# Improving Calibration Results for Optical Spectrometers

Brian Rohrback, Infometrix, Inc.

Infometrix

PEFTEC Conference November 2015

# Abstract

A project has been undertaken to assess how to reduce the amount of effort devoted to maintaining and optimizing spectroscopic model performance in support of refinery and chemical plant labs. Over the last five years, a series of algorithmic approaches have been examined with the goal of streamlining the process of chemometric model construction to make the models significantly more robust when put into routine practice. This effort generated the following observations:

1. Even though there are published "Best Practices" for generating chemometric models, these practices are infrequently followed;
2. Recalibration of an optical spectrometer is perceived to be warranted due to changes in crude slates and blending component composition, but may not be required;
3. Even if a calibration was performed properly during initial installation, staffing changes and lack of training undermines subsequent recalibrations; and
4. It is of benefit to minimize software maintenance frequency to control product giveaway.

# Abstract, continued

In order to use optical analyzers effectively, the analyst needs to consider the limitations that ultimately determine the eventual success and life-expectancy of any calibration. There are many areas that have an impact on the ultimate quality of a chemometrics calibration: outlier detection, selection of the number of factors, and even choice of algorithm. This presentation reports on a multi-organization effort leading to an improvement in calibration procedures for on-line and laboratory multiwavelength spectrometers. Here the process of calibration is examined in detail and a path is outlined for building calibrations that are more reliable and less sensitive to process shifts. Much of this improvement can be attained without requiring replacement of either the hardware or software in place. Additional improvement in calibration quality is available through the use of well-referenced methods that constitute the best technologies available.

# 38 Years of Chemometrics

What problems do we see that create the most problems in building a chemometric system?

1. **Poorly characterized standards**
2. **Groupings in the data**
3. **Overfitting the inferential model**
4. **Non-optimal calibration set**
5. Changing protocol
6. Complex samples
7. Poor instrument stability

# Analytical Chemistry Basics - I

When looking at an analyzer technology for a specific application, you need to first look at the sources of variation, typically instrumental and process (from the feedstock and the process parameters themselves).

You would like to try for "Hard Models" as they will be more stable over time, but often, instruments designed to work with Soft Modeling techniques are favored as, in the case with optical spectroscopy, they can often provide their measurements more quickly.

The penalty you pay for Soft Models is that they will require more frequent recalibration.

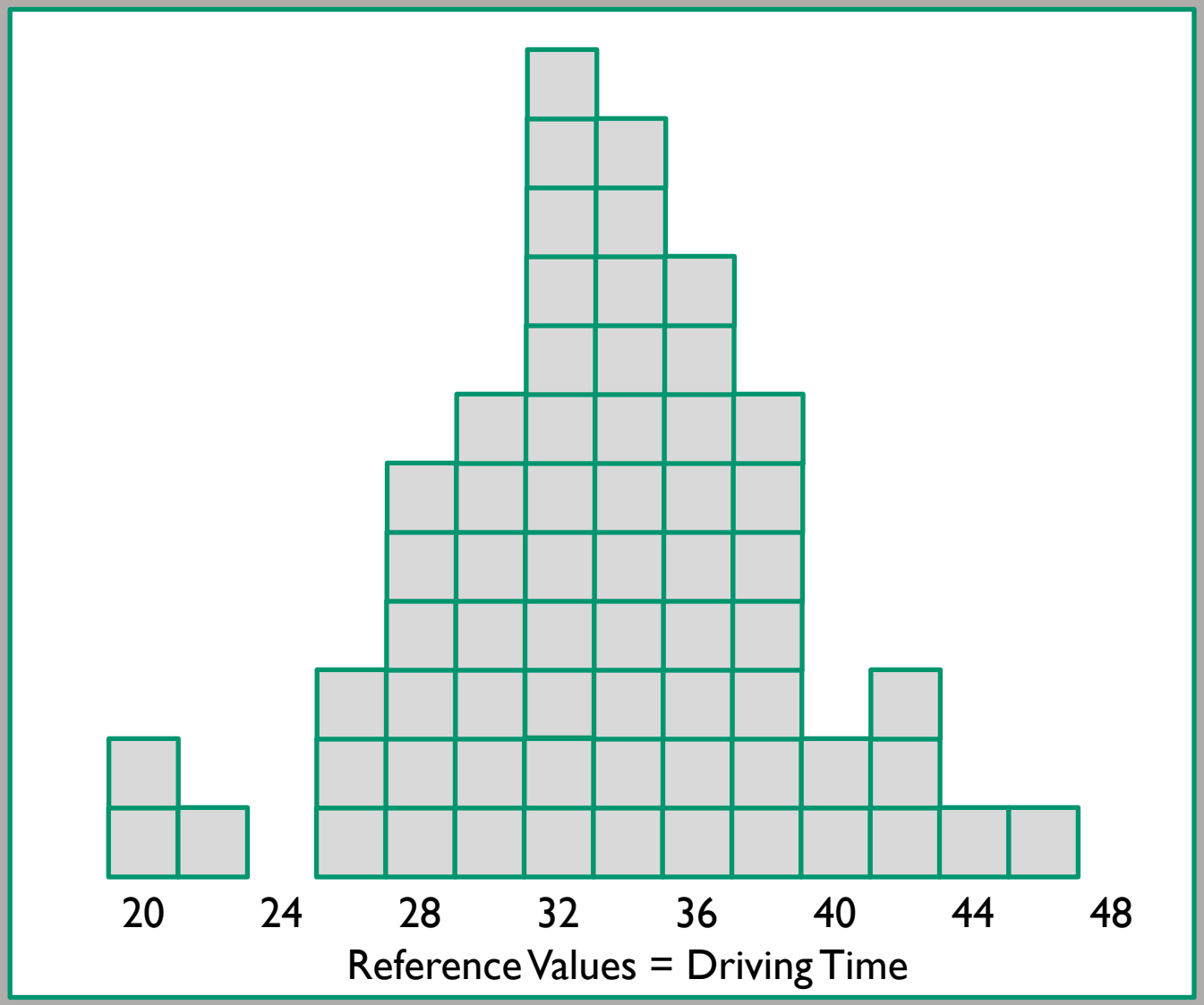
# Analytical Chemistry Basics - 2

- What is a Hard Model?
  - GC for natural gas: separate all of the components, look up the associated heat of combustion for each of the components, multiply and add to get BTU content
  - Hydroxyl value of polyurethane polyol prepolymers
- What is a Soft Model?
  - Application of NIR on natural gas samples representing different compositions: measure the BTU content of the gases as a whole using a reference method; apply PLS to calibrate the NIR response to the reference
  - Inferential
- Inferential models employ chemometrics, but not all chemometric models are inferential

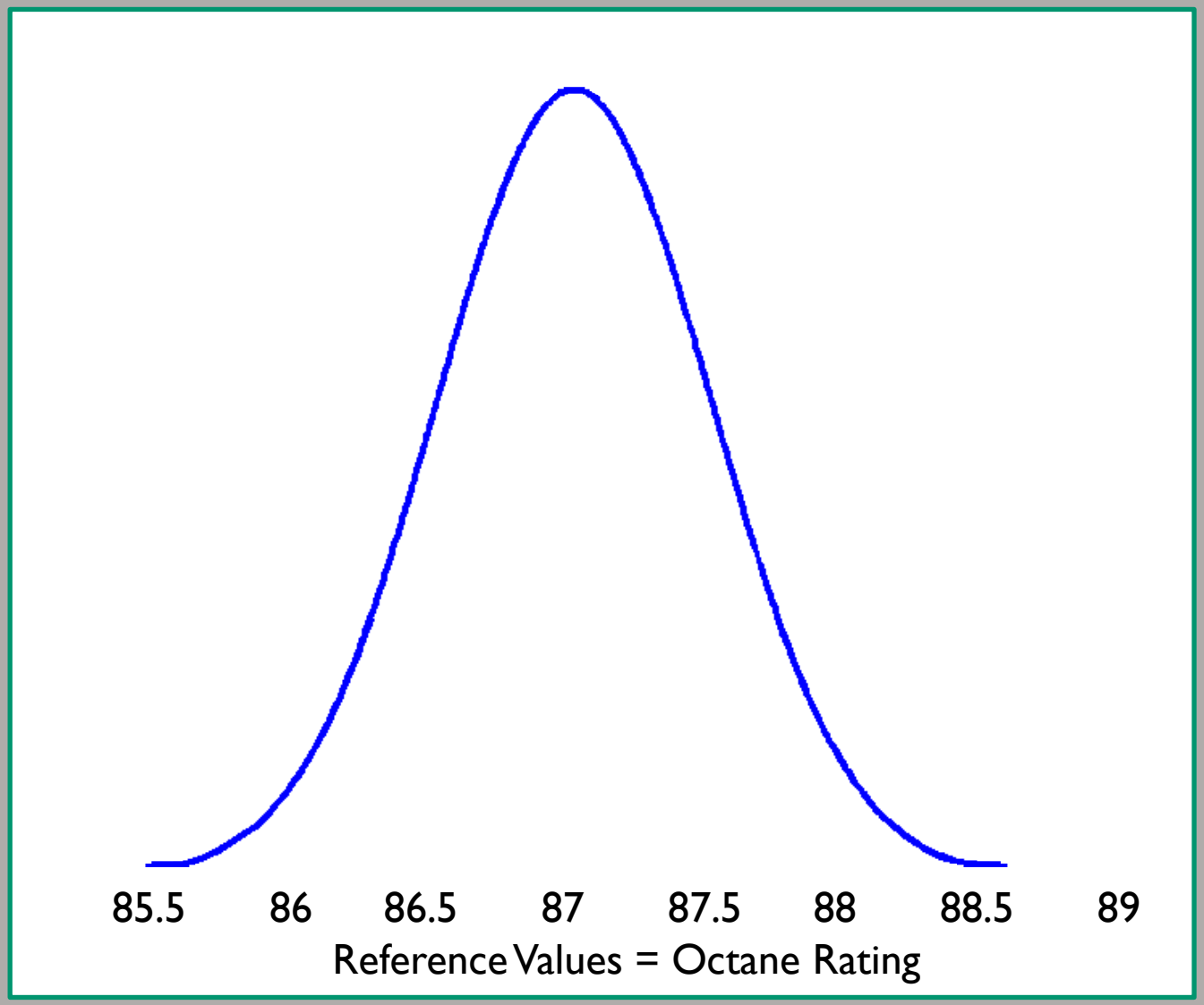
# Consider …

- New colleague says 'Hey, I live near you; how long does it take to get to work?"
- You answer "Today it took me 45 minutes, but there was an accident. Yesterday was a breeze; left a little early and I made all the lights and got here in 20."
- But, Fridays you stop for donuts, Thursdays you usually fill the tank, and sometimes you drop your kids off along the way; there's always something.
- So, what is the real driving time?
  - It depends on many factors
  - We can develop some feeling by getting multiple measures, over a period of time
  - A single day's result is not necessarily indicative of future expectation

# Building an understanding of Error



# Building an understanding of Error



# Sources of Error in All Chemometric Models Built for Spectrometers

- **Reference Error**
  - The reference method is not perfect
  - Nevertheless, it is considered to be true (accurate)
  - Error assessed in terms of precision and/or reproducibility
- **"Spectrometric" Error**
  - Instrument error + error in chemometric model
  - Not perfect
  - Precision is typically much better than reference method

# Understanding RMSEP

How are Reference Error, Spectrometric Error, and RMSEP related?

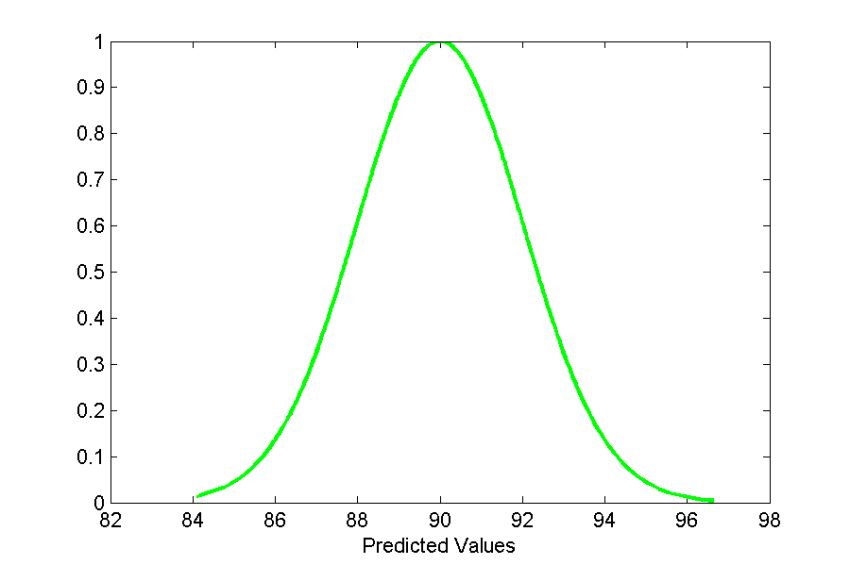
$$\mathbf{RMSEP} = \sqrt{(\mathbf{RefError}^2 + \mathbf{SpectrometricError}^2)}$$

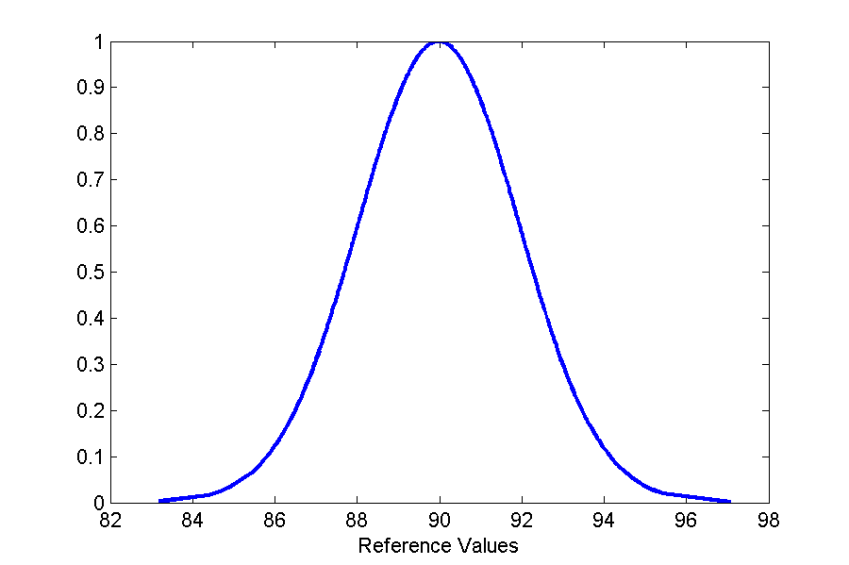
Two examples to illustrate

- Example A looks at a single sample
- Example B considers a population of samples

# Case A1

## Reference and Spectrometric Error are Similar



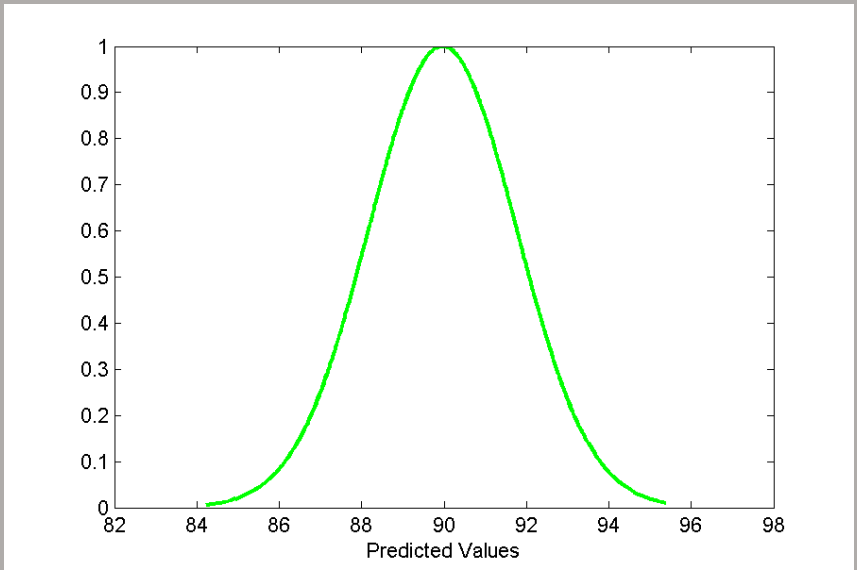



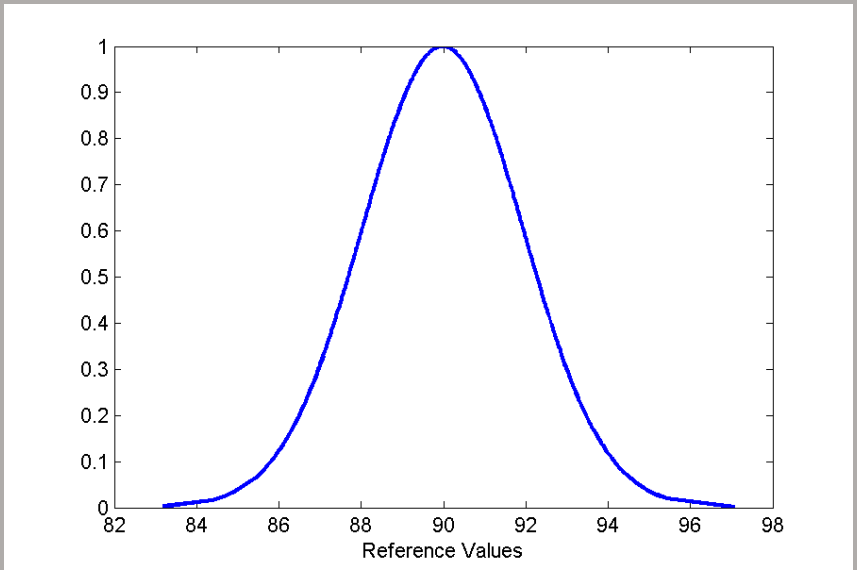


Infometrix

# Case A2

## Spectrometric Error reduced



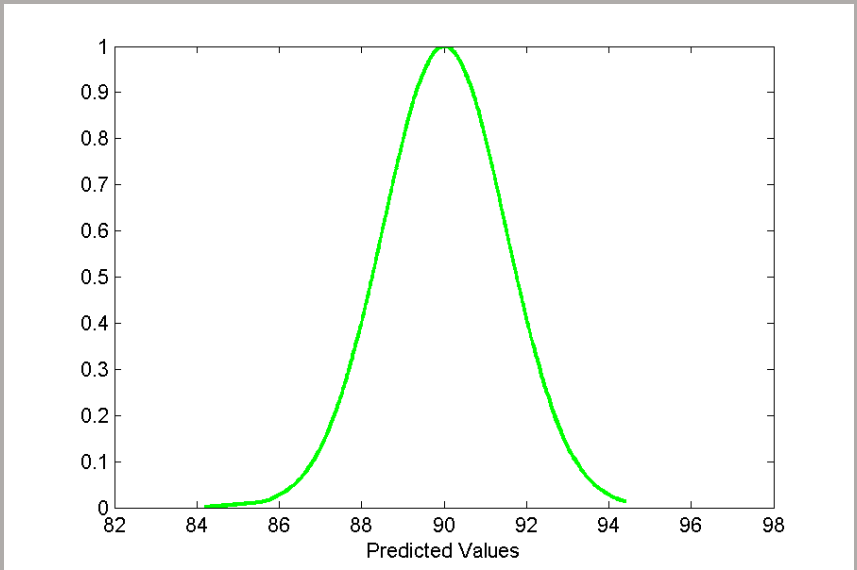



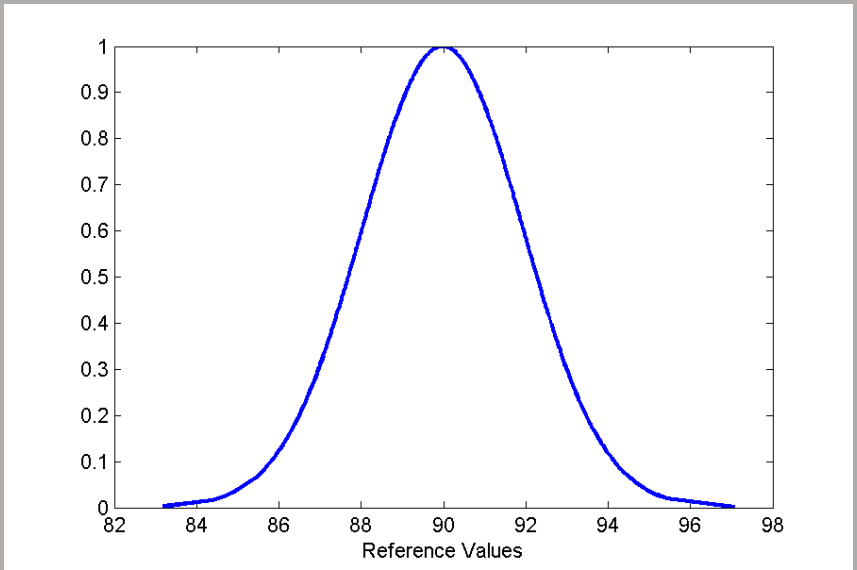


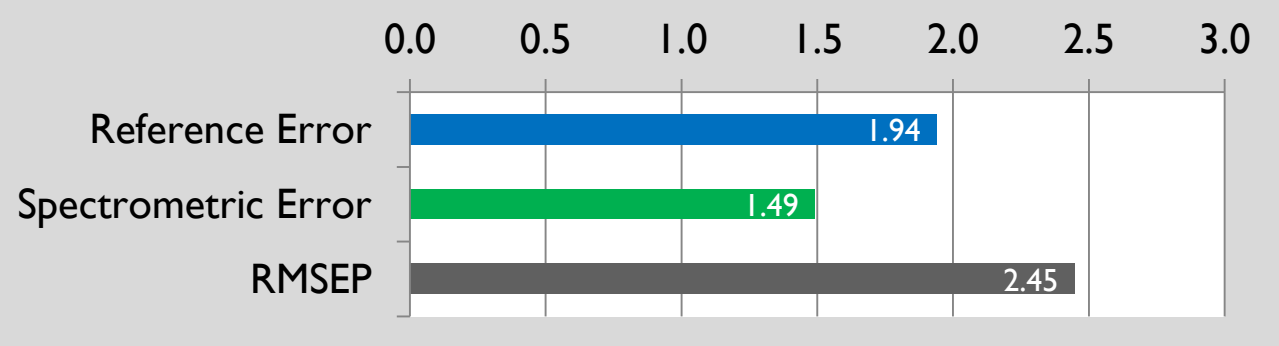
Infometrix

# Case A3

## Spectrometric Error reduced further…



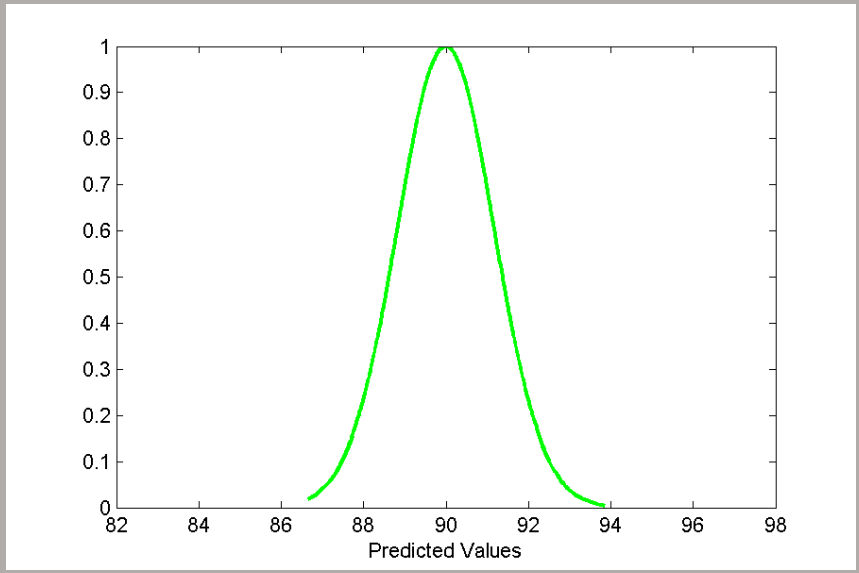



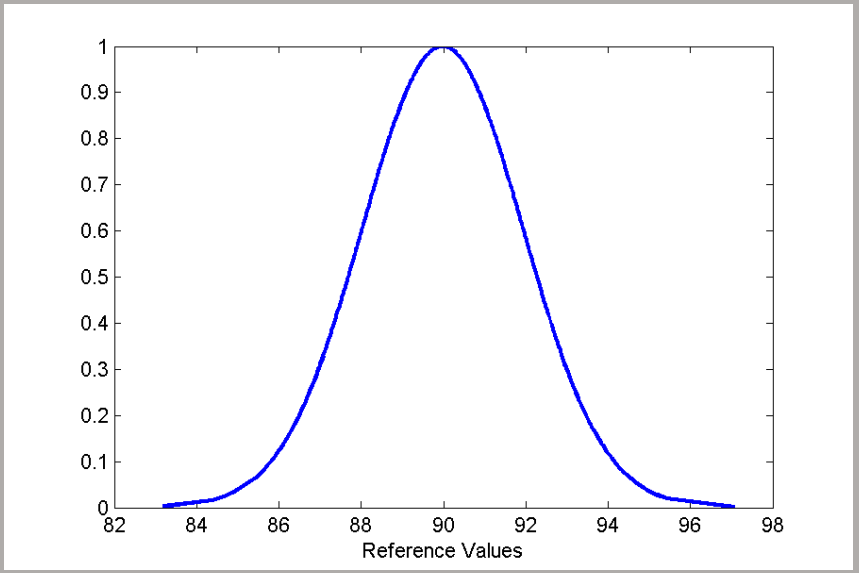



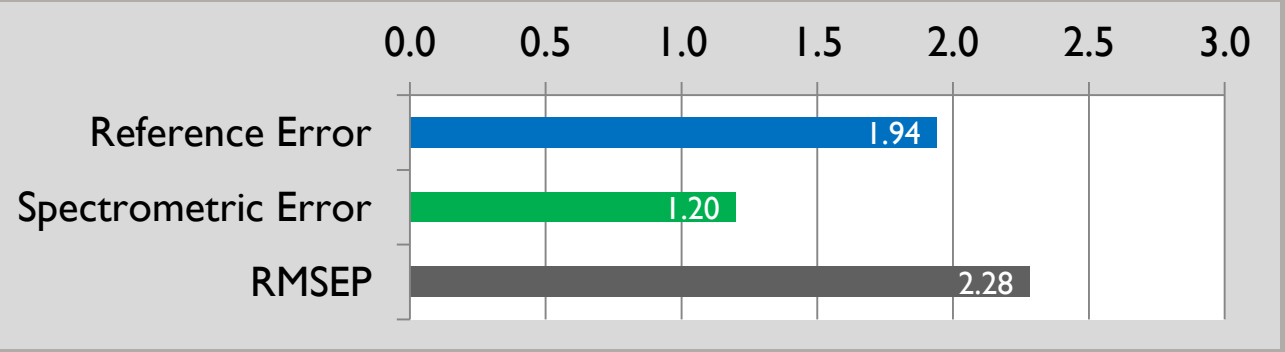

Infometrix

# Case A4

## …and further…



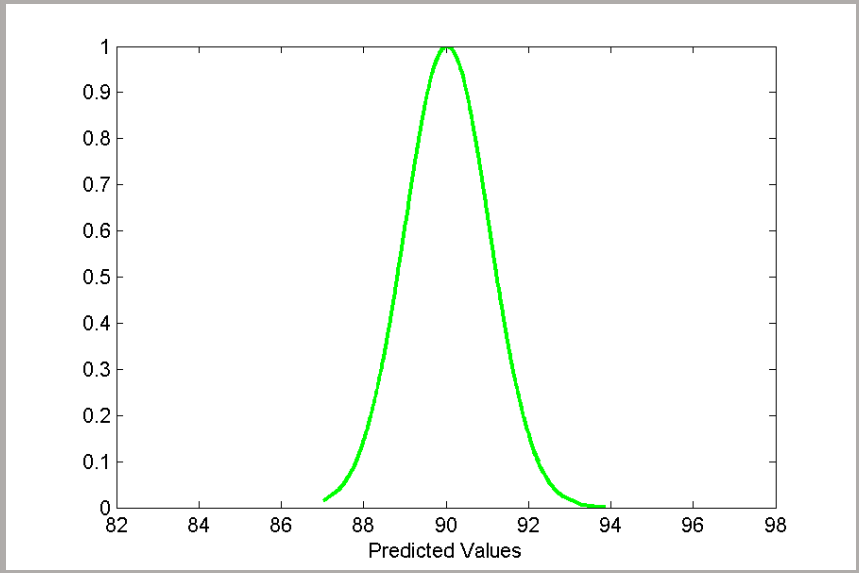



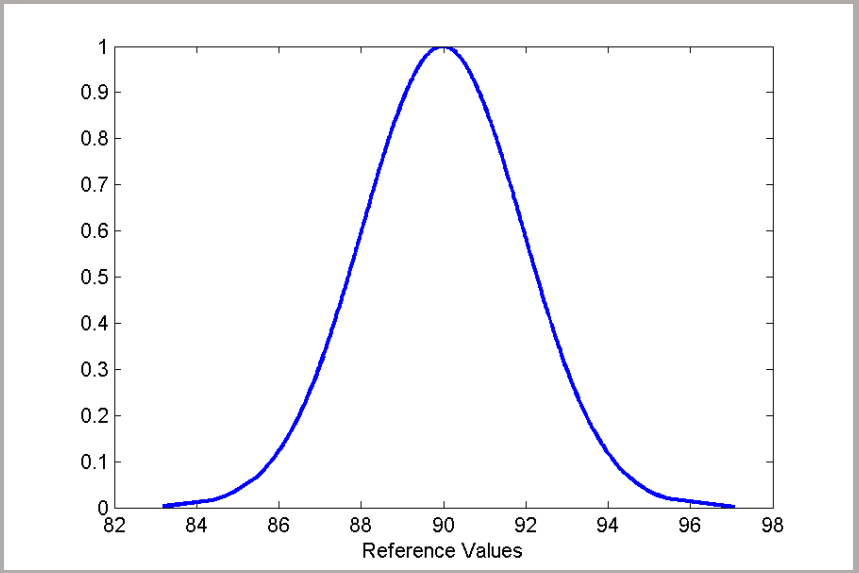



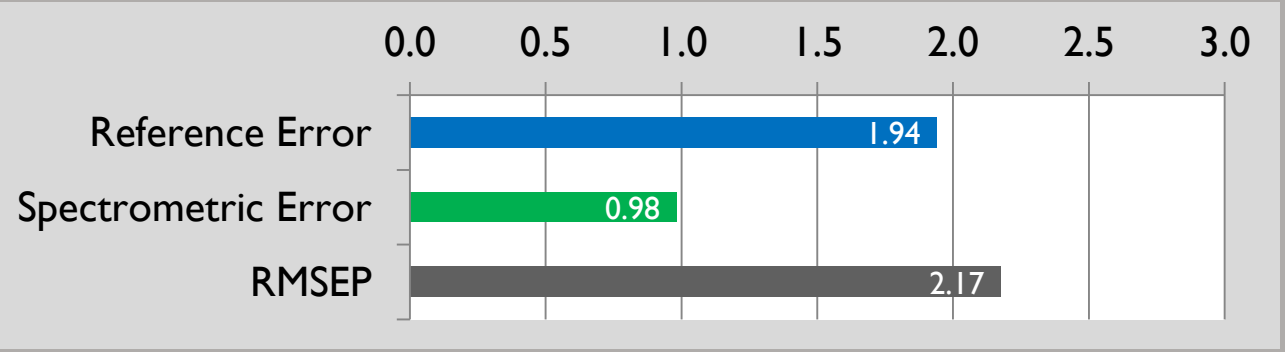

Infometrix

# Case A5

## …and further…



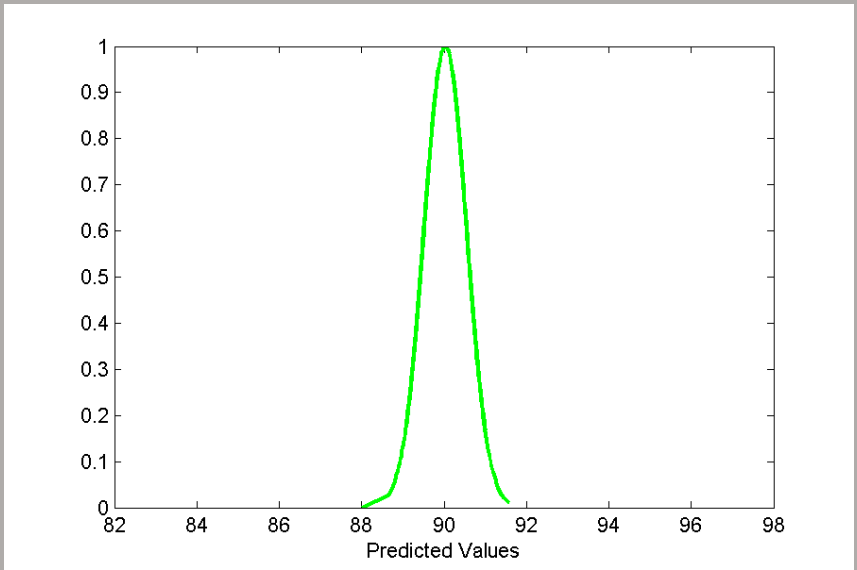



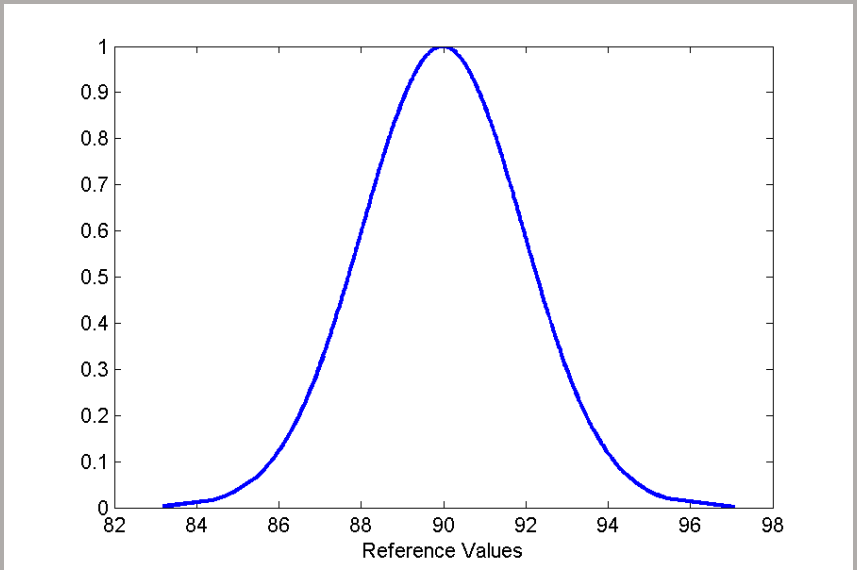



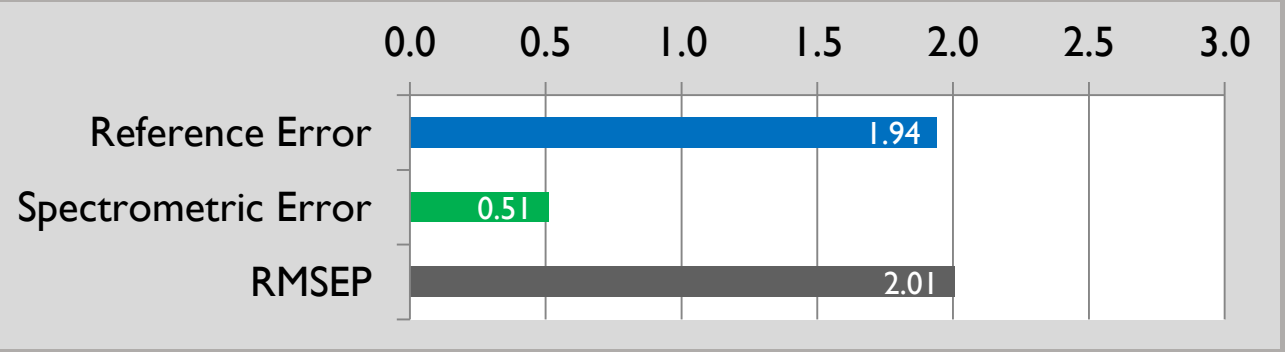

Infometrix

# Case A6

## …and further…



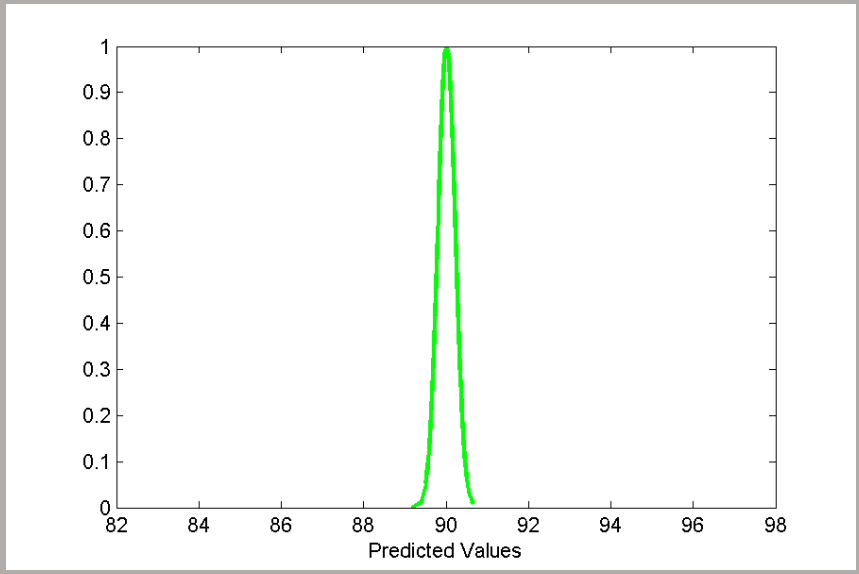



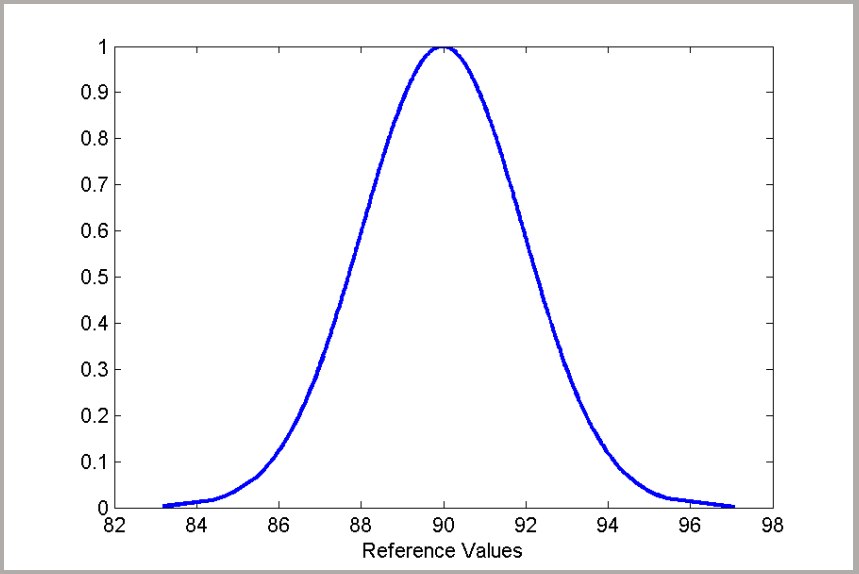



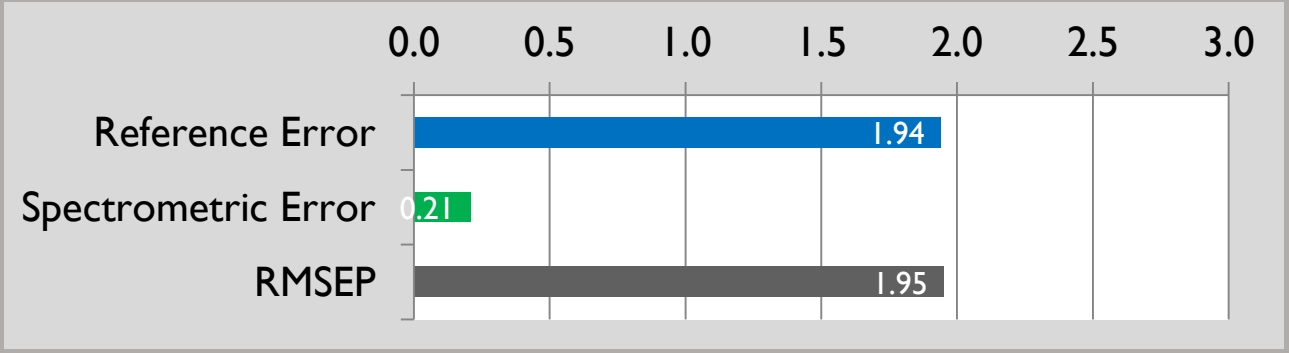

Infometrix

# Case A7

## …and further…



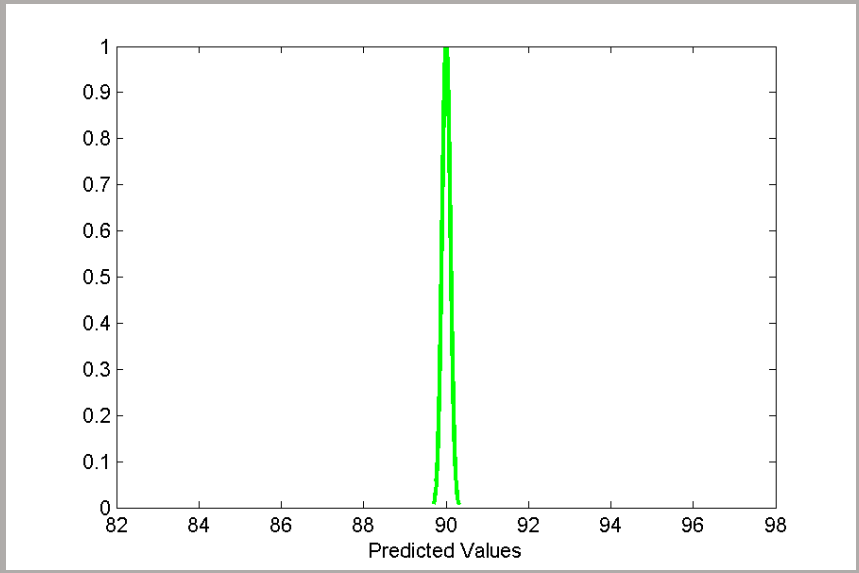



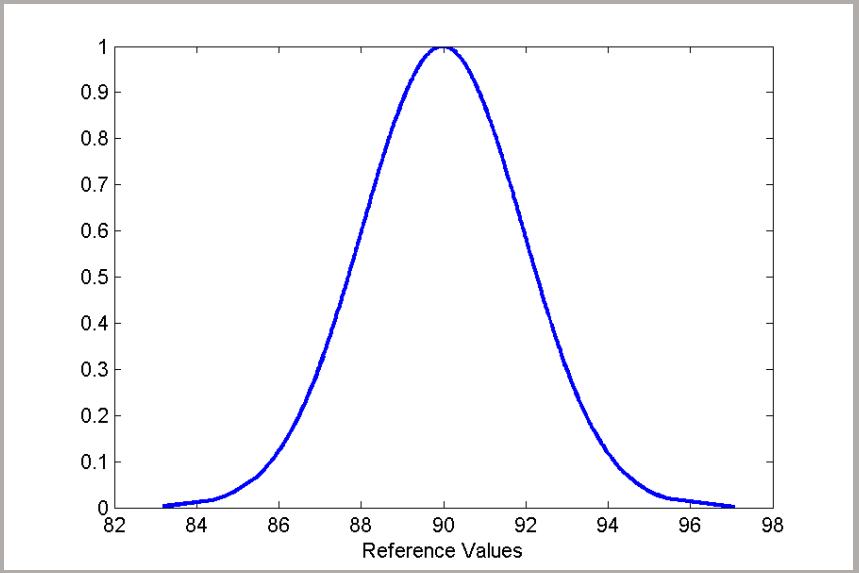



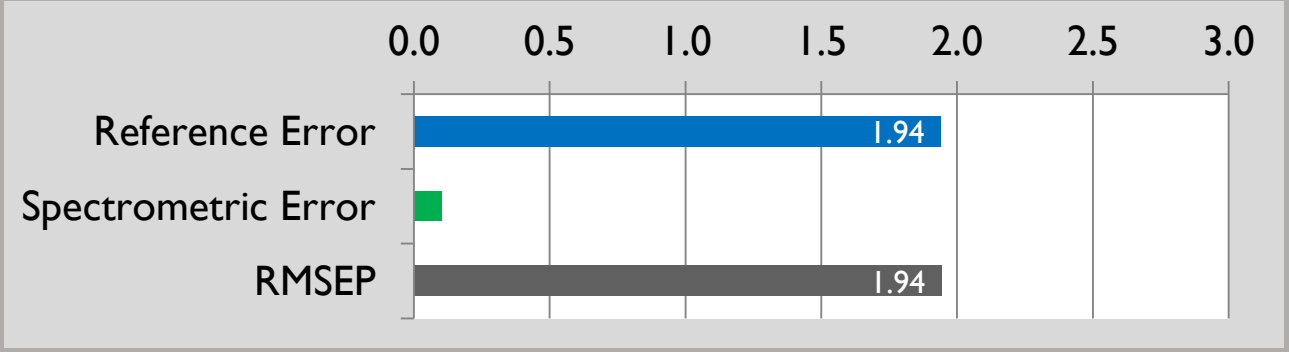

Infometrix

# Case A8

## …and further…



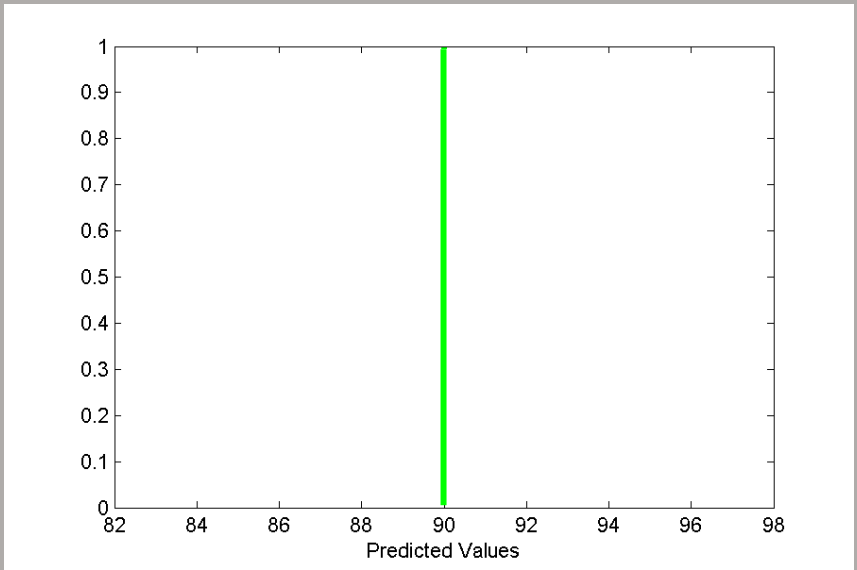



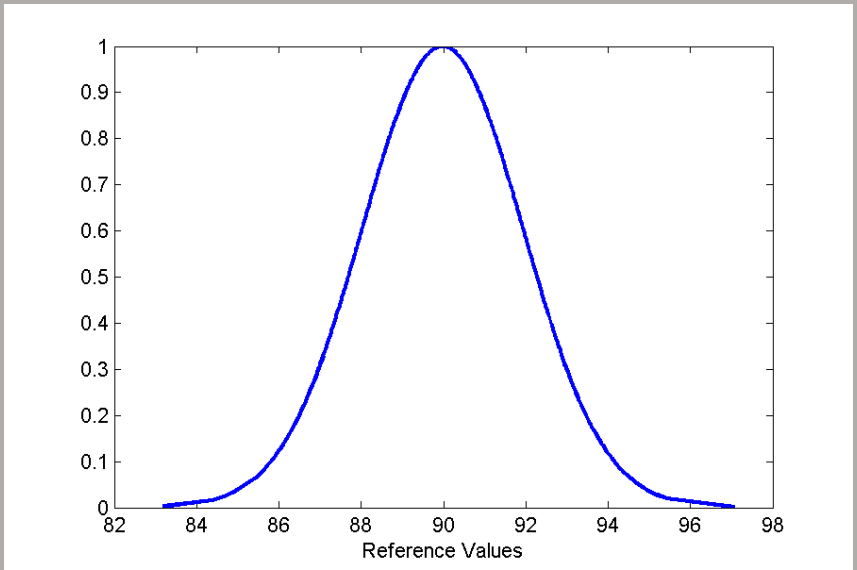



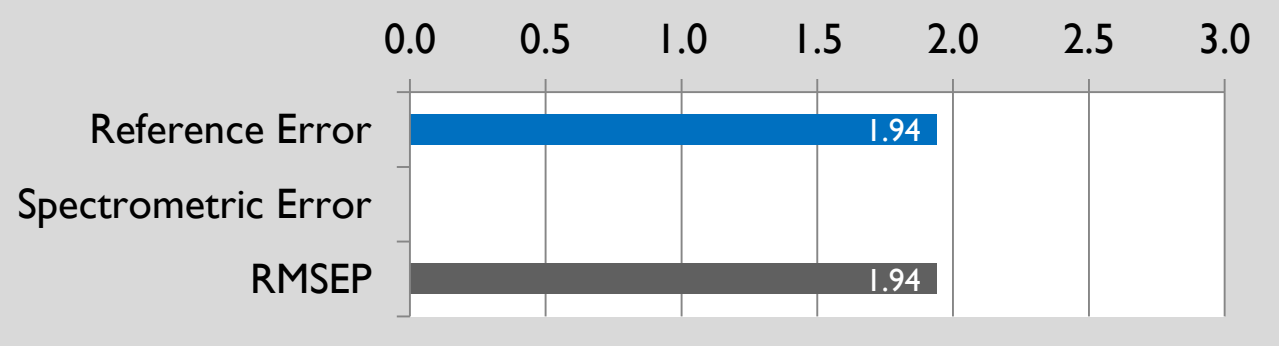

Infometrix

# Case A9

## …until the Spectrometric Error is zero



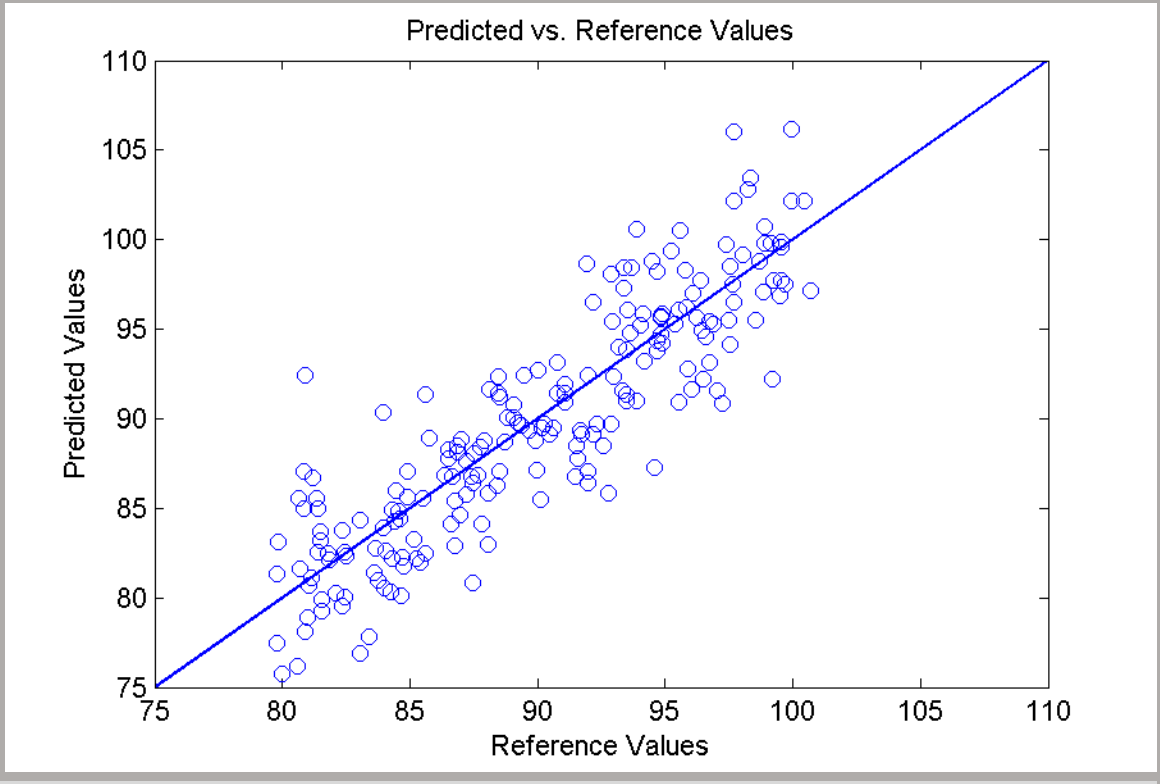



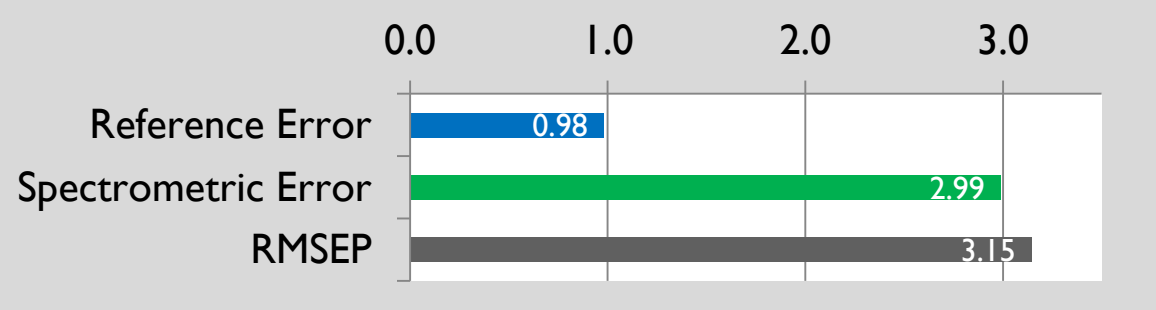




Infometrix

# Case B1

Spectrometric Error >> Reference Error

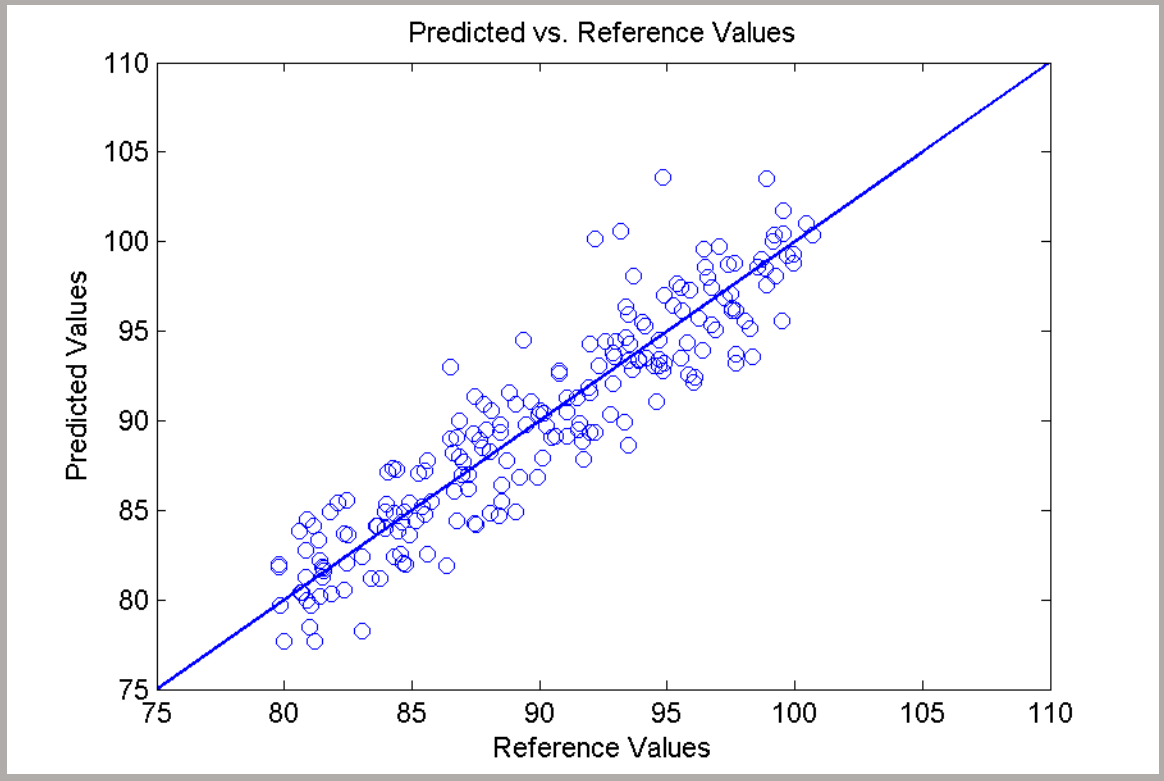


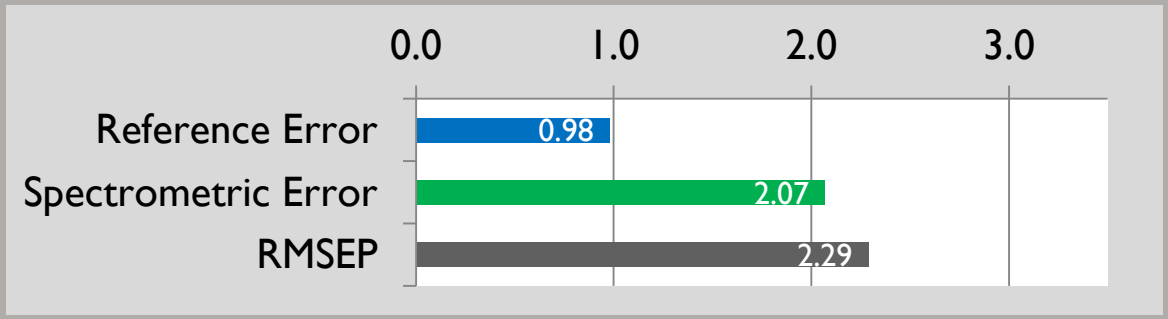




Infometrix

# Case B2

Spectrometric Error decreases by ~1/3, RMSEP decreases by < 1/3

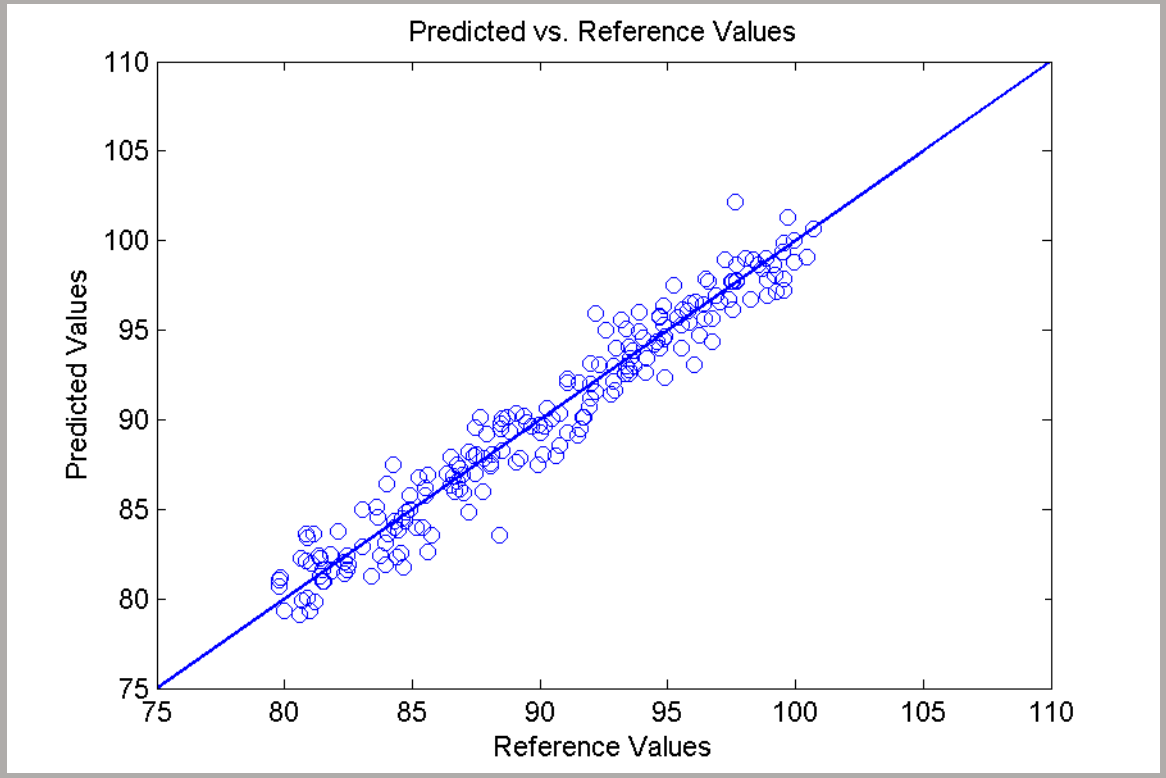


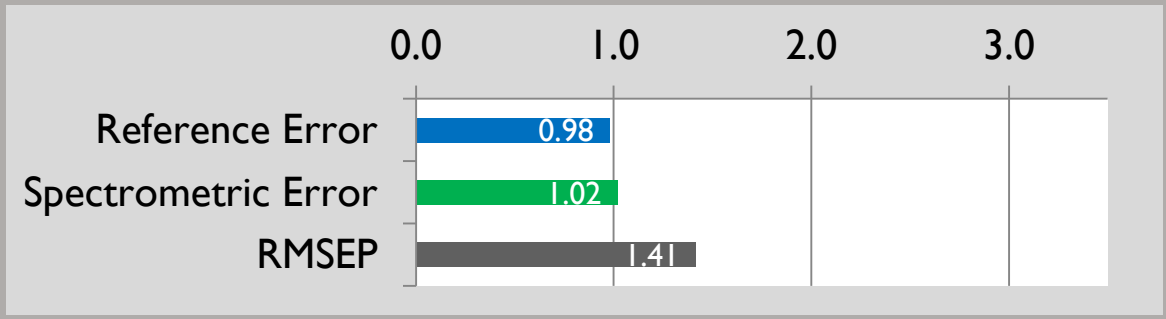




Infometrix

# Case B3

Spectrometric and Reference Errors are about equal

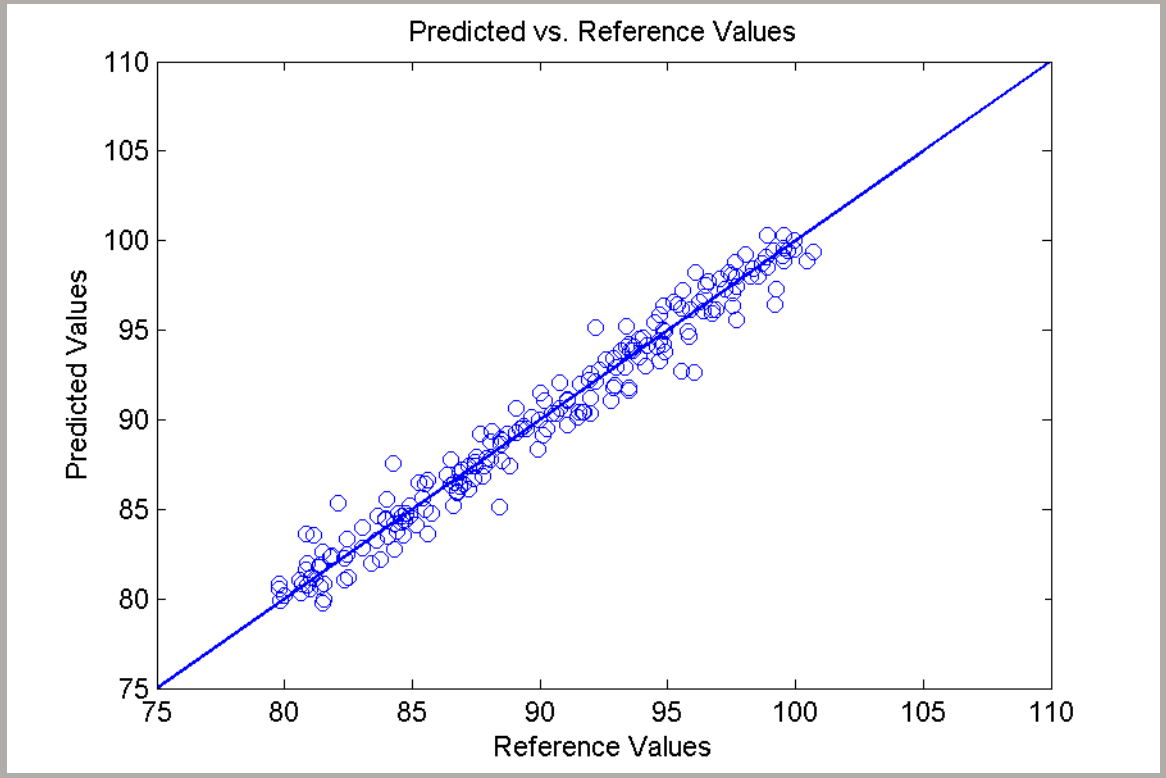


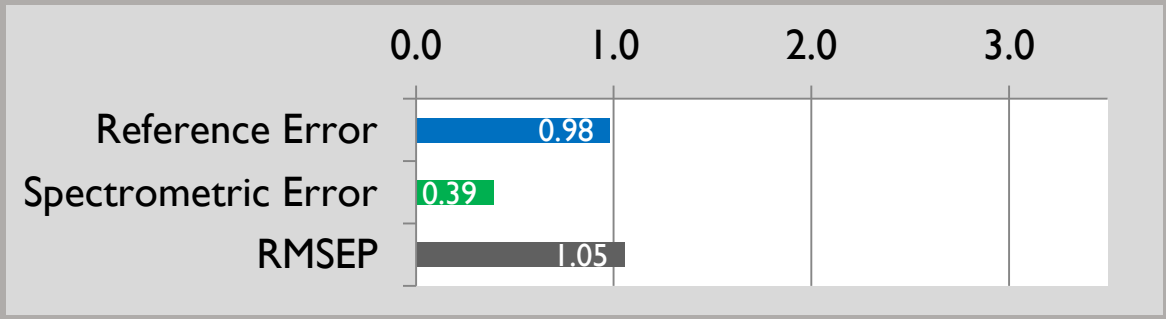




Infometrix

# Case B4

Spectrometric Error is about 1/3 the Reference Error

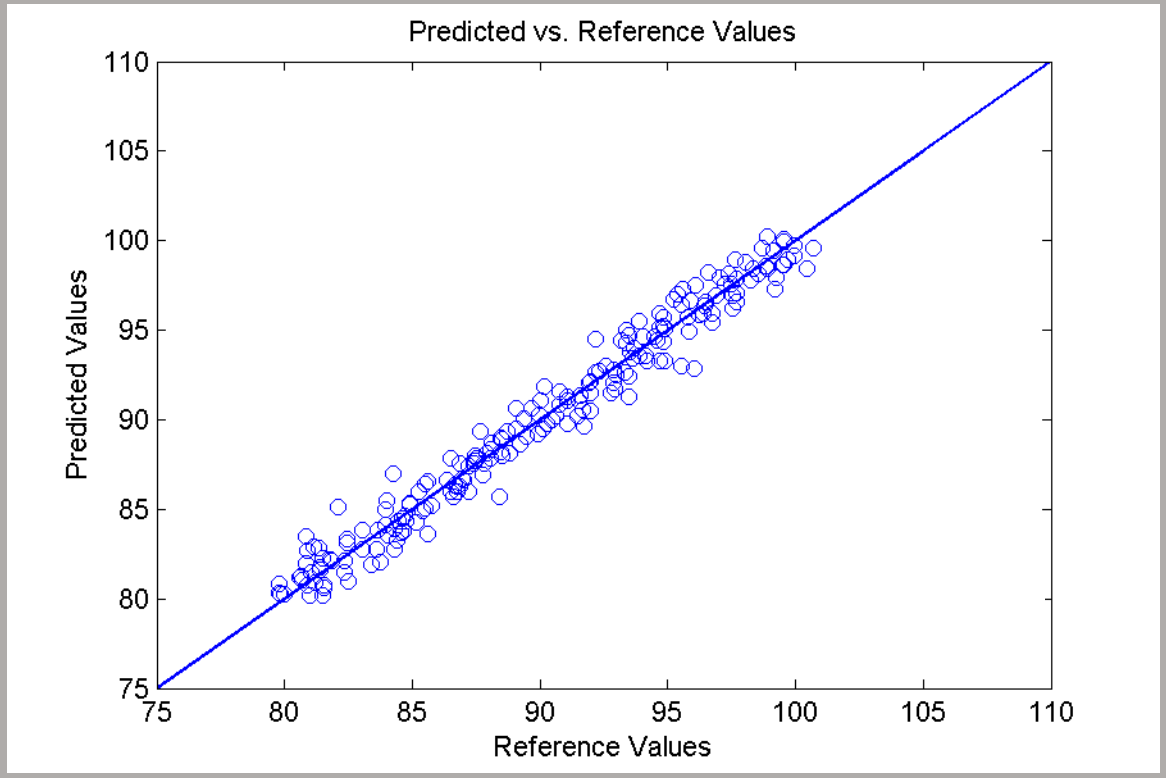


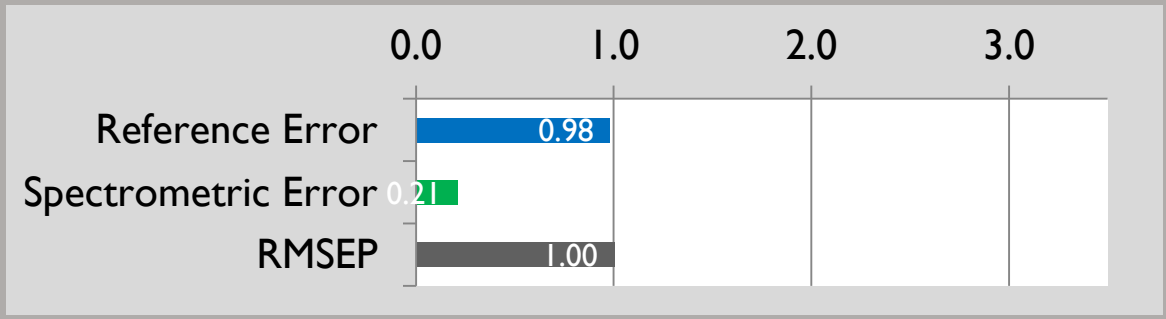




RMSEP is approaching Reference Error

# Case B5

Spectrometric Error reduced by ~50% vs Case B4, little effect on RMSEP

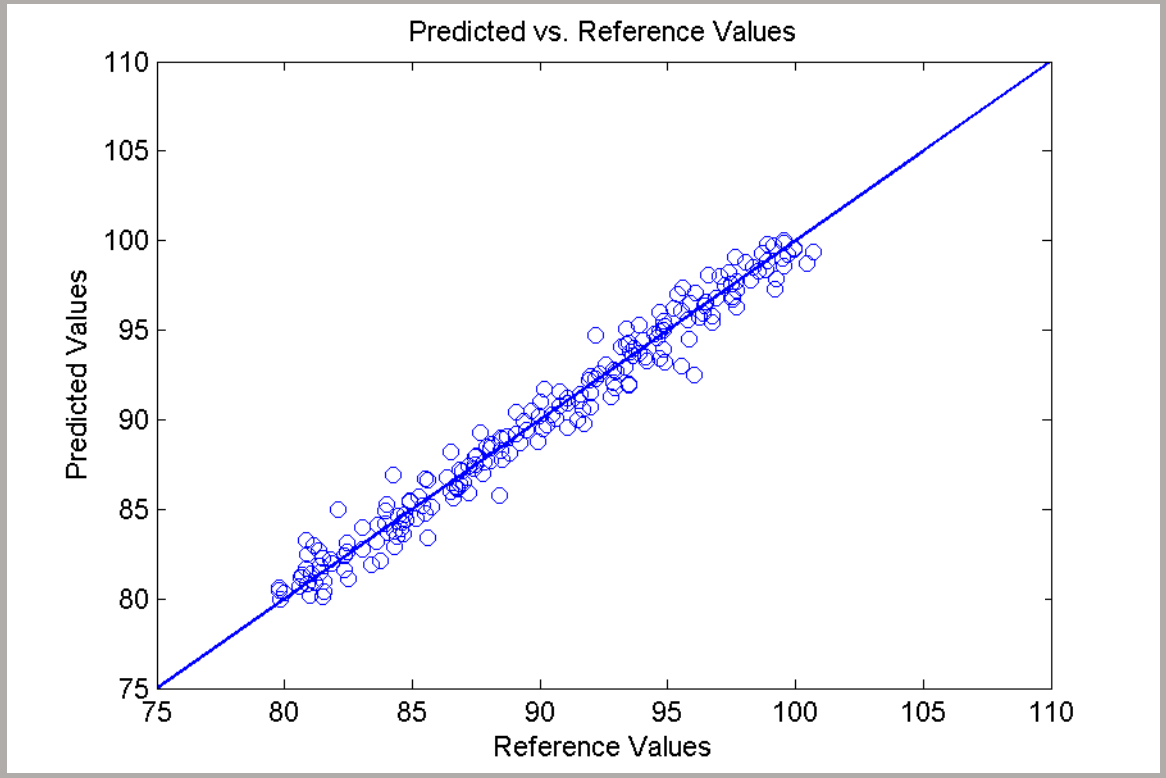


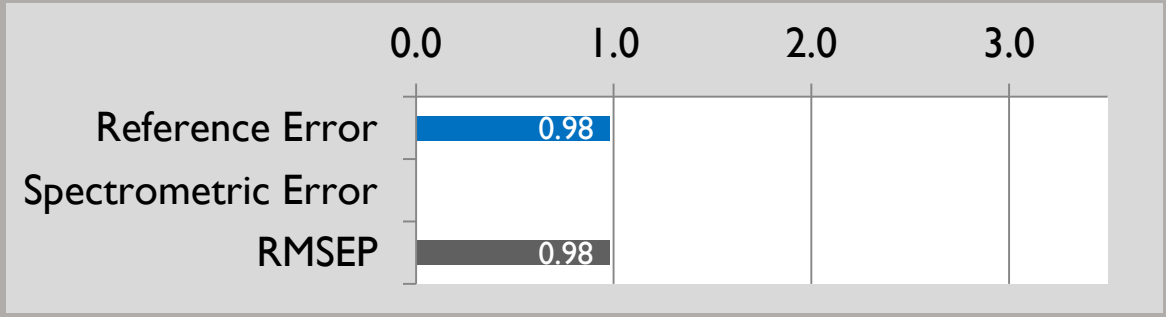




RMSEP is about the same as Reference Error

# Case B6

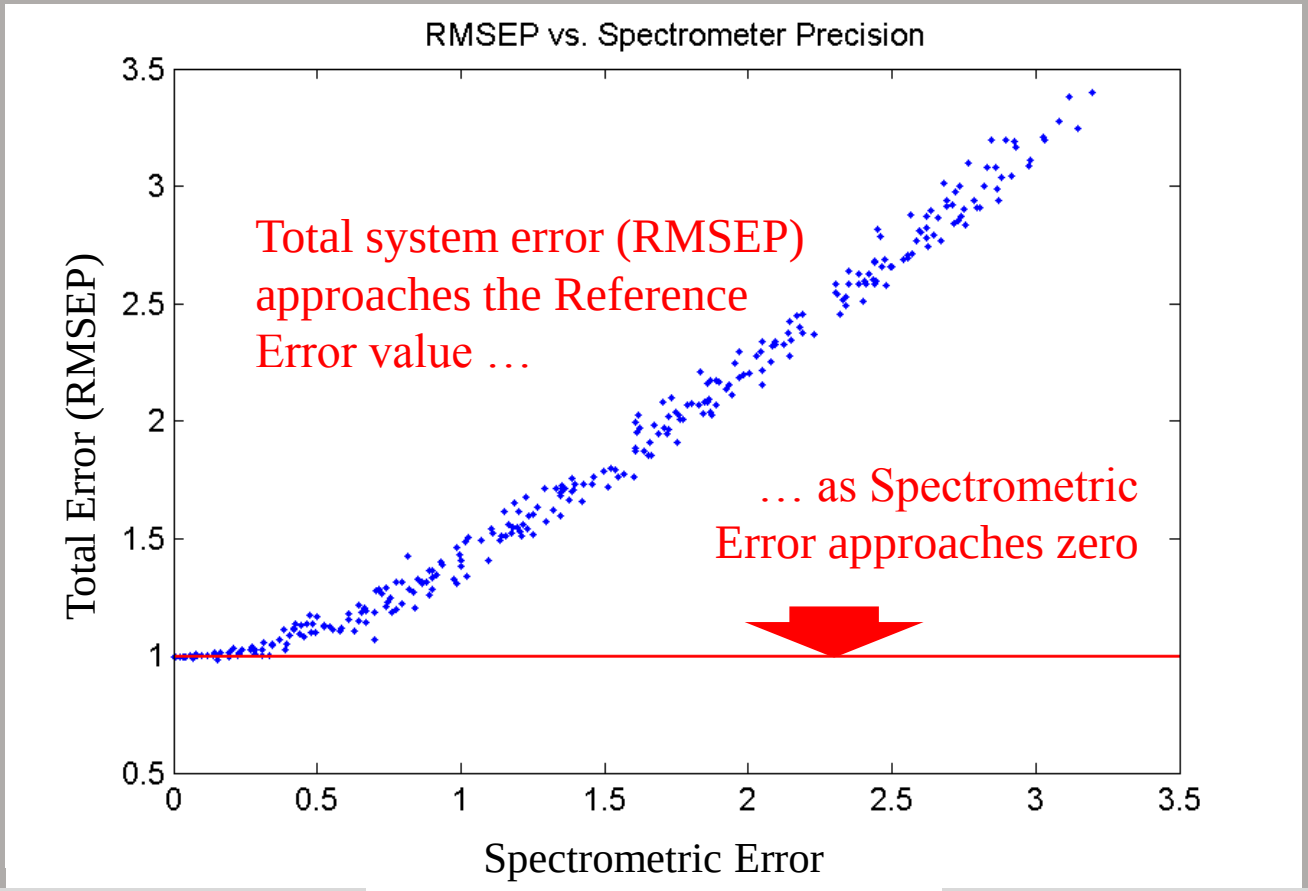
Spectrometric Error is zero, RMSEP equals Reference Error



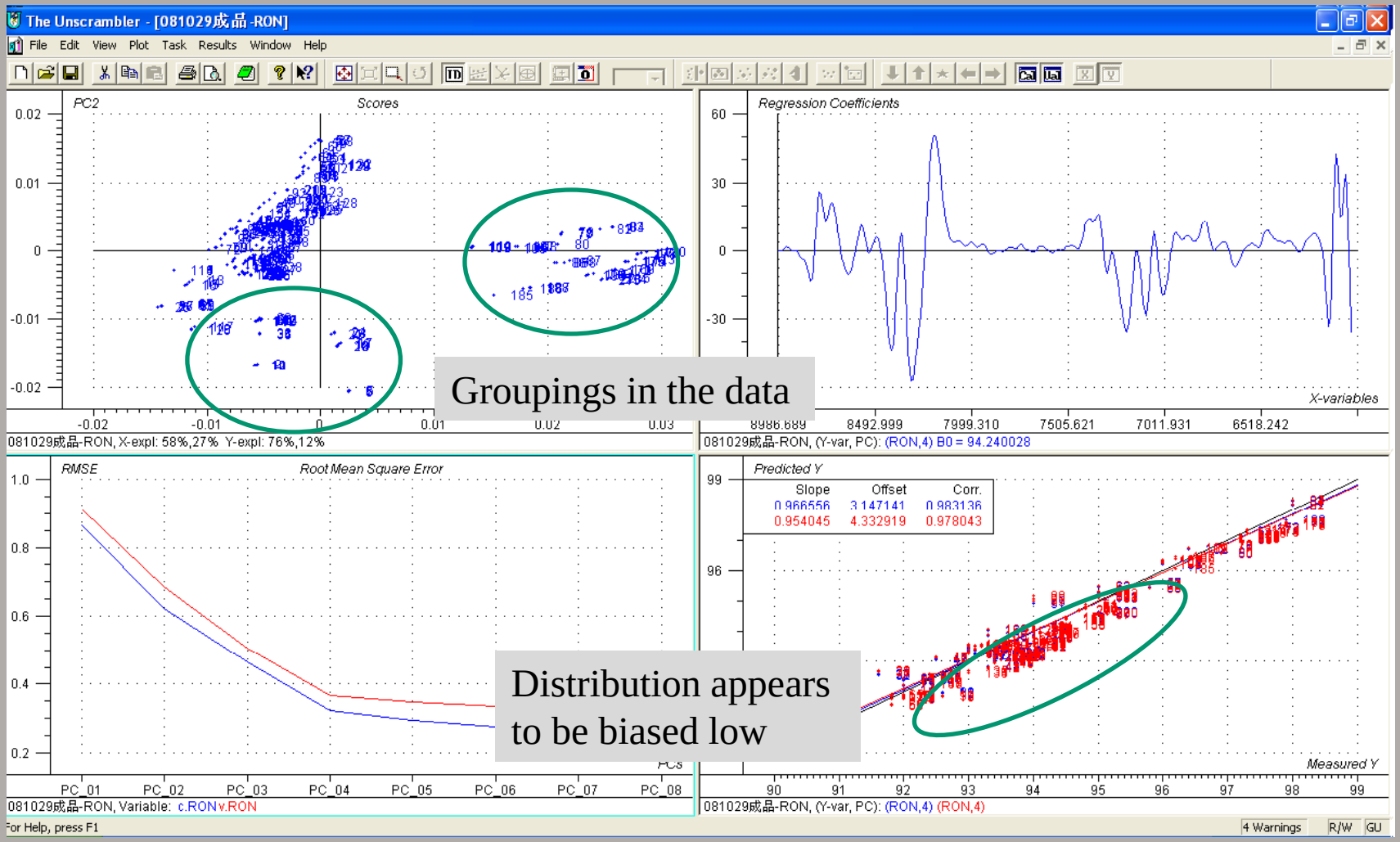

Infometrix

# RMSEP is Limited by Reference Error




At Spectrometric Error = 0, the incorrect conclusion is that "the model cannot be better than the reference method"

# Model Construction Issues




Groupings in the data

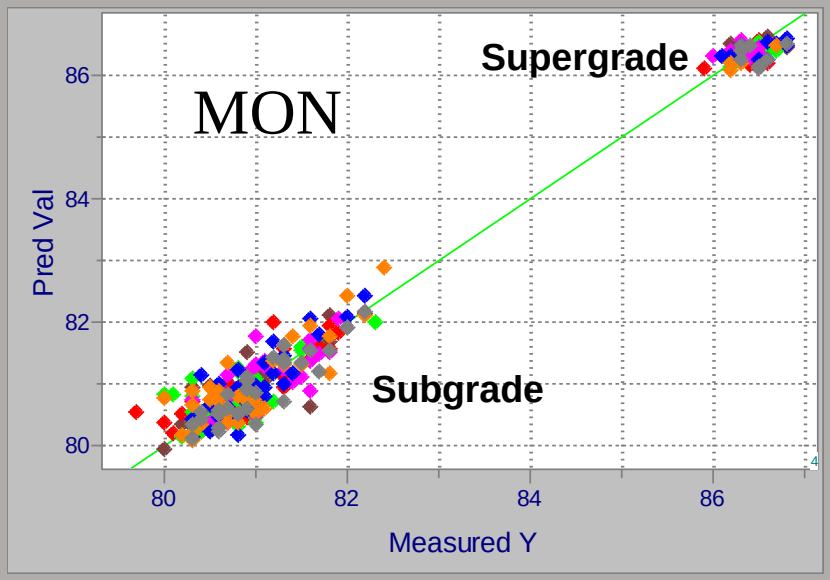
Distribution appears to be biased low

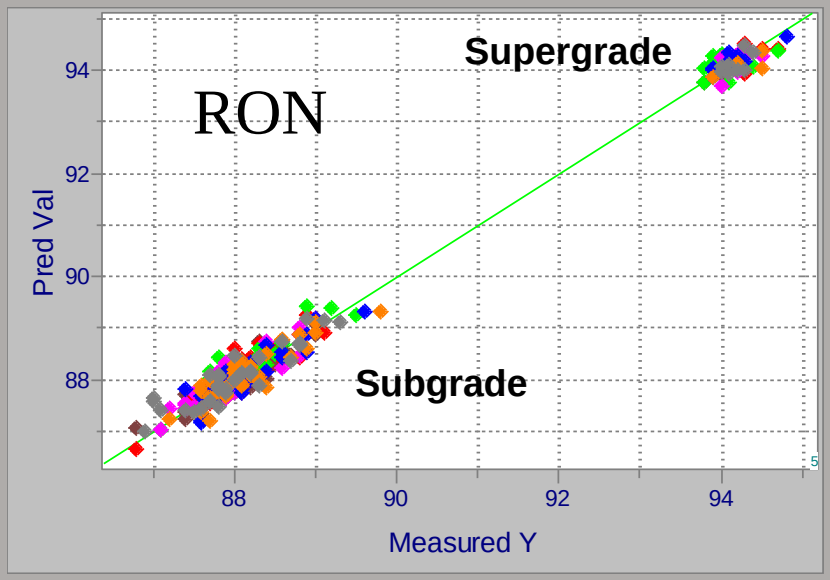
***A clear case of not handling the data properly***

# Segmented Predictions

- When samples do not group into a single homogeneous cluster, breaking the calibration problem into multiple regressions might improve the overall prediction quality.
- In the case of motor fuel properties, the chemical composition of sub-grade and super-grade fuels is sufficiently different that a hierarchical, or segmented, approach could be attempted:
  - Use one model to classify the grade
  - Use a second model, one for each category found by the first model, to determine the quality rating

# Sub- and Super-Grade Fuels




Feeding-forward information on the blend reduces the property range that is being modeled and can provide a better model and more precise measurement of fuel properties.

# Comparison of Octane Models

- For the data set that includes both grades, the standard error of cross validation for RON is 0.22 and for MON is 0.25.
- If we build a model for the two ranges of octane measurements separately, there is just a small change in the evaluation of sub-grade gasoline, but there is a significant drop in the error associated with premium.

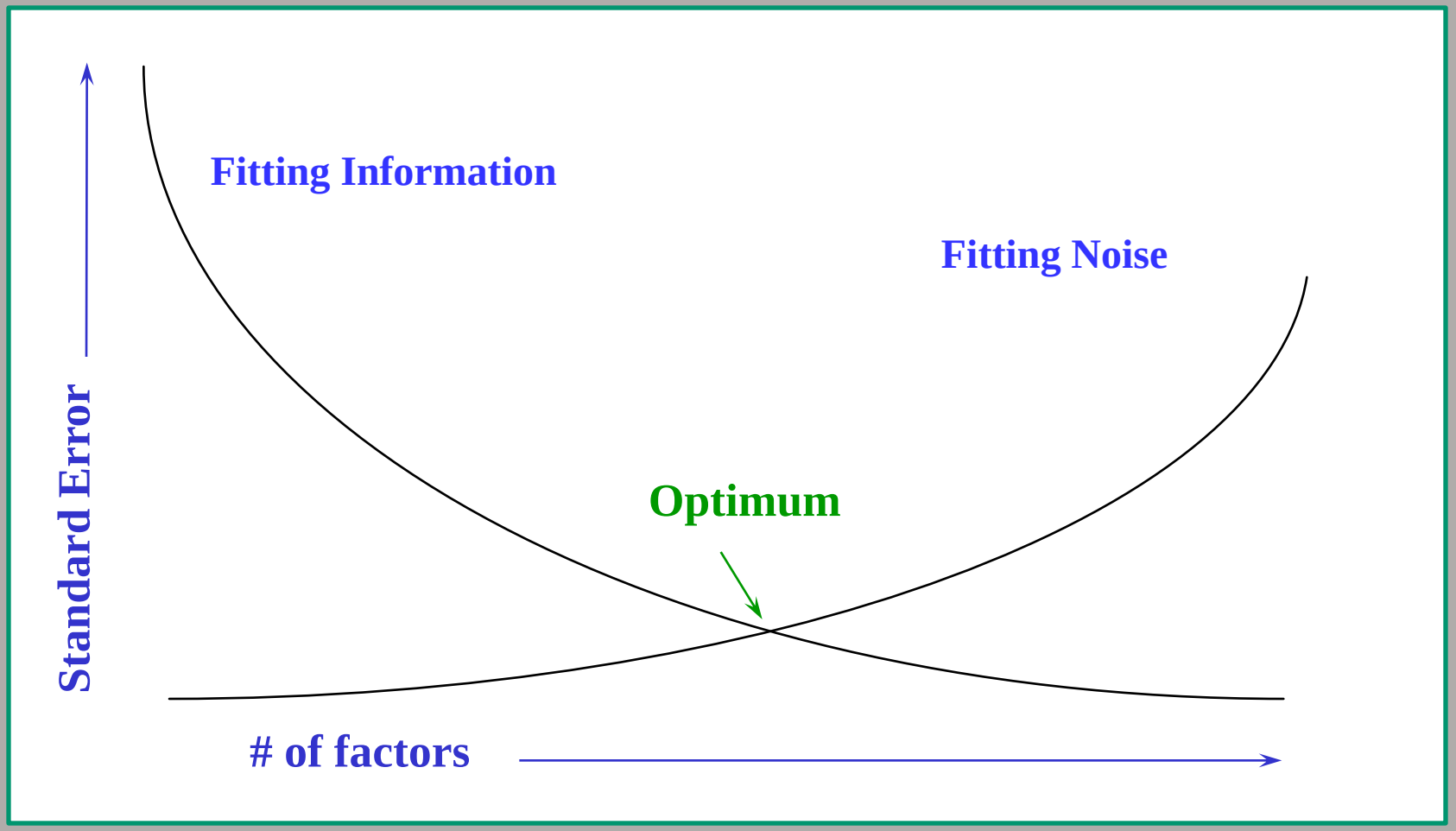
| | All Samples | Subgrade | Premium |
|---|---|---|---|
| RON | 0.22 | 0.22 | 0.15 |
| MON | 0.25 | 0.23 | 0.20 |

# Result: a More Parsimonious Model

- Looking at several properties, the value of integrating readily-available information improves the fuel property assessment considerably.
- Note that the number of factors required to build the PLS model is significantly reduced and likely provides more stability over time.

| | All Samples | | Subgrade | | Premium | |
|---|---|---|---|---|---|---|
| | Outliers Removed | | Outliers Removed | | Outliers Removed | |
| | SECV | Factors | SECV | Factors | SECV | Factors |
| Aro | 0.87 | 7 | 0.76 | 3 | 0.84 | 2 |
| Ole | 0.58 | 7 | 0.57 | 4 | 0.32 | 4 |
| RON | 0.23 | 7 | 0.22 | 5 | 0.15 | 3 |
| MON | 0.24 | 7 | 0.23 | 5 | 0.20 | 5 |
| Benz | 0.04 | 8 | 0.02 | 5 | 0.02 | 4 |
| IBP | 1.99 | 7 | 1.83 | 3 | 1.55 | 6 |
| d50 | 1.54 | 8 | 0.98 | 7 | 1.18 | 4 |
| d90 | 3.69 | 9 | 3.66 | 4 | 1.65 | 4 |

# Choosing Best # of Factors

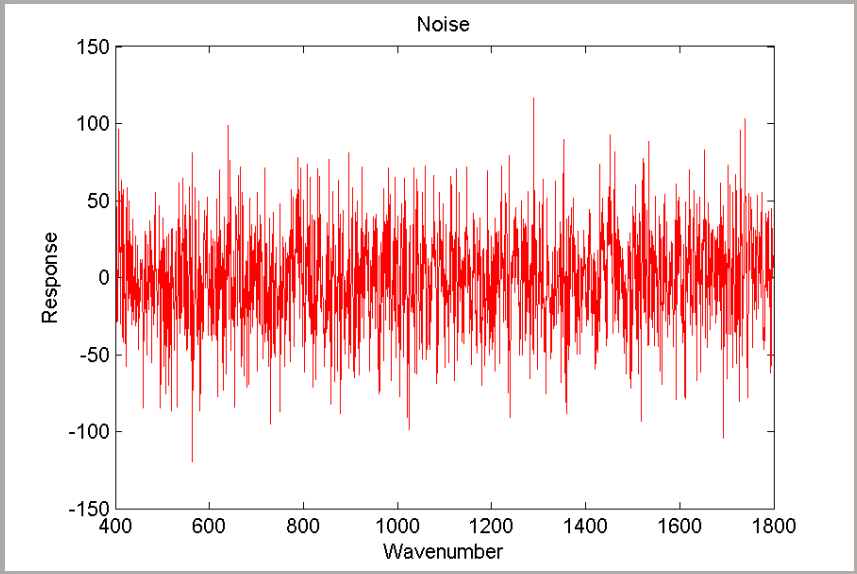


# Avoiding Overfitting

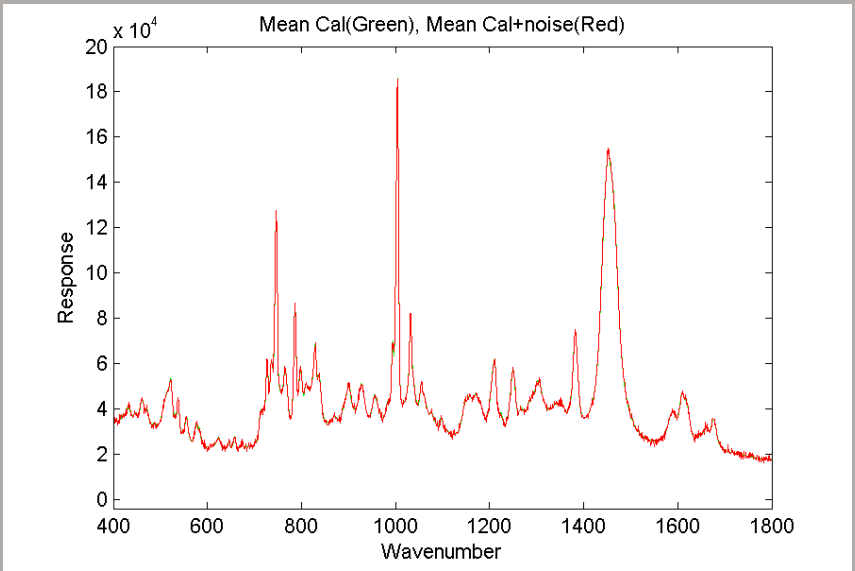
- As more factors are added
  - calibration error is reduced
  - noise variation in the training data are built into the model
- Prediction data will have different noise; after all, noise is random
- A model with more factors than necessary will try to apply the noise portion of the model to prediction samples – and fail – resulting in increased error

# Noise added to mean spectrum

## Noise, std dev = 35

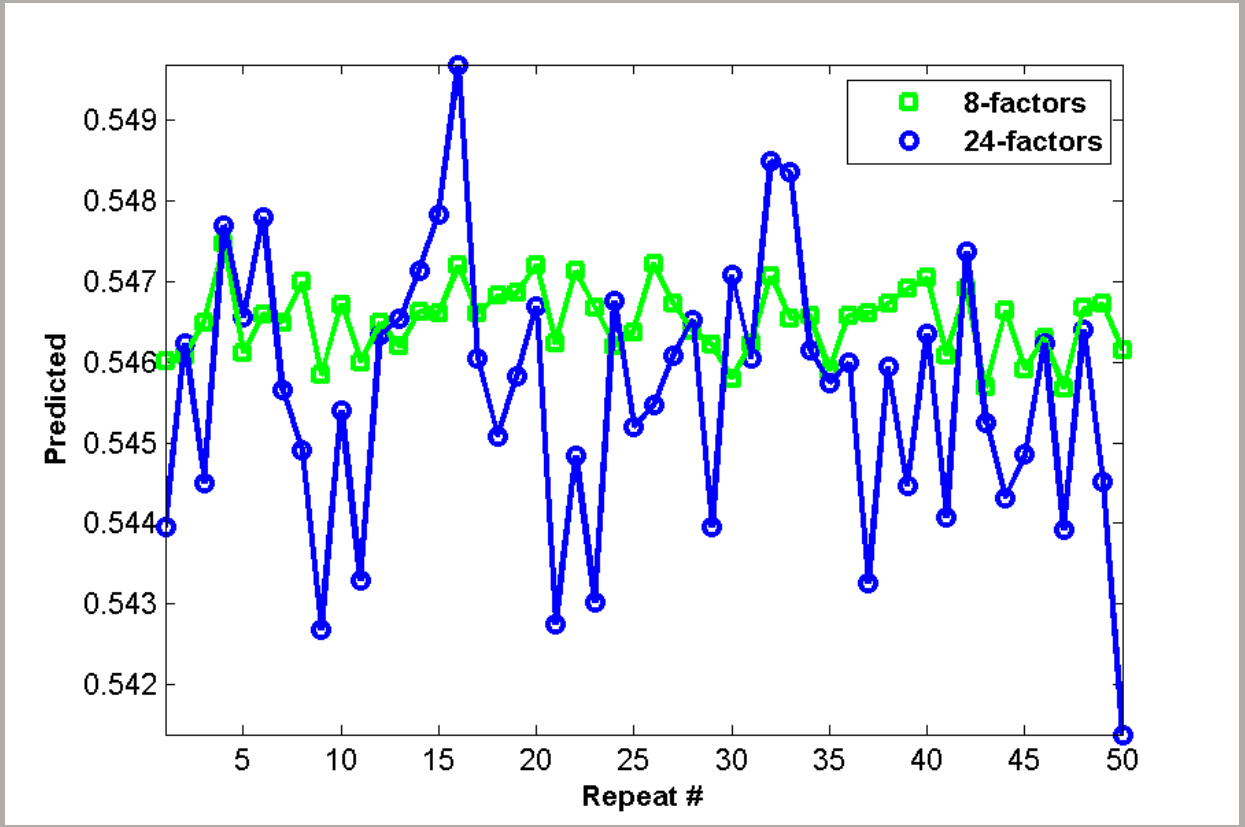



## Mean spectrum + noise




Infometrix

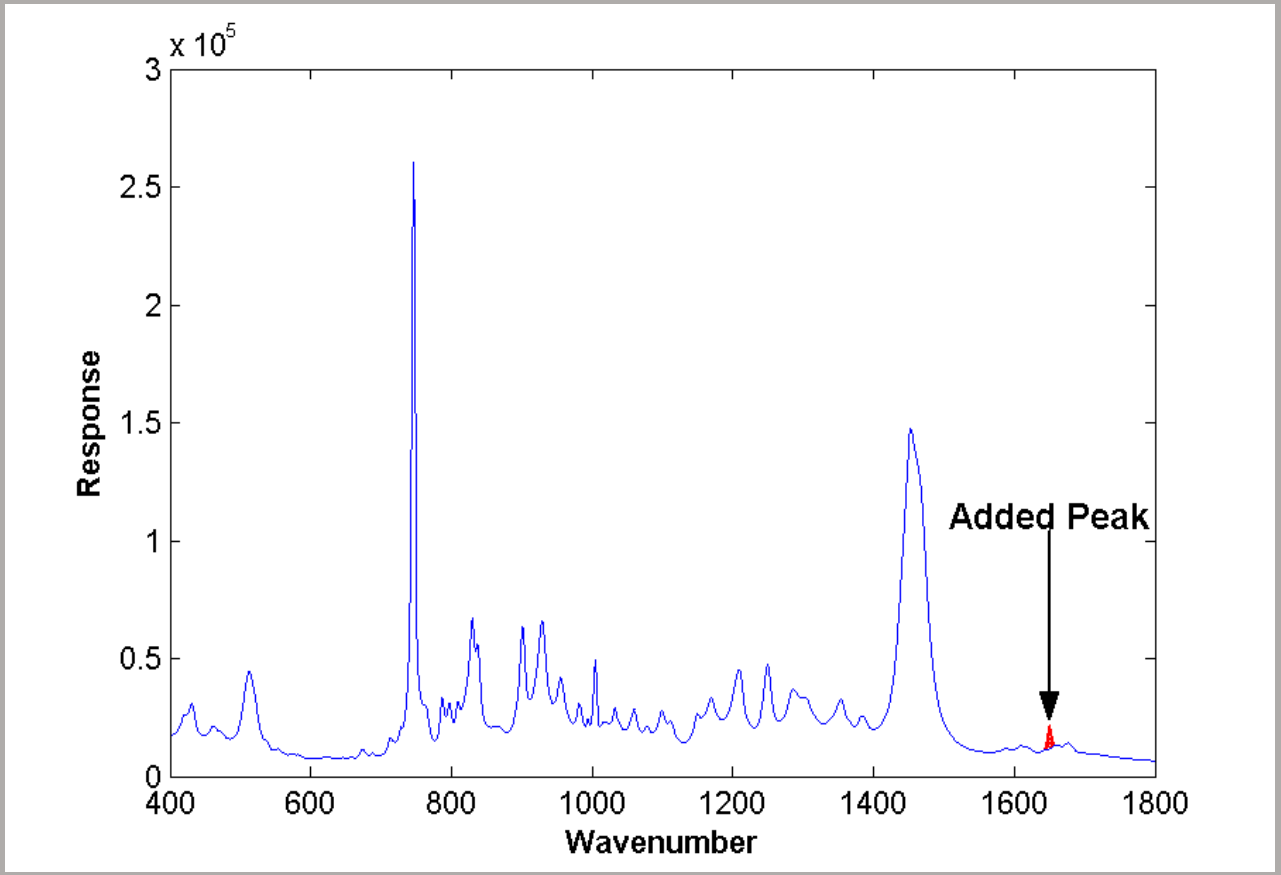
# Impact of Noise on Predictions



# Avoiding Overfitting 2

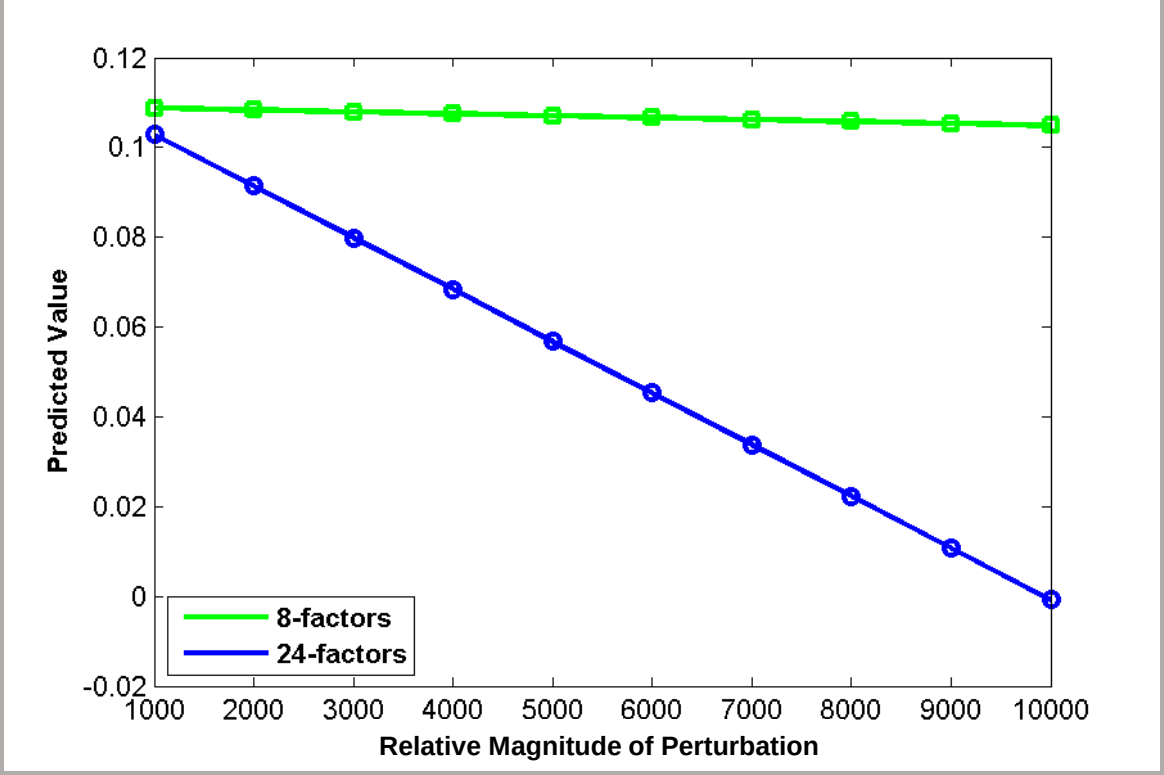
- As more factors are added
  - calibration error is reduced
  - information from interferent not correlated to property may be added to model
- Prediction data will have different levels of interferent
- A model with more factors than necessary will try to apply the uncorrelated portion of the model to prediction samples – and fail – resulting in increased error

# Addition of a Small Perturbation




Added peak intensity: 1,000 to 10,000

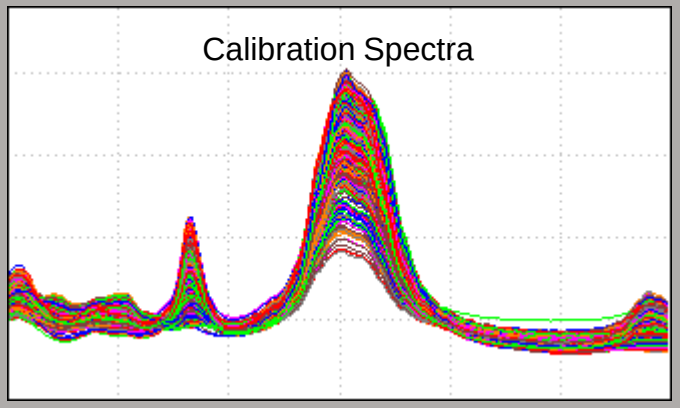
# Overfitting's Impact on Predictions

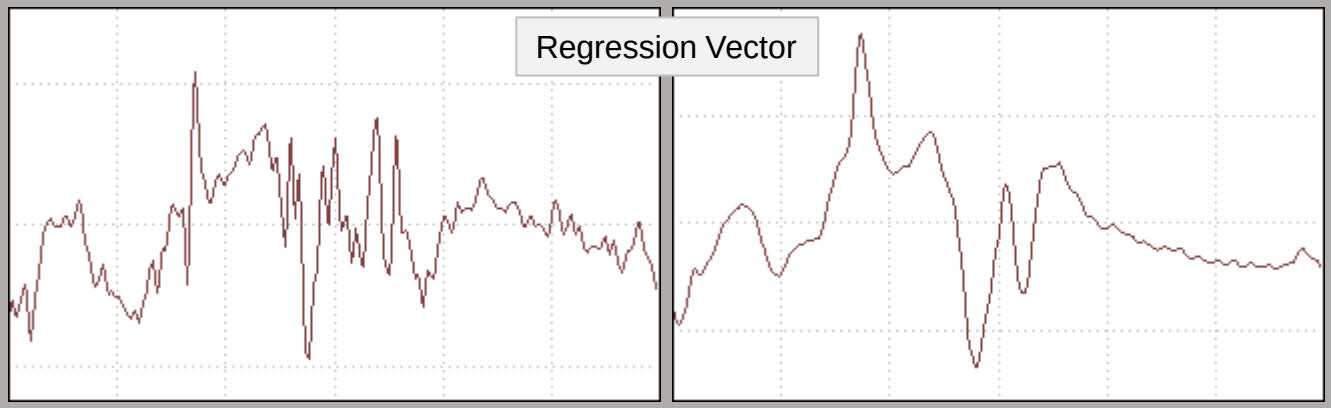


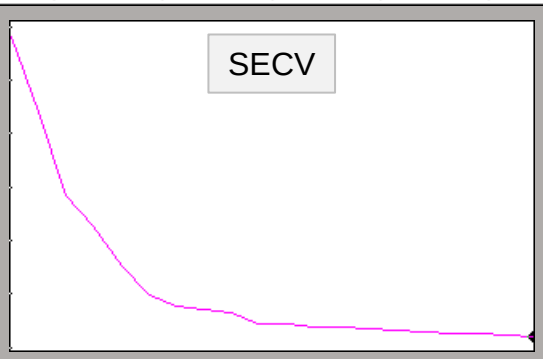
# Jaggedness for Model Complexity

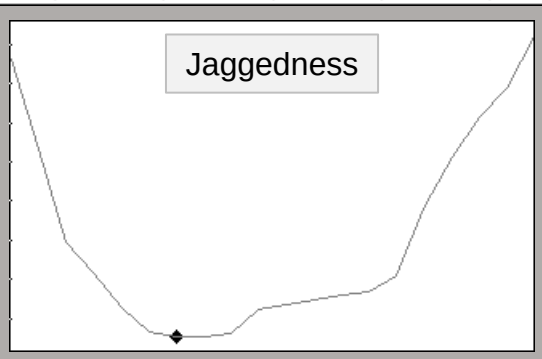
- SECV may not always indicate optimal number of factors for a regression model
- Regression vector usually shows noise structure overlay when overfitting
- Quantify regression vector 'shape' → Jaggedness

# Jaggedness Demonstrated



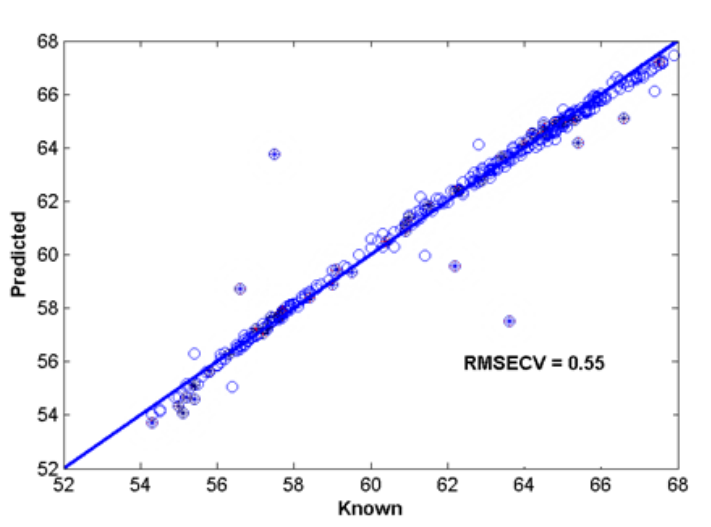



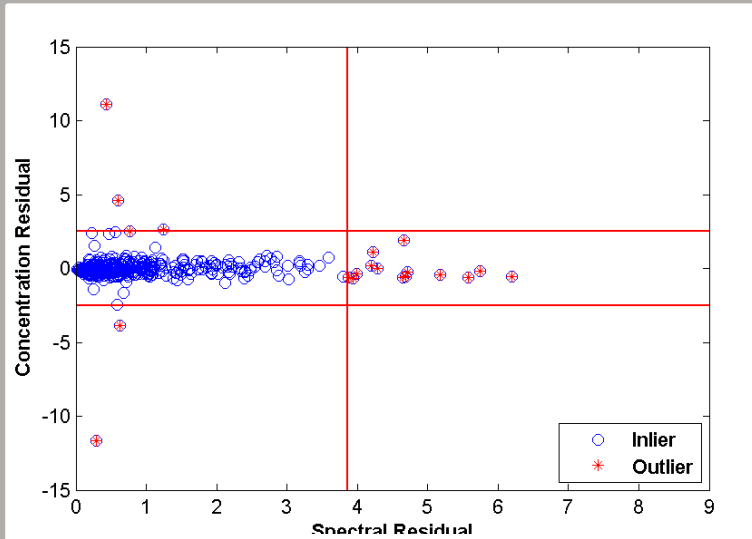

Infometrix

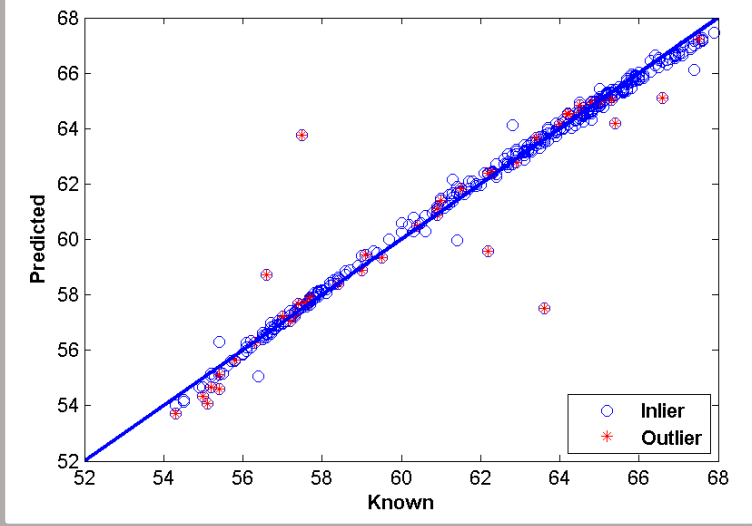
# PLS Model Has Outliers




First PLS Model

Use diagnostics to flag outliers

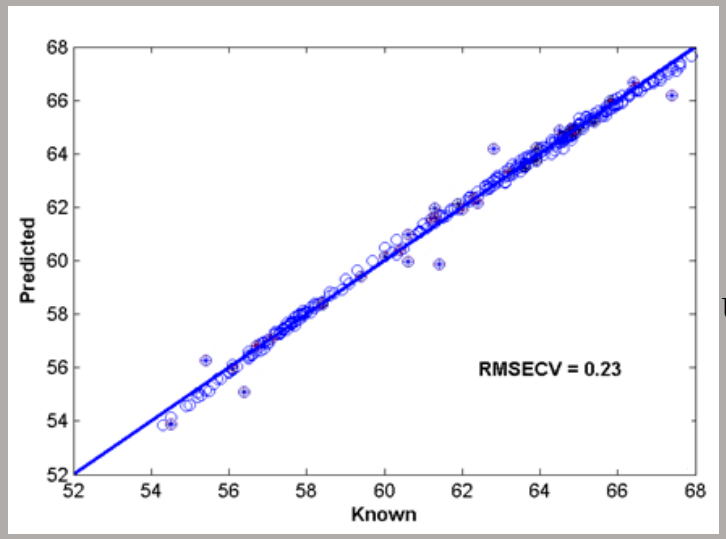



RMSECV = 0.55

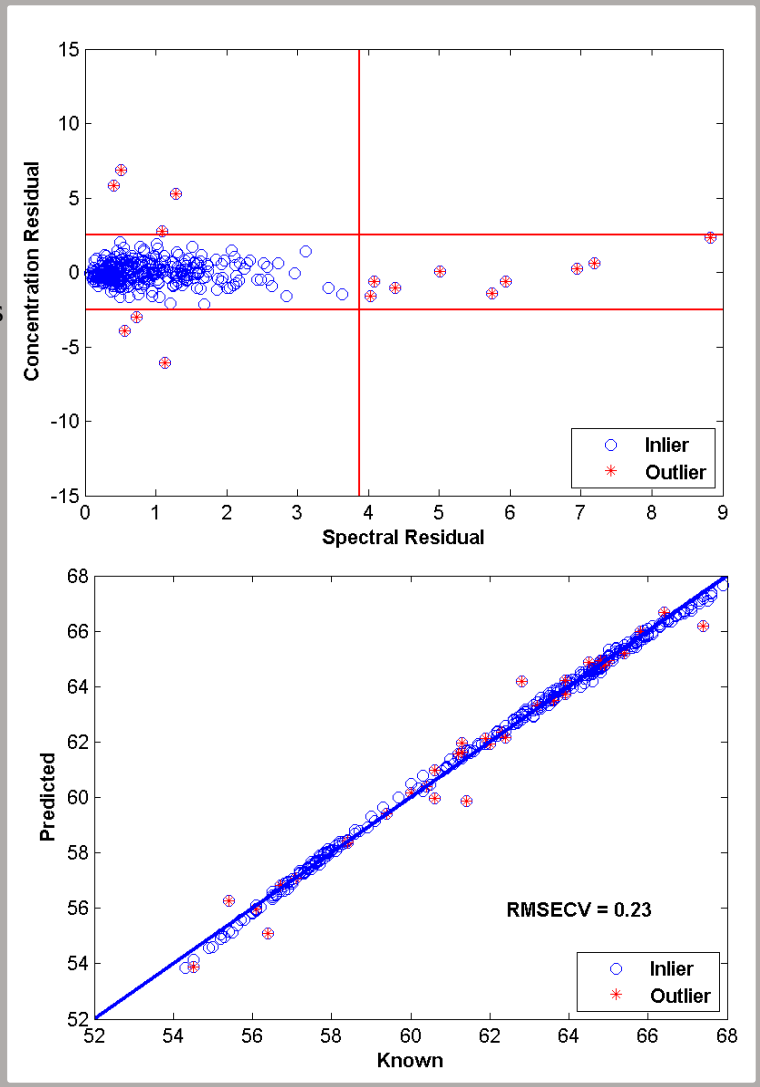
Infometrix

# Second PLS – Additional Outliers




**Second PLS Model**

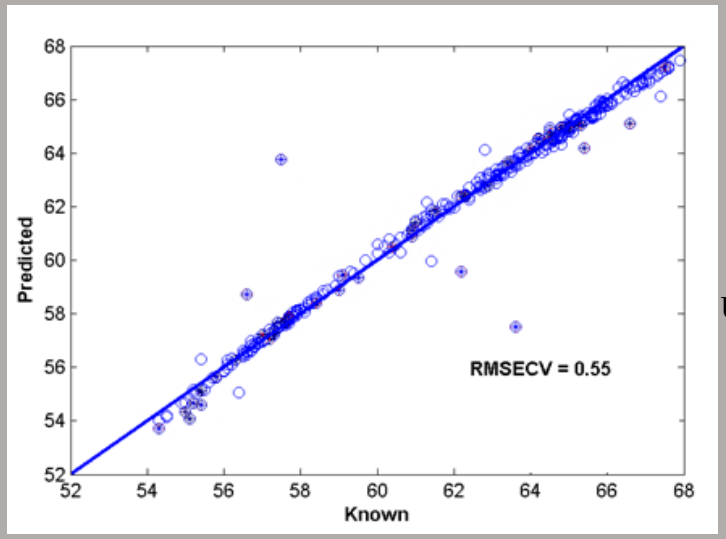
Use diagnostics to flag outliers




RMSECV = 0.23

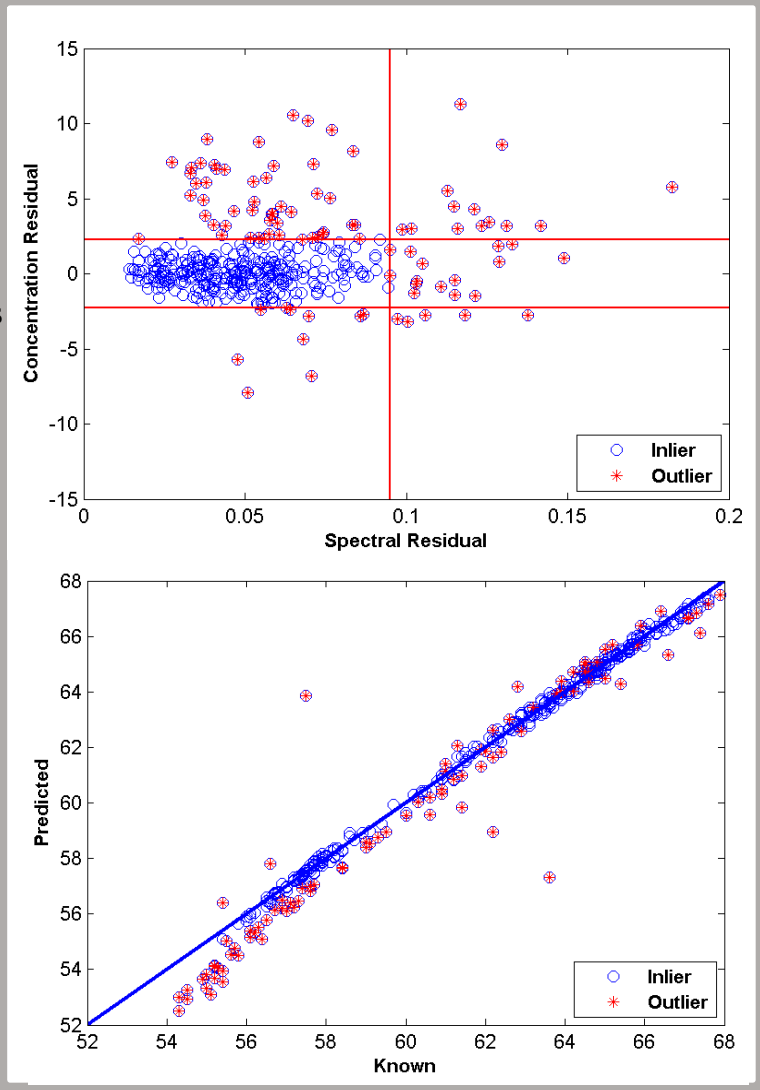
Infometrix

# Robust PLS – Just One Pass



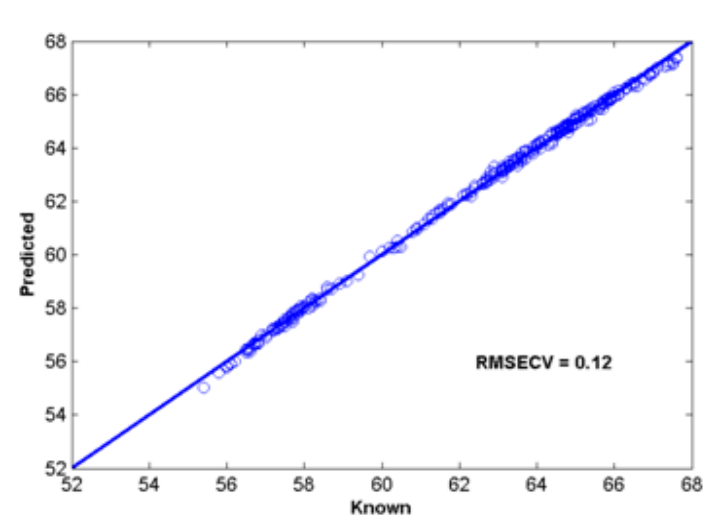

First PLS Model

Use diagnostics to flag outliers



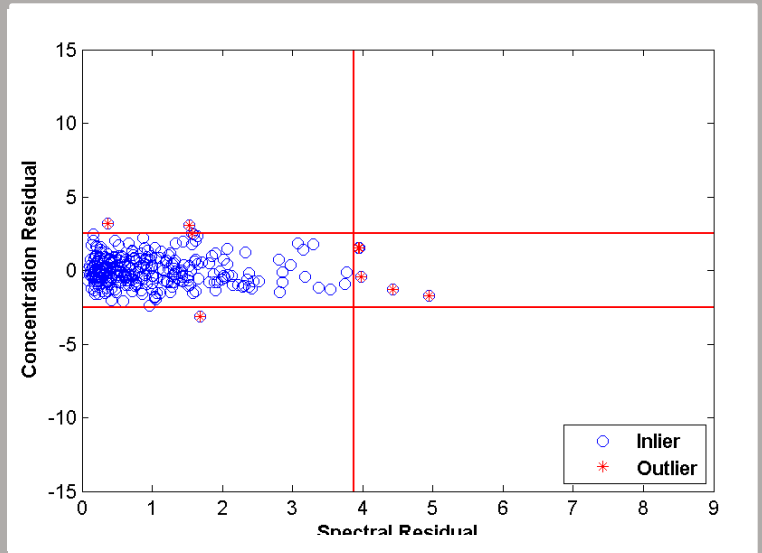

RMSECV = 0.55

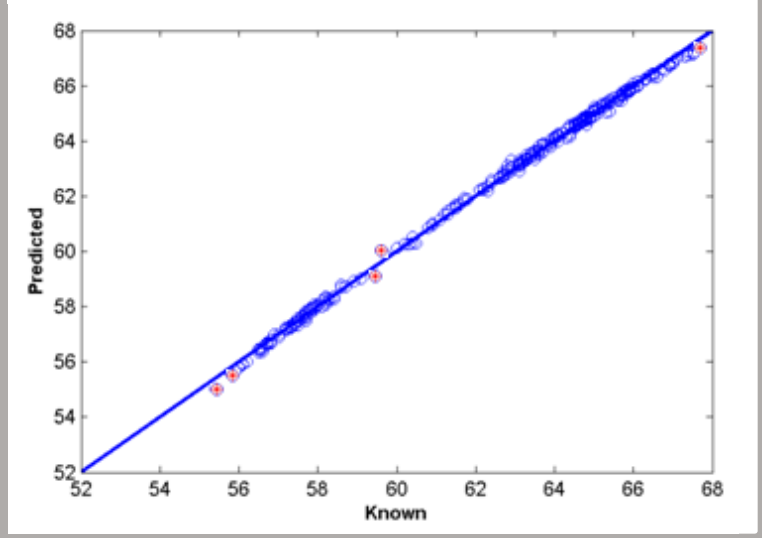
# Robust Model – No More Outliers




**Robust PLS Model**

Use diagnostics to flag outliers




RMSECV = 0.12

# 38 Years of Chemometrics

What problems do we see that create the most problems in building a chemometric system?

1. **Poorly characterized standards**
2. **Groupings in the data**
3. **Overfitting the inferential model**
4. **Non-optimal calibration set**
5. Changing protocol
6. Complex samples
7. Poor instrument stability