



建立定量模型及模型评价



August 14, 2013

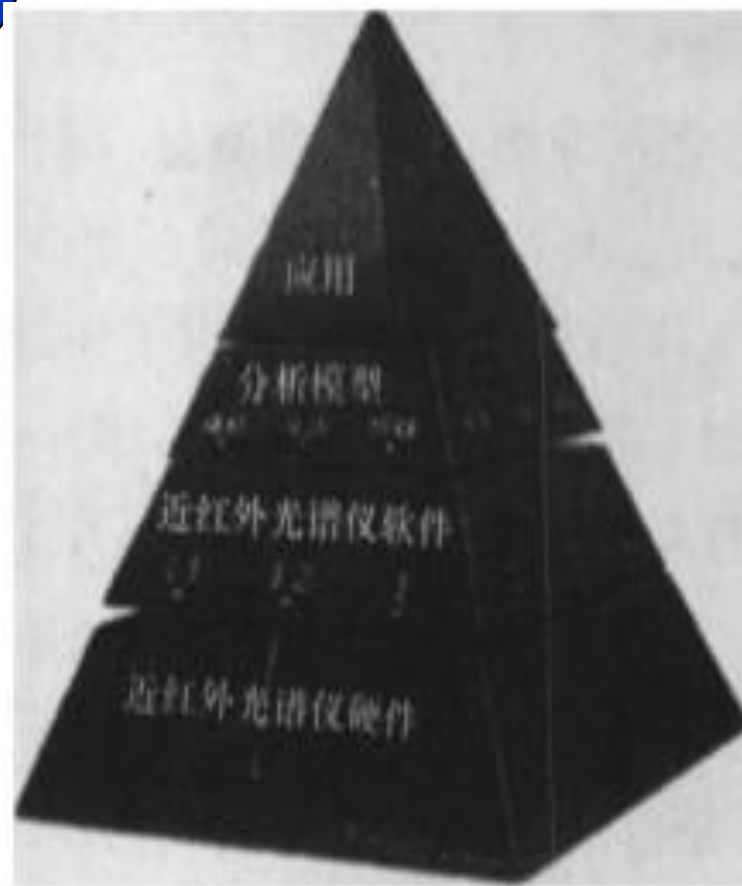
仪器/软件/模型的一体化

模型

化学计量学
软件

光谱仪

硬件的可靠性是模型稳定和技术应用的基础。



怎样建立好的近红外定量分析模型？



前提：可靠的光谱和化学值

1. 选择或配备代表性建模样品集。
2. 模型优化。
3. 模型评价（内部交叉验证和外部验证）。

选择建模样品



选择或配备代表性建模样品集。

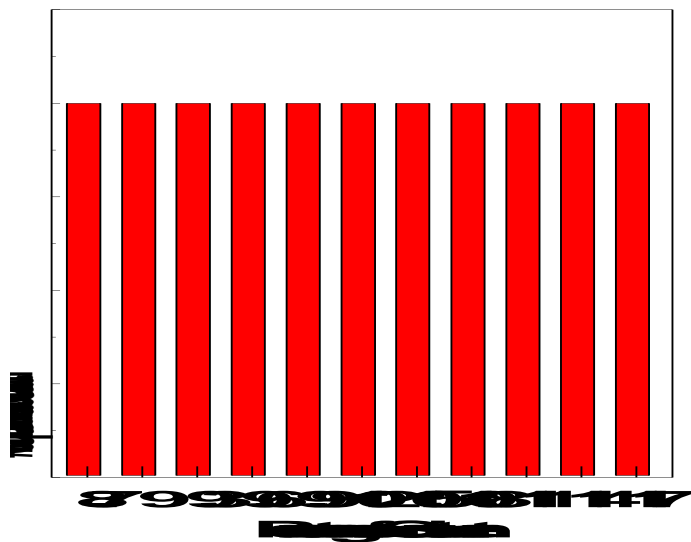
- 天然样品和反应过程中的样品，选择的代表性样品数量比较大。
- 成分已知样品，选择的代表性样品数量比较少。

可使用化学计量学方法对样品进行选择、设计。

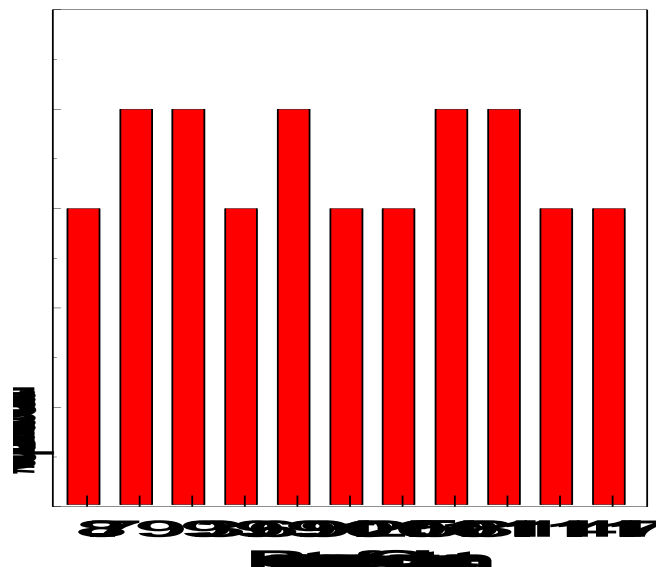
不同含量样品数的选择



不同含量样品数量选择不能以现实的样品变化进行选择，应该在所有含量范围内均匀选择。



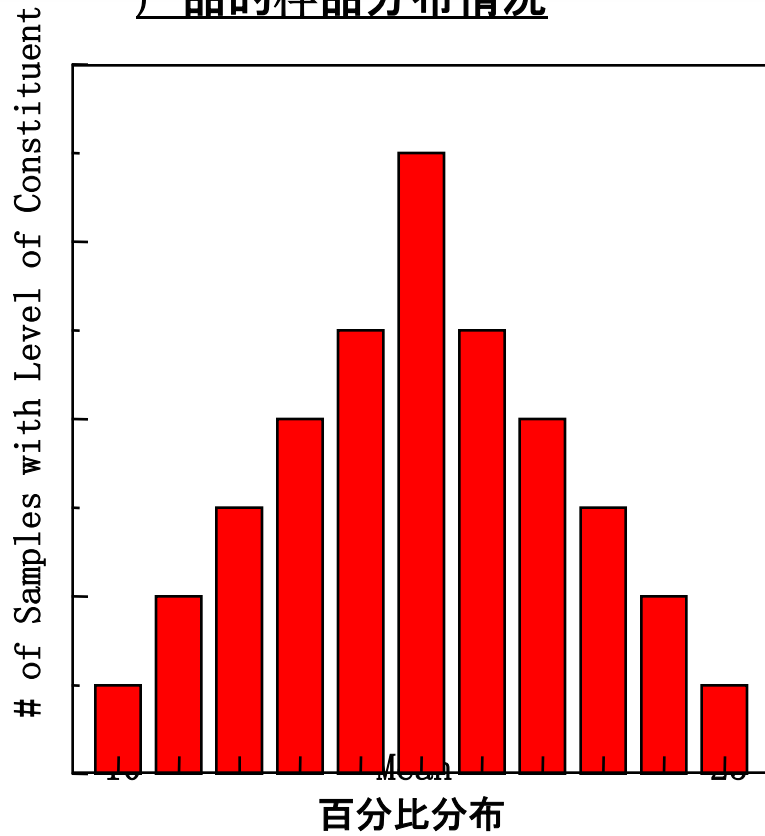
完美的分布
(很少出现的情况)



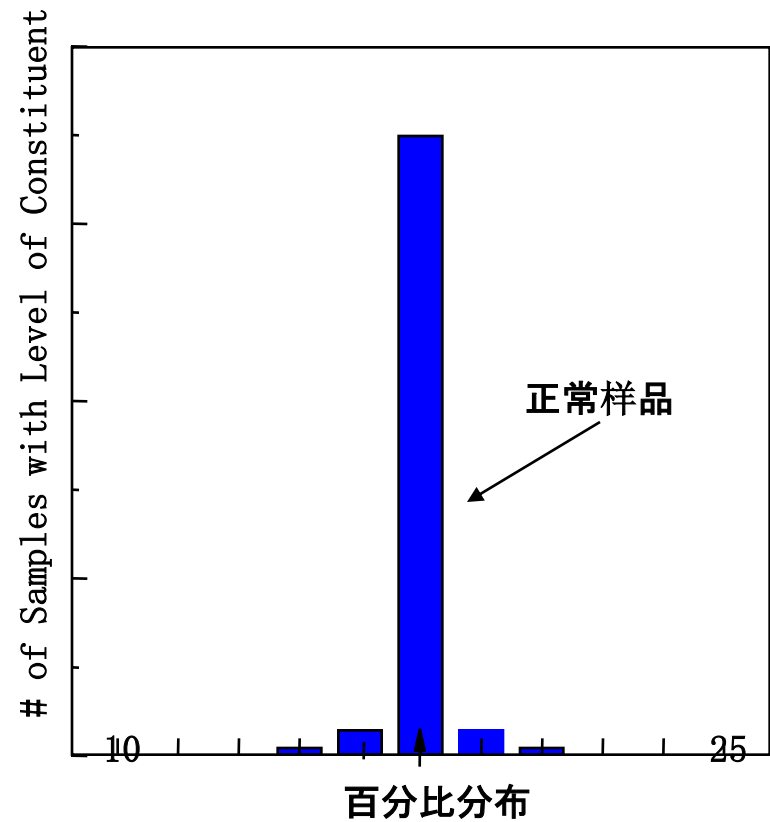
比较合理的分布

样品分布

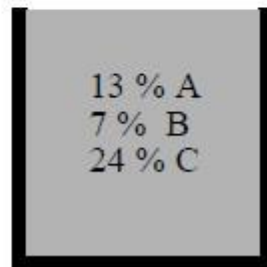
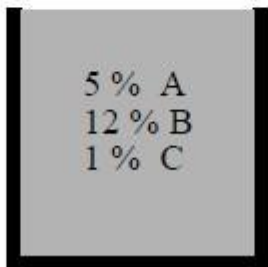
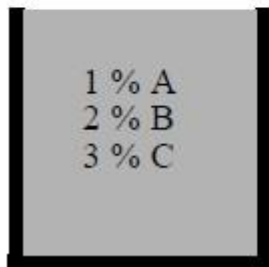
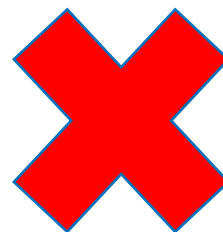
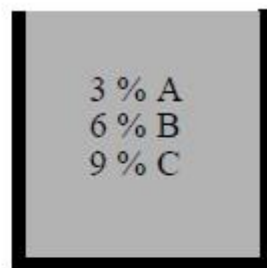
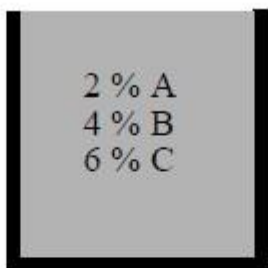
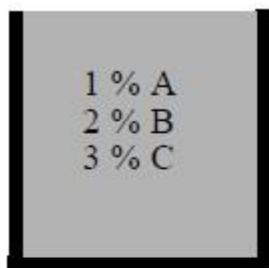
饲料、食品、烟草、化工、炼油等产品的样品分布情况



制药分布



避免样品的共线性

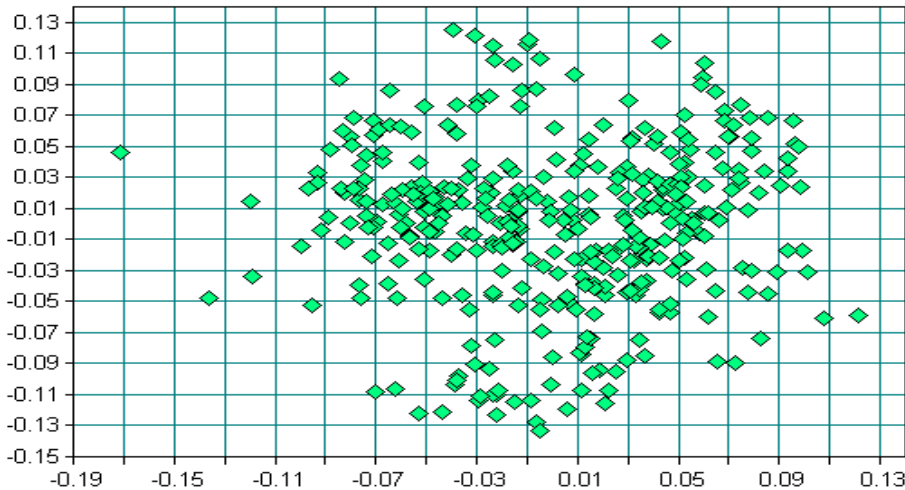


在实验室准备建模样品时，要保证这些样品的浓度不能是线性增加或减少。特别指出：稀释的样品不适合作建模样品。

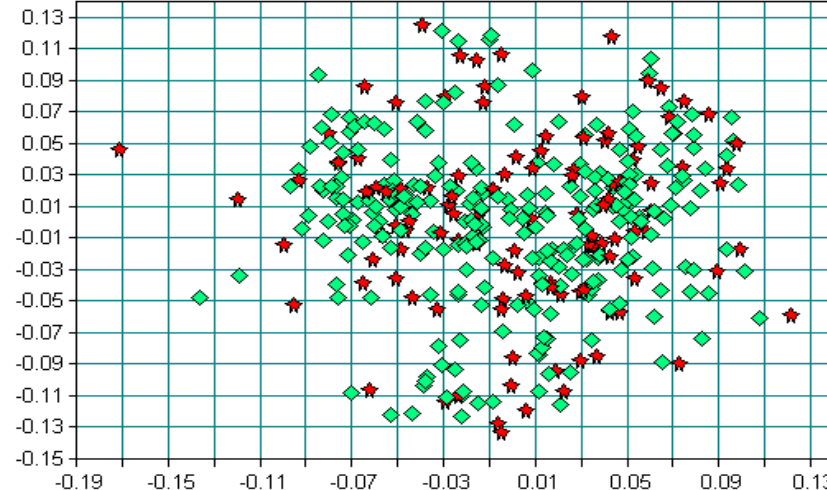
多余样品的挑选 (373选261)



Score 2 vs Score 1

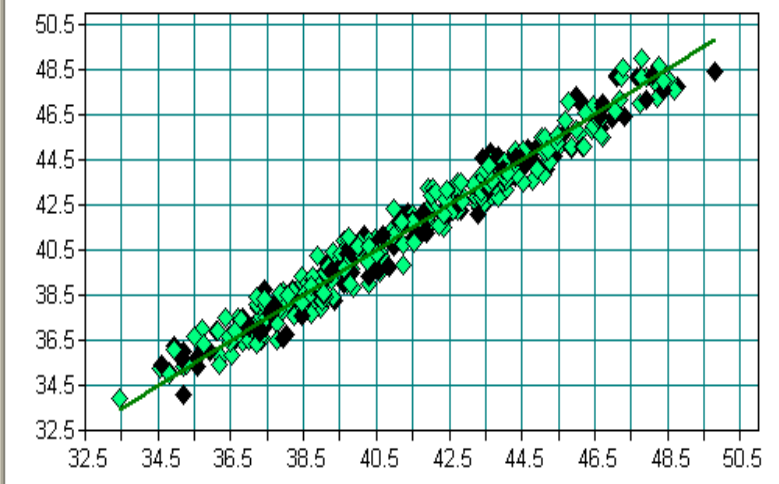


Score 2 vs Score 1



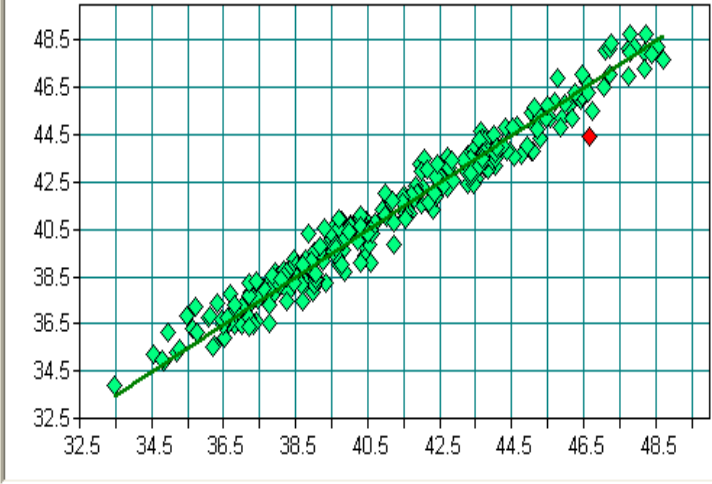
预测值 vs 真值 / 蛋白 [%] / 交叉检验

R²: 96.64
 RMSECV: 0.645
 RPD: 5.45
 偏移: -0.00199



预测值 vs 真值 / 蛋白 [%] / 交叉检验

R²: 96.47
 RMSECV: 0.647
 RPD: 5.32
 偏移: -9.38E-005



识别异常点



异常点：与大多数样品相比，其数据信息在某些方面具有较严重的错误或异常。

“好”异常点：

- 组分含量处于两端极值(最高或最低)：样品具有含量的极值信息，不应盲目删除
- 干扰组分含量处于两端极值(最高或最低)：事先未分析干扰组分
- 样品来源不同：分类、分段建模

“坏”异常点：

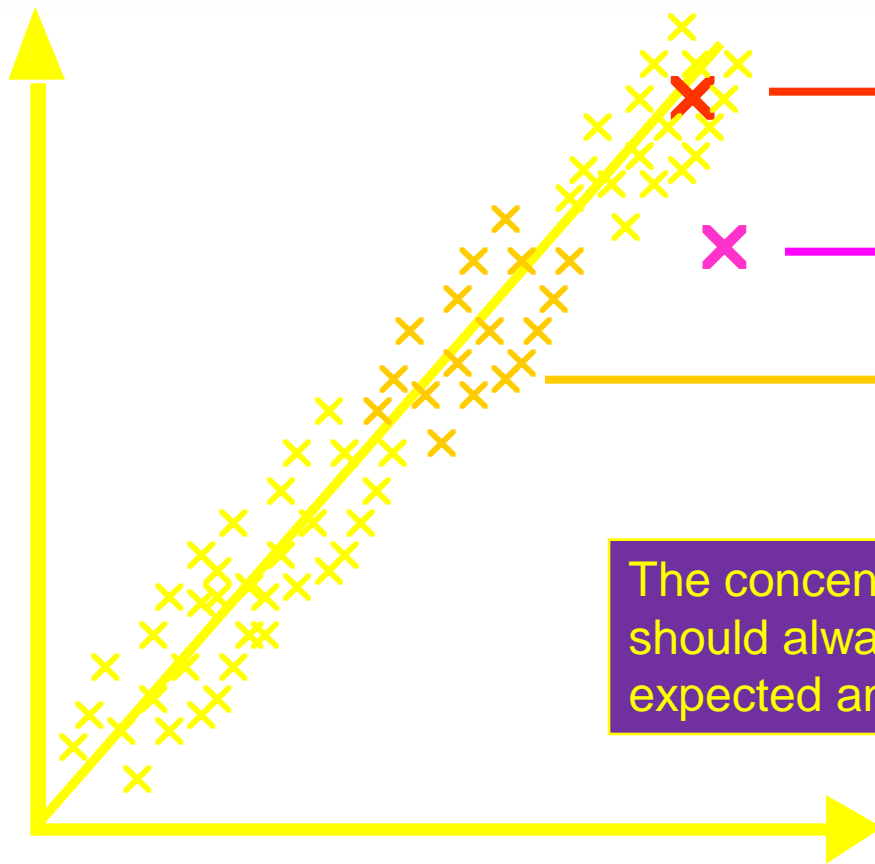
- 数据测试输入错误：
应加以更正或重新测定
- 光谱测量错误：
重新测定样品光谱
- 参照测量错误：
重新测定参照光谱

建模过程中，异常点是删除，还是更正？→应慎重
建模样品数量较少时，应避免过多的剔除异常点

异常样品的判断



Predicted value



**“extreme sample”
or good outlier
Useful for model
update**

**bad outlier, need to
Identify the causes**

**Typical
concentration range**

The concentration range of the calibration should always be larger than the expected analysis range.

Reference value

光谱残留：一个因子的结果永远不能完全描述光谱矩阵和含量矩阵的变量。因子没能描述的剩余部分被称为残留。

光谱残留的平方同整个其他平均值比较， F 值的计算公式如下：

$$FValue_i = \frac{(M - 1)(SpecRes_i)^2}{\sum_{j \neq i} (SpecRes_j)^2}$$

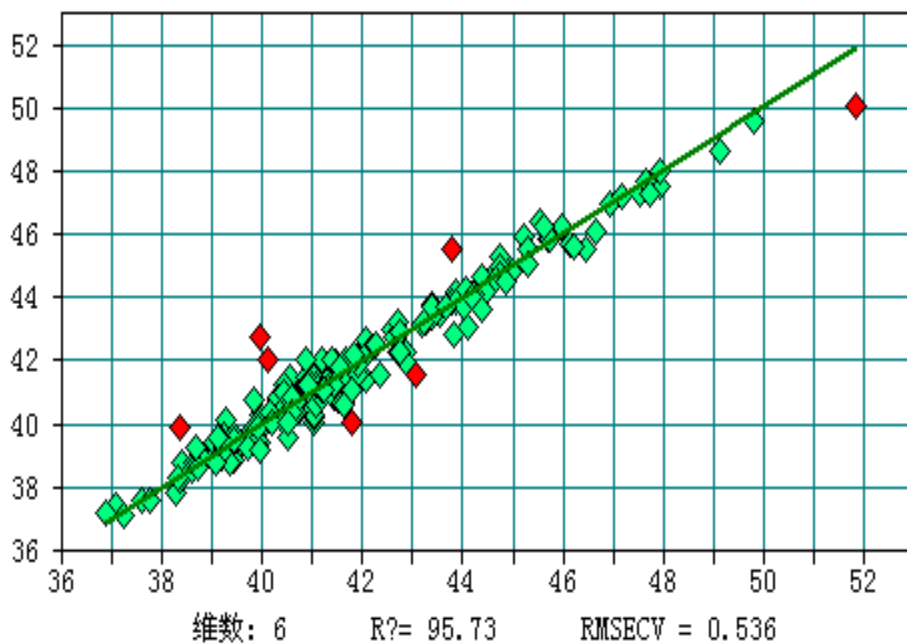
从 F 值和自由度数可以算出 F 概率， F 概率表示标准谱是异常项的概率。

自动侦测异常点的概率值为99%

剔除异常样品



预测值 vs 真值 / 粗蛋白 [%] / 交叉检验



图表页: 双击异常点
由红变黑 (由黑变
红), 以示剔除
(恢复)

怎样建立好的近红外定量分析模型？



1. 选择或配备代表性建模样品集。
2. 模型优化。
3. 模型评价（内部交叉验证和外部验证）。

模型优化

谱区范围的选择

选择信息量相关性最大的谱区。

光谱预处理方法的选择

选择减小光谱漂移、增加信息量、稳定可靠的光谱预处理方法。

最佳Rank（PLS算法）

相同的谱区范围，RMSECV无显著性差别，Rank小的模型好。

设置页：对谱区进行选择

建立定量 2 方法(0) - C:\Users\yanping.wang\Desktop\大豆\模型\数据.Q2

调入方法 组分 光谱 参数 检验 图表 报告 保存方法 优化 设置

图表页

标记点的大小 10

方法保护

在定量 2 方法文件中保存谱图

只有在“放大方法”或“改变参数”模式，想保护方法才使用该选项

选择优化的预处理选项

- 消除常数偏移量
- 减去一条直线
- 矢量归一化 (SNV)
- 最小-最大归一化
- 多元散射校正
- 一阶导数
- 二阶导数
- 一阶导数 + 减去一条直线
- 一阶导数 + 矢量归一化 (SNV)
- 一阶导数 + MSC

平滑点数: 17

最大测试范围:

11998.9 4247

交互式选择范围

在后台运行优化

使用定义的优化范围

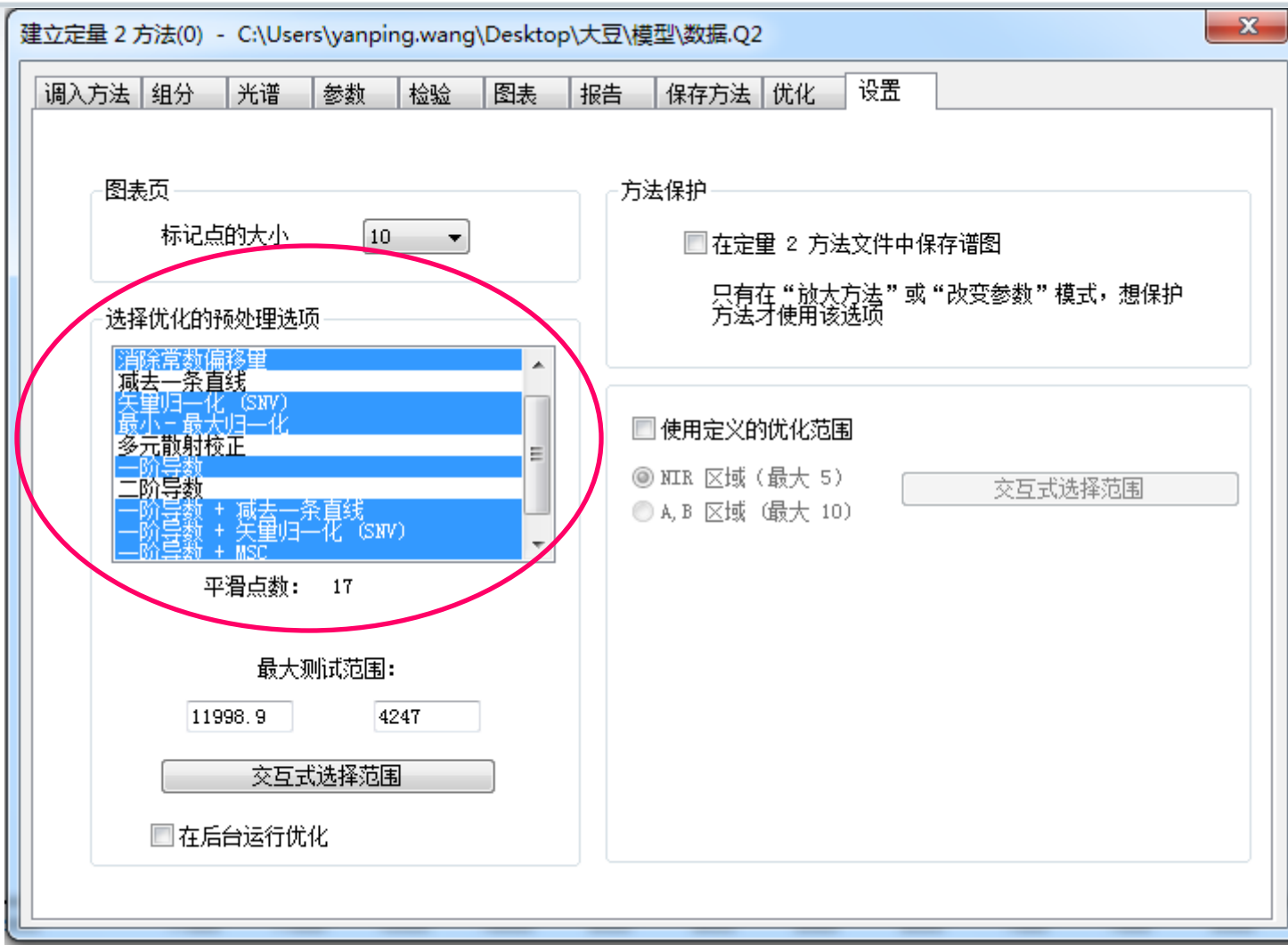
NIR 区域 (最大 5)

A, B 区域 (最大 10)

交互式选择范围

	起点	终点
1	10719.6	8819
2	8584.9	6601.6
3	6078.3	4769.9
4	4604.6	4177.7
5		

设置页：多种数据预处理方法可供选择



建立定量 2 方法(0) - C:\Users\yanping.wang\Desktop\大豆\模型\数据.Q2

调入方法 组分 光谱 参数 检验 图表 报告 保存方法 优化 设置

图表页

标记点的大小 10

选择优化的预处理选项

- 消除常数偏移量
- 减去一条直线
- 变量归一化 (SNV)
- 最小-最大归一化
- 多元散射校正
- 二阶导数
- 二阶导数
- 一阶导数 + 减去一条直线
- 一阶导数 + 变量归一化 (SNV)
- 一阶导数 + MSC

平滑点数: 17

最大测试范围:

11998.9 4247

交互式选择范围

在后台运行优化

方法保护

在定量 2 方法文件中保存谱图

只有在“放大方法”或“改变参数”模式，想保护方法才使用该选项

使用定义的优化范围

NIR 区域 (最大 5)

A, B 区域 (最大 10)

交互式选择范围

优化页：帮助完成自动优化检验

建立定量 2 方法(0) - C:\Users\yanping.wang\Desktop\大豆\模型\数据.Q2

调入方法 组分 光谱 参数 检验 图表 报告 保存方法 优化 设置

使用参数 蛋白质 NIR 优化

数值	RMSECV	维数	范围	预处理
88	0.441	6	7502 - 6098 5450 - 4246.6	最小 - 最大归一化
89	0.449	8	9403.5 - 6098 5450 - 4246.6	最小 - 最大归一化
90	0.433	6	6101.9 - 4246.6	最小 - 最大归一化
91	0.464	5	9403.5 - 7498.1 6101.9 - 4246.6	最小 - 最大归一化
92	0.467	5	7502 - 4246.6	最小 - 最大归一化
93	0.475	5	9403.5 - 4246.6	最小 - 最大归一化
94	0.607	4	9403.5 - 7498.1	一阶导数
95	0.567	3	7502 - 6098	一阶导数
96	0.526	4	9403.5 - 6098	一阶导数
97	0.578	4	6101.9 - 5446.2	一阶导数
98	0.586	3	9403.5 - 7498.1 6101.9 - 5446.2	一阶导数
99	0.538	4	7502 - 5446.2	一阶导数
100	0.534	3	9403.5 - 5446.2	一阶导数
101	0.495	5	5450 - 4597.6	一阶导数

优化状态

Step 1 / 交叉检验



优化页： 将所选的优化组合信息传递到**参数页**

建立定量 2 方法(0) - C:\Users\yanping.wang\Desktop\大豆\模型\数据.Q2

调入方法 组分 光谱 参数 检验 图表 报告 保存方法 优化 设置

使用参数 蛋白质 NIR 优化

数值	RMSECV	维数	范围	预处理
90	0.433	6	6101.9 - 4246.6	最小-最大归一化
28	0.434	7	6101.9 - 4246.6	消除常数偏移量
58	0.438	5	6101.9 - 4246.6	矢量归一化 (SNV)
88	0.441	6	7502 - 6098 5450 - 4246.6	最小-最大归一化
60	0.444	6	9403.5 - 7498.1 6101.9 - 4246.6	矢量归一化 (SNV)
53	0.447	7	7502 - 5446.2 4601.5 - 4246.6	矢量归一化 (SNV)
102	0.449	4	9403.5 - 7498.1 5450 - 4597.6	一阶导数
89	0.449	8	9403.5 - 6098 5450 - 4246.6	最小-最大归一化
57	0.453	5	7502 - 6098 5450 - 4246.6	矢量归一化 (SNV)
164	0.456	4	9403.5 - 7498.1 5450 - 4597.6	一阶导数 + 矢量归一化 (S
104	0.457	4	9403.5 - 6098 5450 - 4597.6	一阶导数
56	0.457	5	9403.5 - 7498.1 5450 - 4246.6	矢量归一化 (SNV)
62	0.458	4	9403.5 - 4246.6	矢量归一化 (SNV)
183	0.458	5	6101.9 - 4246.6	一阶导数 + 矢量归一化 (S

优化状态

优化完成

怎样建立好的近红外定量分析模型？

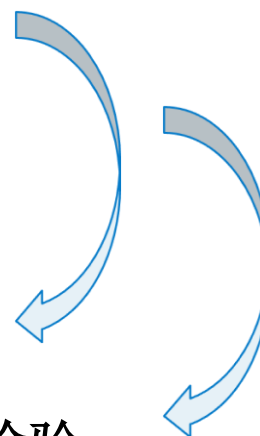


1. 选择或配备代表性建模样品集。
2. 模型优化。
3. 模型评价（内部交叉验证和外部验证）。

评价模型的质量(检验):



1. 建模集 又称 校正集 或 训练集 (Calibration Set)
 2. 检验集 (Validation Set)
- 交叉检验: 对建模和检验使用相同的样品系列
 - 检验集检验: 使用两个样品系列, 分别对应建模和检验



决定系数

$$R^2 = \left(1 - \frac{\sum (Differ_i)^2}{\sum (y_i - y_m)^2} \right) = \left(1 - \frac{\sum (y_i - y)^2}{\sum (y_i - y_m)^2} \right)$$

RPD (Residual Prediction Deviation) 残留预测偏差 = SD/SEP

均方根误差

$$RMSE = \sqrt{\frac{1}{M} \sum (Differ_i)^2}$$

同一组样品、同一组分， R^2 越大，RPD越大，RMSE越小。

决定于模型优化的条件、实验室化学分析水平和近红外仪器的性能。

RMSE最重要，决定预测样品的误差大小。

① 检验集检验(外部检验)

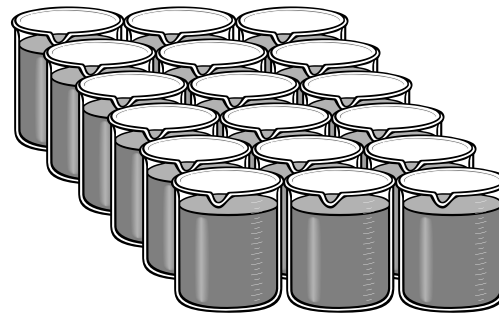
测量两个独立的样品系列，且覆盖整个系统的含量范围



校正集样品

建立基本模型

RMSEE



检验集样品

验证基本模型

RMSEV

Differ₁

Differ₂

Differ₃

.....

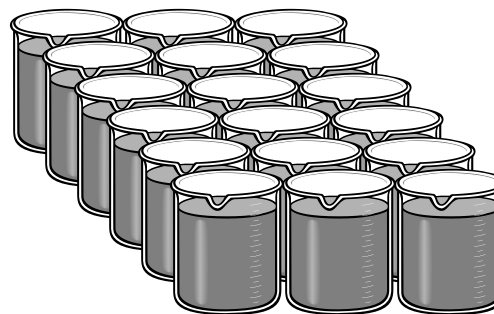
Differ_n

问题： 只有一部分测试样品用于基本模型的建立



校正集样品

建立基本模型



检验集样品

验证基本模型

适于处理大量样品，计算速度快

② 交叉检验

只有一个样品系列用于建模和验证

按照一定取出样品数，逐次分析所有样品，完成检验



校正集样品

建立基本模型

取出



检验集样品

验证基本模型



校正集样品
建立基本模型

RMSEE

分析



检验集样品

RMSECV



第二步，取出另一样品作为
检验集样品，对剩余样品建
立的基本模型进行验证分析

Differ₁

Differ₂

Differ₃

.....

Differ_n

如此重复、循环，直至每一个样品都被检验分析

RMSEE、RMSEC, RMSECV, RMSEP

•RMSEE与RMSECV:

- 同一模型, RMSEE小于RMSECV; 但两者不应存在显著性差异, 否则, 样品代表性不好或模型信息提取不充分。

•RMSECV与RMSEP:

- RMSEP远大于RMSECV, 建模样品的代表性差、模型信息拟和不够或过拟和;
- RMSEP远小于RMSECV, 验证样品代表性差。
- 文章或方法中只给出校正均方差RMSEE, 不能说明任何问题, 要注意。
- 至少要给出交叉检验均方差, 当然最好给出预测均方差。
- RMSECV应该与实验室标准方法的精密度相当。

如何判定近红外定量分析模型的质量



Validation No 2

检验
 校正

不能用于校正数据来优化方法!

拟合值 / 真值 酒精度 维数: 4

R²: 99.99
RMSEE: 0.173

拟合值 vs 真值 / 酒精度 [v/v] / 校正

True Value [v/v]	Predicted Value [v/v]
35	35
40	40
45	45
50	50
55	55
60	60
65	65
70	70
75	75
80	80
85	85
90	90
95	95

校正

Validation No 2

检验
 校正

预测值 / 真值 酒精度 维数: 4 推荐 4

R²: 99.99
RMSECV: 0.18

预测值 vs 真值 / 酒精度 [v/v] / 交叉检验

True Value [v/v]	Predicted Value [v/v]
35	35
40	40
45	45
50	50
55	55
60	60
65	65
70	70
75	75
80	80
85	85
90	90
95	95

交叉检验

Validation No 3

检验
 校正

预测值 / 真值 酒精度 维数: 4 推荐 4

R²: 99.99
RMSEP: 0.172

预测值 vs 真值 / 酒精度 [v/v] / 检验集检验

True Value [v/v]	Predicted Value [v/v]
35	35
40	40
45	45
50	50
55	55
60	60
65	65
70	70
75	75
80	80
85	85
90	90
95	95

外部检验

- 窗口
- 打印
- 保存
- 光谱
- 载荷
- 特殊...

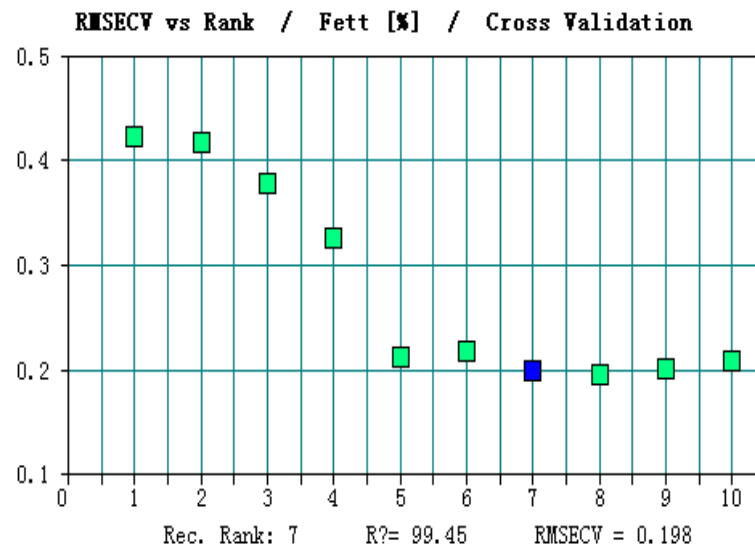
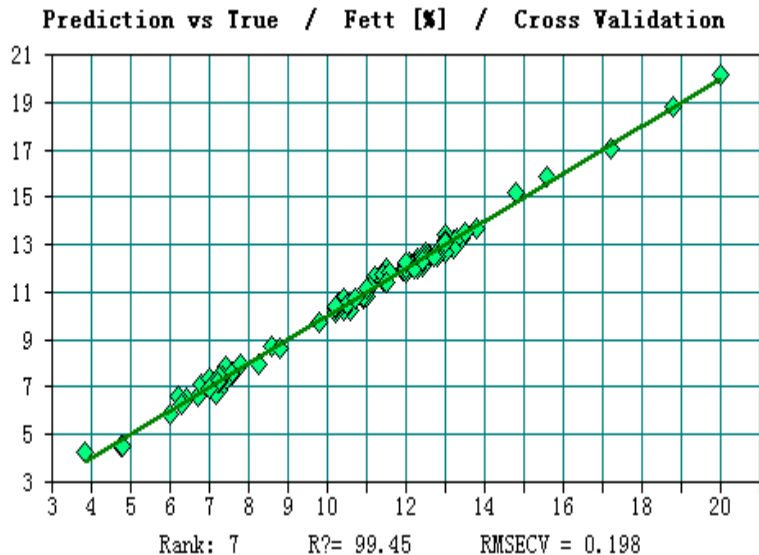
最佳主因子数选择

- 1) **欠拟合**：如果PLS所用的因子数太少，那么光谱中一些有用的信息就没有包含在模型中，那么校正的结果和预测的结果都不会很好。
- 2) **最佳拟合**：应该使校正的结果和预测的结果都很好。
- 3) **过拟合**：如果PLS所用的因子数太多，那么虽然对校正集样品的校正结果很好，但是对没有在校正集中样品的预测结果会很差。

模型内部交叉检验



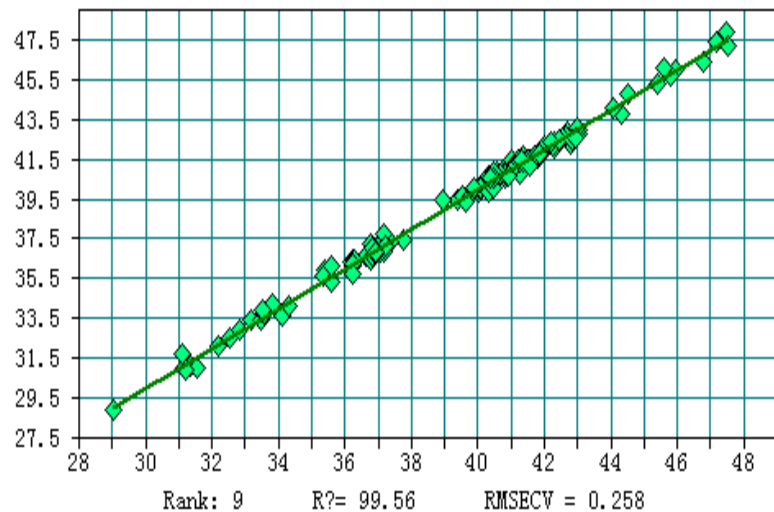
RMSECV vs Rank 图表



结果不好!

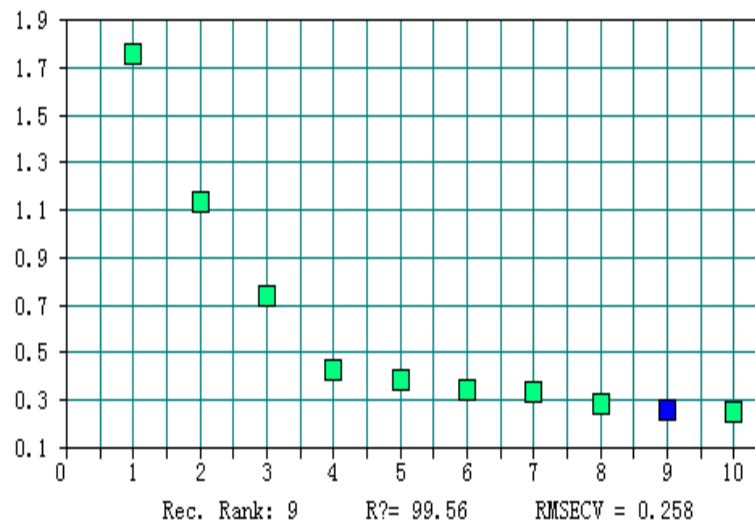
RMSECV vs Rank 图表

Prediction vs True / Trockenmasse [%] / Cross Validation

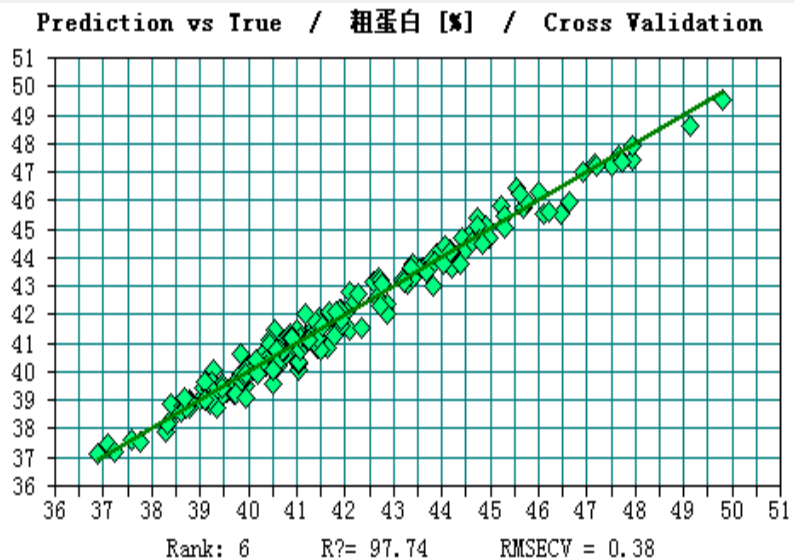


结果一般！

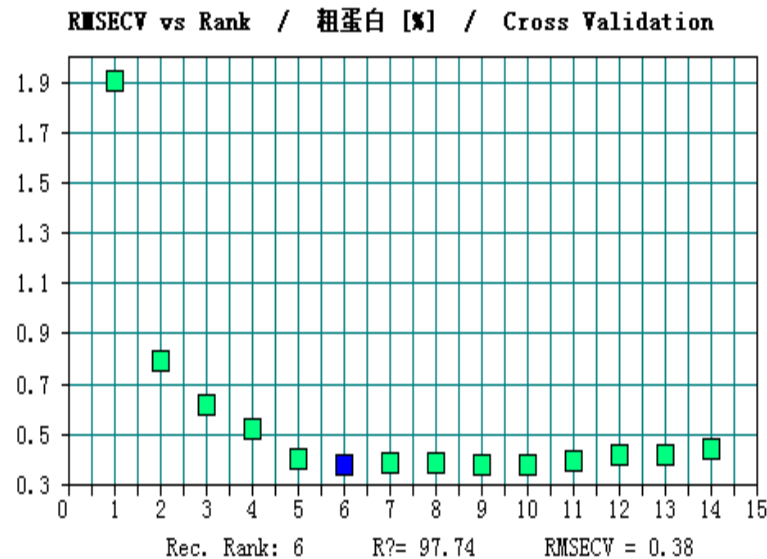
RMSECV vs Rank / Trockenmasse [%] / Cross Validation



RMSECV vs Rank 图表



好的结果!

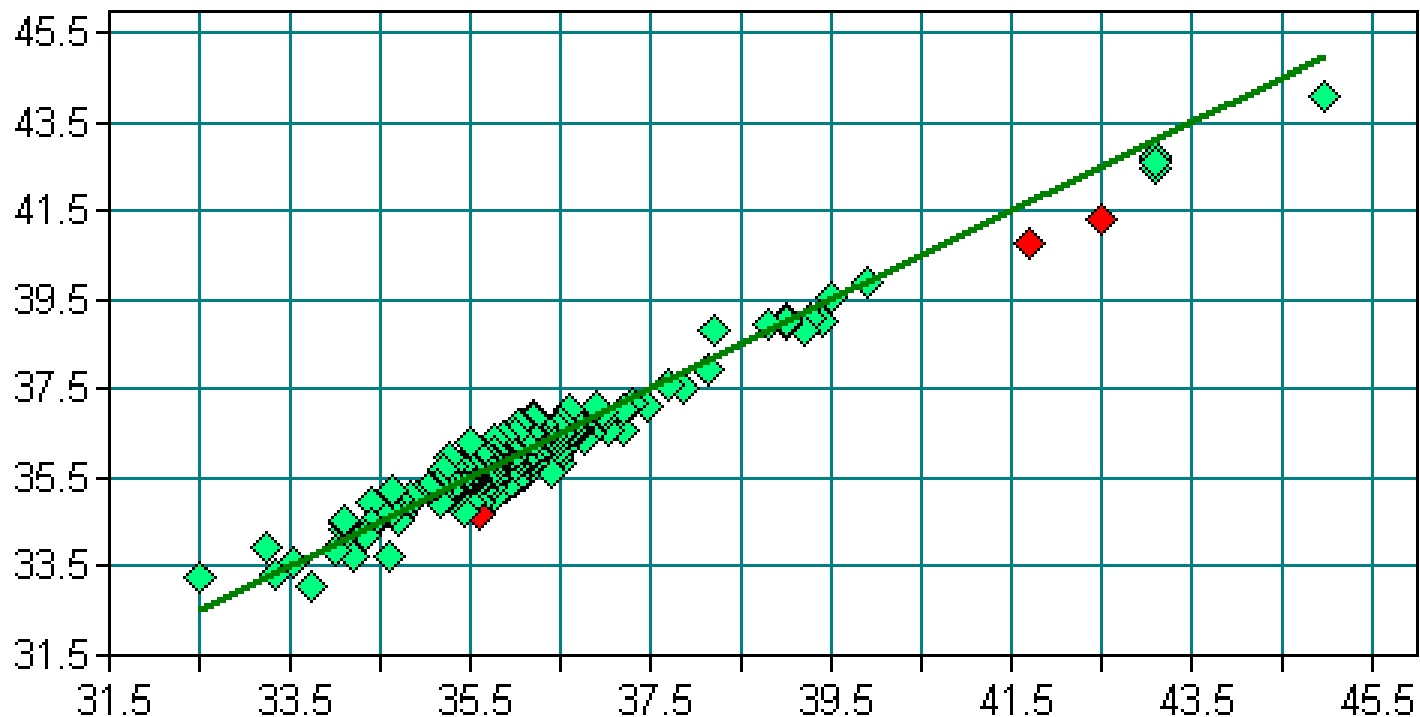


OPUS Quant2 (PLS) Report



Detection of gaps in calibration range by density values

Vorhersage vs Wahr / Moisture [% m/m] / Kreuzvalidierung

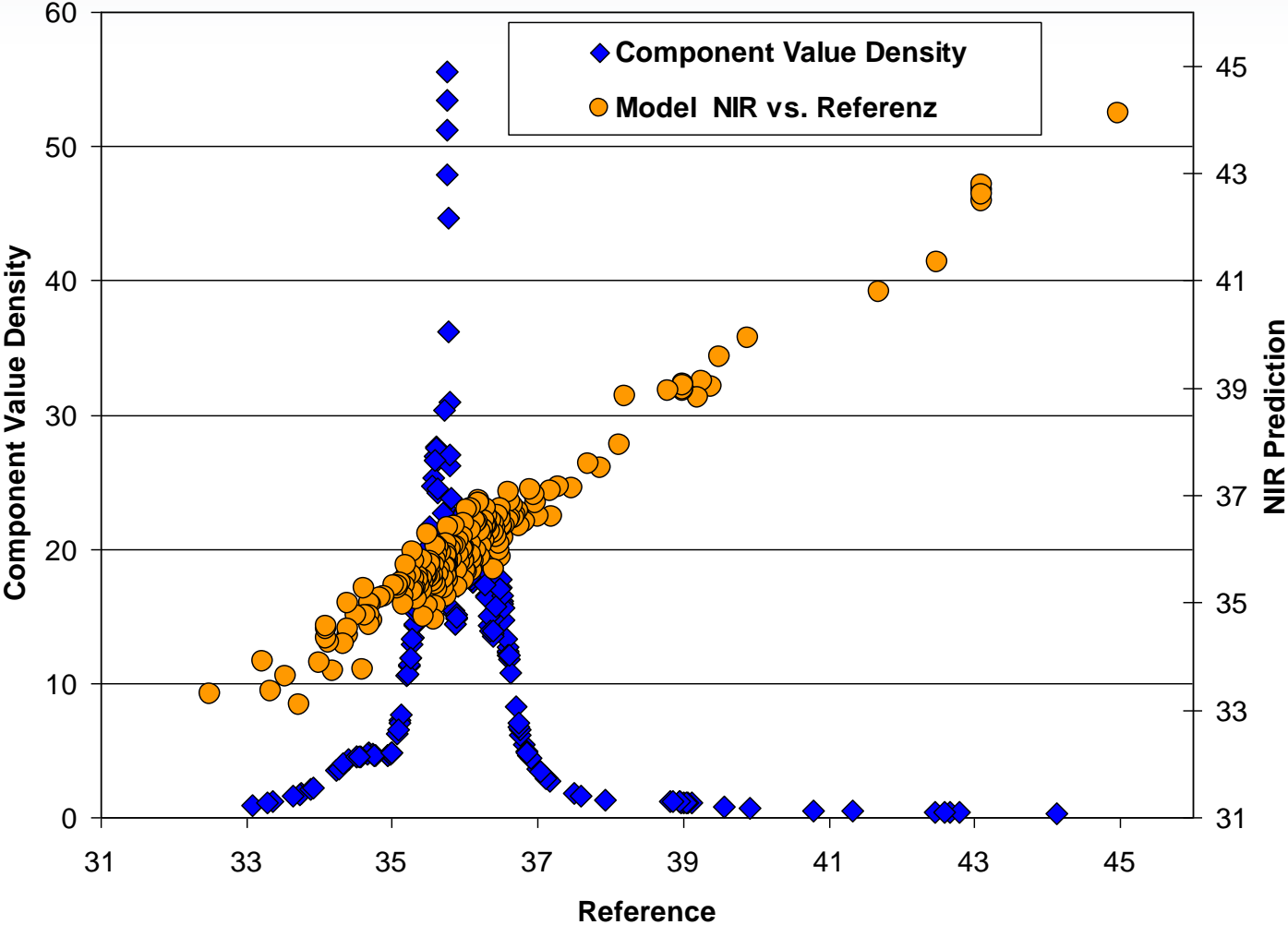


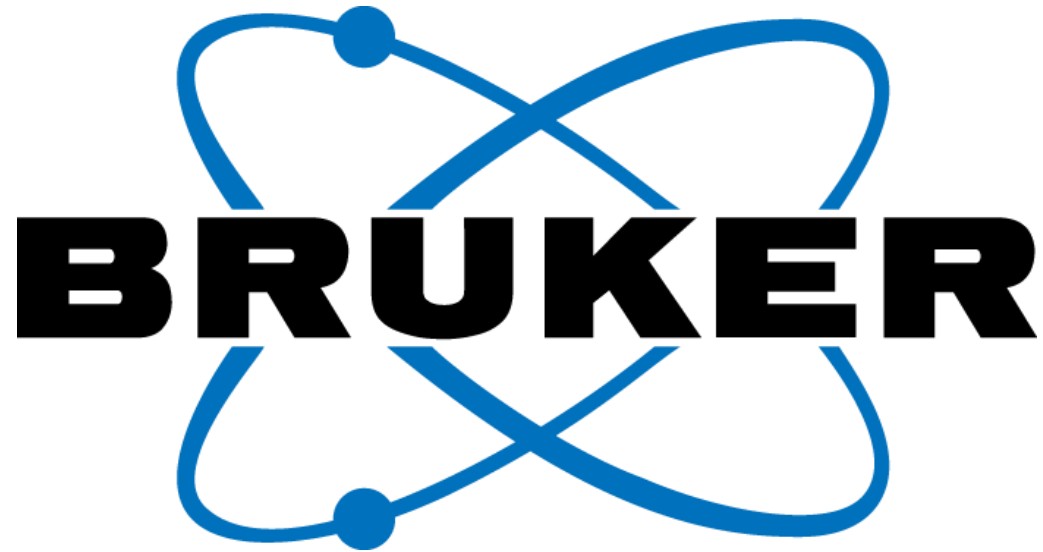
Rang: 6 $R^2 = 95.86$ RMSECV = 0.352 Bias: 0.0156 RPD: 4.92
Validation No 1 Cheese.q2

OPUS Quant2 (PLS) Report



Detection of gaps in calibration range by density values





www.bruker.com