Normalization and Baseline Removal Effects on Univariate and Multivariate Hydrogen Prediction Accuracy using Laser-Induced Breakdown Spectroscopy

Caroline R. Ytsma[1], M. Darby Dyar[2,3], Kate H. Lepore[2], Carlie M. Wagoner[2], and Avery E. Hanlon[2]

[1]Department of Chemistry, Smith College, 1 Chapin Way, Northampton, MA 01063 (ytsmacr@gmail.com), [2]Department of Astronomy, Mount Holyoke College, 50 College Street, South Hadley, MA 01075, [3]Planetay Science Institute, 1700 East Fort Lowell, Suite 106 * Tucson, AZ 85719

Overview

Laser-induced breakdown spectroscopy (LIBS) is used for remote quantification of rock and soil compositions. Hydrogen is particularly important because hydrous minerals indicate the presence of water on Mars at the time the sample was formed.

LIBS is useful analyzing H because strong H emission lines occur in the visible region. Unfortunately few geologic standards with known H are available for use in calibrating its concentration. Even for geological standards, reported chemical analyses are given as LOI (loss on ignition) results that lump all volatiles (e.g., H_2O, CO_2 etc.) together, rendering them useless for H calibrations.

Our laboratory collections include many H standards characterized by the uranium extraction technique for previous studies (e.g. [1-3]) of structural/internal H**.** Materials cover a range of minerals, rocks, and glasses with varying concentrations of H, as seen below in **Figure 1**, and are used here to develop a robust calibration for H in LIBS data.

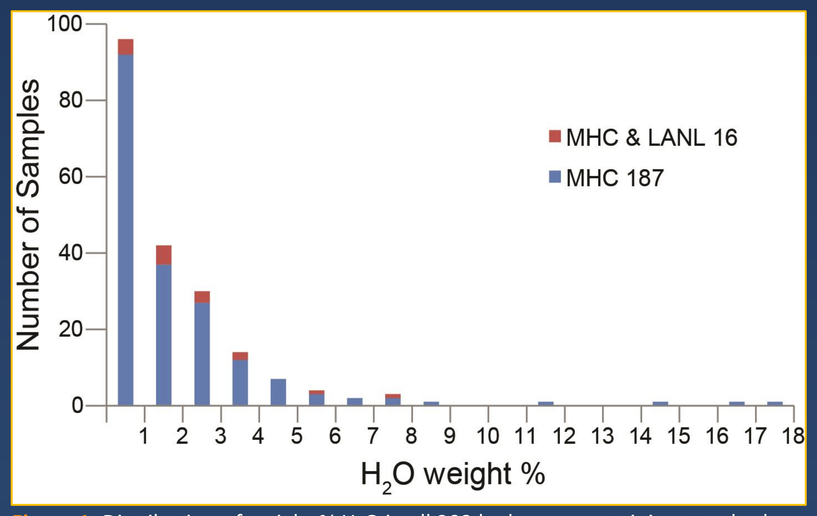
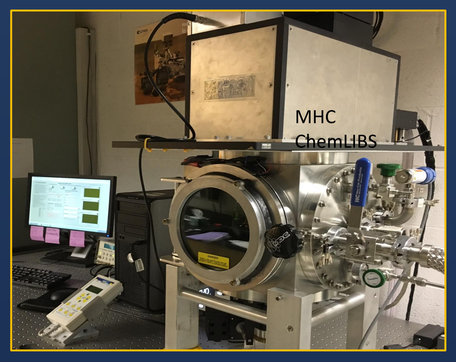


Figure 1. Distribution of weight % H_2O in all 203 hydrogen-containing standards.

Data collected at Mount Holyoke College (MHC) are compared with laboratory data on some of the same standards collected on the ChemCam engineering model at Los Alamos National Laboratory (LANL). This study thus compares three sample sets:

1. **16** standards run at **LANL** for which H analyses were available
2. the same **16** samples run at **MHC** at 5% laser power
3. all **203 MHC** standards also run at 5% laser power

Methods



Powders of all 203 geologic standards were pressed into discs for analysis with the MHC LIBS, which has more limited sensitivity because its spectrometers utilize 1D CCD detectors rather than the 2D detectors used on the ChemCam flight instrument. However, the current MHC ChemLIBS instrument is typical of (or better than) most commercially available LIBS units and has the capability of adjustable laser power, allowing spectral power to be matched to Mars in terms of plasma temperature [e.g., 6]. Comparisons between these two instruments allow some of the variables affecting quantitative analysis of hydrogen and other elements to be better understood. **Figure 2** illustrates variations in detectors, as both peaks are the result of the H-alpha Balmer emission line (656.6 nm).

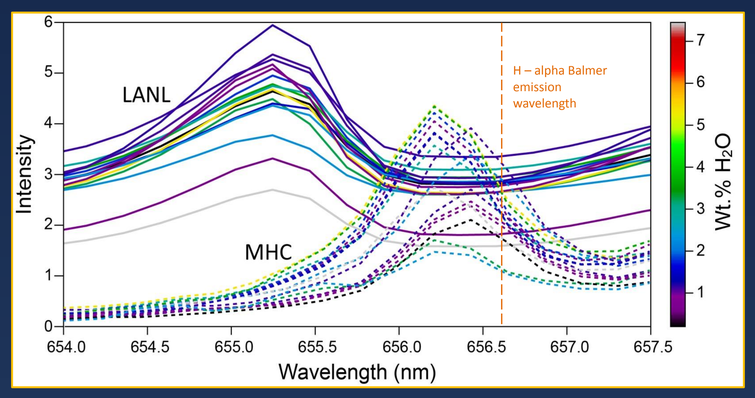


Figure 2. Spectra of the same 16 hydrogen containing samples for LANL and MHC, displaying intensity × $1e^3$ and $1e^9$, respectively.

Analyses

Univariate analyses were undertaken by regressing the intensity at the centroid of this H-alpha Balmer emission line in the samples, as used by [4], against known H concentration expressed as wt% H_2O (**Table 1**). This was repeated for nine variations to modify spectral intensity by normalizing to:

1. Norm3 - total intensity over energy range of each of three spectrometers *individually*
2. Cumulative - total intensity over energy range of each of three spectrometers *collectively*
3. Max - intensity of highest peak in entire spectral range
4. L1 - the sum of absolute values
5. L2 - the sum of squared values
6. H peak at 656.6 nm
7. C peak at 247.8 nm [5]
8. C peak at 657.7 nm
9. O peak at 777.5 nm.

Table 1. R^2 values of normalized 656.6 nm intensity vs. H_2O wt%. No baseline removal added.

Method	LANL 16	MHC 16	MHC 203
Cumulative	0.22	0.08	0.04
Norm3	0.00	0.10	0.01
Max	0.04	0.08	0.25
L1	0.34	0.09	0.08
L2	0.28	0.08	0.14
H 656.6 nm	0.04	0.00	0.25
C 657.7 nm	0.04	0.03	0.32
C 247.8 nm	no peak	0.23	0.02
O 777.5 nm	no peak	0.73	0.34

Multivariate analyses used a range of wavelengths that extend beyond the intense 656 nm peak and captured possible predictive information from other features. Prediction accuracy is expressed as root mean square error (RMSE) in units of wt.% H_2O (**Figure 3**). These regression methods consisted of partial least squares (PLS) and least absolute shrinkage and selection operator [e.g., 8] (Lasso).

Table 2. RMSE values on unnormalized, non-baseline removed data.

Range	LANL Lasso	LANL PLS	MHC 16 Lasso	MHC 16 PLS	MHC 203 Lasso	MHC 203 PLS
243-253 nm	3.97E-04	1.102	1.13E-10	1.419	1.394	2.285
650-675 nm	1.36E-02	1.000	2.50E-10	0.1005	1.452	2.419
773-783 nm	1.25	1.619	4.26E-09	0.7799	2.738	2.405
All 3	6.13E-09	0.3426	4.37E-11	1.419	1.506	2.278
200-300 nm	2.06E-02	0.5988	8.54E-11	0.8096	0.1336	2.285
600-800 nm	3.93E-04	1.243	4.66E-09	0.9724	0.2038	1.714
Both	3.81E-03	1.115	1.68E-10	0.8086	0.2509	2.208

Results

- *Univariate* analyses of LANL data were only slightly more accurate than MHC's on average (**Table 1**), implying that detector sensitivity is not a limiting factor.
- Normalization methods greatly improve prediction accuracy for raw data, but the R^2 values on the *univariate* regressions are much too low to use for definitive H quantification.
- *Univariate* methods are likely poor performers because of chemical matrix effects that change the 656.6 nm peak's intensity. There is only a weak correlation between intensity and H in both LANL and MHC spectra (**Figure 2**).
- Baseline removal has negligible effects on prediction accuracy and is therefore not included in any comparisons on this poster.
- Use of *multivariate* comparisons, H quantification accuracy increases dramatically (note low RMSE values in **Figure 3**).

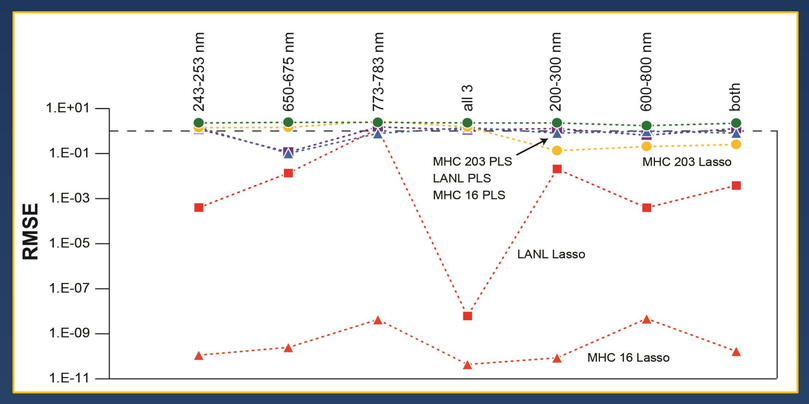


Figure 3. RMSE values from **Table 2.** Note logarithmic y-axis, with RMSE values over 1 showing relatively poor prediction results and smaller values indicating more accurate analyses.

Figure 4 shows the distribution of channels chosen by the multivariate methods to predict H. Raw data were used for both plots. Wavelengths chosen by the lasso methods were indexed against the NIST atomic emission database. Interestingly, most of the lines chosen for H predictions from the larger MHC data set used Fe II emission lines.

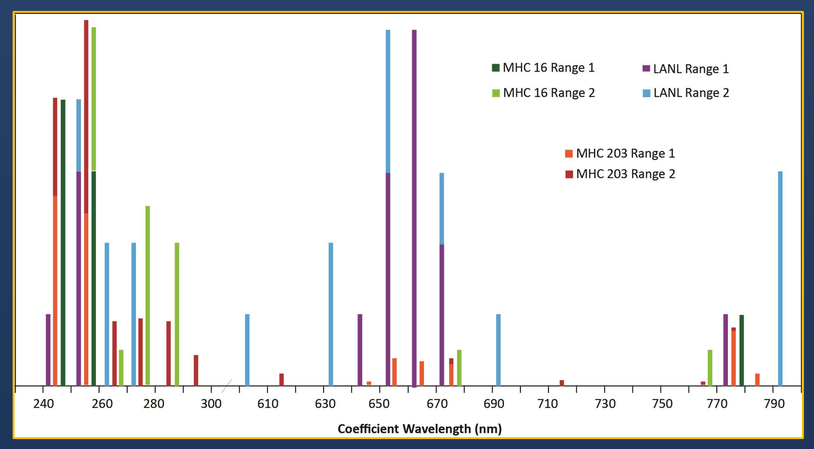


Figure 4. Histogram of occurrences of peaks in wavelengths ranges used by different methods shown in Table 2. Range 1 = 243-253, 650-675, and 773.783 nm. Range 2 covers 200-300 and 600-800 nm.

Summary

Multivariate lasso analysis is a dependable method of H prediction for the range of LIBS instrument sensitivities tested, with best prediction accuracy of ±0.13 wt% H_2O from all 203 MHC standards. Univariate analysis is far less accurate and slightly more dependent on specific instrument characteristics.

Acknowledgments: Research supported by NASA grants NNX14AG56G and NNX15AC82G and NSF grants IIS-1564083, CHE-1306133 and CHE-1307179.

References: [1] Dyar M. D. et al. (1991) *Geology*, *19*, 1029-1032. [2] Dyar M. D. et al. (1993) *Geochim. Cosmochim. Acta*, *57*, 2913-2918. [3] Woods S. et al. (2000) *Amer. Mineral.*, *85*, 480-487. [4] Thomas N.H. et al. (2015) *LPS XLVI*, Abstract #2119. [5] Schröder, S. et al. (2015) *Icarus, 249*, 43-61. [6] Tokar R. et al. (2015) *LPS XLVI*, Abstract #1369. [7] Dyar M. D. et al. (2016) *Amer. Mineral., 101*, 744-747.