-	0.000	-	-	-	--	--
01	0.569	0.569	11.949	0.479	0.479	0.266	0.097	0.266	R1	9
02	0.406	0.975	8.530	0.346	0.825	0.611	0.097	0.715	R1	3
03	0.020	0.996	0.430	0.154	0.979	0.874	0.097	0.964	R1	3

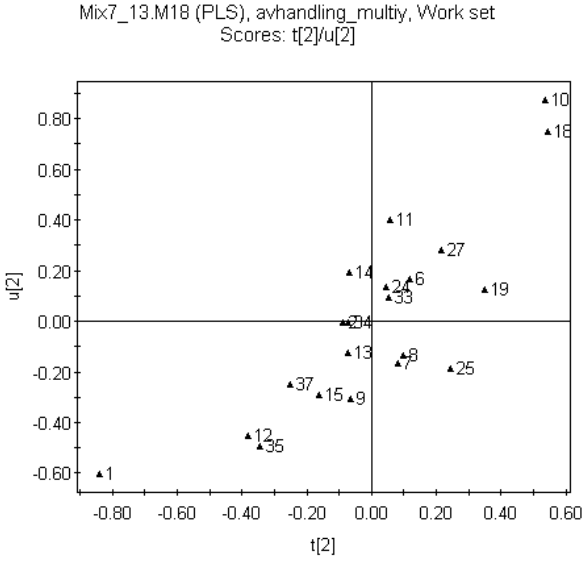


- Cross validation gives 3 significant components
- R2X= 0.996, R2Y= 0.979, Q2= 0.874
- Q2 increases significantly for every new component added

Interpretation of “scores” for the PLS model

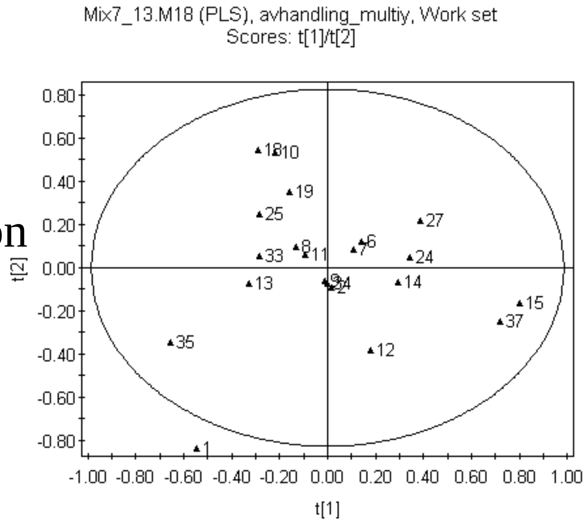
t1/u1
correlation X/Y
in 1st comp.

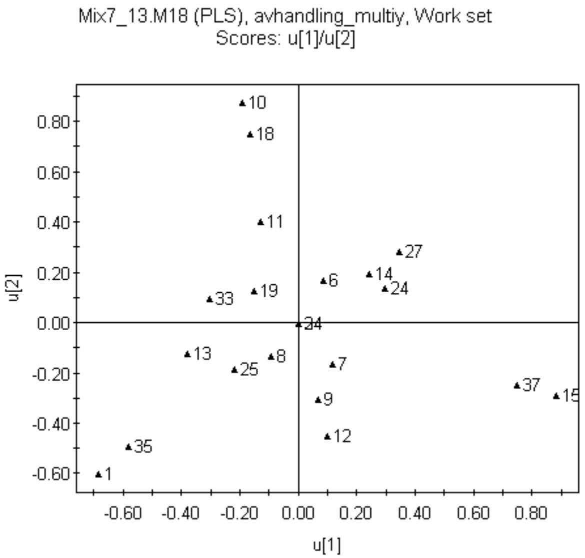




t1/u1
correlation X/Y
in 2nd comp.

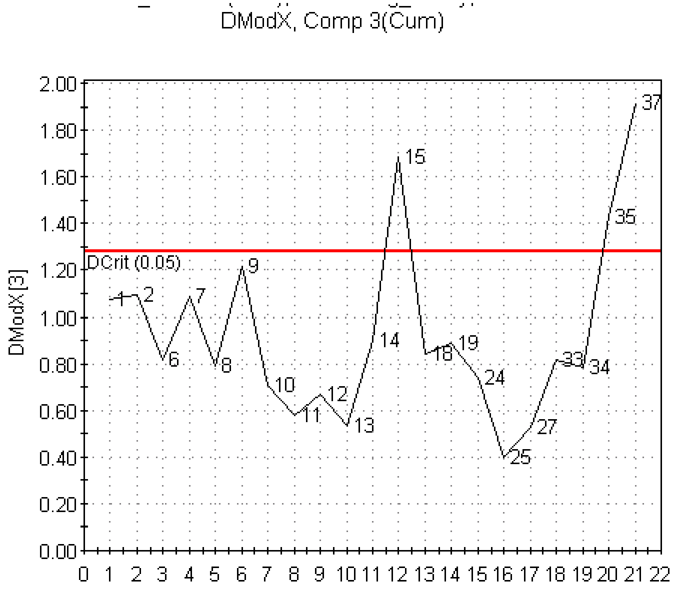
t1/t2
overview of
sample variation
in X.

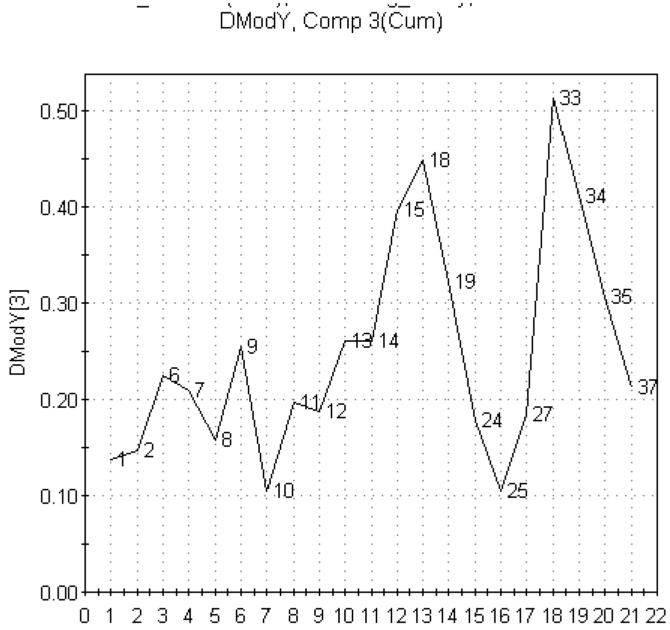




u1/u2
overview of sample
variation in Y.

DModX, DModY for calibration samples

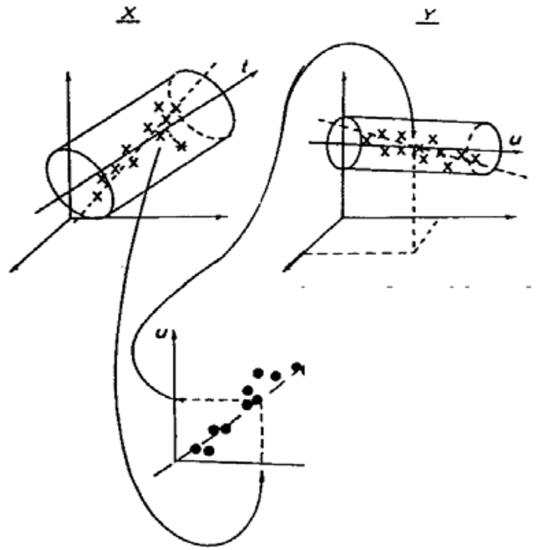




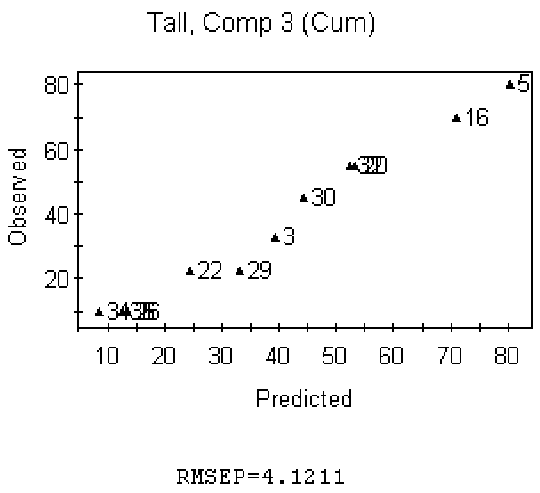
No extreme outliers in X nor Y space!

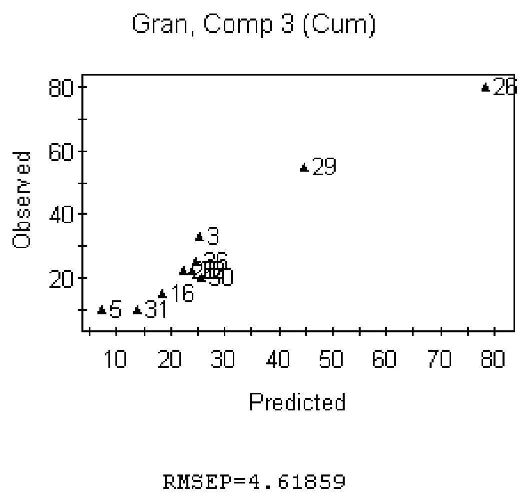
Prediction of test samples

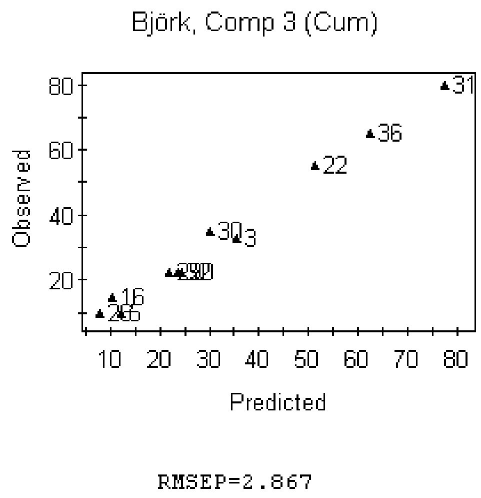
- Validation of the model by prediction of the 10 test samples
- RMSEP is the average prediction error in the same unit as Y.



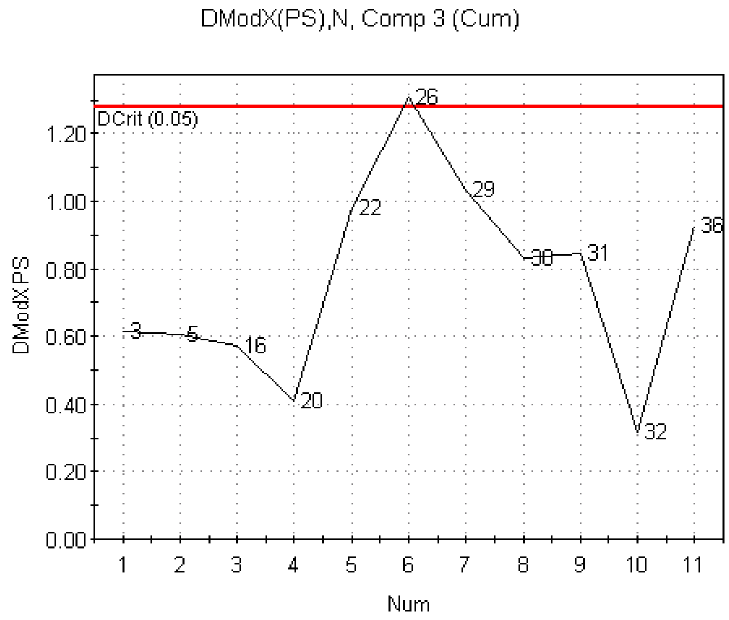
RMSEP = sqrt(PRESS/N)

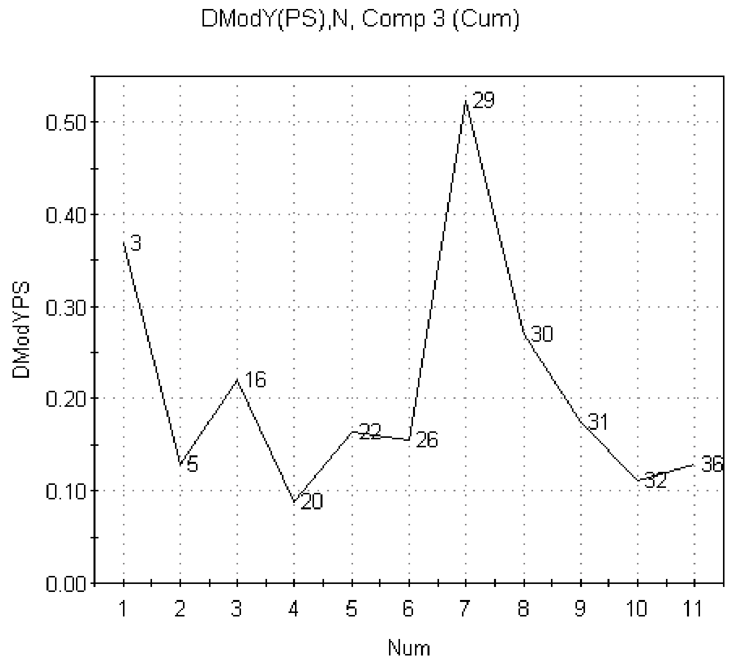






DModX, DModY for test samples





No extreme outliers in X nor Y space!

Summary of Calibration Model

- High R2, Q2
- Good correlation between X and Y (t/u)
- No outliers (“scores”, DModX,Y)
- Good predictions of external samples (test set)

Conclusion

We have a model that can be used for prediction of unknown samples (within the model limits).

Prediction of unknown samples

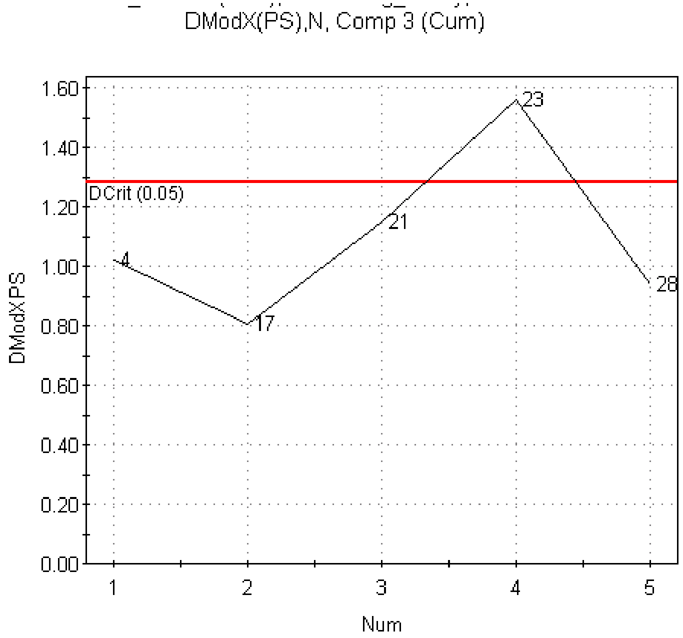
Spectra for five unknown samples were used to predict the mixtures of the three wood species (Pine, Spruce, Birch)

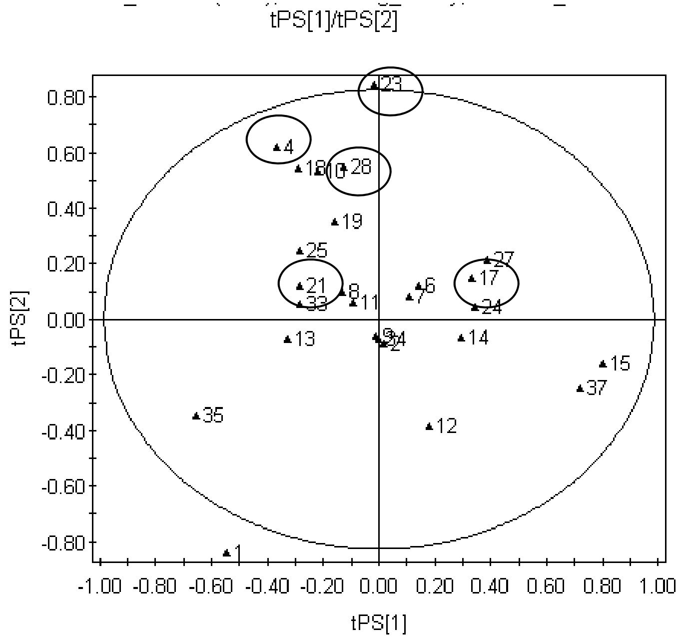
Predictions

Obs	Tall(pred)	Gran(pred)	Björk(pred)	T+G+B
40K	13.6697	92.5024	-6.22493	99.94717
170K	14.8937	25.5495	59.4891	99.9323
210K	50.626	30.6325	18.6762	99.9347
230K	-1.3846	74.4088	26.9754	99.9996
280K	5.71936	81.644	12.5795	99.94286

The sum of the predicted values for the three wood species is close to 100 %. This comes from the properties of the experimental design (closure).

DModX and Scores for unknown samples

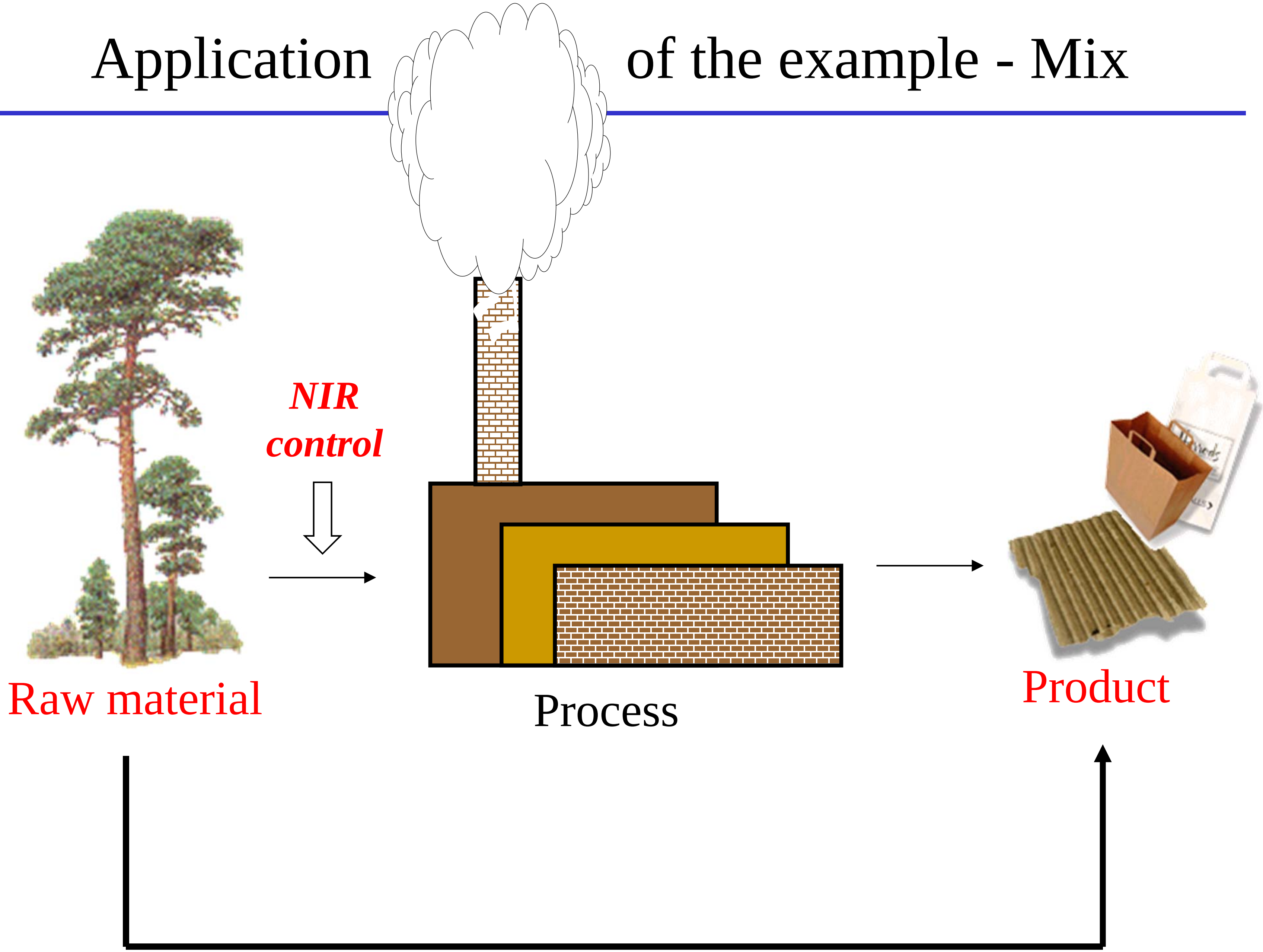




DModX + “scores imply that sample 23 doesn’t really fit the model.

Care should be taken in terms of the reliability of the prediction of sample 23.

Application of the example - Mix



Conclusion - Multivariate calibration

- Multivariate calibration gives robust models that can separate systematic variation from noise.

- Multivariate calibration uses many variables for calibration.

- Multivariate calibration is based on projection methods (PCA, PLS)

- Replace “traditional method” with a new faster, simpler, cheaper, …. method (spectroscopy).

- Selection of calibration and test samples (PCA)

- Correlation X/Y (PLS)

- An absolute must to validate model with external samples

- Prediction of unknowns once the model has been validated and is reliable.