ERROR ANALYSIS FOR REMOTE LASER-INDUCED BREAKDOWN SPECTROSCOPY ANALYSIS USING COMBINATIONS OF IGNEOUS, SEDIMENTARY, AND PHYLLOSILICATE SAMPLES. M. D. Dyar[1], M. L. Carmosino[1], J. M. Tucker[2], E.A. Speicher[1], E. B. Brown[1], S. M. Clegg[3], R. C. Wiens[3], J. E. Barefield[3], J. S. Delaney[4], G. M. Ashley[4], and S. G. Driese[5]. [1]Dept. of Astronomy, Mount Holyoke College, mdyar@mtholyoke.edu, [2]Dept. of Earth and Planetary Sciences, Harvard Univ., 20 Oxford Street, Cambridge MA 02138, [3]Los Alamos National Laboratory, P.O. Box 1663, MS J565, Los Alamos, NM 87545, [4]Dept. of Earth and Planetary Sci., Rutgers University, [5]Dept. of Geology, Baylor Univ., One Bear Place #97354, Waco, TX 76798.

Introduction: Laser-induced breakdown spectroscopy (LIBS) will be used by the ChemCam instrument on MSL to obtain chemical analyses of martian geology with ~175-500 μm spot sizes at standoff distances of 1.5-7 m. To interpret data from diverse martian rock types, multivariate statistical methods are needed to overcome chemical matrix effects in LIBS and correct for variations in peak areas caused by interactions in the plasma that are functions of chemical composition.

Previous studies have used small data sets [1] or ones with restricted ranges of rock type and composition [2] to demonstrate the usefulness of partial least-squares (PLS) techniques in mitigating matrix effects and reducing errors on major elemental analyses. To apply PLS and other multivariate techniques to routine analyses of martian samples, a training set of spectra from samples with known chemical compositions is used to build predictive models. An ideal training set should span the smallest compositional range possible but must also fully represent the variation in unknown samples, as demonstrated in [2] within a suite of igneous rocks. This study tests those conclusions further by comparing prediction accuracies for combinations of different rock types and minerals (igneous, sedimentary, and phyllosilicates) to better understand how the samples represented in a training set used for calibration affect PLS elemental predictions for unknowns.

Experimental methods and samples: Laboratory conditions mimic ChemCam and Mars conditions as closely as possible [2]. We utilized three LIBS sample sets with disparate compositions (Table 1). The "century set" (Cset) consists of 100 igneous rock samples [2]; the majority are basalts by composition, but samples of higher and lower silica content are included to represent a wide range of naturally-occurring igneous rock compositions. The sedimentary rock suite (Seds) includes 17 samples from Olduvai Gorge, Tanzania in the East African Ridge [3]. The 17 phyllosilicate samples (Phyllo) [4] were obtained from the CMS Source Clays Repository, the CRPG, the Czech Geological Survey, and the Harvard Mineralogical Museum.

Statistical methods: Data were analyzed using software written in GNU R [5] by M.L.C. This customized R software wraps routines from several packages including hyperSpec [6], Peaks [7], and PLS [8] and specifically applies them to LIBS data sets. Both PLS-1, which regresses a single response variable (element) against the predictor variables (spectra), and PLS-2, which regresses multiple responses against the predictors, were used to explain the variance in **X** and **Y**. The tests used only the major elements Si, Al, Ti, Fe, Mg, Mn, Ca, Na, P, and K for **Y** variables; **X** variables were the 6144 channels of the three detectors.

Table 1. Compositions of Three Data Sets Used

	Century Set of Igneous Rocks		Phyllosilicate		Sedimentary Rocks	
	mean	s.d.	mean	s.d.	mean	s.d.
SiO_2	52.53	9.69	53.22	9.41	51.89	9.84
Al_2O_3	12.92	3.43	16.48	12.28	13.72	2.90
TiO_2	2.04	1.40	0.42	0.61	1.26	0.34
Fe_2O_3T	10.79	4.23	9.51	13.86	8.54	2.07
MgO	8.91	7.40	5.38	6.78	5.07	1.35
MnO	0.17	0.06	0.03	0.04	0.25	0.06
CaO	8.12	3.66	3.18	5.92	3.22	7.70
K_2O	1.35	1.60	1.29	2.78	2.13	0.49
Na_2O	2.56	1.15	0.40	0.69	4.50	1.14
P_2O_5	0.43	0.56	0.09	0.21	0.20	0.04

s.d. = standard deviation

Internal validation was used to tune parameters of the model. External validation used a test set related to the unknowns to apply derived parameters and estimate how generalizable the data will be with those fixed parameters. Root mean square error predictions were used to compare results (RMSEP). The number of components to be used in the models was individually chosen for each element using the first local minimum value of RMSEP or the number of components that gives the absolute lowest value of RMSEP for each element (global model). Simple leave-one-out, internally- and cross-validated models were also calculated for each of the three data sets individually.

Predictions of igneous rock compositions: Although every sample in the Cset is an igneous rock, there is intentionally a great deal of compositional diversity represented. Thus it is unsurprising that there is little difference in predicting compositions of Cset samples from ½Cset alone vs. ½Cset+Phyllo+Seds (Table 2). Also as expected, using 34 Sed+Phyllo spectra to predict 100 Cset samples is unsuccessful.

Predictions of phyllosilicate compositions: An externally-validated model using global minima and only the Cset provides the best RMSEP values across the major elements (shaded column in Table 2); using only the Cset rather than the Cset+Seds yielded the lowest RMSEP values overall. This suggests that ei-

ther there is no additional compositional diversity within the sediments that contributes anything to the Cset model alone or there is so much compositional diversity in the Seds and they are so different from the Phyllo samples that they are not useful to the model.

Predictions of sedimentary rock compositions: Use of external validation coupled with the 1st local minimum in the RMSEP provides optimal results in this data set (shaded column inTable 2). For nearly all elements, including both Phyllo and the Cset in the training set yields better or comparable prediction errors than using a specialized training set of Seds alone.

Discussion: These results demonstrate the importance of training set selection on expected test errors. However, it is very difficult to discern any guiding principles for subset selection from our results. For example, in the Cset some elements (e.g., MgO and CaO) are better predicted by the largest possible training set, while others (e.g. SiO_2 and Fe_2O_3) are optimized when only other igneous rocks are used. In [2], using training set samples deliberately-chosen to reflect composition extremes gave slightly better predicitions for all elements except Na_2O. Using smaller data sets with closely-related compositions [2] yielded lower prediction errors for high and low-Si igneous sample groups. However, this latter trend (tighter compositional range in training set reduces prediction errors) is not broadly applicable because it is not observed for the other rock and mineral types studied here. Prediction errors for Phyllo are lowest when solely the Cset is used to train, but Seds are best predicted by a larger training set with both Cset+Phyllo.

All these results underscore the importance of training set selection in producing optimal LIBS prediction errors, and suggest a need for an automated methodology to choose specialized training sets for predictions of individual unknowns. The ad-hoc attempts presented here and in [2] show that using geological reasoning to manually select training set samples can only go so far in optimization of prediction errors. More rigorous (less subjective) statistical procedures for training set selection are needed. Unfortunately, PLS is not designed to select specific samples for training sets. PLS provides high-dimensional regression capabilities and is thus useful for LIBS analysis in some applications, but it provides no framework for taking advantage of structural similarities in the data. Other high-dimensional regression techniques that combine the ability to shrink the number of input variables using projection with automatic selection of similar samples for predictions will be needed.

Acknowledgments: This work was supported by NASA MFRP grants NNG06GH35G and NNX09AL21G, as well as funding from the Massachusetts Space Grant Consortium.

References: [1] Clegg S.M. et al. (2009) *Spectrochim. Acta Part B: Atom. Spectr.,* 64, 79-88. [2] Tucker J. M. et al. (2010) *Chem. Geol.*, 277, 137-148. [3] Ashley G. M. and Driese S.G. (2000) *J. Sed. Res.*, 70, 1065-1080 and Clegg S.M. et al. (2008) *LPSC XXXIX*, Abstract #2107. [4] Tucker J. M. et al. (2008) *Wkshp on Martian Phyllo.,* Abstract #7028. [5] R Development Core Team (2010) *R: A language and environment for statistical computing.* R Found. Stat. l Comput., Vienna, Austria, URL http://www.R-project.org. [6] Beleites and Sergo, in prep. hyperSpec: a package to handle hyperspectral data sets in R', *J. Stats. Software*, R package v. 0.95. URL: http://hyperspec.r-forge.r-project.org. [7] Morháč, M. et al. (2008) Peaks: Peaks. R package vs. 0.2. http://www.slac.stanford.edu/comp/unix/package/cernroot/22312/TSpectrum.html and Morháč, M. et al. (2009) *Nucl. Instr. Methods Phys. Res. A*,, 600, 478-487. [8] Wehrens, R. and Mevik, B.-H. (2007) R package 2.1-0. http://mevik.net/work/software/pls.html [9] Hastie T. et al. (2009) *The Elements of Statistical Learning*, 2nd Ed., Springer Science, New York, 745 pp.

Table 2. Best Root Mean Square Errors of Prediction (RMSEP) Values for Three Data Sets (in units of ±wt.%)

Goal:	**Predicting Igneous Rocks**				**Predicting Phyllosilicates**				**Predicting Sedimentary Rocks**			
Method:	PLS-2 Internal Cross-Validation	Internal validation, PLS-2 Global Minimum			PLS-2 Internal Cross-Validation	External validation, PLS-2 Global Minimum			PLS-2 Internal Cross-Validation	External validation PLS-1 1st local Minimum		
Training set used:	**Cset****	**½Cset**	**Seds + Phyllo**	**Seds + Phyllo+½Cset**	**Phyllo****	**Seds**	**Cset**	**Seds+ Cset**	**Seds****	**Phyllo**	**Cset**	**Phyllo +Cset**
SiO_2	3.89-7.44	3.78	18.26	4.11	10.91-17.18	4.65	3.26	3.41	2.22-2.93	4.02	3.26	1.88
Al_2O_3	1.76-4.67	3.28	7.04	3.08	2.41-9.73	0.53	0.76	1.33	0.62-1.10	1.78	0.71	0.57
TiO_2	0.41-1.18	0.52	0.55	0.55	1.16-1.61	1.40	0.29	0.92	0.23-0.31	0.41	0.25	0.37
Fe_2O_3T	3.23-5.56	2.06	12.24	3.59	8.20-11.64	5.00	1.43	6.31	1.02-1.50	2.36	1.33	1.21
MgO	1.86-4.47	3.51	3.83	3.25	1.50-3.39	1.29	5.20	2.12	0.71-1.24	0.87	0.94	1.00
MnO	0.05-0.11	0.03	0.12	0.04	0.16-0.26	0.11	0.07	0.07	0.04-0.07	0.11	0.07	0.07
CaO	1.50-1.80	2.00	4.67	1.25	4.47-5.98	2.03	0.87	0.83	0.43-1.67	0.48	0.82	0.64
K_2O	0.77-1.80	1.01	1.43	0.71	1.88-3.61	1.24	0.89	1.22	0.36-0.48	0.55	0.34	0.46
Na_2O	0.64-1.23	1.08	2.11	0.66	0.60-1.09	0.87	0.90	1.13	0.43-0.98	1.07	0.54	0.71
P_2O_5	0.27-0.48	0.29	0.34	0.36	0.23-0.60	0.53	0.06	0.03	0.02-0.04	0.09	0.07	0.06
Sum*	**n.a.**	**17.56**	**50.59**	**17.61**	**n.a.**	**17.65**	**13.72**	**17.38**	**n.a.**	**11.74**	**8.33**	**6.96**

* The sum of RMSEP values provides an arbitrary means of comparing different models, though it has no statistical use.

**The optimal number of components in each of these models could not be chosen because of the small size of Seds and Phyllo data sets (the same calculation was also used for the Cser to allow comparisons among them). However, the prediction errors will fall in the ranges given.