MULTIVARIATE LIBS ANALYSIS OF GEOLOGIC MATERIALS. J. M. Tucker[1], M. D. Dyar[1], M. W. Schaefer[2], S. M. Clegg[3], and R. C. Wiens[3], [1]Department of Astronomy, Mount Holyoke College, 50 College St., South Hadley, MA 01075, jtucker@mtholyoke.edu; mdyar@mtholyoke.edu. [2]Department of Geology and Geophysics, E235 Howe-Russel, Louisiana State University, Baton Rouge, LA 70803, mws@lsu.edu. [3]Los Alamos National Laboratory, P.O. Box 1663, Los Alamos, NM 87545, sclegg@lanl.gov, rwiens@lanl.gov.

Introduction: The ChemCam instrument aboard the Mars Science Laboratory rover *Curiosity* will use remote LIBS (laser-induced breakdown spectroscopy) to assess the chemistry of rock and soil samples on Mars. LIBS spectra of geological samples spanning from the near-UV to near-IR typically consist of dozens to hundreds of atomic emission lines. Extensive testing and experimentation in LIBS laboratories around the world is underway to determine the most effective means of extracting quantitative information about the sample chemistry from the LIBS spectra.

Because of the dependence of peak intensity on elemental abundance, regression analyses are used to model spectral response and predict chemical compositions of unknown samples. Multivariate analyses are employed to exploit the wealth of compositional data available in LIBS spectra and to compensate for matrix effects. Partial least squares regression (PLS) is well suited to situations where there are many predictor variables (**X**; spectra) and few response variables (**Y**; elemental abundances). In this study, PLS regression is employed with varying spectral treatments and data manipulations to determine the most successful methodology for predicting unknown sample compositions.

Experiment: 100 igneous rocks with a wide range of compositions were chosen for this study. All had been analyzed for major and minor elements in Michael Rhodes' XRF lab at the Univ. of Massachusetts using standard operating procedures [1]. Samples were powdered to mitigate inhomogeneity and equalize grain size and porosity, and pressed into pellets. LIBS analyses were performed at Los Alamos National Laboratory under conditions configured to mimic those of ChemCam. Samples were placed at a 9 m standoff distance in a sealed chamber evacuated and filled with 7 Torr CO_2. A Nd:YAG laser operating at 1064 nm is set to 17 mJ per shot. Spectra were recorded with Ocean Optics spectrometers with nearly continuous coverage from 220 nm to 930 nm. We probed 5 spots on each sample, each with 50 laser shots, for a total of 250 shots per sample. Spectra were normalized by dividing each wavelength channel by the total intensity in the spectrum, which compensates for variability in laser power. The five normalized spectra were then averaged to produce a single spectrum per sample.

PLS regressions and predictions were performed using The Unscrambler, a commercially-available software product. Spectra of the 100 samples were split randomly into a training set for calibration and a test set for prediction, each with 50 samples. Except when noted, the same training and test sets were used in every comparative analysis to facilitate direct comparisons. Major elements considered in this study are Si, Al, Ti, Fe, Mg, Mn, Ca, Na, P, and K.

Results: Several of the protocols for processing LIBS spectra were evaluated. Results were quantified by plotting predicted values of the elemental concentrations in the test set *vs*. their known values. A value of R^2 (coefficient of determination) was calculated for each plot and the values of R^2 compared to assess the relative predictive power of each regression model.

Atomic fraction vs. wt.% oxide. Regardless of how they are performed, geochemical analyses are reported in elemental wt.% oxides. However, spectral peak intensities should be directly correlated with atomic fractions, so these should produce better predictions. Our results show that atomic fraction predictions are no better than wt.% oxide predictions, likely because of the issues of error propagation. Calculating the atomic fraction of a single element incorporates the error from all wt.% oxide values, and is especially affected by incomplete analyses (e.g., lacking H_2O or CO_2). Thus there is more inherent error in atomic fraction values than wt.% oxides in the training set. Further, when using atomic fractions in the training set, predictions in the test set are returned as atomic fraction; converting them back to wt.% oxide incorporates all the errors in the predictions of all elements.

PLS1 vs. PLS2. Because it utilizes all available predictors and eliminates multicollinearity, PLS analysis can is routinely more successful than standard univariate analysis at predicting elemental concentrations. PLS1 regresses a single response variable (element) against the predictor variables (spectra). PLS2 regresses multiple responses against the predictors and explains the variance in both **X** and **Y**, taking advantage of the natural well-known correlations between elements in igneous rocks. Our results show consistently that PLS2 is more successful at predicting elemental concentrations than PLS1 for this sample set.

*Scaled **Y** variables.* Because a PLS2 regression attempts to minimize variance in the **Y** variables simultaneously, it will predict elements with larger ranges better and ignore to a greater extent those whose

ranges are relatively constant. To eliminate this effect, various methods of rescaling the elemental concentrations were tested. Simple statistical normalization was determined to be most effective: subtracting the mean and dividing by the standard deviation of all samples in the training set for each element individually. Prediction of elements with small compositional ranges in these samples, such as Ti, Mn, Na, and P, show significant improvements with scaled **Y**. It must be noted that rescaling is only appropriate when the original elemental distribution is approximately Gaussian.

Redistributed training set: Any kind of regression works best when the training set fully encompasses the range of variation encountered in the test set. Hence, a new training set of 50 samples was selected that included samples exhibiting the highest and lowest values of all 10 major elements, and regularly spaced samples in between. Predictions using this training set are not significantly better than those using the 50 random samples in previous examples. This validates the assumption that the original training set encompassed the variation of the test set fairly well.

Normalization and background subtraction. In another set of analyses, an additional procedure [2] for further spectral normalization was employed. After spectra were normalized to total intensity and combined, they were scaled back up by a factor of the sum of the sums of all original spectral intensities. A baseline was then fitted to the spectrum and removed. Our results show that PLS2 regression performed at this point predicts elemental concentrations with accuracies comparable to the original method. Perhaps the baseline, which is primarily caused by bremßtrahlung, is relatively consistent and systematic (at a given standoff distance), and thus its removal does not improve the quantitative nature of the spectral peaks.

Peak areas vs. channel intensities. Theoretically, peak areas should correlate directly with elemental abundances. With a baseline removed, lineshapes can be fitted to individual peaks in the spectra. Ideally Voigt profiles should be fit, but at this time we can only fit Gaussian and Lorentzian profiles. An automated routine fit all peaks in the spectra above a signal-to-noise threshold. PLS2 regressions were performed using the areas of these peaks as **X**. The results of the peak area analyses are comparable to those from the channel-intensity analyses. The area analyses suffer from a problem of matching—that is, the same peak in two spectra can be fit at slightly different energies, and it can be difficult to match them back up when binning the results. Work is underway to mitigate this problem and it is thought that a successful solution will greatly improve the results.

Minor elements. PLS2 regressions were also run including minor and trace elements among the **Y** variables. Some minor elements were predicted well, especially Ba, Cr, Ni, Rb, and Sr. In some cases, the highest PLS regression coefficients correspond to peaks from that element, such as 407 nm and 421 nm for Sr. In the case of Ba, the highest regression coefficient corresponds to the Ba peak at 455 nm, but this peak is strongly overlapped by Ti [3]. Recognizing this, the regression subtracts out the effect of Ti by a highly negative regression coefficient corresponding to the Ti peak at 501 nm. This is a clear example of why PLS can be dramatically better than univariate analysis.

Other elements were predicted well because of the well-known correlations between major and minor elements in geological samples; e.g., Rb readily assumes the crystallographic role of K by a simple substitution. PLS does not care which element produces a signal, so long as it correlates with an elemental abundance, and the highest regression coefficients predicting Rb correspond to K lines. This issue also arises among the major elements. For example, Si lines are few, but in igneous rocks Si correlates with many other elements (e.g. negatively with Fe, Mg). Thus, features due to other elements are used in predicting Si.

Elemental correlations: Using elemental correlations to predict abundances may be quite useful. A better calibration set than the randomly chosen one in the above analyses could be compiled either by choosing samples in which correlations are minimized, or those in which they are exploited maximally. The latter should be used with caution, because exploiting geochemical correlations in a training set assumes these correlations exist in the unknown samples. To this end, the 100 samples were split into a high-silica group and a low-silica group, with the cutoff at 52 wt% SiO_2 (between basalt and basaltic andesite). The high silica group shows a strong correlation between SiO_2 and Na_2O+K_2O (R^2=0.79) and the low silica group exhibits no such correlation (R^2=0). These two groups were themselves split in half into test sets and training sets, and separate PLS regressions were used to predict compositions of the two new test sets. Results show significant and sometimes dramatic improvement in the predictive power of the two smaller training sets over the single 50-sample training set used in all previous regressions, showing the promise of this approach.

Acknowledgments: We are grateful for support from NASA grants NNG06GH35G and NNX09AL21G.

References: [1] Rhodes J.M. and Vollinger M.J. (2004) *Geochem. Geophys. Geosyst.*, 5 Q03G13. [2] Schaefer et al. (2008) *LPSC XXXIX*, Abstract #2171. [3] Tucker et al. (2009) *LPSC XXXX*, Abstract #2024.