The changes made are written in the colour green to make them easier to follow.

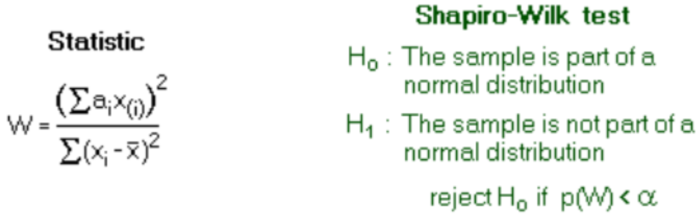
Problem 1: Statistical Tests
Report

Problem description:
We had the weekly yield of milk production of two farms whit 10 cows (Farm A) and 15 cows (Farm B). We want to check if there is a difference between the two herds at the 5% and 1% level of significance.

Outcome:
When we want to compare the mean of two different populations, in this case the two Farms, we use the 2-sample T-Test. For the 2-sample T-Test it is needed that some secondary condition be fulfilled. These conditions are:

- The samples have to be independent of each other. In independent samples, the subjects of one group do not provide information about the subjects of the other groups. Each group contains different subjects. → Two farms with different cows (=subjects).
- The sample must be following a normal distribution. → Shapiro-Wilk test
- The variances must be equal. → 2-sample F-Test otherwise Welch-Test

Results of the Shapiro-Wilk test for normality:
Toolbar “Math” → Tests… → Colum 1 & 2 “marking the data” red/blue → Shapiro-Wilk



Statistic

$$W = \frac{\left(\sum a_i x_{(i)}\right)^2}{\sum (x_i - \bar{x})^2}$$

Shapiro-Wilk test

H_0 : The sample is part of a normal distribution

H_1 : The sample is not part of a normal distribution

reject H_0 if $p(W) < \alpha$

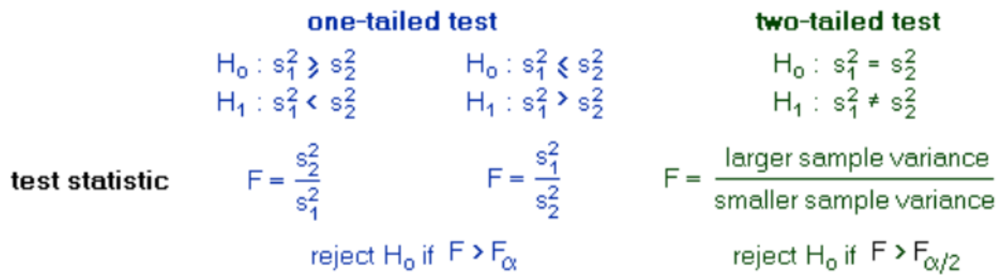
Farm A	Farm B

`Data of Block A (red):`
`-----------------------------------`
`No. of Data: 10`
`Mean: 141.9600`
`Std. Dev.: 19.0247`
`W-statistic: 0.9631`
`p(W):` **`0.8205`**
	`Data of Block B (blue):`
	`-----------------------------------`
	`No. of Data: 15`
	`Mean: 166.3533`
	`Std. Dev.: 23.7825`
	`W-statistic: 0.9349`
	`p(W):` **`0.3230`**

reject H_0 if $p(W) < α$	
α=0.01 0.8205 > 0.01 H_0 not rejected → Normal distribution α=0.05 0.8205 > 0.05 H_0 not rejected → Normal distribution	α=0.01 0.3230 > 0.01 H_0 not rejected → Normal distribution α=0.05 0.3230 > 0.05 H_0 not rejected → Normal distribution

Results of the two-sample F-test:

Toolbar “Math” → Tests… → Colum 1 & 2 “marking the data” red/blue → 2-Sample F-Test



one-tailed test **two-tailed test**

$H_0 : s_1^2 \geq s_2^2$ $H_0 : s_1^2 \leq s_2^2$ $H_0 : s_1^2 = s_2^2$

$H_1 : s_1^2 < s_2^2$ $H_1 : s_1^2 > s_2^2$ $H_1 : s_1^2 \neq s_2^2$

test statistic $F = \frac{s_2^2}{s_1^2}$ $F = \frac{s_1^2}{s_2^2}$ $F = \frac{\text{larger sample variance}}{\text{smaller sample variance}}$

reject H_0 if $F > F_\alpha$ reject H_0 if $F > F_{\alpha/2}$

<table>
<tr><th>Farm A</th><th>Farm B</th></tr>
<tr><td>Data of Block A (red):
No. of Data: 10
Mean: 141.96000
Std. Dev.: 19.02473</td><td>Data of Block B (blue):
No. of Data: 15
Mean: 166.35333
Std. Dev.: 23.78249</td></tr>
<tr><td colspan="2">reject H_0 if $F > F_{\alpha/2}$</td></tr>
<tr><td colspan="2">$F_{0.01/2}$=6.0887
1.5627 < 6.0887
H_0 not rejected → Variance is equal, no Welch-Test needed
$F_{0.05/2}$=3.7980
1.5627 < 3.7980
H_0 not rejected → Variance is equal, no Welch-Test needed</td></tr>
</table>

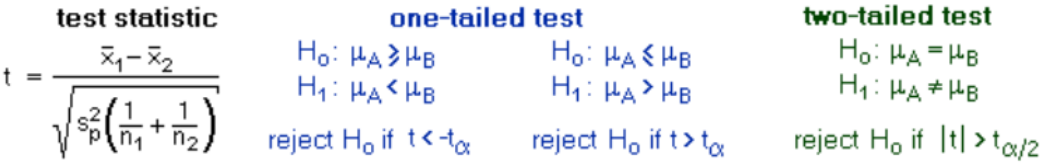
```
F-statistic: 1.5627
p(F), 1-sided: 0.2532
deg. freedom (numerator): 14
deg. freedom (denominator): 9

critical F-values:
```

α	F(α)
0.001	9.3337
0.002	7.7953
0.005	**6.0887**
0.01	5.0052
0.02	4.0709
0.025	**3.7980**
0.05	3.0255

Results of the two-sample t-test:

Toolbar “Math” → Tests… → Colum 1 & 2 “marking the data” red/blue → 2-Sample t-Test



test statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

one-tailed test

$H_0: \mu_A \geqslant \mu_B$
$H_1: \mu_A < \mu_B$
reject H_0 if $t < -t_\alpha$

$H_0: \mu_A \leqslant \mu_B$
$H_1: \mu_A > \mu_B$
reject H_0 if $t > t_\alpha$

two-tailed test

$H_0: \mu_A = \mu_B$
$H_1: \mu_A \neq \mu_B$
reject H_0 if $|t| > t_{\alpha/2}$

Farm A	Farm B
Data of Block A (red): No. of Data: 10 Mean: 141.96000 Std. Dev.: 19.02473	Data of Block B (blue): No. of Data: 15 Mean: 166.35333 Std. Dev.: 23.78249
reject H_0 if $\|t\| > t_{\alpha/2}$	
$t_{0.01/2}$=2.8073 2.7106 < 2.8073 H_0 not rejected → The means of the yield are equal. $t_{0.05/2}$=2.0687 2.7106 > 2.0687 H_0 rejected → The means of the yield are different.	

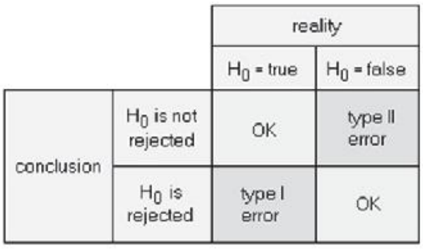
```
t-statistic: -2.7106
p(t), 1-sided: 0.0062
pooled stdv.: 22.0434
deg. freedom: 23

critical t-values:
```

α	t(α)
0.001	3.4850
0.002	3.1978
0.005	**2.8073**
0.01	2.4999
0.02	2.1770
0.025	**2.0687**
0.05	1.7139

Conclusion:

With an α of 1%, it is recognizable that there is a difference between the two herds. With an α of 5%, there is no difference visible.



		reality	
		H_0 = true	H_0 = false
conclusion	H_0 is not rejected	OK	type II error
	H_0 is rejected	type I error	OK

Figure 1: Error types

The changes made are written in the colour green to make them easier to follow.

Problem 1: Statistical Tests

Report

Problem description:

We had four different biofilms they are used to recut new serpulid worms. We should check if there was a effect of the various biofilms on the recruitment of the worms.

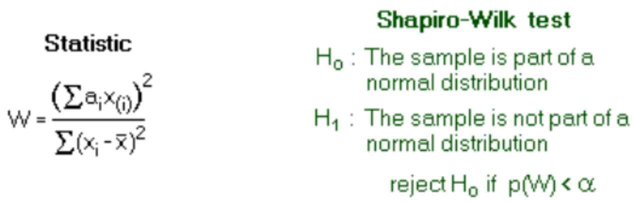
Outcome:

The differences between the means of Populations, in this case the worms, are reflected in the Variances of the samples. Thus, the analysis of variances (ANOVA) can make statements about the means. When performing an ANOVA the following basic assumptions have to be met:

- The data has to be normally distributed → Shapiro-Wilk test
- The variances must be equal for all samples → Levene's-Test

Results of the normality distribution:

Toolbar "Math" → Tests… → Colum 1 & 2 "marking the data" red/blue → Shapiro-Wilk



Statistic

$$W = \frac{(\sum a_i x_{(i)})^2}{\sum (x_i - \bar{x})^2}$$

Shapiro-Wilk test

H_0 : The sample is part of a normal distribution

H_1 : The sample is not part of a normal distribution

reject H_0 if $p(W) < \alpha$

Biofilm SL	Biofilm UL	Biofilm NL	Biofilm NF
Data of Block A (red): ----------------- No. of Data: 7 Mean: 95.2857 Std. Dev.: 24.23 W-statistic: 0.9142 p(W): **0.4258**	Data of Block B (blue): ----------------- No. of Data: 7 Mean: 141.0000 Std. Dev.: 38.05 W-statistic: 0.9379 p(W): **0.6199**	Data of Block A (red): ----------------- No. of Data: 7 Mean: 154.7143 Std. Dev.: 30.40 W-statistic: 0.9153 p(W): **0.4338**	Data of Block B (blue): ----------------- No. of Data: 7 Mean: 133.2857 Std. Dev.: 35.48 W-statistic: 0.9750 p(W): **0.9318**
reject H_0 if $p(W) < \alpha$			
$\alpha=0.01$ $0.4258 > 0.01$ H_0 not rejected → Normal distribution $\alpha=0.05$ $0.4258 > 0.05$ H_0 not rejected → Normal distribution	$\alpha=0.01$ $0.6199 > 0.01$ H_0 not rejected → Normal distribution $\alpha=0.05$ $0.6199 > 0.05$ H_0 not rejected → Normal distribution	$\alpha=0.01$ $0.4338 > 0.01$ H_0 not rejected → Normal distribution $\alpha=0.05$ $0.4338 > 0.05$ H_0 not rejected → Normal distribution	$\alpha=0.01$ $0.9318 > 0.01$ H_0 not rejected → Normal distribution $\alpha=0.05$ $0.9318 > 0.05$ H_0 not rejected → Normal distribution

Results of the Levene's Test (based on the absolute differences):

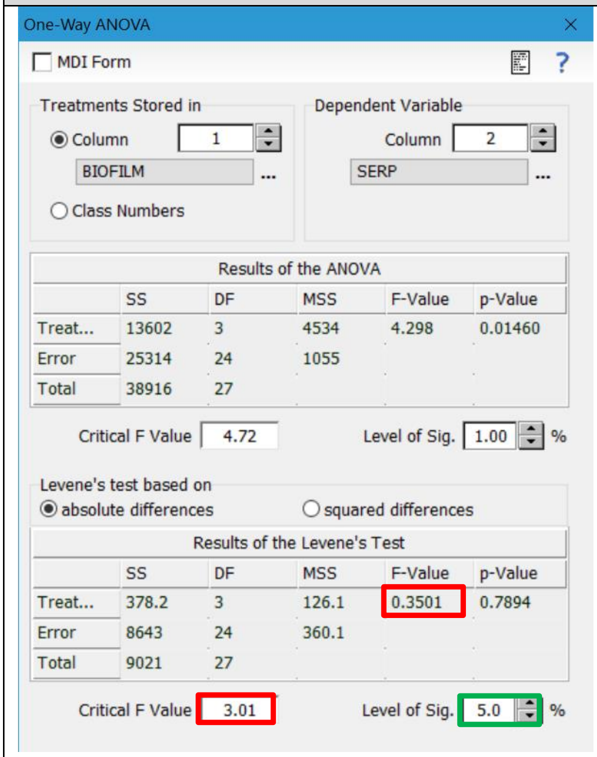
Toolbar "Math" → ANNOVA → One Factor… → Set correct columns → Change "Level of Sig." to the needed values

$$L = \frac{\frac{1}{k-1}\sum_{j=1}^{k} n_j(\bar{Y}_j - \bar{Y})^2}{\frac{1}{n-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ji} - \bar{Y}_j)^2}$$

Nullhypothese: $H_0: \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$

Alternativhypothese: $H_1: \sigma_i^2 \neq \sigma_j^2$ für *mindestens ein* Gruppenpaar i, j mit $i \neq j$

α=1%



One-Way ANOVA

□ MDI Form

Treatments Stored in: ● Column 1, BIOFILM; ○ Class Numbers

Dependent Variable: Column 2, SERP

Results of the ANOVA

	SS	DF	MSS	F-Value	p-Value
Treat...	13602	3	4534	4.298	0.01460
Error	25314	24	1055		
Total	38916	27			

Critical F Value 4.72 Level of Sig. 1.00 %

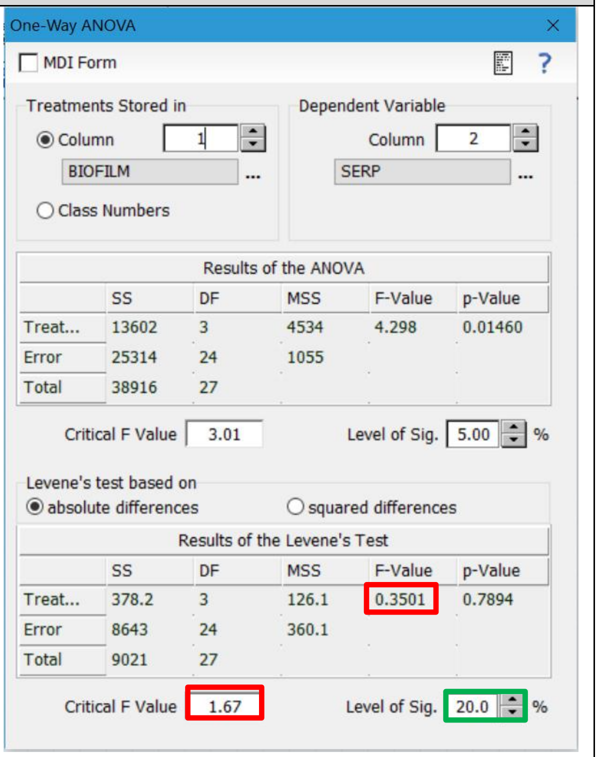
Levene's test based on ● absolute differences ○ squared differences

Results of the Levene's Test

	SS	DF	MSS	F-Value	p-Value
Treat...	378.2	3	126.1	0.3501	0.7894
Error	8643	24	360.1		
Total	9021	27			

Critical F Value 3.01 Level of Sig. 5.0 %

α=5%



One-Way ANOVA

□ MDI Form

Treatments Stored in: ● Column 1, BIOFILM; ○ Class Numbers

Dependent Variable: Column 2, SERP

Results of the ANOVA

	SS	DF	MSS	F-Value	p-Value
Treat...	13602	3	4534	4.298	0.01460
Error	25314	24	1055		
Total	38916	27			

Critical F Value 3.01 Level of Sig. 5.00 %

Levene's test based on ● absolute differences ○ squared differences

Results of the Levene's Test

	SS	DF	MSS	F-Value	p-Value
Treat...	378.2	3	126.1	0.3501	0.7894
Error	8643	24	360.1		
Total	9021	27			

Critical F Value 1.67 Level of Sig. 20.0 %

reject $H_0: \sigma_1^2 = \ldots = \sigma_k^2$

α=1%	α=5%
Level of significance (5*1% = 5%) 0.3501 < 3.01 H_0 not rejected → equal variance	Level of significance (4*5% = 20%) 0.3501 < 1.67 H_0 not rejected → equal variance

Results of the One-way ANOVA:

Toolbar “Math” → ANNOVA → One Factor… → Set correct columns → Change “Level of Sig.” to the needed values

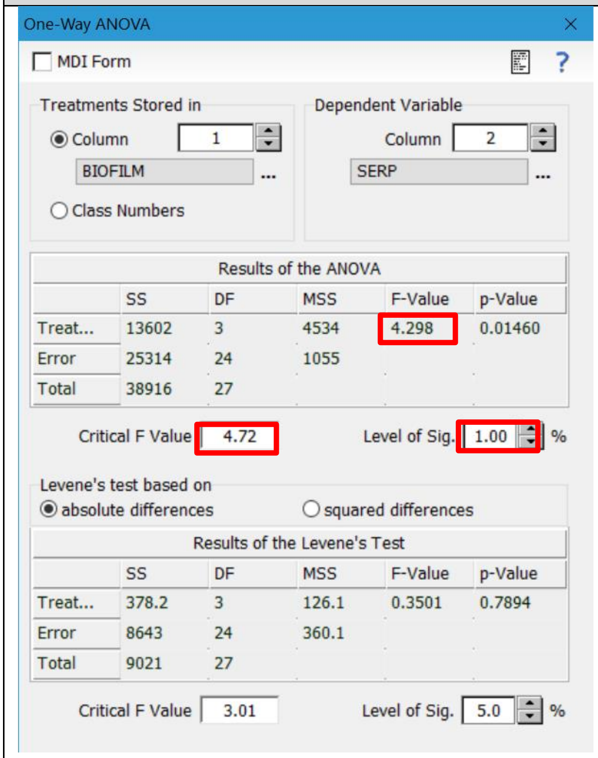
$$F \equiv \frac{MQA}{MQR} = \frac{\sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2/(k-1)}{\sum_{i=1}^{k}(n_i-1)s_i^2/(N-k)}.$$

$$H_0\colon \mu_1 = \mu_2 = \ldots = \mu_k$$

Die Alternativhypothese lautet:

$$H_1\colon \exists i,j:\ \mu_i \neq \mu_j$$

α=1%



One-Way ANOVA

□ MDI Form

Treatments Stored in
● Column 1
BIOFILM
○ Class Numbers

Dependent Variable
Column 2
SERP

Results of the ANOVA

	SS	DF	MSS	F-Value	p-Value
Treat...	13602	3	4534	4.298	0.01460
Error	25314	24	1055		
Total	38916	27			

Critical F Value 4.72 Level of Sig. 1.00 %

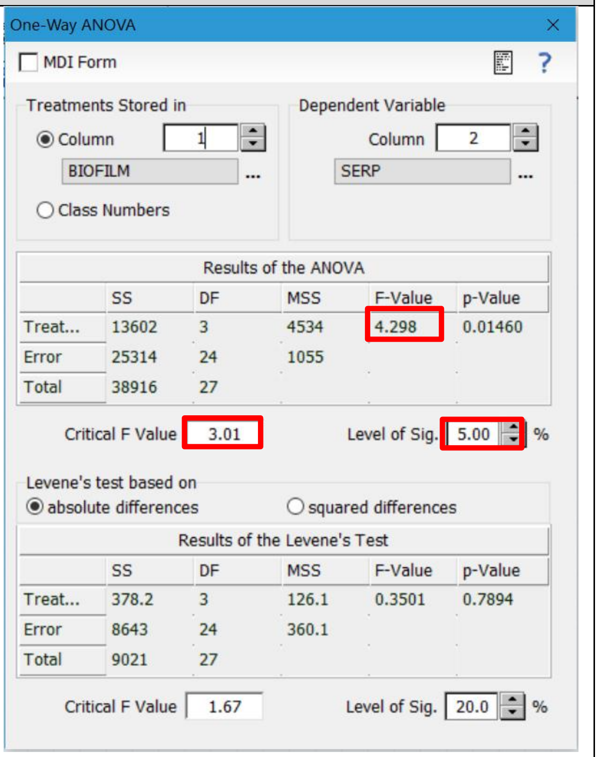
Levene's test based on
● absolute differences ○ squared differences

Results of the Levene's Test

	SS	DF	MSS	F-Value	p-Value
Treat...	378.2	3	126.1	0.3501	0.7894
Error	8643	24	360.1		
Total	9021	27			

Critical F Value 3.01 Level of Sig. 5.0 %

α=5%



One-Way ANOVA

□ MDI Form

Treatments Stored in
● Column 1
BIOFILM
○ Class Numbers

Dependent Variable
Column 2
SERP

Results of the ANOVA

	SS	DF	MSS	F-Value	p-Value
Treat...	13602	3	4534	4.298	0.01460
Error	25314	24	1055		
Total	38916	27			

Critical F Value 3.01 Level of Sig. 5.00 %

Levene's test based on
● absolute differences ○ squared differences

Results of the Levene's Test

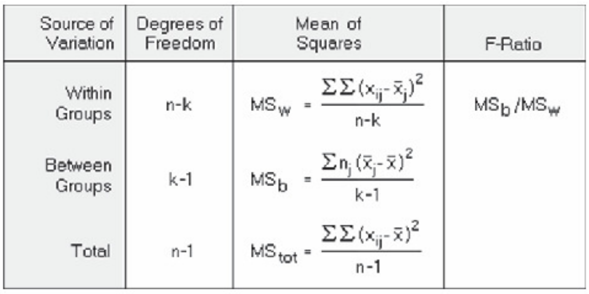
	SS	DF	MSS	F-Value	p-Value
Treat...	378.2	3	126.1	0.3501	0.7894
Error	8643	24	360.1		
Total	9021	27			

Critical F Value 1.67 Level of Sig. 20.0 %

reject H_0:μ_1=…= μ_k	
α=1% 4.298 < 4.72 H_0 not rejected → equal means	α=5% 4.298 > 3.01 H_0 rejected → difference of the means

Conclusion:

When the F-value is greater than the critical F-value there is a significant difference between the means. For the 5% level of significance this is the case. The 1% level of significance becomes no time rejected.



Source of Variation	Degrees of Freedom	Mean of Squares	F-Ratio
Within Groups	n-k	$MS_w = \frac{\Sigma\Sigma(x_{ij}-\bar{x}_j)^2}{n-k}$	MS_b/MS_w
Between Groups	k-1	$MS_b = \frac{\Sigma n_j(\bar{x}_j-\bar{x})^2}{k-1}$	
Total	n-1	$MS_{tot} = \frac{\Sigma\Sigma(x_{ij}-\bar{x})^2}{n-1}$	

Figure 1: Analysis of Variance

Problem 3: Linear Regression Report

Problem description:

We have a size of n=86 data, 43 for the weight increase [kg/day] and 43 protein intake [g/kg LWT/day]. We want to show in a linear model which function (line, parabolic, logarithmic) for the model fitted best and describe why it is the best. Also, the increase of the daily protein intake gets increase and we should look how the used model change.

Outcome:

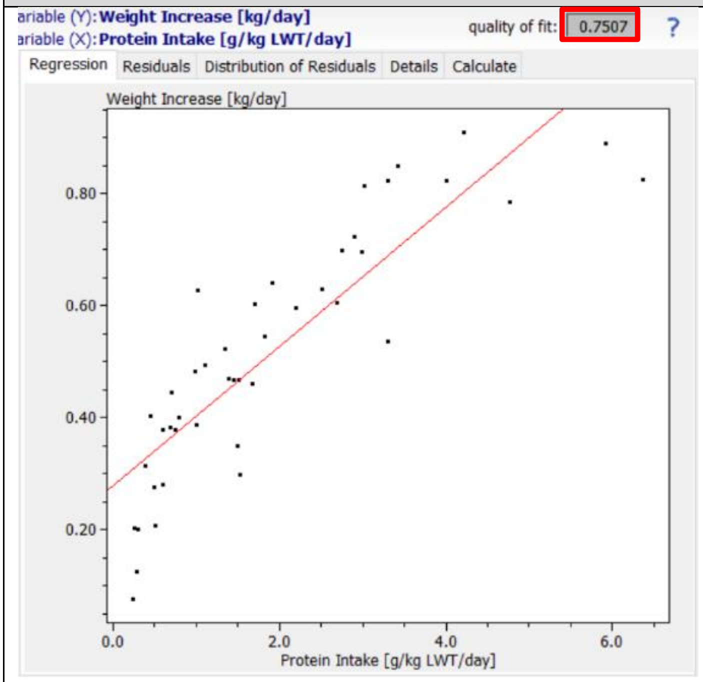
The simplest type of analysis, for very small data sets, is a simple regression. To make statements about the data, it is advisable to create a graph, as it is easy to create and already contains a lot of information in a visual form. However, before the data can be analysed at all, it is necessary to check that the following conditions are met:

- All measurements are independent of each other
- For each X the Y-values are normally distributed
- For each X, the Y-distribution has the same variance

Which of the proposed models would you select as the best one?

Toolbar "Math" → Simple regression

Linear Regression

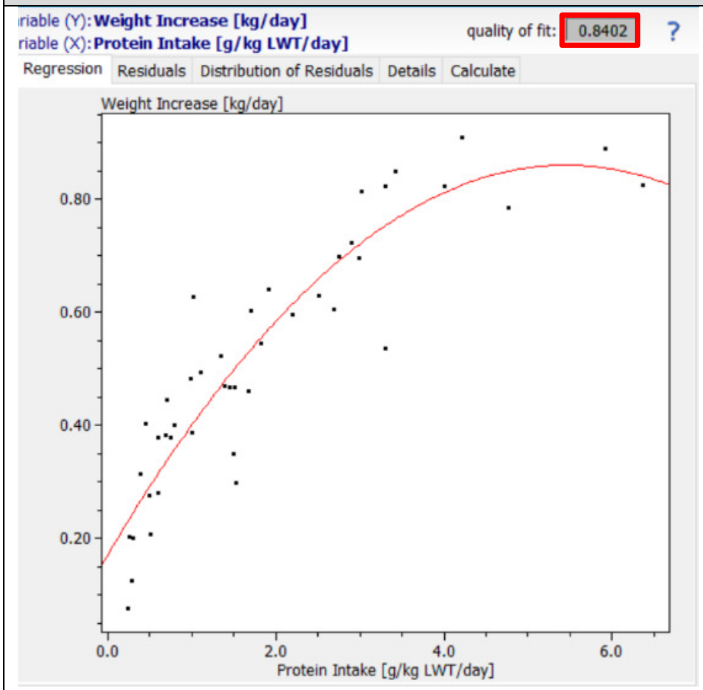


ariable (Y):**Weight Increase [kg/day]**
ariable (X):**Protein Intake [g/kg LWT/day]**
quality of fit: 0.7507

```
-----------------------------
Source          F      quality
                       of fit
-----------------------------
Regression   123.45    0.7507

Function:
y = k*x + d
k =  1.24244E-01
d =  2.78507E-01
```

Parabolic Regression



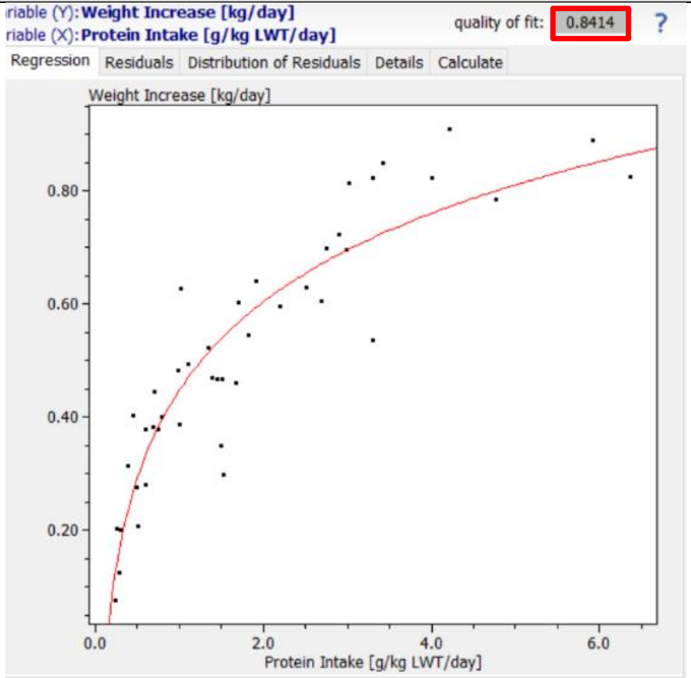
riable (Y): **Weight Increase [kg/day]**
riable (X): **Protein Intake [g/kg LWT/day]**

quality of fit: 0.8402

```
-----------------------------
Source          F       quality
                        of fit
-----------------------------
Regression    105.18     0.7507

Function:
y = k0 + k1*x + k2*x^2
k0 =  1.70404E-01
k1 =  2.53164E-01
k2 = -2.32185E-02
```

Logarithmic Regression



riable (Y): **Weight Increase [kg/day]**
riable (X): **Protein Intake [g/kg LWT/day]**

quality of fit: 0.8414

```
-----------------------------
Source          F       quality
                        of fit
-----------------------------
Regression    217.46     0.7507

Function:
y = k0 + k1*ln(x)
k0 =  4.47135E-01
k1 =  2.25901E-01
```

When we are looking through the graphs, we see that the logarithmic regression shows the highest fit to the data and consider all the data. Also, by checking of the F-value (information about reliable of the model) and quality of fit (R²) we see that the logarithmic evaluation has the highest F-value and quality of fit.

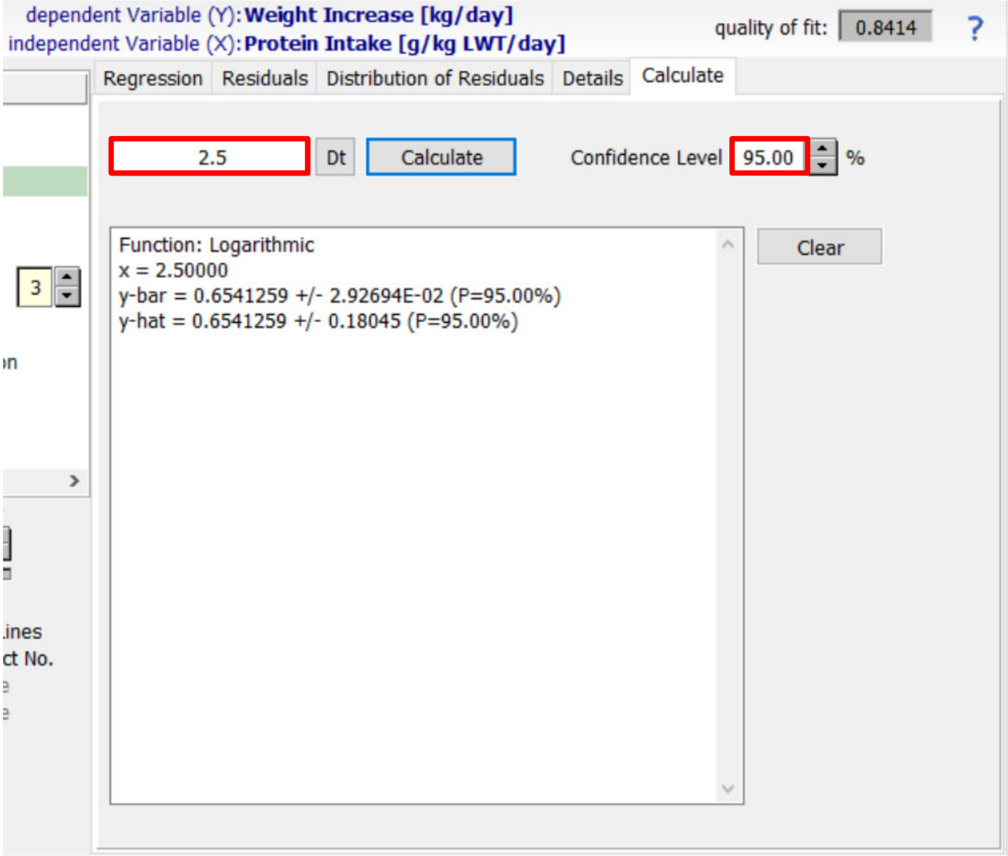
The conclusion is that this logarithmic model is the best one for this data.

If you look at the parabolic model: what are the arguments against it of in favor of it?

If we look at the curve shape, we see that with increasing protein supply at 5.0 g/kg, weight decreases. This assumption cannot be correct from a biological and physiological point of view. The F-value and the quality of fit are both very high and would therefore also reflect a good agreement with the data. However, the slope of the curve from 5.0 g/kg onwards excludes this model as ideal.

Using the selected "best" model: assuming that the cattle's daily intake of additional protein is 2.5 g/kg, what is the expected weight increase per day? Do not forget to specify the confidence interval of your estimate (95%).

Toolbar “Math” → Simple regression → Registry card “Calculate” → input Dt → Calculate



dependent Variable (Y): **Weight Increase [kg/day]**
independent Variable (X): **Protein Intake [g/kg LWT/day]**
quality of fit: 0.8414

Dt: 2.5
Confidence Level: 95.00 %

```
Function: Logarithmic
x = 2.50000
y-bar = 0.6541259 +/- 2.92694E-02 (P=95.00%)
y-hat = 0.6541259 +/- 0.18045 (P=95.00%)
```

We take as best model the logarithmic once. By looking at y-hat, used for single values other ones for population values, we get for the expected weight increase per day:

```
y [kg/day] = 0.6541259 ∓ 0.18045
```

Problem 4: Multiple Regression Report

Problem description:

We have a data set of n=2436 values. This is divided into 14 columns, each containing 174 measurements of different body proportions (e.g. hip circumference, neck diameter, etc.). The analysis of this data set is to determine which of these measurements give an indication of the body fat percentage.

Outcome:

Multiple linear regression differs from linear regression in that more than one input variable (xi) is used. The assumptions and prerequisites are the same as for simple regression. If you want to check whether there is a correlation between the values of several variables, you can do a regression analysis. The correlation is tested by means of a comparison. A method based on this procedure is stepwise regression.

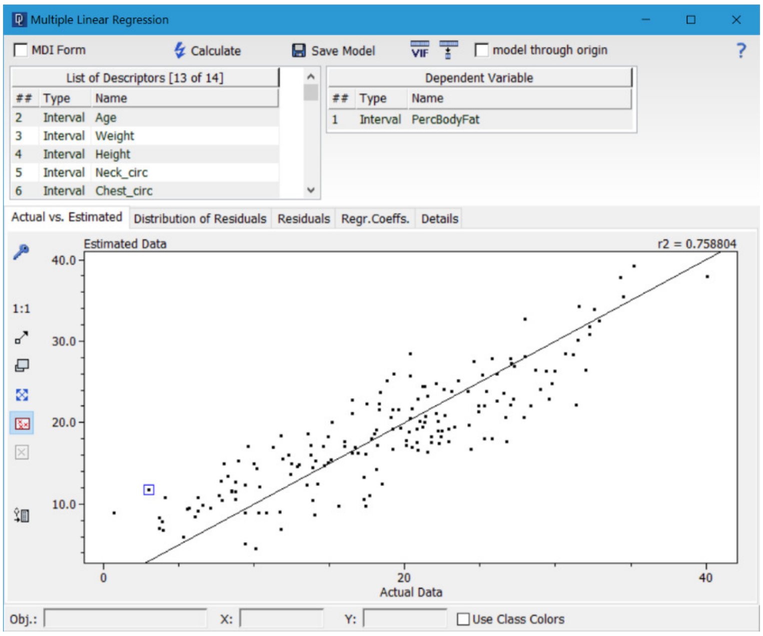
Calculate an MLR model using all variables (model 1) and check which of the variables contribute significantly to the model (at the 5% level of significance).

Toolbar MLR → select “Dependent Variable” and “List of Descriptors” → Calculate → Registry card “Details”

```
------------------------------------------------------------
ANOVA          DF     sum of squares     mean square       F
------------------------------------------------------------
Regression     13        8.44996E+03     6.49997E+02    38.720
Residual      160        2.68593E+03     1.67870E+01
Total         173        1.11359E+04
------------------------------------------------------------
```

Level of significance 5% = 0.05 **α** < 0.05

```
Regression coefficients:
Col  Var-Name          Coefficient +/- Std.Err.(coeff)     t-Test      alpha
 -----------------------------------------------------------------------
-    INTERCEPT        -9.9223942E+00 +/- 1.9678816E+01     -0.504     0.6148
2    Age               1.0548101E-01 +/- 3.6638422E-02      2.879     0.0045
3    Weight           -1.6210032E-01 +/- 1.3011933E-01     -1.246     0.2147
4    Height           -3.0286819E-02 +/- 3.8723038E-02     -0.782     0.4353
5    Neck_circ        -5.6864228E-01 +/- 2.5083560E-01     -2.267     0.0247
6    Chest_circ        3.6349376E-02 +/- 1.1441980E-01      0.318     0.7511
7    Abdomen_circ      9.0856834E-01 +/- 9.7672758E-02      9.302     0.0000
8    Hip_circ         -3.2827084E-01 +/- 1.6733579E-01     -1.962     0.0515
9     Thigh_circ       3.1089535E-01 +/- 1.6819169E-01      1.848     0.0664
10   Knee_circ        -7.3027302E-02 +/- 2.9299244E-01     -0.249     0.8035
11   Ankle_circ        3.2688615E-01 +/- 2.2782441E-01      1.435     0.1533
12   Biceps_circ       1.2324685E-01 +/- 1.9056066E-01      0.647     0.5187
13   Forearm_circ      8.0422926E-01 +/- 2.7056556E-01      2.972     0.0034
14   Wrist_circ       -2.2151694E+00 +/- 6.3644439E-01     -3.481     0.0006
```



List of Descriptors [13 of 14]

##	Type	Name
2	Interval	Age
3	Interval	Weight
4	Interval	Height
5	Interval	Neck_circ
6	Interval	Chest_circ

Dependent Variable

##	Type	Name
1	Interval	PercBodyFat

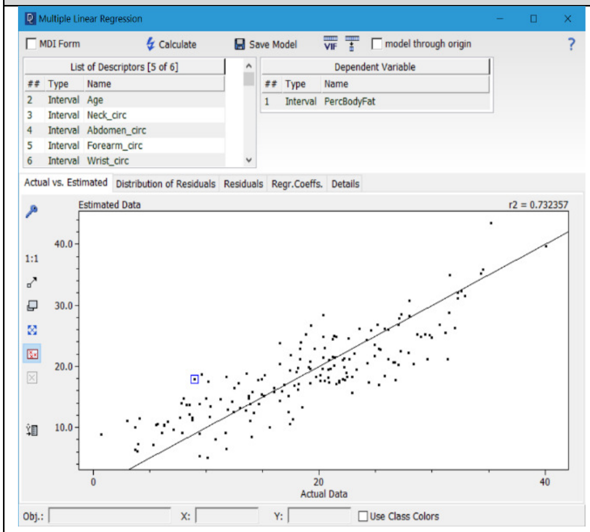
Figure 1: Multiple linear regression whit all variables (= Model 1)

Discard all non-significant variables and recalculate the model (model2). Compare this model with a model obtained by stepwise regression (model 3).

Window “Editor” → right click “Delete Colum” → Toolbar MLR → select “Dependent Variable” and “List of Descriptors” → Calculate → Registry card “Details”

Toolbar “Math” → Multiple Regression → Multiple Linear Regression → Variable Selection… → Chose by Selection Mode “Stepwise Regression” → Chose variables in “List of Variables” Incl./Target → Start → MLR → Calculate → Registry card “Details”

Model 2


Multiple Linear Regression

□ MDI Form Calculate Save Model □ model through origin

List of Descriptors [5 of 6]

##	Type	Name
2	Interval	Age
3	Interval	Neck_circ
4	Interval	Abdomen_circ
5	Interval	Forearm_circ
6	Interval	Wrist_circ

Dependent Variable

##	Type	Name
1	Interval	PercBodyFat

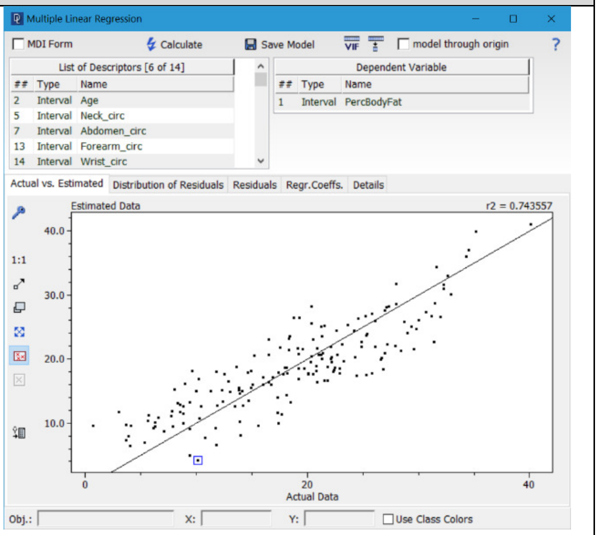
Actual vs. Estimated | Distribution of Residuals | Residuals | Regr.Coeffs. | Details

Obj.: X: Y: □ Use Class Colors

```
----------------------------
ANOVA           DF          F
----------------------------
Regression       5     91.941
Residual       168
Total          173
----------------------------


Regression coefficients:
Col  Var-Name          alpha
---------------------------
-    INTERCEPT        0.4650
2    Age              0.0000
3    Neck_circ        0.0039
4    Abdomen_circ     0.0000
5    Forearm_circ     0.0003
6    Wrist_circ       0.0000
```

Model 3


Multiple Linear Regression

□ MDI Form Calculate Save Model □ model through origin

List of Descriptors [6 of 14]

##	Type	Name
2	Interval	Age
5	Interval	Neck_circ
7	Interval	Abdomen_circ
13	Interval	Forearm_circ
14	Interval	Wrist_circ

Dependent Variable

##	Type	Name
1	Interval	PercBodyFat

Actual vs. Estimated | Distribution of Residuals | Residuals | Regr.Coeffs. | Details

Obj.: X: Y: □ Use Class Colors

```
------------------------------
ANOVA           DF            F
------------------------------
Regression       6       80.703
Residual       167
Total          173
------------------------------


Regression coefficients:
Col  Var-Name          alpha
---------------------------
-    INTERCEPT        0.0146
2    Age              0.0057
5    Neck_circ        0.0432
7    Abdomen_circ     0.0000
13   Forearm_circ     0.0001
14   Wrist_circ       0.0007
3    Weight           0.0076
```

Explain the differences between model 2 and model 3. Which of the two models is better? Explain why.

In Model 2, only the alpha value is used for exclusion. Those variables that are smaller than the α= 5% value are taken into account for the calculation of the new multiple linear regression. It should be noted that all variables with a high alpha value are excluded immediately; there is no stepwise exclusion with a new check of the alpha value.

In Model 3, a stepwise regression is carried out on the basis of the preselection of Model 2, in which the programme adds or removes variables until it finds a model that best fits the specified criterion. DataLab looks at the parameters:

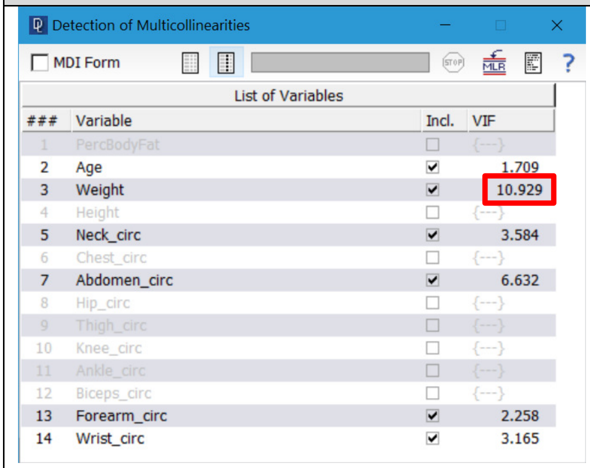
- smallest absolute t-statistic of the coefficients of the model
- Akaike Information Criterion
- Bayes Information Criterion
- F statistic obtained from the ANOVA of the model (F-value)
- goodness of fit of the model (R^2)

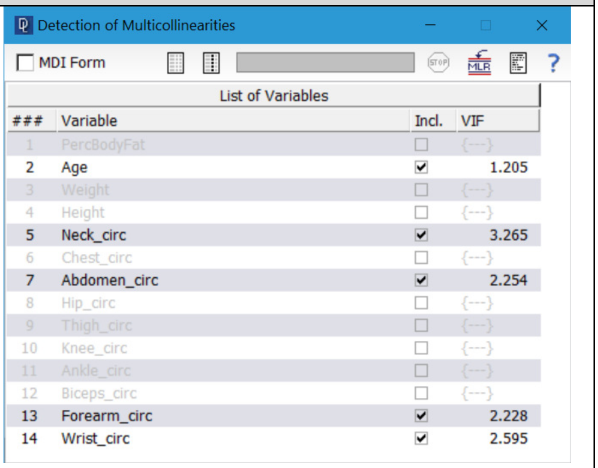
	Model 1	Model 2	Model 3
F-value	38.720	91.941	80.703
biggest alpha	-	0.0039	0.0432

Model 2 and 3 both have an alpha value less than 5%, so this is not very decisive for my assessment. When we are looking at the F-value we see that model 2 has the highest one. By looking at the used variables two of them are interesting. The Weight and the Agee. As a medically trained person, I know that weight is not related to body fat - for example, muscle mass has a greater dead weight than fat cells. Age is a realistic variable because as the body ages, metabolic processes decrease and more "energy" is stored. The model 2 has the wight excluded. So, my personal opinion is to use model 2 as the best one.

Check for multicollinearities in model 3.

Toolbar "Math" → Multiple Regression → Multiple Linear Regression → Variable Selection… → Chose by Selection Mode "Stepwise Regression" → Chose variables in "List of Variables" Incl./Target → Start → MLR → Calculate → VIF

<table>
<tr><th>Model 3</th><th>Model 2</th></tr>
<tr><td>

<table>
<tr><th>###</th><th>Variable</th><th>Incl.</th><th>VIF</th></tr>
<tr><td>1</td><td>PercBodyFat</td><td>□</td><td>{---}</td></tr>
<tr><td>2</td><td>Age</td><td>☑</td><td>1.709</td></tr>
<tr><td>3</td><td>Weight</td><td>☑</td><td>10.929</td></tr>
<tr><td>4</td><td>Height</td><td>□</td><td>{---}</td></tr>
<tr><td>5</td><td>Neck_circ</td><td>☑</td><td>3.584</td></tr>
<tr><td>6</td><td>Chest_circ</td><td>□</td><td>{---}</td></tr>
<tr><td>7</td><td>Abdomen_circ</td><td>☑</td><td>6.632</td></tr>
<tr><td>8</td><td>Hip_circ</td><td>□</td><td>{---}</td></tr>
<tr><td>9</td><td>Thigh_circ</td><td>□</td><td>{---}</td></tr>
<tr><td>10</td><td>Knee_circ</td><td>□</td><td>{---}</td></tr>
<tr><td>11</td><td>Ankle_circ</td><td>□</td><td>{---}</td></tr>
<tr><td>12</td><td>Biceps_circ</td><td>□</td><td>{---}</td></tr>
<tr><td>13</td><td>Forearm_circ</td><td>☑</td><td>2.258</td></tr>
<tr><td>14</td><td>Wrist_circ</td><td>☑</td><td>3.165</td></tr>
</table></td>
<td>

<table>
<tr><th>###</th><th>Variable</th><th>Incl.</th><th>VIF</th></tr>
<tr><td>1</td><td>PercBodyFat</td><td>□</td><td>{---}</td></tr>
<tr><td>2</td><td>Age</td><td>☑</td><td>1.205</td></tr>
<tr><td>3</td><td>Weight</td><td>□</td><td>{---}</td></tr>
<tr><td>4</td><td>Height</td><td>□</td><td>{---}</td></tr>
<tr><td>5</td><td>Neck_circ</td><td>☑</td><td>3.265</td></tr>
<tr><td>6</td><td>Chest_circ</td><td>□</td><td>{---}</td></tr>
<tr><td>7</td><td>Abdomen_circ</td><td>☑</td><td>2.254</td></tr>
<tr><td>8</td><td>Hip_circ</td><td>□</td><td>{---}</td></tr>
<tr><td>9</td><td>Thigh_circ</td><td>□</td><td>{---}</td></tr>
<tr><td>10</td><td>Knee_circ</td><td>□</td><td>{---}</td></tr>
<tr><td>11</td><td>Ankle_circ</td><td>□</td><td>{---}</td></tr>
<tr><td>12</td><td>Biceps_circ</td><td>□</td><td>{---}</td></tr>
<tr><td>13</td><td>Forearm_circ</td><td>☑</td><td>2.228</td></tr>
<tr><td>14</td><td>Wrist_circ</td><td>☑</td><td>2.595</td></tr>
</table></td></tr>
<tr><td colspan="2">VIF > 10 → collinearity between the variables</td></tr>
<tr><td>VIF > 10
Weight 10.929 > 10
Collinearity between the variables is given</td><td>VIF > 10
all VIF < 10
No collinearity is given</td></tr>
</table>

If a model is based on highly correlated variables, the estimated regression coefficients become unstable. This renders the coefficients useless for causal interpretation.

Do you think that person "R42" who exhibits an unusual low body height has any influence on model 3?

Toolbar MLR → select "Dependent Variable" and "List of Descriptors" → Calculate → Registry card "Details"

```
Regression results:
Obj-# Name    Target Value    Regr.Result     Difference       Cook's D
 -----------------------------------------------------------------------
42 R-42        3.29000E+01    3.25366E+01       -0.363          1.098E-02
```

Data points with large residuals (outliers) can affect the result and precision of a regression. Cook distance measures the effect of missing out a given observation. Data points with a large Cook distance should be looked at more closely during data analysis.

It is a visible outlier in residuals but does not have a high value in Cook's distance. It has no influence on the model.

The changes made are written in the colour green to make them easier to follow.

Problem 5: PCA of Breast Cancer Data
Report

Problem description:

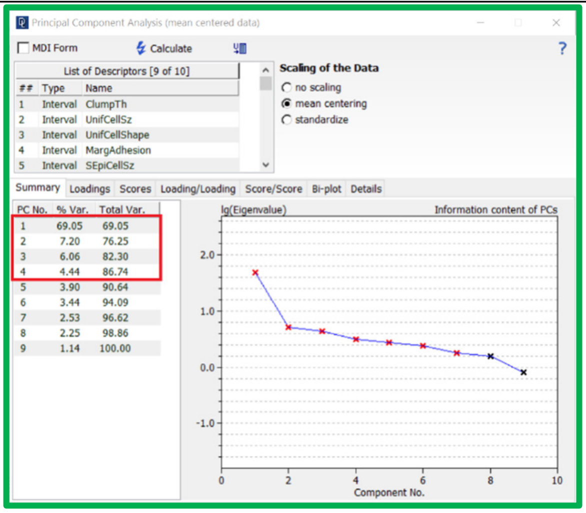
We have a data set of n=6830 values. This is divided into 10 columns, each containing 683 samples of breast tumor mass. The analysis of the data should show that it is possible to make statements about the mortality probability of tumours using statistical evaluation methods.

Outcome:

The central idea behind principal component analysis is to project the high-dimensional data space onto a two-dimensional plane in way that any interesting features of the data will become visible.

How many principle components are sufficient to describe the dataset?

Toolbar PCA → select “List of Descriptors” and “Scaling of the data” → Calculate → Registry card “Summary” → Registry card “Score/Score” → chose PC No.

<table>
<tr><th>Mean centering</th><th>Standardize</th></tr>
<tr>
<td>


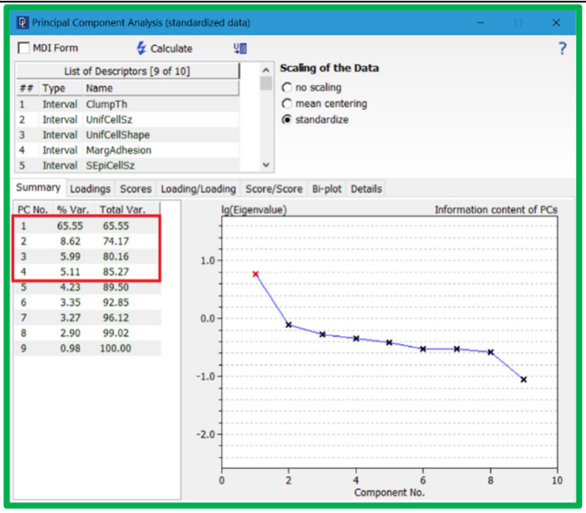
Principal Component Analysis (mean centered data)

□ MDI Form Calculate ?

<table>
<tr><th colspan="3">List of Descriptors [9 of 10]</th></tr>
<tr><th>##</th><th>Type</th><th>Name</th></tr>
<tr><td>1</td><td>Interval</td><td>ClumpTh</td></tr>
<tr><td>2</td><td>Interval</td><td>UnifCellSz</td></tr>
<tr><td>3</td><td>Interval</td><td>UnifCellShape</td></tr>
<tr><td>4</td><td>Interval</td><td>MargAdhesion</td></tr>
<tr><td>5</td><td>Interval</td><td>SEpiCellSz</td></tr>
</table>

Scaling of the Data
○ no scaling
● mean centering
○ standardize

Summary | Loadings | Scores | Loading/Loading | Score/Score | Bi-plot | Details

<table>
<tr><th>PC No.</th><th>% Var.</th><th>Total Var.</th></tr>
<tr><td>1</td><td>69.05</td><td>69.05</td></tr>
<tr><td>2</td><td>7.20</td><td>76.25</td></tr>
<tr><td>3</td><td>6.06</td><td>82.30</td></tr>
<tr><td>4</td><td>4.44</td><td>86.74</td></tr>
<tr><td>5</td><td>3.90</td><td>90.64</td></tr>
<tr><td>6</td><td>3.44</td><td>94.09</td></tr>
<tr><td>7</td><td>2.53</td><td>96.62</td></tr>
<tr><td>8</td><td>2.25</td><td>98.86</td></tr>
<tr><td>9</td><td>1.14</td><td>100.00</td></tr>
</table>
</td>
<td>


Principal Component Analysis (standardized data)

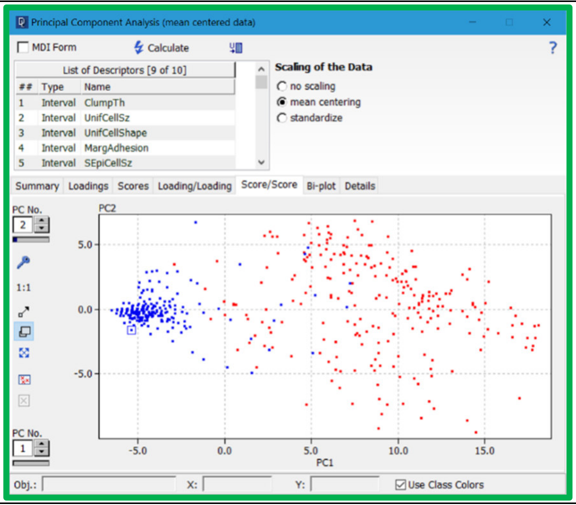
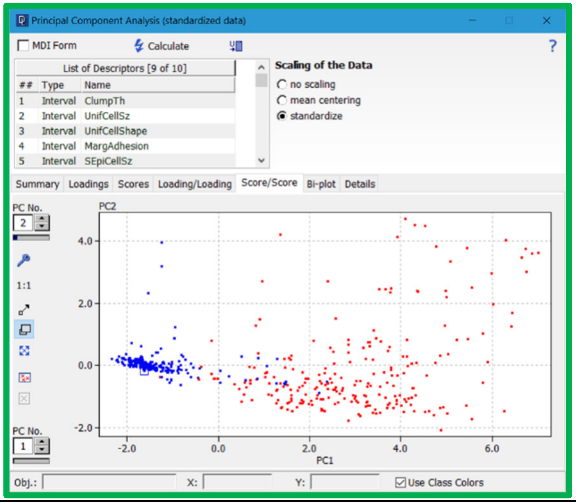
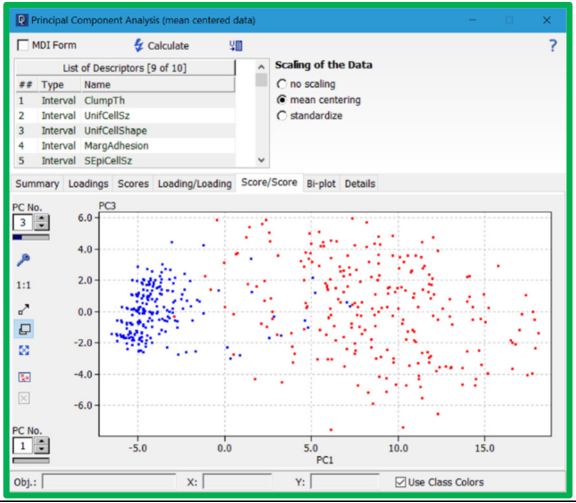
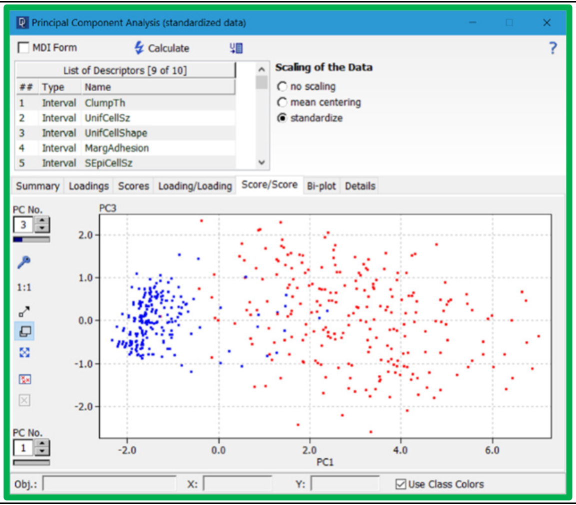
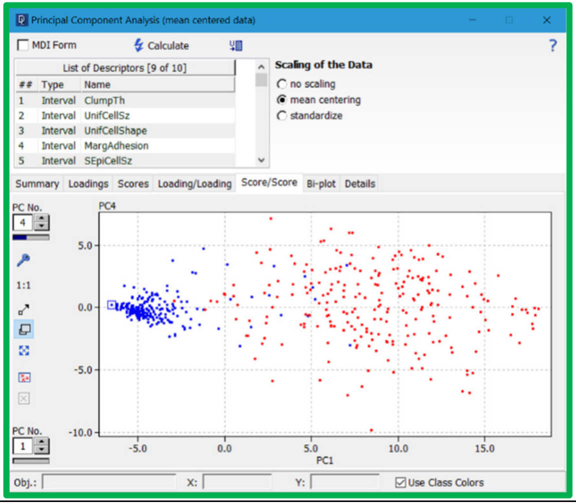
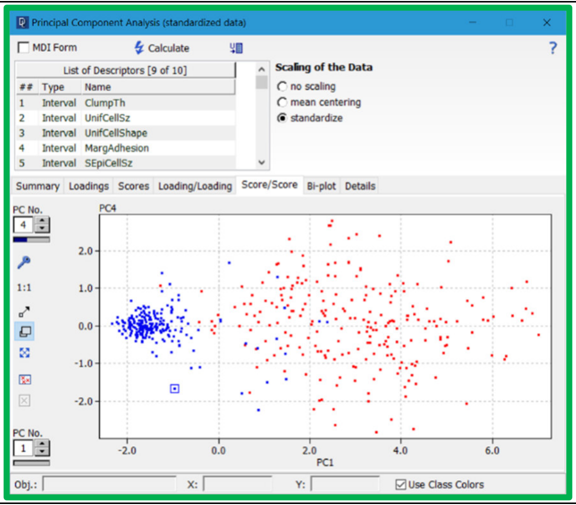
□ MDI Form Calculate ?

<table>
<tr><th colspan="3">List of Descriptors [9 of 10]</th></tr>
<tr><th>##</th><th>Type</th><th>Name</th></tr>
<tr><td>1</td><td>Interval</td><td>ClumpTh</td></tr>
<tr><td>2</td><td>Interval</td><td>UnifCellSz</td></tr>
<tr><td>3</td><td>Interval</td><td>UnifCellShape</td></tr>
<tr><td>4</td><td>Interval</td><td>MargAdhesion</td></tr>
<tr><td>5</td><td>Interval</td><td>SEpiCellSz</td></tr>
</table>

Scaling of the Data
○ no scaling
○ mean centering
● standardize

Summary | Loadings | Scores | Loading/Loading | Score/Score | Bi-plot | Details

<table>
<tr><th>PC No.</th><th>% Var.</th><th>Total Var.</th></tr>
<tr><td>1</td><td>65.55</td><td>65.55</td></tr>
<tr><td>2</td><td>8.62</td><td>74.17</td></tr>
<tr><td>3</td><td>5.99</td><td>80.16</td></tr>
<tr><td>4</td><td>5.11</td><td>85.27</td></tr>
<tr><td>5</td><td>4.23</td><td>89.50</td></tr>
<tr><td>6</td><td>3.35</td><td>92.85</td></tr>
<tr><td>7</td><td>3.27</td><td>96.12</td></tr>
<tr><td>8</td><td>2.90</td><td>99.02</td></tr>
<tr><td>9</td><td>0.98</td><td>100.00</td></tr>
</table>
</td>
</tr>
<tr>
<td>To covered 85% of the total variance we need the first 4 components. The first component contains alone 69.05% of all variations. In summery we have a total variation of 86.74%.</td>
<td>To covered 85% of the total variance we need the first 4 components. The first component contains alone 65.55% of all variations. In summary we have a total variation of 85.27%.</td>
</tr>
</table>

Mean centering	Standardize
Blue Dots: Benign* Red Dots: Malignant*	
 Principal Component Analysis (mean centered data) □ MDI Form Calculate List of Descriptors [9 of 10] ## Type Name 1 Interval ClumpTh 2 Interval UnifCellSz 3 Interval UnifCellShape 4 Interval MargAdhesion 5 Interval SEpiCellSz Scaling of the Data ○ no scaling ● mean centering ○ standardize PC No. 2 PC No. 1 Obj.: X: Y: ☑ Use Class Colors Summary Loadings Scores Loading/Loading Score/Score Bi-plot Details PC2 5.0 0.0 -5.0 -5.0 0.0 5.0 10.0 15.0 PC1	 Principal Component Analysis (standardized data) □ MDI Form Calculate List of Descriptors [9 of 10] ## Type Name 1 Interval ClumpTh 2 Interval UnifCellSz 3 Interval UnifCellShape 4 Interval MargAdhesion 5 Interval SEpiCellSz Scaling of the Data ○ no scaling ○ mean centering ● standardize PC No. 2 PC No. 1 Obj.: X: Y: ☑ Use Class Colors Summary Loadings Scores Loading/Loading Score/Score Bi-plot Details PC2 4.0 2.0 0.0 -2.0 -2.0 0.0 2.0 4.0 6.0 PC1
Score/Score plot of PC 1 and 2 of mean centering.	Score/Score plot of PC 1 and 2 of standardize.
 Principal Component Analysis (mean centered data) □ MDI Form Calculate List of Descriptors [9 of 10] ## Type Name 1 Interval ClumpTh 2 Interval UnifCellSz 3 Interval UnifCellShape 4 Interval MargAdhesion 5 Interval SEpiCellSz Scaling of the Data ○ no scaling ● mean centering ○ standardize PC No. 3 PC No. 1 Obj.: X: Y: ☑ Use Class Colors Summary Loadings Scores Loading/Loading Score/Score Bi-plot Details PC3 6.0 4.0 2.0 0.0 -2.0 -4.0 -6.0 -5.0 0.0 5.0 10.0 15.0 PC1	 Principal Component Analysis (standardized data) □ MDI Form Calculate List of Descriptors [9 of 10] ## Type Name 1 Interval ClumpTh 2 Interval UnifCellSz 3 Interval UnifCellShape 4 Interval MargAdhesion 5 Interval SEpiCellSz Scaling of the Data ○ no scaling ○ mean centering ● standardize PC No. 3 PC No. 1 Obj.: X: Y: ☑ Use Class Colors Summary Loadings Scores Loading/Loading Score/Score Bi-plot Details PC3 2.0 1.0 0.0 -1.0 -2.0 -2.0 0.0 2.0 4.0 6.0 PC1
Score/Score plot of PC 1 and 3 of mean centering.	Score/Score plot of PC 1 and 3 of standardize.
 Principal Component Analysis (mean centered data) □ MDI Form Calculate List of Descriptors [9 of 10] ## Type Name 1 Interval ClumpTh 2 Interval UnifCellSz 3 Interval UnifCellShape 4 Interval MargAdhesion 5 Interval SEpiCellSz Scaling of the Data ○ no scaling ● mean centering ○ standardize PC No. 4 PC No. 1 Obj.: X: Y: ☑ Use Class Colors Summary Loadings Scores Loading/Loading Score/Score Bi-plot Details PC4 5.0 0.0 -5.0 -10.0 -5.0 0.0 5.0 10.0 15.0 PC1	 Principal Component Analysis (standardized data) □ MDI Form Calculate List of Descriptors [9 of 10] ## Type Name 1 Interval ClumpTh 2 Interval UnifCellSz 3 Interval UnifCellShape 4 Interval MargAdhesion 5 Interval SEpiCellSz Scaling of the Data ○ no scaling ○ mean centering ● standardize PC No. 4 PC No. 1 Obj.: X: Y: ☑ Use Class Colors Summary Loadings Scores Loading/Loading Score/Score Bi-plot Details PC4 2.0 1.0 0.0 -1.0 -2.0 -2.0 0.0 2.0 4.0 6.0 PC1
Score/Score plot of PC 1 and 4 of mean centering.	Score/Score plot of PC 1 and 4 of standardize.

*Classification (Dignität) of tumours, learned in the lecture Pathology and Physiology:

Benign	Malign
• Slow growth • Sharply limited • Expansive-displacing • No metastases • Local complications (e.g. pressure) • Highly differentiated • Monomorphic (normal looking) cells • Healing by excision	• Rapid growth • Indistinctly circumscribed • Invasive-destructive • Metastatic • Local AND systemic complications • Lowly differentiated • Polymorphic (dysplastic) cells • Progression/recurrence possible (for cure often surgery/ chemotherapy/r adiation necessary

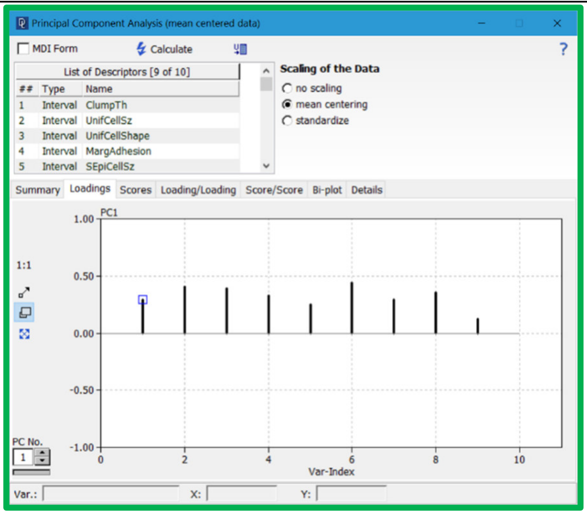
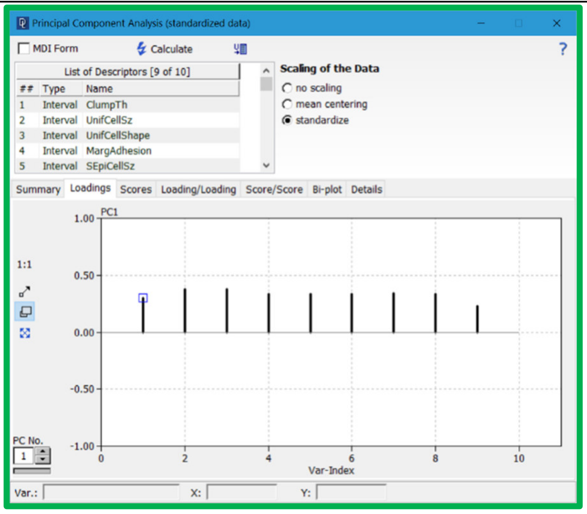
Whit this information we can say that the blue dots show the benign tumours and the red dots the malignant.

The four most important variations were evaluated using the score/score plot. The 1 principal component was always shown on the x-axis. The 1 principal component alone covers the highest percentage. The blue dots stand for benign tumours and the red dots for malignant tumours.

It is evident that the benign tumours are very centrally located, whereas the malignant tumours always have a high dispersion.

What is the reason that there is not much difference between mean centered and standardized data?

Toolbar PCA → select "List of Descriptors" and "Scaling of the data" → Calculate → Registry card "Loading"

Mean centering	Standardize
	

Principal Component Analysis (mean centered data)

□ MDI Form Calculate ?

List of Descriptors [9 of 10]		
##	Type	Name
1	Interval	ClumpTh
2	Interval	UnifCellSz
3	Interval	UnifCellShape
4	Interval	MargAdhesion
5	Interval	SEpiCellSz

Scaling of the Data
○ no scaling
● mean centering
○ standardize

Summary | Loadings | Scores | Loading/Loading | Score/Score | Bi-plot | Details

Principal Component Analysis (standardized data)

□ MDI Form Calculate ?

List of Descriptors [9 of 10]		
##	Type	Name
1	Interval	ClumpTh
2	Interval	UnifCellSz
3	Interval	UnifCellShape
4	Interval	MargAdhesion
5	Interval	SEpiCellSz

Scaling of the Data
○ no scaling
○ mean centering
● standardize

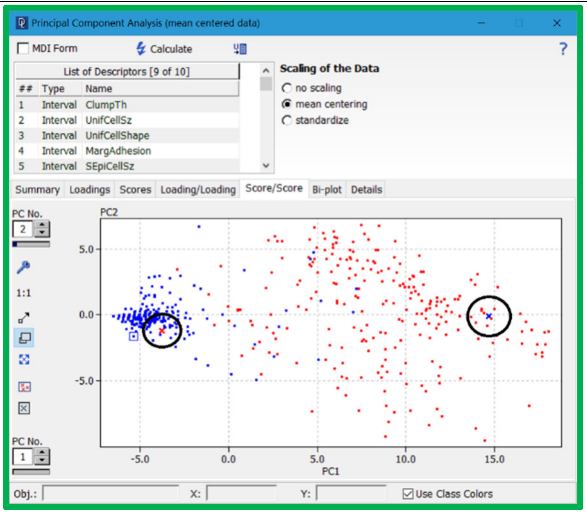
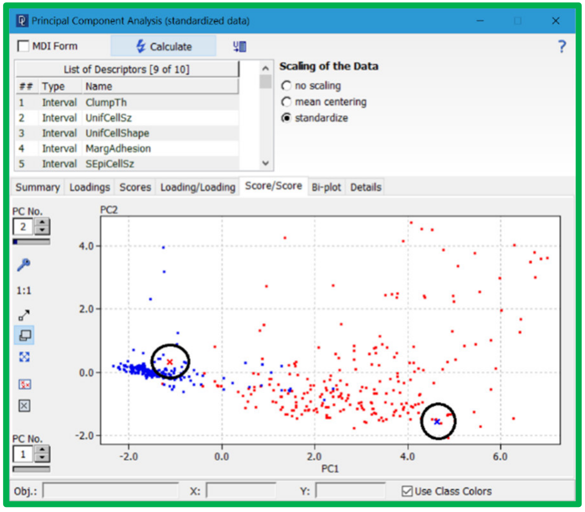
Summary | Loadings | Scores | Loading/Loading | Score/Score | Bi-plot | Details

All variables are on the same side, in the positive range. Thus, all variables correlate with the first principal component. Some variables are equally high, so they contain redudant information compared to each other. For the small variables, there is no such coverage.	All variables are on the same side, in the positive range. Thus, all variables correlate with the first principal component. Since the variables are largely of the same height, one can conclude that the individual variables contain multiple information.

We have no big difference between mean centering and standardize data because of the nature of the data, that they have a range of 1 to 10 and that leads to the point that they show the same standard deviation.

How could we decide whether a particular tissue is malignant? Suppose that you know the parameters of two samples. Which of the two samples is most probably malignant?

Editor → Right mouse klick "Insert Row" → Include parameters of samples → Type A/B "mark selected rows as type A/B" → Toolbar PCA → select "List of Descriptors" and "Scaling of the data" → Calculate → Registry card "Score/Score" → chose PC No.

<table>
<tr><th>Mean centering</th><th>Standardize</th></tr>
<tr><td colspan="2">Blue Dots: Benign*
Red Dots: Malignant*</td></tr>
<tr><td>
Principal Component Analysis (mean centered data)
☐ MDI Form — Calculate
List of Descriptors [9 of 10]
Type Name
1 Interval ClumpTh
2 Interval UnifCellSz
3 Interval UnifCellShape
4 Interval MargAdhesion
5 Interval SEpiCellSz
Scaling of the Data
○ no scaling
● mean centering
○ standardize
Summary | Loadings | Scores | Loading/Loading | Score/Score | Bi-plot | Details
PC No. 2 / PC No. 1
Obj.: X: Y: ☑ Use Class Colors
</td><td>
Principal Component Analysis (standardized data)
☐ MDI Form — Calculate
List of Descriptors [9 of 10]
Type Name
1 Interval ClumpTh
2 Interval UnifCellSz
3 Interval UnifCellShape
4 Interval MargAdhesion
5 Interval SEpiCellSz
Scaling of the Data
○ no scaling
○ mean centering
● standardize
Summary | Loadings | Scores | Loading/Loading | Score/Score | Bi-plot | Details
PC No. 2 / PC No. 1
Obj.: X: Y: ☑ Use Class Colors
</td></tr>
<tr><td>Red cross: Sample 2 → benign tumor
Bue cross: Sample 1 → malignant tumor</td><td>Red cross: Sample 2 → benign tumor
Bue cross: Sample 1 → malignant tumor</td></tr>
</table>

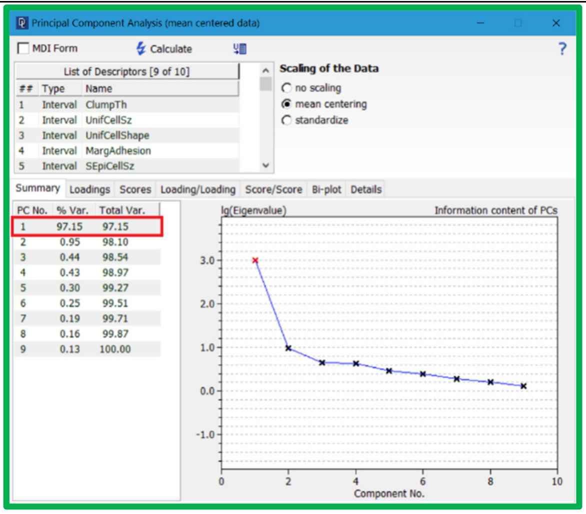
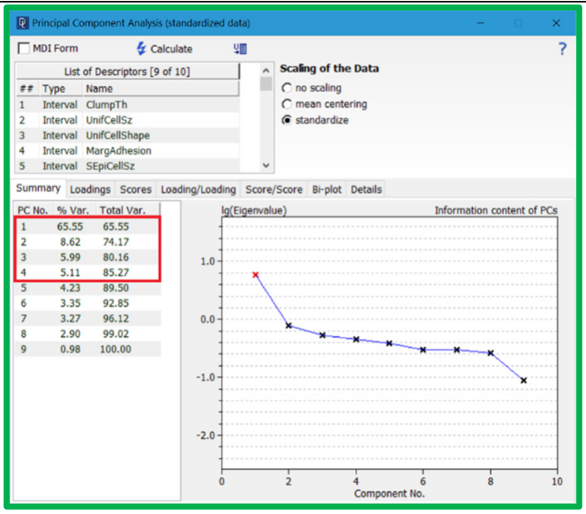
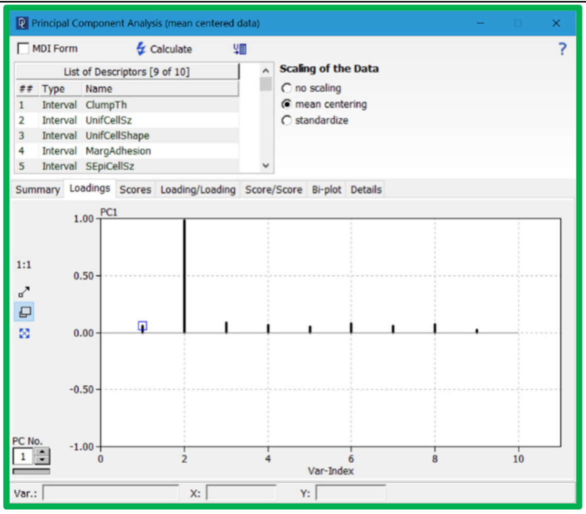
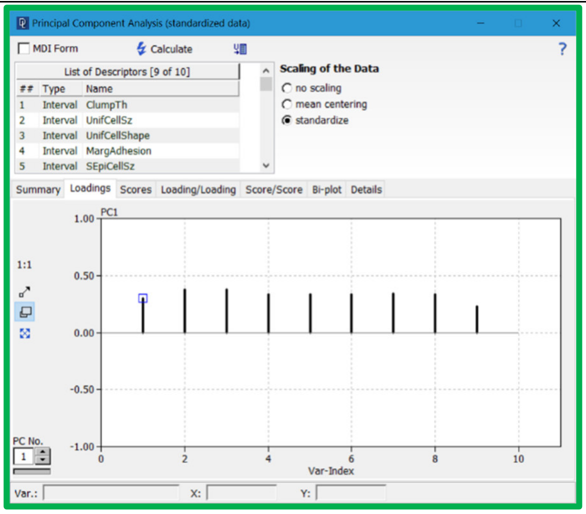
To check whether the two samples were benign or malignant tumours, the data sets were included in the database. When reviewing the score/score graphs, it was found that sample 2 (red cross) was always around the blue dots, which represent benign tumours.

Data set 1 (blue cross) was found to always be in the area of the red points which shows the malignant tumour data.

Multiply the variable 3 (UnifCellSz) by a factor of 10 and repeat the PCA (both for mean centered and standardized data). Which difference you see? Explain your findings.

Editor → Mark column → right mouse click "Copy" → include copy in excel and multiplied whit 10 than copy new counts → Mark column → right mouse click "Paste" → Toolbar PCA → select "List of Descriptors" and "Scaling of the data" → Calculate → Registry card "Summary" → Registry card "Loading"

Editor → Right mouse klick "Insert Row" → Include parameters of samples → Type A/B "mark selected rows as type A/B" → Toolbar PCA → select "List of Descriptors" and "Scaling of the data" → Calculate → Registry card "Score/Score" → chose PC No.

Mean centering	Standardize
Variable 3 (UnifCellSz) were multiplied whit the factor of 10.	
	
In contrast to the first calculation, we need now, to covered 85% of the total variance, only the first component. The first component contains alone 97.15% of all variations.	The standardize dataset shows no differences to the first calculations.
	

Principal Component Analysis (mean centered data)

□ MDI Form Calculate

List of Descriptors [9 of 10]

##	Type	Name
1	Interval	ClumpTh
2	Interval	UnifCellSz
3	Interval	UnifCellShape
4	Interval	MargAdhesion
5	Interval	SEpiCellSz

Scaling of the Data
○ no scaling
● mean centering
○ standardize

Summary | Loadings | Scores | Loading/Loading | Score/Score | Bi-plot | Details

PC No.	% Var.	Total Var.
1	97.15	97.15
2	0.95	98.10
3	0.44	98.54
4	0.43	98.97
5	0.30	99.27
6	0.25	99.51
7	0.19	99.71
8	0.16	99.87
9	0.13	100.00

Principal Component Analysis (standardized data)

□ MDI Form Calculate

List of Descriptors [9 of 10]

##	Type	Name
1	Interval	ClumpTh
2	Interval	UnifCellSz
3	Interval	UnifCellShape
4	Interval	MargAdhesion
5	Interval	SEpiCellSz

Scaling of the Data
○ no scaling
○ mean centering
● standardize

Summary | Loadings | Scores | Loading/Loading | Score/Score | Bi-plot | Details

PC No.	% Var.	Total Var.
1	65.55	65.55
2	8.62	74.17
3	5.99	80.16
4	5.11	85.27
5	4.23	89.50
6	3.35	92.85
7	3.27	96.12
8	2.90	99.02
9	0.98	100.00

Principal Component Analysis (mean centered data)

□ MDI Form Calculate

List of Descriptors [9 of 10]

##	Type	Name
1	Interval	ClumpTh
2	Interval	UnifCellSz
3	Interval	UnifCellShape
4	Interval	MargAdhesion
5	Interval	SEpiCellSz

Scaling of the Data
○ no scaling
● mean centering
○ standardize

Summary | Loadings | Scores | Loading/Loading | Score/Score | Bi-plot | Details

PC No. 1

Var.: X: Y:

Principal Component Analysis (standardized data)

□ MDI Form Calculate

List of Descriptors [9 of 10]

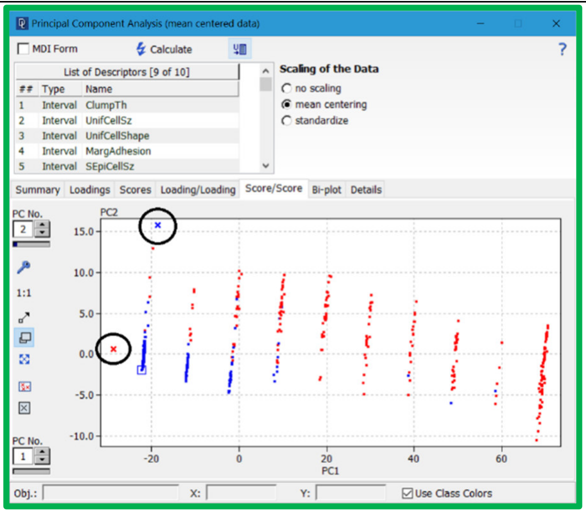
##	Type	Name
1	Interval	ClumpTh
2	Interval	UnifCellSz
3	Interval	UnifCellShape
4	Interval	MargAdhesion
5	Interval	SEpiCellSz

Scaling of the Data
○ no scaling
○ mean centering
● standardize

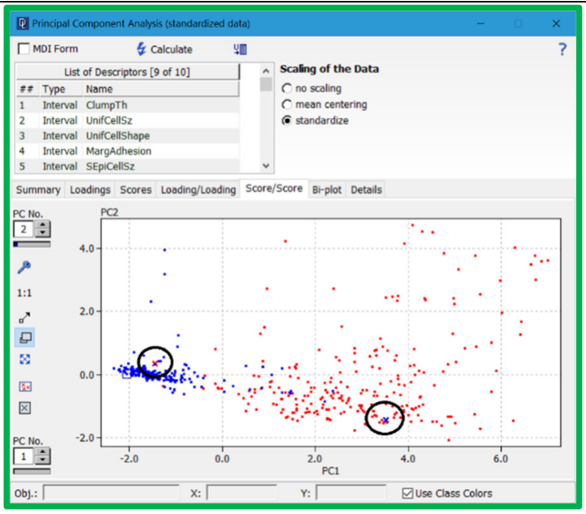
Summary | Loadings | Scores | Loading/Loading | Score/Score | Bi-plot | Details

PC No. 1

Var.: X: Y:

<table>
<tr><td>The second variable is very high. This is this one wo we have multiplied whit the factor of 10. This variable is the reason that between “mean centering” data and “standadize” data we have a difference.</td><td>The standardize dataset shows no differences to the first calculations.</td></tr>
<tr><td colspan="2">Blue Dots: Benign*
Red Dots: Malignant*</td></tr>
<tr><td>

PC2
15.0
10.0
5.0
0.0
-5.0
-10.0
-20
0
20
40
60
PC1
1:1

Principal Component Analysis (mean centered data)
□ MDI Form
Calculate
List of Descriptors [9 of 10]
Type Name
1 Interval ClumpTh
2 Interval UnifCellSz
3 Interval UnifCellShape
4 Interval MargAdhesion
5 Interval SEpiCellSz
Scaling of the Data
○ no scaling
● mean centering
○ standardize
Summary Loadings Scores Loading/Loading Score/Score Bi-plot Details
PC No. 2
PC No. 1
Obj.: X: Y: ☑ Use Class Colors</td><td>

PC2
4.0
2.0
0.0
-2.0
-2.0
0.0
2.0
4.0
6.0
PC1
1:1

Principal Component Analysis (standardized data)
□ MDI Form
Calculate
List of Descriptors [9 of 10]
Type Name
1 Interval ClumpTh
2 Interval UnifCellSz
3 Interval UnifCellShape
4 Interval MargAdhesion
5 Interval SEpiCellSz
Scaling of the Data
○ no scaling
○ mean centering
● standardize
Summary Loadings Scores Loading/Loading Score/Score Bi-plot Details
PC No. 2
PC No. 1
Obj.: X: Y: ☑ Use Class Colors</td></tr>
<tr><td>The score/score evaluation between the principel components shows a completely different graphical evaluation.
If the two samples are added to this data set to test the malignancy of a tissue, no clear statement can be made.</td><td>The standardize dataset shows no differences to the first calculations. Regardless of whether it is only the score/score evaluation or the examination of the malignancy of a tissue.</td></tr>
</table>

If variable 3 is multiplied by a factor of 10, nothing changes in the evaluation method using "standardize". The "mean cantering" shows a completely different behaviour, which only focuses on variable 3. This strong influence of the variable on the entire evaluation method makes it impossible to make statements about the malignancy of certain tissues.

Problem 6: Estimate the moisture content of corn
Report

Problem description:

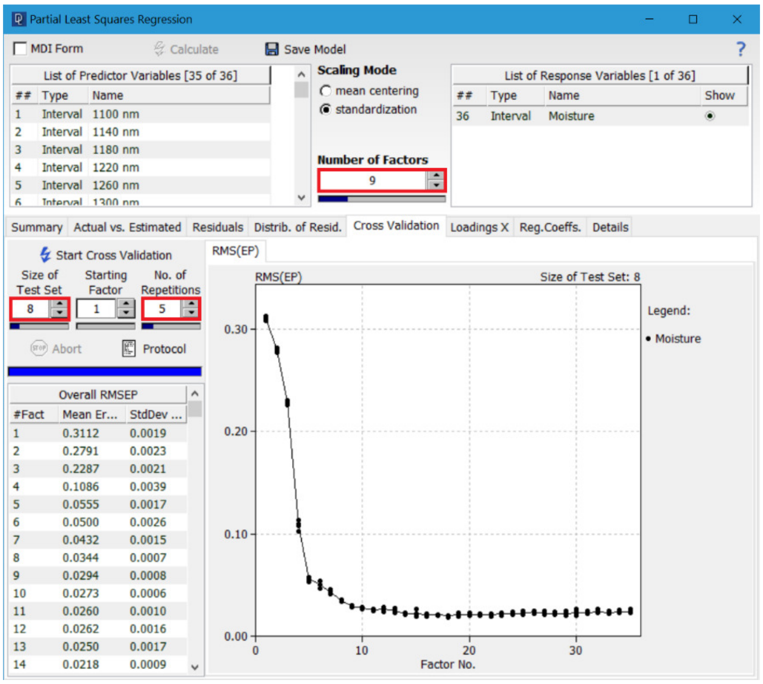
We have a data set of n=2880 values. This is divided into 36 columns, each containing 80 measurements of different moisture content of corn by specific wavelengths. We should use this data to establish a multilinear model for the estimation of the moisture using PLS.

Outcome:

PLS is a linear regression method that projects components in a regression model as new independent variables (explanatory variables, or predictors) in a new space.

Find the optimum number of factors

Toolbar PLS → select "Dependent Variable" and "List of Descriptors" → chose "Scaling Model" "standardization" → Calculate → Registry card "Summary"



Partial Least Squares Regression

□ MDI Form

Scaling Mode
○ mean centering
● standardization

Number of Factors: 9

List of Predictor Variables [35 of 36]

##	Type	Name
1	Interval	1100 nm
2	Interval	1140 nm
3	Interval	1180 nm
4	Interval	1220 nm
5	Interval	1260 nm
6	Interval	1300 nm

List of Response Variables [1 of 36]

##	Type	Name	Show
36	Interval	Moisture	●

Size of Test Set: 8
Starting Factor: 1
No. of Repetitions: 5

Overall RMSEP

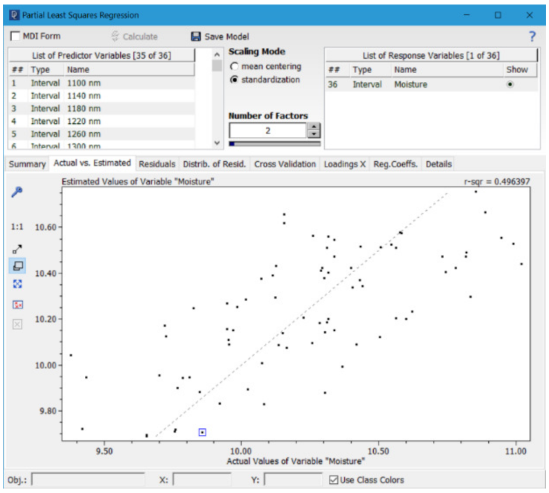
#Fact	Mean Er...	StdDev ...
1	0.3112	0.0019
2	0.2791	0.0023
3	0.2287	0.0021
4	0.1086	0.0039
5	0.0555	0.0017
6	0.0500	0.0026
7	0.0432	0.0015
8	0.0344	0.0007
9	0.0294	0.0008
10	0.0273	0.0006
11	0.0260	0.0010
12	0.0262	0.0016
13	0.0250	0.0017
14	0.0218	0.0009

Figure 1: Cross Validation of the corn data to estimate how many factors are the best.

For the size of test set we take 10% of all the data in this case 8. Also we estimate 5 repetitions. By looking at the graph the best numbers of factors should be 9.

Compare the "actual vs. estimated" plot for 2, 4, 6 and 8 factors. What do you observe when you increase the number of factors?

Toolbar PLS → select "Dependent Variable" and "List of Descriptors" → chose "Scaling Model" "standardization" → Calculate → Registry card "Actual vs. Estimated" → chose No. of PCs



List of Predictor Variables [35 of 36]

##	Type	Name
1	Interval	1100 nm
2	Interval	1140 nm
3	Interval	1180 nm
4	Interval	1220 nm
5	Interval	1260 nm
6	Interval	1300 nm

Scaling Mode
○ mean centering
● standardization

Number of Factors: 2

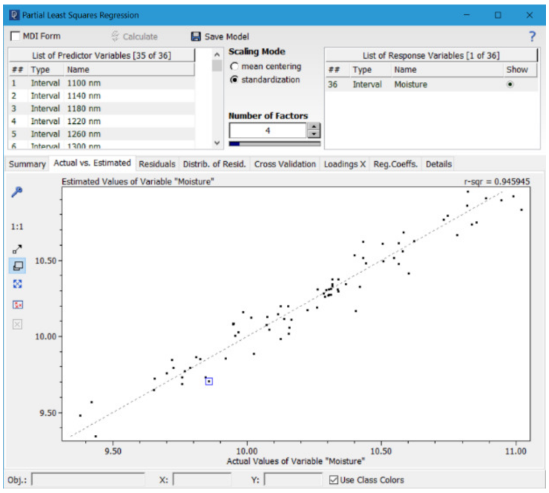
List of Response Variables [1 of 36]

##	Type	Name	Show
36	Interval	Moisture	●

r-sqr = 0.496397

Obj.: X: Y: ☑ Use Class Colors

Figure 2: Actual vs. Estimated graph whit 2 factors.



List of Predictor Variables [35 of 36]

##	Type	Name
1	Interval	1100 nm
2	Interval	1140 nm
3	Interval	1180 nm
4	Interval	1220 nm
5	Interval	1260 nm
6	Interval	1300 nm

Scaling Mode
○ mean centering
● standardization

Number of Factors: 4

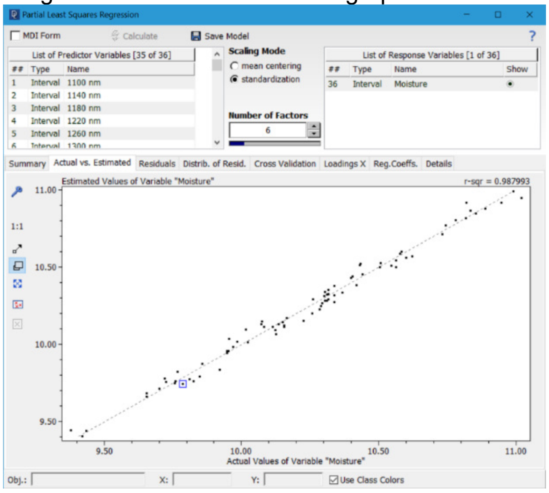
List of Response Variables [1 of 36]

##	Type	Name	Show
36	Interval	Moisture	●

r-sqr = 0.945945

Obj.: X: Y: ☑ Use Class Colors

Figure 3: Actual vs. Estimated graph whit 4 factors.



List of Predictor Variables [35 of 36]

##	Type	Name
1	Interval	1100 nm
2	Interval	1140 nm
3	Interval	1180 nm
4	Interval	1220 nm
5	Interval	1260 nm
6	Interval	1300 nm

Scaling Mode
○ mean centering
● standardization

Number of Factors: 6

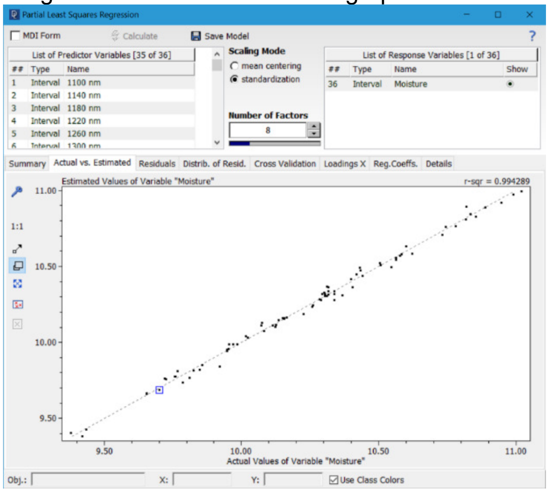
List of Response Variables [1 of 36]

##	Type	Name	Show
36	Interval	Moisture	●

r-sqr = 0.987993

Obj.: X: Y: ☑ Use Class Colors

Figure 4: Actual vs. Estimated graph whit 6 factors.



List of Predictor Variables [35 of 36]

##	Type	Name
1	Interval	1100 nm
2	Interval	1140 nm
3	Interval	1180 nm
4	Interval	1220 nm
5	Interval	1260 nm
6	Interval	1300 nm

Scaling Mode
○ mean centering
● standardization

Number of Factors: 8

List of Response Variables [1 of 36]

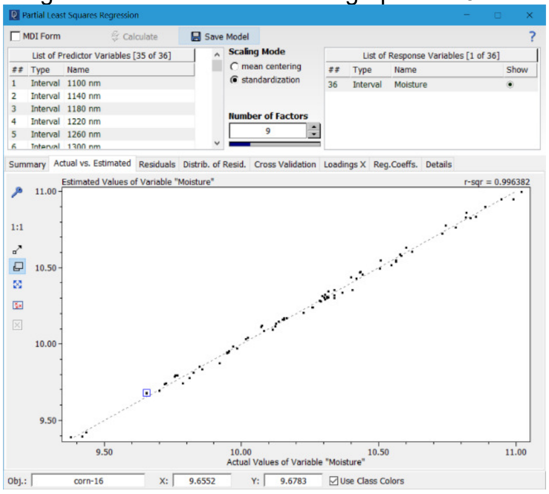
##	Type	Name	Show
36	Interval	Moisture	●

r-sqr = 0.994289

Obj.: X: Y: ☑ Use Class Colors

Figure 5: Actual vs. Estimated graph whit 8 factors.

Whit increasing the number of factors the fitting of the values increases to the diagonal line, which represent the perfect estimation. Between the factors 8 and 9 no high change is more visible.



List of Predictor Variables [35 of 36]

##	Type	Name
1	Interval	1100 nm
2	Interval	1140 nm
3	Interval	1180 nm
4	Interval	1220 nm
5	Interval	1260 nm
6	Interval	1300 nm

Scaling Mode
○ mean centering
● standardization

Number of Factors: 9

List of Response Variables [1 of 36]

##	Type	Name	Show
36	Interval	Moisture	●

r-sqr = 0.996382

Obj.: corn-16 X: 9.6552 Y: 9.6783 ☑ Use Class Colors

Figure 6: Actual vs. Estimated graph whit 9 factors. This number of factors I chose as the best fit.

Compare the regression coefficients of the original variables to the absorption spectrum of water (see for example https://omlc.org/spectra/water/data/palmer74.txt for details). Perform this comparison using 3, 6, 10 and 35 factors.

Menu tab “File” → “Load” → “Simple Text…” → “Load Text File” → chose “Colum headings only” → selected the data wo should be ignore → chose by “Separator: Unknown” → look in the Registry card “Preview” → “Copy to Data Matrix” → Menu tab “Math” → “Simple Regression…” → select “independent variable” and “dependent variable” → click “OK” → chose “Draw connecting lines”

Toolbar PLS → select “Dependent Variable” and “List of Descriptors” → chose “Scaling Model” “standardization” → Calculate → Registry card “Reg. Coeffs.” → chose No. of PCs

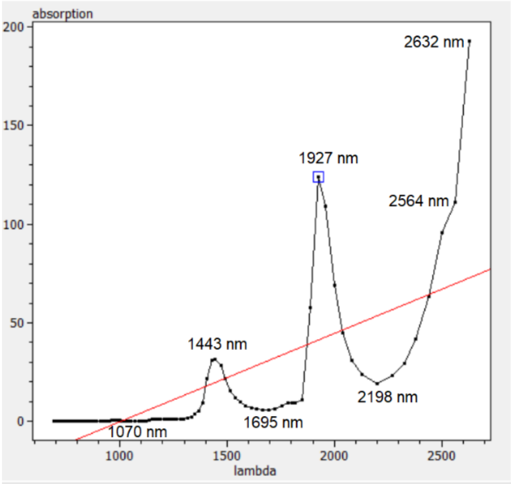
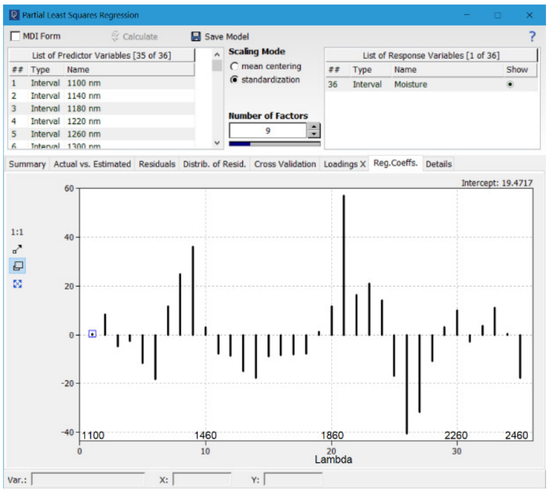


Figure 7: Regression graph from the absorption spectrum of water. Data for the graph are taken from the website: https://omlc.org/spectra/water/data/palmer74.txt



List of Predictor Variables [35 of 36]

##	Type	Name
1	Interval	1100 nm
2	Interval	1140 nm
3	Interval	1180 nm
4	Interval	1220 nm
5	Interval	1260 nm
6	Interval	1300 nm

Scaling Mode
○ mean centering
● standardization

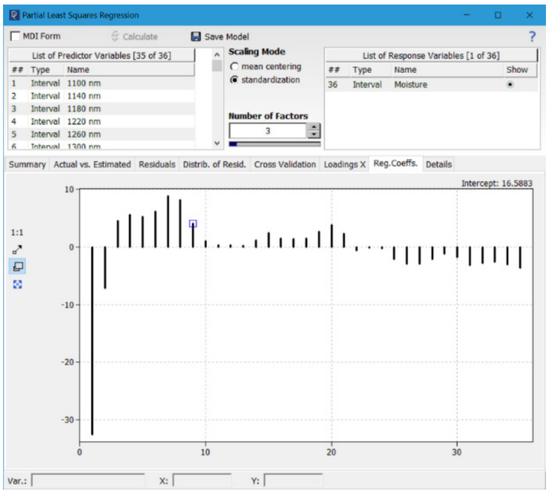
Number of Factors 9

List of Response Variables [1 of 36]

##	Type	Name	Show
36	Interval	Moisture	●

Intercept: 19.4717

Figure 8: Regressions Coefficient of the variable’s whit 9 factors. This number of factors I chose as the best fit. The first pike on the left symbols the 1100 nm variables, the pike on the far-right symbols the 2460 nm.



List of Predictor Variables [35 of 36]

##	Type	Name
1	Interval	1100 nm
2	Interval	1140 nm
3	Interval	1180 nm
4	Interval	1220 nm
5	Interval	1260 nm
6	Interval	1300 nm

Scaling Mode
○ mean centering
● standardization

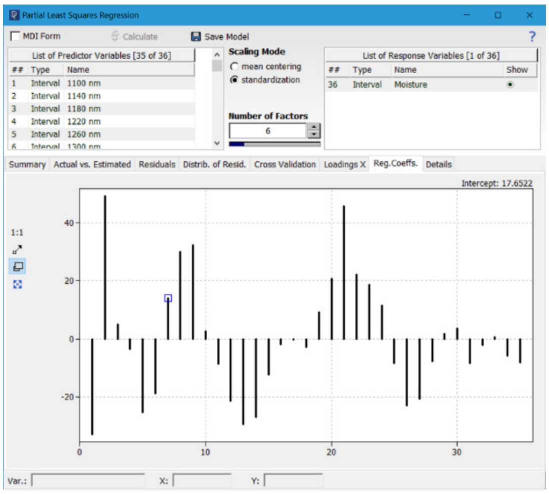
Number of Factors 3

List of Response Variables [1 of 36]

##	Type	Name	Show
36	Interval	Moisture	●

Intercept: 16.5883

Figure 9: Regressions Coefficient of the variable’s whit 3 factors. The first pike on the left symbols the 1100 nm variables, the pike on the far-right symbols the 2460 nm.



List of Predictor Variables [35 of 36]

##	Type	Name
1	Interval	1100 nm
2	Interval	1140 nm
3	Interval	1180 nm
4	Interval	1220 nm
5	Interval	1260 nm
6	Interval	1300 nm

Scaling Mode
○ mean centering
● standardization

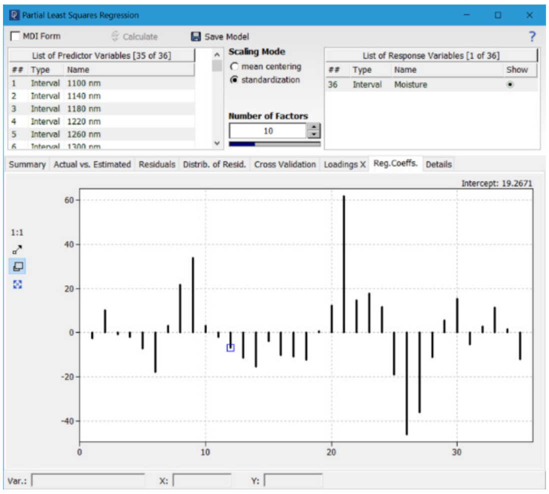
Number of Factors 6

List of Response Variables [1 of 36]

##	Type	Name	Show
36	Interval	Moisture	●

Intercept: 17.6522

Figure 10: Regressions Coefficient of the variable’s whit 6 factors. The first pike on the left symbols the 1100 nm variables, the pike on the far-right symbols the 2460 nm.



Partial Least Squares Regression

☐ MDI Form Calculate Save Model ?

List of Predictor Variables [35 of 36]		
##	Type	Name
1	Interval	1100 nm
2	Interval	1140 nm
3	Interval	1180 nm
4	Interval	1220 nm
5	Interval	1260 nm
6	Interval	1300 nm

Scaling Mode
○ mean centering
● standardization

Number of Factors: 10

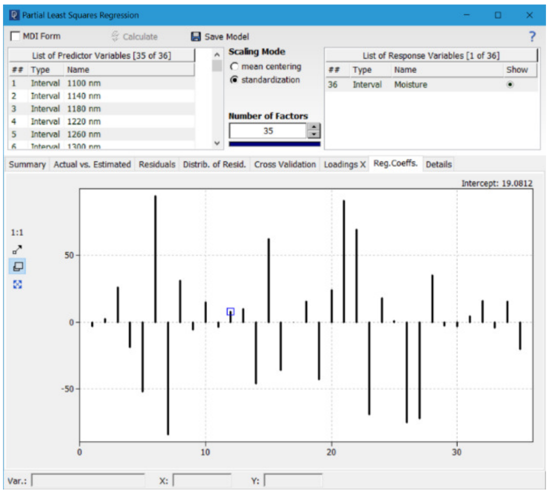
List of Response Variables [1 of 36]			
##	Type	Name	Show
36	Interval	Moisture	●

Summary | Actual vs. Estimated | Residuals | Distrib. of Resid. | Cross Validation | Loadings X | Reg.Coeffs. | Details

Intercept: 19.2671

Var.: X: Y:

Figure 11: Regressions Coefficient of the variable’s whit 10 factors. The first pike on the left symbols the 1100 nm variables, the pike on the far-right symbols the 2460 nm.



Partial Least Squares Regression

☐ MDI Form Calculate Save Model ?

List of Predictor Variables [35 of 36]		
##	Type	Name
1	Interval	1100 nm
2	Interval	1140 nm
3	Interval	1180 nm
4	Interval	1220 nm
5	Interval	1260 nm
6	Interval	1300 nm

Scaling Mode
○ mean centering
● standardization

Number of Factors: 35

List of Response Variables [1 of 36]			
##	Type	Name	Show
36	Interval	Moisture	●

Summary | Actual vs. Estimated | Residuals | Distrib. of Resid. | Cross Validation | Loadings X | Reg.Coeffs. | Details

Intercept: 19.0812

Var.: X: Y:

Figure 12: Regressions Coefficient of the variable’s whit 35 factors. The first pike on the left symbols the 1100 nm variables, the pike on the far-right symbols the 2460 nm.

For the evaluation, it is irrelevant whether the values are positive or negative. For the analysis, all values should be plotted on one side. When looking at the regression coefficient graphs in comparison to the spectrum of water, it can be seen that the worst agreement exists with a factor of 3. As the number of factors increases, the agreement increases continuously, reaching its maximum at a factor of 9. With a further increase of the factor, the quality of the match decreases again.

Why is it a bad idea to use MLR instead of PLS for modelling the moisture content?

PLS is better for this data set as it requires fewer factors. MLR would pick up all 35 factors, but this would have the problem that the picks of the regression coefficient evaluation would overlap.

Is there something special about a model using 35 factors?

When you use to many factors you get a higher instability. If you include all the factors, in this case the 35 factors, the system degenerates in an MLR.

Problem 7: Recognizing Dry Beans
Report

Problem description:

We have a data set of n=16000 values. This is divided into 7 columns where each containing 2000 geometrical measurements of seven different dry beans. The last column (8 column) includes the name of the bean cultivars. We should use this dataset to look for a processing to classify these beans. For this we make Partial Least Squares Discriminant Analysis (PLS/DA) and Random Forest (RF) and looking which system working best for the classify.

Outcome:

PLS is a linear regression method that projects components of a regression model as new independent variables (explanatory variables or predictors) into a new space. Both the X and Y data are projected onto new spaces. PLS-DA is a variant used when the Y data are categorical.
A Random Forest is a classification and regression procedure consisting of several uncorrelated decision trees. All decision trees are trained under a certain type of randomisation during the learning process. For a classification, each tree in that forest is allowed to make a vote and the class with the most votes decides the final classification.

Preparation of the data set

Menu tab “Tools” → “Create Indicator Variables…” → Select “Cultivar” → Create Variables

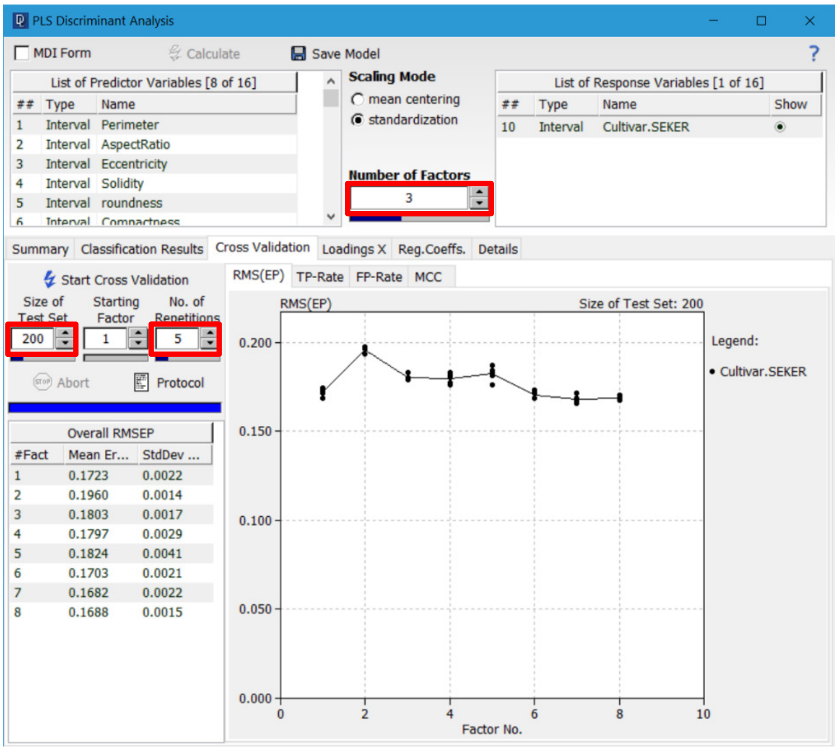
Table 1: Extract from the table with the measurement data after inserting the additional columns with the values 1 and 0 per bean type.

object nam	Cultivar	Culti-var.S	Culti-var.B	Culti-var.B	Culti-var.C	Culti-var.H	Culti-var.S	Culti-var.D
R-1269	SEKER	1	0	0	0	0	0	0
R-2390	BARBUNYA	0	1	0	0	0	0	0
R-3857	BOMBAY	0	0	1	0	0	0	0
R-4169	CALI	0	0	0	1	0	0	0
R-6835	HOROZ	0	0	0	0	1	0	0
R-8280	SIRA	0	0	0	0	0	1	0
R-11826	DERMASON	0	0	0	0	0	0	1

To generate a column once for each bean type where it has the value 1 and all other bean types have the value 0. Otherwise, you get an error message with PLS/DA.

Find the optimum number of factors

Toolbar DPLS → select “List of Predictors Variables” and “List of Response Variables” → chose “Scaling Model” “standardization” → Calculate → Registry card “Cross Validation” → Registry card “RMS(EP)” → “Start Cross Validation”



PLS Discriminant Analysis

□ MDI Form | Calculate | Save Model

List of Predictor Variables [8 of 16]

##	Type	Name
1	Interval	Perimeter
2	Interval	AspectRatio
3	Interval	Eccentricity
4	Interval	Solidity
5	Interval	roundness
6	Interval	Compactness

Scaling Mode
○ mean centering
● standardization

Number of Factors: 3

List of Response Variables [1 of 16]

##	Type	Name	Show
10	Interval	Cultivar.SEKER	●

Summary | Classification Results | Cross Validation | Loadings X | Reg.Coeffs. | Details

Start Cross Validation

Size of Test Set: 200
Starting Factor: 1
No. of Repetitions: 5

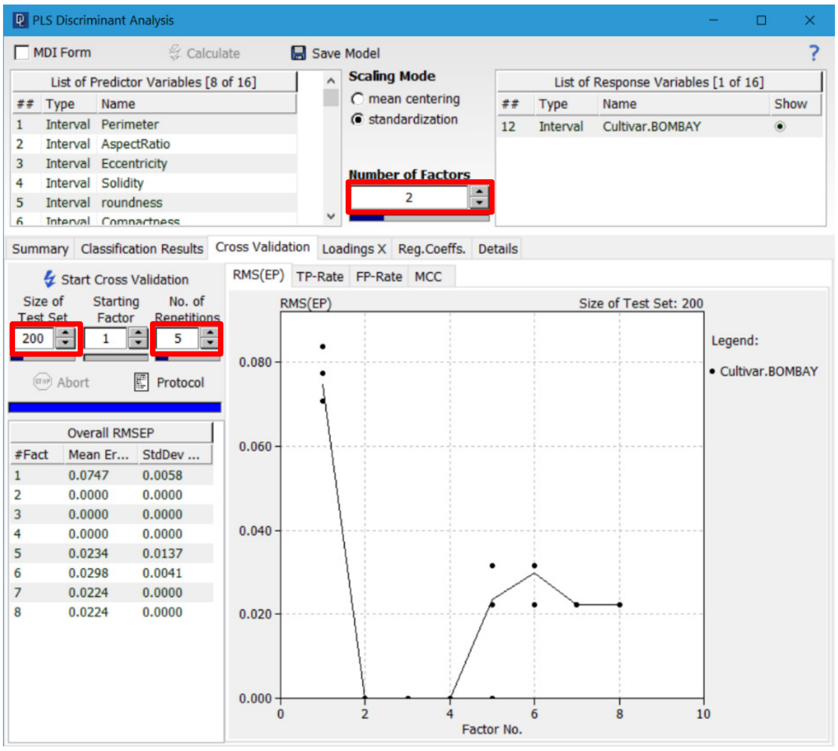
Abort | Protocol

Overall RMSEP

#Fact	Mean Er...	StdDev ...
1	0.1723	0.0022
2	0.1960	0.0014
3	0.1803	0.0017
4	0.1797	0.0029
5	0.1824	0.0041
6	0.1703	0.0021
7	0.1682	0.0022
8	0.1688	0.0015

Figure 1: Cross Validation of the bean data SEKER to estimate how many factors are the best.

For the size of test set we take 10% of all the data in this case 200. Also we estimate 5 repetitions. By looking at the graph the best numbers of factors should be 3.



PLS Discriminant Analysis

□ MDI Form | Calculate | Save Model

List of Predictor Variables [8 of 16]

##	Type	Name
1	Interval	Perimeter
2	Interval	AspectRatio
3	Interval	Eccentricity
4	Interval	Solidity
5	Interval	roundness
6	Interval	Compactness

Scaling Mode
○ mean centering
● standardization

Number of Factors: 2

List of Response Variables [1 of 16]

##	Type	Name	Show
12	Interval	Cultivar.BOMBAY	●

Summary | Classification Results | Cross Validation | Loadings X | Reg.Coeffs. | Details

Start Cross Validation

Size of Test Set: 200 | Starting Factor: 1 | No. of Repetitions: 5

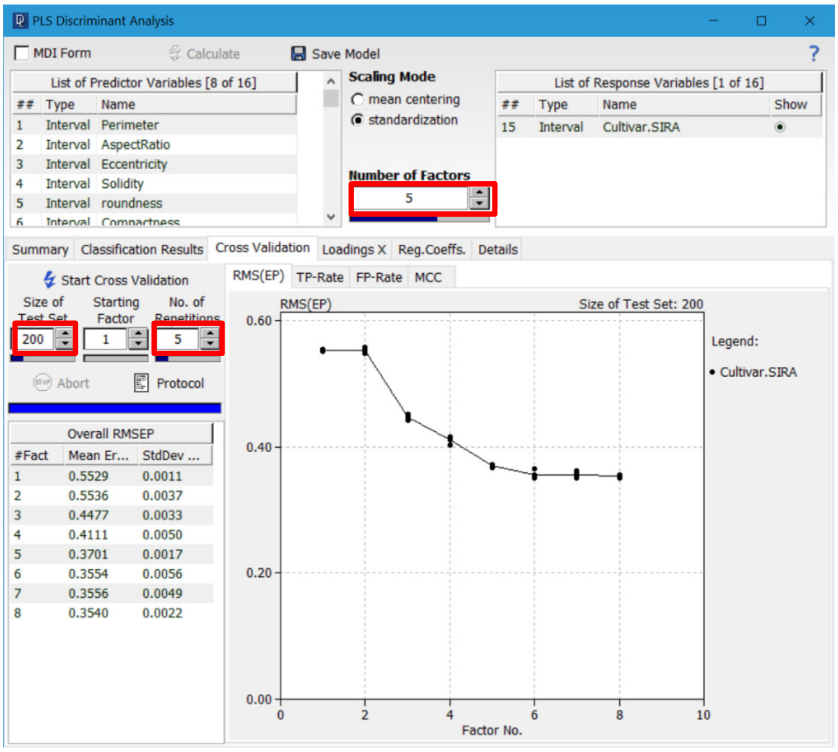
Abort | Protocol

Overall RMSEP

#Fact	Mean Er...	StdDev ...
1	0.0747	0.0058
2	0.0000	0.0000
3	0.0000	0.0000
4	0.0000	0.0000
5	0.0234	0.0137
6	0.0298	0.0041
7	0.0224	0.0000
8	0.0224	0.0000

Figure 2: Cross Validation of the bean data BOMBAY to estimate how many factors are the best.

For the size of test set we take 10% of all the data in this case 200. Also we estimate 5 repetitions. By looking at the graph the best numbers of factors should be 2.



PLS Discriminant Analysis

□ MDI Form | Calculate | Save Model

List of Predictor Variables [8 of 16]

##	Type	Name
1	Interval	Perimeter
2	Interval	AspectRatio
3	Interval	Eccentricity
4	Interval	Solidity
5	Interval	roundness
6	Interval	Compactness

Scaling Mode
○ mean centering
● standardization

Number of Factors: 5

List of Response Variables [1 of 16]

##	Type	Name	Show
15	Interval	Cultivar.SIRA	●

Summary | Classification Results | Cross Validation | Loadings X | Reg.Coeffs. | Details

Start Cross Validation

Size of Test Set: 200 | Starting Factor: 1 | No. of Repetitions: 5

Abort | Protocol

Overall RMSEP

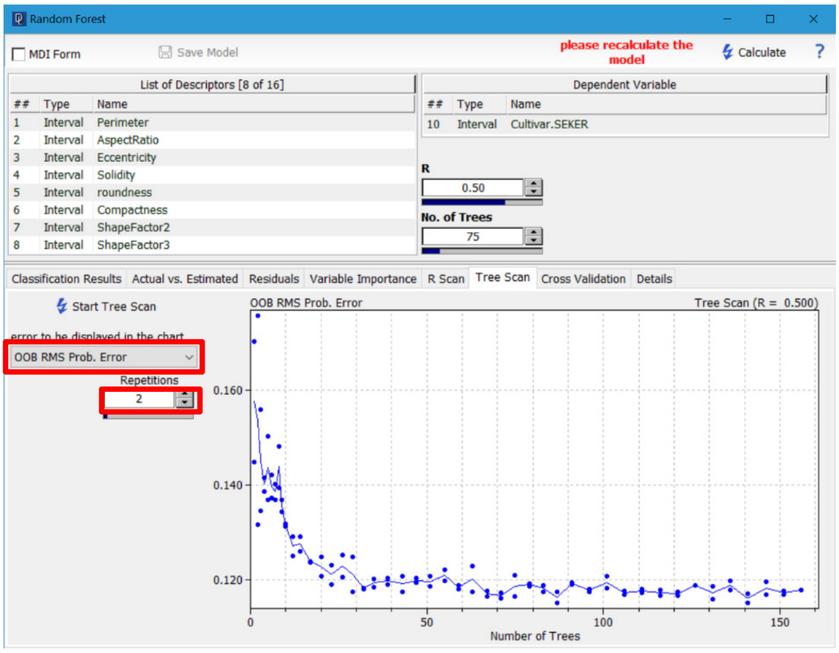
#Fact	Mean Er...	StdDev ...
1	0.5529	0.0011
2	0.5536	0.0037
3	0.4477	0.0033
4	0.4111	0.0050
5	0.3701	0.0017
6	0.3554	0.0056
7	0.3556	0.0049
8	0.3540	0.0022

Figure 3: Cross Validation of the bean data SIRA to estimate how many factors are the best.

For the size of test set we take 10% of all the data in this case 200. Also we estimate 5 repetitions. By looking at the graph the best numbers of factors should be 5.

Find the optimum number of trees

Menu tab "Tools" → "Random Forest" → select "List of Independent Variables" and "Dependent Variable" → Calculate → Registry card "Tree Scan" → "Start Tree scan" → after some time click "Abort"



List of Descriptors [8 of 16]

##	Type	Name
1	Interval	Perimeter
2	Interval	AspectRatio
3	Interval	Eccentricity
4	Interval	Solidity
5	Interval	roundness
6	Interval	Compactness
7	Interval	ShapeFactor2
8	Interval	ShapeFactor3

Dependent Variable

##	Type	Name
10	Interval	Cultivar.SEKER

R: 0.50

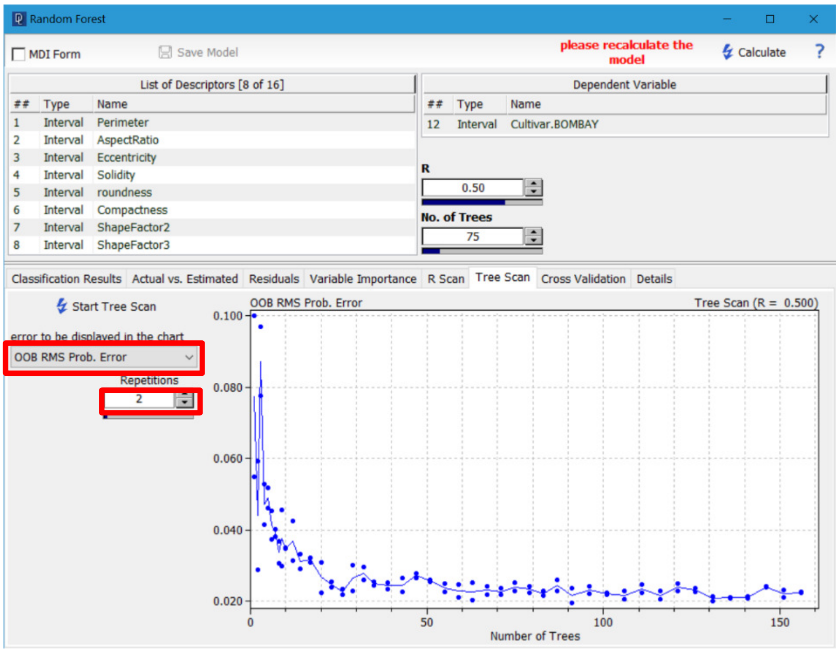
No. of Trees: 75

error to be displayed in the chart: OOB RMS Prob. Error

Repetitions: 2

Figure 4: Tree Scan of the bean data SEKER to estimate how many Trees are the best.

The settings R and No. of Trees were taken from the system settings. The number of repetitions was reduced to 2, otherwise the programme would take forever to calculate. The "Error to be displayed in the chart" was set to "OOB RMS Prob. Error" was used. By looking at the graph the best numbers of trees should be 70.



List of Descriptors [8 of 16]

##	Type	Name
1	Interval	Perimeter
2	Interval	AspectRatio
3	Interval	Eccentricity
4	Interval	Solidity
5	Interval	roundness
6	Interval	Compactness
7	Interval	ShapeFactor2
8	Interval	ShapeFactor3

Dependent Variable

##	Type	Name
12	Interval	Cultivar.BOMBAY

R: 0.50

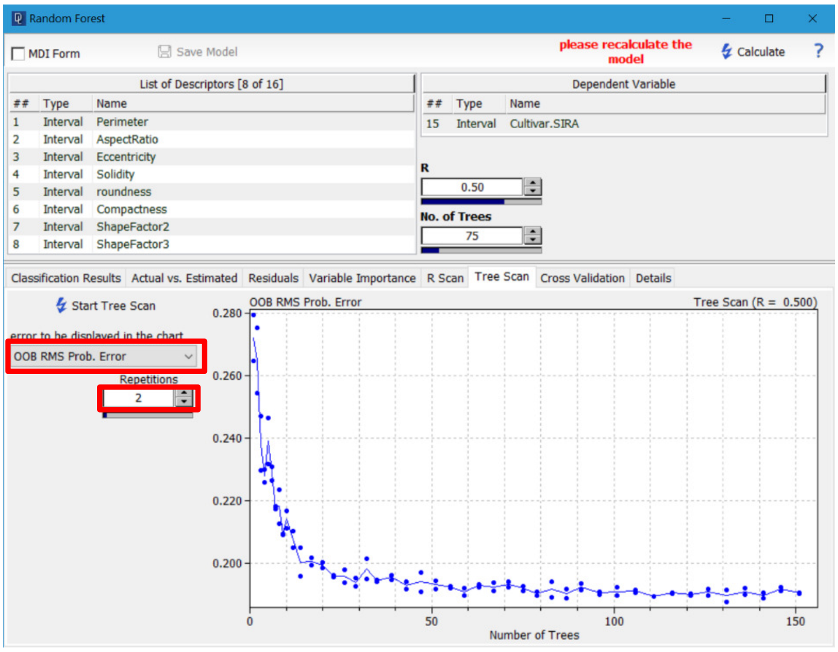
No. of Trees: 75

error to be displayed in the chart: OOB RMS Prob. Error

Repetitions: 2

Figure 5: Tree Scan of the bean data BOMBAY to estimate how many Trees are the best.

The settings R and No. of Trees were taken from the system settings. The number of repetitions was reduced to 2, otherwise the programme would take forever to calculate. The "Error to be displayed in the chart" was set to "OOB RMS Prob. Error" was used. By looking at the graph the best numbers of trees should be 60.



List of Descriptors [8 of 16]

##	Type	Name
1	Interval	Perimeter
2	Interval	AspectRatio
3	Interval	Eccentricity
4	Interval	Solidity
5	Interval	roundness
6	Interval	Compactness
7	Interval	ShapeFactor2
8	Interval	ShapeFactor3

Dependent Variable

##	Type	Name
15	Interval	Cultivar.SIRA

R: 0.50

No. of Trees: 75

error to be displayed in the chart: OOB RMS Prob. Error

Repetitions: 2

Figure 6: Tree Scan of the bean data SIRA to estimate how many Trees are the best.

The settings R and No. of Trees were taken from the system settings. The number of repetitions was reduced to 2, otherwise the programme would take forever to calculate. The "Error to be displayed in the chart" was set to "OOB RMS Prob. Error" was used. By looking at the graph the best numbers of trees should be 80.

Comparison of the six classifiers based on the Matthews Correlation Coefficient (MCC) value

Toolbar DPLS → select “List of Predictors Variables” and “List of Response Variables” → chose “Scaling Model” “standardization” → Calculate → Registry card “Cross Validation” → Registry card “RMS(EP)” → “Start Cross Validation” → Registry card “MCC”

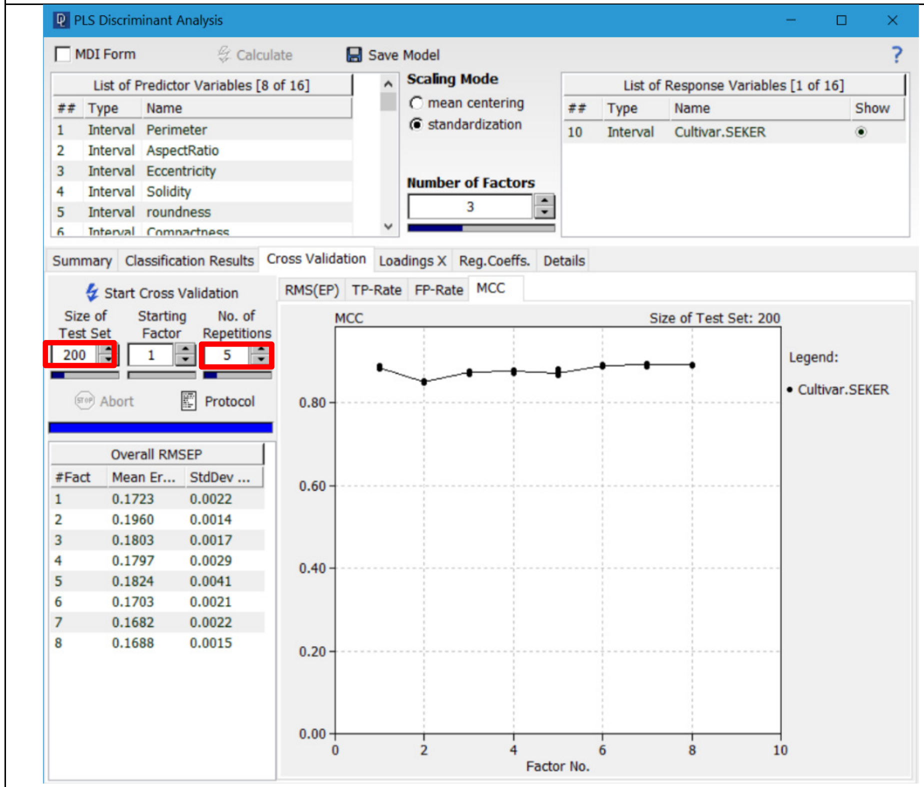
Menu tab “Tools” → “Random Forest” → select “List of Independent Variables” and “Dependent Variable” → Calculate → Registry card “Cross Validation” → “Start Cross Validation”

PLS/DA	RF

MCC = +1 → perfect forecast
MCC = 0 → random prediction
MCC = -1 → complete disagreement

$\mathrm{MCC} \neq \begin{cases} +1 \\ 0 \\ -1 \end{cases}$ → this is not a reliable indicator of how similar a predictor is to random guessing, since MCC depends on the data set.

PLS/DA



PLS Discriminant Analysis

☐ MDI Form   Calculate   Save Model

List of Predictor Variables [8 of 16]

##	Type	Name
1	Interval	Perimeter
2	Interval	AspectRatio
3	Interval	Eccentricity
4	Interval	Solidity
5	Interval	roundness
6	Interval	Compactness

Scaling Mode
○ mean centering
● standardization

Number of Factors: 3

List of Response Variables [1 of 16]

##	Type	Name	Show
10	Interval	Cultivar.SEKER	●

Summary | Classification Results | Cross Validation | Loadings X | Reg.Coeffs. | Details

Start Cross Validation

Size of Test Set: 200
Starting Factor: 1
No. of Repetitions: 5

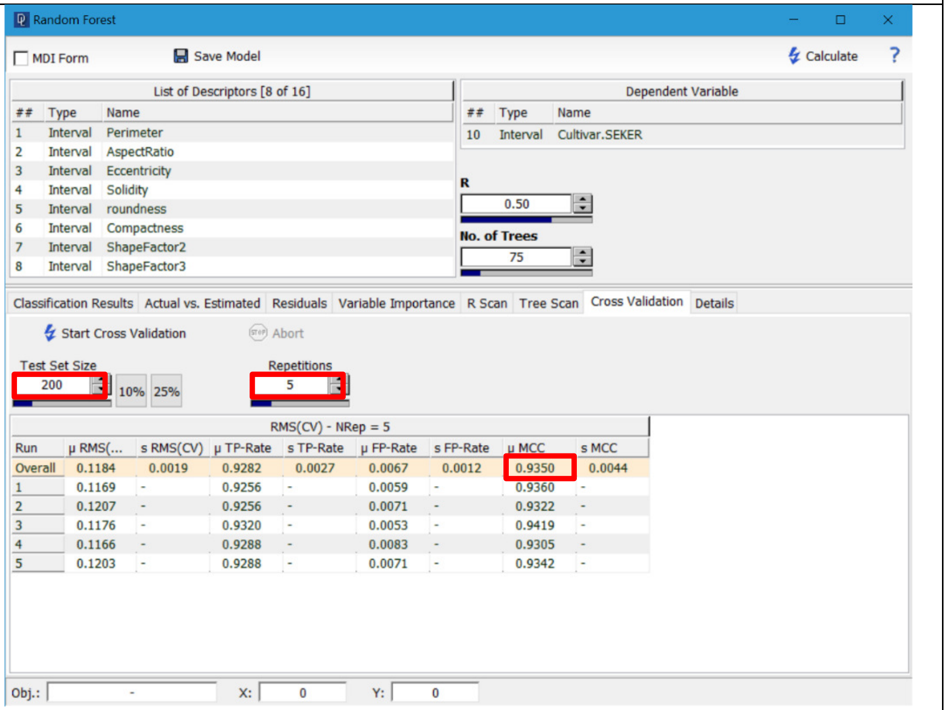
Abort   Protocol

Overall RMSEP

#Fact	Mean Er...	StdDev ...
1	0.1723	0.0022
2	0.1960	0.0014
3	0.1803	0.0017
4	0.1797	0.0029
5	0.1824	0.0041
6	0.1703	0.0021
7	0.1682	0.0022
8	0.1688	0.0015

Cultivars SEKER shows for a test set of 10% (200) and 5 repetition an MCC of 0.8909 by calculation whit PLS/DA.

RF



Random Forest

☐ MDI Form   Save Model   Calculate

List of Descriptors [8 of 16]

##	Type	Name
1	Interval	Perimeter
2	Interval	AspectRatio
3	Interval	Eccentricity
4	Interval	Solidity
5	Interval	roundness
6	Interval	Compactness
7	Interval	ShapeFactor2
8	Interval	ShapeFactor3

Dependent Variable

##	Type	Name
10	Interval	Cultivar.SEKER

R: 0.50
No. of Trees: 75

Classification Results | Actual vs. Estimated | Residuals | Variable Importance | R Scan | Tree Scan | Cross Validation | Details

Start Cross Validation   Abort

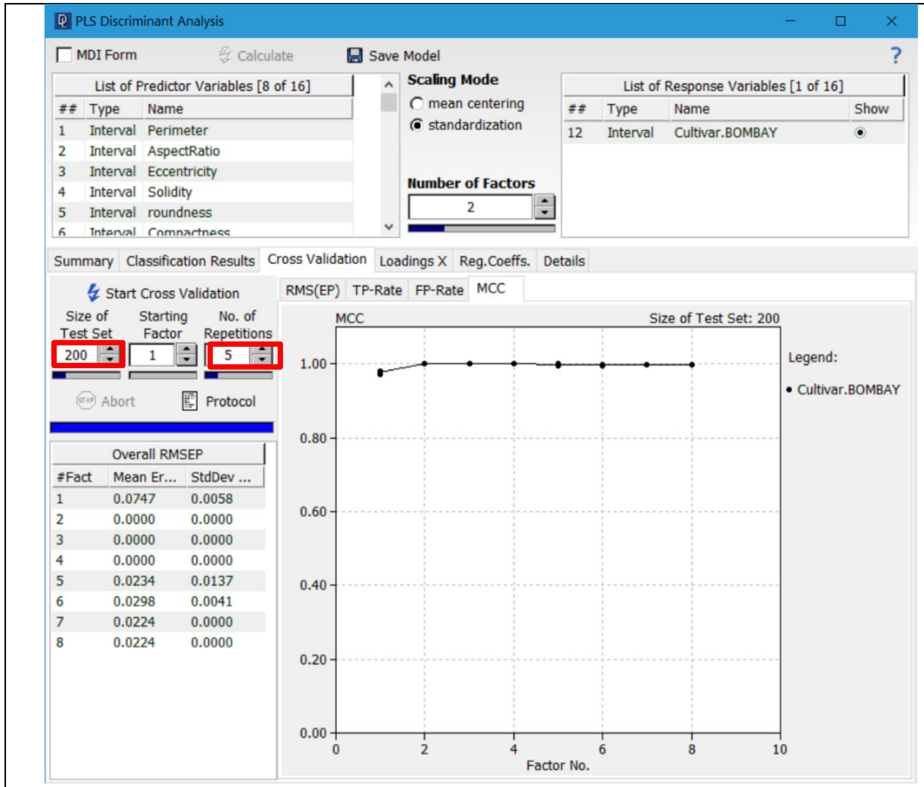
Test Set Size: 200   10%   25%
Repetitions: 5

RMS(CV) - NRep = 5

Run	μ RMS(...	s RMS(CV)	μ TP-Rate	s TP-Rate	μ FP-Rate	s FP-Rate	μ MCC	s MCC
Overall	0.1184	0.0019	0.9282	0.0027	0.0067	0.0012	0.9350	0.0044
1	0.1169	-	0.9256	-	0.0059	-	0.9360	-
2	0.1207	-	0.9256	-	0.0071	-	0.9322	-
3	0.1176	-	0.9320	-	0.0053	-	0.9419	-
4	0.1166	-	0.9288	-	0.0083	-	0.9305	-
5	0.1203	-	0.9288	-	0.0071	-	0.9342	-

Obj.: -   X: 0   Y: 0

Cultivars SEKER shows for a test set of 10% (200) and 5 repetition an MCC of 0.9350 by calculation whit RF.



PLS Discriminant Analysis

□ MDI Form   Calculate   Save Model   ?

List of Predictor Variables [8 of 16]		
##	Type	Name
1	Interval	Perimeter
2	Interval	AspectRatio
3	Interval	Eccentricity
4	Interval	Solidity
5	Interval	roundness
6	Interval	Compactness

Scaling Mode
○ mean centering
● standardization

Number of Factors: 2

List of Response Variables [1 of 16]			
##	Type	Name	Show
12	Interval	Cultivar.BOMBAY	●

Summary | Classification Results | Cross Validation | Loadings X | Reg.Coeffs. | Details

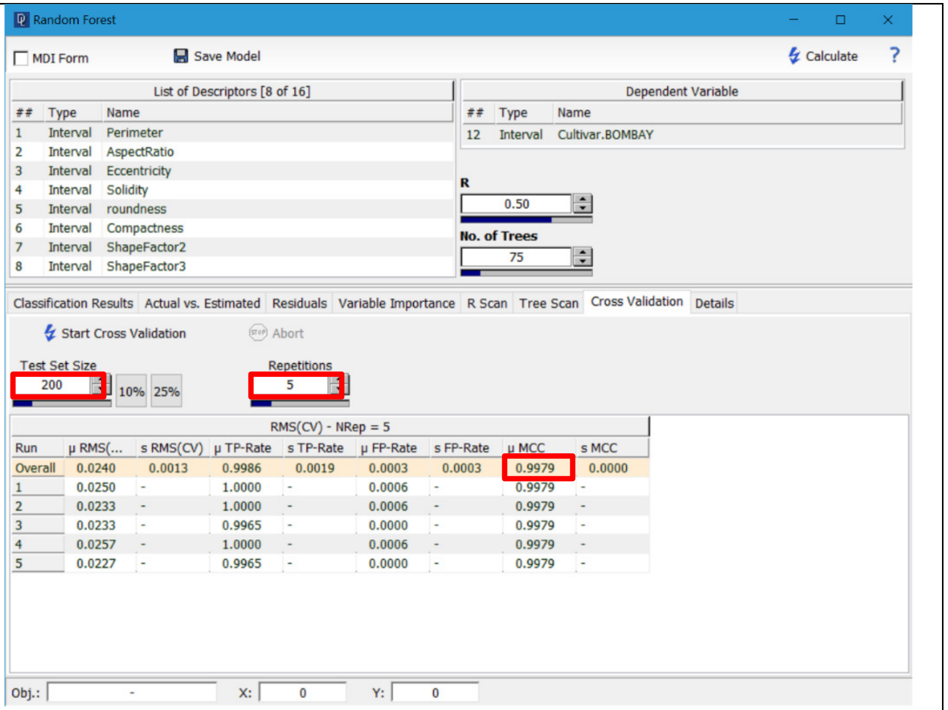
Start Cross Validation

Size of Test Set: 200
Starting Factor: 1
No. of Repetitions: 5

Abort   Protocol

Overall RMSEP		
#Fact	Mean Er...	StdDev ...
1	0.0747	0.0058
2	0.0000	0.0000
3	0.0000	0.0000
4	0.0000	0.0000
5	0.0234	0.0137
6	0.0298	0.0041
7	0.0224	0.0000
8	0.0224	0.0000

Cultivars BOMBAY shows for a test set of 10% (200) and 5 repetition an MCC of 0.9979 by calculation whit PLS/DA.



Random Forest

□ MDI Form   Save Model   Calculate   ?

List of Descriptors [8 of 16]		
##	Type	Name
1	Interval	Perimeter
2	Interval	AspectRatio
3	Interval	Eccentricity
4	Interval	Solidity
5	Interval	roundness
6	Interval	Compactness
7	Interval	ShapeFactor2
8	Interval	ShapeFactor3

Dependent Variable		
##	Type	Name
12	Interval	Cultivar.BOMBAY

R: 0.50
No. of Trees: 75

Classification Results | Actual vs. Estimated | Residuals | Variable Importance | R Scan | Tree Scan | Cross Validation | Details

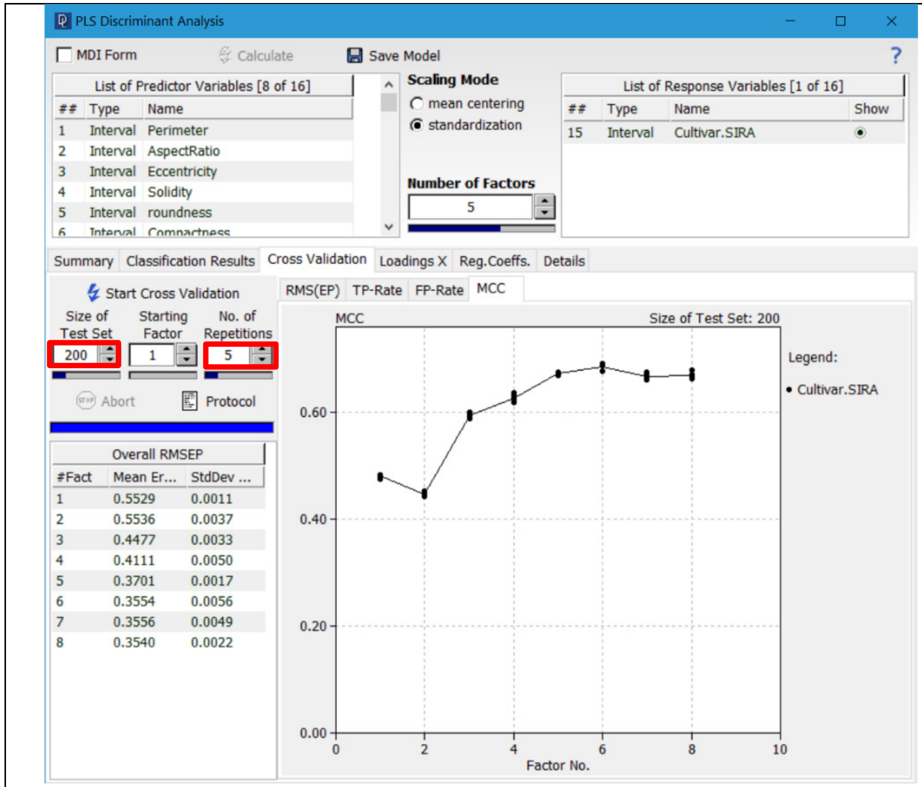
Start Cross Validation   Abort

Test Set Size: 200   10%   25%
Repetitions: 5

RMS(CV) - NRep = 5								
Run	μ RMS(...	s RMS(CV)	μ TP-Rate	s TP-Rate	μ FP-Rate	s FP-Rate	μ MCC	s MCC
Overall	0.0240	0.0013	0.9986	0.0019	0.0003	0.0003	0.9979	0.0000
1	0.0250	-	1.0000	-	0.0006	-	0.9979	-
2	0.0233	-	1.0000	-	0.0006	-	0.9979	-
3	0.0233	-	0.9965	-	0.0000	-	0.9979	-
4	0.0257	-	1.0000	-	0.0006	-	0.9979	-
5	0.0227	-	0.9965	-	0.0000	-	0.9979	-

Obj.: -   X: 0   Y: 0

Cultivars BOMBAY shows for a test set of 10% (200) and 5 repetition an MCC of 0.9979 by calculation whit RF.



PLS Discriminant Analysis

□ MDI Form | Calculate | Save Model | ?

List of Predictor Variables [8 of 16]

##	Type	Name
1	Interval	Perimeter
2	Interval	AspectRatio
3	Interval	Eccentricity
4	Interval	Solidity
5	Interval	roundness
6	Interval	Compactness

Scaling Mode
○ mean centering
● standardization

Number of Factors: 5

List of Response Variables [1 of 16]

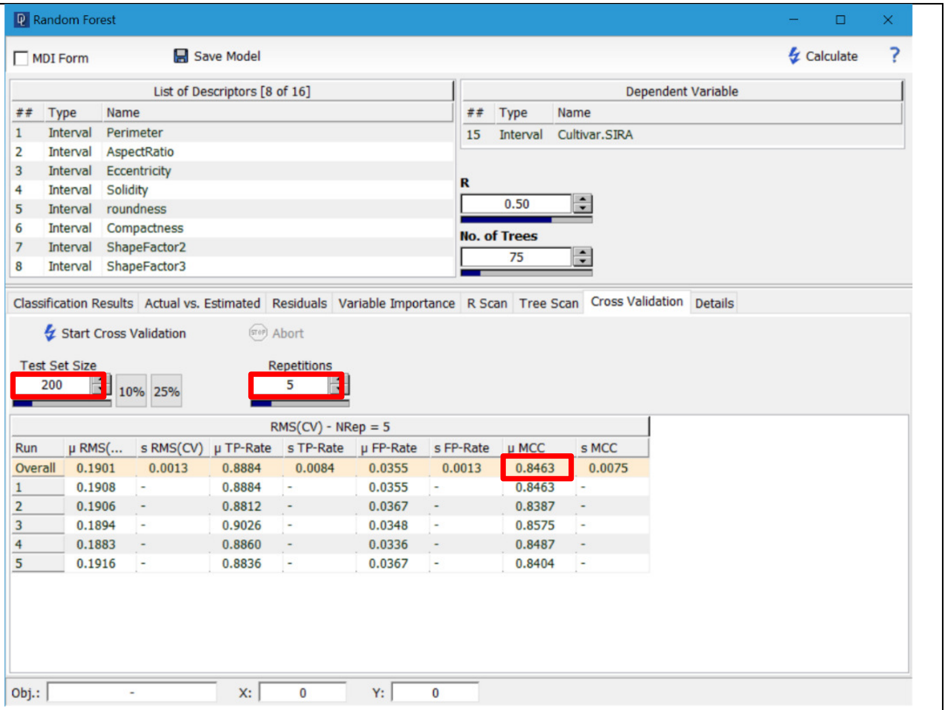
##	Type	Name	Show
15	Interval	Cultivar.SIRA	●

Size of Test Set: 200
Starting Factor: 1
No. of Repetitions: 5

Overall RMSEP

#Fact	Mean Er...	StdDev ...
1	0.5529	0.0011
2	0.5536	0.0037
3	0.4477	0.0033
4	0.4111	0.0050
5	0.3701	0.0017
6	0.3554	0.0056
7	0.3556	0.0049
8	0.3540	0.0022

Cultivars SIRA shows for a test set of 10% (200) and 5 repetition an MCC of 0.6715 by calculation whit PLS/DA.



Random Forest

□ MDI Form | Save Model | Calculate | ?

List of Descriptors [8 of 16]

##	Type	Name
1	Interval	Perimeter
2	Interval	AspectRatio
3	Interval	Eccentricity
4	Interval	Solidity
5	Interval	roundness
6	Interval	Compactness
7	Interval	ShapeFactor2
8	Interval	ShapeFactor3

Dependent Variable

##	Type	Name
15	Interval	Cultivar.SIRA

R: 0.50
No. of Trees: 75

Test Set Size: 200 | 10% | 25%
Repetitions: 5

RMS(CV) - NRep = 5

Run	μ RMS(...	s RMS(CV)	μ TP-Rate	s TP-Rate	μ FP-Rate	s FP-Rate	μ MCC	s MCC
Overall	0.1901	0.0013	0.8884	0.0084	0.0355	0.0013	0.8463	0.0075
1	0.1908	-	0.8884	-	0.0355	-	0.8463	-
2	0.1906	-	0.8812	-	0.0367	-	0.8387	-
3	0.1894	-	0.9026	-	0.0348	-	0.8575	-
4	0.1883	-	0.8860	-	0.0336	-	0.8487	-
5	0.1916	-	0.8836	-	0.0367	-	0.8404	-

Obj.: - X: 0 Y: 0

Cultivars SIRA shows for a test set of 10% (200) and 5 repetition an MCC of 0.8463 by calculation whit RF.

Why is RF based classifier for class 6 and 1 (cultivars SIRA and SEKER) considerably better than the corresponding PLS/DA classifier?

Based on the MCC value, classification by RF is more suitable for SEKER and SIRA. As there is a high MCC value for these two bean types compared to PLS/DA.
When examining the geometric attributes of the SIRA and SEKER bean crops, it is evident that they have very high overlap. This means that PLA/DA cannot be used to clearly sort between these beans. However, since RF carries out several sorting with different thresholds, these similarities can be better differentiated and the two bean types can be better classified.

Why performs PLS/DA better than RF for class 3 (cultivar BOMBAY)?

When comparing the MCC for PLS/DA and RF, no significant difference can be found. The only reason why PLS/DA is better for the BOMBAY variety is that this type of bean is very distinctive due to its geometric shapes and can therefore be easily separated from the others. PLS/DA is therefore a less intensive and simple variant for the classification of this bean species.

What are the drawback of using too many or too few trees in an RF classifier?

If too few trees are used, it can happen that the beans are not sorted by variety. Therefore, many beans of a different variety are classified as e.g. SEKER.

If too many trees are used, the calculations take a very long time and we run in a “over fitting”, but further optimisation is no longer possible.

Problem 8: Clustering of Red Wines
Report

Problem description:

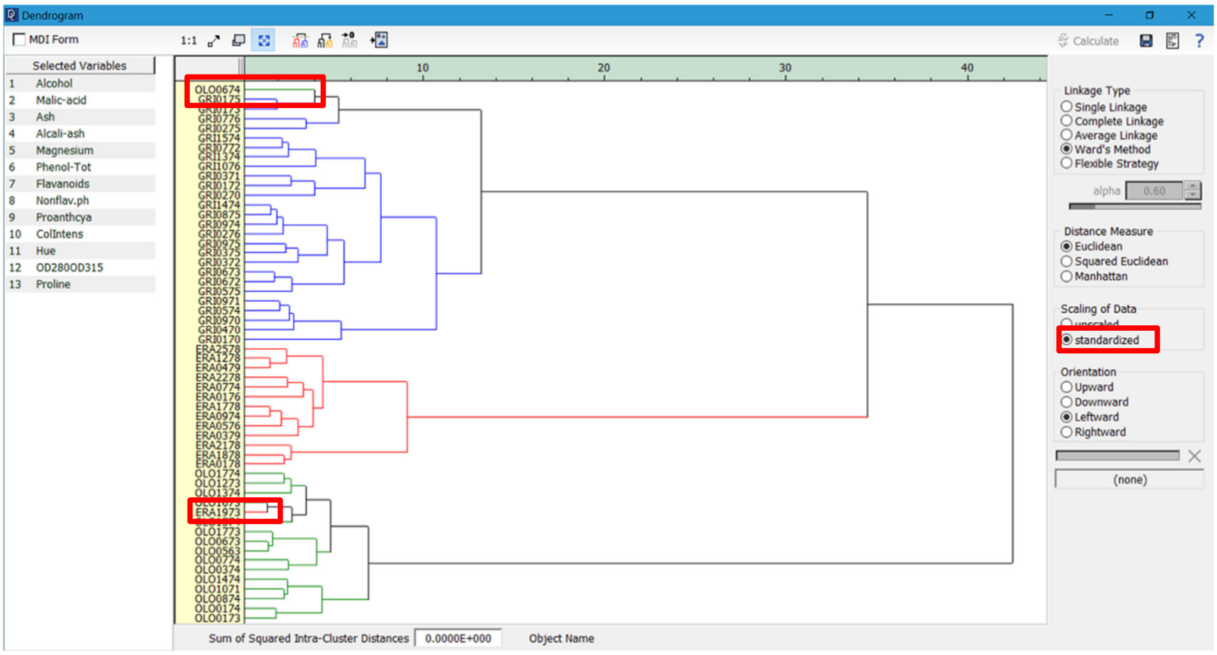
We have a data set of n=728 values. This is divided into 13 columns. Each column represents a specific ingredient that occurs in a certain concentration in different wines. 56 wines were measured in terms of their composition. We should use this data to perform a clustering to detect adulterated wines.

Outcome:

Dendrograms are often used for displaying relationships among clusters. A dendrogram shows the multidimensional distances between objects in a tree-like structure. Objects which are closest to each other in the multidimensional data space are connected by a horizontal line, forming a cluster which can be regarded as a "new" object. The new cluster and the remaining original data are again searched for the closest pair, and so on. The distance of the particular pair of objects (or clusters) is reflected in the height of the horizontal line. Dendrograms are heavily dependant upon the measure used to calculate the distances between the objects.

Do you find any misclassification (i.e. a wine which belongs of the "wrong" subtree of the dendrogram)?

Toolbar "hierarchical cluster analysis" → select "Ward's Method", "Euclidean", "standardized" and "Leftward" → select "selected Variables → Calculate



Dendrogram

□ MDI Form

Selected Variables	
1	Alcohol
2	Malic-acid
3	Ash
4	Alcali-ash
5	Magnesium
6	Phenol-Tot
7	Flavanoids
8	Nonflav.ph
9	Proanthcya
10	ColIntens
11	Hue
12	OD280OD315
13	Proline

Linkage Type
○ Single Linkage
○ Complete Linkage
○ Average Linkage
● Ward's Method
○ Flexible Strategy

alpha 0.60

Distance Measure
● Euclidean
○ Squared Euclidean
○ Manhattan

Scaling of Data
○ unscaled
● standardized

Orientation
○ Upward
○ Downward
● Leftward
○ Rightward

(none)

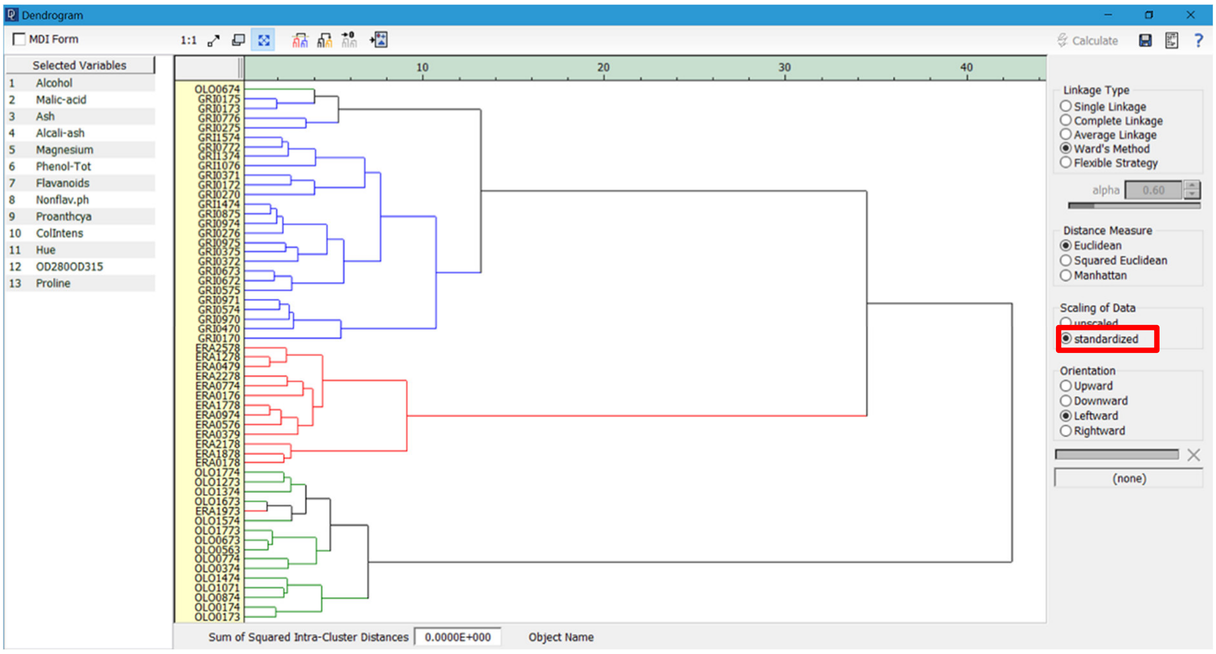
Sum of Squared Intra-Cluster Distances 0.0000E+000 Object Name

Figure 1: Dendrogram fort the wines data set. The scaling of the data is set as "standardized".

Two misclassifications can be identified from the dendrogram (see Figure 1). One is the wine OLO0674, which belongs to the Borolo wines but is classified among the Grinolino. Secondly, the wine ERA1973 which belongs to the Barolo wines but is a Barbera wine.

Compare the dendrograms obtained from standardized data and unscaled data. Why are there so many misclassifications in the case of unscaled data?

Toolbar "hierarchical cluster analysis" → select "Ward's Method", "Euclidean", "unscaled" and "Leftward" → select "selected Variables → Calculate



Dendrogram

☐ MDI Form

Selected Variables
1 Alcohol
2 Malic-acid
3 Ash
4 Alcali-ash
5 Magnesium
6 Phenol-Tot
7 Flavanoids
8 Nonflav.ph
9 Proanthcya
10 ColIntens
11 Hue
12 OD280OD315
13 Proline

Linkage Type
○ Single Linkage
○ Complete Linkage
○ Average Linkage
● Ward's Method
○ Flexible Strategy

alpha 0.60

Distance Measure
● Euclidean
○ Squared Euclidean
○ Manhattan

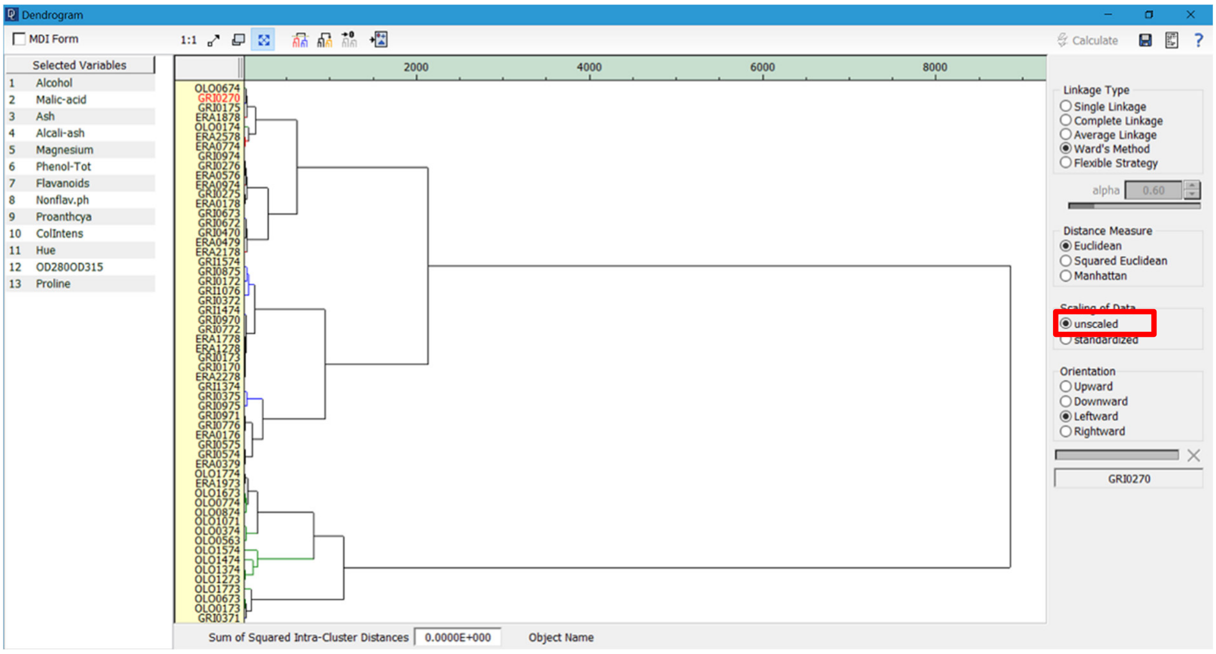
Scaling of Data
○ unscaled
● standardized

Orientation
○ Upward
○ Downward
● Leftward
○ Rightward

(none)

Sum of Squared Intra-Cluster Distances 0.0000E+000 Object Name

Figure 2: Dendrogram fort the wines data set. The scaling of the data is set as "standardized".



Dendrogram

☐ MDI Form

Selected Variables
1 Alcohol
2 Malic-acid
3 Ash
4 Alcali-ash
5 Magnesium
6 Phenol-Tot
7 Flavanoids
8 Nonflav.ph
9 Proanthcya
10 ColIntens
11 Hue
12 OD280OD315
13 Proline

Linkage Type
○ Single Linkage
○ Complete Linkage
○ Average Linkage
● Ward's Method
○ Flexible Strategy

alpha 0.60

Distance Measure
● Euclidean
○ Squared Euclidean
○ Manhattan

Scaling of Data
● unscaled
○ standardized

Orientation
○ Upward
○ Downward
● Leftward
○ Rightward

GRI0270

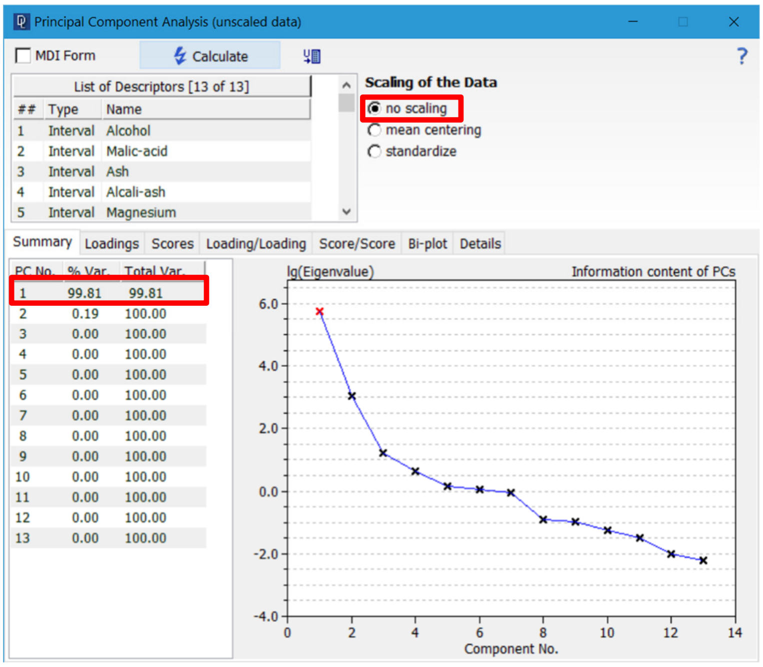
Sum of Squared Intra-Cluster Distances 0.0000E+000 Object Name

Figure 3: Dendrogram fort the wines data set. The scaling of the data is set as "unscaled".

Most misclassifications are made with Barbera wines, although other wines are also very often misclassified. The reason for the large number of misclassifications in the unscaled case is that variables with large values are dominating the calculation of the Euclidean distance and thus the clustering.

What do you think are the most influential variables when using unscaled data?

Toolbar PCA → select “List of Descriptors” → select “no scaling” → Calculate → Registry card “Summary” → Registry card “Loadings”



Principal Component Analysis (unscaled data)

□ MDI Form Calculate

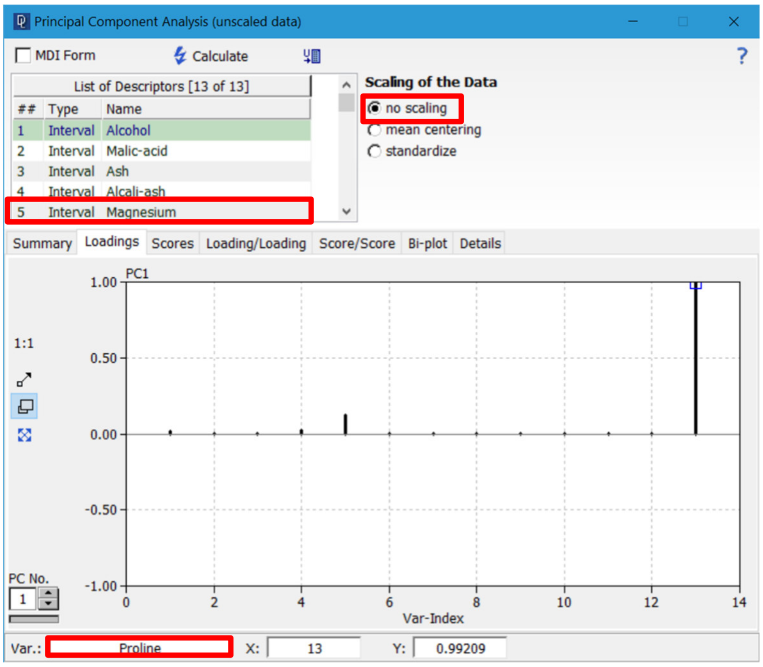
List of Descriptors [13 of 13]

##	Type	Name
1	Interval	Alcohol
2	Interval	Malic-acid
3	Interval	Ash
4	Interval	Alcali-ash
5	Interval	Magnesium

Scaling of the Data
● no scaling
○ mean centering
○ standardize

PC No.	% Var.	Total Var.
1	99.81	99.81
2	0.19	100.00
3	0.00	100.00
4	0.00	100.00
5	0.00	100.00
6	0.00	100.00
7	0.00	100.00
8	0.00	100.00
9	0.00	100.00
10	0.00	100.00
11	0.00	100.00
12	0.00	100.00
13	0.00	100.00

Figure 4: PCA for the wine data.



Principal Component Analysis (unscaled data)

□ MDI Form Calculate

List of Descriptors [13 of 13]

##	Type	Name
1	Interval	Alcohol
2	Interval	Malic-acid
3	Interval	Ash
4	Interval	Alcali-ash
5	Interval	Magnesium

Scaling of the Data
● no scaling
○ mean centering
○ standardize

Var.: Proline X: 13 Y: 0.99209

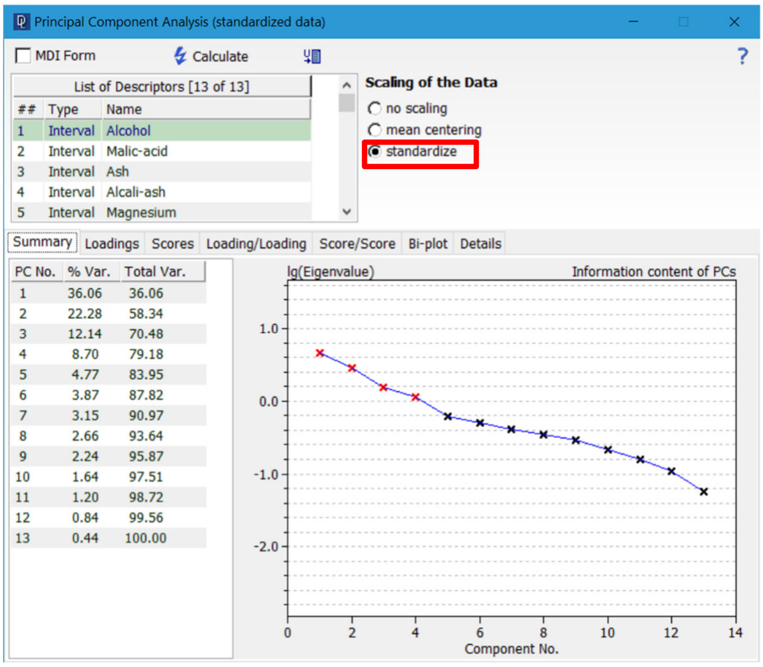
Figure 5: Loadings of the PCA analysis.

When looking at the evaluation of the PCA, it can be seen that PC No. 1 is already responsible for 99.81% of the deviations. If we now look at the "loadings", we can see that PC No. 1 is "Proline". Also PC No. 5 "Magnesium" have a high influence.

Thus, it can be said that "proline" and "magnesium", both variables with large values, make a significant contribution to the calculation of the Euclidean distance and thus dominate the clustering. Thus, the largest misclassifications occur in the unscaled case.

Try to use PCA to get an idea how severe the misclassification is. What are your findings?

Toolbar PCA → select "List of Descriptors" → select "standardize" → Calculate → Registry card "Summary" → Registry card "Loadings"



Principal Component Analysis (standardized data)

□ MDI Form Calculate

List of Descriptors [13 of 13]

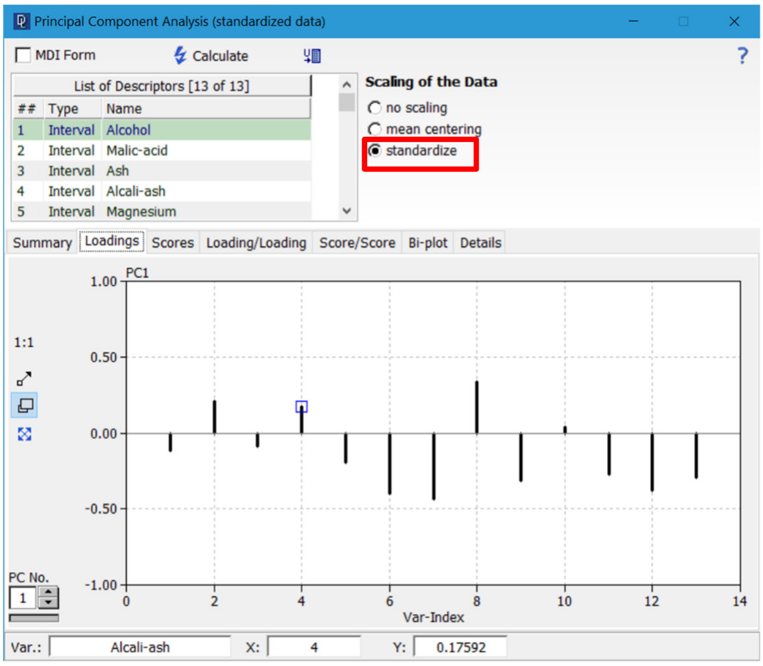
##	Type	Name
1	Interval	Alcohol
2	Interval	Malic-acid
3	Interval	Ash
4	Interval	Alcali-ash
5	Interval	Magnesium

Scaling of the Data

○ no scaling
○ mean centering
● standardize

PC No.	% Var.	Total Var.
1	36.06	36.06
2	22.28	58.34
3	12.14	70.48
4	8.70	79.18
5	4.77	83.95
6	3.87	87.82
7	3.15	90.97
8	2.66	93.64
9	2.24	95.87
10	1.64	97.51
11	1.20	98.72
12	0.84	99.56
13	0.44	100.00

Figure 6: PCA for the wine data.



Principal Component Analysis (standardized data)

□ MDI Form Calculate

List of Descriptors [13 of 13]

##	Type	Name
1	Interval	Alcohol
2	Interval	Malic-acid
3	Interval	Ash
4	Interval	Alcali-ash
5	Interval	Magnesium

Scaling of the Data

○ no scaling
○ mean centering
● standardize

Summary | Loadings | Scores | Loading/Loading | Score/Score | Bi-plot | Details

Var.: Alcali-ash X: 4 Y: 0.17592

Figure 7: Loadings of the PCA analysis.

If we now look at the PCA for standardize, we can see that all variables are weighted approximately equally in terms of their influence on the Euclidean distance and thus there is very little misclassification.

Problem 9: Metal inhibition of oxygen uptake
Report

Problem description:

We have a data set of n=80 values. This is divided into 5 columns. The 5 columns represent 4 metal concentrations in mg/L and the resulting O_2 uptake. In each case 16 measurements were taken with varying metal concentrations. We should use these data to find out which of the metals or metal combinations have an influence on the O_2 uptake of sludge.

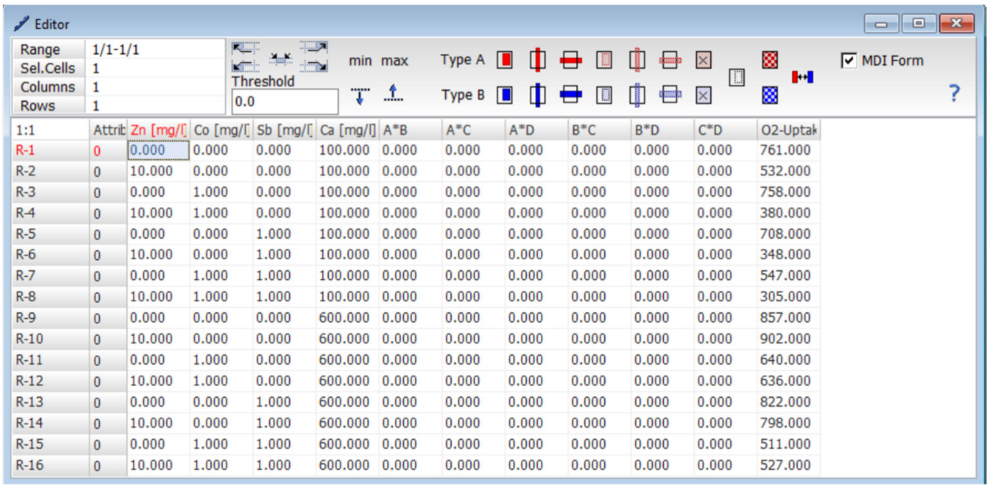
Outcome:

Factorial experimental designs are used to identify interactions. A basic distinction is made between single factorial and multifactorial designs.

Calculate and interpret the main and interaction effects of the four metals.

Editor → Menu tab "Tools" → "Dichotomization…" → set parameter and calculate → open "Editor" → right Mouse klick "Insert Column" → right Mouse klick "Edit Heading" → Toolbar f(x) → set in the Formula → Executed → Toolbar MLR → select "Dependent Variable" and "List of Descriptors" → Calculate → Registry card "Details"

Metals	Treatment [mg/L]	
	-1	+1
A = Zn (Zinc)	0	10
B = Co (Cobalt)	0	1
C = Sb (Antimony)	0	1
D = Ca (Calcium)	100	600
Calculations rules		
-1 * -1 = +1 +1 * +1 = +1	-1 * +1 = -1 +1 * -1 = -1	



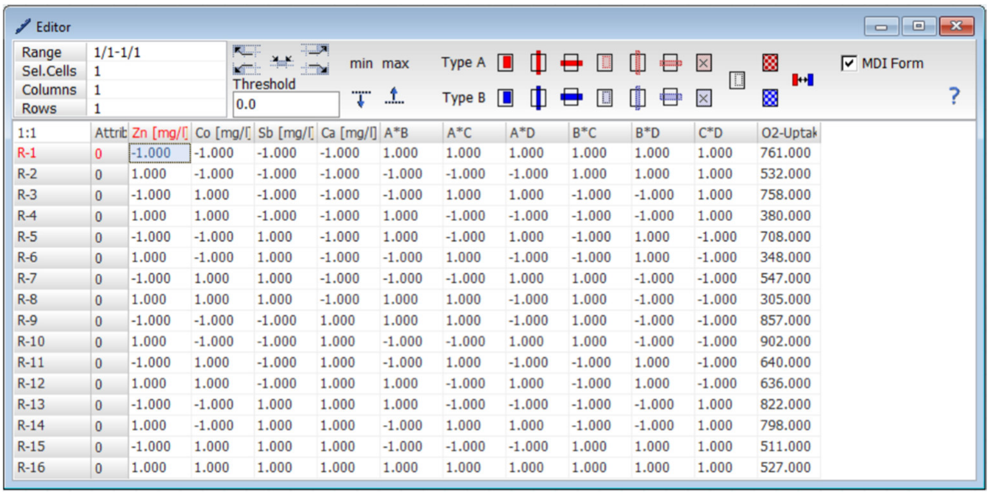
Range	1/1-1/1
Sel.Cells	1
Columns	1
Rows	1

Threshold 0.0

☑ MDI Form

1:1	Attrib	Zn [mg/l]	Co [mg/l]	Sb [mg/l]	Ca [mg/l]	A*B	A*C	A*D	B*C	B*D	C*D	O2-Uptak
R-1	0	0.000	0.000	0.000	100.000	0.000	0.000	0.000	0.000	0.000	0.000	761.000
R-2	0	10.000	0.000	0.000	100.000	0.000	0.000	0.000	0.000	0.000	0.000	532.000
R-3	0	0.000	1.000	0.000	100.000	0.000	0.000	0.000	0.000	0.000	0.000	758.000
R-4	0	10.000	1.000	0.000	100.000	0.000	0.000	0.000	0.000	0.000	0.000	380.000
R-5	0	0.000	0.000	1.000	100.000	0.000	0.000	0.000	0.000	0.000	0.000	708.000
R-6	0	10.000	0.000	1.000	100.000	0.000	0.000	0.000	0.000	0.000	0.000	348.000
R-7	0	0.000	1.000	1.000	100.000	0.000	0.000	0.000	0.000	0.000	0.000	547.000
R-8	0	10.000	1.000	1.000	100.000	0.000	0.000	0.000	0.000	0.000	0.000	305.000
R-9	0	0.000	0.000	0.000	600.000	0.000	0.000	0.000	0.000	0.000	0.000	857.000
R-10	0	10.000	0.000	0.000	600.000	0.000	0.000	0.000	0.000	0.000	0.000	902.000
R-11	0	0.000	1.000	0.000	600.000	0.000	0.000	0.000	0.000	0.000	0.000	640.000
R-12	0	10.000	1.000	0.000	600.000	0.000	0.000	0.000	0.000	0.000	0.000	636.000
R-13	0	0.000	0.000	1.000	600.000	0.000	0.000	0.000	0.000	0.000	0.000	822.000
R-14	0	10.000	0.000	1.000	600.000	0.000	0.000	0.000	0.000	0.000	0.000	798.000
R-15	0	0.000	1.000	1.000	600.000	0.000	0.000	0.000	0.000	0.000	0.000	511.000
R-16	0	10.000	1.000	1.000	600.000	0.000	0.000	0.000	0.000	0.000	0.000	527.000

Figure 1: Table of the Data bevor Dichotomization.



Range	1/1-1/1
Sel.Cells	1
Columns	1
Rows	1

Threshold: 0.0

☑ MDI Form

1:1	Attrib	Zn [mg/l]	Co [mg/l]	Sb [mg/l]	Ca [mg/l]	A*B	A*C	A*D	B*C	B*D	C*D	O2-Uptak
R-1	0	-1.000	-1.000	-1.000	-1.000	1.000	1.000	1.000	1.000	1.000	1.000	761.000
R-2	0	1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	1.000	1.000	1.000	532.000
R-3	0	-1.000	1.000	-1.000	-1.000	-1.000	1.000	1.000	-1.000	-1.000	1.000	758.000
R-4	0	1.000	1.000	-1.000	-1.000	1.000	-1.000	-1.000	-1.000	-1.000	1.000	380.000
R-5	0	-1.000	-1.000	1.000	-1.000	1.000	-1.000	1.000	-1.000	1.000	-1.000	708.000
R-6	0	1.000	-1.000	1.000	-1.000	-1.000	1.000	-1.000	-1.000	1.000	-1.000	348.000
R-7	0	-1.000	1.000	1.000	-1.000	-1.000	-1.000	1.000	1.000	-1.000	-1.000	547.000
R-8	0	1.000	1.000	1.000	-1.000	1.000	1.000	-1.000	1.000	-1.000	-1.000	305.000
R-9	0	-1.000	-1.000	-1.000	1.000	1.000	1.000	-1.000	1.000	-1.000	-1.000	857.000
R-10	0	1.000	-1.000	-1.000	1.000	-1.000	-1.000	1.000	1.000	-1.000	-1.000	902.000
R-11	0	-1.000	1.000	-1.000	1.000	-1.000	1.000	-1.000	-1.000	1.000	-1.000	640.000
R-12	0	1.000	1.000	-1.000	1.000	1.000	-1.000	1.000	-1.000	1.000	-1.000	636.000
R-13	0	-1.000	-1.000	1.000	1.000	1.000	-1.000	-1.000	-1.000	-1.000	1.000	822.000
R-14	0	1.000	-1.000	1.000	1.000	-1.000	1.000	1.000	-1.000	-1.000	1.000	798.000
R-15	0	-1.000	1.000	1.000	1.000	-1.000	-1.000	-1.000	1.000	1.000	1.000	511.000
R-16	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	527.000

Figure 2: Table of the Data after Dichotomization.

Level of significance 5% = 0.05 **α** < 0.05

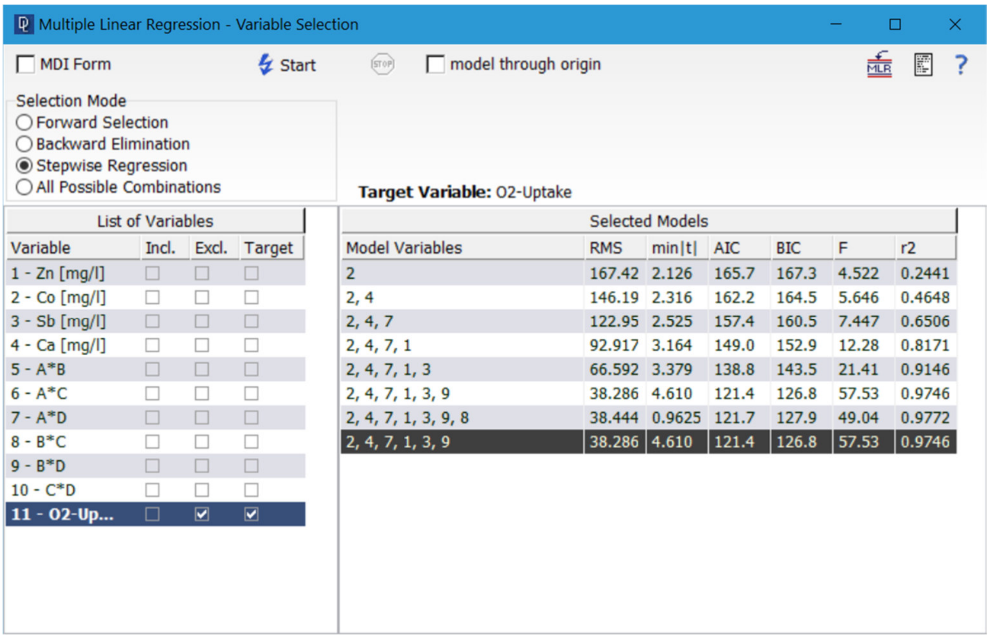
```
Regression coefficients:
Col  Var-Name          Coefficient +/- Std.Err.(coeff)     t-Test      alpha
 -------------------------------------------------------------------------
-    INTERCEPT         6.2700000E+02 +/- 1.1330269E+01     55.338     0.0000
1    Zn [mg/l]        -7.3500000E+01 +/- 1.1330269E+01     -6.487     0.0013
2    Co [mg/l]        -8.9000000E+01 +/- 1.1330269E+01     -7.855     0.0005
3    Sb [mg/l]        -5.6250000E+01 +/- 1.1330269E+01     -4.965     0.0042
4    Ca [mg/l]         8.4625000E+01 +/- 1.1330269E+01      7.469     0.0007
5    A*B              -2.5000000E+00 +/- 1.1330269E+01     -0.221     0.8341
6    A*C              -2.7500000E+00 +/- 1.1330269E+01     -0.243     0.8179
7    A*D               7.7625000E+01 +/- 1.1330269E+01      6.851     0.0010
8    B*C              -9.2500000E+00 +/- 1.1330269E+01     -0.816     0.4514
9    B*D              -4.4125000E+01 +/- 1.1330269E+01     -3.894     0.0115
10   C*D               9.1250000E+00 +/- 1.1330269E+01      0.805     0.4572
```

```
Regression coefficients:
Col  Var-Name          Coefficient +/- Std.Err.(coeff)     t-Test      alpha
 -------------------------------------------------------------------------
-    INTERCEPT         6.2700000E+02 +/- 9.5714600E+00     65.507     0.0000
1    Zn [mg/l]        -7.3500000E+01 +/- 9.5714600E+00     -7.679     0.0000
2    Co [mg/l]        -8.9000000E+01 +/- 9.5714600E+00     -9.298     0.0000
3    Sb [mg/l]        -5.6250000E+01 +/- 9.5714600E+00     -5.877     0.0002
4    Ca [mg/l]         8.4625000E+01 +/- 9.5714600E+00      8.841     0.0000
7    A*D               7.7625000E+01 +/- 9.5714600E+00      8.110     0.0000
9    B*D              -4.4125000E+01 +/- 9.5714600E+00     -4.610     0.0013
```

We perform a backward elimination by hand. To do this, we exclude the descriptor with the highest alpha and recalculate it. We repeat this until we reach an alpha value of 0.05.
The result contains only Zn, Co, Sb, Ca, A*D and B*D.

Which of the four metals increase the oxygen uptake?

Toolbar MLR → select "Dependent Variable" and "List of Descriptors" → Calculate → click on "automatic variable selection" → chose "All Possible Combinations" → Start


Multiple Linear Regression - Variable Selection

□ MDI Form   Start   □ model through origin

Selection Mode
○ Forward Selection
○ Backward Elimination
● Stepwise Regression
○ All Possible Combinations

Target Variable: O2-Uptake

List of Variables

Variable	Incl.	Excl.	Target
1 - Zn [mg/l]	□	□	□
2 - Co [mg/l]	□	□	□
3 - Sb [mg/l]	□	□	□
4 - Ca [mg/l]	□	□	□
5 - A*B	□	□	□
6 - A*C	□	□	□
7 - A*D	□	□	□
8 - B*C	□	□	□
9 - B*D	□	□	□
10 - C*D	□	□	□
11 - O2-Up...	□	☑	☑

Selected Models

Model Variables	RMS	min|t|	AIC	BIC	F	r2
2	167.42	2.126	165.7	167.3	4.522	0.2441
2, 4	146.19	2.316	162.2	164.5	5.646	0.4648
2, 4, 7	122.95	2.525	157.4	160.5	7.447	0.6506
2, 4, 7, 1	92.917	3.164	149.0	152.9	12.28	0.8171
2, 4, 7, 1, 3	66.592	3.379	138.8	143.5	21.41	0.9146
2, 4, 7, 1, 3, 9	38.286	4.610	121.4	126.8	57.53	0.9746
2, 4, 7, 1, 3, 9, 8	38.444	0.9625	121.7	127.9	49.04	0.9772
2, 4, 7, 1, 3, 9	38.286	4.610	121.4	126.8	57.53	0.9746

Figure 3: MLR – Variable Selection for the Metals.

In this calculation, it is assumed that these 6 variables (Zn, Co, Sb, Ca, A*D and B*D) have the greatest influence on oxygen uptake. This includes the Metals Zinc, Cobalt, Antimony and Calcium. So all metals have a high influence on the oxygenation uptake.

What would you recommend the manager of the sewage plant?

My recommendation would be to use Sb (Antimony) and add it. On the one hand this is because it has a great influence on the O_2 uptake, on the other hand we only need 1mg/L of this metal. Thus, we achieve a large effect with a relatively small amount, which of course also represents a reduction in material costs.

Zn (Zinc), Co (Cobalt) and Ca (Calcium) would also allow an increase in oxygenation, which would have the positive effect that the two metals work both individually and in combination (A*D and B*D). However, 100 - 600 mg/L of calcium will be needed, so this is not so recommend because of the high material costs.