Accuracy at Reduced Cost for Near Infrared Measurements of Grain Quality

Nanning Cao, Charles R. Hurburgh

Department of Agricultural and Biosystems Engineering
Iowa State University

February 12, 2013
Challenges for Agriculture Products

- Samples are natural products
- Experimental design is impossible
- New source of data variation
  - New product variety
  - Climate
  - Process conditions
- Difficult or impossible to foresee some data variations
- Reference method analysis could be time consuming and expensive
Challenges for NIR

- Ultra large databases (n>3,000)
- Less samples for calibration
- Small spectral effects
- Database creation and data sharing
- Model maintenance and update
- Uniformity across units, within brand
- Uniformity across brands
- Standardized calibration evaluations
Objectives

- Examine the effect of different calibration sets on the calibration transfer process
- Develop strategies to produce robust calibration models from small data subsets that is transferable to instruments in the network
- Reduce the number of samples need to be scanned on multiple instruments
- Reduce the cost for reference analysis
Materials

- Whole soybean samples (crop year 2001-2009)
  - Protein (N>1,000)
  - Oil (N>1,000)
  - Multiple reference values on four instruments

- Validation set
  From crop year 2011 (Independent, Same range of variation)

- Calibration transfer set
  20 standardization samples
Instruments in the network

- Bruins OmegAnalyzerG (S/N: 106110 & 106118)
- Infratec Grain Analyzer (S/N: 1229_553075 & 1241_0350)
- 850-1048 nm with 2nm interval
- Transmittance
- Room temperature

Bruins OmegAnalyzerG
(Bruins Instrument, Puchheim, Germany)

Infratec Grain Analyzer
(Foss, Puchheim, Denmark)
http://www.foss.dk/industry-solution/products/infratec-1241
Calibration

- Sample selection method
  - D-optimal algorithm

- PLS regression
  - Matlab R2011a
  - PLS_Toolbox v.6.2.1

- Pretreatment methods
  - 2\textsuperscript{nd} derivative (15-point window, 3\textsuperscript{rd} order polynomial)
  - Standard Normal Variate (SNV)
  - Mean center
Procedures

• Sample selection
  – Using D-optimal algorithm to select samples
  – Three calibration sets for each instrument

• Comparison of calibration models
  – Standard Error of Calibration (SEC)
  – Standard Error of Prediction (SEP)
  – Bias
  – Root-Mean-Square Error of Prediction (RMSEP)
  – Relative Predictive Determinant (RPD)
  – RPD = Standard deviation/ SEP

Higher RPD = Better prediction
Samples selected by D-optimal

<table>
<thead>
<tr>
<th>Number of samples in calibration set</th>
<th>RPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.52</td>
</tr>
<tr>
<td>200</td>
<td>0.54</td>
</tr>
<tr>
<td>300</td>
<td>0.56</td>
</tr>
<tr>
<td>400</td>
<td>0.58</td>
</tr>
<tr>
<td>500</td>
<td>0.60</td>
</tr>
<tr>
<td>600</td>
<td>0.62</td>
</tr>
<tr>
<td>700</td>
<td></td>
</tr>
<tr>
<td>800</td>
<td></td>
</tr>
</tbody>
</table>

- **protein**
- **oil**
## Summary statistics of protein (13% moisture basis)

<table>
<thead>
<tr>
<th>Instrument</th>
<th>N</th>
<th>Mean (%)</th>
<th>Range (%)</th>
<th>Standard deviation (%pts)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>OmegG (106110)</strong></td>
<td>Cal: 1100</td>
<td>36.25</td>
<td>24.72-46.89</td>
<td>3.41</td>
</tr>
<tr>
<td></td>
<td>Val: 154</td>
<td>34.92</td>
<td>28.93-43.87</td>
<td>3.48</td>
</tr>
<tr>
<td></td>
<td>Trans: 20</td>
<td>36.38</td>
<td>24.21-45.38</td>
<td>4.65</td>
</tr>
<tr>
<td><strong>OmegG (106118)</strong></td>
<td>Cal: 1100</td>
<td>36.25</td>
<td>24.72-46.89</td>
<td>3.41</td>
</tr>
<tr>
<td></td>
<td>Val: 154</td>
<td>34.92</td>
<td>28.93-43.87</td>
<td>3.48</td>
</tr>
<tr>
<td></td>
<td>Trans: 19</td>
<td>36.38</td>
<td>24.21-45.38</td>
<td>4.78</td>
</tr>
<tr>
<td><strong>Infratec 1241</strong></td>
<td>Cal: 1101</td>
<td>36.27</td>
<td>24.72-46.50</td>
<td>3.42</td>
</tr>
<tr>
<td>(12410350)</td>
<td>Val: 154</td>
<td>34.92</td>
<td>28.93-43.87</td>
<td>3.48</td>
</tr>
<tr>
<td></td>
<td>Trans: 20</td>
<td>36.49</td>
<td>26.40-45.70</td>
<td>4.39</td>
</tr>
<tr>
<td><strong>Infratec 1229</strong></td>
<td>Cal: 1101</td>
<td>36.27</td>
<td>24.72-46.50</td>
<td>3.42</td>
</tr>
<tr>
<td>(553075)</td>
<td>Val: 154</td>
<td>34.92</td>
<td>28.93-43.87</td>
<td>3.48</td>
</tr>
<tr>
<td></td>
<td>Trans: 20</td>
<td>36.49</td>
<td>26.40-45.70</td>
<td>4.39</td>
</tr>
</tbody>
</table>
Spectral differences within brand
Standardization methods

- **No standardization**
  - Directly apply calibration models built on master instrument to secondary unit

- **Robust model**
  - Combine measurements on both master and secondary instruments

- **Slope and bias correction**
  - Build linear regression on calibration transfer set
  - Post correct predicted values for validation set
Results on Protein

- Calibrations built on subsets with less samples obtained reasonable prediction performances.
- Standardization methods improved transferability of calibration models built on subsets.
Conclusions

- Smaller calibration sets with proper selection could obtain same or even better validation performance than using the whole data pool.
- Slope and bias correction is a simple and efficient standardization method for similar instruments.
- Cost on reference analysis could be reduced by selecting representative calibration subsets.
Thank you!

www.iowagrain.org