TSNAs in tobacco and smoke : PLS regression for prediction of smoke contents according to chemical and physical characteristics

B. Vidal, M. Bouzige, R. Laroche

ALTADIS
R&D - France

Material

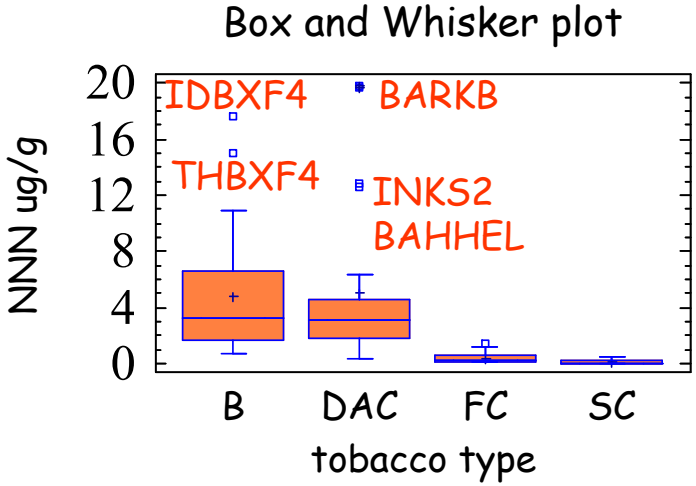
- 87 tobaccos from different types
  - Flue-cured : 23 tobacco grades, 1 blend, 6 countries
  - Burley : 21 tobacco grades, 3 blends, 9 countries
  - Sun-cured : 15 tobacco grades, 4 countries
  - Dark air cured : 16 tobacco grades, 8 blends, 6 countries
- Samples are representative of tobacco market (origins, stalk positions, quality, chemical and physical characteristics)

TSNA analysis

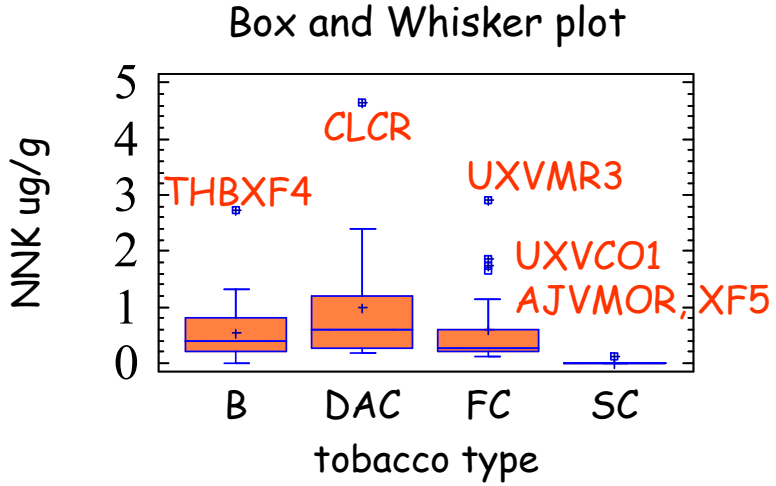
- Analysis of tobacco powders
- Production of cigarette for each tobacco sample
  - All the cigarettes are made with the same NTM
  - The cigarettes have the same draw resistance
- Analysis of the TSNA in the smoke

TSNAs in tobacco

BARKBC 54.1



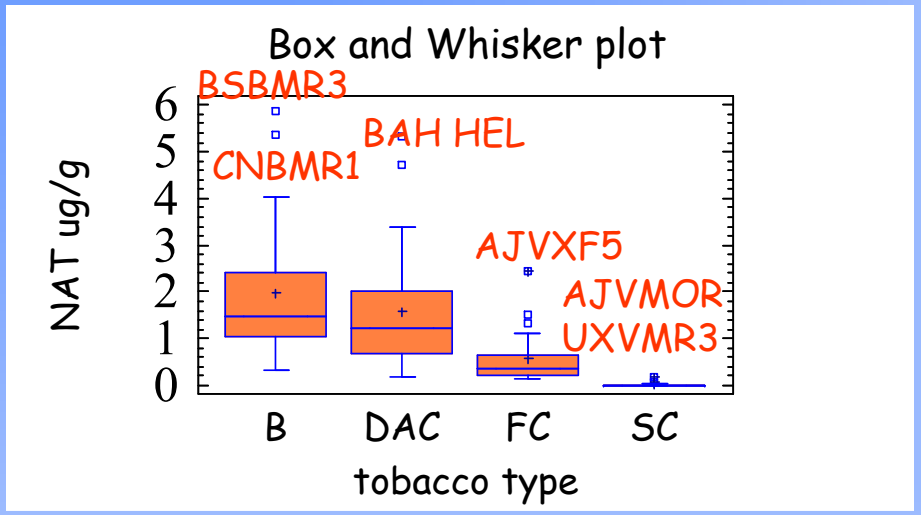
BARKBC 15.5



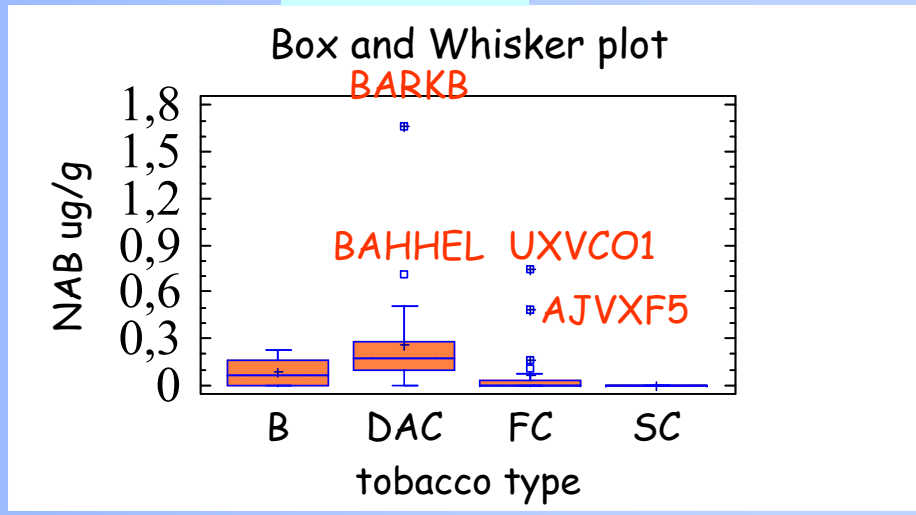
TSNA in tobacco µg/g	B	DAC	FC	SC
NNN				
Min	0.66	0.36	0.12	0
Max	17.6	19.77 (54.1)	1.47	0.48
Mean	4.83	5.01 (7.06)	0.41	0.10
NNK				
Min	0	0.17	0.12	0
Max	2.74	4.66 (15.51)	2.91	0.11
Mean	0.55	1.00 (1.60)	0.60	0.01

TSNA in tobacco

BARKBC 24.3



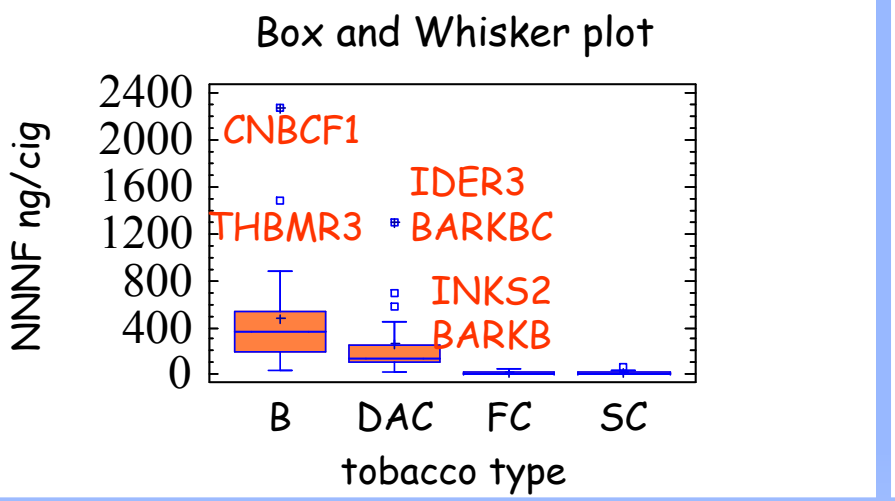
BARKBC 3.8

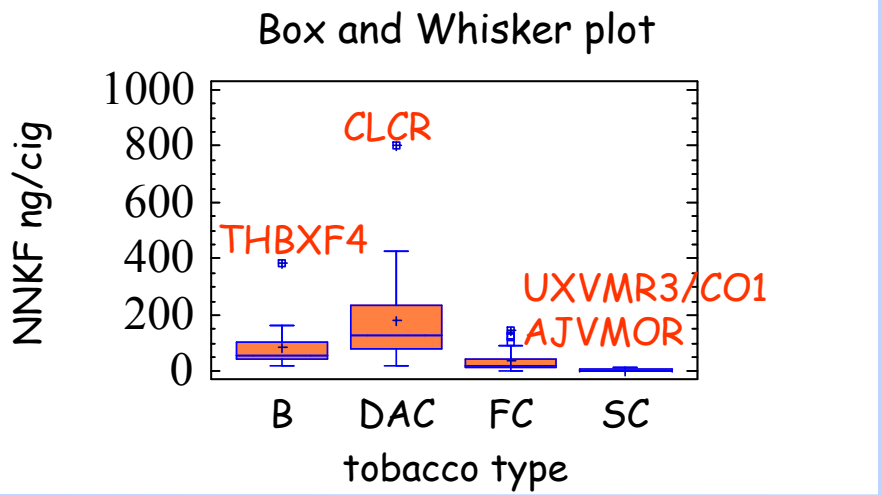


TSNA in tobacco µg/g	B	DAC	FC	SC
NAT				
Min	0.33	0.17	0.16	0
Max	5.87	5.33 (24.46)	2.45	0.18
Mean	1.98	1.57 (2.53)	0.57	0.02
NAB				
Min	0	0	0	0
Max	0.23	1.66 (3.84)	0.75	0
Mean	0.08	0.26 (0.41)	0.03	0

TSNA in smoke

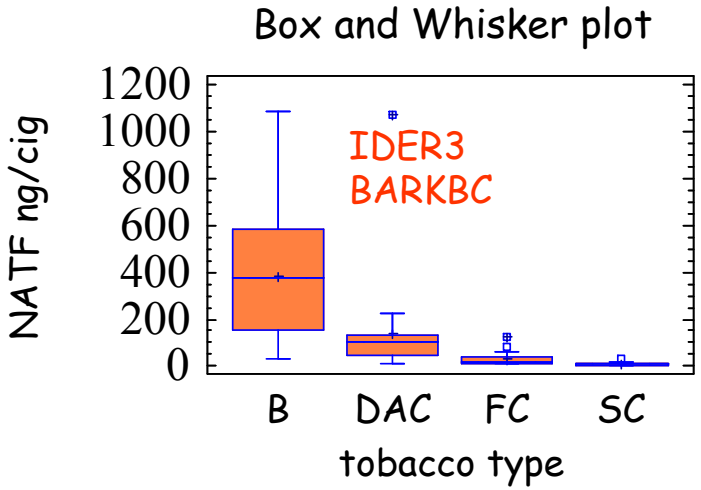
BARKBC 2201

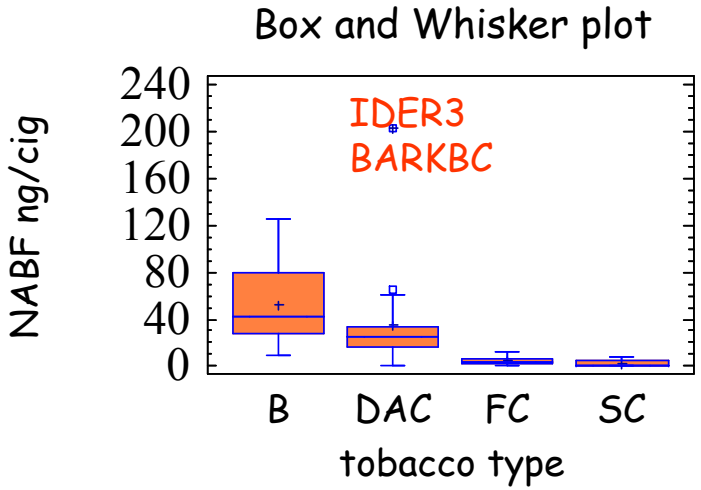




TSNA in smoke ng/cig	B	DAC	FC	SC
NNNF				
Min	30.5	18	3.2	0
Max	2269	1347	49	67.2
Mean	481.87	302.4	17.37	16.97
NNKF				
Min	22.5	18.7	4.8	0
Max	382	804 (2201)	145	15.2
Mean	84.35	183.11 (267.19)	40.60	3.91

TSNA in smoke





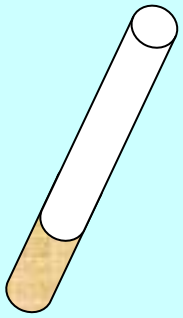
TSNA in smoke ng/cig	B	DAC	FC	SC
NATF				
Min	28.5	9.5	8.0	0
Max	1085	1067	126	30.4
Mean	383.14	162.79	30.81	8.33
NABF				
Min	9.3	0	0	0
Max	125	202	12.2	8
Mean	51.73	40.24	4.7	2.0

Prediction of TSNA in smoke

Tobacco

Chemical parameters

TSNAs
NNN, NNK, NAB, NAT



Physical parameters

Smoking results

Linear Regression

TSNAs in smoke
NNNF, NNKF,
NABF, NATF

X

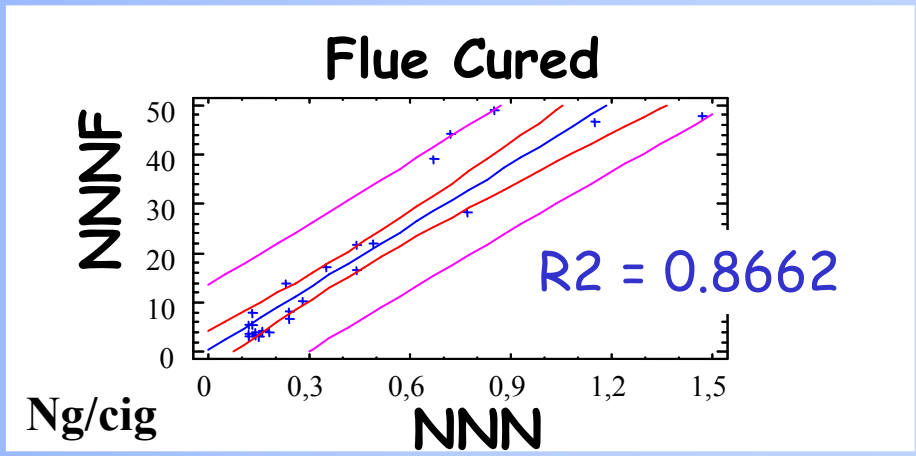
PLS Regression

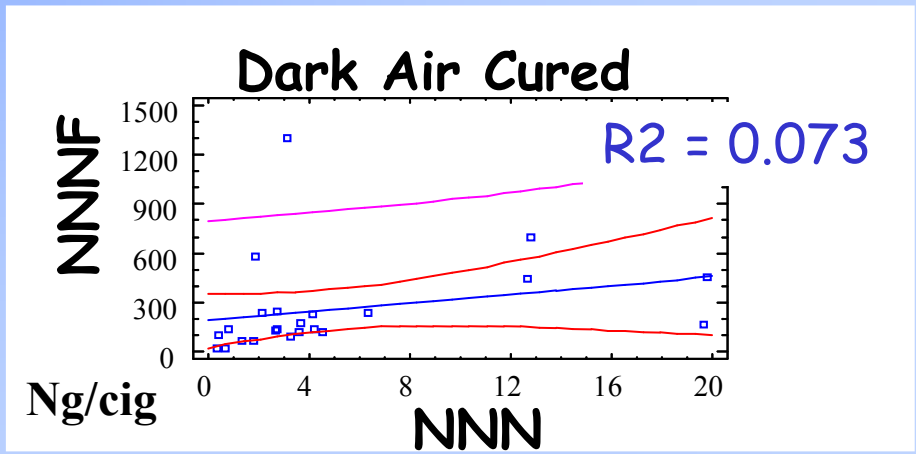
Y

Linear Regression

- Linear relationship between tobacco and smoke can be explained by transfer

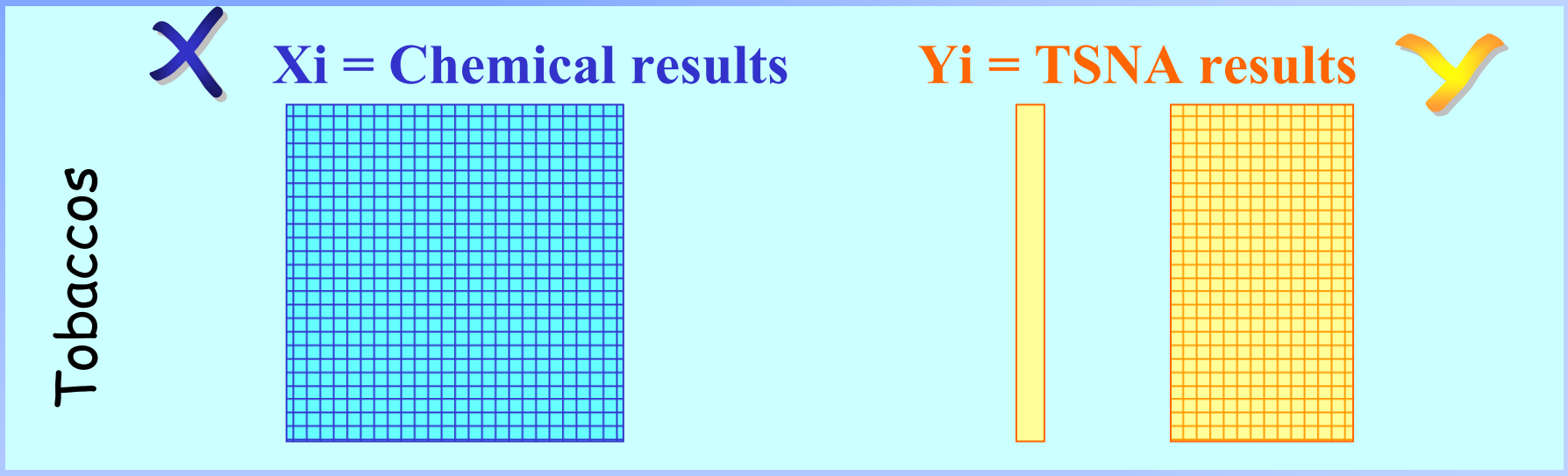
- BUT…
  - Linear regression is not always a good model for TSNA prediction in smoke
  - What to do if no value for TSNA in tobacco?





Partial Least Square Regression ???

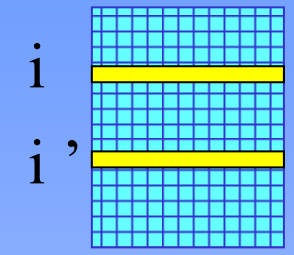
- Link a block of explanatory variables and one or many variables to be explained



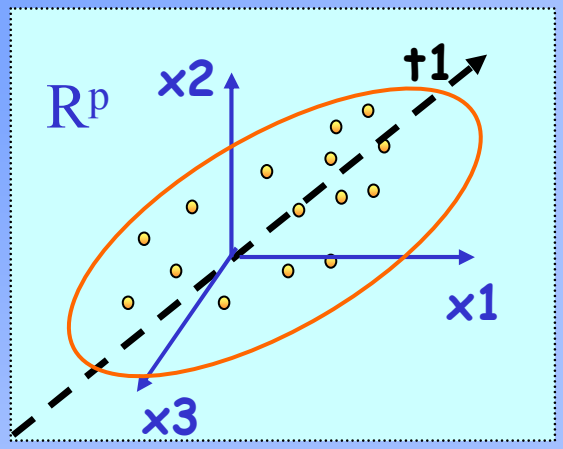
- Variables can be highly correlated and more numerous than the observations
- There can be missing values

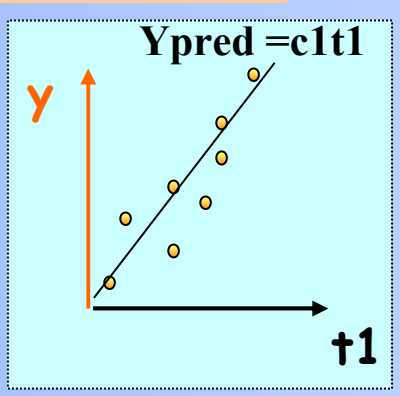
How does it work? Case of only one Y

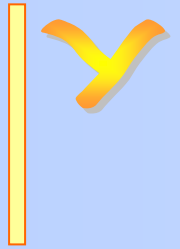
n observations



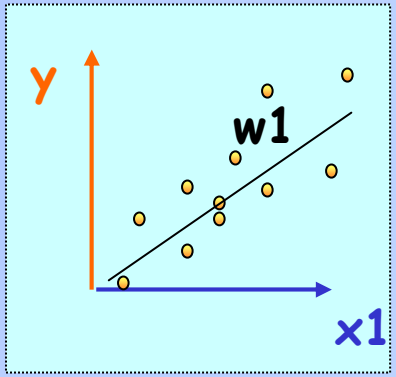
Find t1 with two constraints





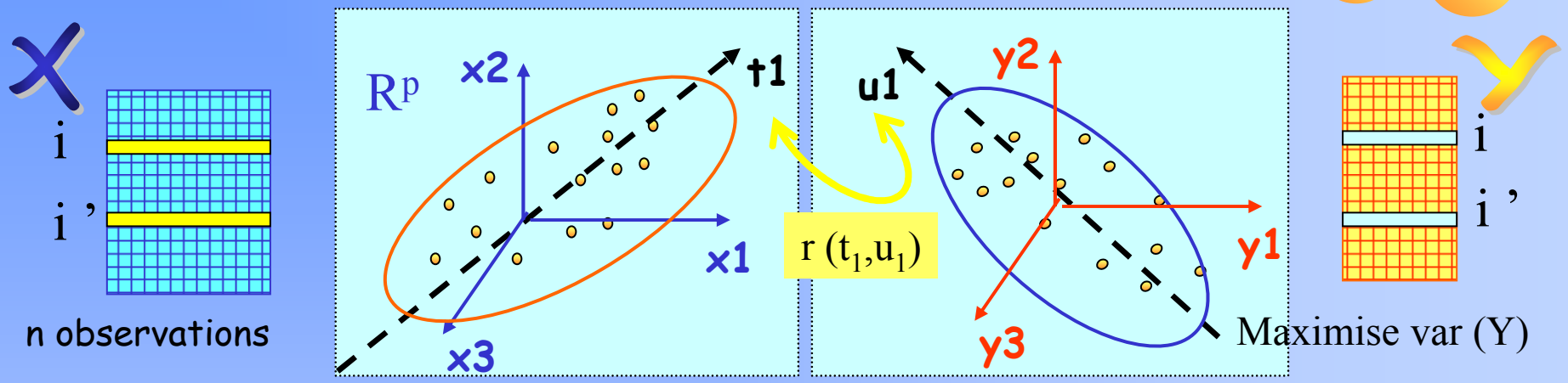


- Algorithm
  - t1 = w1x1 + w2x2 +…wpxp
  - ypredicted = c1t1 = c1w1x1 + …..c1wpxp
  - t2 = w21x11+…+w2px1p= v1x1 +…+vpxp
  - ypredicted = c1t1 + c2t2 = (c1w1 + v1c2)x1+….



- Coefficients are meaningful

More than one Y



$$Cov\ (t_1,u_1) = r(t_1,u_1) * var(X) * var(Y)$$

- Find the first PLS component t_1, a line in the space of X and a line u_1 in the space of Y which are calculated for :
  - Variance explained of X & Y by t & u is large
  - Relationship between t & u is maximize
  - Successives t & u must be orthogonal

Indicators

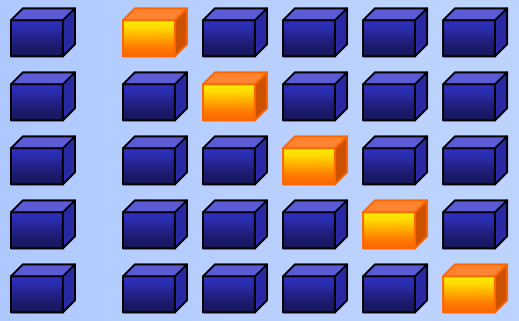
- Cross validation

Prediction error

$$Q^2_{cum} = 1 - \prod_{a=1}^{h} \frac{PRESSa}{RESSa_{-1}} \approx 1 - \frac{PRESS_h}{\sum_i (y_i - \bar{y})^2}$$

Estimation error

Standard deviation



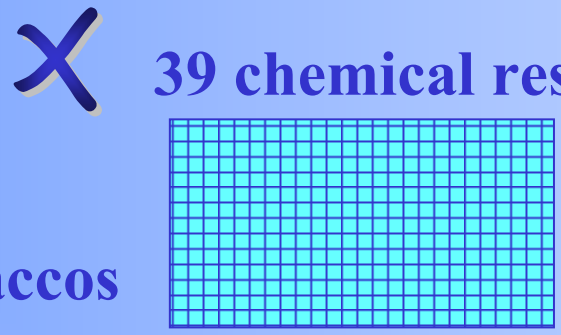
Robustness of the model

- R2Ycum=1-RESSh/var(y) Adjustment quality

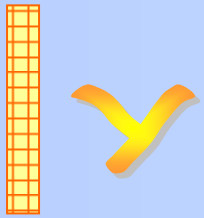
- VIP : Variable importance in the projection

Applications : Burleys

- Initial model :



NNKF

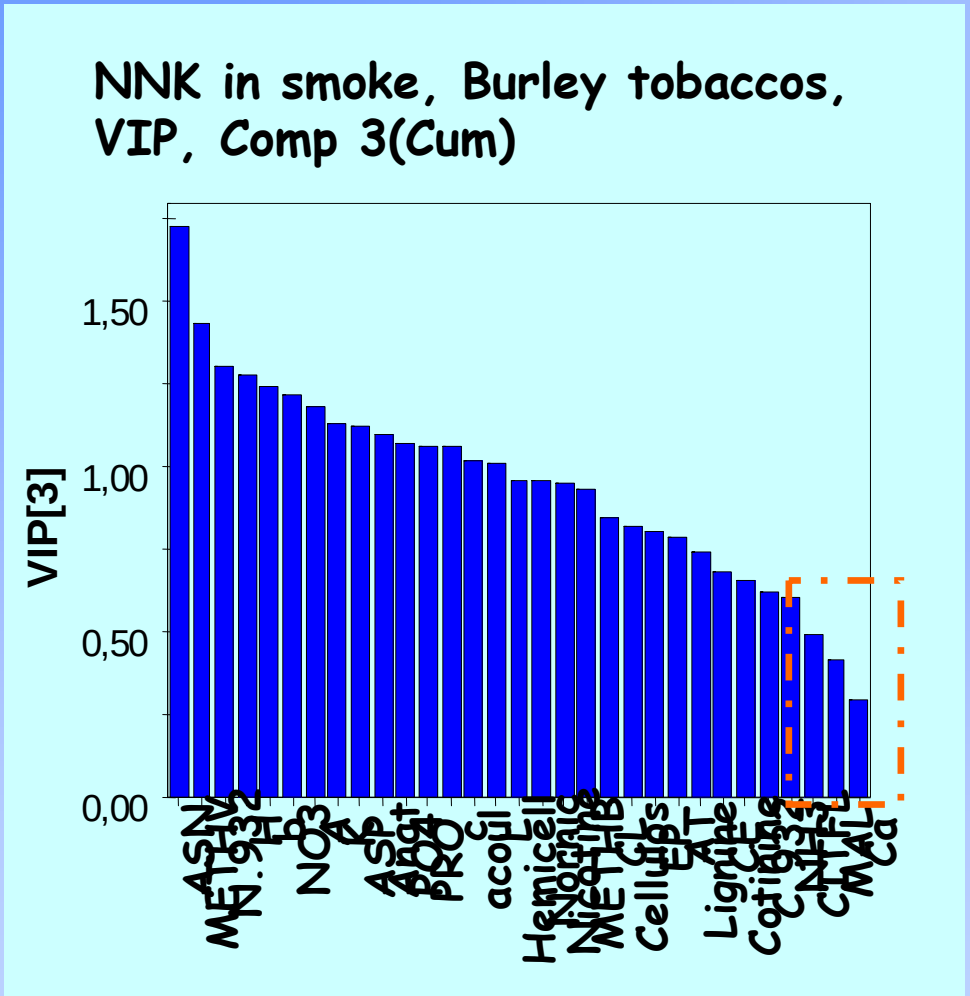


No results for TSNA in tobacco in the X matrix

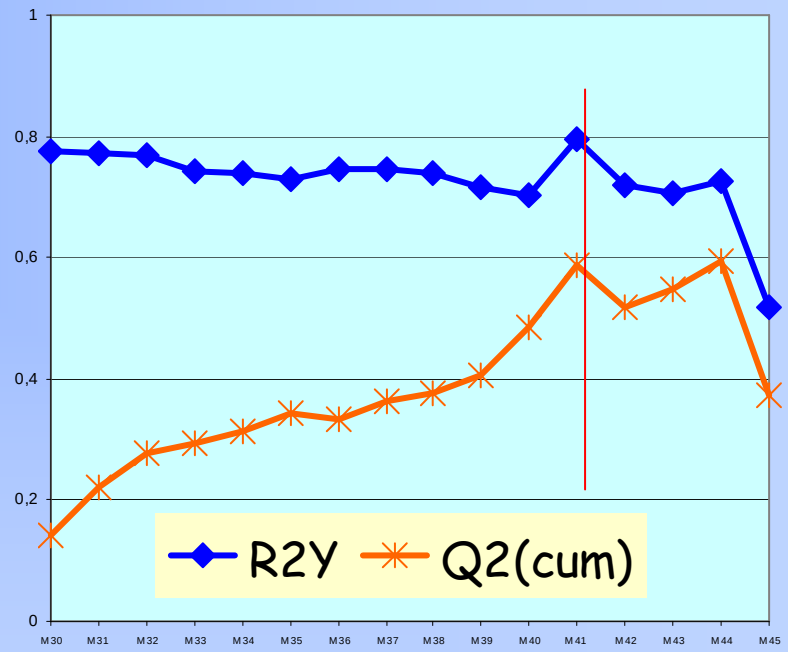
- Model M1 : 3 components
  - R2Y = 0.775 Q2cum=0.143

Poor Q2, the model needs to be improved

Improvement of the model



- Close look at the coefficients for each variable

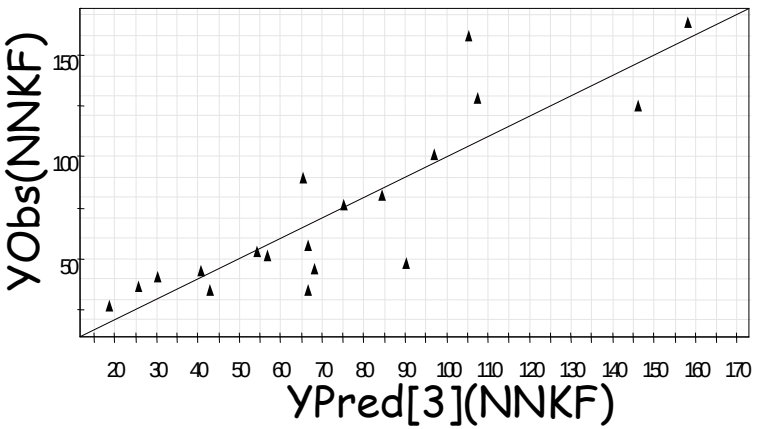


- Elimination of « noisy » variables
- Selection of the best R2Y, Q2cum

Final model

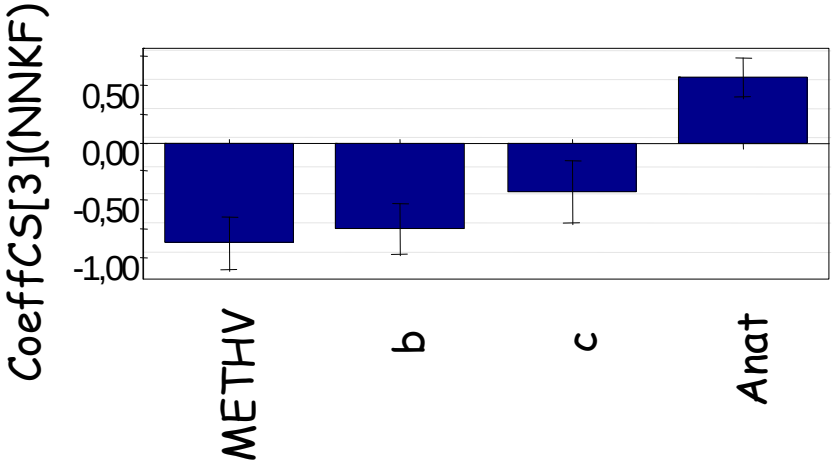
- 3 components, 4 variables, R2Y=0.75, Q2cum=0.64

Observed vs predicted



RMSEE = 23.8 (mean = 84)

Coefficients



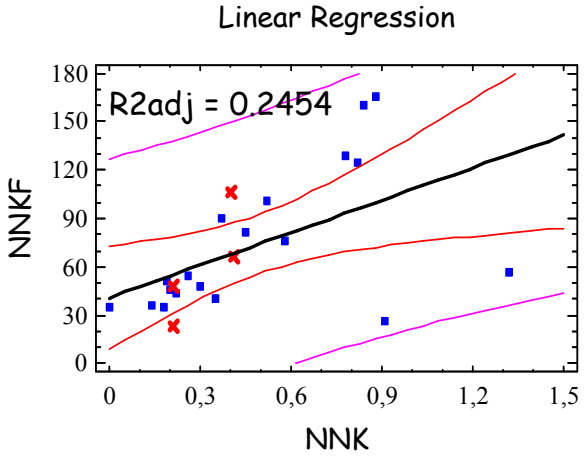
TSNAChimttTabttesVar.M70 (PLS), Untitled

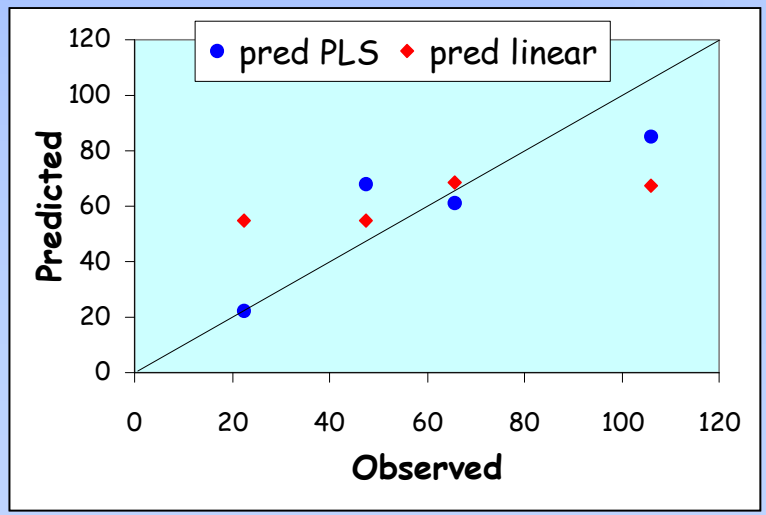
NNKF (ng/cig) = 1458 - 0.96 MEV - 35.9 b + 16.8 c + 270.7 Anat

Validation

- Prediction of NNKF levels for new tobaccos

	Observed	Predicted	
		PLS	Linear
CNBM	65,8	61	68,35
FRBM	47,5	67,8	54,9
SRBC	22,5	22,4	54,9
BY	106	85,2	67,6
	RMSEP	14,73	25,42

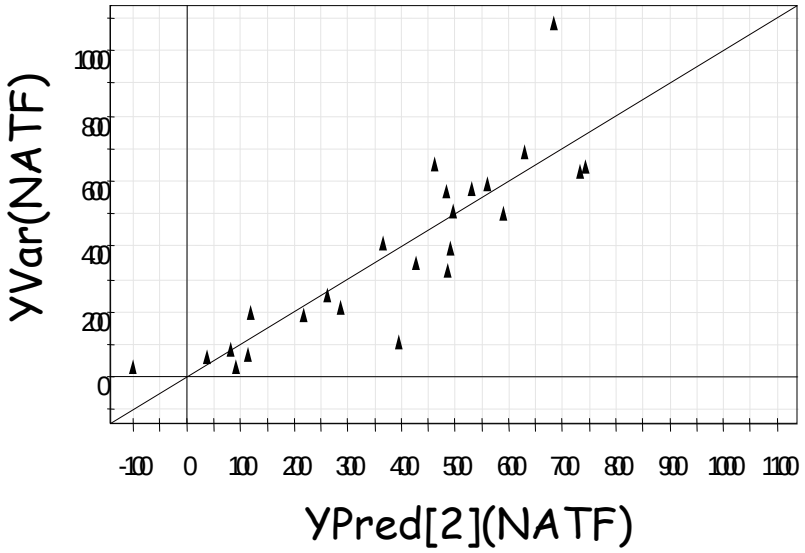




Prediction of NAT in smoke (Burley)

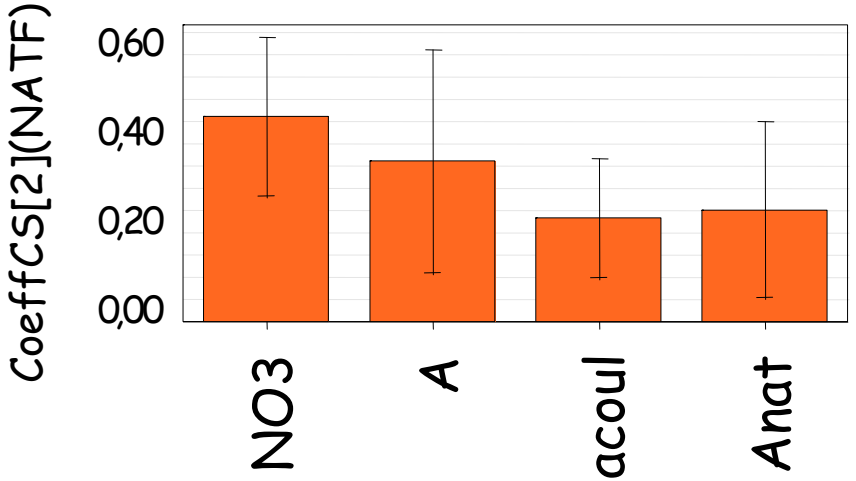
- Final model : R2Y = 0.762, Q2cum=0.693

Observed vs predicted



RMSEE = 138 (84,5)

Coefficients

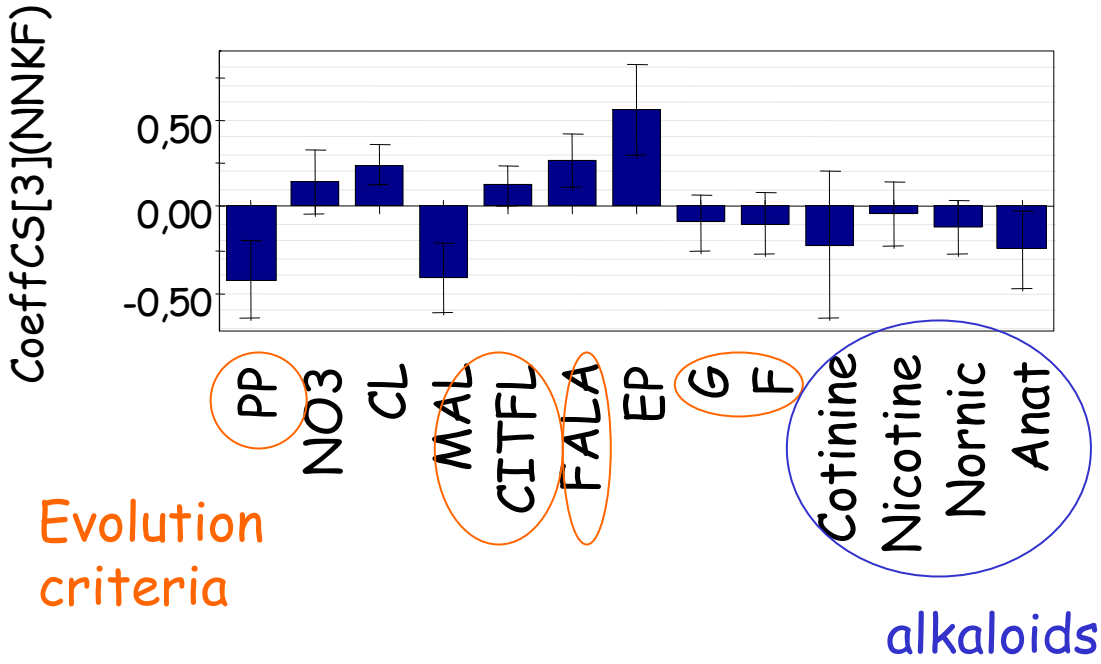


TSNAChimttTabttesVar.M43 (PLS), Untitled

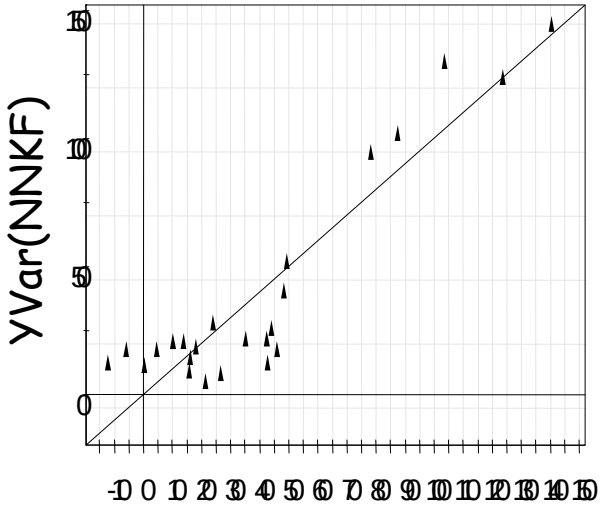
Prediction of NNN in smoke for FC

- Final model : R2Y = 0.855, Q2cum=0.652

Coefficients



Observed vs predicted



RMSEE = 17.8

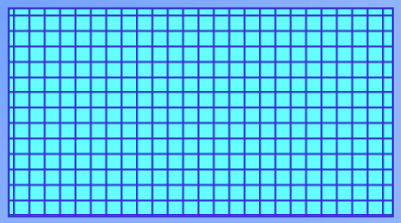
TSNAChimttTabttesVar.M55 (PLS), Untitled

Prediction of 4 TSNAs for DAC

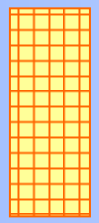
X

39 chemical results

24 tobaccos



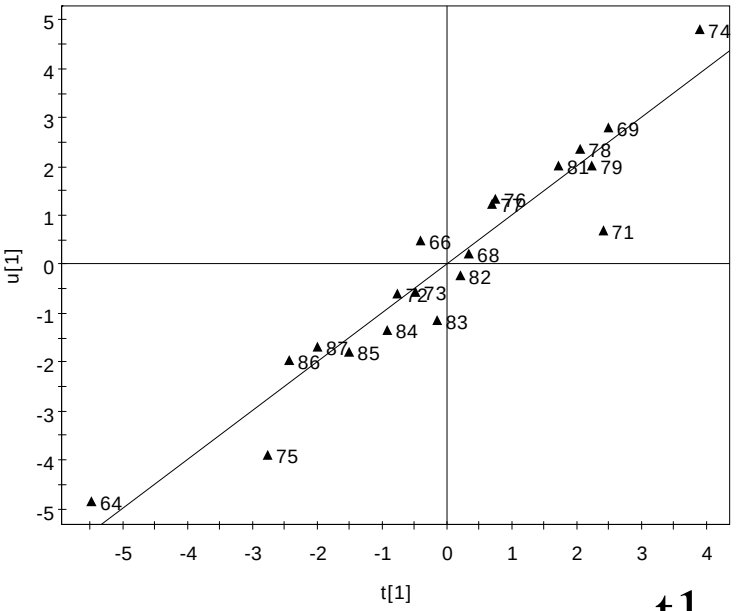
NNKF, NATF, NNNF, NABF



Y

- Final model :
- R2 = 0.86, Q2cum = 0.6
- 12 variables
- 5 components

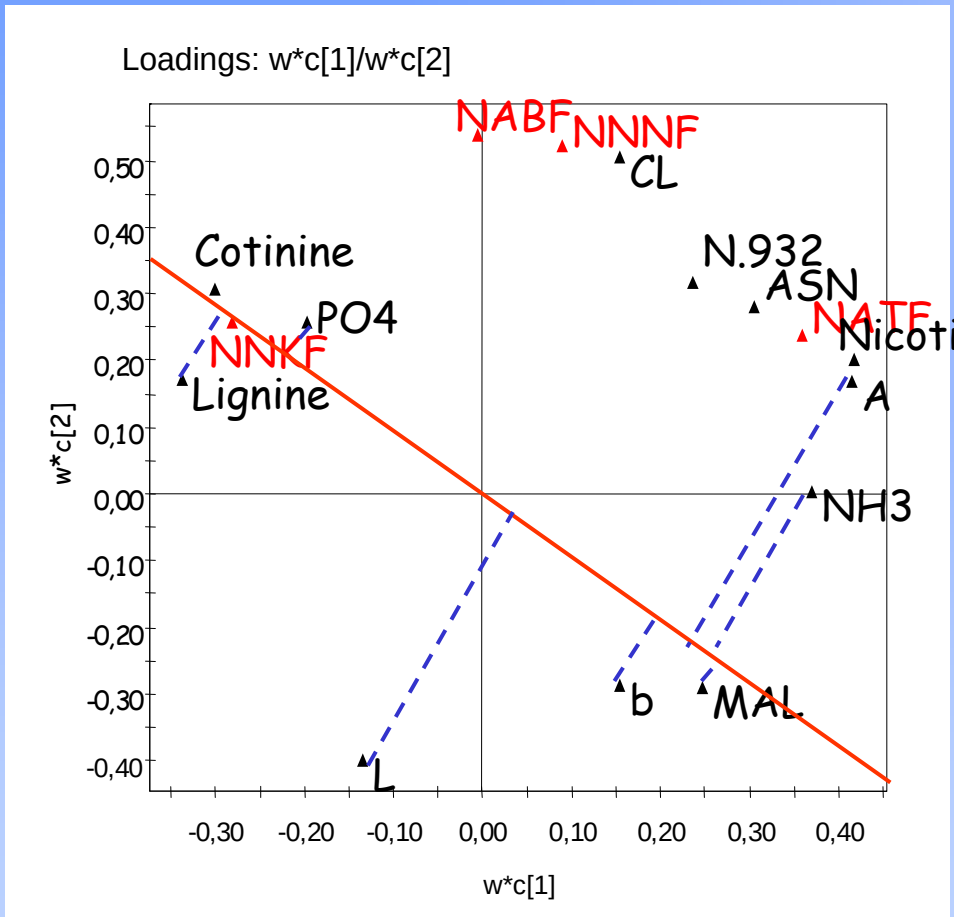
U1

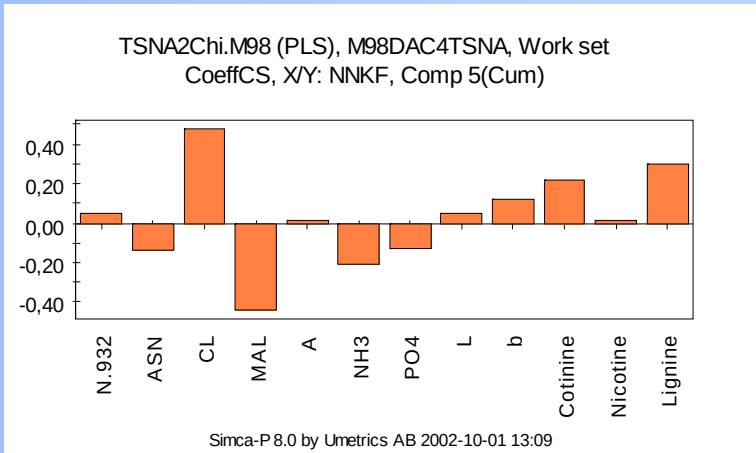


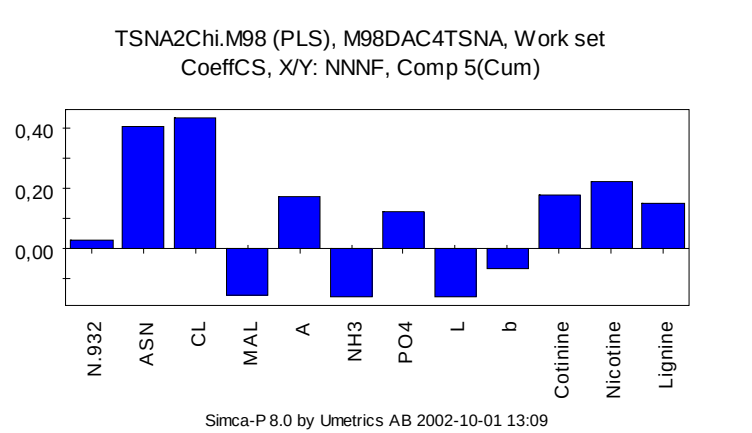
t1

DAC : prediction of 4 TSNA

4 models, same variables, different coefficients







Other predictions

TSNA tobacco in the models

	Flue cured			Burley			Sun cured			Dark Air Cured		
	R2Ycum	Q2cum	RMSEE	R2Ycum	Q2cum	RMSEE	R2Ycum	Q2cum	RMSEE	R2Ycum	Q2cum	RMSEE
4 TSNA	0,94	0,859	0,8 - 7,7	0,802	0,554	17 - 171				0.794	0.611	7-82
NNNF ng/cig	0,952	0,871	3,5	0,928	0,818	142,7	0,995	0,962	1,23	0,806	0,687	4,79
NNKF ng/cig	0,989	0,961	4,39	0,947	0,889	16,7				0,853		40,9
NATF ng/cig	0,86	0,815	9,9	0,834	0,781	107,6				0,781	0,599	26,7
NABF ng/cig	0,948	0,908	0,75	0,882	0,804	12				0,784	0,617	8,4

No TSNA tobacco in the models

	Flue cured			Burley			Sun cured			Dark Air Cured		
	R2Ycum	Q2cum	RMSEE	R2Ycum	Q2cum	RMSEE	R2Ycum	Q2cum	RMSEE	R2Ycum	Q2cum	RMSEE
4 TSNA	0,699	0,582	1,7 - 9	0,719	0,438	20-187				0.862	0.586	6-69
NNNF ng/cig	0,83	0,683	6,56	0,878	0,585	169,5	0,949	0,904	4,1	0,92	0,61	3
NNKF ng/cig	0,85	0,65	17,8	0,75	0,64	23,8				0,7	0,66	61,7
NATF ng/cig	0,879	0,527	9,7	0,762	0,69	137,9				0,806	0,593	25,1
NABF ng/cig	0,905	0,745	1,13	0,805	0,594	14,46				0,752	0,613	9

Conclusion and Prospects

- PLS regression is an interesting method to quickly forecast TSNA potential of a tobacco
- Variables used in the model can give clues for a better understanding of TSNA formation or raise questions

- Validation of the models on new observations are on progress
- Near Infrared Spectroscopy is now tested
- Application to cigarettes with different NTM : use of PLS to predict TSNAs in smoke according to tobacco characteristics and NTM physical properties