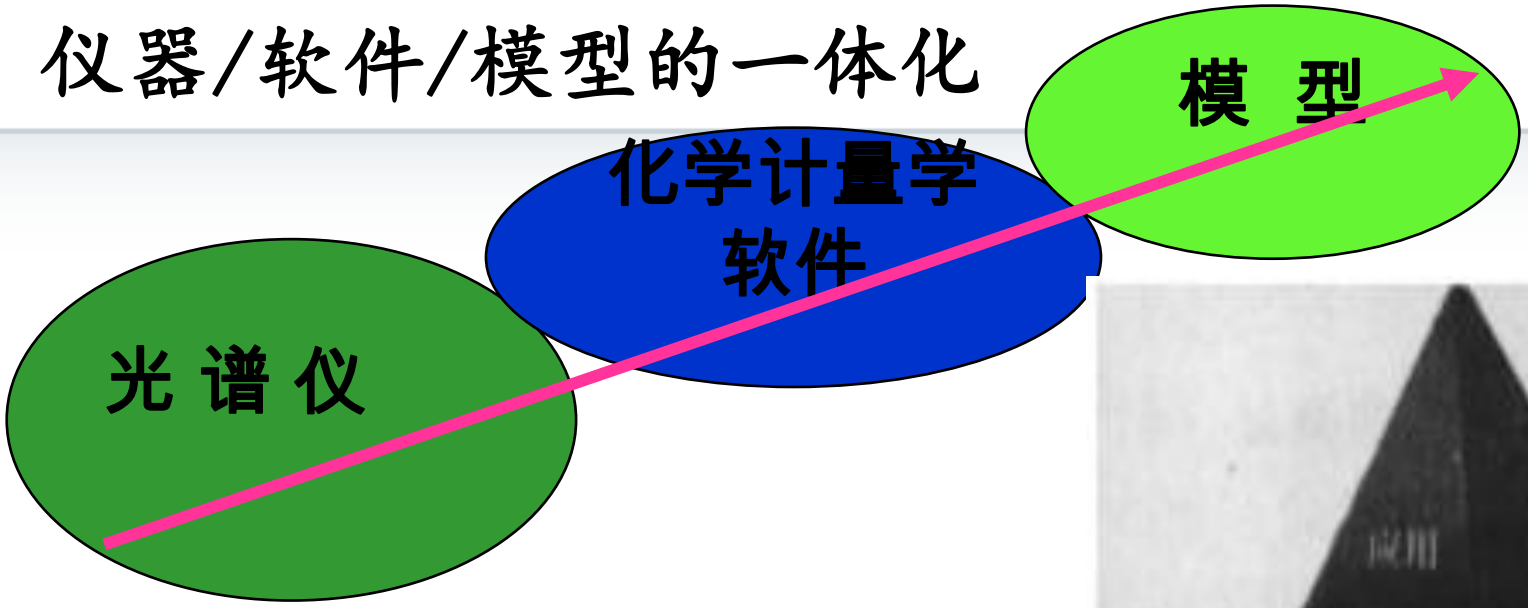
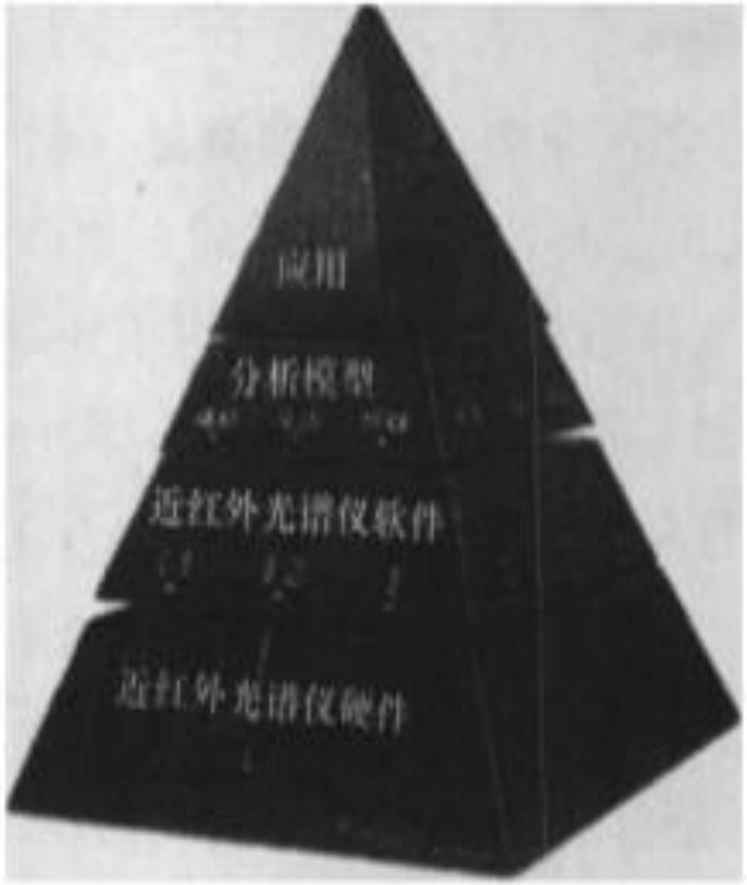
建立定量模型及模型评价

仪器/软件/模型的一体化



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

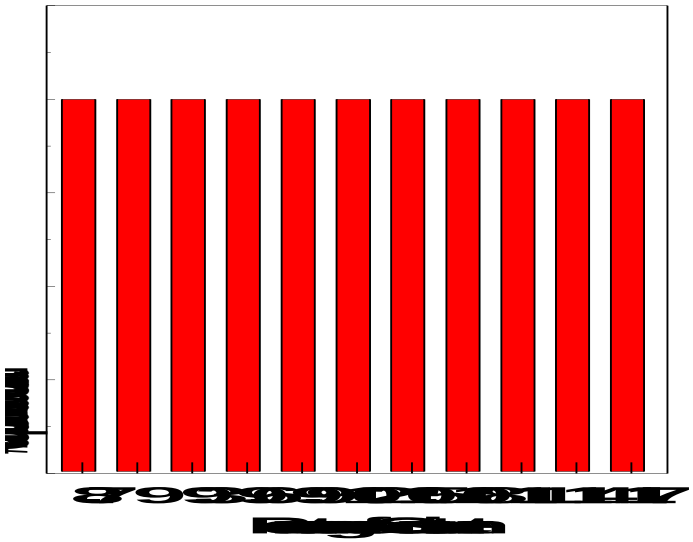
选择或配备代表性建模样品集。

- 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。

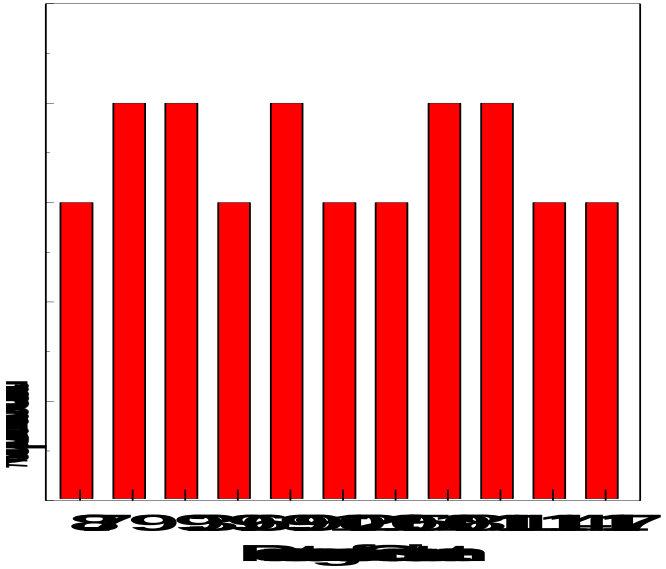
可使用化学计量学方法对样品进行选择、设计。

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。



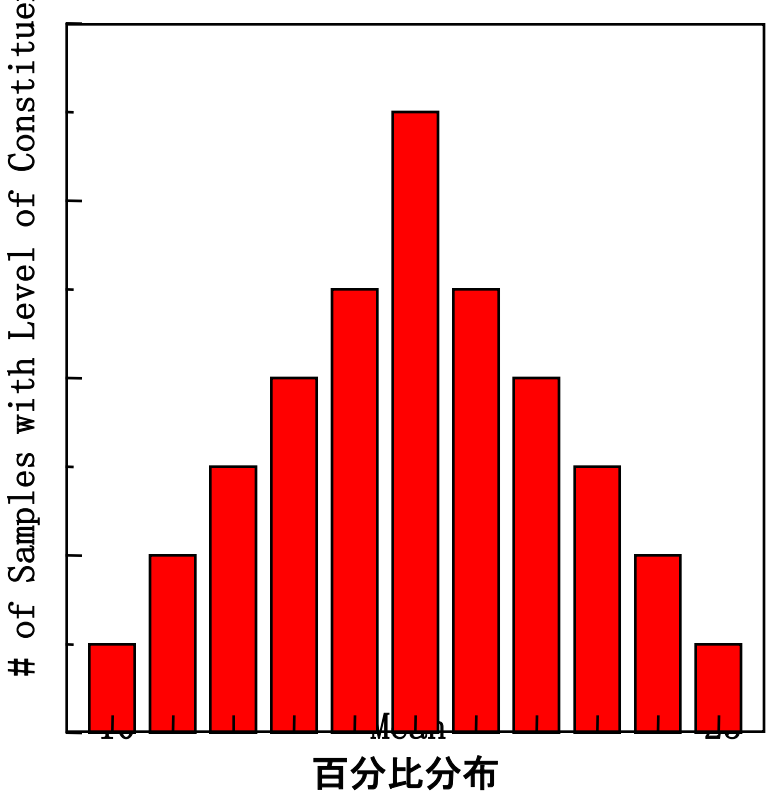
完美的分布
(很少出现的情况)



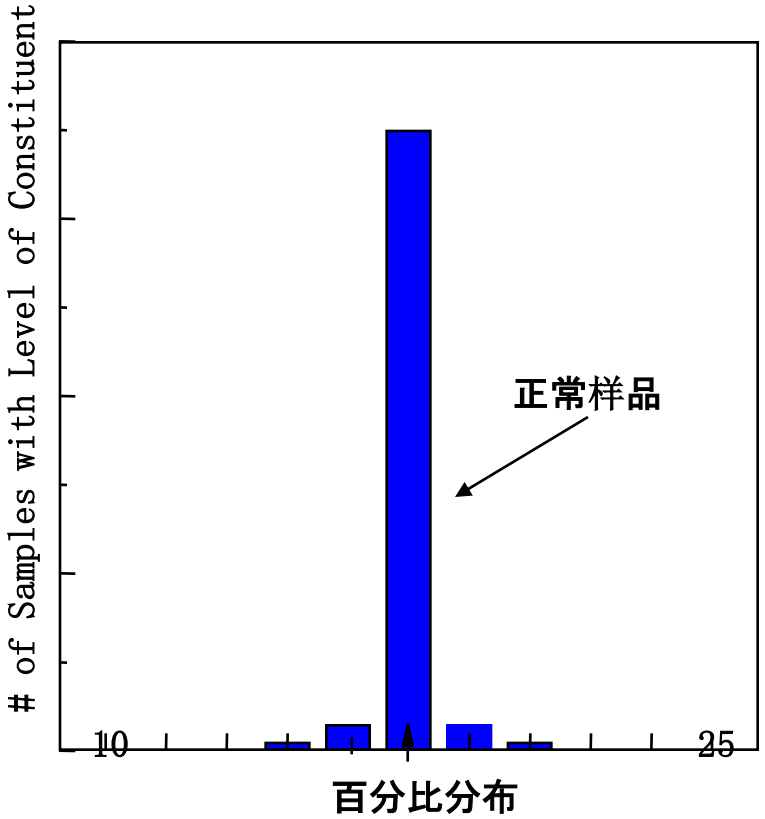
比较合理的分布

样品分布

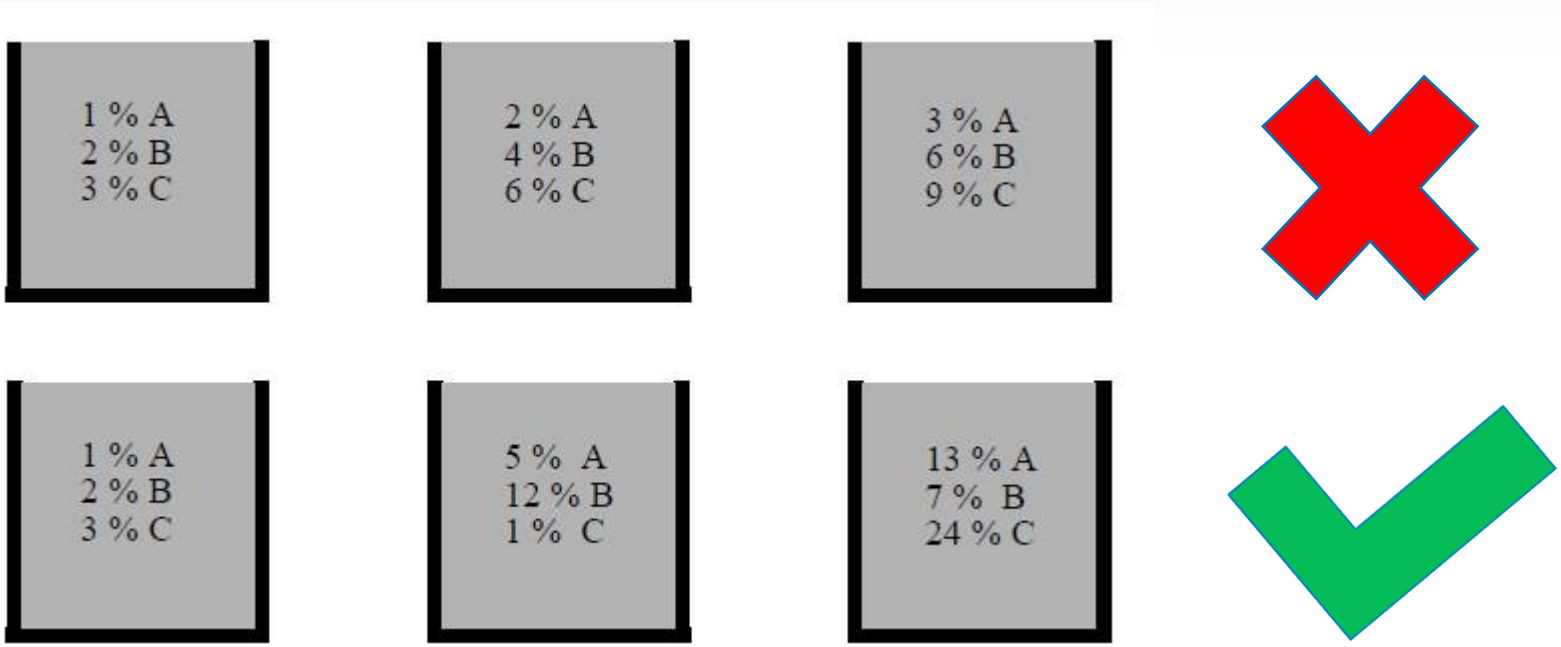
饲料、食品、烟草、化工、炼油等产品的样品分布情况



制药分布

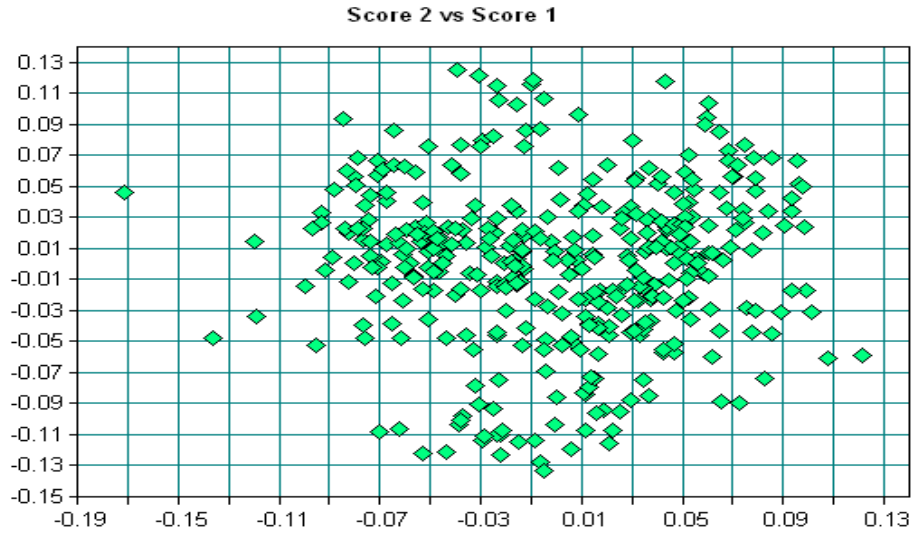


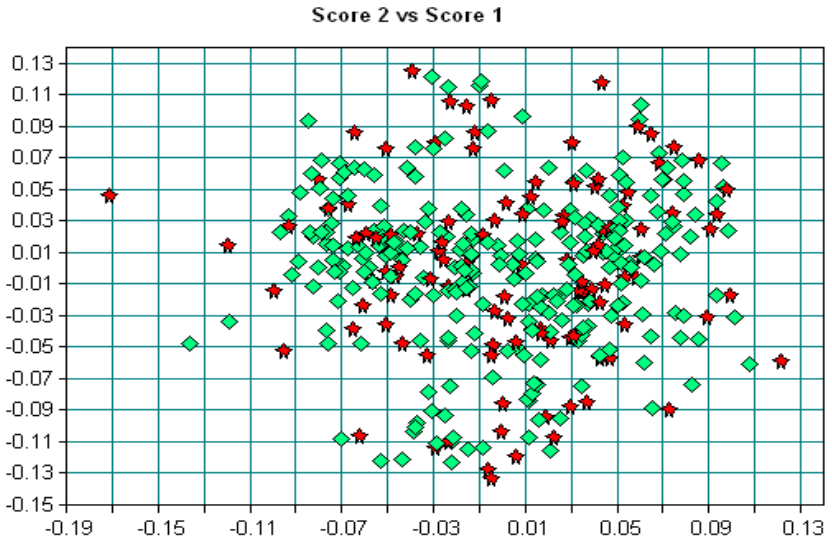
避免样品的共线性

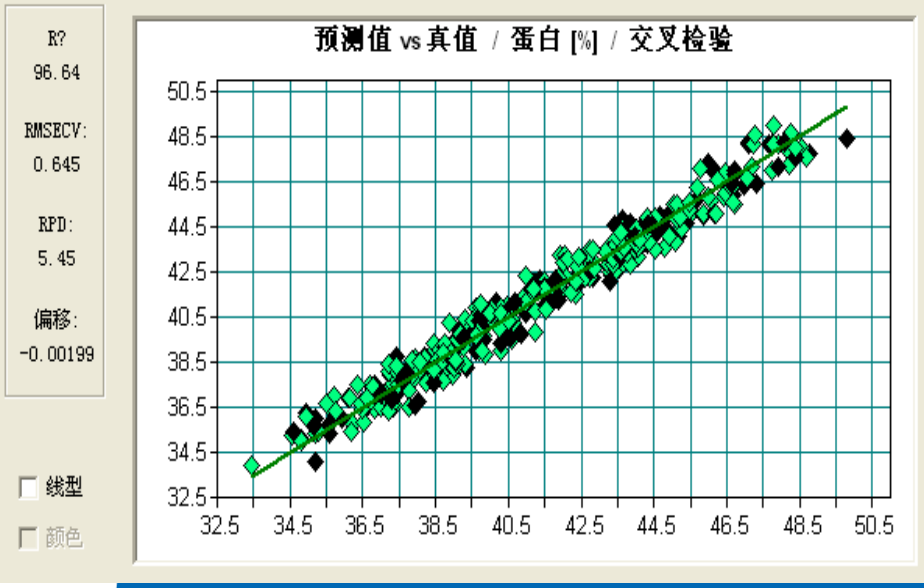


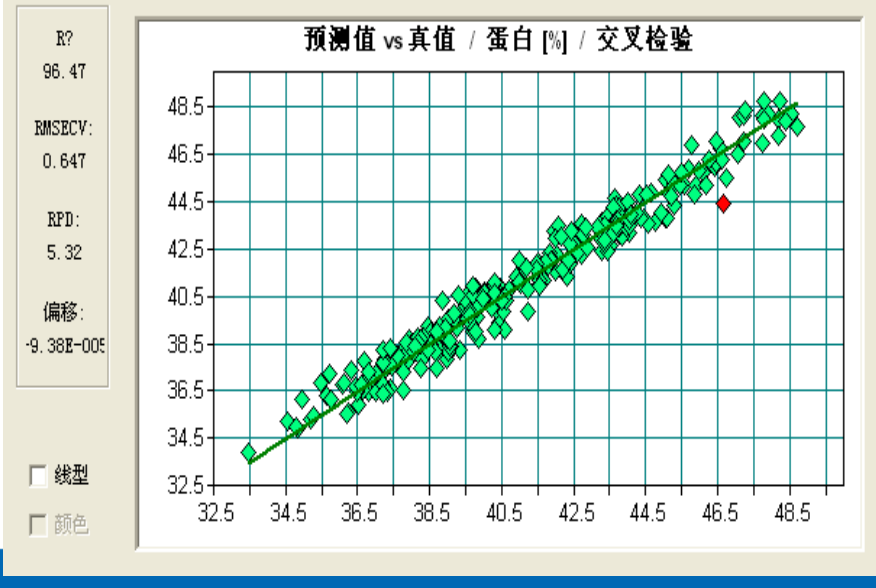
在实验室准备建模样品时，要保证这些样品的浓度不能是线性增加或减少。特别指出：稀释的样品不适合作建模样品。

多余样品的挑选（３７３选２６１）









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：

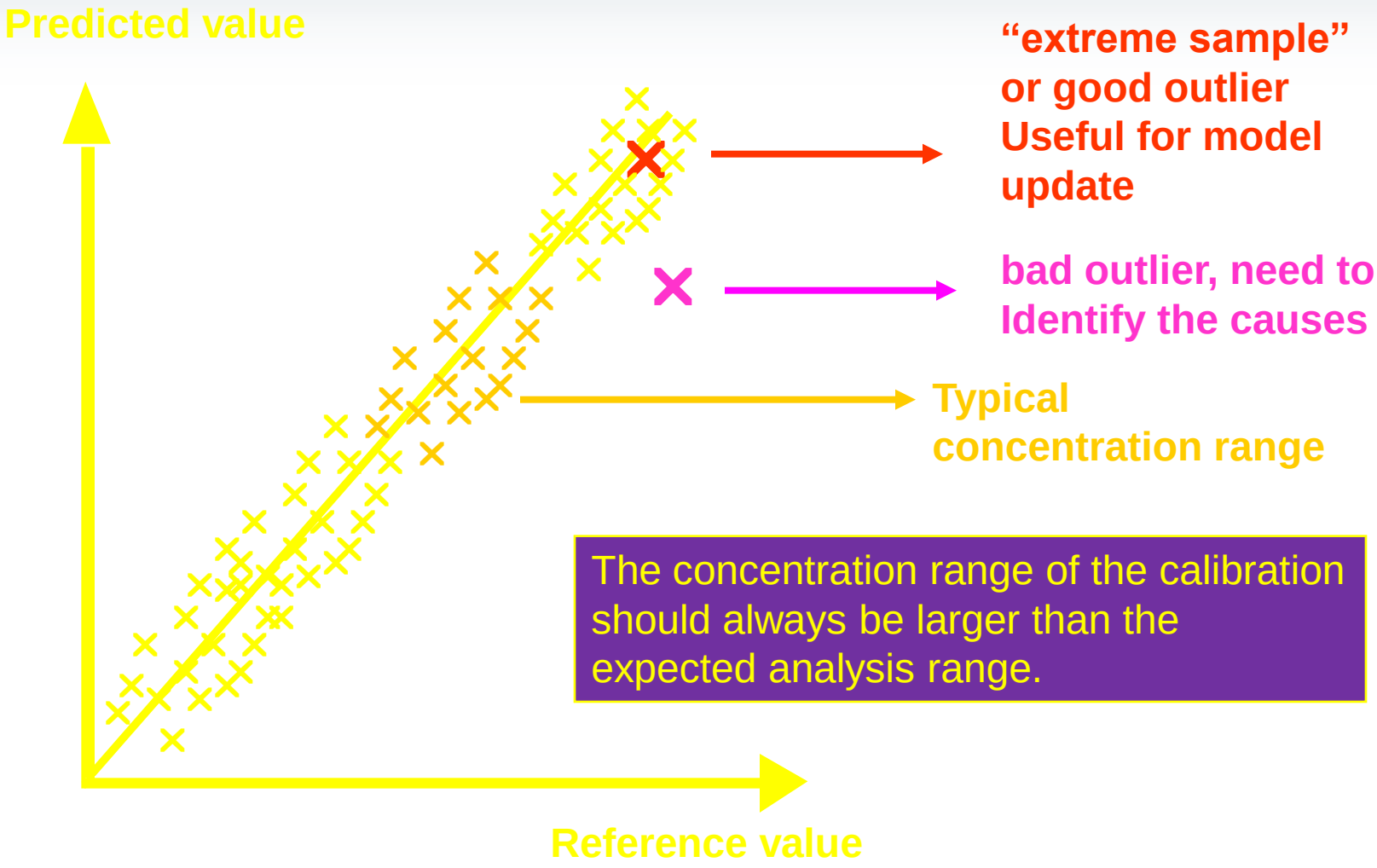
- 组分含量处于两端极值(最高或最低)：样品具有含量的极值信息，不应盲目删除
- 干扰组分含量处于两端极值(最高或最低)：事先未分析干扰组分
- 样品来源不同：分类、分段建模

“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

异常样品的判断

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。

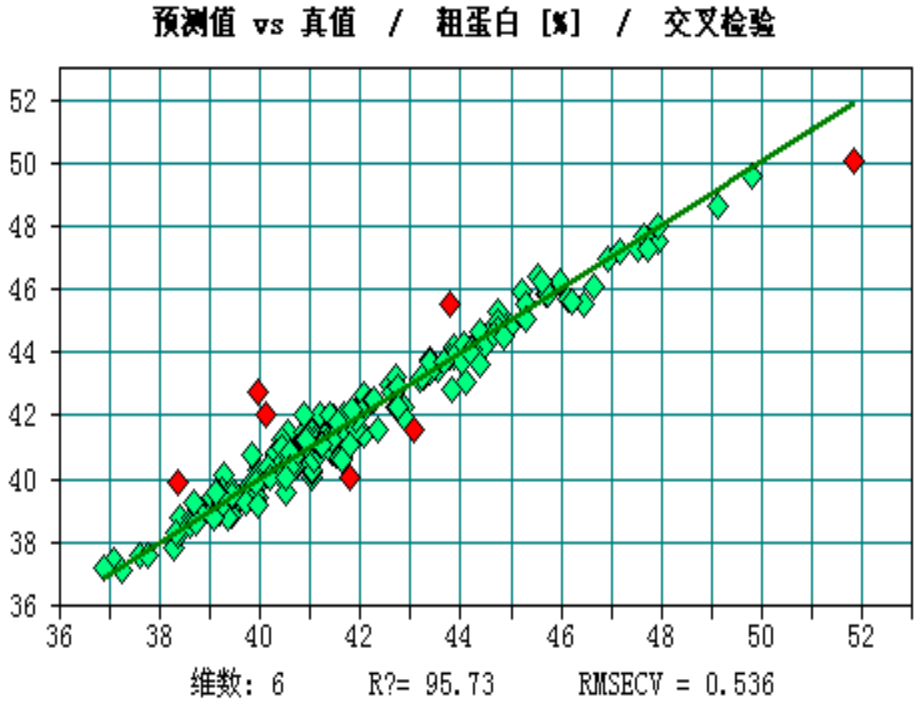
光谱残留的平方同整个其他平均值比较，*F*值的计算公式如下：

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

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



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

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

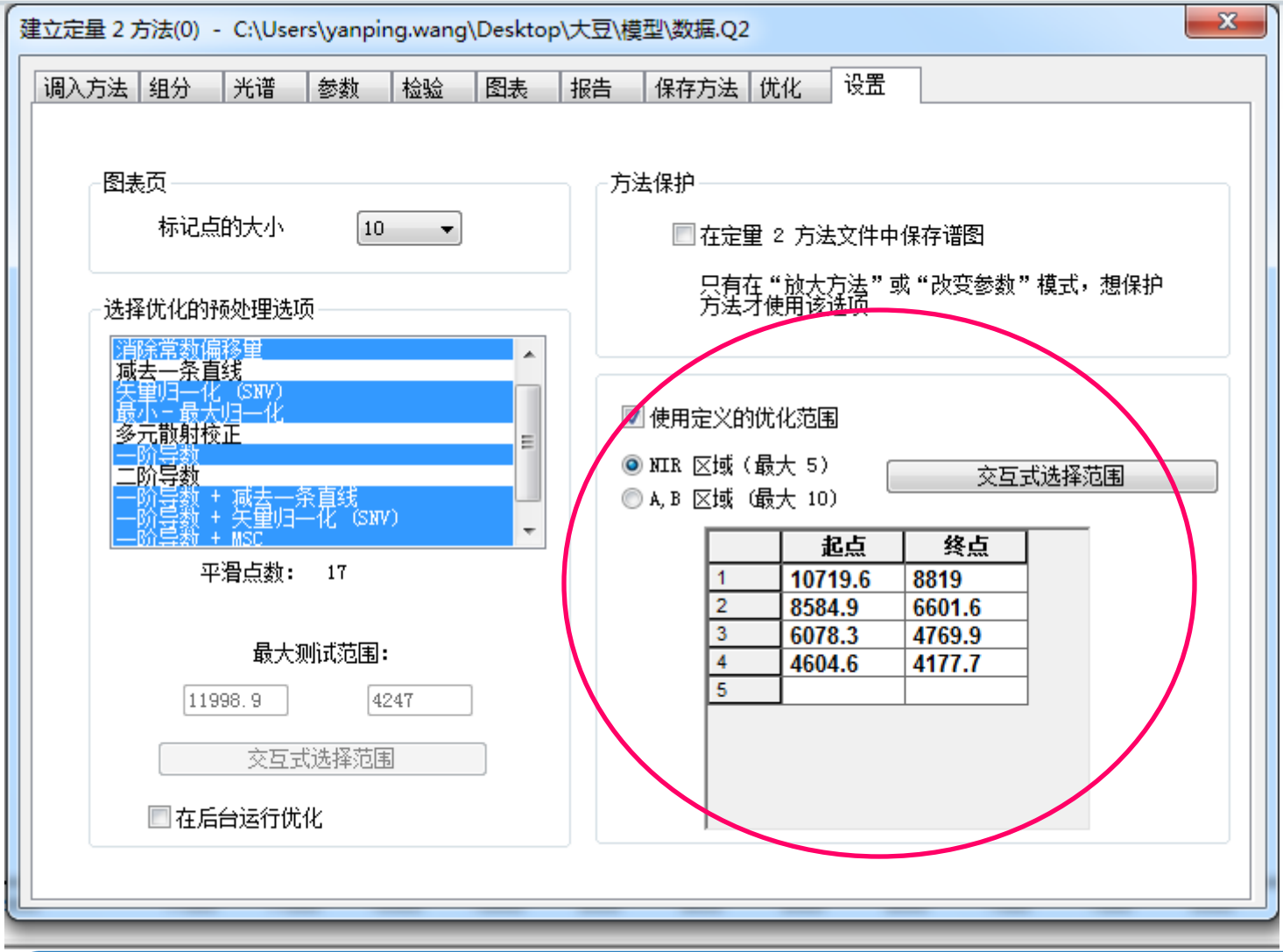
选择减小光谱漂移、增加信息量、稳定可靠的光谱预处理方法。

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。

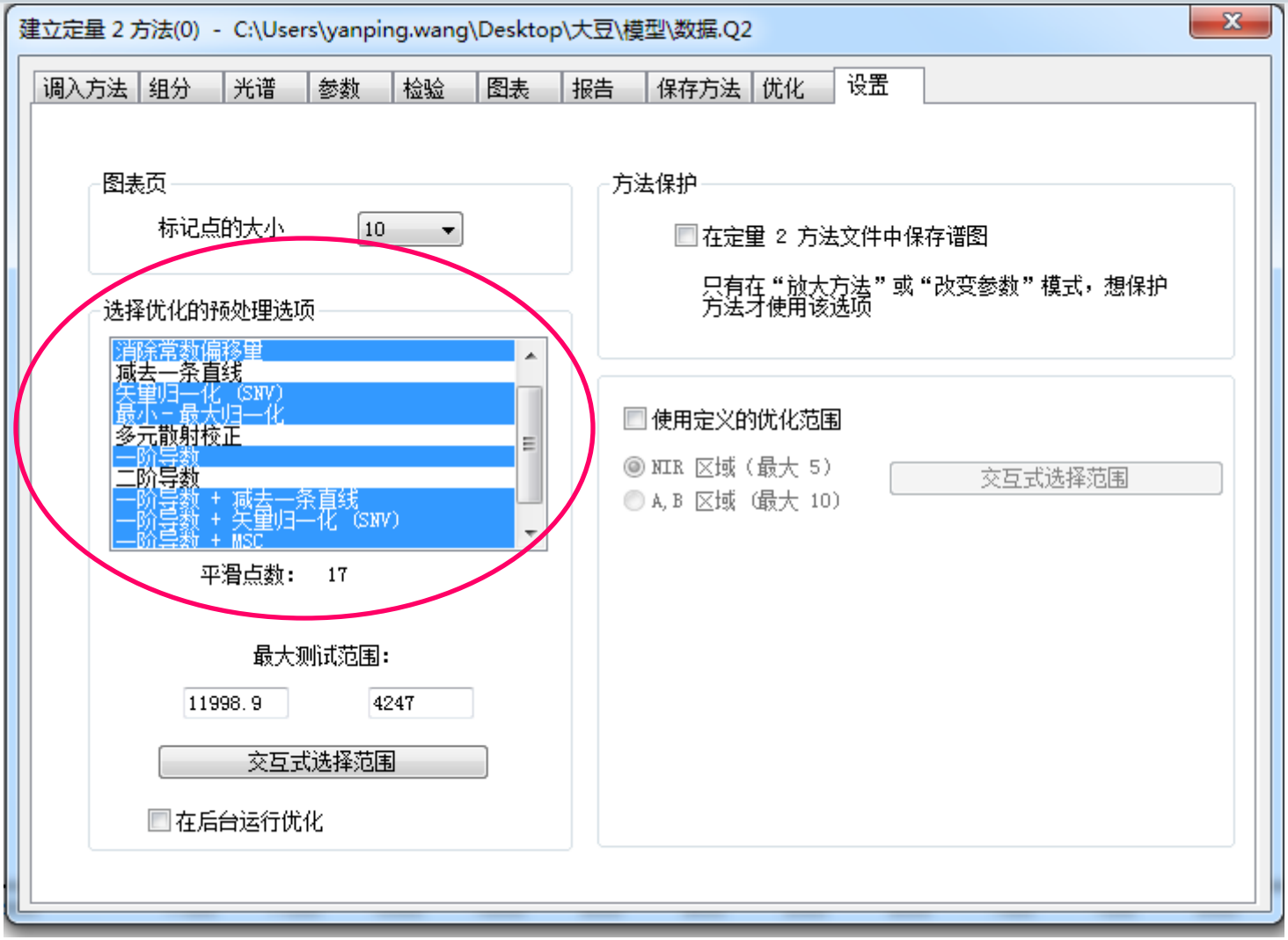
BRUKER

*设置*页：对谱区进行选择



BRUKER

*设置*页：多种数据预处理方法可供选择



BRUKER

*优化*页：帮助完成自动优化检验

建立模型 ➤ ➤ ➤ ➤ ➤ ➤优化模型

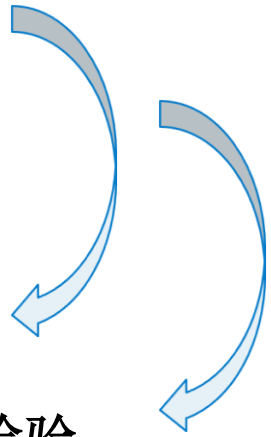
*优化*页：将所选的优化组合信息传递到*参数*页

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

评价指标

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

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

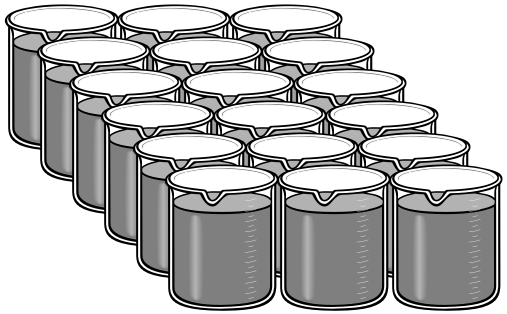
同一组样品、同一组分，R^2越大， RPD越大， RMSE越小。

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。

❶ 检验集检验(外部检验)

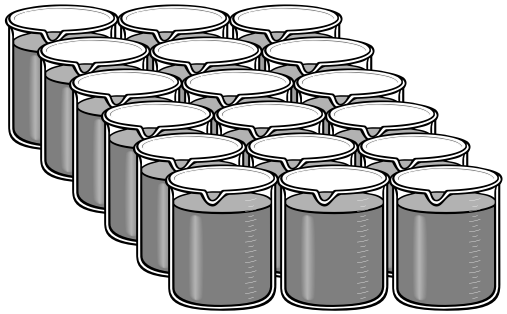
测量两个独立的样品系列，且覆盖整个系统的含量范围



校正集样品

建立基本模型

RMSEE



检验集样品

验证基本模型

RMSEV

$Differ_1$

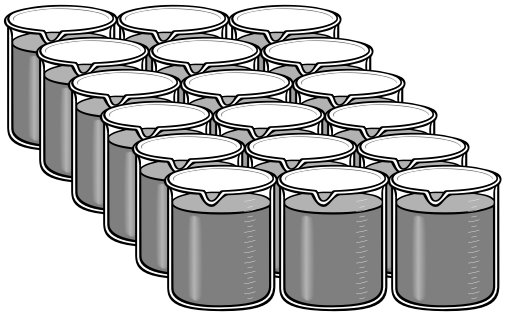
$Differ_2$

$Differ_3$

……

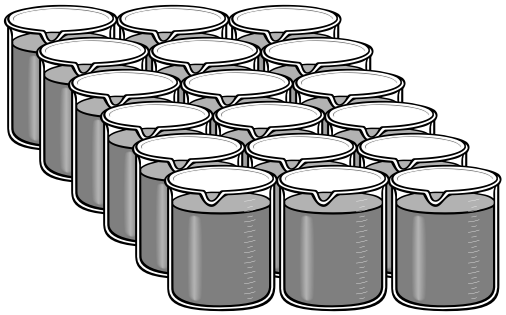
$Differ_n$

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

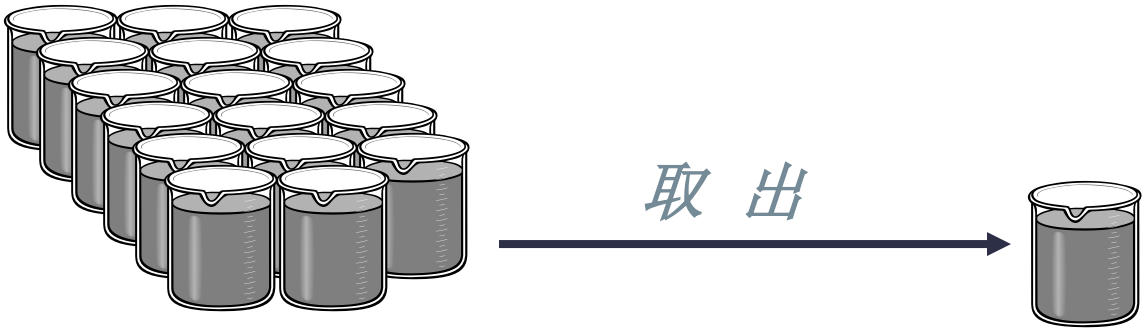
验证基本模型

适于处理大量样品，计算速度快

❷ 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

检验集样品

验证基本模型

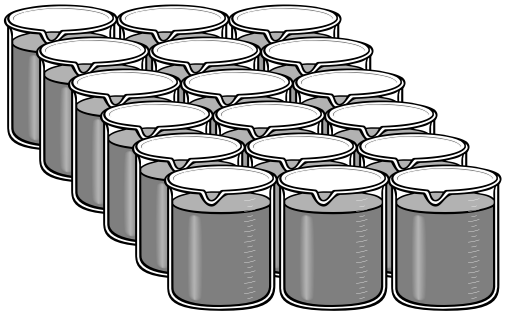
分析

校正集样品
建立基本模型

RMSEE

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

$Differ_1$

$Differ_2$

$Differ_3$

……

$Differ_n$

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；

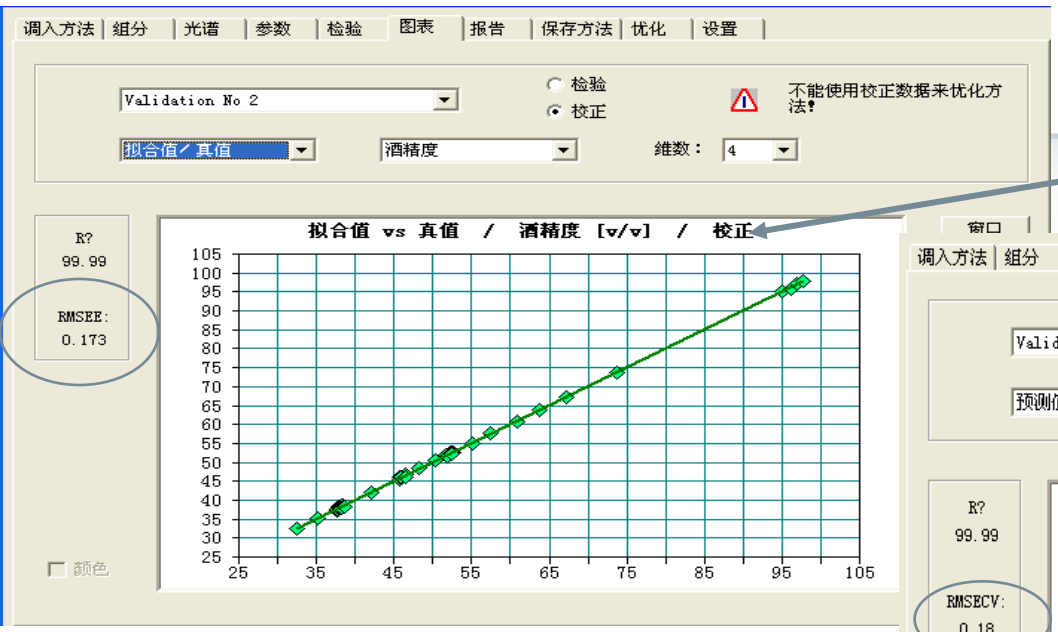
•RMSEP远小于RMSECV，验证样品代表性差。

•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。

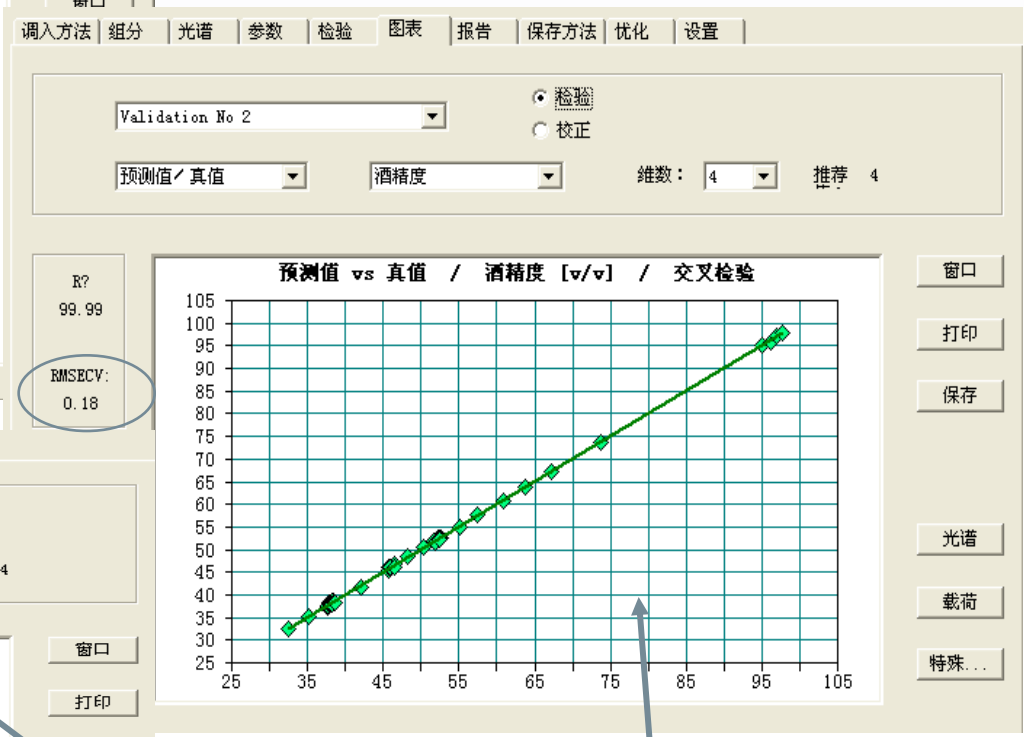
•至少要给出交叉检验均方差，当然最好给出预测均方差。

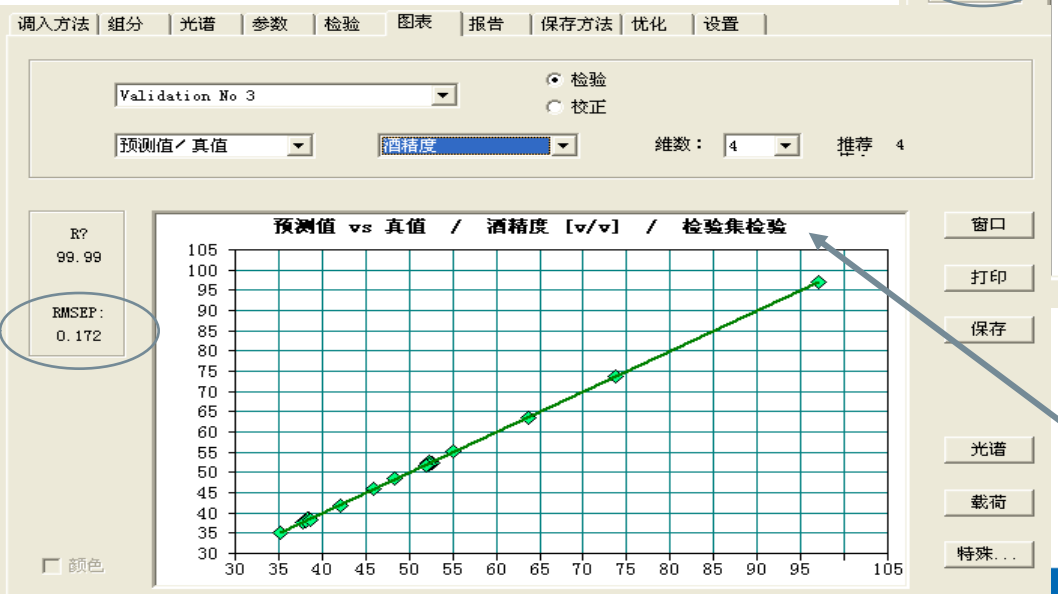
•RMSECV应该与实验室标准方法的精密度相当。

如何判定近红外定量分析模型的质量



校正





交叉检验

外部检验

最佳主因子数选择

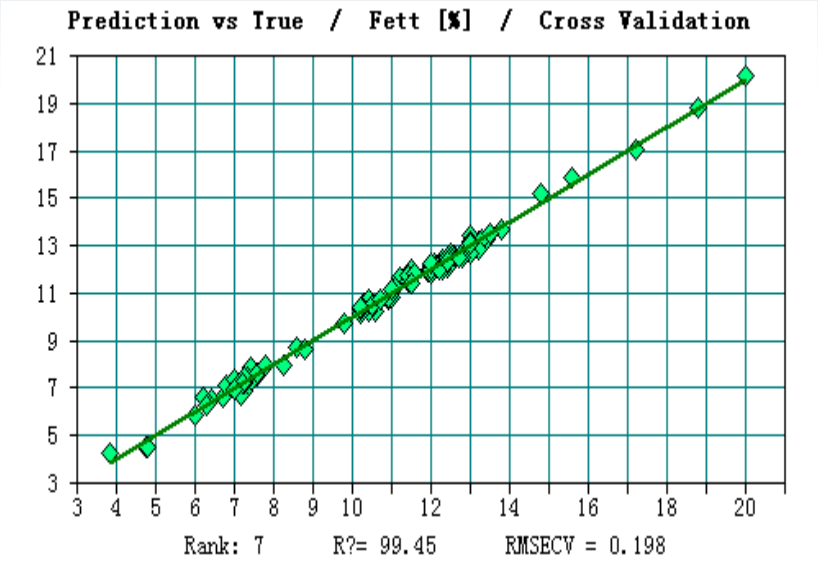
1）欠拟合：如果PLS所用的因子数太少，那么光谱中一些有用的信息就没有包含在模型中，那么校正的结果和预测的结果都不会很好。

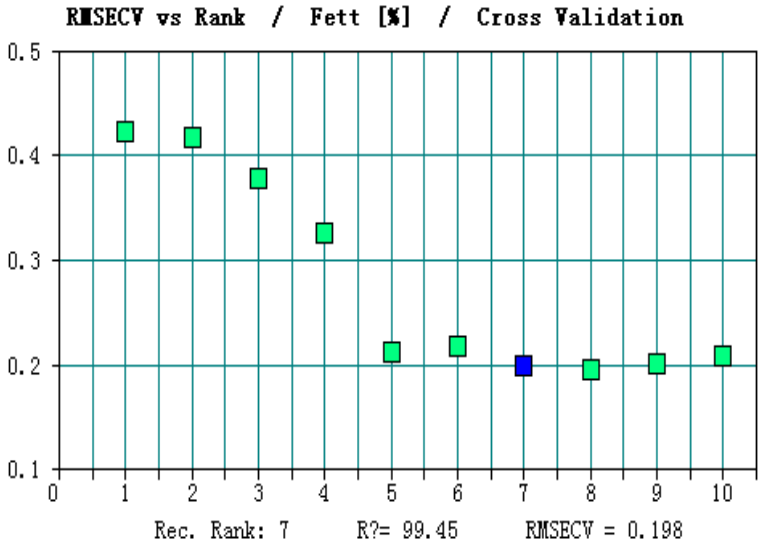
2）最佳拟合：应该使校正的结果和预测的结果都很好。

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

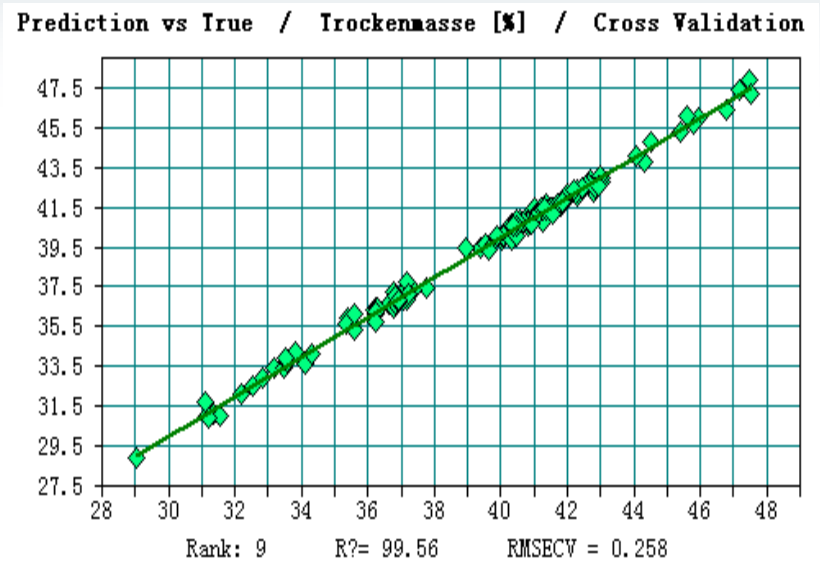


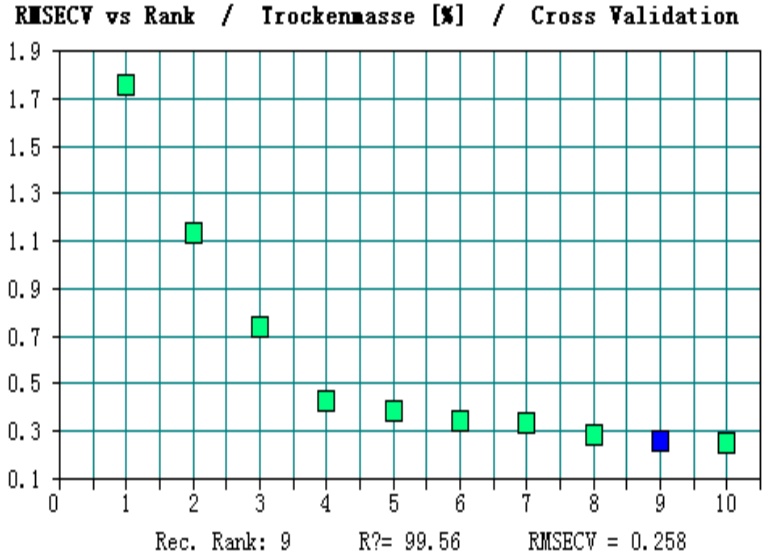


结果不好！

模型内部交叉检验

RMSECV vs Rank图表

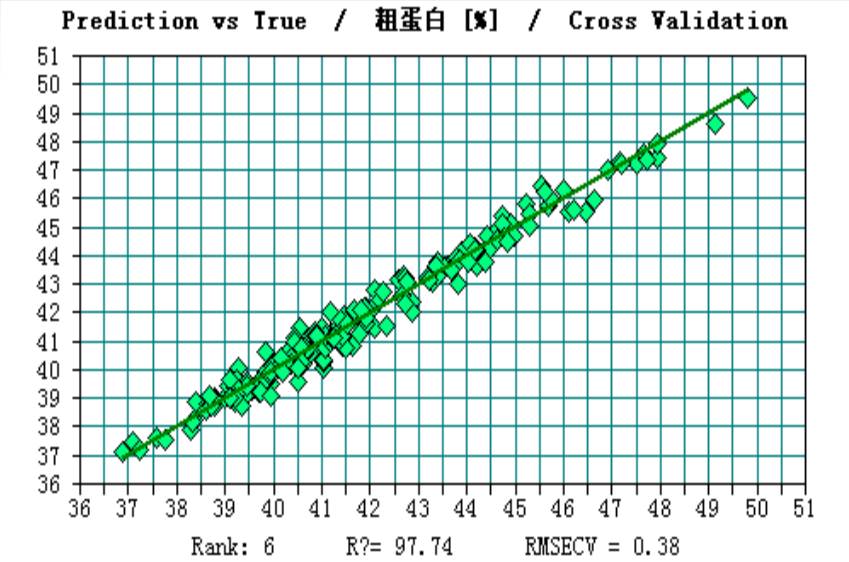




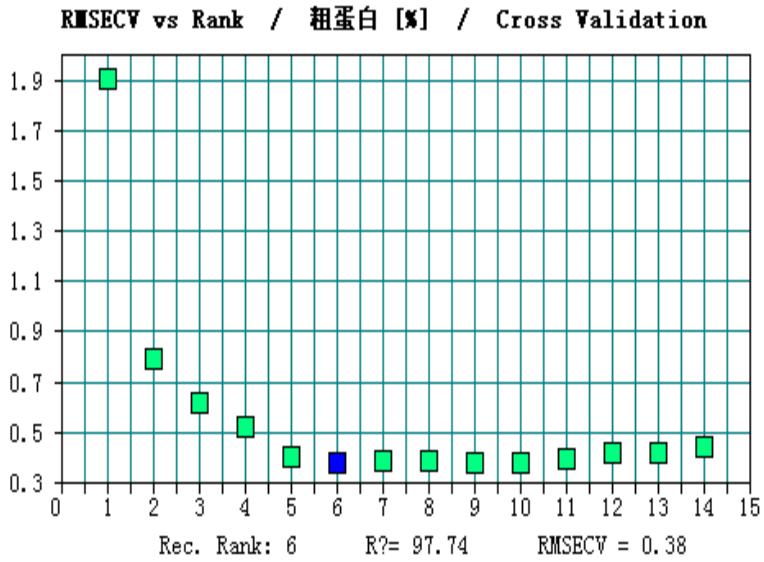
结果一般！

模型内部交叉检验

RMSECV vs Rank图表

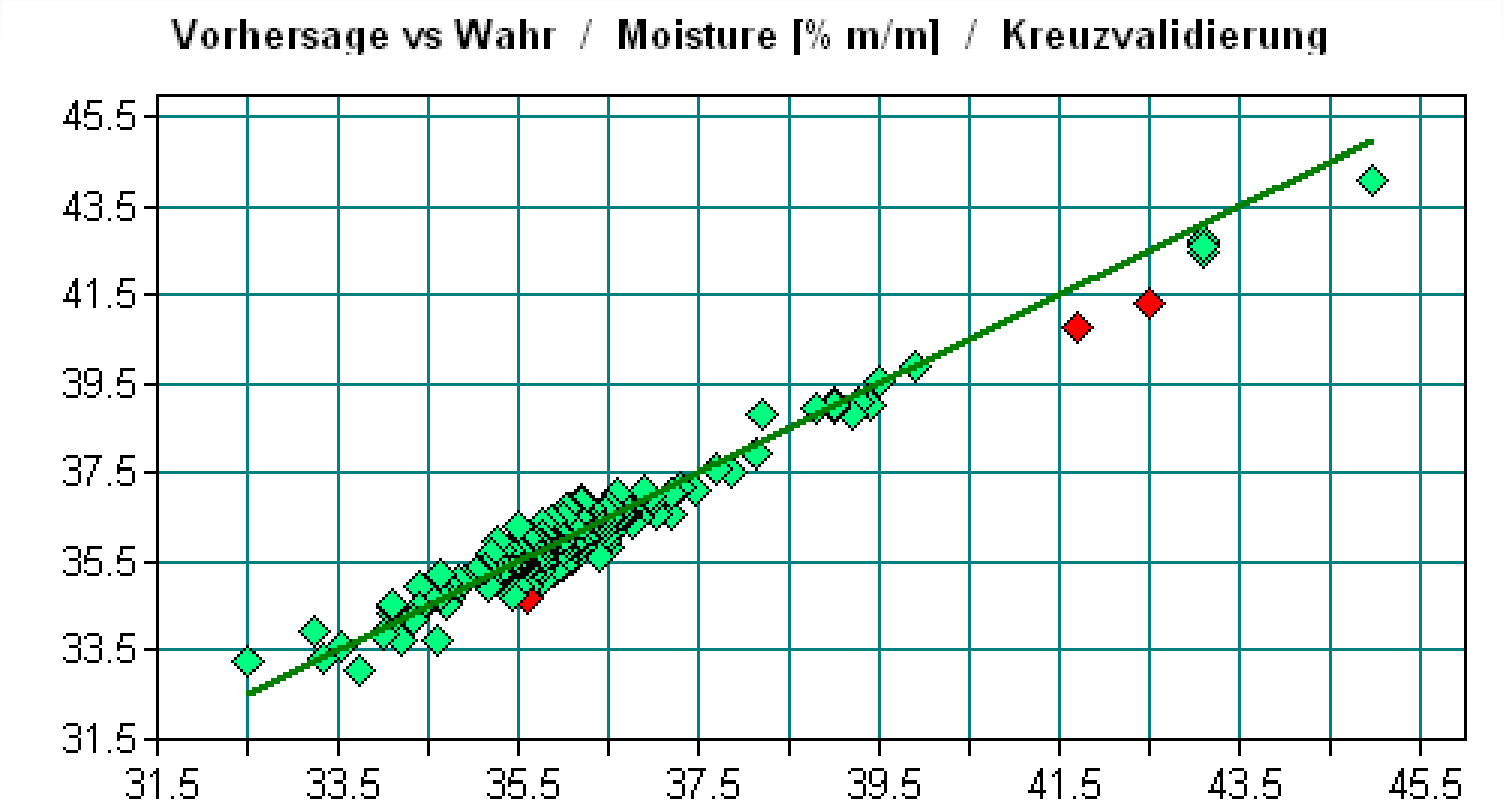


好的结果！



OPUS Quant2 (PLS) Report

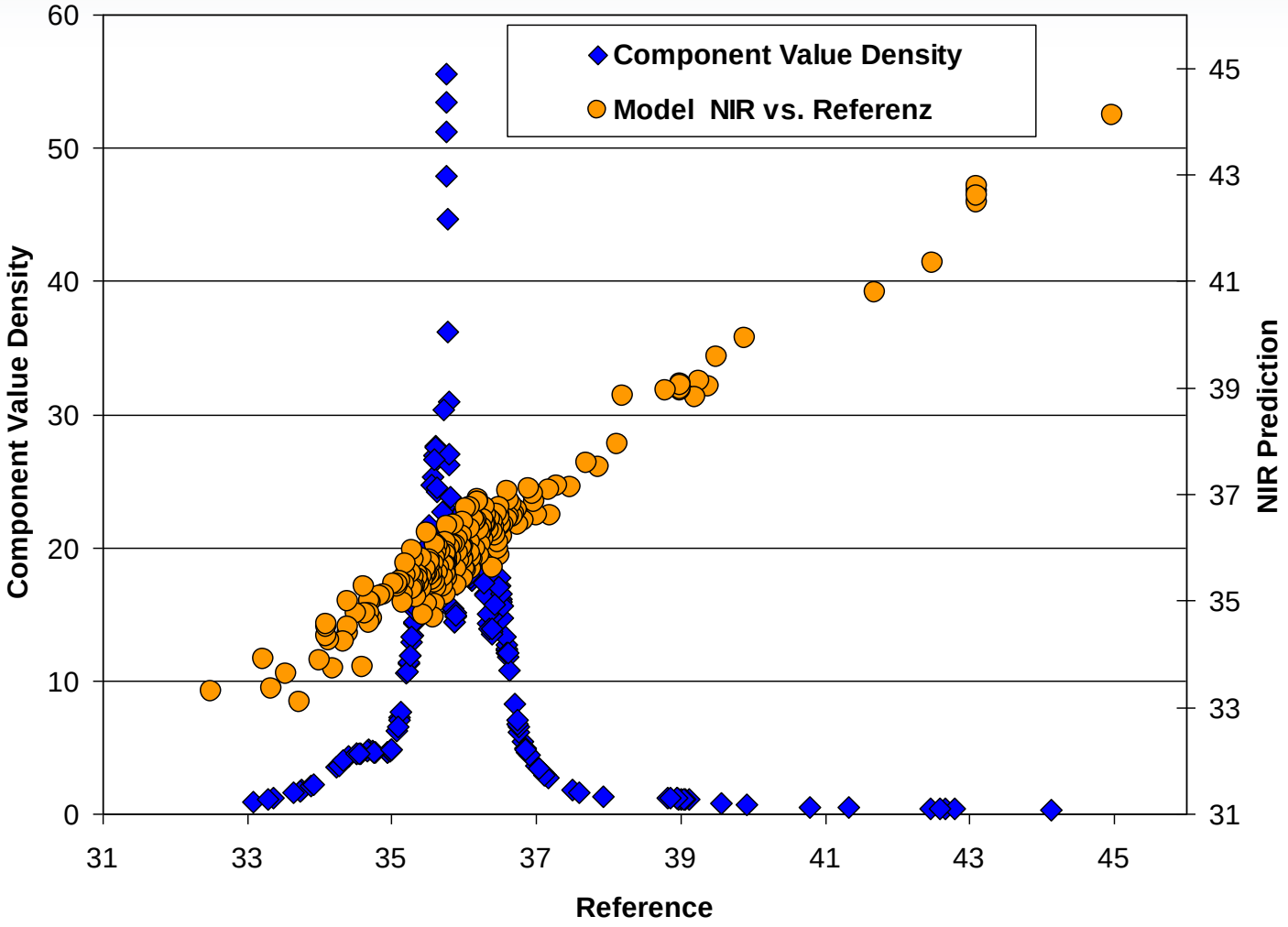
Detection of gaps in calibration range by density values

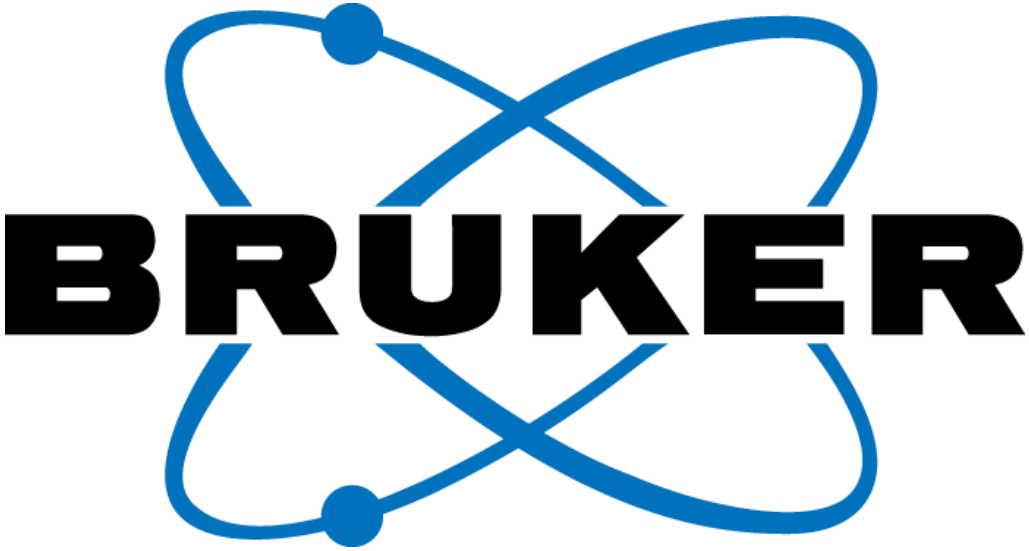


Rang: 6 R^2 = 95.86 RMSECV = 0.352 Bias: 0.0156 RPD: 4.92
Validation No 1 Cheese.q2

OPUS Quant2 (PLS) Report

Detection of gaps in calibration range by density values





www.bruker.com