APPLICATION OF MULTIVARIATE ANALYSIS TECHNIQUES FOR THE IDENTIFICATION OF SULFATES FROM RAMAN SPECTRA – IMPLICATIONS FOR EXOMARS.

G. Lopez-Reyes[1], P. Sobron[2], C. Lefevbre[2] and F. Rull[1], [1]Unidad Asociada UVA-CSIC-Center of Astrobiology (Valladolid, Spain. lopezrge@cab.inta-csic.es), [2] Space Science and Technology. Canadian Space Agency (St. Hubert, Canada).

Introduction: The Raman instrument (RLS) onboard the 2018 ExoMars rover will determine the structural and compositional features of Mars' surface and subsurface samples [1]. In its current configuration, samples will be collected using a drill, then crushed and finally delivered to the analytical laboratory [1]. While such concept of operation will enable adequate management of collected samples, geological and morphological context will be lost; thus, identification of the mineral phases present in the geological targets and quantification of their abundance with RLS will rely on spectral data alone. Innovative spectral treatment methods must be developed that will enable unambigous qualitative and quantitative identification.

Multivariate analysis (MVA) techniques for the quantitative analysis of the Raman spectra of several minerals and rocks show, for instance, that principal component analysis (PCA) is capable of differentiating mineral species such as carbonates, sulfates, oxides and silicates in geological samples [2]. Partial least squares (PLS) has been used to determine the quality of biodiesel fuels [3]. Artificial neural networks (ANN) have been designed for the identification and quantification of inorganic salts in water solutions [4]. Also, Combinations of this techniques for chemometrical analysis from Raman spectra have been attempted [5].

The aim of this study is to evaluate the aforementioned MVA techniques for the analysis of mineral samples in the framework of the operation of the RLS instrument. Sulfates are used because they are one of the two major types of secondary minerals found on Mars that may provide potentially habitable environments. Due to the association of sulfate salts with ancient aqueous environments in which life might have thrived, it is expected that sulfates will become priority targets for the ExoMars mission.

Spectra set: A total set of 17 spectra of sulfates were used as input for training all the three techniques. The set was divided in four subsets, each grouping sulfates with different hydration state: $FeSO_4$ (1, 4 and 7w), $MgSO_4$ (1-7w, and 12w), $CaSO_4$ (0, ½ and 2w) and Na_2SO_4 (0 and 10w). A set of mixed spectra was synthetically generated by computing linear combinations of the original spectra, parameterized with the expected proportion of the mixture and the cross-section of the mixed materials. Random noise was added to the synthetic spectra to guarantee differentiation. This set of mixtures was used to validate the models.

PCA model: PCA analysis of the samples showed that the first two components, PC1 and PC2, represent ~80% of the variance in the data set: 61% and 17%, respectively. The scores are plotted in Figure 1. Dehydrated or poorly hydrated salts are somewhat differentiated from the highly hydrated salts, for both the training and test samples.

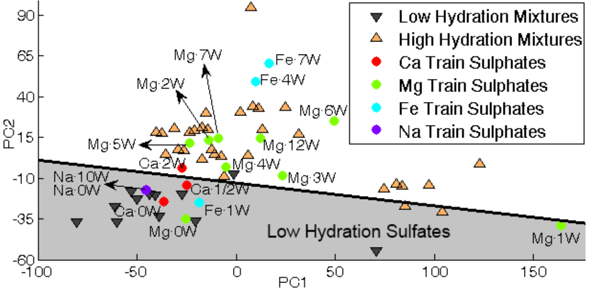

Figure 1. Scatter plot of PC2 vs. PC1 for training (colored) and validation sets (black) with PCA.

PLS model: The PLS responses were chosen to indicate the hydration level and the relative abundance of the various cations. The criterion to optimize the model behavior was to evaluate the Mean Square Error (MSE) of the difference between the expected regression responses and those fitted by the model.

Model 1: When using all the points of the spectrum as input variables, the optimum number of components in the PLSR model was found to be 7. In this case, a 93% mean correlation coefficient were achieved for the pure samples spectra, and a 89% for the mixtures.

Model 2: Alternatively, a different model was created based only on selected peaks from the spectra. In this case, the optimum number of components was 4, and also provided good prediction: linear unitary slope and 99% mean correlation coefficients for all the responses of the pure samples, and 95% mean correlation for the 1:1 proportion mixtures spectra (Figure 2). However, the model provided somewhat more inaccurate results for mixtures with unbalanced proportions.

ANN model: We have designed a three-layer Feed-Forward Back-Propagation network with 33 neurons in the input layer, 31 neurons in the hidden layer and 17 outputs corresponding to each of the samples, with log sigmoid neuron transfer functions. The input data for the ANN consisted on 33 Raman intensities at selected wavenumbers of non-overlapping peaks (the same as in the second PLS model previously described) to improve performance [6]. When using all the spectra points as input, no positive results were obtained.

Three different input sets were fed into the network: (1) pure sulfates for the training spectra set, (2) different pure sulfates spectra for validation and over-fitting avoidance, and (3) spectra of pure sulfates, and different proportion binary mixtures of them, for the test set.

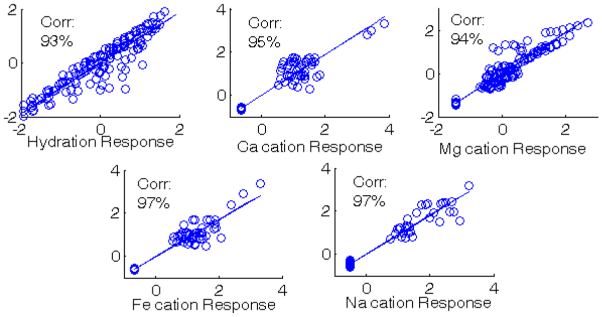

Figure 2. PLS model 2 responses for the mixtures.

Sample identification. By considering the highest outputs of the network as detected materials, the test spectra set showed that the network is able to achieve very high levels of accuracy in the detection of materials. The pure sulfates of the validation set were identified with 100% accuracy. The spectral noise of these spectra had to be increased 10 times in order to start reducing the accuracy of the network detection for the pure salts. The mixtures showed varying results depending on the relative abundances of the mixture (Figure 3). These preliminary results allow defining a detection window in which we achieve a positive identification ratio between 75% and 95% when the sulfates are mixed in relative abundances varying from 0.1:0.9 to 0.9:0.1. The more balanced the relative abundances, the higher the identification ratio. It is important to note that this figures were obtained when taking the highest output value as the only mineral present for pure materials, and the two highest for binary mixtures, which limits the model performace.

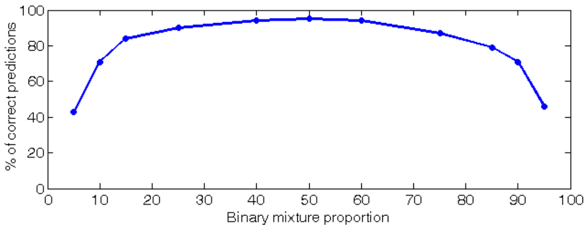

Figure 3. Identification accuracy of the ANN model.

Abundance quantification. In addition, the network outputs provide information on the relative abundance of the materials. Figure 4 represents the predicted proportions for a mixture of calcium and sodium sulfates in different relative abundances. It shows a certain degree of correlation between the modeled values and the actual abundances of the mixtures. However, this figure also shows that the model fails to correctly detect the 0.9·$CaSO_4$ + 0.1·Na_2SO_4 mixture.

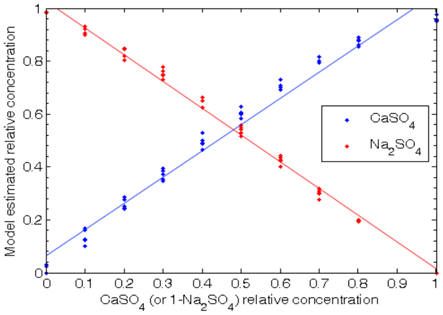

Figure 4. Detected relative abundances with ANN for a $CaSO_4$ + Na_2SO_4 mixture with different proportions.

Summary and conclusions: We have compared the performance of three MVA mineral ID techniques. The results presented here are encouraging, and demonstrate that these techniques can provide critical information for the identification and quantification of mineral phases in geological samples relevant for ExoMars from Raman spectra. PCA differentiated hydrated sulfates from dehydrated ones. PLS proved to be a good chemomentric tool for the detection and quantification of binary mixtures, especially with well-balanced relative abundances. ANN has proven to be a powerful tool for the identification of sulfates: by training only with pure samples, the model proved to be able to detect and somewhat quantify the abundance of both elements in binary mixtures with promising figures and quite unbalanced abundances in the mixtures. However, information on the number of mineral phases present on the sample is desirable to improve the performance without defining "arbitrary" decision thresholds. Also, the selection of the input variable is critical for the model performance.

The application of unsupervised multivariate techniques for the processing of RLS products is a must in order to provide science support to the ExoMars mission. These techniques will provide fast identification of materials and quantification of mineral species during the tactical operations of the rover-based mission.

Future work will focus on incorporating additional mineral samples (phyllosilicates, carbonates, igneous rocks, etc.) to our models, and in exploring additional techniques such as Kernel-PLS or MESMA.

Acknowledgements: A. Wang provided the Raman spectra used in this work. A. Koujelev provided helpful insights on the implementation of ANN. GLR acknowledges the University of Valladolid for providing funding for the project. PS and CL acknowledge support from NSERC.

References: [1] Rull, F. et al. *42 LPSC*, #2400. [2] Lafuente, B. et al. (2012). *Georaman X^{th}*, 149-150. [3] Ghesti et al. (2007) *Energy & Fuels*, 21 (5), 2475-2480. [4] Burikov et al. (2010) *Optical Memory & Neural Networks*, 19, 140-148. [5] Dorfer, T. et al. (2010). *JRS* 41(6), 684-689. [6] Koujelev, A. et al. (2010). Planet. Space Sci. 58(4): 682-690.