Leveraging Multiple Linear Regression for Wavelength Selection

Tony Lemos, John H. Kalivas
Department of Chemistry
Idaho State University
921 S. 8th Avenue, STOP 8023 Pocatello, ID 83209, USA
lemoton2@isu.edu, kalijohn@isu.edu

Abstract

In multivariate calibration, wavelengths selection is often used to lower prediction errors of sample properties. As a result, many methods have been created to select wavelengths. Several of the wavelength selection methods involve many tuning parameters that are typically complex or difficult to work with. The purpose of this poster is to show an easy way to select wavelengths while using few simple tuning parameters. The proposed method uses multiple linear regression (MLR) as an indicator to which wavelengths should be used to create a model. From a collection of random MLR models, those models with an acceptable bias/variance balance are evaluated to determine the wavelengths most frequently used. Portions of the most frequently selected wavelengths are chosen as the final MLR selected wavelengths. These MLR selected wavelengths are used to produce a calibration model by the method of partial least squares (PLS). This proposed wavelength selection method is compared to PLS models containing all wavelengths using several near infrared data sets. The PLS models with the selected wavelengths show an improvement in prediction error, suggesting this method as a simple way to select wavelengths.

Objectives

- Create a simple wavelength selection method that lowers prediction errors
- Minimize the number of tuning parameters

Approach

Two multivariate calibration methods are used

- Multiple Linear Regression (MLR)

$$\mathbf{y} = \mathbf{Xb} \rightarrow \hat{\mathbf{b}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$$

  - Models are formed using MLR
  - Wavelengths of filtered models are collected

- Partial Least Squares (PLS)

$$\mathbf{y} = \mathbf{Xb} \rightarrow \hat{\mathbf{b}} = \mathbf{X}^{+}\mathbf{y}$$

  - PLS models are formed using selected wavelengths

Measures of Model Quality

$$\text{RMSEC} = \sqrt{\sum_{j=1}^{n}(\hat{y}_j - y_j)^2 / n} \qquad \|\hat{\mathbf{b}}\| = \sqrt{\sum_{k=1}^{w} b_k^2}$$

$$\text{RMSEP} = \sqrt{\sum_{i=1}^{m}(\hat{y}_i - y_i)^2 / m} \qquad R^2$$

Experimental Design

- MLR models are plotted with bias/variance measures
- A percentage of MLR model with low $\|\mathbf{b}\|$ and RMSEC are selected

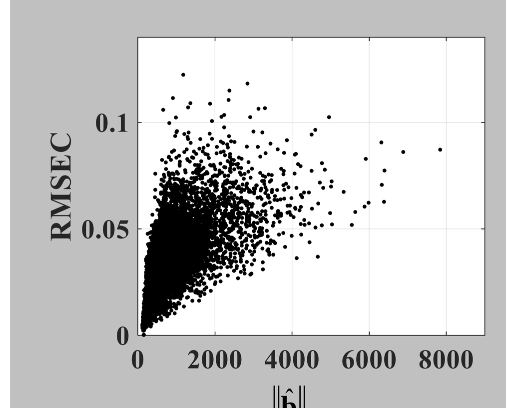

Figure 1 – 10,000 MLR models

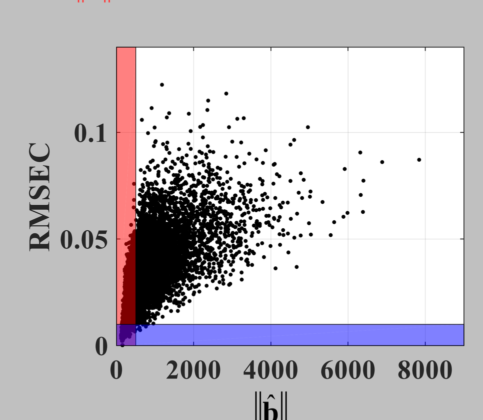

Figure 2 – 30% of the lowest $\|\mathbf{b}\|$ and RMSEC

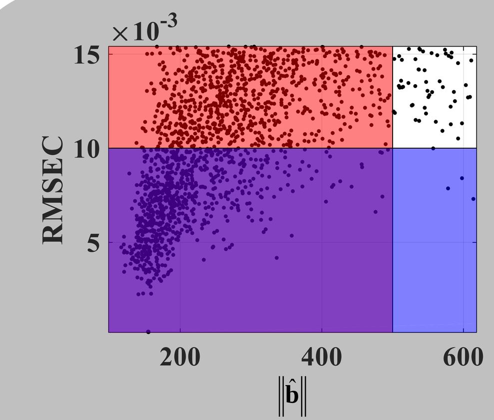

Figure 3 – Intersected models

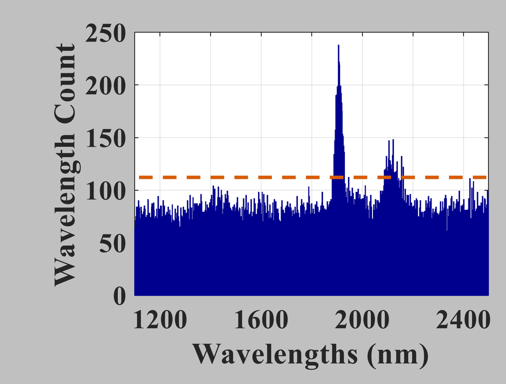

Figure 4 – Wavelengths from intersected models

- Wavelengths of intersected models are retained
  - More intersected models are created and retained to distinguish better wavelengths
- Wavelengths are selected for the final collection of wavelengths
  - Number of wavelengths is based on the rank of calibration set
- PLS models are created from selected wavelengths
  - Compared against all wavelength PLS
- Measures of model quality
  - RMSEP
  - R^2_{pred}

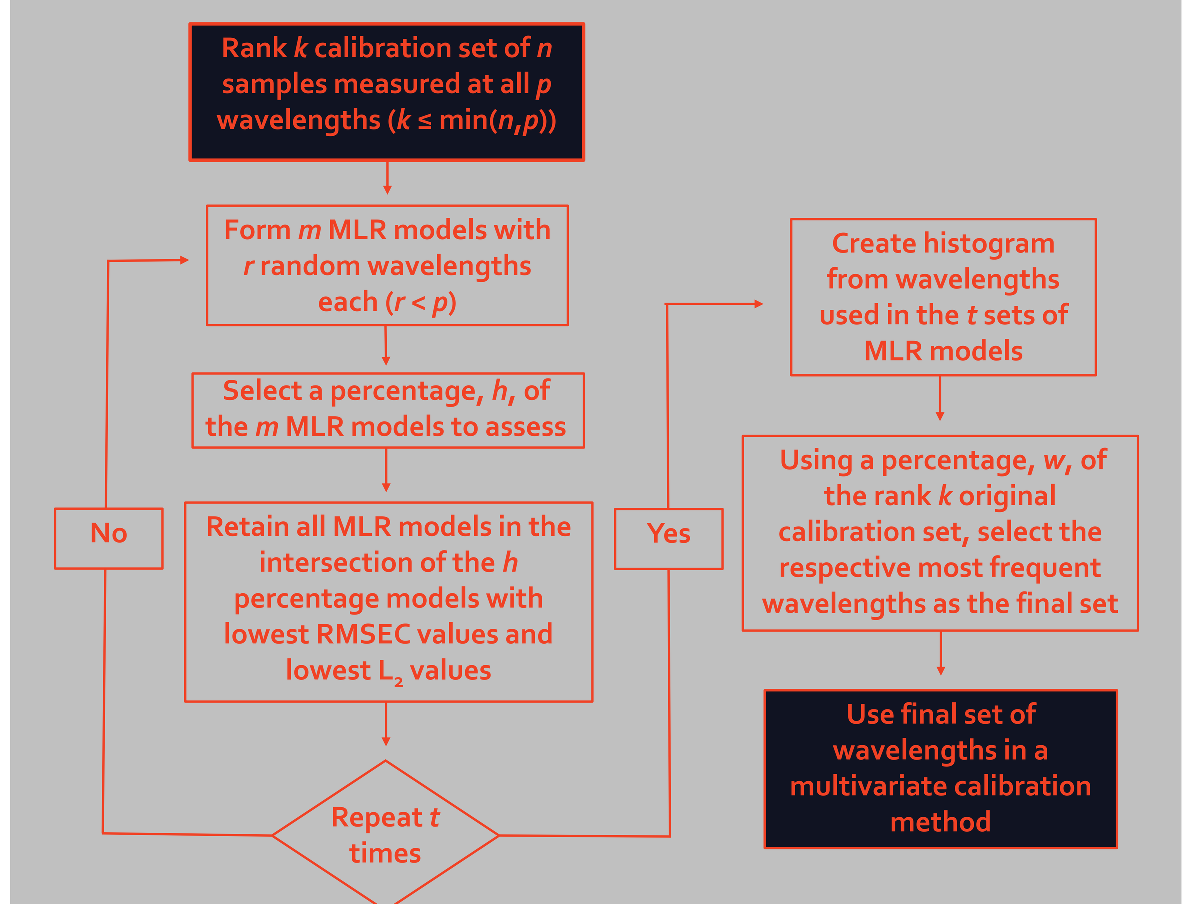

Figure 5 – Flow chart of the MLR wavelength selection method

NIR Data Sets

- Corn – 80 samples measured at 700 wavelengths on 1 instrument (m5) for the prediction properties moisture and oil
- Sugar – 125 samples measured on 700 wavelengths for the prediction property sucrose
- Gasoline – 55 samples measured at 401 wavelengths for the prediction property octane number

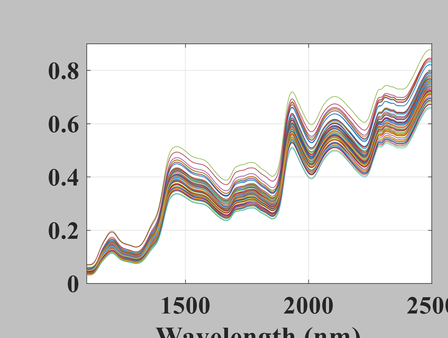

Figure 6 – Spectra for corn

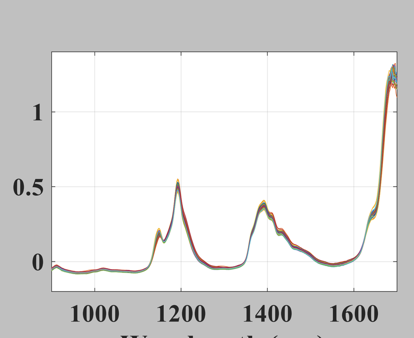

Figure 8 – Spectra for gasoline

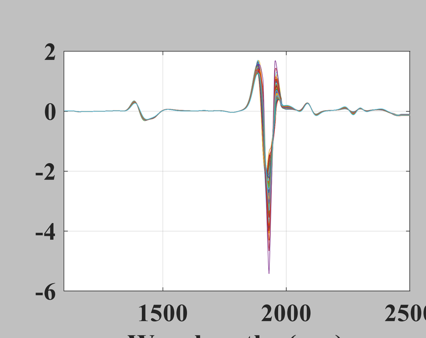

Figure 7 – Spectra for sugar

Results

Tuning Parameters

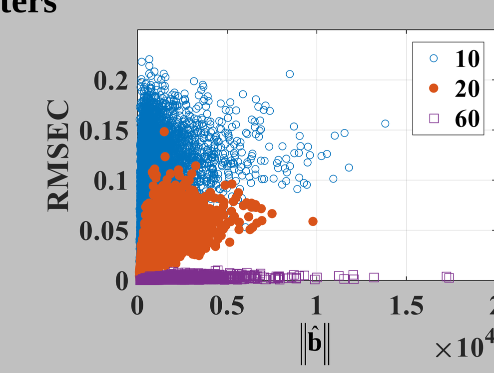

Figure 9 – Effects of changing r using 10,000 models

Number of wavelengths for each MLR model (r)
- Wavelengths will effect which models are intersected
  - More wavelengths, lower RMSEC
  - Less wavelengths, lower L_2 norm
  - Need wavelengths in between
- For this study, r is set to 20 wavelengths

Number of MLR models (m)
- Need to have enough to represent the range of MLR models
  - Small amounts do not show which wavelength to choose
  - More models that are formed, the more likely the selected wavelength is useful
- m is set to 10,000 models

Percentage of MLR models with low RMSEC and L_2 norm (h)
- The intersection allows to inspect models that are neither over-fitter or under-fitted
  - A large h will allow poor models in the intersection
  - A Small h will not show which wavelengths are useful
- After using m = 10,000, h is set to 30%

The number of intersection sets (t)
- More than one intersection is needed to create a good histogram
  - One intersection lets more random wavelength to be chose
  - More intersection allows more dominant wavelengths to appear more obvious
- t is set to 50
  - The histograms converges at t = 50

Percentage of selected wavelengths (w)
- w is based on the percentage of the rank of the calibration set
  - A higher percentage, the more the PLS model appear like all wavelength PLS
  - A lower percentage, worse the PLS model performs
- w = 80% is chosen for this study

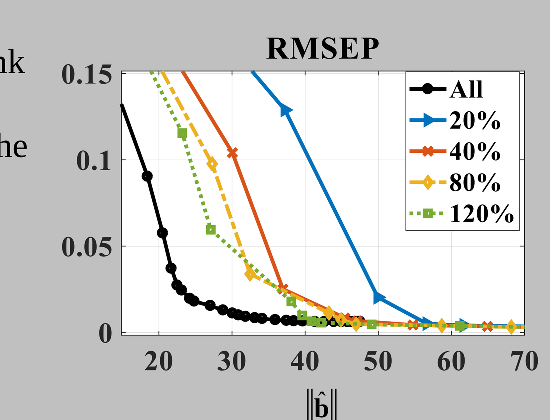

Figure 10 – Effects of changing the value for w

Corn - m5 Moisture

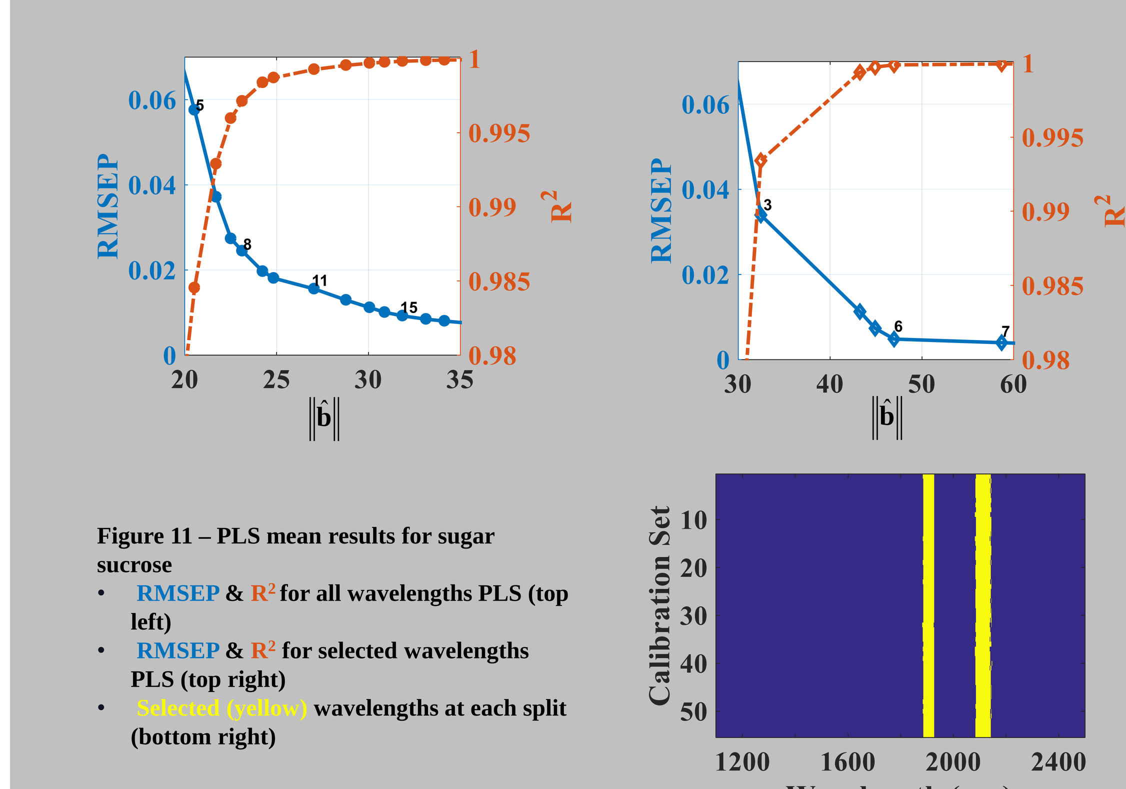
Figure 11 – PLS mean results for sugar sucrose
- RMSEP & R^2 for all wavelengths PLS (top left)
- RMSEP & R^2 for selected wavelengths PLS (top right)
- Selected (yellow) wavelengths at each split (bottom right)

Corn - m5 Oil

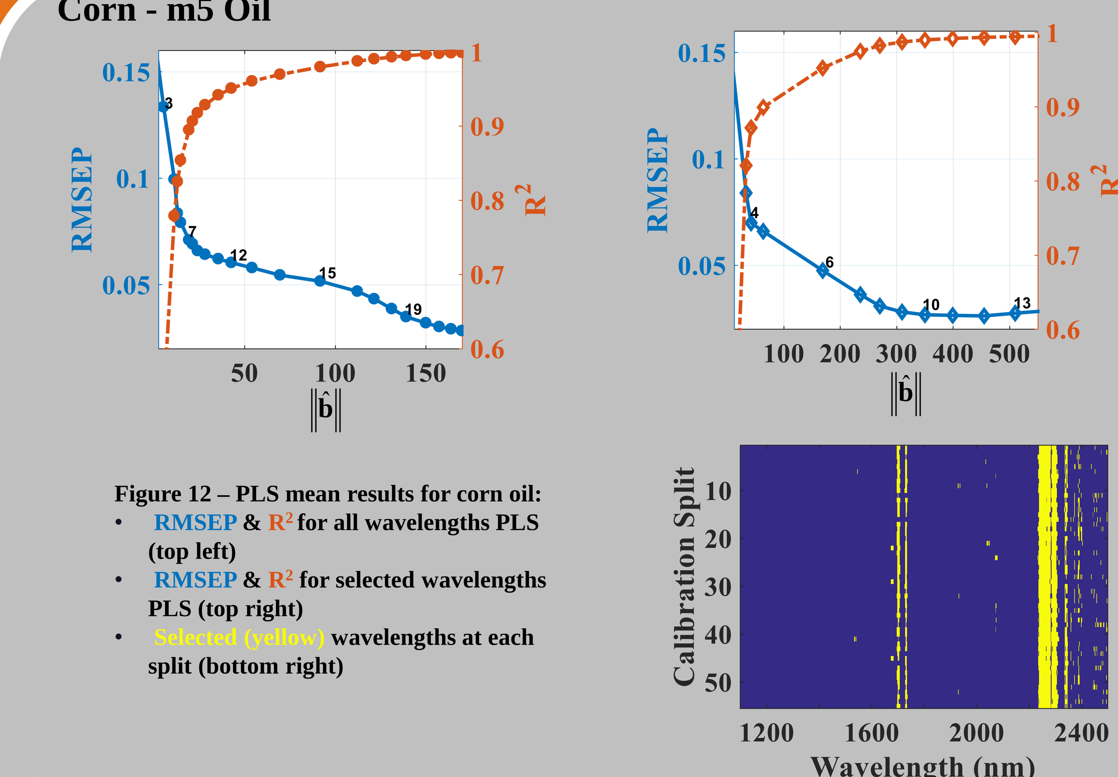
Figure 12 – PLS mean results for corn oil:
- RMSEP & R^2 for all wavelengths PLS (top left)
- RMSEP & R^2 for selected wavelengths PLS (top right)
- Selected (yellow) wavelengths at each split (bottom right)

Sugar – Sucrose

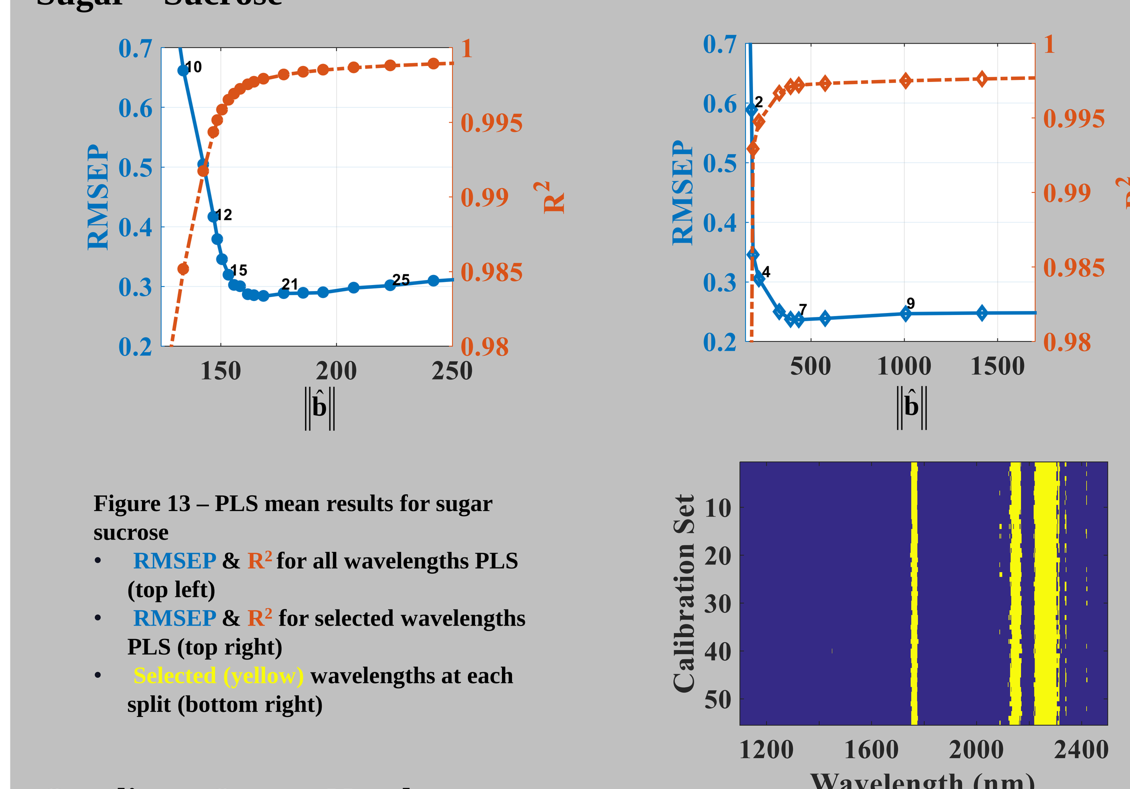
Figure 13 – PLS mean results for sugar sucrose
- RMSEP & R^2 for all wavelengths PLS (top left)
- RMSEP & R^2 for selected wavelengths PLS (top right)
- Selected (yellow) wavelengths at each split (bottom right)

Gasoline – Octane Number

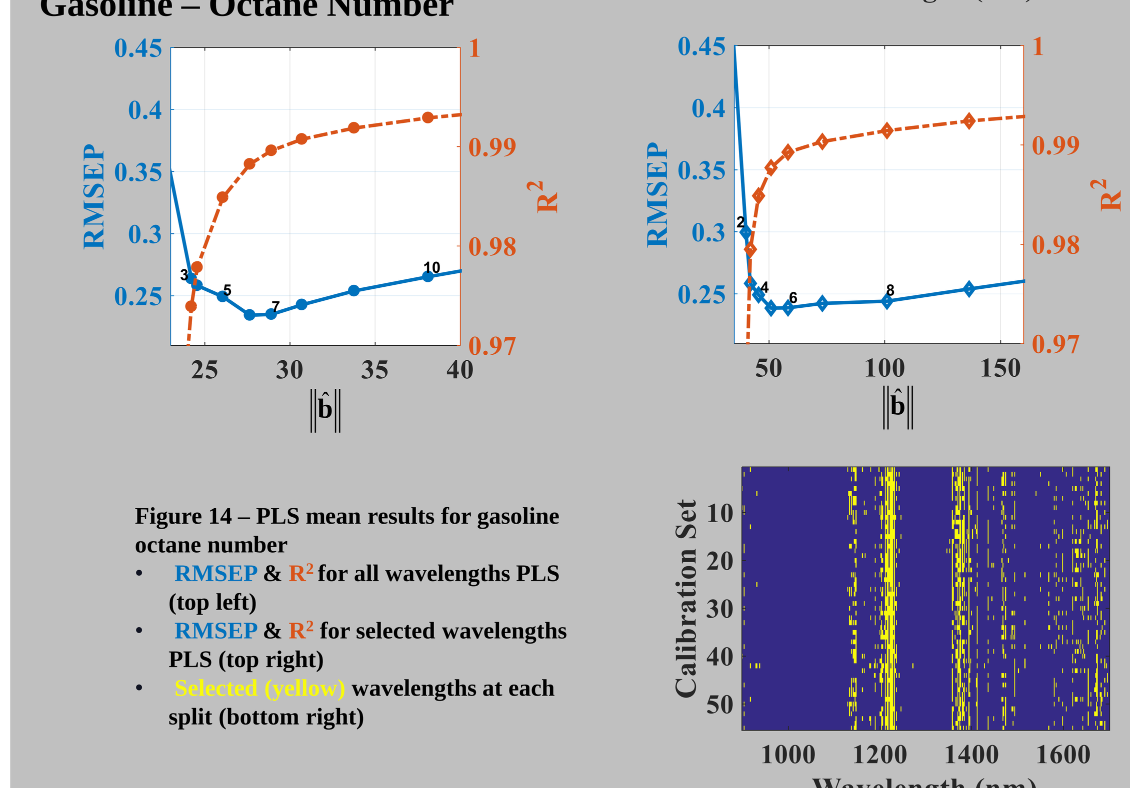
Figure 14 – PLS mean results for gasoline octane number
- RMSEP & R^2 for all wavelengths PLS (top left)
- RMSEP & R^2 for selected wavelengths PLS (top right)
- Selected (yellow) wavelengths at each split (bottom right)

Conclusions

- MLR wavelength selection helps from improved calibration models
  - Generally does better than all wavelength PLS
  - Most datasets choses banded wavelengths
    - Gasoline did not
  - Larger L_2 norm
- Tuning parameters
  - Goal was to limit the number of parameters
  - Out of the five, only two can be changed
    - Gasoline needs adjustment to improve
- The proposed method is successful and can be used for wavelength selection

Tuning Parameters	
r	Adjust to get 'cone' shape
m	10,000 models
h	30%
t	50 intersections
w	Adjust to get improved performance

Acknowledgements: Work supported by the National Science Foundation under grant No. CHE-1506417 (co-funded by CDS and E Programs) and is gratefully acknowledged by the authors.