CHEMCAM LIBS Multivariate Regression Models Accuracy Assessment, J. Lasue[1], S. M. Clegg[2], O. Forni[1], R. B. Anderson[3], M. D. Dyar[4], C. Fabre[5], O. Gasnault[1], E. Lewin[6], S. Maurice[1], R. L. Tokar[7], R. C. Wiens[2], and the MSL Science team. [1]IRAP-OMP, CNRS-UPS, Toulouse, France (jlasue@irap.omp.eu) [2]Los Alamos National Laboratory, NM, USA, [3]US Geological Survey, Flagstaf f, AZ, USA [4]Mount Holyoke College, South Hadley, MA, USA, [5]G2R-Georessources, Nancy, France, [6]ISTerre, Grenoble, France, [7]Planetary Science Institute, Tucson, AZ, USA

Introduction: ChemCam is a Laser-Induced Breakdown Spectroscopy (LIBS) instrument on-board the Mars Science Laboratory (MSL). It can analyze the chemical composition of geological samples at a distance by detecting the light emission of constituent elements [1, 2]. The instrument has now successfully acquired >100000 *in situ* spectra. Each spectrum consists of intensity signals distributed over 6144 channels. The structure of the spectral data can be handled by multivariate data analysis techniques for classification [3] as well as quantification purposes. The ChemCam team uses a Partial Least Squares (PLS) algorithm, regressed against a database of 66 standards measured by ChemCam prior to flight, to provide rapid elemental abundances on the tactical (day-to-day) timeline [4, 5, 6]. PLS can correct for chemical matrix effects which can obscure a linear relationship between abundance and peak areas [*e.g.* 7, 8]. Since MSL landing and the start of operations, an updated database for chemical predictions has been generated in the laboratory [9] that has been shown to improve significantly the PLS prediction accuracies [10]. In this work, we compare the PLS accuracy results [6, 10] with those obtained through comparable techniques to consolidate the accuracy and optimization of the prediction models for the current and updated databases.

ChemCam data quantification procedure:

One of the most used and most accurate techniques for LIBS instruments is the PLS regression [4, 5]. Using a database of standards of known compositions analyzed under laboratory conditions replicating Mars, it is possible to use the PLS regression on the spectra obtained on martian targets to accurately predict their chemical composition [6, 7]. The algorithm currently used (PLS1) predicts the abundance of one chemical element at a time to prevent inferring false correlations amongst elements present in the training set.

In order to determine the accuracy, a classic "leave-one-out" (LOO) cross-validation procedure was used, where one standard is removed at a time and predicted using the remaining standards as training set. The accuracy of the model is obtained using the root mean square of error prediction: $RMSEP = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_{ip} - y_i\right)^2}$, where n is the number of standards, y_i the composition of the sample for a given element and y_{ip} is the prediction of that composition for the LOO model. The PLS algorithm is closely related to the Principal Components Analysis technique, where the eigenvectors and eigenvalues (also called loadings and Principal Components, PC) of the variance-covariance matrix of the dataset are computed as they correspond to the directions of largest variance in the dataset and allow to describe the data set with a smaller number of dimensions. The RMSEP value describing the accuracy of the model for each element varies as a function of the number of dimensions used to describe the dataset. It typically decreases as the number of principal components (NP) increases before reaching a minimum and an asymptotic behavior. The variation of the RMSEP as a function of NP is used to select the best training model for the predictions by taking the NC for which the RMSEP value is about 1σ above its minimum [see *e.g.* 6, 11].

The RMSEP values obtained for the current PLS algorithm, applied to the current database of 66 geological standards used to analyse ChemCam data are given in table 1 [6]. The best number of components (NC) is shown and differs from element to element based on the optimization done.

Several techniques have been used to validate the PLS algorithm with ChemCam data. The loadings used to predict the elements have been checked to be correlated with the emission lines of the relevant elements in the spectra. The PLS components also compare qualitatively well with the Independent Components Analysis classification of the LIBS spectra based on the separation of the emission lines of each chemical elements [5]. Furthermore, the ChemCam calibration targets on-board the MSL rover have been used to verify the PLS predictions. The table also gives the precision of the instrument based on measurement repetability on targets from [12] and the distribution of the chemical compositions of the standards in the database by indicating the minimum, median, maximum and 1st and 3rd quartile values for each oxide.

Analysis using other multivariate regression methods:

Other multivariate regression methods can potentially be more robust to non-linear behavior than PLS. Sparse methods such as the Least Absolute Shrinkage and Selection Operator (LASSO) or the elastic net will only regress with selected values of the spectra that are most correlated with the element to predict, which makes for more robust regression models. Their models and accuracies are quite comparable to the ones obtained using the PLS technique [13].

Recent multivariate data analysis developments have included another family of algorithms based on

the so-called Support Vector Machines (SVM). These techniques initially developed for multivariate data clustering can also be for regression modeling. In terms of clustering, the principle of the technique is to reduce the relevant number of data points from each cluster to the few data points whose position best defines the hyperplane that separate the data clusters and optimizes the margins of clusters separation [14]. Reducing the number of data points in such a way makes for a robust and fast computation of the clustering separation. Moreover, non-linear borders between data points can be defined by introducing specific kernel functions that map the initial data set into spaces where the hyperplanes are more easily calculated.

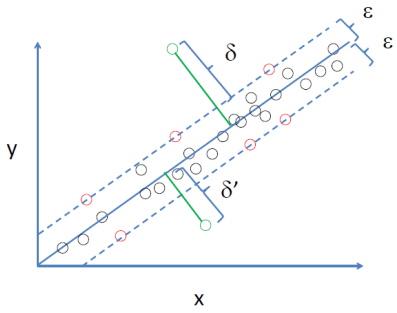

Figure 1: Illustration of the SVM regression principle. Instead of minimizing the distance between the regression line and each data point, a subset of points (the Support Vector Machines, in red) that defines the hyperplane is used to optimize the distance ε. Some outliers (green) can be included in the model with δ and δ' weights.

The adaptation of the technique to the regression problem is similar in scope to the clustering technique by finding the hyperplanes that follow the regression of data points and optimizing their position using few data points located at the extreme of the regression cluster. Similarly to the clustering problem, in the case of nonlinear regression, the hyperplane can be best defined after mapping the data point by an appropriate kernel function optimized to give the best results. This method has been shown to compare favorably with the more common PLS technique [15, 16].

We have generated multivariate predictive models using the SVM technique and based on the initial ChemCam database of 66 geological standards. The parameters used in this procedure were a cost of 1 and the best kernel function chosen amongst linear, radial and polynomial models. The cross-validation used in this case was to randomly select a test set of one third of the samples, and a training set of two-thirds of the samples, 40 different times. The training set is used to generate the predictive regression model and the test set predictions are used to calculated the RMSEP for this model. The results of these calculations are shown in Table 1. The optimized PLS1 algorithm gives better accuracies than the SVM method while both sets of values are of the same order of magnitude. We can therefore expect to confirm in the same way the strong improvements obtained on the accuracy of the PLS predictive models based on the updated ChemCam database [9, 10].

	SiO2	TiO2	Al2O3	FeOT	MgO	CaO	Na2O	K2O
TRAIN MIN.	0.2	0	0	0	0	0.1	0	0
TRAIN 1ST QUART.	40.8	0.27	5	2.7	0.8	2.5	0.3	0.3
TRAIN MEDIAN	48.6	0.68	13.1	6	2.2	7.1	2.4	0.8
TRAIN 3RD QUART.	59.3	1.47	16.1	12.1	6.4	12.8	3.4	1.8
TRAIN MAX.	75.4	5.9	38.8	36.2	49.2	54.9	5.9	6.4
NORM	3	1	1	1	1	3	1	3
NC PLS	8	10	4	7	8	8	10	4
RMSEP PLS	7.1	0.55	3.7	4	3	3.3	0.7	0.9
RMSEP SVM	8.2	0.7	4.5	5	4.6	4.8	1	0.9
PRECISION	1.5	0.14	0.57	1.8	0.49	0.42	0.49	0.14

Table 1: Comparison between RMSEP values obtained at the best NC for the PLS method and the SVM regression method and instrument precision. All values in wt.%. Upper part of the table correspond to the training data set quartiles values for each oxide composition. (adapted from [5, 12])

Conclusion: All these tests give us confidence in the fact that the PLS results give optimized predictions for the chemical content of the martian targets, and that in the few cases of absolute value biases, at least the trends calculated follow an actual composition change between targets. Similar to APXS, the precision of the method is normally much lower than the predictive models accuracy as indicated by the RMSEP. We have verified the fact that PLS and other multivariate techniques generate consistent accuracies for the models predicting the chemical composition of Mars targets from ChemCam LIBS data.

Acknowledgments: The instrument was developed in the US under the NASA Mars Program Office support to MSL. The instrument was supported in France at IRAP, under funding from CNES.

References: [1] Maurice S. et al (2012) Space Sci. Rev., DOI 10.1007/s11214-012-9902-4. [2] Wiens R.C. et al. (2012) Space Sci. Rev., DOI 10.1007/s11214-012-9912-2. [3] Forni O. et al. (2013) Spect. Acta B, DOI: 10.1016/j.sab.2013.05.003, [4] Lasue J. et al. (2013) LPSC, [5] Wiens R.C. et al., (2013), Spect. Acta B, DOI : 10.1016/j.sab.2013.02.003, [6] Anderson R. et al. (2014) this meeting, [7] Clegg S.M. et al. (2009) SpectroChim. Acta B, 64, 79. [8] Tucker J.M. et al. (2010) Chem. Geol. 277, 137, [9] Ehlmann B. et al. (2013) LPSC, [10] Clegg S.M. et al. (2014) LPSC, [11] Geladi P. and Kowalski P.R. (1986) Analyt. Chim. Acta 185, 1. [12] Blaney, D.L. et al. 2014. JGR (submitted). [13] Dyar D. et al. (2013) LPSC, [14] Boser, B.E., Guyon, I.M., Vapnik, V.N. (1992) COLT proceedings. [15] Thissen, U., et al. (2004), Chemomet. Int. Lab. Syst. [16] Hernandez, N., et al. (2009) Analyt. Chim. Acta.