Soil Spectroscopy: Principle and Applications

Prof. Eyal Ben Dor
Department of Geography and Human Environment

Brno Czech Republic, June 25-26

european social fund in the czech republic
EUROPEAN UNION
MINISTRY OF EDUCATION, YOUTH AND SPORTS
OP Education for Competitiveness
INVESTMENTS IN EDUCATION DEVELOPMENT

SPeR – A
(Chemometrics)

Basic Theory

Lesson 5

Some information and a suggestion

NIRA – Near Infrared Analysis – First paper by Ben Gerah and Norris 1967 (NIR- 1-2.5um)

Also can be found as Sper-ANear Infrared Spectroscopy

Non of these terms, as well as the vis-NIR reflects what we are realy doing: chmometrics based on (reflectance) spectroscopy

As we measure Reflectance and do Spectral Analysis the correct term should be :

SpeR - A (Spectral Reflectance - Analysis). It can be done in the VNIR-SWIR, in the VNIR or in the SWIR

It is also a **spare** method for the wet chemistry

Chemomatrics

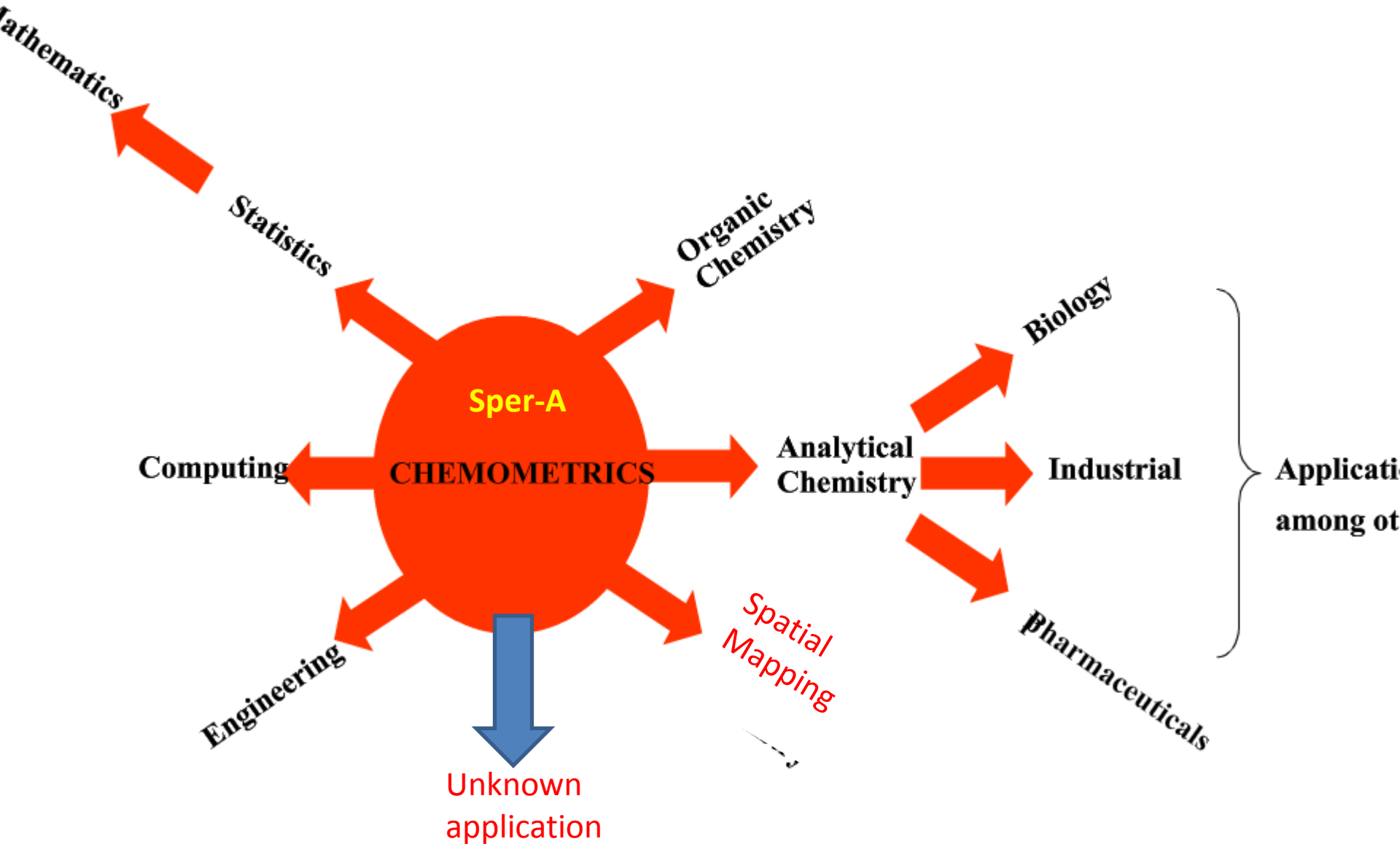
Chemometrics is the science of extracting information from chemical systems by data-mining means.

It is a highly interfacial discipline, using methods frequently employed in core data-analytic disciplines such as multivariate statistics, applied mathematics, and computer science, in order to address problems in chemistry, Spectroscopy , biochemistry, medicine, biology and chemical engineering.

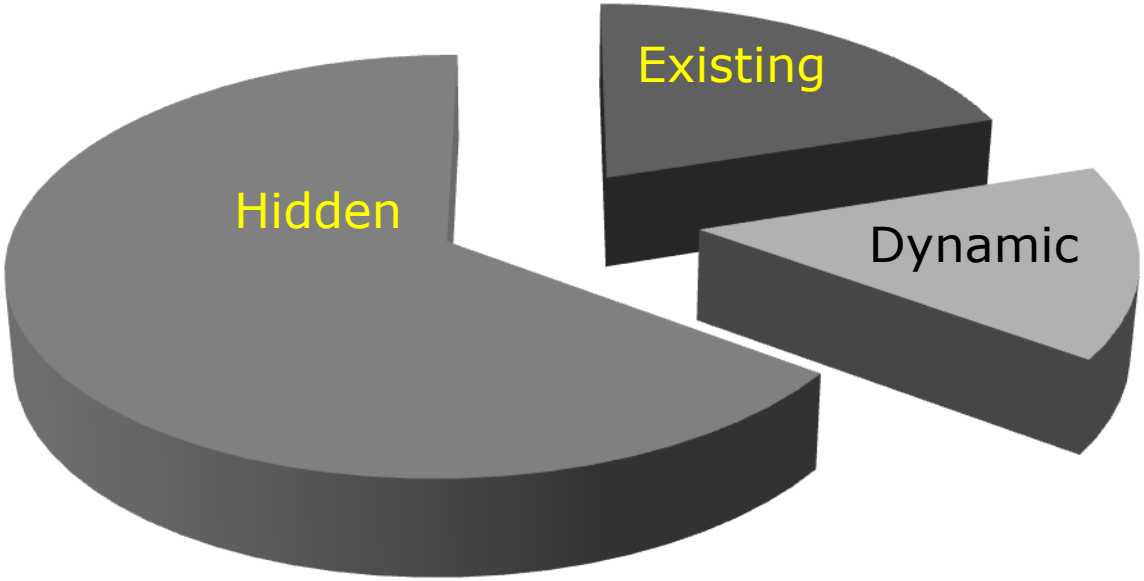
Spectral – Chemomatrics
Sper-A

- Using spectral data to predict chemical AND physical information

- Spectral- data mining (mathematics and statistics)


athematics
Statistics
Organic Chemistry
Sper-A
Computing
CHEMOMETRICS
Analytical Chemistry
Biology
Industrial
Applicati
among ot
Pharmaceuticals
Spatial Mapping
Engineering
Unknown application

Many applications are still NOT used in Sper-A



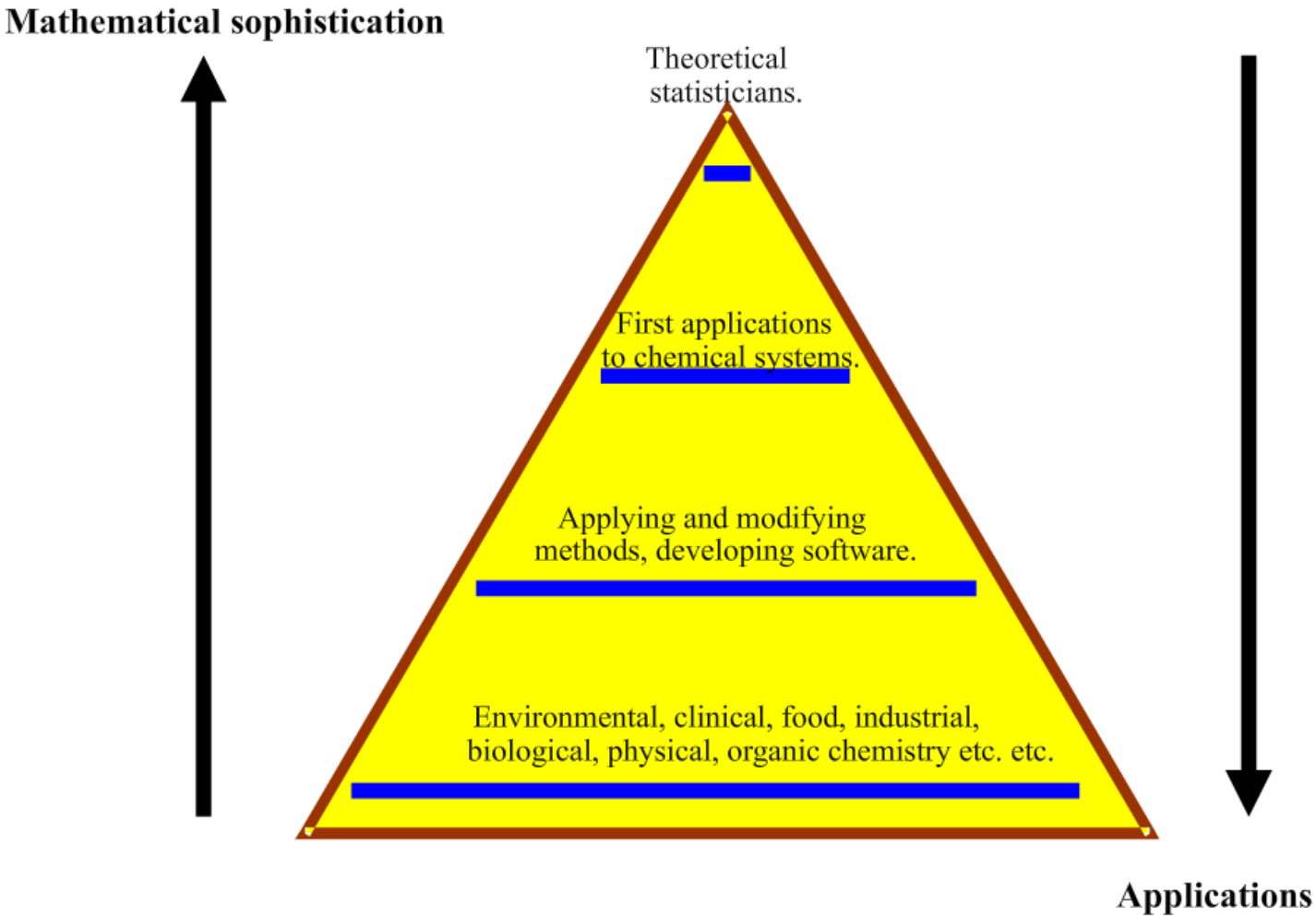
ASD 20 anniversary workshop, Boulder US, October 2009

THE ULTIMATE DREAM

A lower level where no knowledge of chemometrics is required – good software.

- **E.g. technician in warehouse looking at quality of drug**
- **Nurse in hospital looking at diagnosis**
- **Operator in manufacturing plant looking at whether product is OK.**

HIERARCHY OF USERS



Tools for Sper-A

1. **Methods.** (Ways to analyze the data)

2. **Software.** (Means to analyze the data)

3. **Instrumental techniques.** (Means to collect the data)

4. **Applications.** (Utilization the analyzed data)

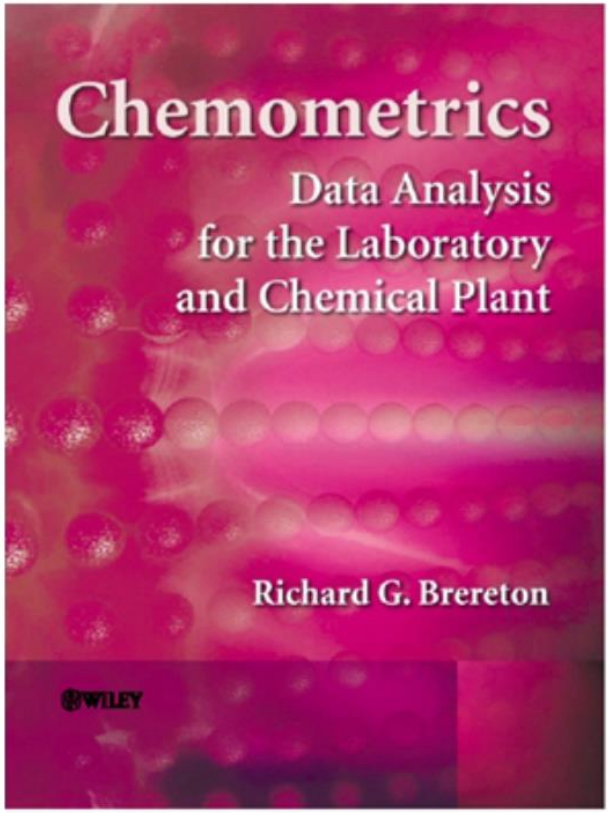
METHODS

- Experimental design
- Pattern recognition
- Calibration

R.G.Brereton, Chemometrics : Data Analysis for the Laboratory and Chemical Plant, Wiley, Chichester, 2003 and 2004

www.spectroscopynow.com Website

Chemometrics Channel



http://www.powershow.com/view/11d5b9-MTYzN/INTRODUCTION_TO_CHEMOMETRICS_powerpoint_ppt_presentation

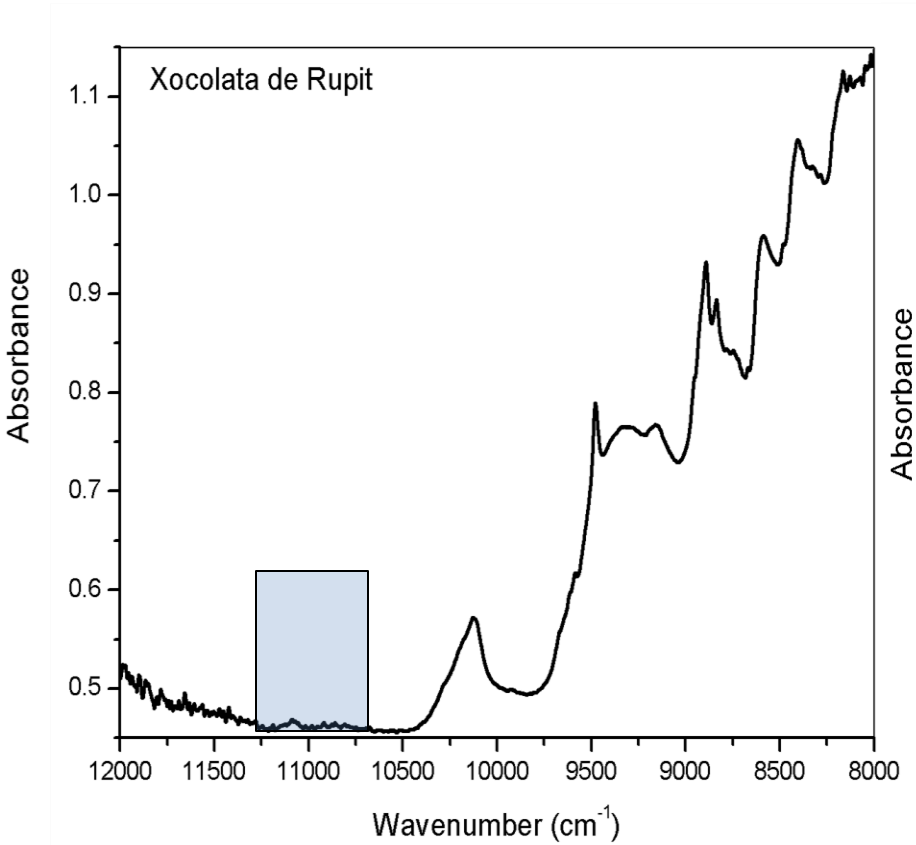
There are Two Basic Ways

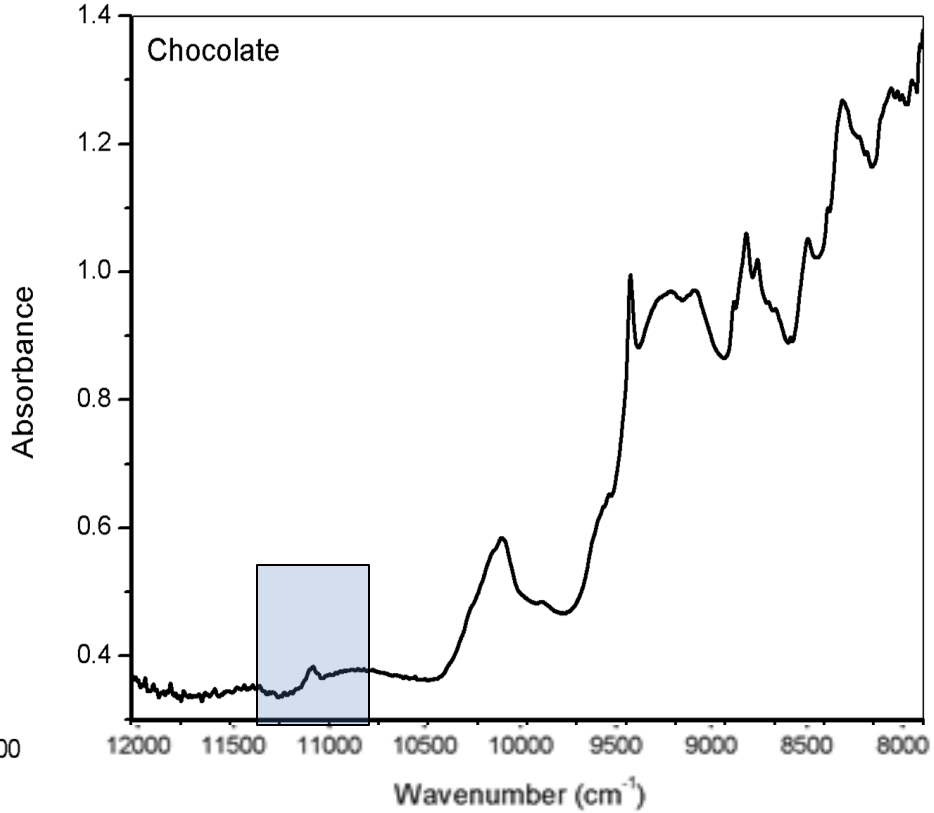
- Supervised: Known features with significant changes

- Unsupervised: No spectral features are known for the application, no spectral features are seen by naked eyes

For the unsupervised – sophisticated "data mining" too is needed (chemometrics approach)

Supervised



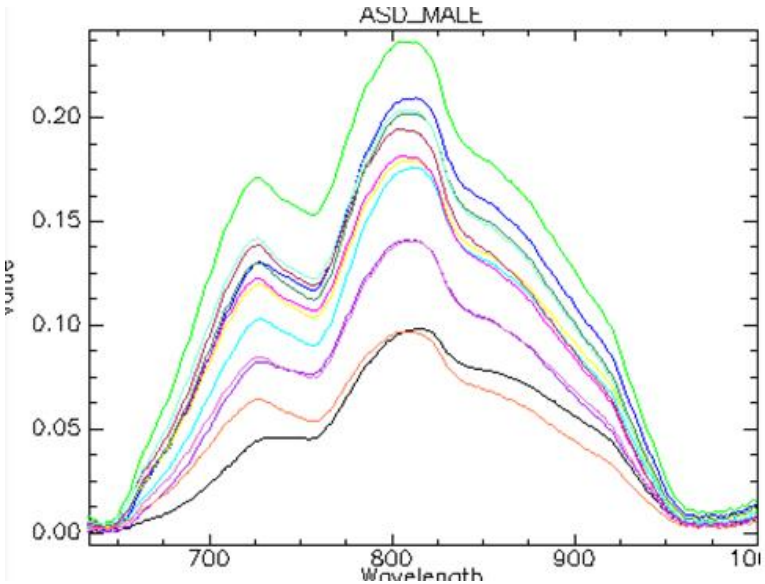


Difference: Some times visible, some time s not

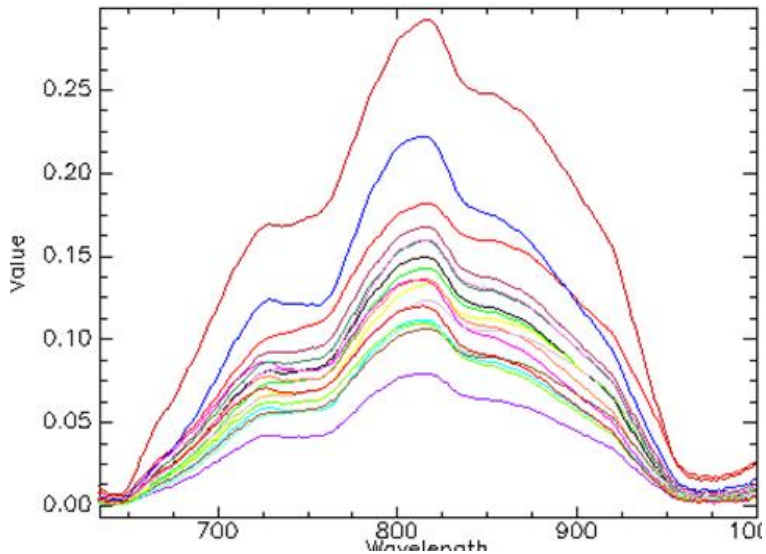
Un supervised

No apriori knowledge is known

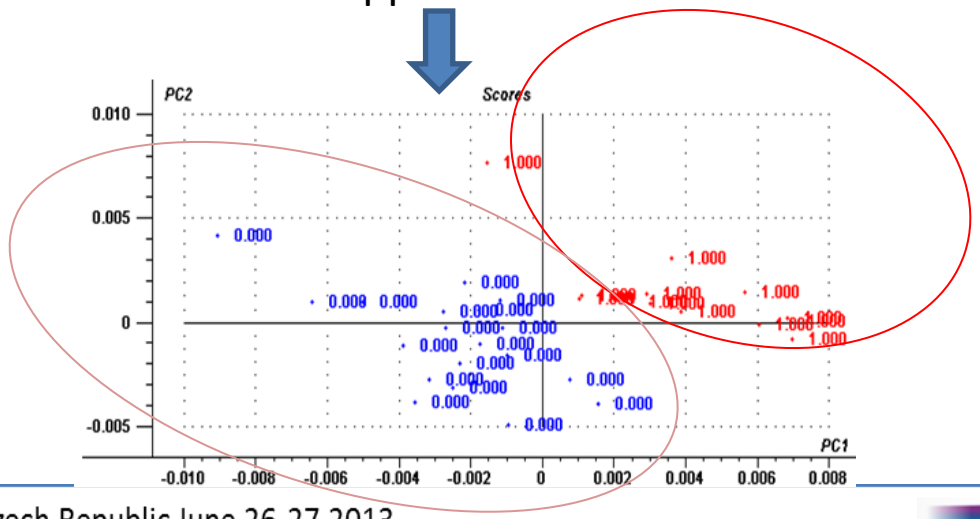
1



0



Data mining approach

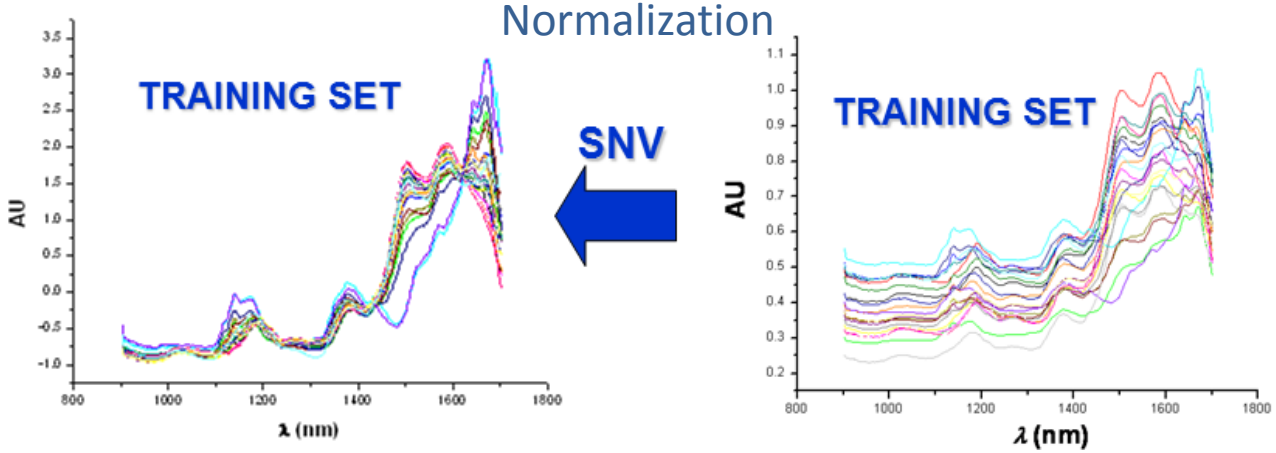


Methods for the Un supervised (Examples)

- Multivariate Regression (MLR)
- PCA
- PCR
- PLSR
- Neural Net Work

Pre processing stage:
any method that orthogonally applies to all variables data set

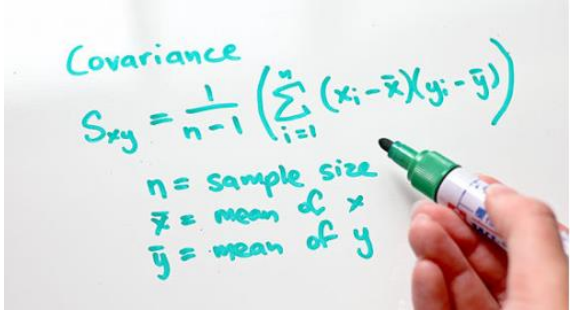
Normalization



Examples:

- Smoothing
- Derivation
- Normalization
- A = log (1/R)
- Others

Linear Regression



Correlation is the association between two variables - the amount by which they covary. The most frequently used measure is Pearson's product moment correlation coefficient, which is a parametric measure of linear association. It is defined as the ratio of the covariance between two variables to the square root of the product of the two variances.

$$\rho = \frac{Cov(X_1,X_2)}{(\sigma_1^2.\sigma_2^2)^{\frac{1}{2}}} = \frac{Cov(X_1,X_2)}{\sigma_1.\sigma_2} \qquad (1)$$

Working with sample data, r, an unbiassed estimate of the population correlation coefficient ρ, becomes

$$r = \frac{Cov(X_1,X_2)}{s_1.s_2} = \frac{N\Sigma X_1X_2 - \Sigma X_1.\Sigma X_2}{((N\Sigma X_1^2 - (\Sigma X_1)^2).(N\Sigma X_2^2 - (\Sigma X_2)^2))^{\frac{1}{2}}} \qquad (2)$$

It varies from +1.0 (perfect positive correlation) through zero (no correlation) to -1.0 (perfect negative correlation). If there are two uncorrelated variables, each with a mean of zero and unit variance, with a bivariate normal distribution, then the equiprobability contours (lines enclosing an area where there is a given probability of a value occurring) are circular. The axes interest at 90°, and cos 90° = r = 0. Equiprobability contours for p = 0.5, 0.75, 0.95, 0.99, 0.999 are shown in fig. 1.

r= cos (α)

No correlation

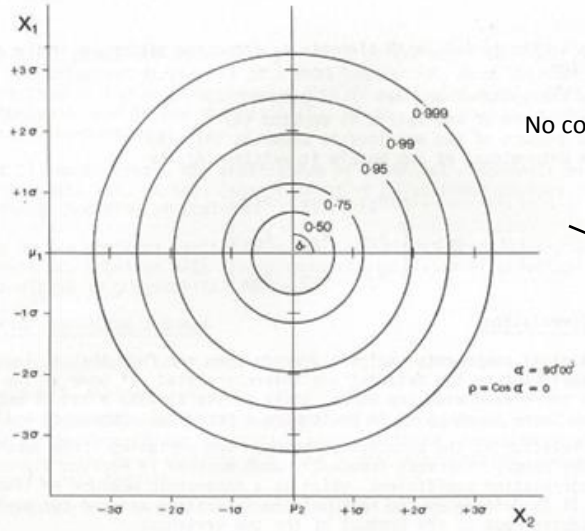


Fig. 1 Equiprobability contours for two uncorrelated variables

Medium correlation

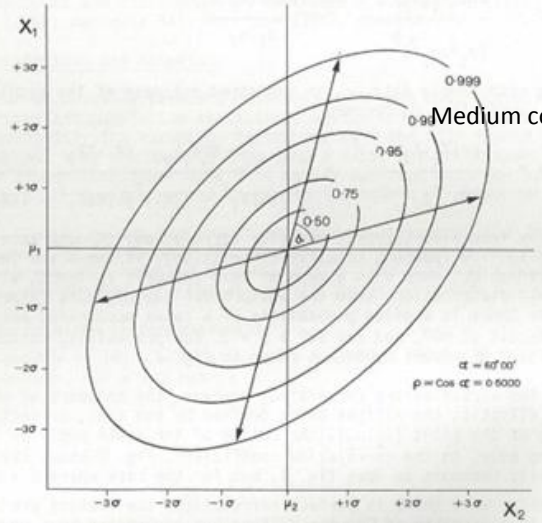


Fig. 2 Equiprobability contours for two correlated variables

6

High correlation

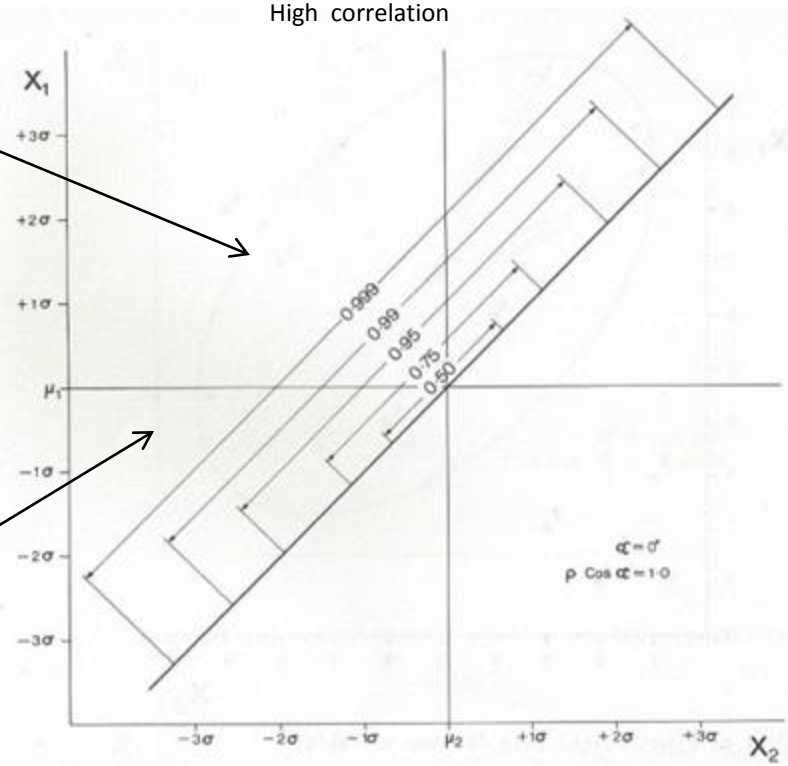


Fig. 3 Equiprobability trace for two perfectly correlated variables

line on which each member of the bivariate normal population lies, and thus one variable is completely defining the other. This case is illustrated in fig. 3.

Stepwise multiple regression

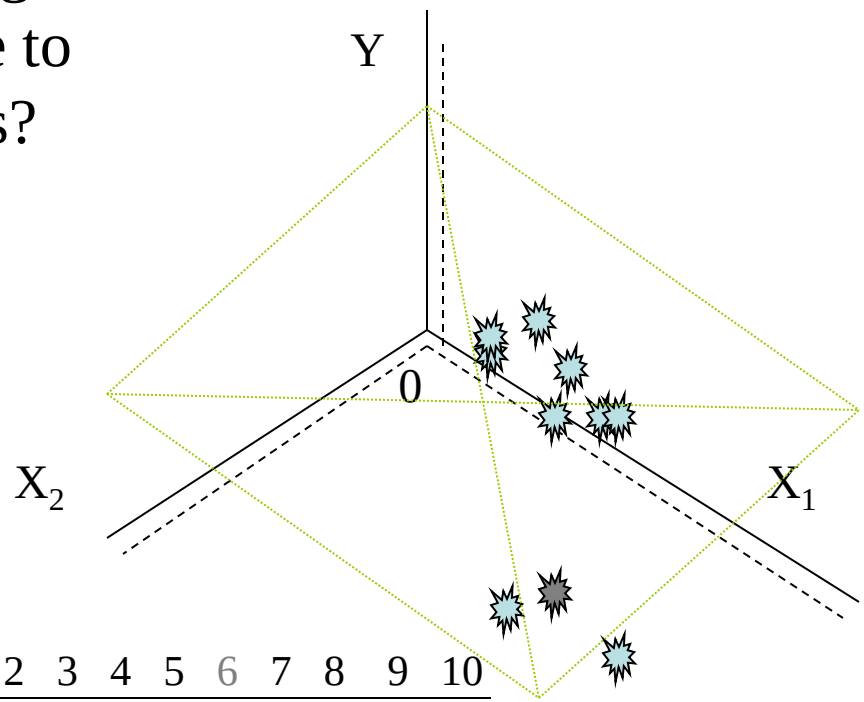
- Stepwise regression is designed to find the most parsimonious set of predictors that are most effective in predicting the dependent variable.

- Variables are added to the regression equation one at a time, using the statistical criterion of maximizing the R^2 of the included variables.

- When none of the possible addition can make a statistically significant improvement in R^2, the analysis stops.

Multiple Regression

What multiple regression does it fit a plane to these coordinates?



Case:	1	2	3	4	5	6	7	8	9	10
Children (Y):	2	5	1	9	6	3	0	3	7	7
Education (X_1)	12	16	20	12	9	18	16	14	9	12
Income 1=$10K ($X_2$):	3	4	9	5	4	12	10	1	4	3

Multiple Regression

- Mathematically, that plane is:

  $Y^{\wedge} = a + b_1X_1 + b_2X_2$

  a = y-intercept, where X's equal zero

  b=coefficient or slope for each variable

  For our problem, the equation is:

  $Y^{\wedge} = 11.8 - .36X_1 - .40X_2$

  Expected # of Children = 11.8 - .36*Educ - .40*Income

Multiple Regression

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.757[a]	.573	.534	2.33785

a. Predictors: (Constant), Income

57% of the variation in number of children is explained by education and income!

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	161.518	2	80.759	14.776	.000[a]
	Residual	120.242	22	5.466		
	Total	281.760	24			

a. Predictors: (Constant), Income, Education

b. Dependent Variable: Children

$$\hat{Y} = 11.8 - .36X_1 - .40X_2$$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	11.770	1.734		6.787	.000
	Education	-.364	.173	-.412	-2.105	.047
	Income	-.403	.194	-.408	-2.084	.049

a. Dependent Variable: Children

Multiple Regression

r^2

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.757[a]	.573	.534	2.33785

a. Predictors: (Constant), Income

$$\frac{\Sigma (Y - \bar{Y})^2 - \Sigma (Y - \hat{Y})^2}{\Sigma (Y - \bar{Y})^2}$$

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	161.518	2	80.759	14.776	.000[a]
	Residual	120.242	22	5.466		
	Total	281.760	24			

a. Predictors: (Constant), Income, Education

b. Dependent Variable: Children

$$\hat{Y} = 11.8 - .36X_1 - .40X_2$$

161.518 ÷ 261.76 = .573

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	11.770	1.734		6.787	.000
	Education	-.364	.173	-.412	-2.105	.047
	Income	-.403	.194	-.408	-2.084	.049

a. Dependent Variable: Children

Multi Linear Regression (MLR) for Sper-A

The number of Independent Variable (spectral information) must be equal or lower than the samples' number (10% is recommended).

Example: A spectrometer has 1000 wavelengths (independent variables)

To run MLR for any attribute the number of soil samples has to be 100,000 !

As this is not realistic- a method to compress the wavelengths into meaningful number is needed

Solution:

Finding from the 1000 wavelength the **few that are highly correlated** to the property in question and only then, run MLR.

If we have 40 samples we must select 4> wavelengths

How we do that?

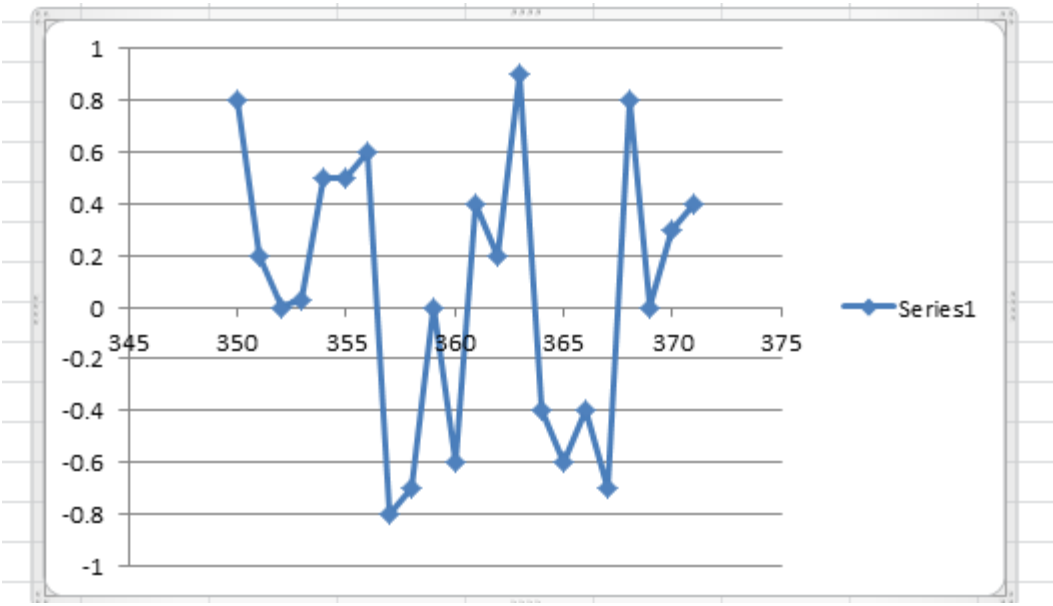
Correlogram: Spectrum of λ against R

X vs. Y(λ, 1,2,3,4,5.6,7) → R

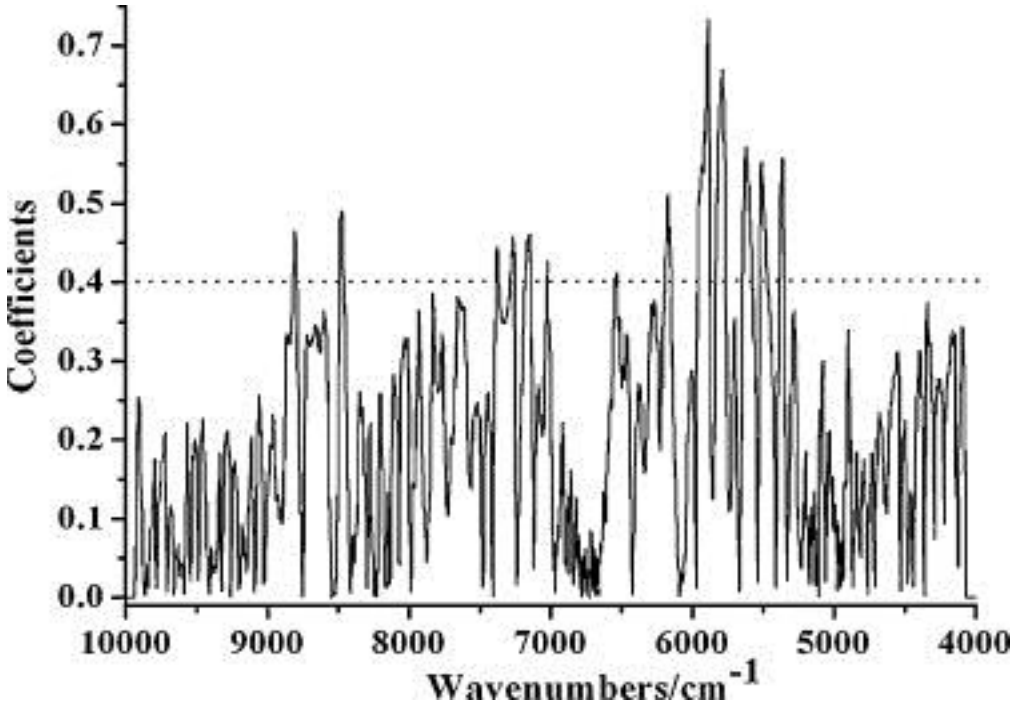
Attributes (x) | Spectral (y) | Sample | R

A	B	C	D	E	F	G	H	I	J	K
Wavelength	Clay	1	2	3	4	5	6	7		R
350	20	0.036597	0.075812	0.042233	0.067764	0.044539	0.084978	0.040354		0.8
351	32	0.037566	0.07866	0.041297	0.068742	0.042737	0.077825	0.032997		0.2
352	45	0.039534	0.075939	0.040123	0.065941	0.040151	0.075078	0.029863		0
353	32	0.041243	0.072183	0.040731	0.063077	0.038358	0.076869	0.032891		0.03
354	67	0.040161	0.073291	0.042836	0.064507	0.036743	0.077427	0.036693		0.5
355	55	0.041081	0.072857	0.046125	0.066654	0.041941	0.080258	0.037634		0.5
356	88	0.043806	0.07418	0.047051	0.068508	0.045377	0.082281	0.038051		0.6
357	23	0.046124	0.078963	0.043976	0.069362	0.041095	0.081242	0.039751		-0.8
358	44	0.04065	0.07851	0.043402	0.068178	0.040155	0.080866	0.039031		-0.7
359	54	0.036009	0.074677	0.041282	0.064074	0.03769	0.079304	0.036201		0
360	67	0.037128	0.07154	0.037754	0.059634	0.034072	0.077133	0.033179		-0.6
361	56	0.044817	0.07871	0.04182	0.065099	0.041206	0.080463	0.035983		0.4
362	44	0.039041	0.076669	0.039498	0.063749	0.039887	0.077967	0.029918		0.2
363	12	0.02679	0.069045	0.033855	0.05777	0.032453	0.072566	0.021016		0.9
364	4	0.033287	0.074241	0.038831	0.061219	0.034768	0.076412	0.029136		-0.4
365	55	0.038166	0.077927	0.041239	0.064212	0.035108	0.080229	0.033523		-0.6
366	22	0.040581	0.07958	0.042114	0.066724	0.036304	0.08221	0.034597		-0.4
367	23	0.043701	0.081406	0.044926	0.070078	0.042514	0.082508	0.037109		-0.7
368	43	0.04293	0.079317	0.042208	0.067472	0.039667	0.081978	0.034157		0.8
369	52	0.038207	0.073806	0.038647	0.061776	0.034788	0.07901	0.029572		0
370	35	0.03209	0.06813	0.038058	0.057491	0.033977	0.07458	0.027347		0.3
371	46	0.03335	0.072875	0.03873	0.061967	0.035109	0.078316	0.029684		0.4

Correlogram



Examples from the Literature (high spectral resolution)



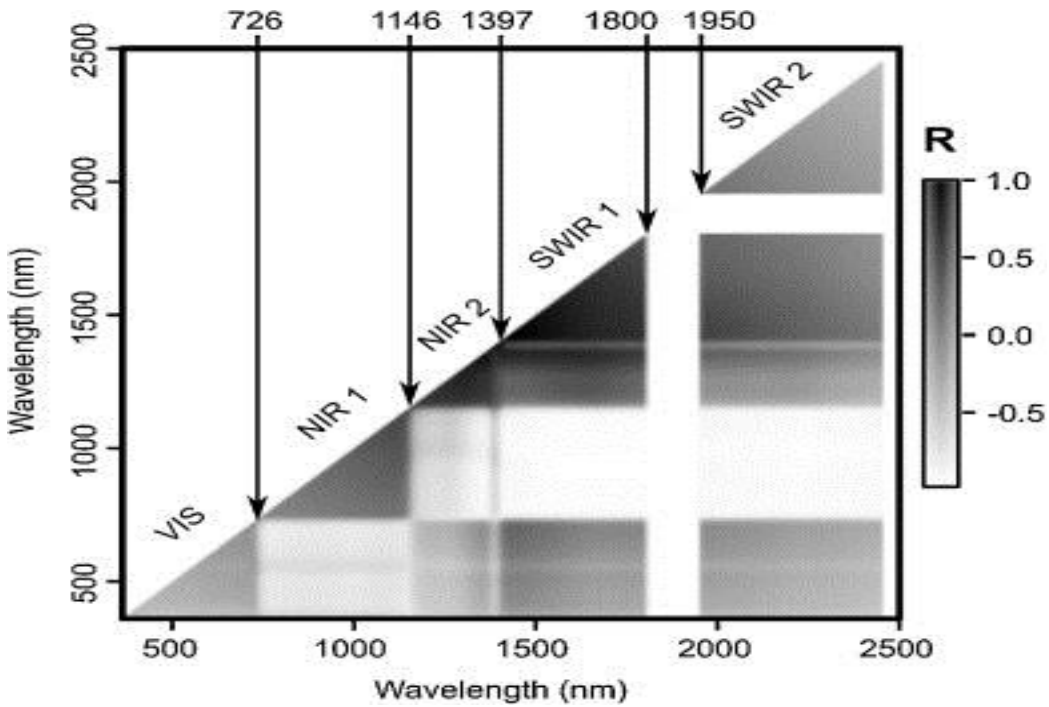
Once the wavelengths are selected

Run

- MLR
- PCA
- PLSR

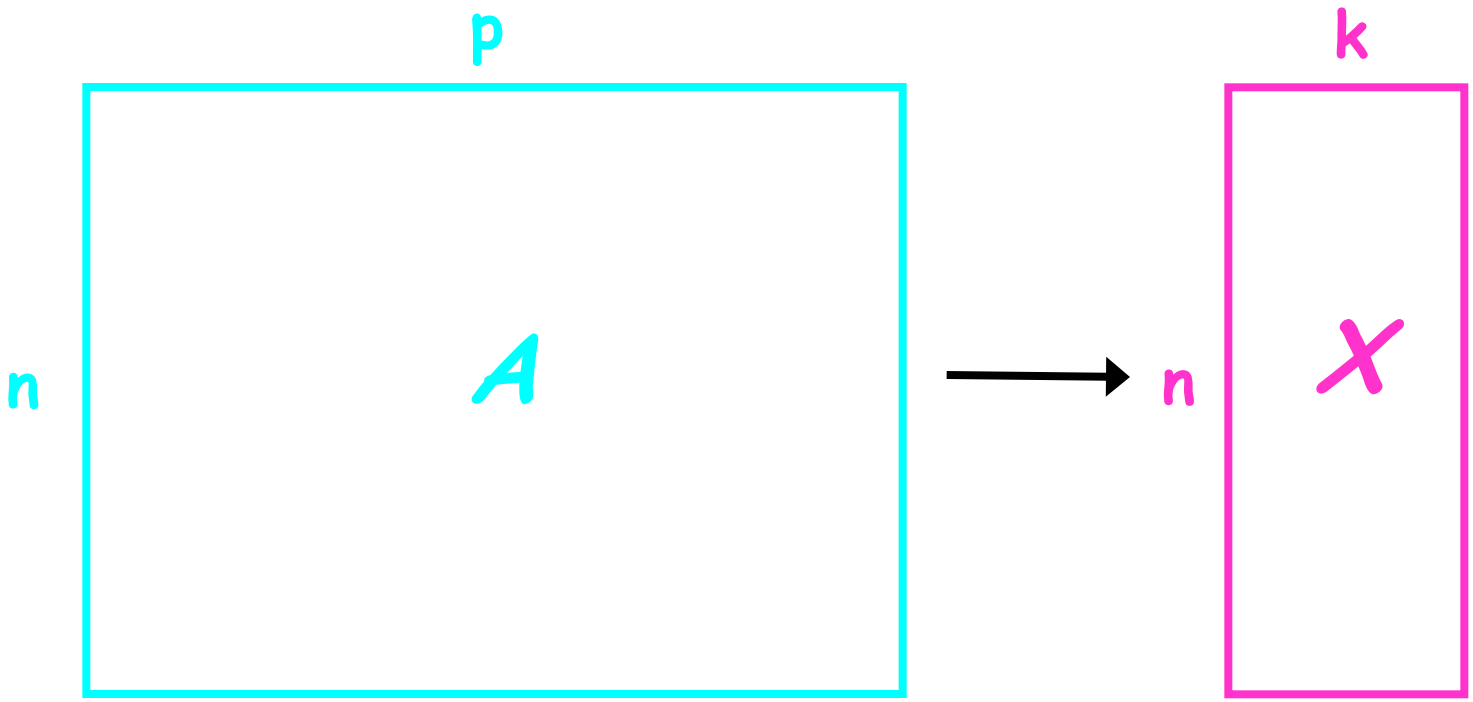
Principal Component Analysis (PCA)

Redundant spectral information: PCA reduction



Data Reduction

- summarization of data with many (p) variables by a smaller set of (k) derived (synthetic, composite) variables.



Data Reduction

- "Residual" variation is information in *A* that is not retained in *X*
- balancing act between
  - clarity of representation, ease of understanding
  - oversimplification: loss of important or relevant information.

Principal Component Analysis (PCA)

- probably the most widely-used and well-known of the "standard" multivariate methods
- invented by Pearson (1901) and Hotelling (1933)
- first applied in ecology by Goodall (1954) under the name "factor analysis" ("principal factor analysis" is a synonym of PCA).

Principal Component Analysis (PCA)

- takes a data matrix of *n* objects by *p* variables, which may be correlated, and summarizes it by uncorrelated axes (principal components or principal axes) that are linear combinations of the original *p* variables
- the first *k* components display as much as possible of the variation among objects.

Geometric Rationale of PCA

- objects are represented as a cloud of *n* points in a multidimensional space with an axis for each of the *p* variables
- the centroid of the points is defined by the mean of each variable
- the variance of each variable is the average squared deviation of its *n* values around the mean of that variable.

$$V_i = \frac{1}{n-1}\sum_{m=1}^{n}\left(X_{im} - \overline{X}_i\right)^2$$

Geometric Rationale of PCA

- degree to which the variables are linearly correlated is represented by their covariances.

$$C_{ij} = \frac{1}{n-1}\sum_{m=1}^{n}\left(X_{im} - \overline{X}_i\right)\left(X_{jm} - \overline{X}_i\right)$$

Covariance of variables *i* and *j*

Sum over all *n* objects

Value of variable *i* in object *m*

Mean of variable *i*

Value of variable *j* in object *m*

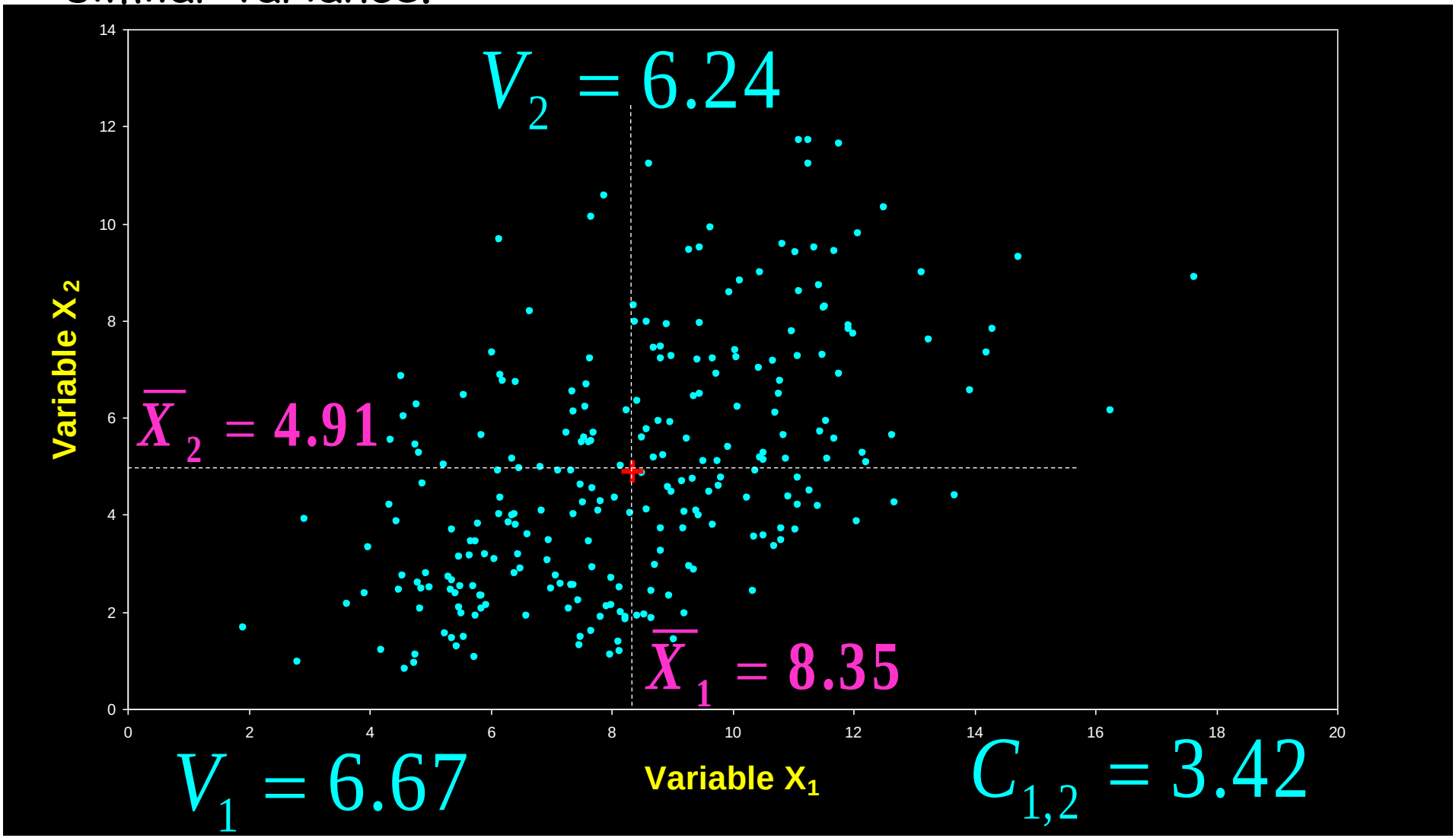
Mean of variable *j*

Geometric Rationale of PCA

- objective of PCA is to rigidly rotate the axes of this *p*-dimensional space to new positions (principal axes) that have the following properties:
  - ordered such that principal axis 1 has the highest variance, axis 2 has the next highest variance, ... , and axis *p* has the lowest variance
  - covariance among each pair of the principal axes is zero (the principal axes are uncorrelated).

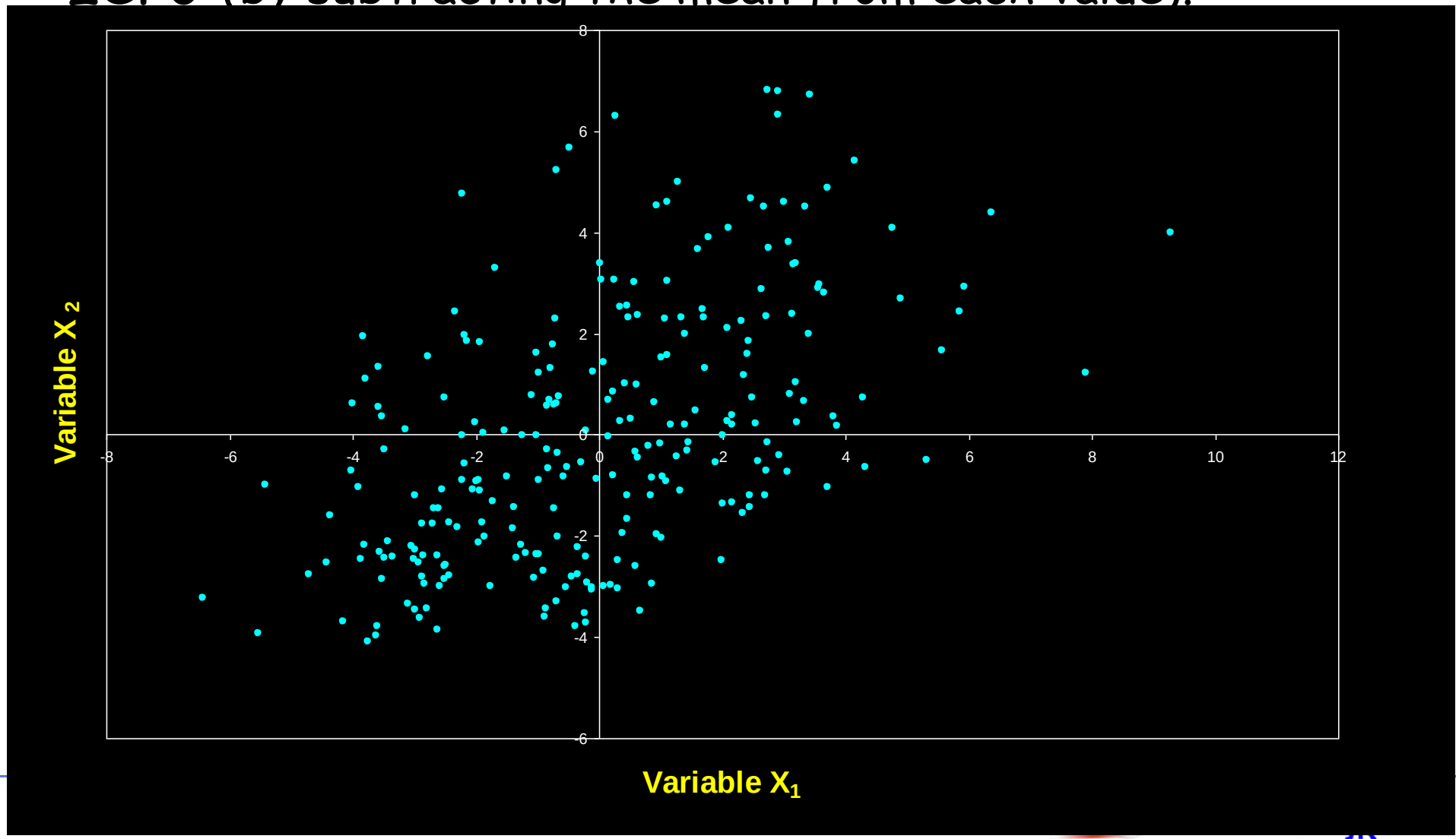
2D Example of PCA

- variables X_1 and X_2 have positive covariance & each has a similar variance.



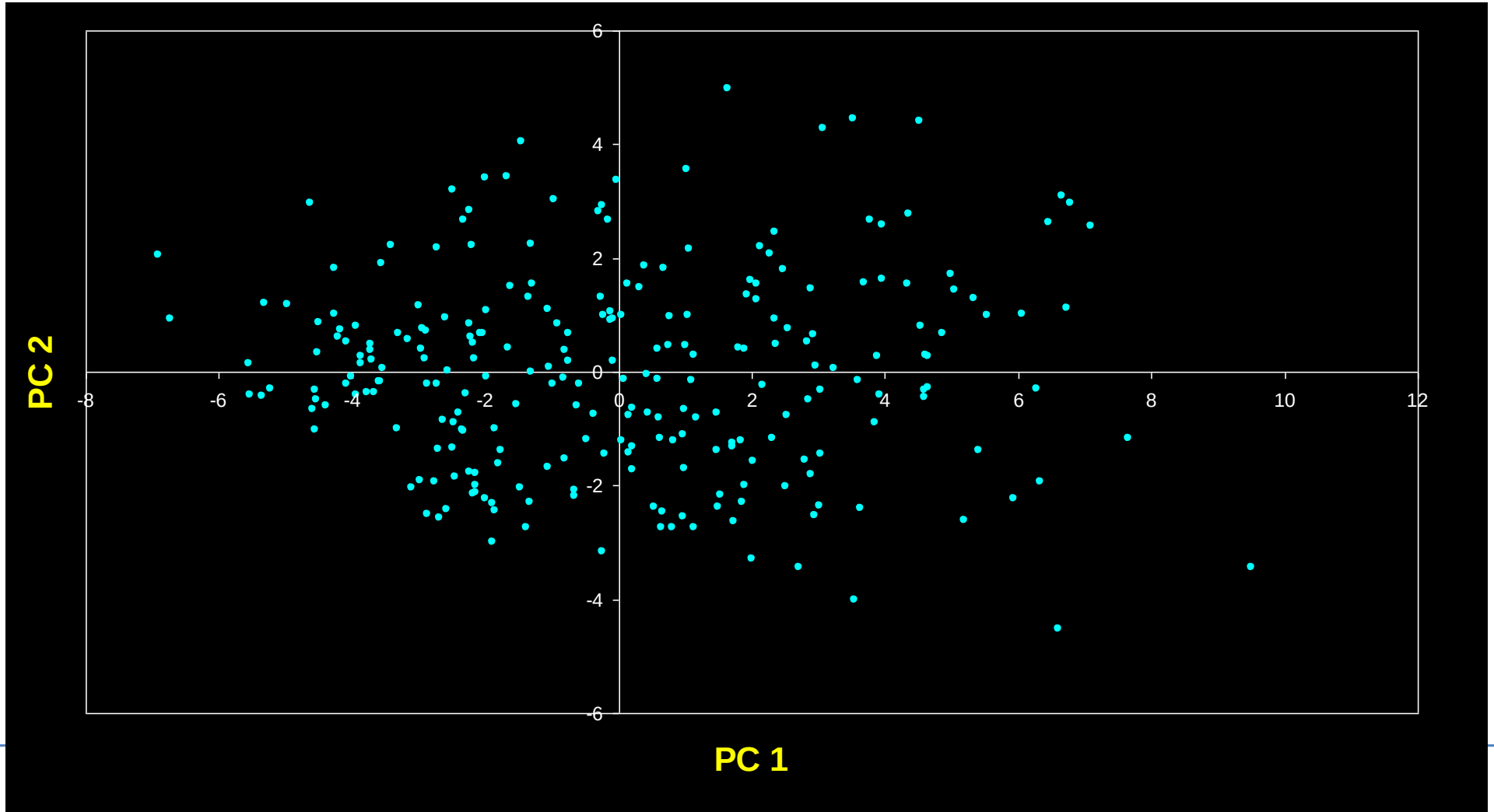
Configuration is Centered

- each variable is adjusted to a mean of zero (by subtracting the mean from each value).

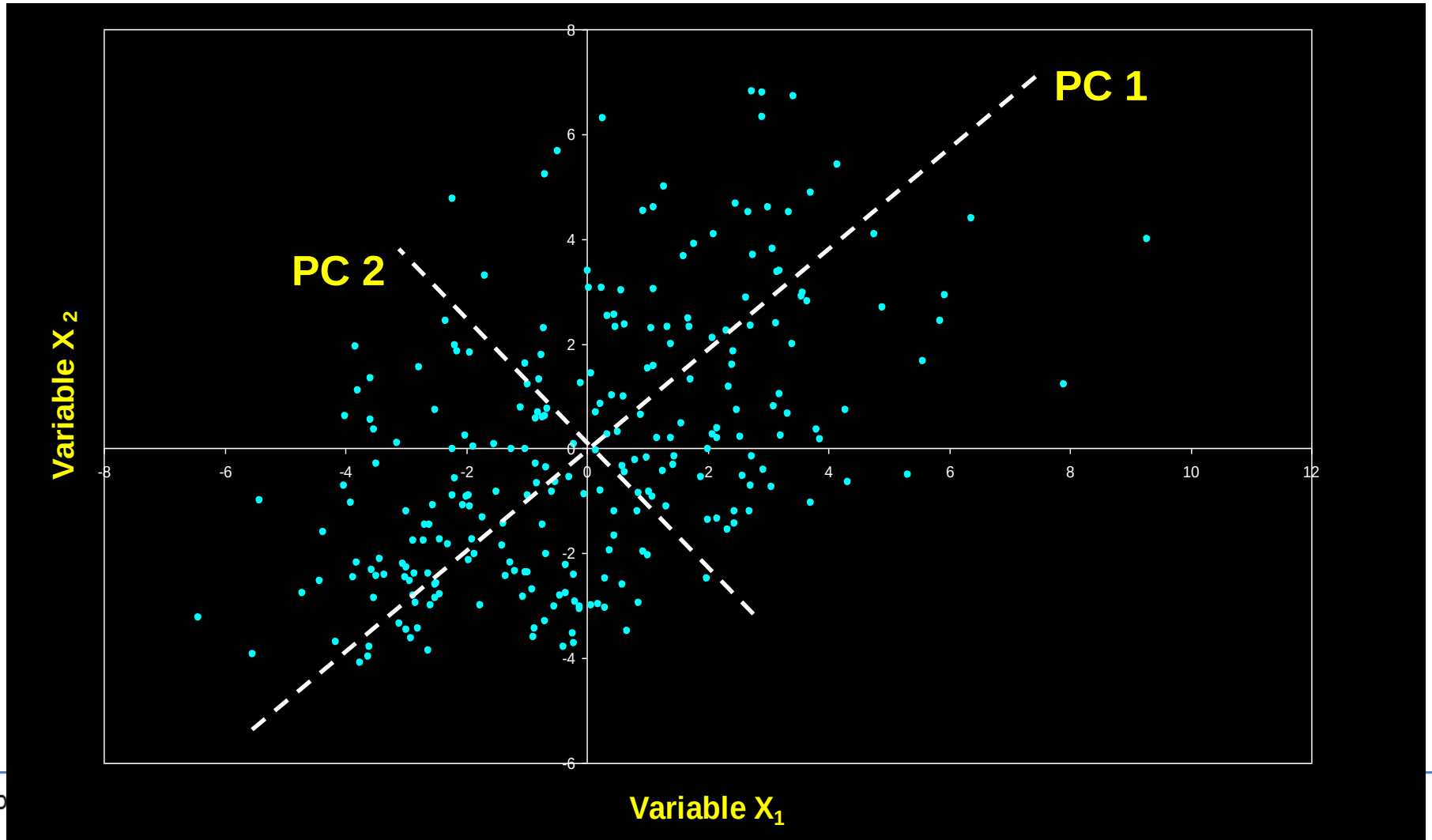


Principal Components are Computed

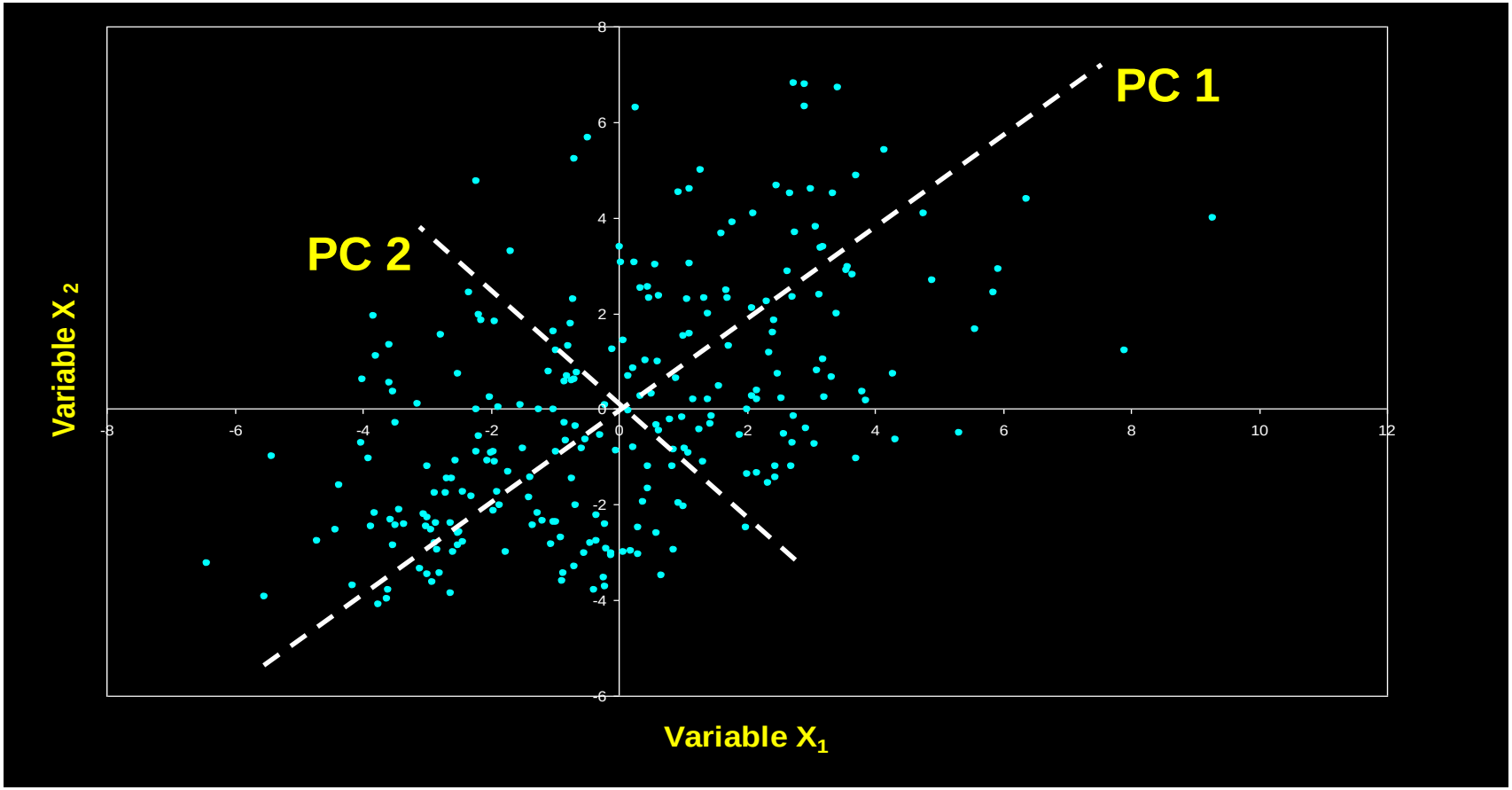
- PC 1 has the highest possible variance (9.88)
- PC 2 has a variance of 3.03
- PC 1 and PC 2 have zero covariance.



- each principal axis is a linear combination of the original two variables
- $PC_j = a_{i1}Y_1 + a_{i2}Y_2 + \ldots a_{in}Y_n$
- a_{ij}'s are the coefficients for factor i, multiplied by the measured value for variable j



- PC axes are a rigid rotation of the original variables
- PC 1 is simultaneously the direction of maximum variance and a least-squares "line of best fit" (squared distances of points away from PC 1 are minimized).



Example for PCA

Data point	X_1	X_2	X_3
1.	1	2	3
2.	2	3	4
3.	3	1	1
4.	4	6	7
5.	5	5	5
6.	6	4	2
7.	7	8	9
8.	8	9	8
9.	9	7	6

$$\Sigma X_1 = \Sigma X_2 = \Sigma X_3 = 45$$

$$\Sigma X_1^2 = \Sigma X_2^2 = \Sigma X_3^2 = 285 \qquad N = 9$$

$$\Sigma X_1 X_2 = 275;\ \Sigma X_1 X_3 = 260;\ \Sigma X_2 X_3 = 280$$

Table 2. Scores on three hypothetical variables, X_1, X_2 and X_3.

then a 3 x 3 matrix of correlation coefficients can be calculated.

	X_1	X_2	X_3
X_1	1.0000	0.8333	0.5833
X_2	0.8333	1.0000	0.9167
X_3	0.5833	0.9167	1.0000

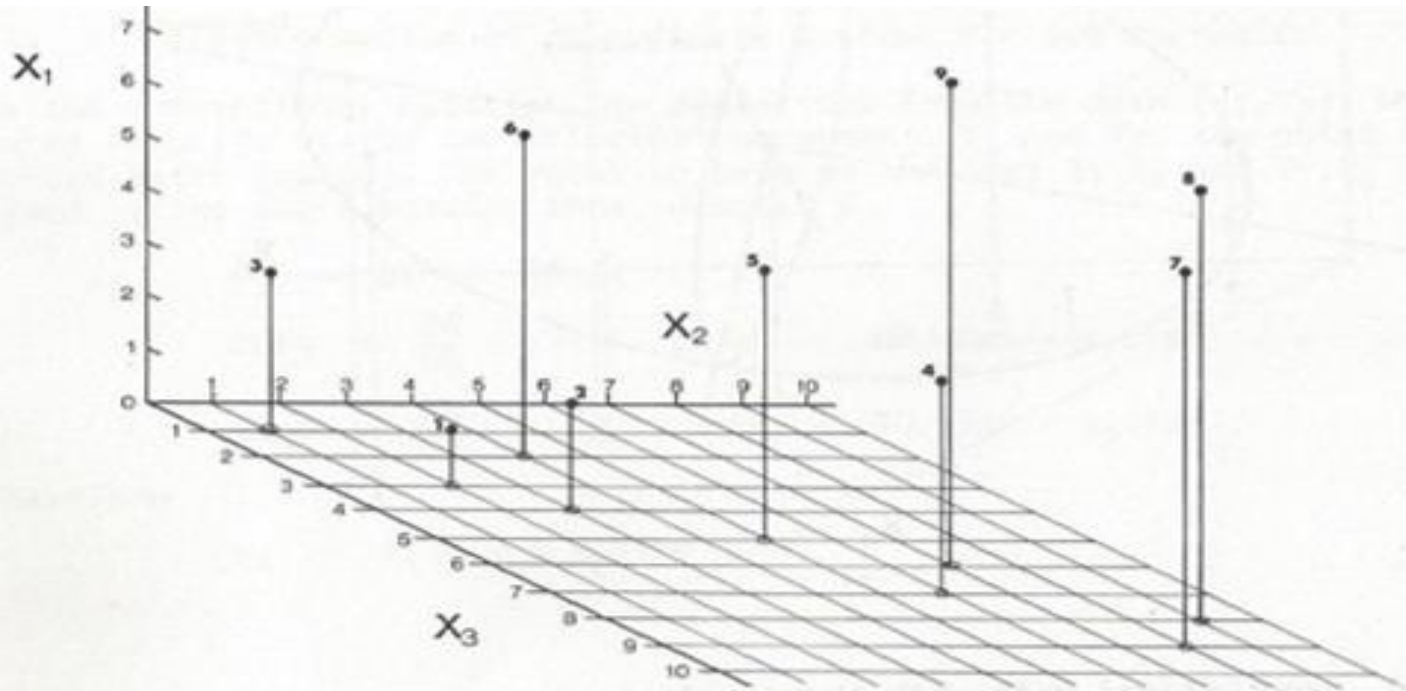


Fig. 8 Plot of hypothetical data for three variables

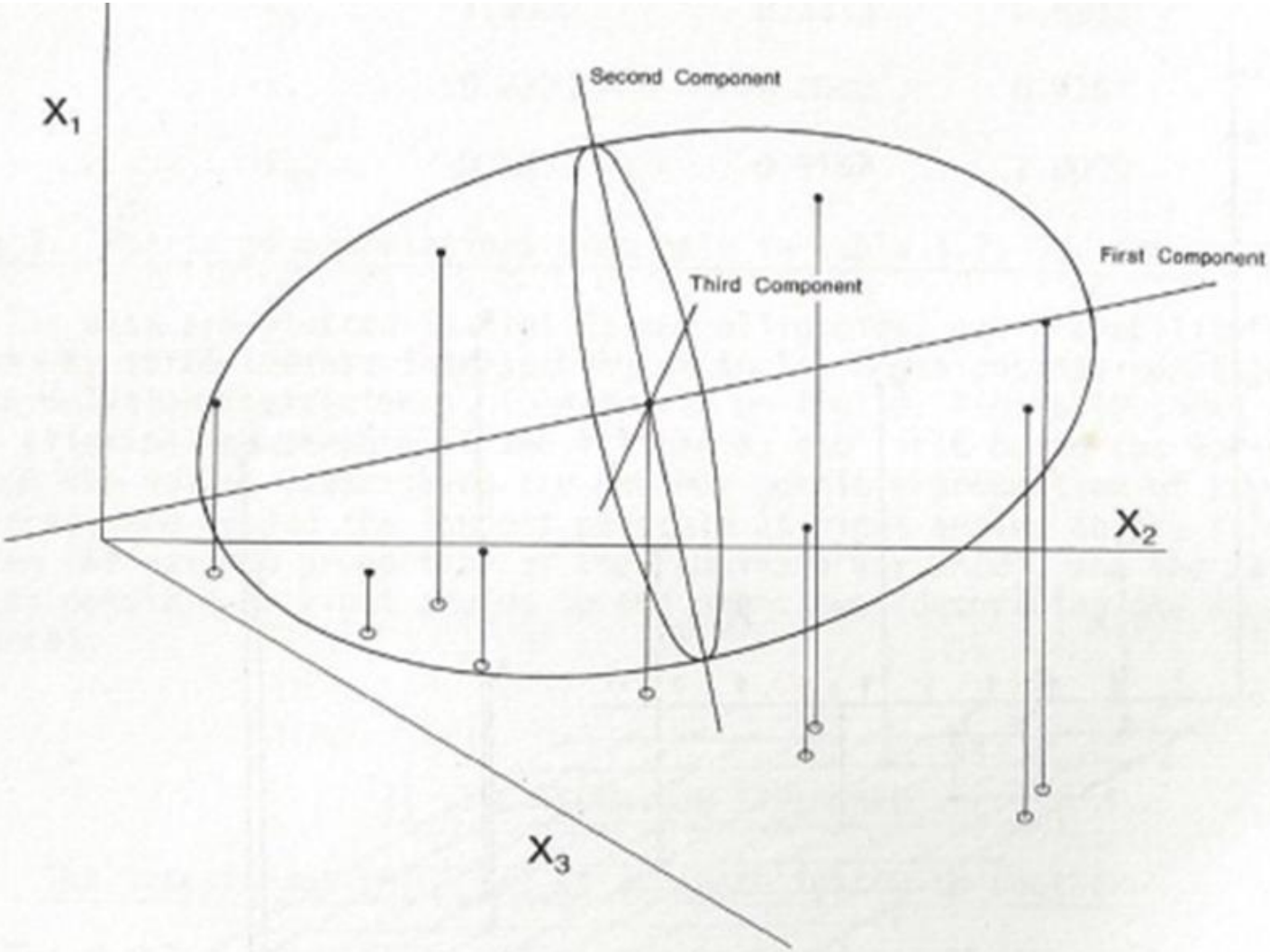
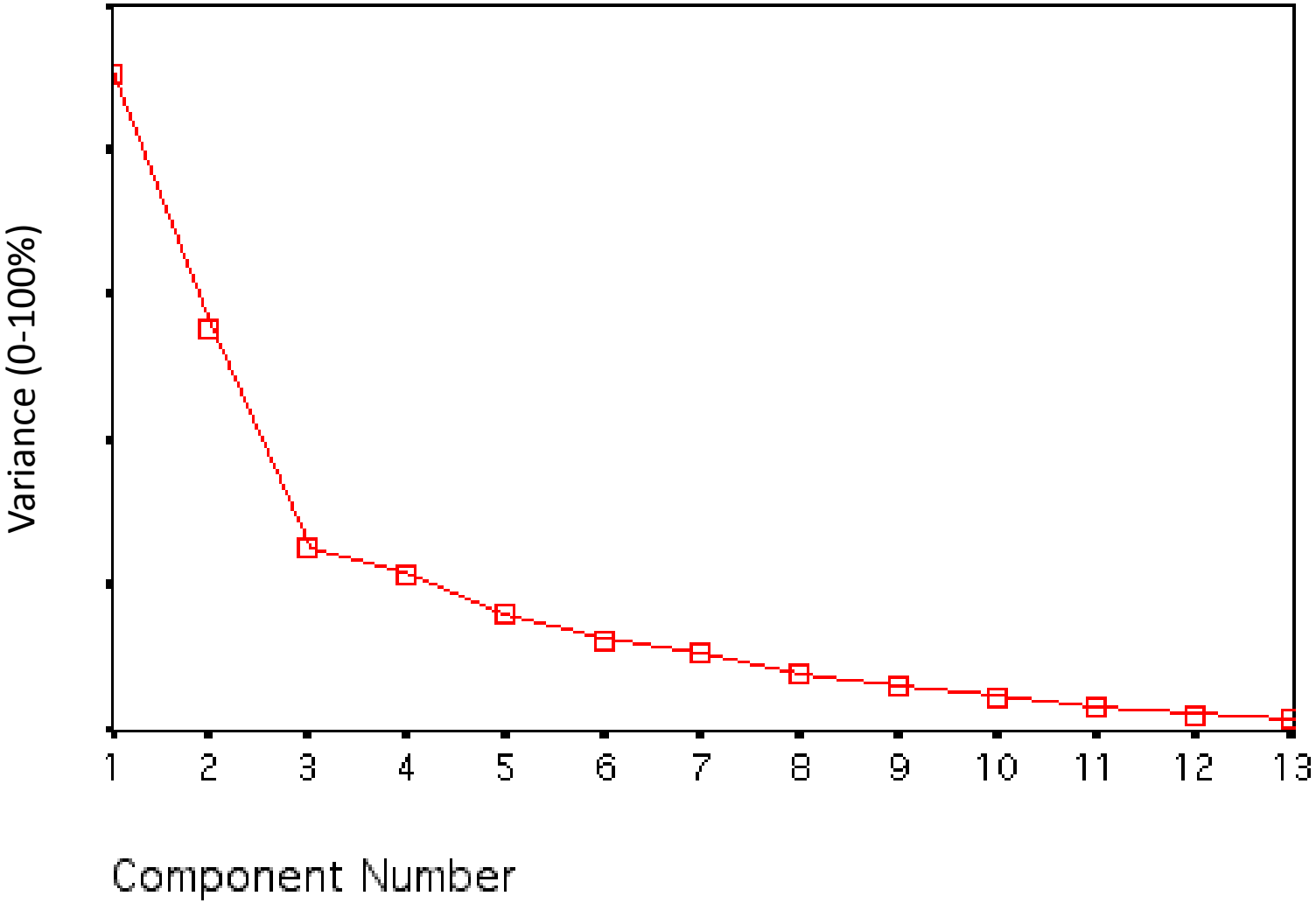


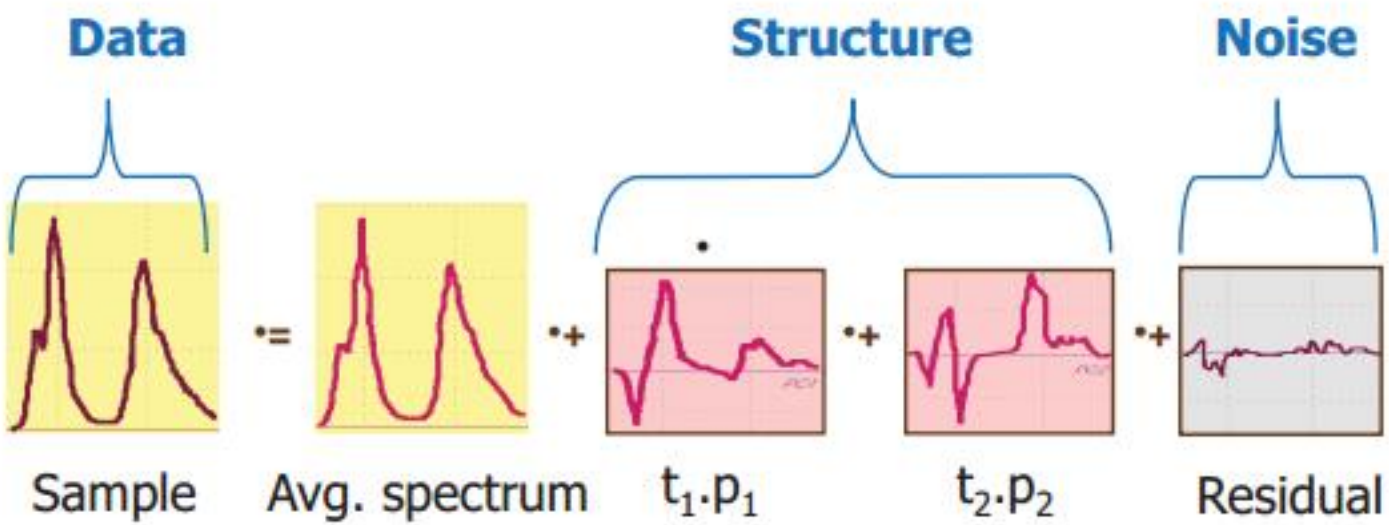
Fig. 10 Hypothetical data with three principal components and equiprobability surface

Eigenvalue



Scores and Loadings

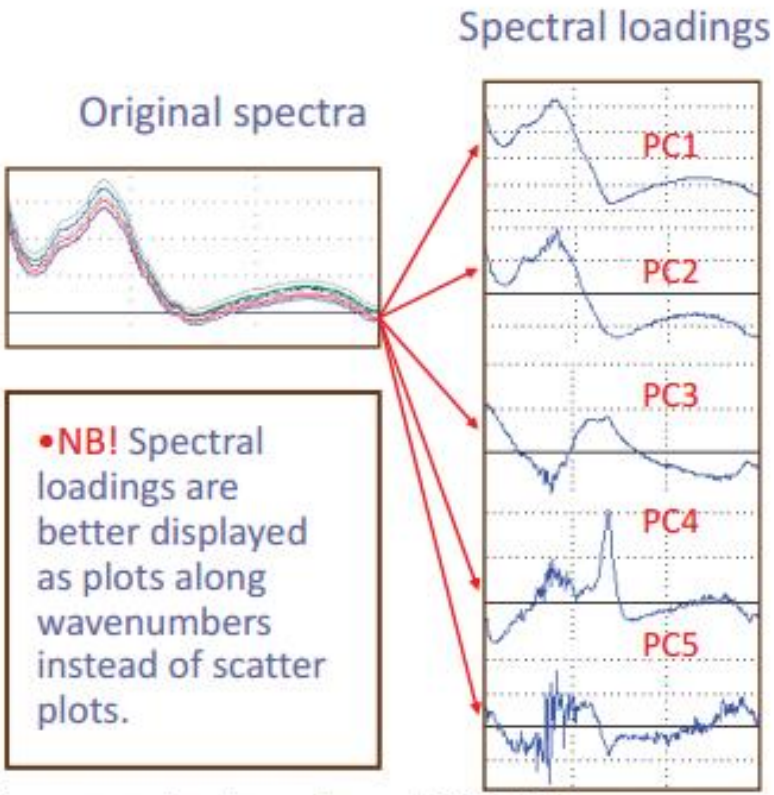
Each sample is represented as a linear combination of the principal components



t_1, t_2 = score vectors

p_1, p_2 = loading vectors

- **Loadings contain spectral features of the data**
- **Peaks in the loading plots indicate important spectral regions for the particular component.**
- **Individual loading vectors do not necessarily correspond to physical constituents.**



•NB! Spectral loadings are better displayed as plots along wavenumbers instead of scatter plots.

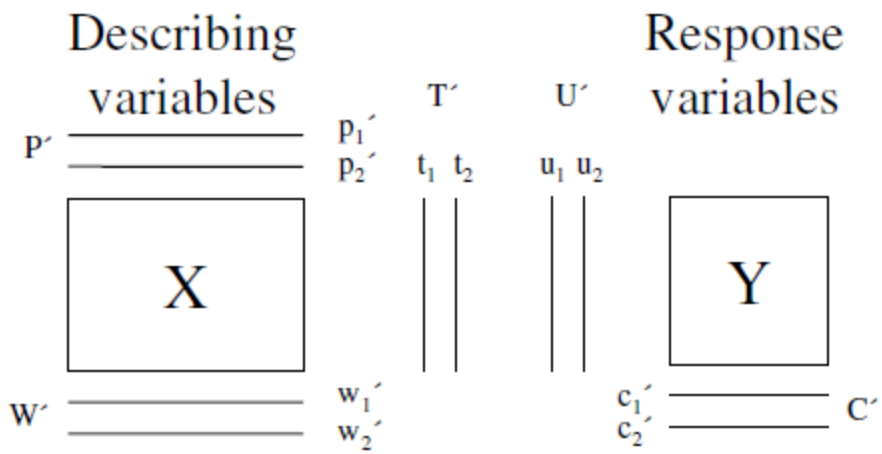
Source: NIR Transmittance spectra of pharmaceutical samples – 700-2500 nm

Inter corelation between variables helps to better correlates: PLSR 1

(i) "CORN"	1.00														
(ii) "ROOT"	0.70	1.00													
(iii) "PASTURE"	-0.71	-0.41	1.00												
(iv) "HAY"	-0.08	-0.15	0.11	1.00											
(v) "OTHER"	0.10	0.04	-0.37	-0.46	1.00										
(vi) "WHEAT"	0.27	0.05	-0.32	0.21	0.15	1.00									
(vii) "OATS"	-0.80	-0.64	0.55	-0.02	0.09	-0.40	1.00								
(viii) "MILCH"	0.00	0.06	0.12	0.45	-0.66	0.31	-0.29	1.00							
(ix) "P-MILCH"	0.60	0.38	-0.54	-0.14	-0.02	0.28	-0.53	0.34	1.00						
(x) "SHEEP"	0.43	0.25	-0.24	-0.00	-0.16	0.49	-0.41	0.47	0.42	1.00					
(xi) "1-10 HO"	0.66	0.46	-0.38	0.03	-0.16	0.27	-0.73	0.17	0.37	0.21	1.00				
(xii) "50-100"	-0.64	-0.67	0.40	-0.13	0.17	-0.09	0.80	-0.08	-0.29	-0.04	-0.78	1.00			
(xiii) "COMBINE"	0.42	0.36	-0.46	0.12	0.00	0.54	-0.77	0.34	0.25	0.41	0.51	-0.61	1.00		
(xiv) "MILKMAC"	0.41	0.18	-0.51	0.28	-0.08	0.38	-0.58	0.01	0.09	0.00	0.49	-0.61	0.60	1.00	
(xv) "MALES"	-0.15	0.07	0.20	0.41	-0.84	-0.12	-0.01	0.70	0.00	0.10	0.07	-0.07	0.01	0.02	1.00
	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)	(ix)	(x)	(xi)	(xii)	(xiii)	(xiv)	(xv)

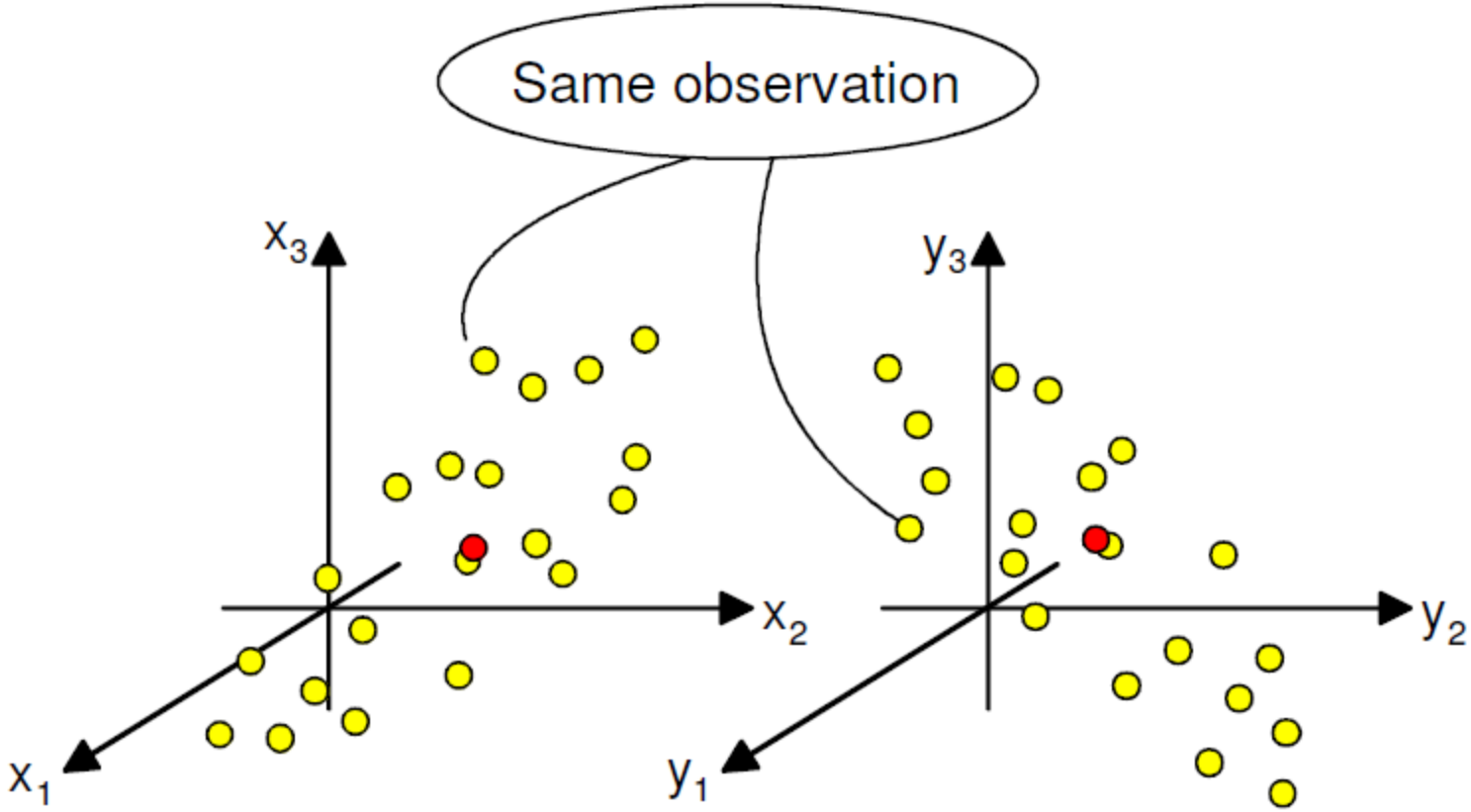
Table 7. Lower triangle of correlation matrix for 15 variables, first analysis.

PLS



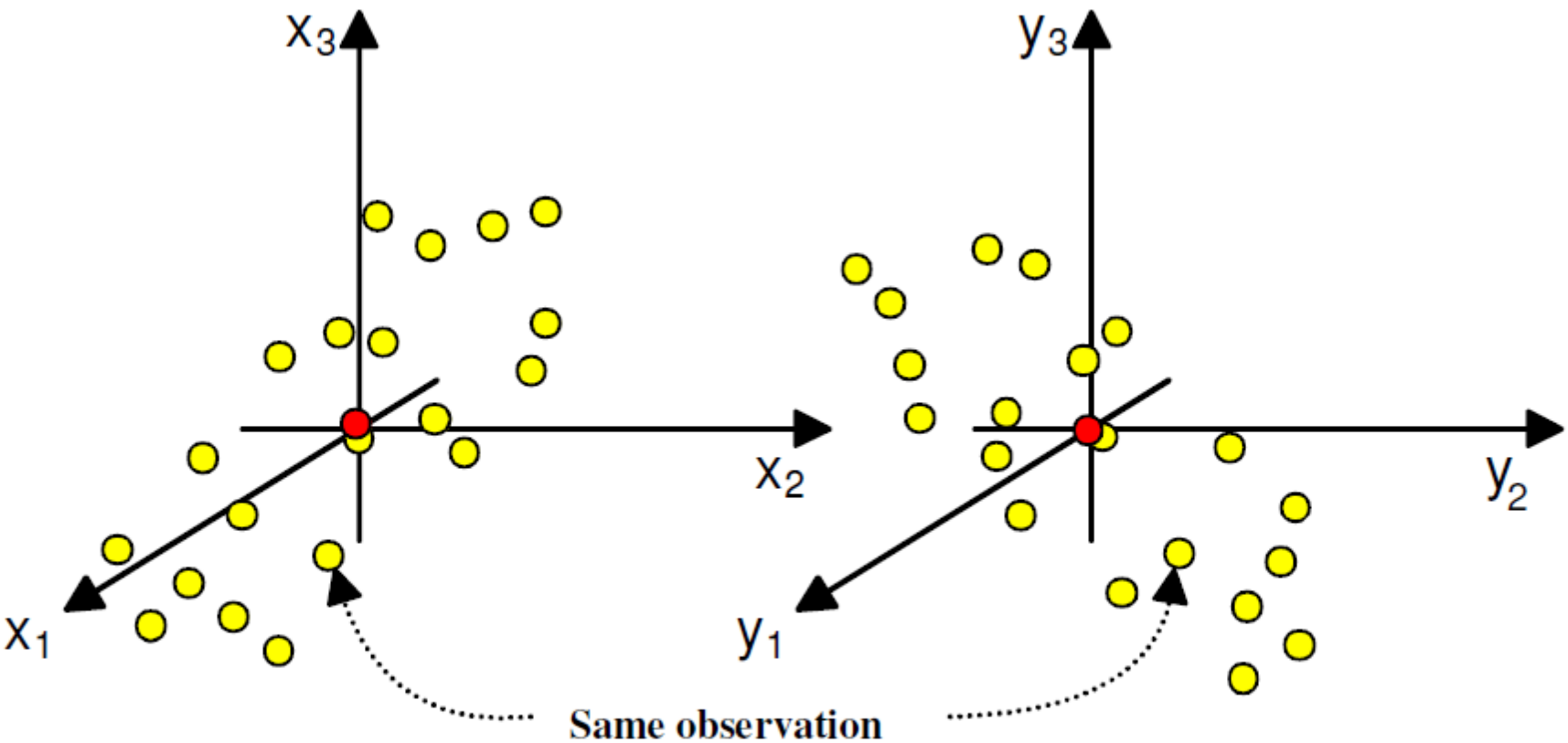
- Can handle many noisy collinear variables (compare with MLR)
- Tolerate moderate amounts of missing data (X and Y)
- Multiple responses modelled at the same time
- The result can be graphically visualized i.e. score plots and

PLS - Geometric Interpretation



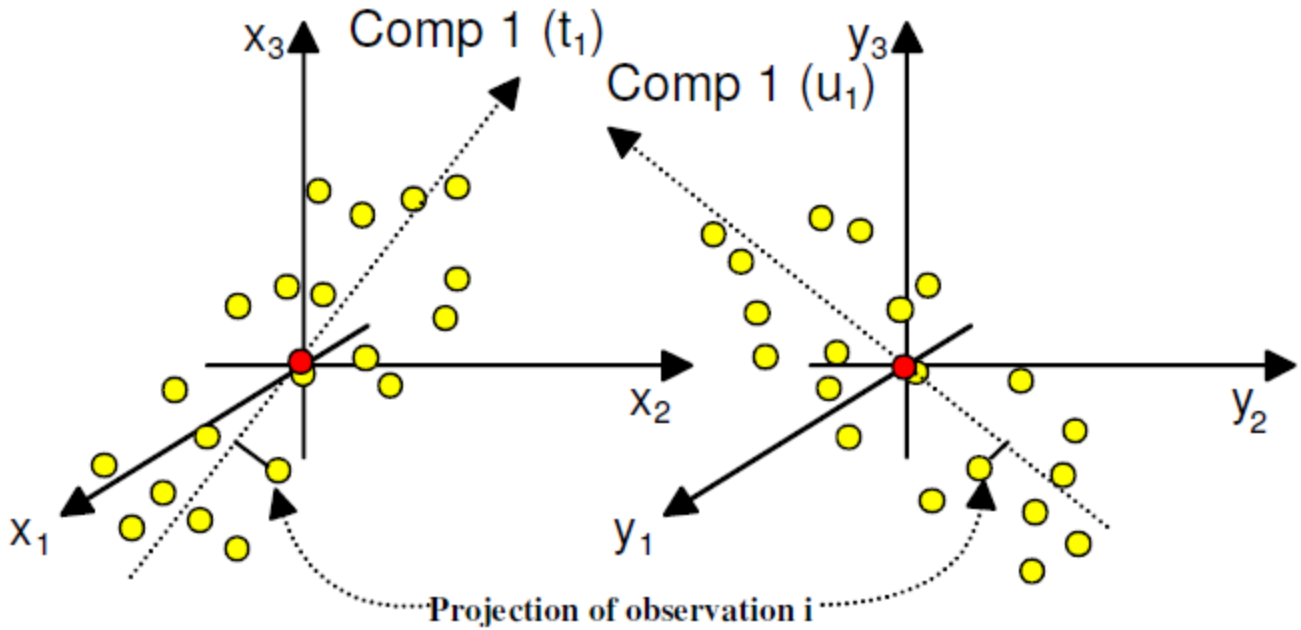
- Each observation is represented by one point in the X-space and one in the Y-space
- As in PCA, the initial step is to calculate and subtract the averages; this corresponds to moving the coordinate systems

PLS - Geometric Interpretation



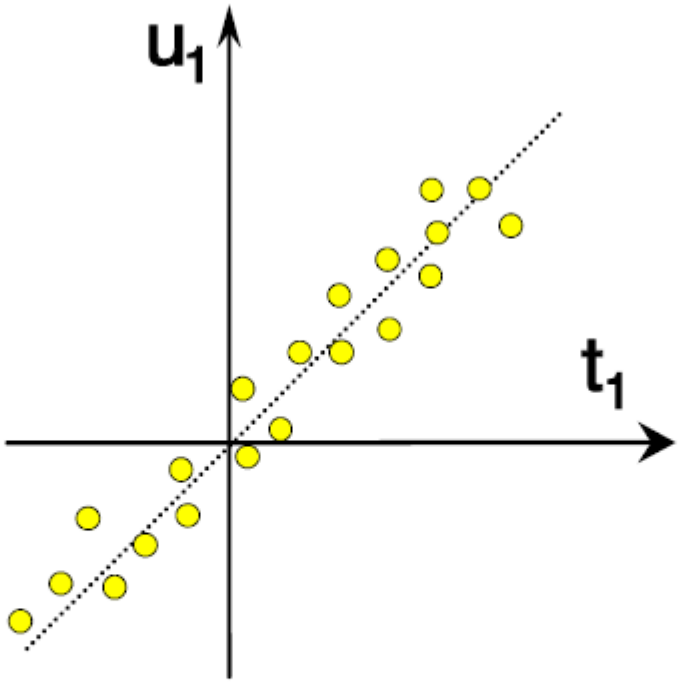
- The mean-centering procedure implies that the origin of each coordinate system is re-positioned

PLS - Geometric Interpretation



- The first PLS-component is a line in the X-space and a line in the Y-space, calculated to
  a) approximate the point-swarms well in X and Y
  b) provide a good correlation between the projections (t_1 and u_1)
- Directions are w_1 and c_1 and co-ordinates along these vectors

PLS - Geometric Interpretation



- The projection coordinates, t_1 and u_1, in the two spaces, X and Y, are connected and correlated through the **inner relation**
  $u_{i1} = t_{i1} + h_i$ (h_i is a residual)

PLSR2

Taking into account ALL variables (and not one variable) and wavelengths

		variables			Wavelengths							
		H (CaCl2	Ntot %	Cox %	350	351	352	353	354	355	356	357
		1	2	3	4	5	6	7	8	9	10	11
1-H2	1	3.4200	0.1860	3.5500	7.5812e–02	7.8660e–02	7.5939e–02	7.2183e–02	7.3291e–02	7.2857e–02	7.4180e–02	7.8963
2-H2	2	3.1600	0.3050	5.9200	6.7764e–02	6.8742e–02	6.5941e–02	6.3077e–02	6.4507e–02	6.6654e–02	6.8508e–02	6.9362
3-H2	3	3.5200	0.2170	3.9000	8.4978e–02	7.7825e–02	7.5078e–02	7.6869e–02	7.7427e–02	8.0258e–02	8.2281e–02	8.1242
4-H2	4	3.4000	0.2420	4.2800	6.3061e–02	5.9753e–02	5.7598e–02	5.8526e–02	5.9966e–02	6.2146e–02	6.2425e–02	5.9424
5-H2	5	3.3300	0.2600	4.8200	5.5647e–02	5.7697e–02	5.8126e–02	5.8575e–02	6.1101e–02	6.0767e–02	5.9853e–02	6.0564
6-H2	6	3.4100	0.2390	3.9700	7.7118e–02	8.1708e–02	8.4066e–02	8.3470e–02	8.1504e–02	7.6186e–02	7.2653e–02	7.4251
7-H2	7	3.3000	0.2400	4.1300	9.0307e–02	8.8831e–02	9.2090e–02	9.6331e–02	9.4941e–02	9.1973e–02	8.7676e–02	8.2700
8-H2	8	3.2700	0.2640	4.8200	7.8177e–02	7.7864e–02	7.9271e–02	8.3268e–02	8.8739e–02	8.3659e–02	7.6078e–02	7.3768
9-H2	9	3.4400	0.1700	3.1600	9.0659e–02	9.2013e–02	9.1548e–02	9.1889e–02	9.6653e–02	9.0379e–02	8.2906e–02	8.3167
10-H2	10	3.5400	0.1960	2.7800	8.2668e–02	8.1480e–02	8.3849e–02	8.8166e–02	9.0274e–02	8.9105e–02	8.5262e–02	8.0365
11-H2	11	3.5500	0.1530	2.7400	0.1054	0.1069	0.1022	9.9786e–02	0.1090	0.1040	9.7862e–02	0.
12-H2	12	3.3900	0.1150	1.8900	9.5994e–02	9.4586e–02	9.0226e–02	8.9133e–02	9.7811e–02	9.6249e–02	9.2702e–02	9.5169
13-H2	13	3.2800	0.1260	2.0100	0.1139	0.1145	0.1098	0.1076	0.1174	0.1131	0.1079	0
14-H2	14	3.5400	0.1090	1.7000	0.1359	0.1314	0.1301	0.1331	0.1389	0.1341	0.1300	0.
15-H2	15	3.0900	0.1380	2.3900	5.2701e–02	5.7808e–02	5.4761e–02	5.1519e–02	6.1359e–02	5.9657e–02	5.5406e–02	5.7463
16-H2	16	3.3700	9.2000e–02	1.3900	0.1191	0.1056	0.1070	0.1157	0.1148	0.1143	0.1125	0.
17-H2	17	3.3000	0.1460	2.3900	0.1227	0.1126	0.1134	0.1195	0.1192	0.1227	0.1240	0
18-H2	18	3.3600	0.1350	2.1600	8.6688e–02	7.9889e–02	8.0846e–02	8.5568e–02	8.7059e–02	9.0212e–02	9.1081e–02	8.7271
19-H2	19	3.4700	0.1580	2.9700	0.1053	0.1028	0.1056	0.1091	0.1072	0.1088	0.1089	0.
20-H2	20	3.3100	0.1460	2.2800	0.1255	0.1135	0.1130	0.1192	0.1194	0.1200	0.1199	0
21-H2	21	3.2000	0.1960	2.9300	9.4911e–02	9.7565e–02	9.9859e–02	9.9101e–02	9.5255e–02	9.2029e–02	9.2249e–02	9.6505
22-H2	22	3.4500	0.1270	2.0800	0.1057	0.1010	0.1008	0.1043	0.1072	0.1023	9.8925e–02	0.
23-H2	23	3.4200	0.1850	3.1600	8.3049e–02	8.1640e–02	8.0518e–02	8.0717e–02	8.3455e–02	8.0279e–02	7.7703e–02	8.1053
24-H2	24	3.4300	0.2120	3.2800	9.9271e–02	9.9942e–02	0.1005	0.1007	9.8387e–02	9.4745e–02	9.3870e–02	9.7490
25-H2	25	3.5100	0.1480	2.0800	9.9715e–02	0.1004	0.1008	0.1011	0.1009	9.3165e–02	8.8576e–02	9.4614

PLSR

Home made (Matlab) ⟵ ***Software*** ⟶ *Commercial (The Unscrambler)*

The Unscrambler - [Analisi dati scopolamina]

File Edit View Plot Modify Task Results Window Help

		SCO%	LID%	FEN%	MAN%	903	905.25	907.5	909.75	912	914.25	916.5	918.75	921	923.25
		1'	2'	3'	4'	5'	6'	7'	8'	9'	10'	11'	12'	13'	14'
21	1	100.0000	0.0000	0.0000	0.0000	−0.1876	−0.4791	−0.5889	−0.5959	−0.6136	−0.6254	−0.6313	−0.6384	−0.6396	−0.6408
10	2	14.6538	34.5397	19.8859	30.9205	−0.8358	−0.9011	−0.9229	−0.9208	−0.9208	−0.9242	−0.9290	−0.9371	−0.9419	−0.9426
16	3	44.9270	0.0000	55.0730	0.0000	−0.3990	−0.7076	−0.8180	−0.8343	−0.8568	−0.8701	−0.8864	−0.8834	−0.8793	−0.8783
8	4	0.0000	49.8008	50.1992	0.0000	−0.8496	−0.9369	−0.9669	−0.9712	−0.9841	−0.9970	−1.0156	−1.0299	−1.0385	−1.0456
6	5	0.0000	45.0100	54.9900	0.0000	−0.7649	−0.8985	−0.9388	−0.9388	−0.9577	−0.9779	−0.9943	−1.0069	−1.0157	−1.0220
9	6	0.0000	100.0000	0.0000	0.0000	−0.8979	−0.9311	−0.9387	−0.9296	−0.9356	−0.9447	−0.9669	−0.9859	−1.0012	−1.0137
4	7	0.0000	0.0000	100.0000	0.0000	−0.5356	−0.7903	−0.8932	−0.9214	−0.9312	−0.9471	−0.9495	−0.9495	−0.9544	−0.9532
14	8	35.0569	0.0000	34.9970	29.9461	−0.6692	−0.8318	−0.8959	−0.9004	−0.9038	−0.9078	−0.9129	−0.9135	−0.9146	−0.9174
13	9	30.0020	35.0190	34.9790	0.0000	−0.7127	−0.8619	−0.9154	−0.9254	−0.9376	−0.9521	−0.9688	−0.9733	−0.9755	−0.9833
1	10	0.0000	0.0000	0.0000	100.0000	−0.6667	−0.7814	−0.8209	−0.8151	−0.8168	−0.8192	−0.8188	−0.8213	−0.8213	−0.8250
3	11	0.0000	0.0000	50.0399	49.9601	−0.6956	−0.8208	−0.8715	−0.8772	−0.8811	−0.8891	−0.8885	−0.8908	−0.8925	−0.8897
7	12	0.0000	50.0400	0.0000	49.9600	−0.8831	−0.9131	−0.9215	−0.9183	−0.9183	−0.9229	−0.9294	−0.9333	−0.9391	−0.9450
2	13	0.0000	0.0000	44.9550	55.0450	−0.7368	−0.8359	−0.8742	−0.8762	−0.8816	−0.8870	−0.8880	−0.8924	−0.8929	−0.8939
19	14	50.1196	0.0000	49.8804	0.0000	−0.5617	−0.7865	−0.8823	−0.9058	−0.9380	−0.9556	−0.9605	−0.9605	−0.9566	−0.9556
11	15	29.9581	14.9391	30.1778	24.9251	−0.7655	−0.8671	−0.9135	−0.9210	−0.9326	−0.9428	−0.9469	−0.9490	−0.9510	−0.9551
12	16	30.0938	34.9830	0.0000	34.9232	−0.8212	−0.8924	−0.9161	−0.9154	−0.9154	−0.9210	−0.9329	−0.9398	−0.9468	−0.9454
17	17	44.9640	55.0360	0.0000	0.0000	−0.9439	−0.9793	−0.9915	−0.9842	−0.9854	−0.9964	−1.0171	−1.0269	−1.0367	−1.0367
15	18	35.0120	30.0160	0.0000	34.9720	−0.8285	−0.8688	−0.8810	−0.8791	−0.8803	−0.8880	−0.8970	−0.9008	−0.9034	−0.9053
18	19	49.9301	0.0000	0.0000	50.0699	−0.6593	−0.7635	−0.8022	−0.8064	−0.8095	−0.8111	−0.8163	−0.8174	−0.8216	−0.8289
20	20	50.8717	49.1283	0.0000	0.0000	−0.9850	−1.0041	−1.0140	−1.0155	−1.0235	−1.0332	−1.0420	−1.0522	−1.0623	−1.0735
5	21	0.0000	30.0000	34.9600	35.0400	−0.8145	−0.8529	−0.8664	−0.8679	−0.8732	−0.8784	−0.8860	−0.8920	−0.8958	−0.9025
37	22	0.0000	40.1118	39.9321	19.9561	−0.9902	−0.9890	−0.9871	−0.9886	−0.9961	−1.0008	−1.0069	−1.0175	−1.0281	−1.0398

For Help, press F1 | Value: 100.0000 | Size: 53 x 360 | R/W | GU

Dependent variables

3 o 4 concentration values

Independent variables

256 absorbance values

Original variables

↓

Principal Components (PC)

$$X \longrightarrow u_m$$

$$\downarrow$$

$$Y \rightleftarrows t_m$$

Building a Calibration Model

Calibration Samples

- How many depends on the samples and the model; more is typically better; can be ~1000-1000
- Calibration samples should reflect the composition and variance expected in test samples

Chemical Characterization

- Controls precision and accuracy of calibration model

Multivariate Analysis Tools

- Translates spectroscopic data into compositional data

QA/QC

- Calibration checks (well characterized "blind" samples or standard reference materials)
- Outlier flag(s)
- Measure(s) of uncertainty

Model types

- Fundamental models
  (hard models, “global models”)

$$E = mc^2 \qquad y = y_0 e^{-kt} \qquad U = IR$$

- Empirical models
  (soft models, “local models”)
  Taylor expansions (polynomials of different complexity)

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_{12} x_{12} + e$$

Rapid Sper-A model Building & Use

Wet Chemistry

+

NIR Spectra

+

Chemometrics

=

Calibration Model

Calibration Model

+

More NIR Spectra

=

Rapid Predictions

Statistics for QA/QI in NIRS

$$RMSEP = \sqrt{\frac{\sum (Ypred - Yref)^2}{n}}$$

$$Bias = \overline{Y}pred - \overline{Y}ref$$

$$SEE = \sqrt{\frac{n}{n-1}(RMSEP^2 - Bias^2)}$$

$$RPD = \frac{STD_{ref}}{SE_{pred}}$$

$$PA = \frac{Yref(Max)-Yref\ (Min)}{SEP}$$

$$SE(chem) = \frac{\sum [SD(i)]}{n-1}$$

$$GAM = \frac{SE(chem)}{SEP}$$

$0 < GAM < 1$

GAM (General Accuracy Measure)

$$SEC = \sqrt{\frac{\sum (Ypred-Yref)^2}{n}}$$

$$RMSEP = \sqrt{\frac{\sum (Ypred-Yref)^2}{n-1-p}}$$

$$SEP = \sqrt{\frac{\sum (Ypred-Yref)^2 - Bias^2}{n-1}}$$

SEC > RMSEP

RPD > 1.6

$R^2c > R^2p$

Sper -Analysis Basic Rules

2 (3) groups

- Calibration
- Validation

Y(max-min) *cal* = Y(max-min) *val*

- Examination (test)

Two ways:

1) Running a model on the Cal set → applying the model on the Val set
2) Running a cross calibration modeling on all samples → applying the model on examination set

Analysis can be done by

Using existing statistics software : MATHLAB, SPSS, SAS

Specific software dedicated for Sper – A

- Unscrambler
- Paracuda

The Unscrambler®
A Handy Tool for Doing Chemometrics

Steps to use Unscrabmer

Excel

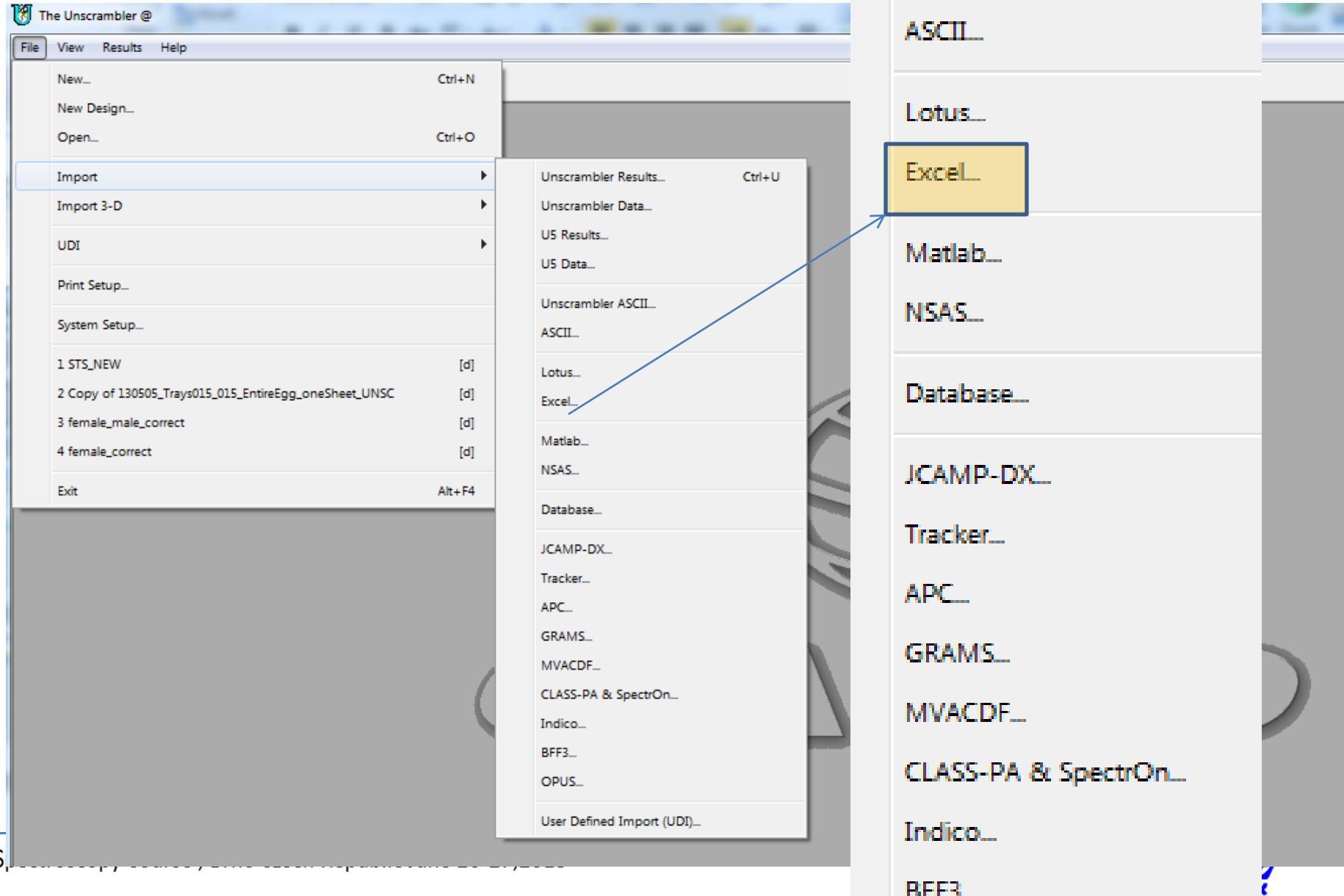
Copy of Loess Diesel [Compatibility Mode] - Microsoft Excel

File Home Insert Page Layout Formulas Data Review View

A2 350

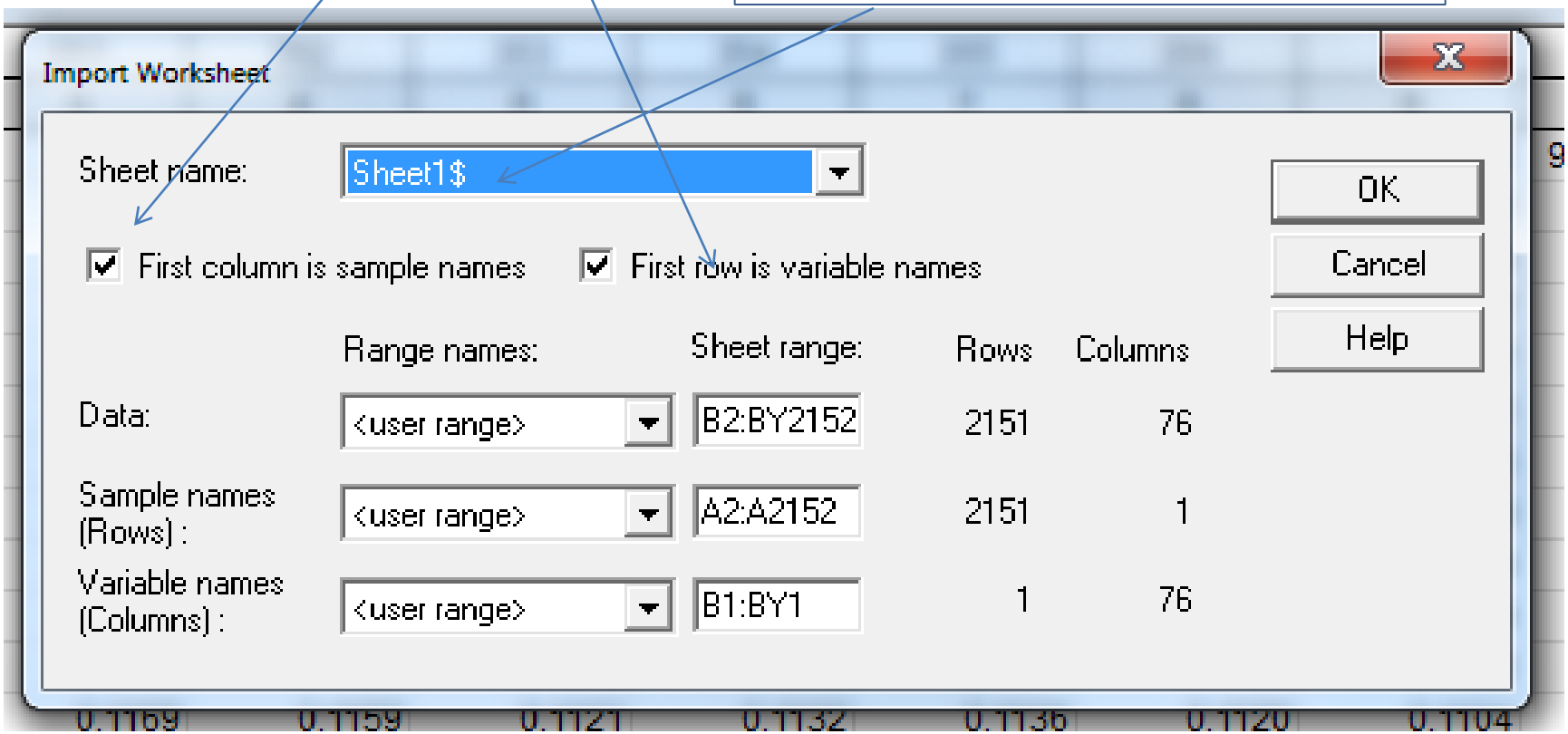
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Sample No	1	2	3	4	5	6	7	8	9	10	11	12
2	350	0.109554	0.1122109	0.1172517	0.1181262	0.1120101	0.114604	0.106207	0.1092466	0.1040154	0.1043194	0.1034673	0.117967
3	351	0.1090609	0.110128	0.1134719	0.1165864	0.1109623	0.1131335	0.1068466	0.1123147	0.1076239	0.107467	0.1065232	0.1169326
4	352	0.1062434	0.1053687	0.1069504	0.111979	0.1076976	0.1078607	0.1064636	0.1123243	0.109267	0.1094098	0.1079751	0.1158651
5	353	0.1022577	0.1013658	0.1082565	0.1094107	0.1065925	0.1058679	0.109565	0.1095998	0.1072613	0.1069507	0.107226	0.1121172
6	354	0.09951617	0.1002378	0.1078123	0.1106486	0.1075815	0.1065014	0.1071751	0.1082621	0.1049285	0.1036334	0.1040438	0.1132183
7	355	0.09980478	0.1019405	0.1094354	0.1129633	0.1099951	0.1087798	0.105702	0.108261	0.1042618	0.1030548	0.1033706	0.1136314
8	356	0.1036036	0.1065044	0.114116	0.1157863	0.113764	0.1125606	0.1066828	0.1096197	0.105774	0.1061384	0.1064434	0.1120332
9	357	0.1002391	0.1055055	0.1078948	0.1096817	0.1087413	0.1071391	0.1025995	0.102878	0.101429	0.1017569	0.1022199	0.110383
10	358	0.09901481	0.1037362	0.1055611	0.1075457	0.1065086	0.1045678	0.1059169	0.106537	0.1051087	0.1054841	0.105332	0.1098302
11	359	0.09996973	0.1022202	0.1062616	0.1082212	0.1063894	0.1043538	0.1125609	0.1152939	0.1124308	0.1127907	0.1117473	0.1099293
12	360	0.104174	0.1054184	0.107048	0.10737	0.105803	0.1048919	0.1060905	0.1081196	0.1057354	0.1051464	0.1049735	0.1090584
13	361	0.1055276	0.1063933	0.1065437	0.1067644	0.1065517	0.1043501	0.1047141	0.1067865	0.1038305	0.1027005	0.1040925	0.1082377
14	362	0.1053406	0.1059467	0.1052187	0.1062122	0.1067414	0.1035937	0.105385	0.1079217	0.1041345	0.102895	0.1047201	0.1085895
15	363	0.1041666	0.1041894	0.1031324	0.1055369	0.1042887	0.1033498	0.1060504	0.1093526	0.1051076	0.1043987	0.1032638	0.1117261
16	364	0.1045722	0.1061242	0.104199	0.1067486	0.1055908	0.1047046	0.1062513	0.1084737	0.1044657	0.1047438	0.1034429	0.1101902
17	365	0.103628	0.1058873	0.1063197	0.1086868	0.1073913	0.1062009	0.1058967	0.1073878	0.1041182	0.1052316	0.1040976	0.1086543
18	366	0.1005694	0.1020208	0.1091167	0.1111554	0.1090194	0.1075351	0.1049231	0.106535	0.1045047	0.1061615	0.1050123	0.1081324
19	367	0.1036887	0.1044917	0.1073315	0.1101712	0.10684	0.1073166	0.103958	0.1047504	0.10281	0.1039172	0.1036436	0.1063277
20	368	0.1043254	0.1047097	0.1070963	0.1093705	0.1059665	0.1068001	0.1034557	0.1048269	0.1028736	0.1023593	0.102392	0.1067168
21	369	0.1035261	0.1039329	0.1078113	0.108732	0.1061183	0.1063475	0.1034656	0.1058774	0.1038614	0.1016207	0.1015275	0.108526
22	370	0.1047737	0.1068714	0.1075083	0.108235	0.1067264	0.1075412	0.1045362	0.1047622	0.1028275	0.1027881	0.1024373	0.109433
23	371	0.1046701	0.1055754	0.1071625	0.1083571	0.1071198	0.107068	0.1068236	0.1062215	0.1038109	0.1046044	0.1052067	0.1079477
24	372	0.1038639	0.1040094	0.1070068	0.1084542	0.1071388	0.1061206	0.1082295	0.1081087	0.1050505	0.1060904	0.1073137	0.1069357

Wavelength


The Unscrambler @
File
View
Results
Help
New... Ctrl+N
New Design...
Open... Ctrl+O
Import
Import 3-D
UDI
Print Setup...
System Setup...
1 STS_NEW [d]
2 Copy of 130505_Trays015_015_EntireEgg_oneSheet_UNSC [d]
3 female_male_correct [d]
4 female_correct [d]
Exit Alt+F4
Unscrambler Results... Ctrl+U
Unscrambler Data...
U5 Results...
U5 Data...
Unscrambler ASCII...
ASCII...
Lotus...
Excel...
Matlab...
NSAS...
Database...
JCAMP-DX...
Tracker...
APC...
GRAMS...
MVACDF...
CLASS-PA & SpectrOn...
Indico...
BFF3...
OPUS...
User Defined Import (UDI)...
U5 Data...
Unscrambler ASCII...
ASCII...
Lotus...
Excel...
Matlab...
NSAS...
Database...
JCAMP-DX...
Tracker...
APC...
GRAMS...
MVACDF...
CLASS-PA & SpectrOn...
Indico...
BFF3

The data not for calculation : Wavelengths, sample number (if exists in Excel)

Select the correct sheet in the Excel open file!!!



The Excel data in Unscrambler environment

Copy of Loess Diesel

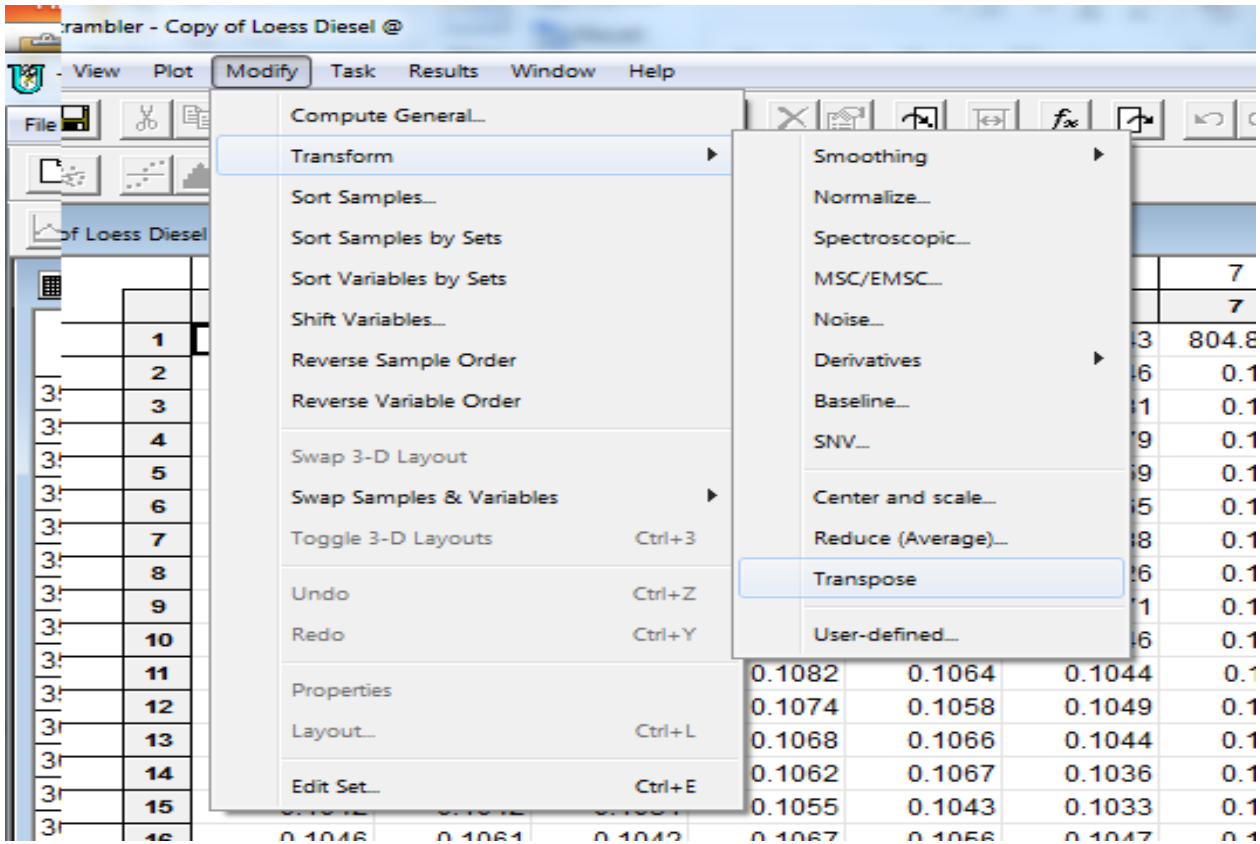
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
350	**1**	0.1096	0.1122	0.1173	0.1181	0.1120	0.1146	0.1062	0.1092	0.1040	0.1043	0.1035	0.1180	0.1164	0.1103	0.1184	0.1195
351	**2**	0.1091	0.1101	0.1135	0.1166	0.1110	0.1131	0.1068	0.1123	0.1076	0.1075	0.1065	0.1169	0.1157	0.1099	0.1127	0.1131
352	**3**	0.1062	0.1054	0.1070	0.1120	0.1077	0.1079	0.1065	0.1123	0.1093	0.1094	0.1080	0.1159	0.1153	0.1105	0.1088	0.1076
353	**4**	0.1023	0.1014	0.1083	0.1094	0.1066	0.1059	0.1096	0.1096	0.1073	0.1070	0.1072	0.1121	0.1121	0.1073	0.1132	0.1112
354	**5**	9.9516e–02	0.1002	0.1078	0.1106	0.1076	0.1065	0.1072	0.1083	0.1049	0.1036	0.1040	0.1132	0.1142	0.1100	0.1148	0.1144
355	**6**	9.9805e–02	0.1019	0.1094	0.1130	0.1100	0.1088	0.1057	0.1083	0.1043	0.1031	0.1034	0.1136	0.1147	0.1113	0.1137	0.1141
356	**7**	0.1036	0.1065	0.1141	0.1158	0.1138	0.1126	0.1067	0.1096	0.1058	0.1061	0.1064	0.1120	0.1119	0.1092	0.1098	0.1095
357	**8**	0.1002	0.1055	0.1079	0.1097	0.1087	0.1071	0.1026	0.1029	0.1014	0.1018	0.1022	0.1104	0.1095	0.1048	0.1099	0.1081
358	**9**	9.9015e–02	0.1037	0.1056	0.1075	0.1065	0.1046	0.1059	0.1065	0.1051	0.1055	0.1053	0.1098	0.1087	0.1037	0.1086	0.1075
359	**10**	9.9970e–02	0.1022	0.1063	0.1082	0.1064	0.1044	0.1126	0.1153	0.1124	0.1128	0.1117	0.1099	0.1089	0.1050	0.1070	0.1077
360	**11**	0.1042	0.1054	0.1070	0.1074	0.1058	0.1049	0.1061	0.1081	0.1057	0.1051	0.1050	0.1091	0.1082	0.1064	0.1097	0.1106
361	**12**	0.1055	0.1064	0.1065	0.1068	0.1066	0.1044	0.1047	0.1068	0.1038	0.1027	0.1041	0.1082	0.1081	0.1055	0.1106	0.1105
362	**13**	0.1053	0.1059	0.1052	0.1062	0.1067	0.1036	0.1054	0.1079	0.1041	0.1029	0.1047	0.1086	0.1085	0.1052	0.1107	0.1100
363	**14**	0.1042	0.1042	0.1031	0.1055	0.1043	0.1033	0.1061	0.1094	0.1051	0.1044	0.1033	0.1117	0.1097	0.1082	0.1110	0.1115
364	**15**	0.1046	0.1061	0.1042	0.1067	0.1056	0.1047	0.1063	0.1085	0.1045	0.1047	0.1034	0.1102	0.1093	0.1061	0.1096	0.1085
365	**16**	0.1036	0.1059	0.1063	0.1087	0.1074	0.1062	0.1059	0.1074	0.1041	0.1052	0.1041	0.1087	0.1089	0.1051	0.1090	0.1075
366	**17**	0.1006	0.1020	0.1091	0.1112	0.1090	0.1075	0.1049	0.1065	0.1045	0.1062	0.1050	0.1081	0.1089	0.1068	0.1099	0.1100

The yellow column and raw are NOT for calculation!

Preparing Unscramble data for processing:

$$U = E^{T}$$

Unscrambler Excel



Unscramble Transposed

Chemi

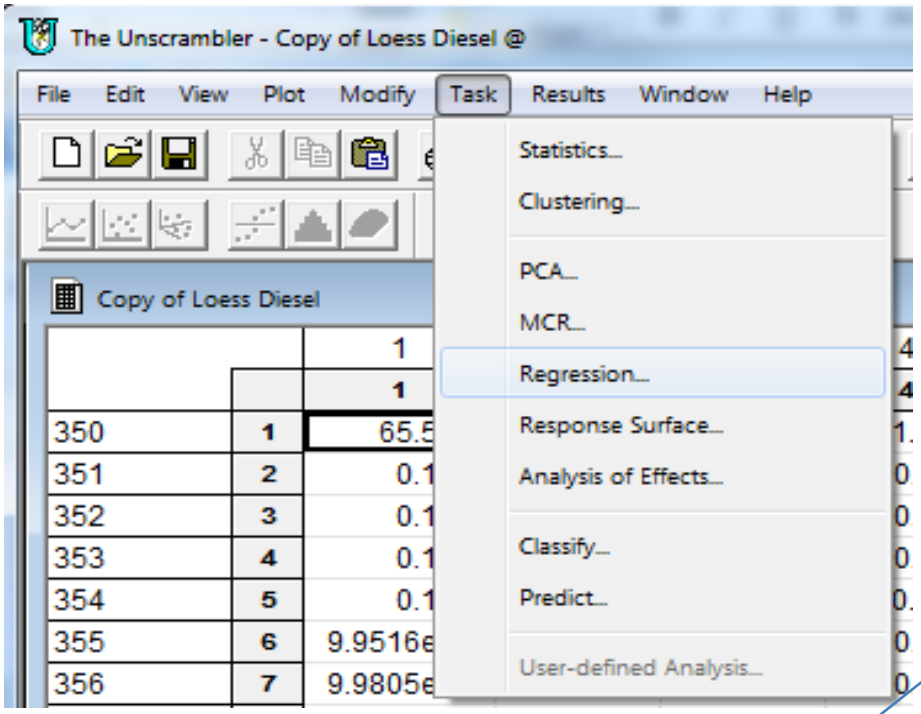
Copy of Loess Diesel

		350	351	352	353	354	355	356	357	358
		1	2	3	4	5	6	7	8	9
1	1	0.1096	0.1091	0.1062	0.1023	9.9516e–02	9.9805e–02	0.1036	0.1002	9.9015
2	2	0.1122	0.1101	0.1054	0.1014	0.1002	0.1019	0.1065	0.1055	0.
3	3	0.1173	0.1135	0.1070	0.1083	0.1078	0.1094	0.1141	0.1079	0.
4	4	0.1181	0.1166	0.1120	0.1094	0.1106	0.1130	0.1158	0.1097	0.
5	5	0.1120	0.1110	0.1077	0.1066	0.1076	0.1100	0.1138	0.1087	0.
6	6	0.1146	0.1131	0.1079	0.1059	0.1065	0.1088	0.1126	0.1071	0.
7	7	0.1062	0.1068	0.1065	0.1096	0.1072	0.1057	0.1067	0.1026	0.
8	8	0.1092	0.1123	0.1123	0.1096	0.1083	0.1083	0.1096	0.1029	0.
9	9	0.1040	0.1076	0.1093	0.1073	0.1049	0.1043	0.1058	0.1014	0.
10	10	0.1043	0.1075	0.1094	0.1070	0.1036	0.1031	0.1061	0.1018	0.
11	11	0.1035	0.1065	0.1080	0.1072	0.1040	0.1034	0.1064	0.1022	0.
12	12	0.1180	0.1169	0.1159	0.1121	0.1132	0.1136	0.1120	0.1104	0.
13	13	0.1164	0.1157	0.1153	0.1121	0.1142	0.1147	0.1119	0.1095	0.
14	14	0.1103	0.1099	0.1105	0.1073	0.1100	0.1113	0.1092	0.1048	0.
15	15	0.1184	0.1127	0.1088	0.1132	0.1148	0.1137	0.1098	0.1099	0.
16	16	0.1195	0.1131	0.1076	0.1112	0.1144	0.1141	0.1095	0.1081	0.
17	17	0.1156	0.1093	0.1050	0.1109	0.1120	0.1097	0.1042	0.1036	0.

Sample Number

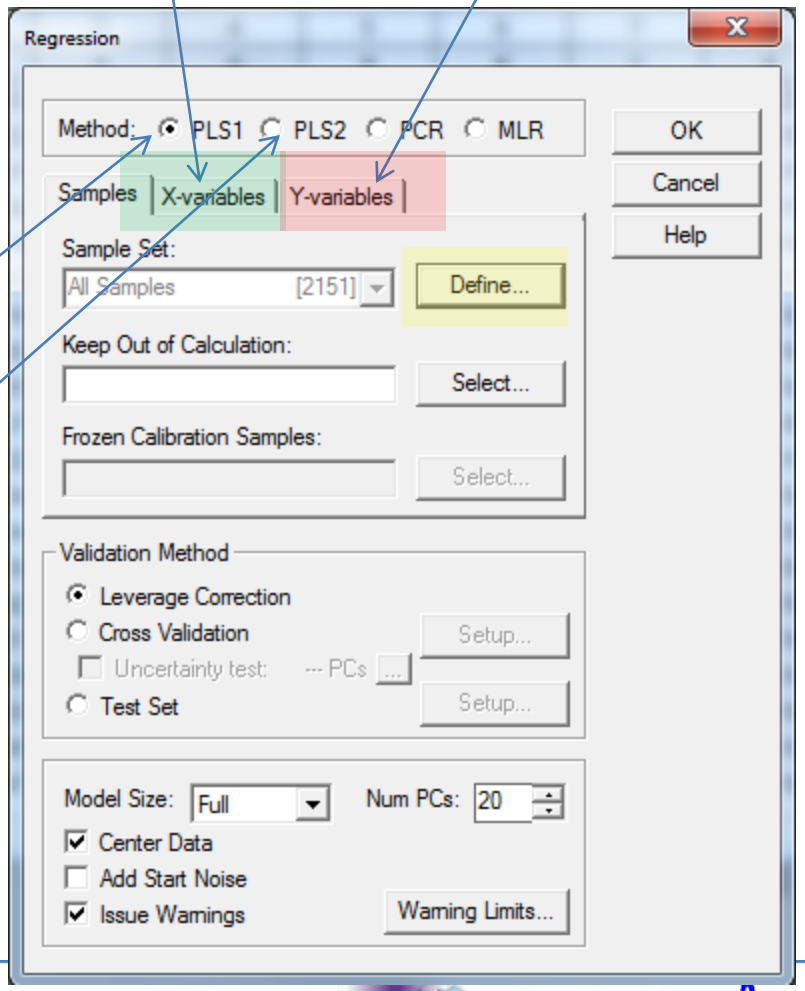
Copy chemistry from Excel

		TPH	350	351	352	353	354	355	356	357	
		1	2	3	4	5	6	7	8	9	
*	1	65.5693	0.1096	0.1091	0.1062	0.1023	9.9516e–02	9.9805e–02	0.1036	0.1002	9.!
*	2	132.7743	0.1122	0.1101	0.1054	0.1014	0.1002	0.1019	0.1065	0.1055	
*	3	267.1843	0.1173	0.1135	0.1070	0.1083	0.1078	0.1094	0.1141	0.1079	
*	4	401.5943	0.1181	0.1166	0.1120	0.1094	0.1106	0.1130	0.1158	0.1097	
*	5	536.0043	0.1120	0.1110	0.1077	0.1066	0.1076	0.1100	0.1138	0.1087	
*	6	670.4143	0.1146	0.1131	0.1079	0.1059	0.1065	0.1088	0.1126	0.1071	
*	7	804.8243	0.1062	0.1068	0.1065	0.1096	0.1072	0.1057	0.1067	0.1026	
*	8	939.2343	0.1092	0.1123	0.1123	0.1096	0.1083	0.1083	0.1096	0.1029	
*	9	1.0736e+03	0.1040	0.1076	0.1093	0.1073	0.1049	0.1043	0.1058	0.1014	
*	10	1.2081e+03	0.1043	0.1075	0.1094	0.1070	0.1036	0.1031	0.1061	0.1018	
*	11	1.3425e+03	0.1035	0.1065	0.1080	0.1072	0.1040	0.1034	0.1064	0.1022	
*	12	1.4769e+03	0.1180	0.1169	0.1159	0.1121	0.1132	0.1136	0.1120	0.1104	
*	13	1.6113e+03	0.1164	0.1157	0.1153	0.1121	0.1142	0.1147	0.1119	0.1095	
*	14	1.7457e+03	0.1103	0.1099	0.1105	0.1073	0.1100	0.1113	0.1092	0.1048	
*	15	1.8801e+03	0.1184	0.1127	0.1088	0.1132	0.1148	0.1137	0.1098	0.1099	
*	16	2.0145e+03	0.1195	0.1131	0.1076	0.1112	0.1144	0.1141	0.1095	0.1081	
*	17	2.1489e+03	0.1156	0.1093	0.1050	0.1109	0.1120	0.1097	0.1042	0.1036	
*	18	2.2833e+03	0.1186	0.1127	0.1085	0.1130	0.1140	0.1121	0.1075	0.1049	
*	19	2.4177e+03	0.1187	0.1172	0.1156	0.1132	0.1130	0.1126	0.1115	0.1125	
*	20	2.5522e+03	0.1163	0.1149	0.1129	0.1084	0.1097	0.1096	0.1063	0.1086	
*	21	2.6866e+03	0.1183	0.1194	0.1201	0.1164	0.1169	0.1158	0.1116	0.1139	
*	22	2.8210e+03	0.1195	0.1195	0.1184	0.1137	0.1145	0.1146	0.1124	0.1130	
*	23	2.9554e+03	0.1197	0.1207	0.1197	0.1144	0.1150	0.1146	0.1111	0.1111	
*	24	3.0898e+03	0.1035	0.1067	0.1094	0.1020	9.7332e–02	9.6085e–02	9.8539e–02	9.9847e–02	



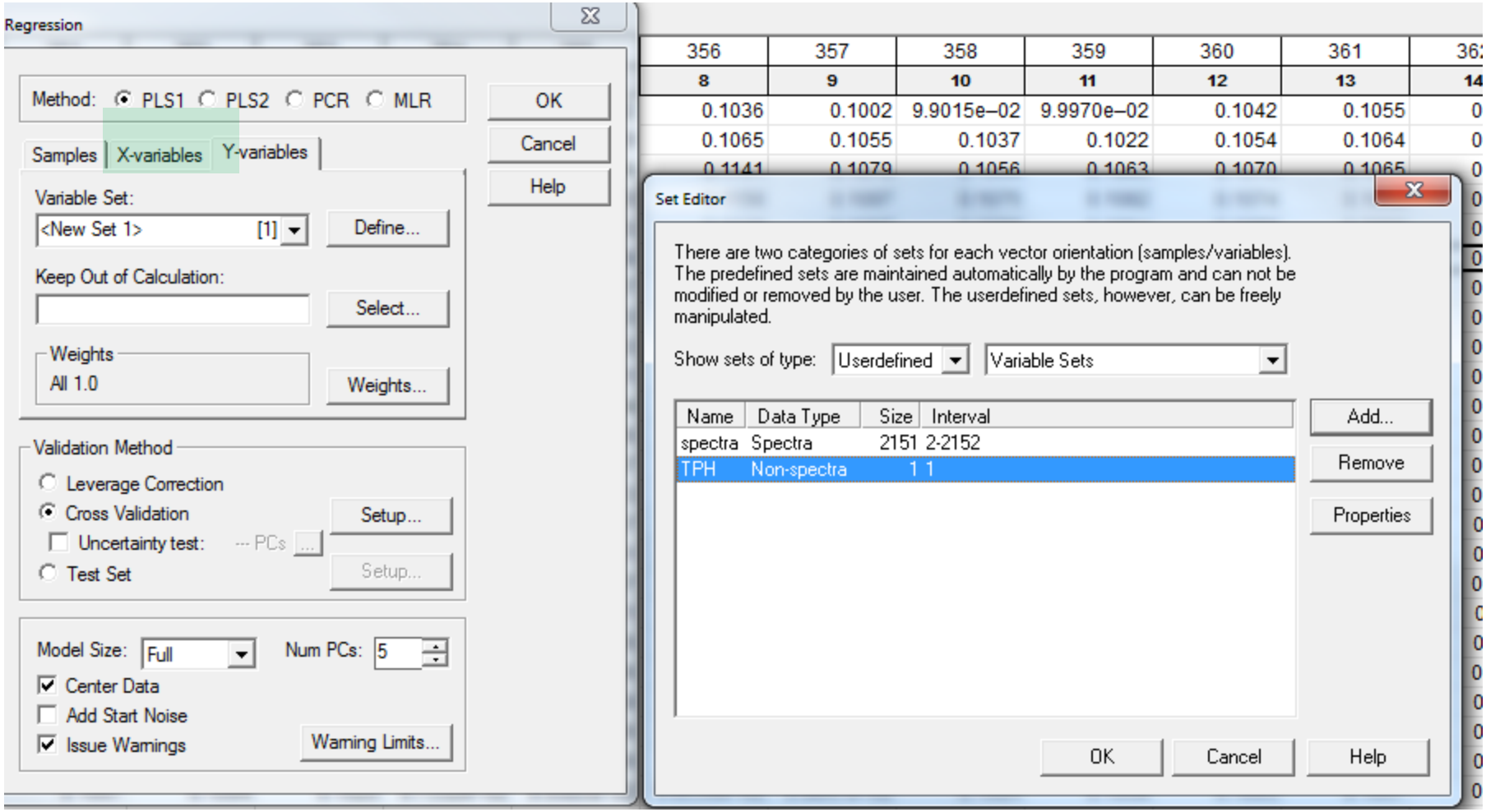
Spectra

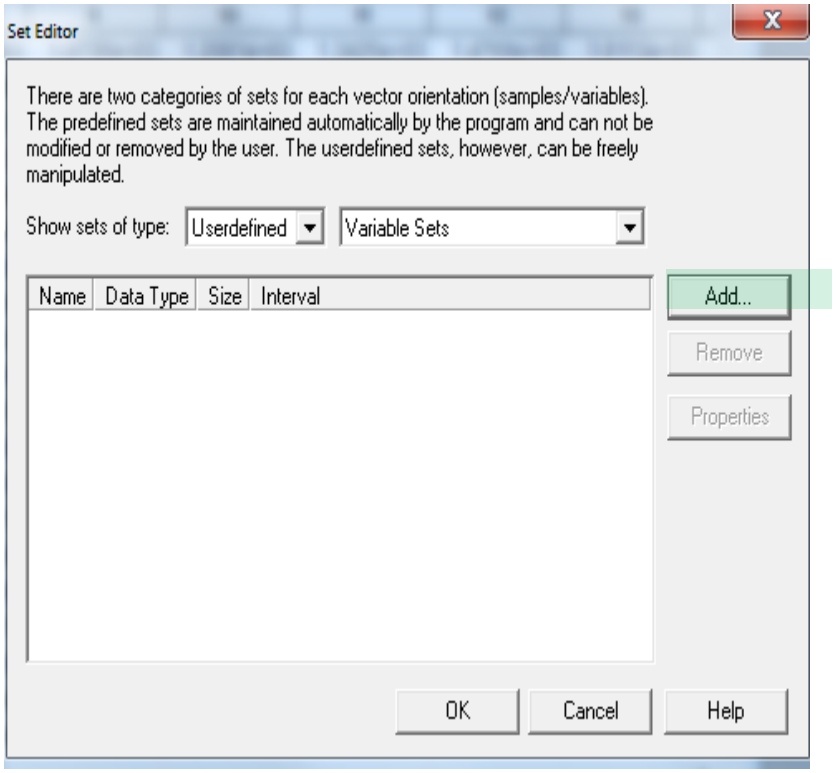
chemistry



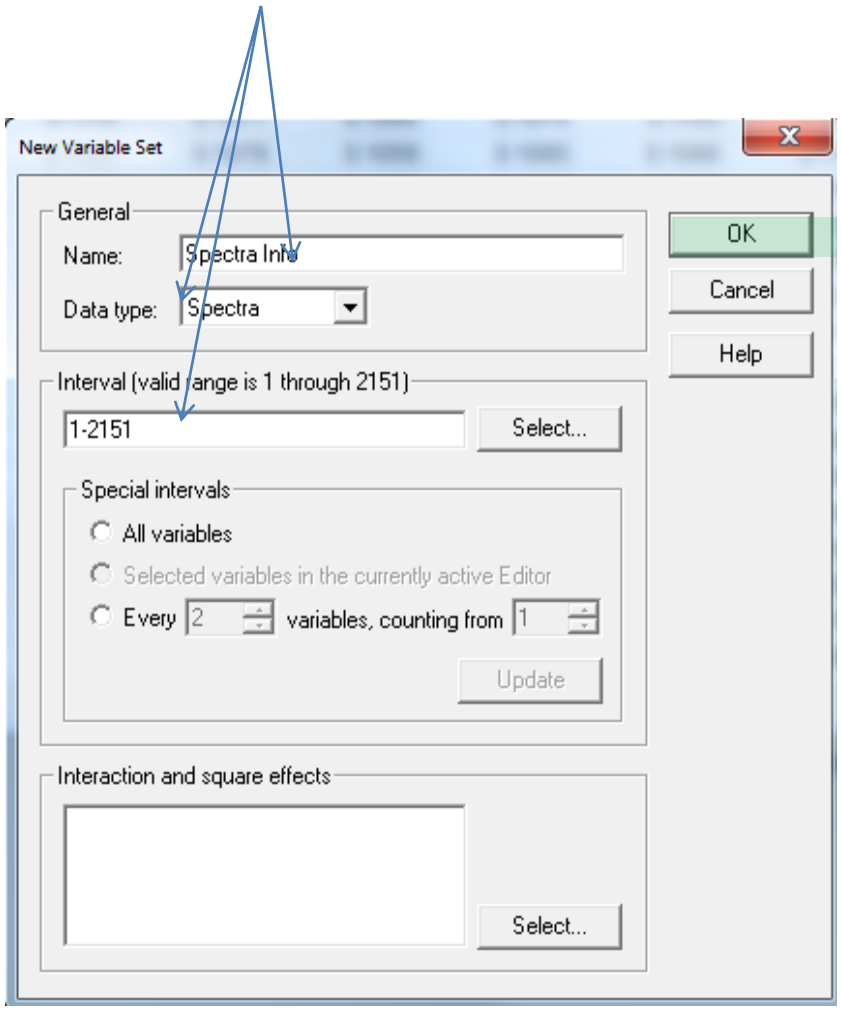
If one attribute is needed (PLS)

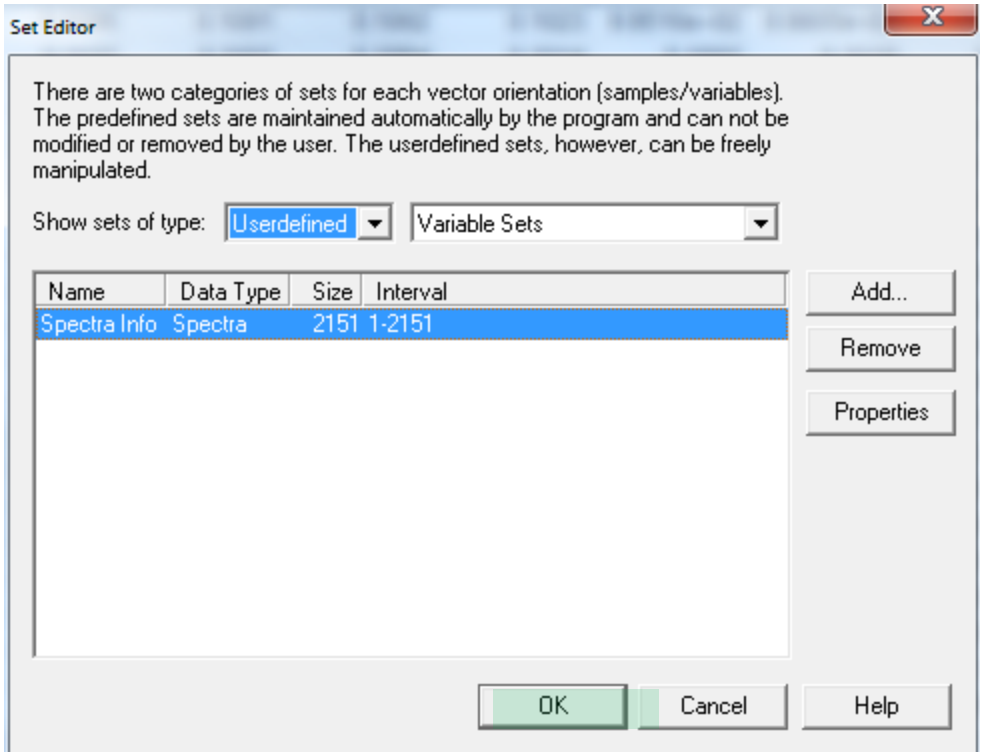
If two or more attribute are used simultaneously (PLS)

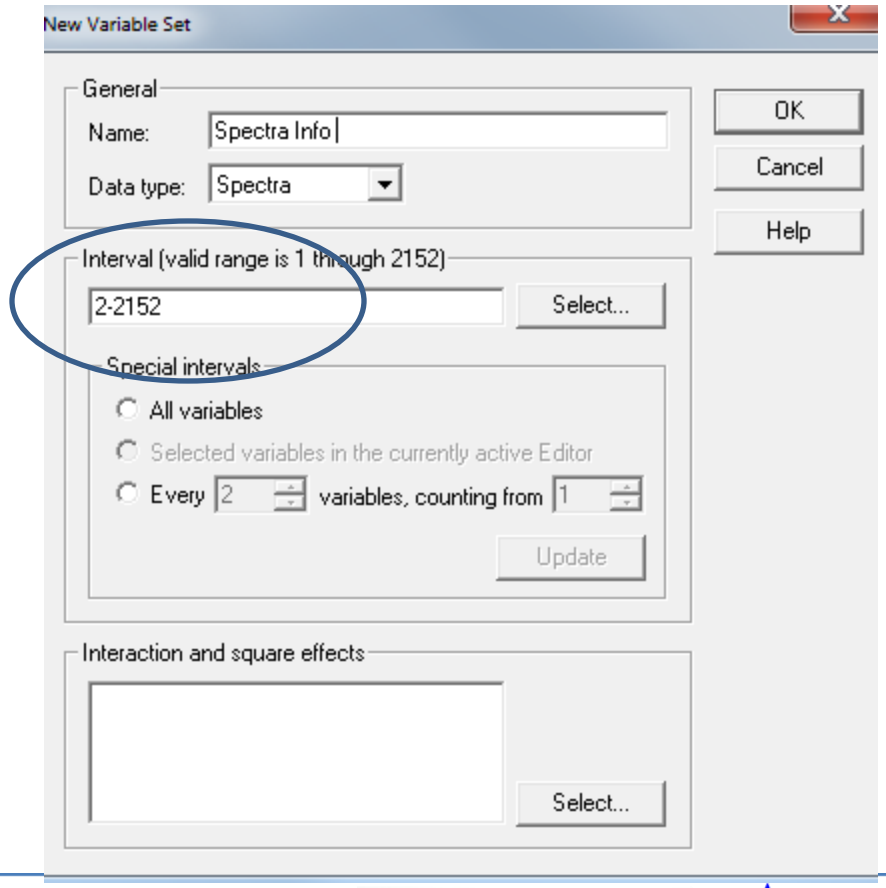

Regression
Method: PLS1 PLS2 PCR MLR
OK
Cancel
Help
Samples X-variables Y-variables
Variable Set:
<New Set 1> [1]
Define...
Keep Out of Calculation:
Select...
Weights
All 1.0
Weights...
Validation Method
Leverage Correction
Cross Validation
Setup...
Uncertainty test: --- PCs
Test Set
Setup...
Model Size: Full
Num PCs: 5
Center Data
Add Start Noise
Issue Warnings
Warning Limits...
356 357 358 359 360 361
8 9 10 11 12 13 14
0.1036 0.1002 9.9015e-02 9.9970e-02 0.1042 0.1055
0.1065 0.1055 0.1037 0.1022 0.1054 0.1064
Set Editor
There are two categories of sets for each vector orientation (samples/variables). The predefined sets are maintained automatically by the program and can not be modified or removed by the user. The userdefined sets, however, can be freely manipulated.
Show sets of type: Userdefined Variable Sets
Name Data Type Size Interval
spectra Spectra 2151 2-2152
TPH Non-spectra 1 1
Add...
Remove
Properties
OK
Cancel
Help

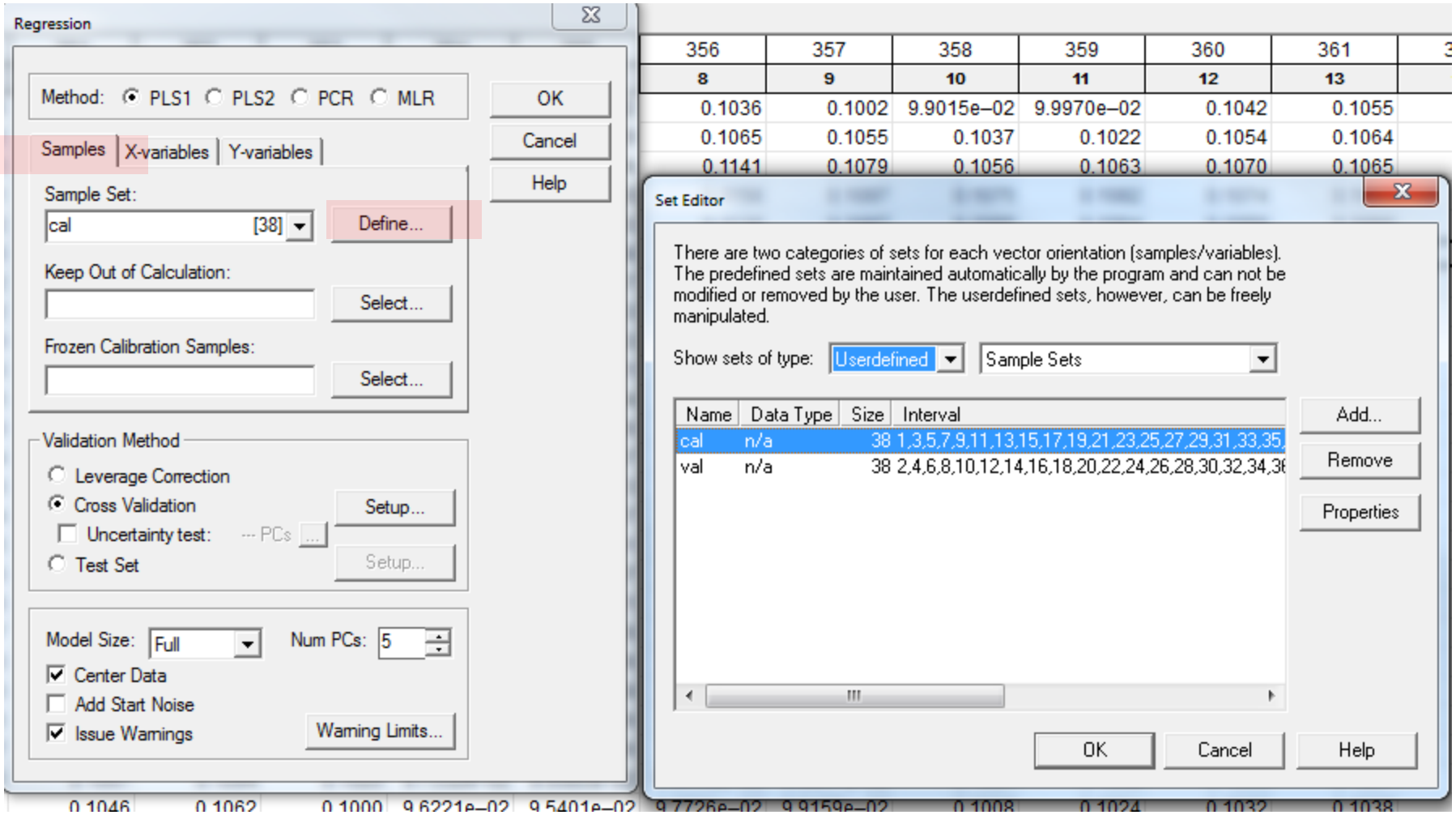

Set Editor
There are two categories of sets for each vector orientation (samples/variables). The predefined sets are maintained automatically by the program and can not be modified or removed by the user. The userdefined sets, however, can be freely manipulated.
Show sets of type:
Userdefined
Variable Sets
Name
Data Type
Size
Interval
Add...
Remove
Properties
OK
Cancel
Help

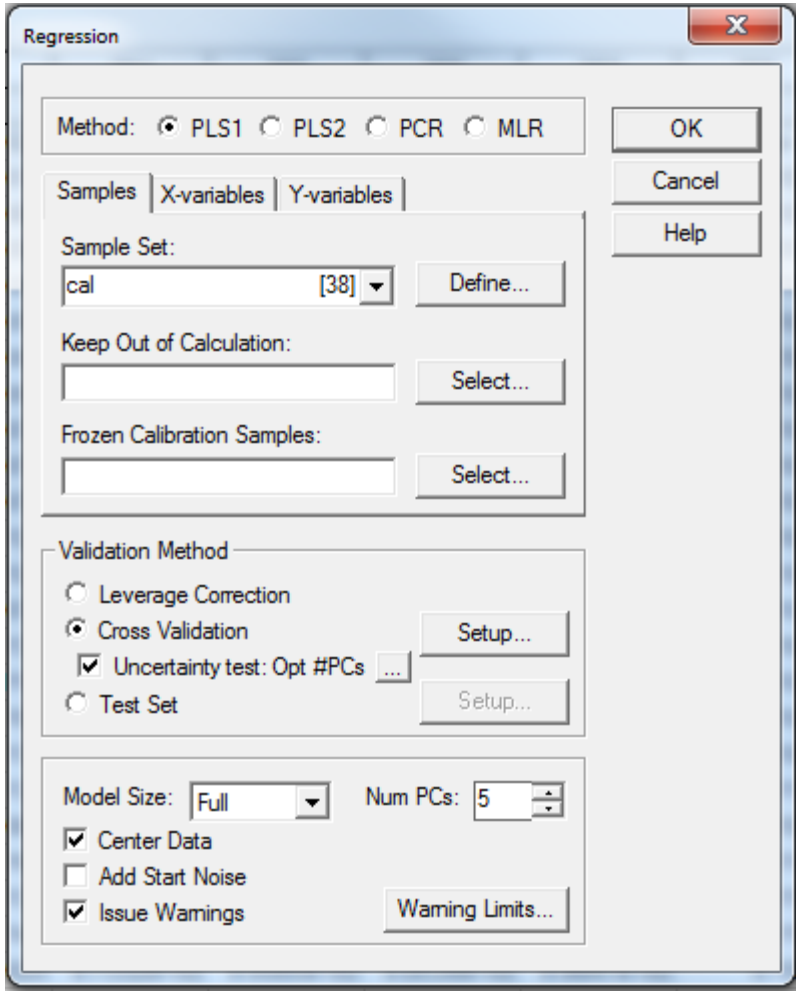
Change


New Variable Set
General
Name:
Spectra Info
Data type:
Spectra
OK
Cancel
Help
Interval (valid range is 1 through 2151)
1-2151
Select...
Special intervals
All variables
Selected variables in the currently active Editor
Every
2
variables, counting from
1
Update
Interaction and square effects
Select...

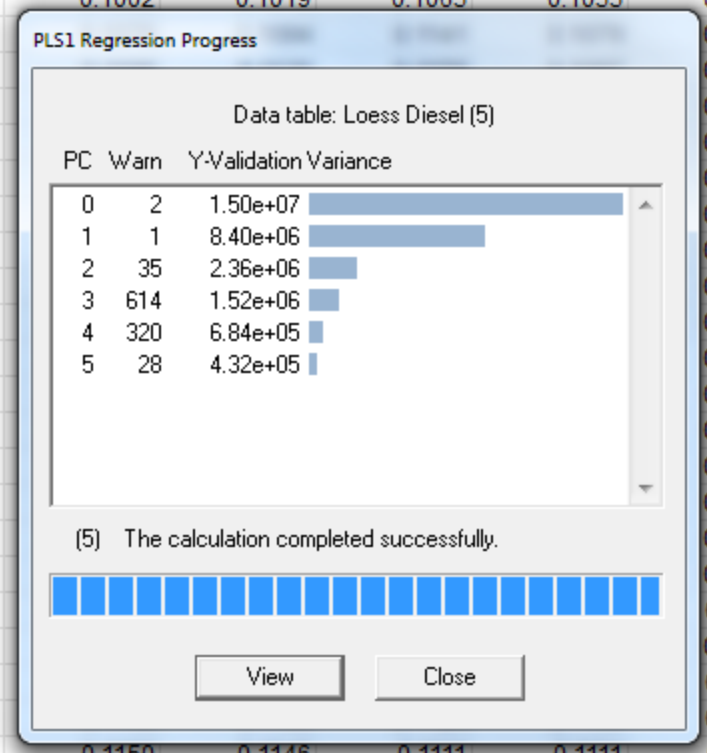

Set Editor
There are two categories of sets for each vector orientation (samples/variables). The predefined sets are maintained automatically by the program and can not be modified or removed by the user. The userdefined sets, however, can be freely manipulated.
Show sets of type:
Userdefined
Variable Sets
Name
Data Type
Size
Interval
Spectra Info
Spectra
2151
1-2151
Add...
Remove
Properties
OK
Cancel
Help

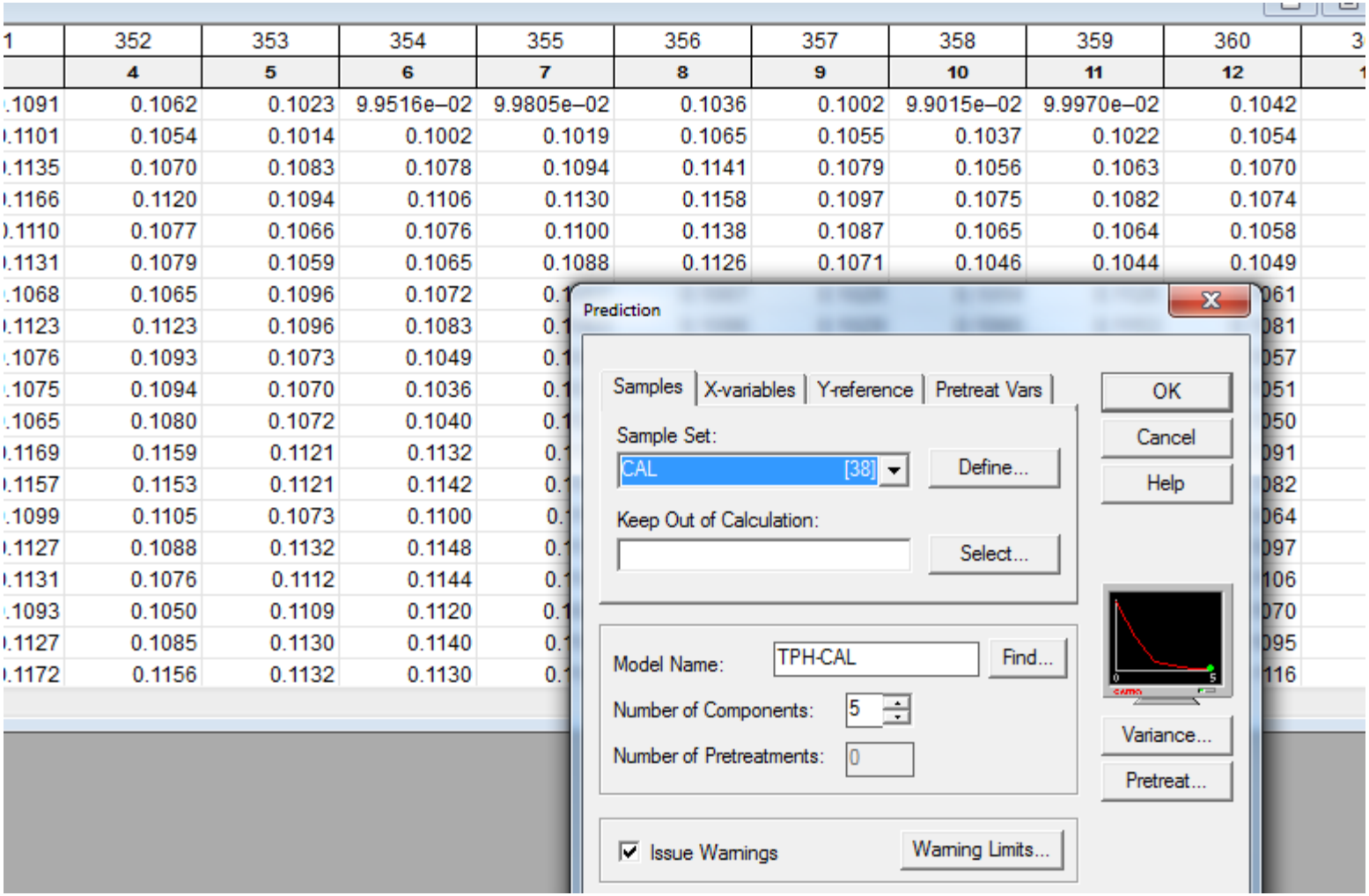

New Variable Set
General
Name:
Spectra Info
Data type:
Spectra
OK
Cancel
Help
Interval (valid range is 1 through 2152)
2-2152
Select...
Special intervals
All variables
Selected variables in the currently active Editor
Every
2
variables, counting from
1
Update
Interaction and square effects
Select...


Regression
Method: PLS1 PLS2 PCR MLR
OK
Cancel
Help
Samples
X-variables
Y-variables
Sample Set:
cal [38]
Define...
Keep Out of Calculation:
Select...
Frozen Calibration Samples:
Select...
Validation Method
Leverage Correction
Cross Validation
Setup...
Uncertainty test:
PCs
Test Set
Setup...
Model Size: Full
Num PCs: 5
Center Data
Add Start Noise
Issue Warnings
Warning Limits...
356 357 358 359 360 361
8 9 10 11 12 13
0.1036 0.1002 9.9015e−02 9.9970e−02 0.1042 0.1055
0.1065 0.1055 0.1037 0.1022 0.1054 0.1064
0.1141 0.1079 0.1056 0.1063 0.1070 0.1065
Set Editor
There are two categories of sets for each vector orientation (samples/variables). The predefined sets are maintained automatically by the program and can not be modified or removed by the user. The userdefined sets, however, can be freely manipulated.
Show sets of type: Userdefined
Sample Sets
Name Data Type Size Interval
cal n/a 38 1,3,5,7,9,11,13,15,17,19,21,23,25,27,29,31,33,35,
val n/a 38 2,4,6,8,10,12,14,16,18,20,22,24,26,28,30,32,34,36
Add...
Remove
Properties
OK
Cancel
Help


Regression
Method: PLS1 PLS2 PCR MLR
OK
Cancel
Help
Samples
X-variables
Y-variables
Sample Set:
cal [38]
Define...
Keep Out of Calculation:
Select...
Frozen Calibration Samples:
Select...
Validation Method
Leverage Correction
Cross Validation
Setup...
Uncertainty test: Opt #PCs
...
Test Set
Setup...
Model Size: Full
Num PCs: 5
Center Data
Add Start Noise
Issue Warnings
Warning Limits...

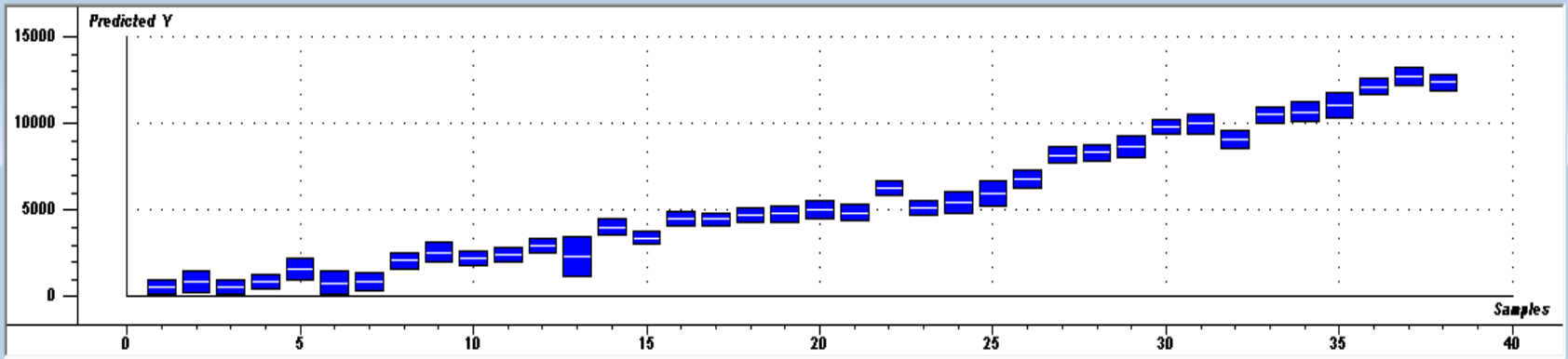
350	351	352	353	354	355	356	357	358	359	360	3
2	3	4	5	6	7	8	9	10	11	12	
0.1096	0.1091	0.1062	0.1023	9.9516e–02	9.9805e–02	0.1036	0.1002	9.9015e–02	9.9970e–02	0.1042	
0.1122	0.1101	0.1054	0.1014	0.1002	0.1019	0.1065	0.1055	0.1037	0.1022	0.1054	
0.1173	0.1135	0.1070	0.1083					0.1056	0.1063	0.1070	
0.1181	0.1166	0.1120	0.1094					0.1075	0.1082	0.1074	
0.1120	0.1110	0.1077	0.1066					0.1065	0.1064	0.1058	
0.1146	0.1131	0.1079	0.1059					0.1046	0.1044	0.1049	
0.1062	0.1068	0.1065	0.1096					0.1059	0.1126	0.1061	
0.1092	0.1123	0.1123	0.1096					0.1065	0.1153	0.1081	
0.1040	0.1076	0.1093	0.1073					0.1051	0.1124	0.1057	
0.1043	0.1075	0.1094	0.1070					0.1055	0.1128	0.1051	
0.1035	0.1065	0.1080	0.1072					0.1053	0.1117	0.1050	
0.1180	0.1169	0.1159	0.1121					0.1098	0.1099	0.1091	
0.1164	0.1157	0.1153	0.1121					0.1087	0.1089	0.1082	
0.1103	0.1099	0.1105	0.1073					0.1037	0.1050	0.1064	
0.1184	0.1127	0.1088	0.1132					0.1086	0.1070	0.1097	
0.1195	0.1131	0.1076	0.1112					0.1075	0.1077	0.1106	
0.1156	0.1093	0.1050	0.1109					0.1044	0.1061	0.1070	
0.1186	0.1127	0.1085	0.1130					0.1060	0.1094	0.1095	
0.1187	0.1172	0.1156	0.1132					0.1132	0.1134	0.1116	
0.1163	0.1149	0.1129	0.1084					0.1090	0.1082	0.1091	
0.1183	0.1194	0.1201	0.1164					0.1148	0.1145	0.1123	
0.1195	0.1195	0.1184	0.1137					0.1128	0.1122	0.1126	
0.1197	0.1207	0.1197	0.1144	0.1150	0.1146	0.1111	0.1111	0.1116	0.1124	0.1135	
0.1035	0.1067	0.1094	0.1020	9.7332e–02	9.6085e–02	9.8539e–02	9.9847e–02	0.1025	0.1058	0.1063	




Prediction
Samples
X-variables
Y-reference
Pretreat Vars
OK
Cancel
Help
Sample Set:
CAL [38]
Define...
Keep Out of Calculation:
Select...
Model Name:
TPH-CAL
Find...
Number of Components:
5
Number of Pretreatments:
0
Variance...
Pretreat...
Issue Warnings
Warning Limits...

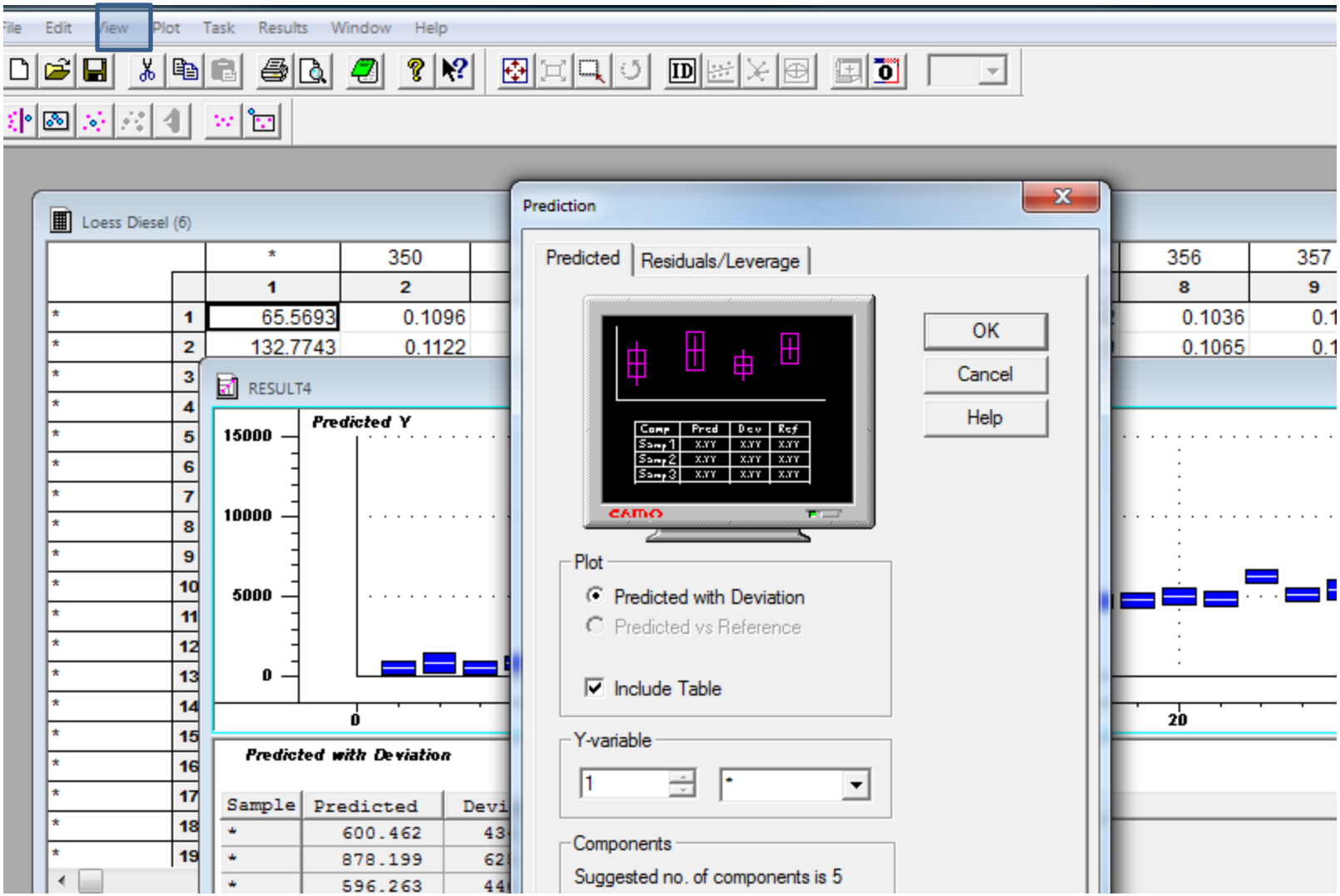
*	350	351	352	353	354	355	356	357	358	359	360	361
1	2	3	4	5	6	7	8	9	10	11	12	13
65.5693	0.1096	0.1091	0.1062	0.1023	9.9516e−02	9.9805e−02	0.1036	0.1002	9.9015e−02	9.9970e−02	0.1042	0.1(
132.7743	0.1122	0.1101	0.1054	0.1014	0.1002	0.1019	0.1065	0.1055	0.1037	0.1022	0.1054	0.1(

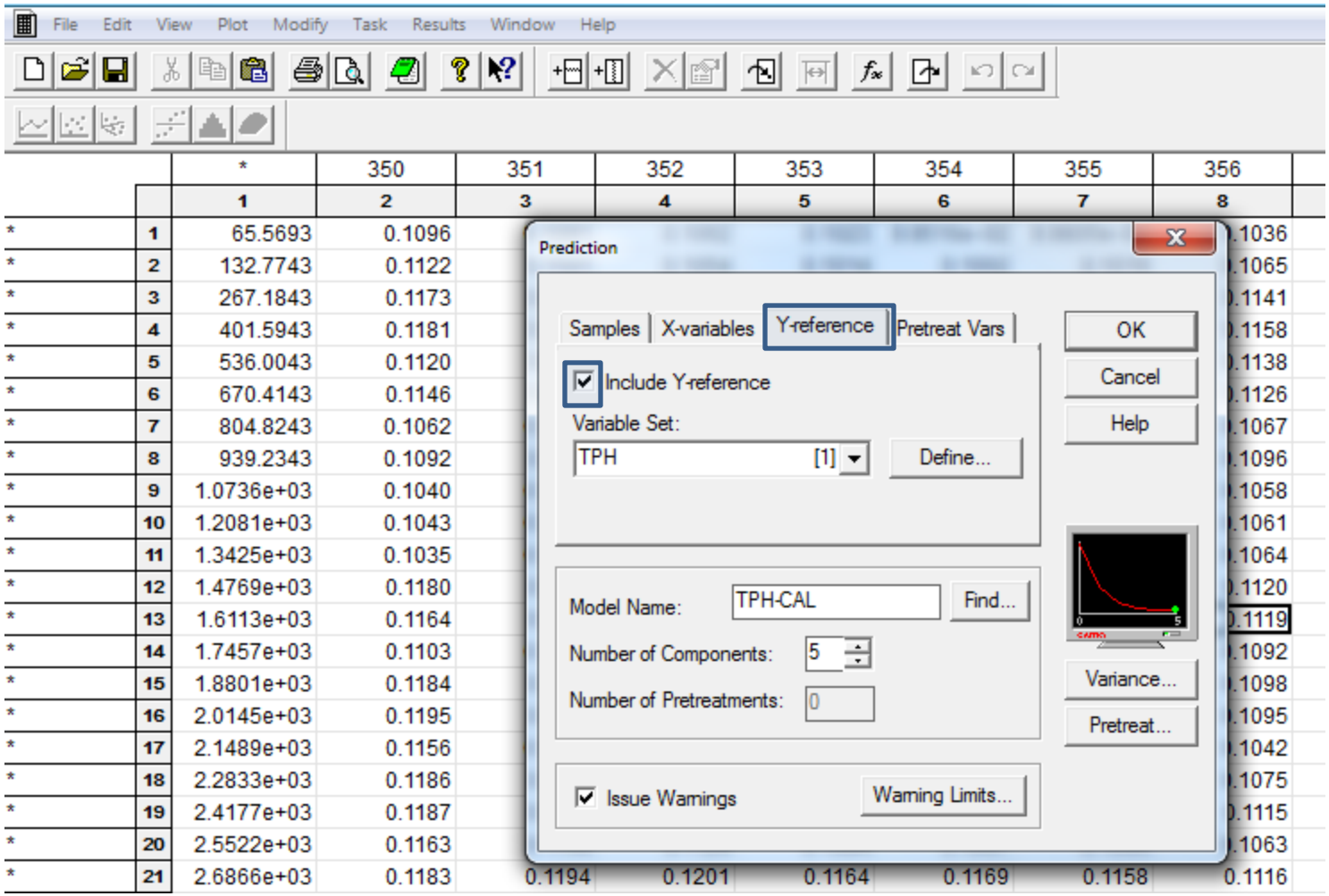
RESULT4

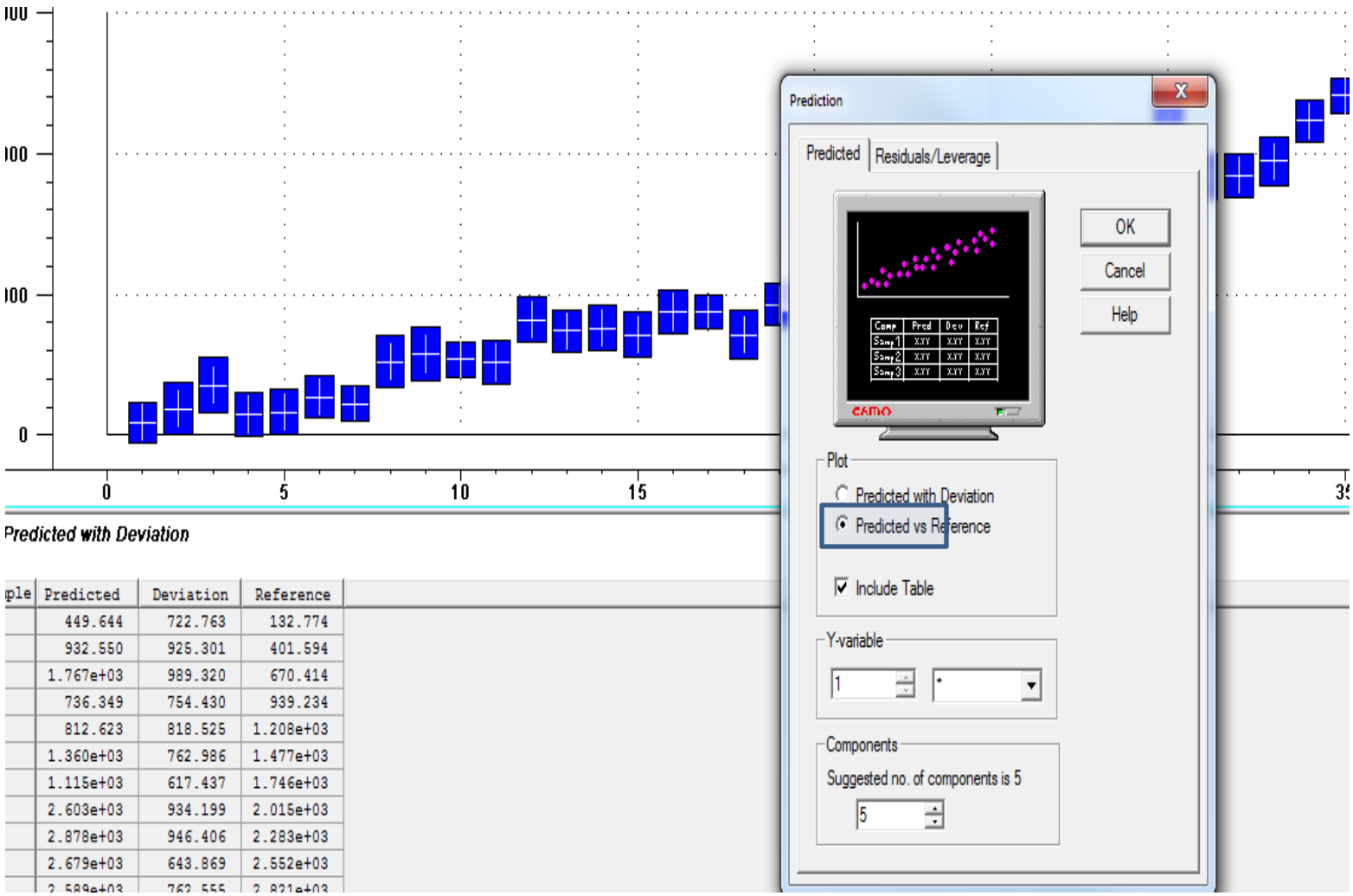


Predicted with Deviation

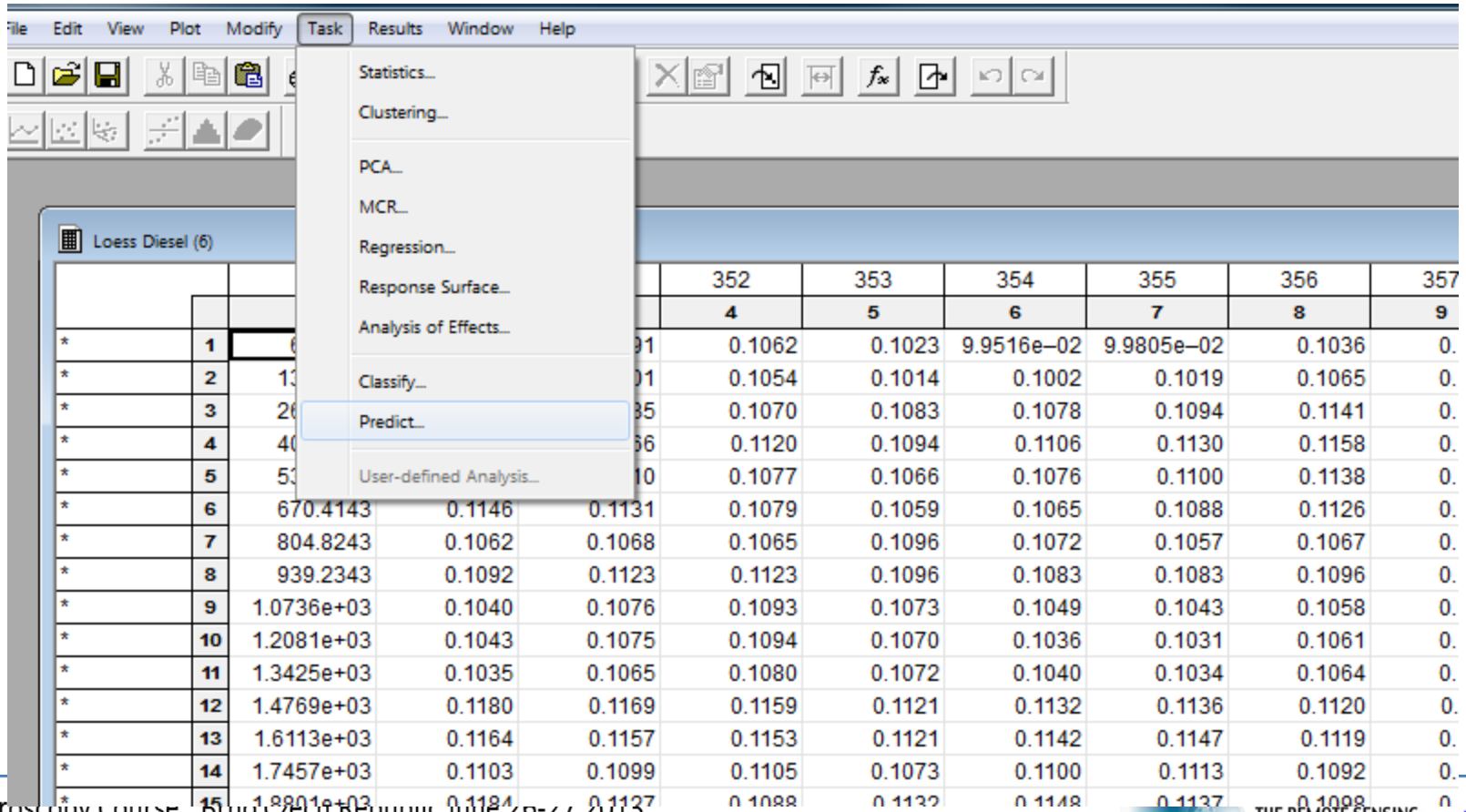
Sample	Predicted	Deviation
*	600.462	434.555
*	878.199	625.050
*	596.263	440.052
*	915.435	426.788
*	1.589e+03	610.382
*	787.546	680.117
*	911.229	519.392
*	2.108e+03	463.972
*	2.550e+03	569.586


File Edit View Plot Task Results Window Help
Loess Diesel (6)
RESULT4
Predicted Y
Predicted with Deviation
Sample Predicted Devi
600.462
878.199
596.263
Prediction
Predicted
Residuals/Leverage
OK
Cancel
Help
Plot
Predicted with Deviation
Predicted vs Reference
Include Table
Y-variable
Components
Suggested no. of components is 5

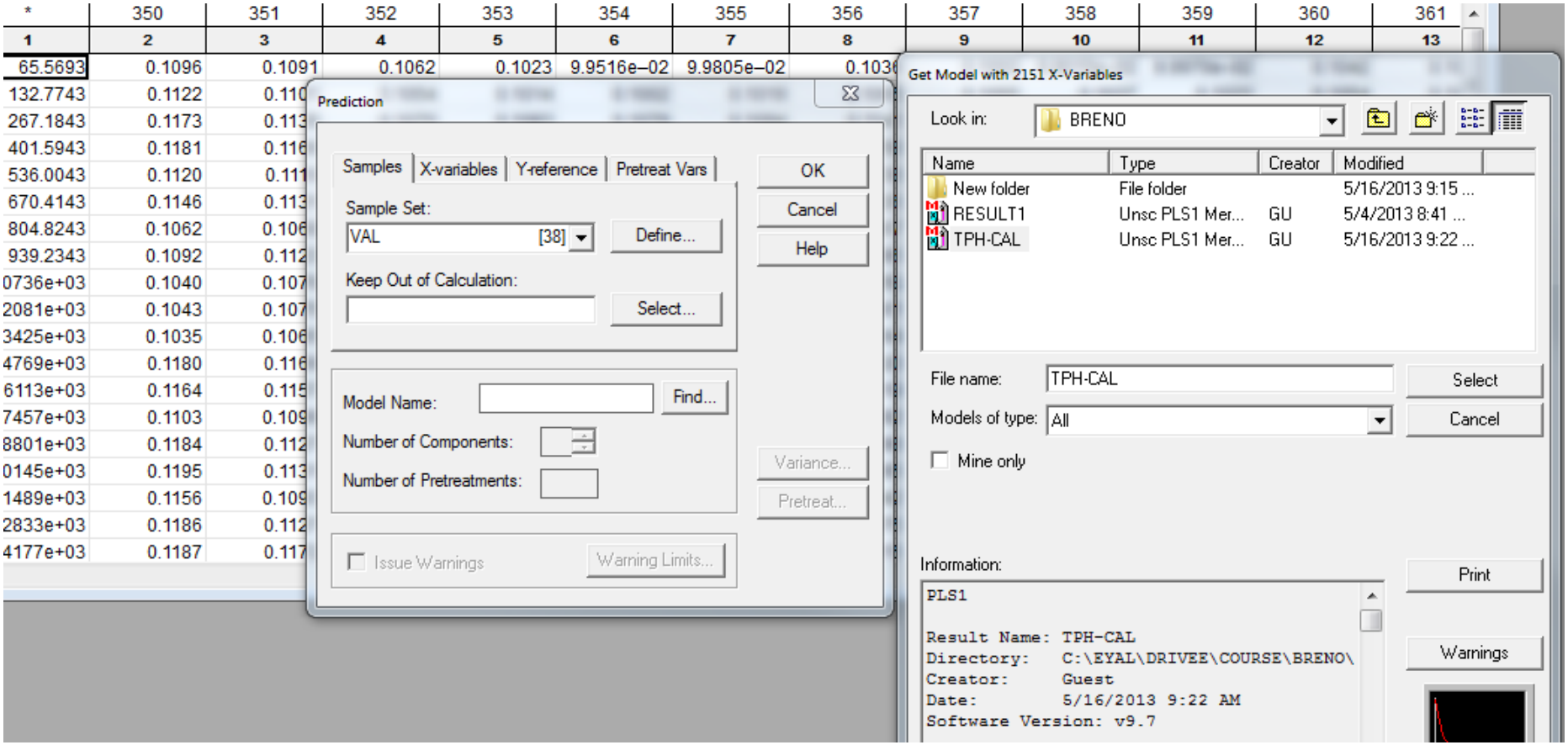

File Edit View Plot Modify Task Results Window Help
Prediction
Samples
X-variables
Y-reference
Pretreat Vars
Include Y-reference
Variable Set:
TPH
[1]
Define...
OK
Cancel
Help
Model Name:
TPH-CAL
Find...
Number of Components:
5
Number of Pretreatments:
0
Variance...
Pretreat...
Issue Warnings
Warning Limits...

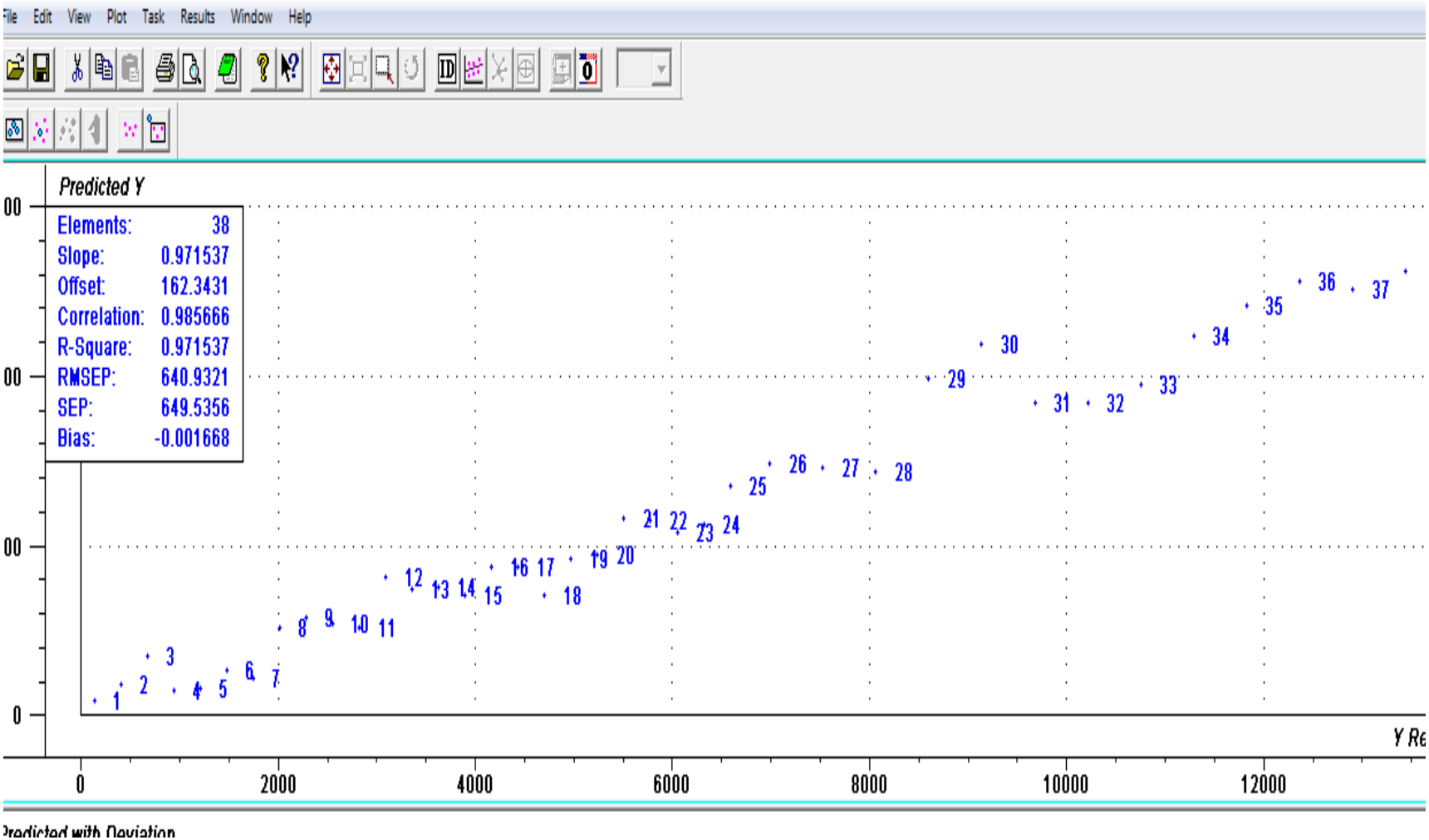

Prediction
Predicted
Residuals/Leverage
OK
Cancel
Help
Plot
Predicted with Deviation
Predicted vs Reference
Include Table
Y-variable
1
Components
Suggested no. of components is 5
5
Predicted with Deviation
ple Predicted Deviation Reference
449.644 722.763 132.774
932.550 925.301 401.594
1.767e+03 989.320 670.414
736.349 754.430 939.234
812.623 818.525 1.208e+03
1.360e+03 762.986 1.477e+03
1.115e+03 617.437 1.746e+03
2.603e+03 934.199 2.015e+03
2.878e+03 946.406 2.283e+03
2.679e+03 643.869 2.552e+03

Save the model and Close the Cal results ,
go to predict

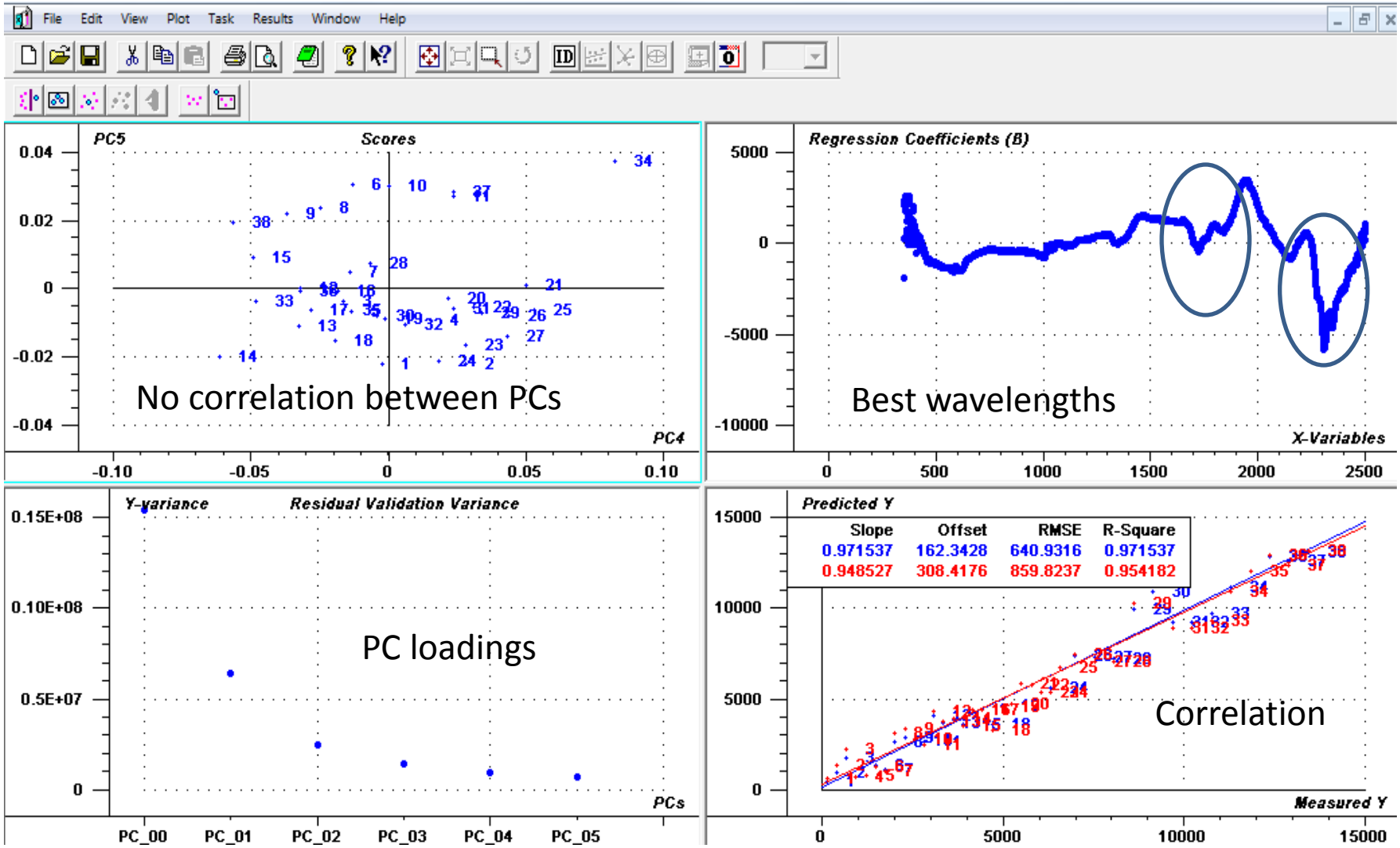


					352	353	354	355	356	357
					4	5	6	7	8	9
*	1	6		91	0.1062	0.1023	9.9516e–02	9.9805e–02	0.1036	0.
*	2	1		01	0.1054	0.1014	0.1002	0.1019	0.1065	0.
*	3	2		35	0.1070	0.1083	0.1078	0.1094	0.1141	0.
*	4	4		66	0.1120	0.1094	0.1106	0.1130	0.1158	0.
*	5	5		10	0.1077	0.1066	0.1076	0.1100	0.1138	0.
*	6	670.4143	0.1146	0.1131	0.1079	0.1059	0.1065	0.1088	0.1126	0.
*	7	804.8243	0.1062	0.1068	0.1065	0.1096	0.1072	0.1057	0.1067	0.
*	8	939.2343	0.1092	0.1123	0.1123	0.1096	0.1083	0.1083	0.1096	0.
*	9	1.0736e+03	0.1040	0.1076	0.1093	0.1073	0.1049	0.1043	0.1058	0.
*	10	1.2081e+03	0.1043	0.1075	0.1094	0.1070	0.1036	0.1031	0.1061	0.
*	11	1.3425e+03	0.1035	0.1065	0.1080	0.1072	0.1040	0.1034	0.1064	0.
*	12	1.4769e+03	0.1180	0.1169	0.1159	0.1121	0.1132	0.1136	0.1120	0.
*	13	1.6113e+03	0.1164	0.1157	0.1153	0.1121	0.1142	0.1147	0.1119	0.
*	14	1.7457e+03	0.1103	0.1099	0.1105	0.1073	0.1100	0.1113	0.1092	0.
*	15	1.8801e+03	0.1184	0.1137	0.1088	0.1132	0.1148	0.1137	0.1098	0


Prediction
Samples
X-variables
Y-reference
Pretreat Vars
Sample Set:
VAL [38]
Define...
Keep Out of Calculation:
Select...
Model Name:
Find...
Number of Components:
Number of Pretreatments:
Issue Warnings
Warning Limits...
OK
Cancel
Help
Variance...
Pretreat...
Get Model with 2151 X-Variables
Look in: BRENO
Name
Type
Creator
Modified
New folder
File folder
5/16/2013 9:15 ...
RESULT1
Unsc PLS1 Mer...
GU
5/4/2013 8:41 ...
TPH-CAL
Unsc PLS1 Mer...
GU
5/16/2013 9:22 ...
File name: TPH-CAL
Models of type: All
Mine only
Select
Cancel
Information:
PLS1
Result Name: TPH-CAL
Directory: C:\EYAL\DRIVEE\COURSE\BRENO\
Creator: Guest
Date: 5/16/2013 9:22 AM
Software Version: v9.7
Print
Warnings


File Edit View Plot Task Results Window Help
Predicted Y
Elements: 38
Slope: 0.971537
Offset: 162.3431
Correlation: 0.985666
R-Square: 0.971537
RMSEP: 640.9321
SEP: 649.5356
Bias: -0.001668
0
0
2000
4000
6000
8000
10000
12000
Predicted with Deviation

PLSR results with 5PC



In any SpeR-A preprocessing is a major stage to go

Preprocessing

Raw reflectance (R) → Manipulated data (M)

(Then running statistical method to perform a model)

Manipulation is any mathematical analysis done on the data base equally (every spectrum treated the same)

Some used examples:

1) Noise reduction (moving average)
2) CR (Continuum Removal)
3) Derivatives (first, second)
4) R → log (1/R)
5) Data reduction (from n wavelength number to n/m number (m > n)
6) Kubelka Munk

As Sper-A is an Empirical Approach there is no way to know which manipulation will lead the best performance

It is possible that more than one mathematical calculation will be used.

Some common Multiple Combinations:

1) Smoothing→ log(1/R) → Derivative → CR
2) log (1/R) → smoothing → CR→ derivative
3) Smoothing→ reduction → Derivative

It is almost impossible to run all preprocessing with all data mining algorithms

The Solution

A program that will do it automatically providing only the “best model”

Modeling

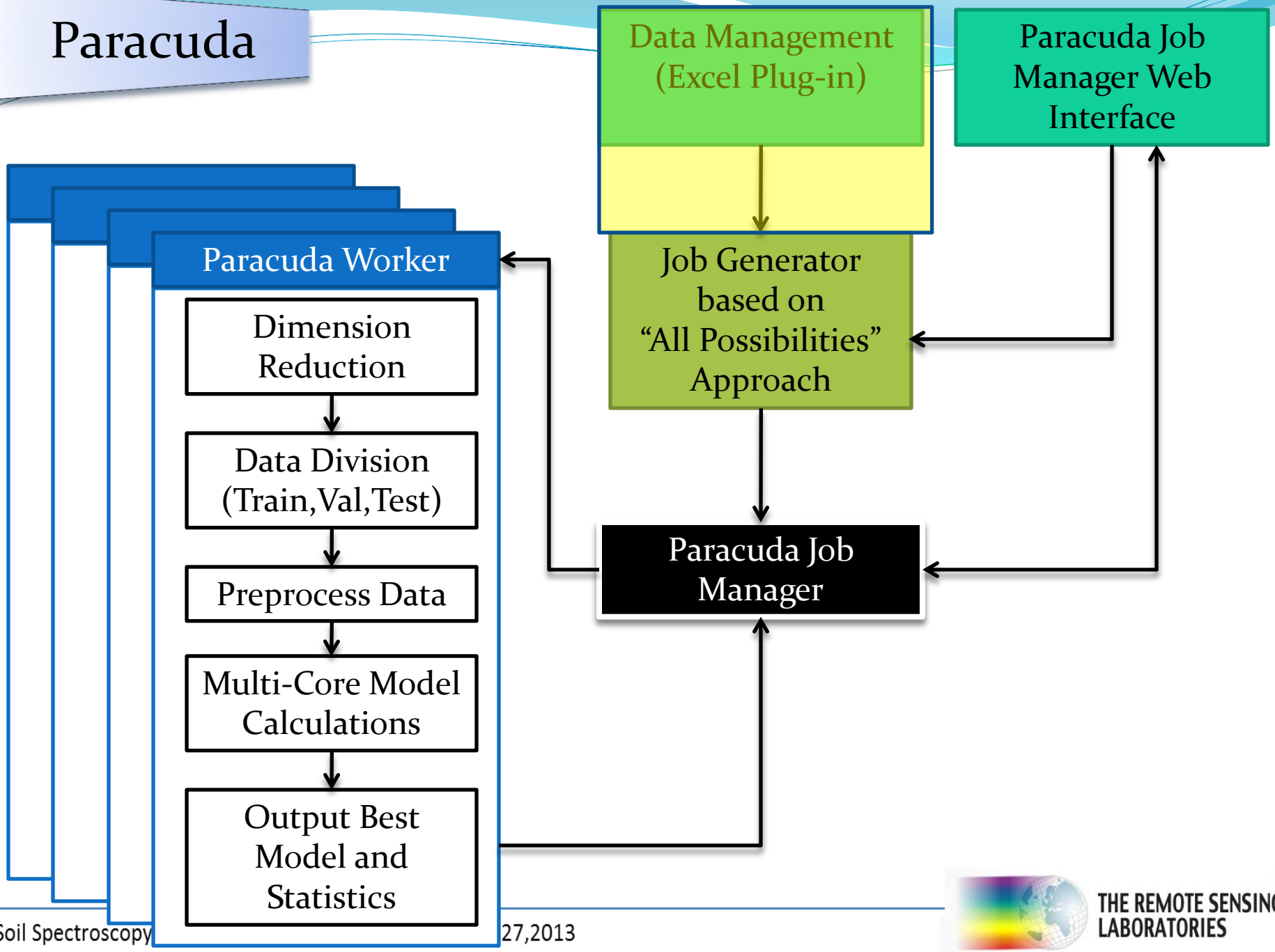
- The Problem:
  - Modeling spectroscopy data is a complicated task due to many preprocessing procedures available.
  - An "All options" approach is the best solution for reliable models, but **very difficult** to implement to many reasons:
    - Computing Power
    - No automated software available
    - Skilled personal
    - Complicated algorithms
    - Limited software capabilities

Paracuda

- The Solution:
  - Design a software suite that will include:
    - All Preprocessing algorithms.
    - All NIRA Statistical approaches.
    - Automated processing system to utilize the “All Possibilities” approach.
    - Distributed computing system for rapid model evaluation.

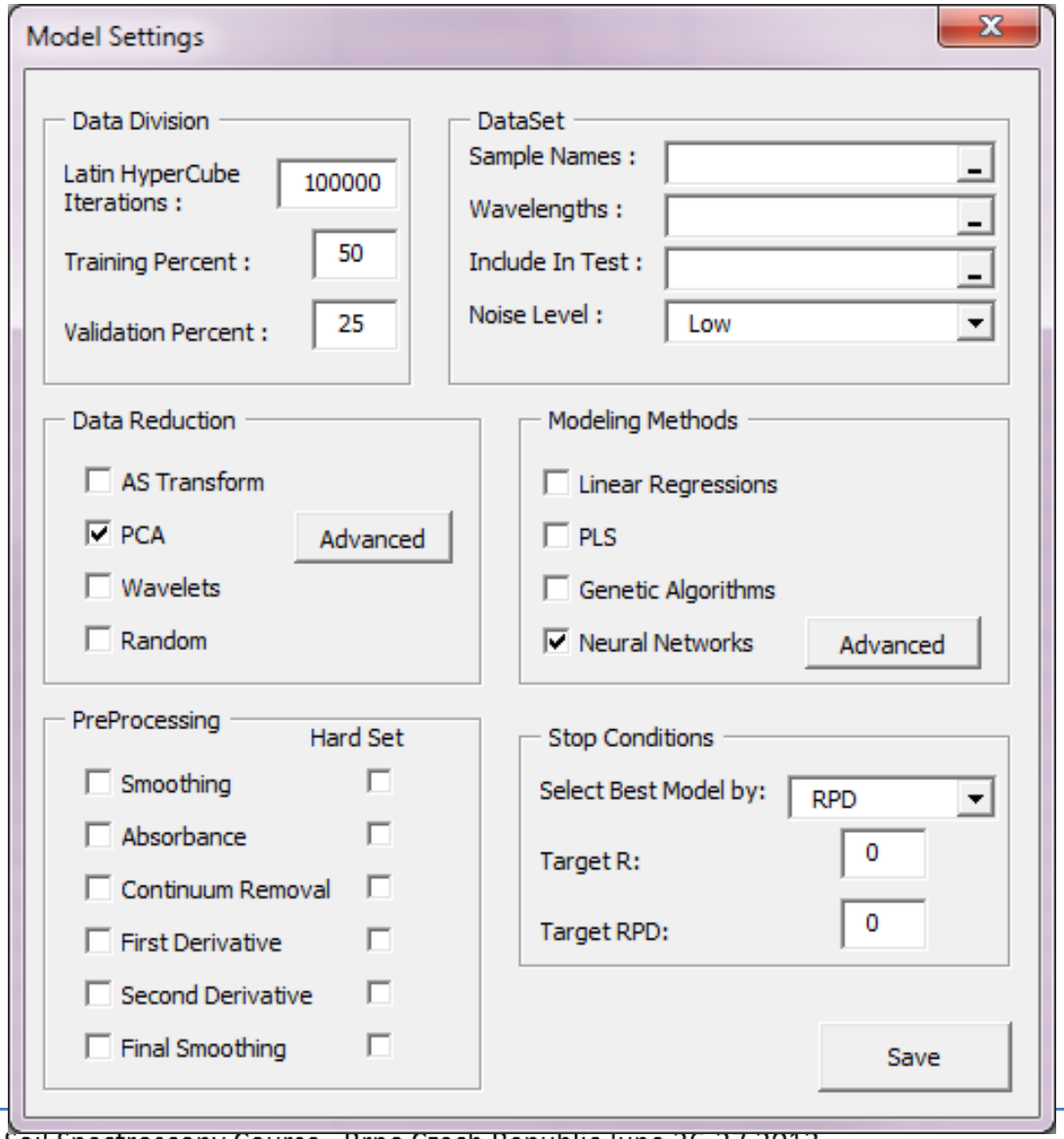
A Simple “One Click” solution

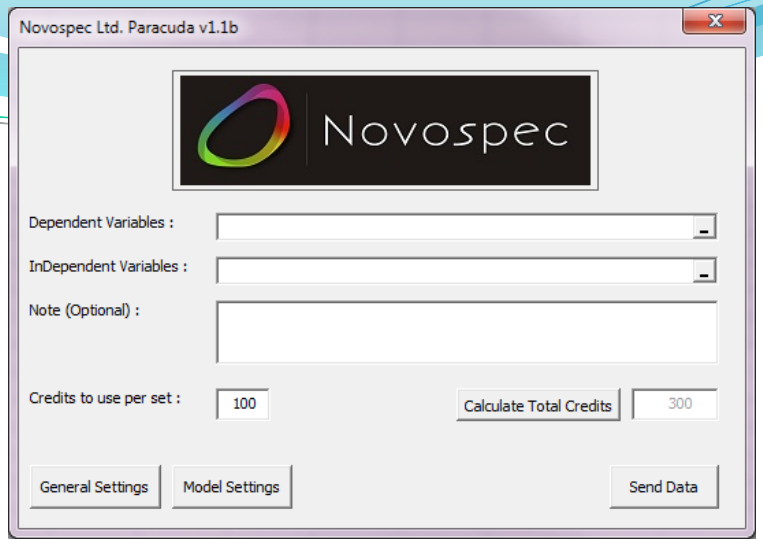
Paracuda



THE REMOTE SENSING LABORATORIES

Paracuda Excel Plug-In





Paracuda Web-Interface

- Statistical Parameters provided to the user:
  - Correlation
- RMSEP
  - SEP
  - Bias
  - SDtY
  - RPD

Welcome, guy@paracuda.com!

You have 87974 credits

Logout

ABOUT
JOBS
PREDICTIONS
STATUS
HELP
CONTACT
NOVOSPEC

All Jobs

Total models calculated: 19091000

Name	Total Models	Average Models Per Minute	Status
RSLMamba	6119000	985	Online
RSLSlave1	3094500	605	Online
RSLMonster	3600000	437	Online
RSLSlave3	353000	522	Online
RSLSlave2	11500	148	Offline
RSLCrow	5913000	515	Busy

Designed by Novospec | www.Novospec.com © 2010

Job Info

Resubmit Job | Refresh

Excel Summery File

Paracuda Matlab File

Model: Model 1 Predict

Delete Library

JobID	Status	Started	Ended
20110331181029-2SACrfS	Done	2011-04-01 12:59:24	2011-04-01 17:04:34

Variables	Samples
2151	82

Neurons	PCs	Variance Explained	Models Tested	Percent Complete
2	6	99.5459	100000	100
LHC Sets	**Training**	**Validation**	**Test**	**Best R/RPD**
100000	50	25	25	0.99122/11.4267

Smoothing		Final Smoothing	
1		1	
Absorbance	**Continuum Removal**	**Derivative**	**Second Derivative**
1	1	0	0

Select	Note
◯	2011-03-31 18:10:29 By guy@paracuda.com Loess with Diesel and Extra 6 Samples from Kerem Shalom

Delete Note

New Note

Parcuda : Comercial Solution for non professional users:

1) By credit (how much CPU time you want)
2) Send raw data (Refinance matrix and attributes)
3) Get back the best model with information on: what manipulation stream yielded the best model, the model to be used on other data bases, statistic parameters,

Advantageous: No need to spend hours to find a model, No need to be professional statistician, No need to learn or purchase sophisticated software, send and forget.