Technical Note - Size matters

Figure 1

Samples: Small dataset

$R^2 = 0.968$

Calibration: Small dataset

SEP = 0.504

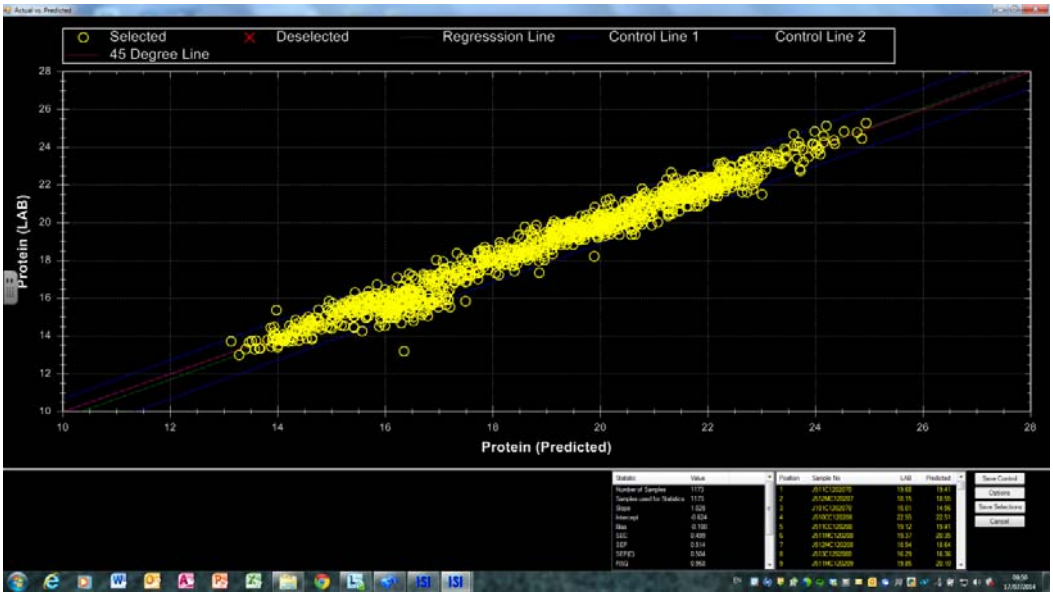


Figure 2

Samples: Small dataset

$R^2 = 0.969$

Calibration: Large dataset (Ingot)

SEP = 0.488

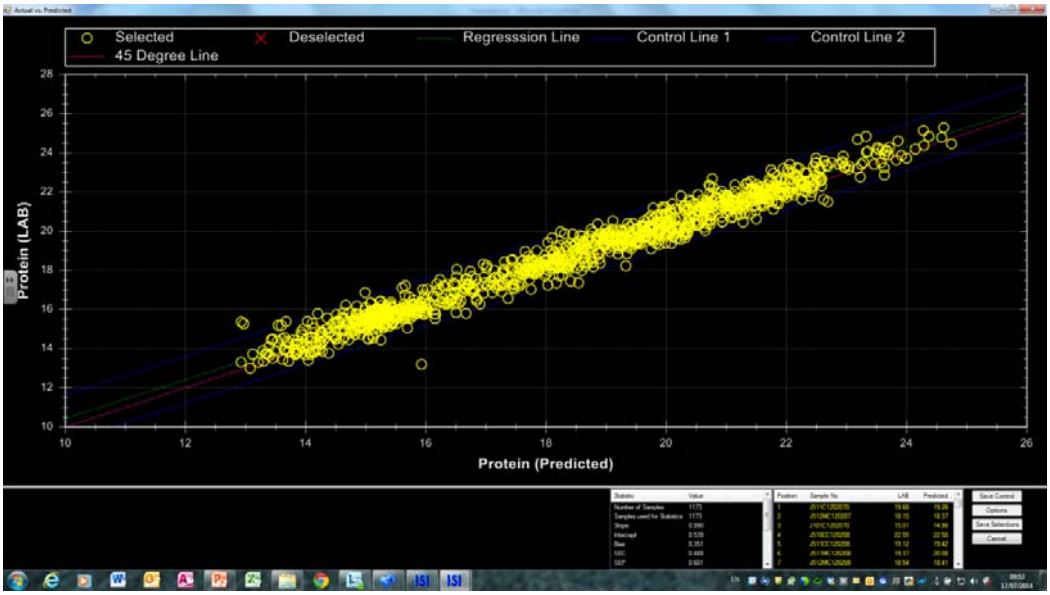


Figure 3

Samples: Large dataset (Ingot) Calibration: Large dataset (Ingot)

R^2 = 0.965 SEP = 0.699

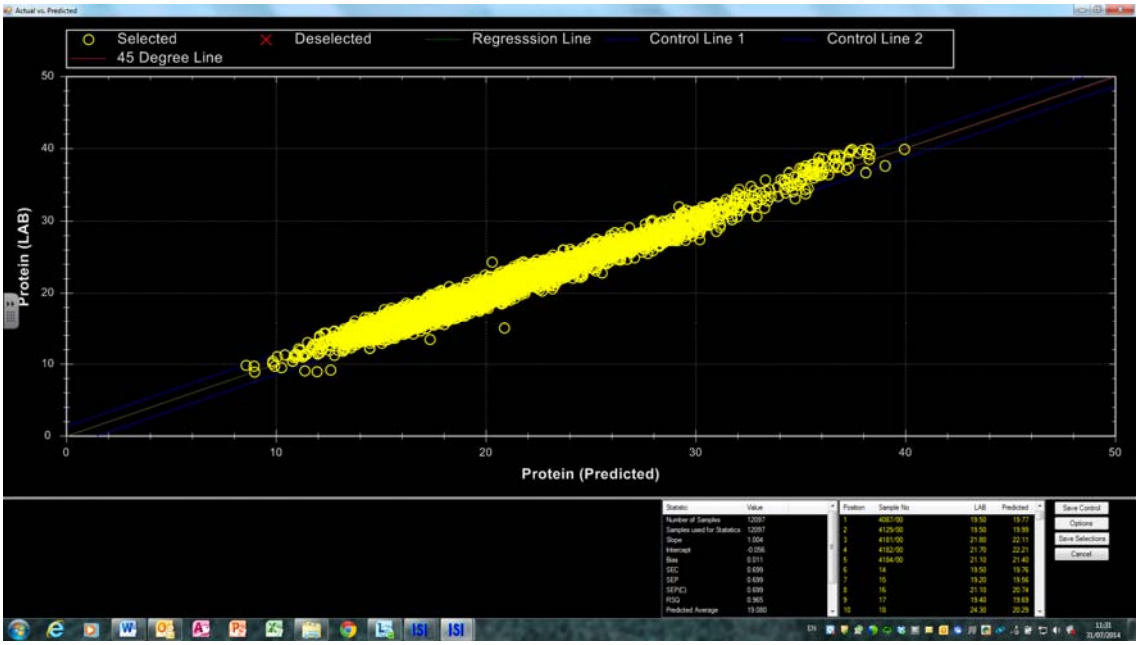
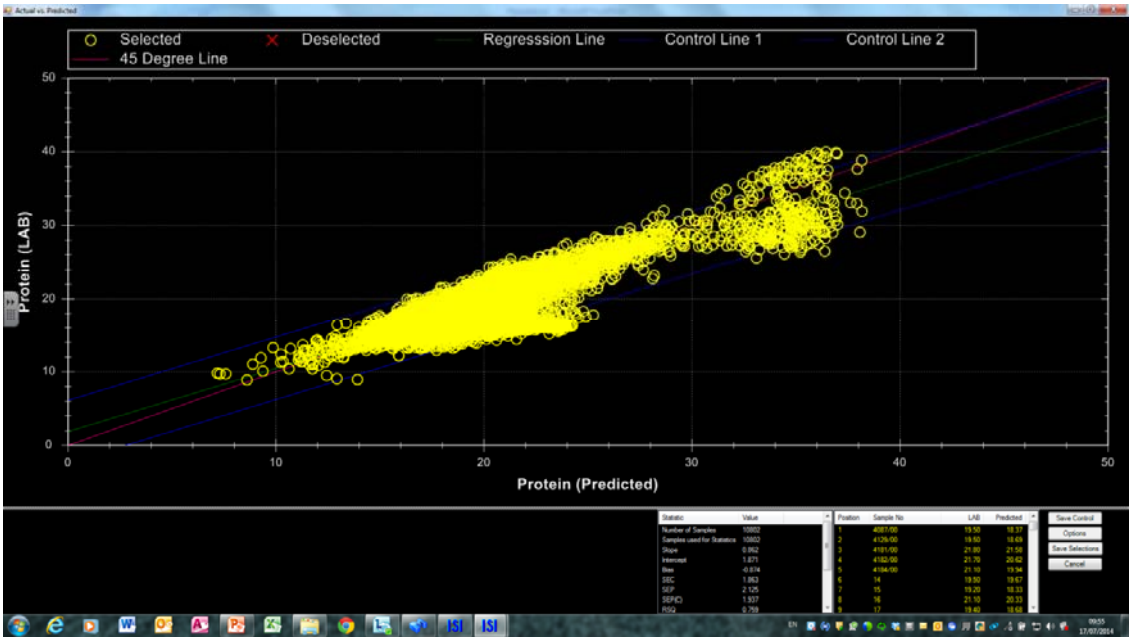


Figure 4

Samples: Large dataset (Ingot) Calibration: Small dataset

R^2 = 0.757 SEP = 1.910



The benefit of a large dataset over a small dataset is the robustness and accuracy of the prediction on unknown samples. The large dataset contains more variation in information on seasonal, geographical and varietal or recipe differences.

Above are graphical representations of the correlation between results for protein obtained from wet chemistry and NIR analysis. Figure 1 shows results from a small set of samples correlated against predictions from an NIR calibration created from this small dataset. The correlation between the two is good. The correlation is also good when the same set of samples is plotted against predictions from an NIR calibration created using a large dataset (Figure 2). Figure 3 shows the correlation between results from a large set of samples and the corresponding NIR predictions based on a calibration created on a large dataset. In Figures 1, 2 and 3, the R^2 value indicates strong correlation. However, plotting a large dataset of samples against predicted results from an NIR calibration developed from a small dataset results in the R^2 value dropping significantly. This highlights the weakness of developing a calibration on a small dataset. A calibration developed from a large dataset will work on both the smaller dataset (Figure 2) even though the samples are not included in the calibration as well as on a large dataset (Figure 3). This highlights the value of the large dataset and how it could offer greater robustness and accuracy than when using smaller, limited datasets.

The correlation (R^2) and Standard Error Predicted (SEP) values demonstrate that a smaller data set and calibration equation will not work on a large data set where the variation of samples has not been included. This should not be unexpected as NIR is a learning technique and if we have not taught the NIR what these samples look like then it will never be able to predict them correctly, hence the poor correlation R^2 value of 0.76 shown in Figure 4.

Calibrations created using large datasets are more robust than those based on smaller datasets. Aunir's Ingot calibrations are based on over 350000 samples gathered from different geographies over 25 years making them both robust and accurate, delivering confidence in results.