REVISION 2 – Correction date Feb. 26, 2014

Multivariate analysis of Raman spectra for the identification of sulfates - Implications for ExoMars

Guillermo Lopez-Reyes[1], Pablo Sobron[2,3,4], Catherine Lefebvre[2] and Fernando Rull[1]

[1]Unidad Asociada UVA-Centro de Astrobiología. Edificio INDITI, Av. Francisco Valles 8, Parque Tecnologico de Boecillo, Parcela 203, Boecillo 47151, Spain.

[2]Space Science & Technology, Canadian Space Agency, 6767. Rte. de l'Aéroport, St. Hubert, Quebec J3Y 8Y9, Canada.

[3]MalaUva Labs, 822 Allen Ave #A, St. Louis, Missouri, 63104, U.S.A.

[4]SETI Institute, 189 Bernardo Ave #100, Mountain View, California 94043, U.S.A.

ABSTRACT

We have built three multivariate analysis mathematical models based on principal component analysis (PCA), partial least squares (PLS), and artificial neural networks (ANN) to detect sulfate minerals in geological samples from laser Raman spectral data. We have critically assessed the potential of the models to automatically detect and quantify the abundance of selected Ca-, Fe-, Na-, and Mg-sulfates in binary mixtures. Samples were analyzed using a laboratory version of the Raman Laser Spectrometer (RLS) instrument onboard the European Space Agency 2018 ExoMars mission. Our results show that PCA and PLS, can be used to quantify to some extent the abundance of mineral phases. PCA separated hydrated from dehydrated mixtures and classified mixtures depending on the phases abundances. PLS provided relatively good calibration curves for these mixtures. Upon spectral pre-processing, ANN provided the most precise qualitative and quantitative results. The detection of mineral phases was 100% accurate for pure samples, as was for binary mixtures where the abundance of mineral phases was > 10%. The outputs of the

ANN were proportional to the phase abundance of the mixture, thus demonstrating the ability of ANN to quantify the abundance of different phases without the need for calibration. Taken together, our findings demonstrate that multivariate analysis provides critical qualitative and quantitative information about the studied sulfate minerals.

Keywords: Sulfates, ExoMars, Raman spectroscopy, Multivariate Analysis, Qualitative, Quantitative

INTRODUCTION

Laser Raman spectroscopy has been proposed as a powerful tool for the identification of minerals in the context of planetary exploration, including Mars (Sharma et al., 2003; Sobron et al., 2013a; Sobron et al., 2008; Wang et al., 2003; Wiens et al., 2007), Europa (Angel et al., 2012; Sobron et al., 2013b), Venus (Lambert et al., 2010), the Moon (Ling et al., 2009), and asteroids (Kong and Wang, 2010). In addition, the feasibility of using laser Raman spectroscopy for the detection of biosignatures in terrestrial analogues to Mars has been demonstrated, e.g. (Bower et al., 2013; Dickensheets et al., 2000; Edwards et al., 2012; Edwards et al., 2011; Edwards et al., 2003; Ellery and Wynn-Williams, 2003; Steele et al., 2010; Wynn-Williams and Edwards, 2000). A Raman Laser Spectrometer (RLS) is part of the science instrument payload of the European Space Agency 2018 ExoMars mission; the RLS instrument will target mineralogical and astrobiological investigations on the surface and subsurface of Mars (Rull et al., 2011a; Rull et al., 2011b).

The current concept of operation of the RLS instrument is a raster analysis of crushed drill-core materials (Rull et al., 2011a). In this configuration, the geological and morphological context of the spots analyzed by RLS will be lost, as the crushing stage will preclude correlation between RLS spectra and the imagery acquired by the rover Close-Up Imager CLUPI (Josset et al., 2012). While the synergy between these instruments in the

current ExoMars payload configuration has been demonstrated (Lopez-Reyes, 2013a), identification of the mineral phases present in the geological targets and quantification of their abundance with RLS will mostly rely on spectral data and not morphology or texture. Therefore, in order to enable unambiguous identification and quantification of phase abundance, robust spectral processing methods are needed.

Mineral identification using Raman spectroscopy is often performed by comparing acquired spectra to reference spectra available from the literature and different databases e.g., the RRUFF project database (Downs, 2006). A number of algorithms have been developed that enable an automated identification of Raman spectra using traditional univariate analysis, *i.e.*, the description of individual variables in a given spectrum (Hermosilla, 2013; Kriesten et al., 2008; Perez-Pueyo et al., 2004; Sobron et al., 2008). These algorithms, however, fail to accurately estimate mineral abundance in complex geological samples, though applications for the quantitative analysis of relatively simple mixtures have been proposed (Lopez-Reyes, 2013a; Schumacher et al., 2011; Vagenas, 2003).

Multivariate Analysis Techniques (MVAT) are statistical techniques which deal with simultaneous measurements on many variables, and aim at understanding the relationships between these many variables to predict the values of important properties not directly measurable (Johnson and Wichern, 2002). Some examples of MVAT applied to the analysis of Raman spectra in the literature show that principal component analysis (PCA) is capable of differentiating mineral species such as carbonates, sulfates, oxides and silicates in geological samples (Lafuente, 2012). Partial least squares (PLS) has been used to determine the quality of biodiesel fuels (Ghesti et al., 2007). Artificial neural networks (ANN) have been designed for the identification and quantification of inorganic salts in

water solutions with Raman spectra (Dolenko et al., 2005). Also, combinations and comparisons of these techniques for chemometrical analysis from Raman spectra (mostly qualitative) have been reported (Dorfer et al., 2010; Ishikawa and Gulick, 2013; Özbalci et al., 2013).

In this work we evaluate the feasibility of using PCA, PLS, and ANN for the identification and quantification of sulfate salts in binary mixtures through analysis of spectra recorded using a breadboard of the flight Raman Laser Spectrometer. The use of binary mixtures is a first step to evaluate these techniques, prior to being tested with more complex samples. We utilized a set of pure sulfates synthetized in the laboratory in order to prepare the binary mixtures. The sulfates we have considered have been proposed as priority targets for astrobiological investigation of Mars (Chou et al., 2013; King and McLennan, 2010; Knoll et al., 2005; Wang et al., 2011) because of their association with liquid water on certain locations on Mars and sulfate reducing bacteria in terrestrial analogue environments (Fernandez-Remolar et al., 2012; Nixon et al., 2013; Sanchez-Andrea et al., 2012). Such environments are known to preserve chemical and morphological fossils (Bonny and Jones, 2003), thus testifying to the importance of the detection and identification of sulfates and the detailed study of their degree of hydration in order to address Mars' hydrologic history and potential for habitability.

In terms of identification, Raman spectroscopy is a very powerful technique for the analysis of the mineralogy of terrestrial sulfate samples analogue to Mars (Sobron et al., 2014; Sobron and Alpers, 2013; Sobron et al., 2009), as well as for the detailed characterization of the hydration states of this kind of salts, which is critical for the rigorous interpretation of the hydrologic history of Mars (Chou et al., 2013; Ling et al., 2008; Wang et al., 2006).

MATERIALS AND METHODS

Samples

A set of 17 Raman spectra of sulfates were used as input for training the PCA, PLS, and ANN MVAT models; hereinafter, we will refer to this set of spectra as training set. Table 1 lists the materials used to record our set of spectra. The materials were synthesized in the laboratory using standard techniques (Ling et al., 2008) -- they were characterized using X-ray diffraction in order to certify their mineralogical composition. The Raman spectra of these materials were acquired using Raman instrumentation described in (Ling et al., 2008; Wang et al., 2006).

Apart from the training set, two more sets of spectra have been defined: the validation and the test sets. A random selection of the three sets of spectra spectra is shown in Figure 1. The training set was divided in four subsets: Ca-, Mg-, Fe-, and Na-sulfates. Each of these four groups included sulfates with several degrees of hydration. The validation set consists on a set of binary mixed spectra with different proportions (0.1:0.9, 0.25:0.75, 0.5:0.5, 0.75:0.25, 0.9:0.1) of all mineral pairs of the samples. It was synthetically generated by computing linear combinations of these spectra, parameterized with the expected proportion of the mixture and the cross-section of the mixed materials. Randomized noise with realistic amplitude was added to the synthetic spectra to guarantee differentiation. Though probably not totally accurate, this set of spectra is expected to behave similarly to weight-proportion mixtures, providing an easy and convenient way to execute the models with hundreds of spectra without the actual need of preparing mixtures and acquiring spectra.

The third set of spectra, the test set, included the Raman spectra of powdered mixtures of anhydrite (CaSO4), thenardite (NaSO4), and MgSO4. This set of spectra is

used to test the models as well as to assess the goodness of using the computed spectra of the validation set as part of the model training/validation procedure. The Raman spectra of anhydrite and MgSO4 show non-overlapping peaks with thenardite, thus facilitating the spectral processing described below. Two subsets of samples (anhydrite + thenardite, and thenardite + MgSO4) with a total of 7 samples each were prepared by mixing these materials in the following weight proportions: 0.01:0.99, 0.1:0.9, 0.25:0.75, 0.5:0.5, 0.75:0.25, 0.9:0.1 and 0.99:0.01. These mixtures were acquired with the RLS instrument, contrary to the training and validation spectra. This way, the models independence with respect to the instruments responses could be assessed. Thirty spectra of each mixture were acquired and averaged, providing one final spectrum of each mixture. These spectra are representative of the overall behavior of the samples, as the average of different acquisitions of 30 spectra have proved to be almost equal when obtained from an ExoMars-type powdered sample with the RLS instrument (Lopez-Reyes. 2013a).

Instrument and Raman spectroscopy

In its current configuration, the ExoMars rover will crush the subsurface drilled samples and provide the instruments a flattened surface of the powdered sample for analysis. The RLS instrument, in its baseline mode operation, will analyze from 20 to 40 points of each sample, with a 50 μm spot size and an irradiance level of 0.6 – 1.2 kW/cm^2 of a 532 nm continuous wave laser (Rull et al., 2011a). In order to test the analytical capabilities of the instrument in an operation-like environment, including fully automated analysis on powdered samples, a RLS ExoMars Simulator has been developed at the University of Valladolid – CSIC – Center of Astrobiology Associated Unit ERICA (Foucher, 2012; Lopez-Reyes, 2013a; Rull, 2011). The Simulator was used to acquire the test set spectra from the mixtures with different phase abundances.

In order to improve the accuracy of the analytical models described below, and given that the aim of the models is to distinguish among different types of sulfates, only spectral regions that are relevant to sulfates were considered. These spectral regions are: (1) the sulfate symmetric stretching ν_1 (950 to 1100 cm^{-1}); (2) the sulfate asymmetric stretching ν_3 (1100 to 1220 cm^{-1}); (3) the sulfate symmetric and asymmetric bending ν_2 and ν_4, respectively (100-750 cm^{-1}); (4) the water bending (1600-1700 cm^{-1}); and (5) the water and OH stretching (2800-3800 cm^{-1}). The spectra were baseline-corrected to remove background contributions (e.g., fluorescence), and normalized in intensity in a way such the maximum peak intensity is 1. In our case, the separation between consecutive points of the spectrum is 0.5 cm^{-1}, which provides a total number of variables of about 4000 (each variable corresponding to the intensity at a determined wavenumber of the selected spectral regions), which is the size of the input data to the different models.

MVAT

This section describes the MVAT that have been used in this work: Principal Component Analysis (PCA), Partial Least-Squares regression (PLS) and Artificial Neural Networks (ANNs). PCA and PLS extract latent variables from the system in order to represent the system in a complexity-reduced variable system. Ideally, the extracted latent variables will respond to physical properties of the model. The difference between PCA and PLS is that, while PCA extracts the variables, PLS also performs a regression on the expected responses of the system for a determined set of inputs. ANNs, on the other hand, is a technique that can model any non-linear function by example-based training of computational networks. While ANNs philosophy makes it possible to have direct outputs of the sample presence and abundance, PCA and PLS provide responses based on the latent variables of the system that need to be classified and/or calibrated to extract the mineral

phases presence/abundance values. The scope of this work does not include the classification step for PCA and PLS responses, but it aims at evaluating the techniques ability to differentiate among such samples. All these techniques are described below in more detail, and were implemented for this work using Mathworks MATLAB.

Principal Component Analysis

PCA is a multivariate technique that computes the variance-covariance structure of a set of variables through a few linear combinations of them, with the objective of reducing the number of data variables (Johnson and Wichern, 2002). PCA calculates new variables called Principal Components (PCs) as linear combinations of the original variables. All the principal components are orthogonal to each other, so the variables in the principal component space do not provide redundant information. The principal components as a whole form an orthogonal basis for the space of the data, which can thus be represented in this new space. There is an infinite number of ways to construct an orthogonal basis to represent the data, so PCA calculates the PCs taking into account that they are uncorrelated among them, while maximizing their variance with coefficient vectors of unit length (as the variance can easily be increased by multiplying by a constant (Johnson and Wichern, 2002)). In other words, the first PC is calculated, among all the possibilities, as the single axis (linear combination of the original variables) in the space that provides the greatest variance by any projection of the data on that axis. The second PC is another single axis that provides the second greatest variance by any projection of the data, and that is orthogonal to the first PC. The third will be also perpendicular to the previous two, and so on. This way, the set of PCs conforms a new set of coordinates which can represent the original data, where most of the variance of the system can be explained with only a few of the PCs. This is due to the fact that, usually, many of the variables of a system are highly

correlated, making it possible to remove some of them and still have a variable set which can account for most of the variability of the system.

The PCA model was trained with the 17 spectra from the pure sulfates (training set), and then applied to the validation and test spectra sets.

Partial Least Squares

Partial Least Squares (PLS) is a common term for a family of multivariate modeling methods that appeared in the 1980's to solve problems in social sciences (e.g. (Wold et al., 1983)), but that can be applied to a large variety of modeling problems (Martens and Naes, 1992), including Raman spectroscopy, as discussed above. The underlying assumption of all PLS methods is, as for PCA, that the observed data is generated by a system which is driven by a small number of latent (not directly observed or measured) variables, called components. In addition, PLS performs a regression of the expected responses for each set of input variables (observations). Thus, this technique is some kind of combination of PCA and linear regression: PLS creates orthogonal score vectors (the latent vectors or components) by maximizing the covariance between two different blocks of variables (Rosipal and Krämer 2006) which correspond to the input variables (predictor variables or observations) and the expected responses (predicted variables or predictions). The higher the number of computed components, the better the model fits the system. In our work, the SIMPLS algorithm (de Jong, 1993) has been used, which calculates the PLS factors directly as linear combinations of the original variables. With SIMPLS, the PLS factors are determined such as to maximize the covariance between the predictor variables and the expected responses, while obeying certain orthogonality and normalization restrictions (de Jong, 1993).

The responses chosen for the analysis of Raman spectra of sulfates were the weight atomic fraction of the sulfate cation (calcium, magnesium, iron and sodium), as well as the weight ratio of the water bound to the sulfate molecule, providing a set of five responses. These responses were chosen to reflect the composition of the samples, thus providing the model with a physical basis for the regression of the input variables. Given that the shifts in the Raman bands are produced as a consequence of the frequency shifts associated to the cation-sulfate group oscillators, the cation ratio is the subjacent physical property that we choose as a response in our model. The spectra were baseline-removed, normalized with respect to the maximum peak height, mean centered and scaled in order to avoid biasing the model.

The model was trained with the spectra from the 17 pure salts (training set). As in every model where input variables are regressed to expected responses, there is a risk of over-fitting. To minimize this effect, we optimized our model using a leave-one-out cross-validation method with the validation and test spectra sets, as outlined in the Results and Discussion section.

Artificial Neural Network

An Artificial Neural Network (ANN) is a mathematical procedure for transforming inputs into desired outputs using highly connected networks of relatively simple processing units called neurons (Johnson and Wichern, 2002). Each neuron performs a mathematical operation that produces an output which is a function (usually non-linear) of a series of biased and weighted inputs coming from other neurons. The underlying principle is that neural networks be modeled after the neural activity in the human brain, in which the interconnection of very simple functional units (the neurons) can solve many complex and non-linear problems in a very fast way. Thus, ANNs consist on parallel computational

models comprised of densely interconnected adaptive processing units (the neurons). These networks provide a tool for modelling nonlinear static or dynamic systems, making use of their adaptive nature, based on "learning by example". This feature makes this kind of computational models very appealing in applications in which the understanding of the problem is little or incomplete, but where training data is readily available (Hassoun, 1995).

ANNs can provide very fast outputs, as it only has to compute a limited number of very simple operations that are easily parallelizable. However, the design and training of the network can be a hard task. ANNs are set in layers of interconnected networks, in which the first layer has as many neurons as inputs in the system, and the output layer has as many neurons as required outputs. All intermediate layers are called hidden layers, and can have any number of neurons (Johnson and Wichern, 2002).

For our design, many different architectures were trained and evaluated to obtain the network with the best performance. In the end, a three-layer network with 33 neurons on the hidden layer was chosen. The input layer was configured with 33 neurons each corresponding to determined spectral positions. The neurons where configured with log-sigmoid transfer functions. The output layer consisted on 17 outputs, each corresponding to one of the sulfates (where each output should take values between 0 and 1, proportional to the abundance of the sample).

As suggested by (Koujelev et al., 2010), and in order to improve the ANN model performance, only a selection of spectral positions is fed into the network as input; the ν_1 peak positions of all the sulfates, as well as the most intense non-overlapping secondary peaks of the pure samples were selected as inputs. This allows the definition of a predefined set of spectral positions that will be extracted from the input spectra and fed to the network. In order to determine these peaks, the input spectra to the ANN have to be pre-preprocessed

so that only the most intense peaks have non-zero intensities, as depicted in Figure 2: from all the inputs to the model, those below a determined threshold will be set to 0. The definition of this threshold depends on the sample noise and can be decided for each spectrum. In addition, a deconvolution of peaks is needed for mixtures where the peaks are partially overlapped.

The training process was performed using a Levenberg-Marquardt back-propagation algorithm (Levenberg, 1944; Marquardt, 1963) that used spectra of pure sulfates (training set) plus some spectra of mixtures with 0.25:0.75, 0.5:0.5 and 0.75:0.25 proportions. The network was trained to provide outputs proportional to the abundance of each sulfate. To avoid over-fitting, the early-stopping technique with the validation set was used. This consists in stopping the iterative training process when the output errors for the validation set increase. Finally, to test the network in a more representative scenario, the test set from the RLS instrument was fed to the network.

RESULTS AND DISCUSSION

PCA

The training of the PCA model showed that the first three components PC1, PC2 and PC3 explain more than 80% of the variance of the training data set, and 90% if the first five components are considered (Fig. 3). This means that 90% of the variance of the system is explained with only 5 of the calculated orthogonal variables. A dendrogram showing the interconnections for the 3 PCs model is depicted in Figure 4. Models with higher number of components (up to ten) do not present better separation among different cations, and worsen the discrimination between low- and high-hydration sulfates. In addition, the lower the number of components, the more general the model can be, so the 3-PC model was selected as the optimal one. The representation of the scores of the training samples in the

new variable system (with only the first two components, for representation convenience) is displayed in Figure 5 as circles. From the validation set, only the calculated scores for mixed spectra of sulfates of the same cation in 50:50 proportions (worst case) are plotted in Figure 5 as triangles. The mixed spectra contain sulfates with different degree of hydration. The scores in Figure 5 show how PCA succeeds in separating, mostly along PC1, the low hydration mixtures from the high hydrated ones, though it fails to distinguish among different types of cations. For example, the model scores the mixture of Fe-sulfates with 4 and 7 water molecules between those two elements, and the same happens with the Ca-sulfates. However, the model fails to correctly separate the Mg-sulfates by hydration level. Therefore, the general conclusion is that PC1 can only be used to separate low hydration from high hydration sulfates, but not to distinguish among different hydration states of same-cation sulfates.

While no additional direct associations between the principal components and the physical properties of the different molecules (as for example the cation ratio) can be inferred, the PCA analysis of sulfates can provide useful information when representing the PC1-PC2 scores. For example, the representation of the scores of the test spectra set is depicted in Figure 6 (for graphical simplicity, only two PCs are represented). This figure shows how the mixtures are placed between the pure components depending on their relative abundance: the higher the abundance of a sulfate of the mixture, the closer to the corresponding pure sulfate score. This would mean that some kind of quantification could be possible based on the PCA model, even when only trained with pure Raman spectra.

PLS

The PLS regression of the pure sulfates spectra to their expected weight cation atomic fractions and water presence was performed for different numbers of components. To

decide the optimal number of components, the Mean Squared Error (MSE) of the predictions for the training, validation and tests sets was calculated. Figure 7 shows some spectra from the training set and the corresponding loadings, where the energies that have the greatest effect on the PLS predictions can be observed. The application of the model to the spectra sets yielded the prediction errors shown in Figure 8 – a. As expected, the prediction error for the training samples (blue line) tends to 0 with increasing number of components, as the model is fitted better. The values of the error for the regression of the test set of spectra (black lines) indicates that the best fitting (minimum prediction error) occurs for a 12-component PLS model. As a general rule, lower numbers of components imply a more general response of the model. Since the spectra of the natural samples were acquired with a different experimental setup than the spectra of the training and validation sets, we interpret the 13-component model (where the minimum prediction error for the validation set is found – see red line) as one in which the 13^{th} component accounts for the spectrometer response. Thus, the 12-component model is considered as the optimum for our set of sulfates spectra; more than 98% of the variance of the system can be explained with this model (Fig. 8 – b), apparently with no over-fitting.

The average absolute prediction error values for the training, validation and test sets are presented in Table 2 for the 12-component model. This table shows how the average error is close to 0 for the training and validation sets, while the model prediction is biased in the prediction values for the test set spectra. The RMSEP (Root Mean Square Error of Prediction) for each test is presented in Table 3. This value represents the prediction error deviation, to compare the prediction accuracy between the different spectra sets. In addition, the predicted vs. expected responses correlations are shown in Table 4, and the responses represented in Figure 9. This figure presents the expected vs. calculated cation

ratios for each spectrum. As each spectrum is represented as one individual point in the graph, the estimation error is defined by the y-axis displacement from the expected value. Thus, the estimation error is 0 when the point is placed on the line with unitary slope (y = x). These data show that a correlation between the predicted and the expected responses is present with this model for all the training, validation and test spectra sets.

The RMSEP results for the training set in Table 3 show that the model has the lower prediction capabilities for the hydration response, which can be explained by the influence of the many close-to-zero values of these responses, as shown in the plot of the training responses in Figure 9 – a. It is important to note how the results of the test samples provide too high values of the iron cation response, especially in Figure 9 – d, when it should always be 0.

For the validation spectra, the prediction accuracy is lower (higher RMSEP) than for the training spectra, as expected, which is also reflected in its lower correlation and higher prediction bias. The model behavior is worse for well-balanced mixtures (0.5:0.5) than for mixtures with unbalanced proportions, as the model was trained with pure samples only. Most of the outlier points observed in Figure 9 – b belong to the mixtures in this proportion. The values in Table 4 for the validation set correspond to the averaged correlation for all the proportions, but these improve between 0.5% and 1% when the 0.5:0.5 mixture is not considered.

The results for the test spectra show better correlations than the training set in the Na-sulfates case, though the RMSEP value is much higher than for the training and validation sets. This can be explained by the fact that the correlation value for the training samples is biased with all the spectra from the rest of the sulfates, which are not represented with the test samples. This can be readily observed in Figure 9 – a, b vs. Figure 9 – c, d; the slopes

of the curves are more or less unitary for the training and validation tests, which indicates a certain accuracy of the model (and which is reflected in similar RMSEP values and low bias), while the test sets show a non-unitary slope linear correlation between the expected and calculated sodium response (with a much higher RMSEP value and prediction bias). This implies that the 12-component PLS model seems to fail to directly predict hydration states and cation abundances for this case. However, it still provides linear calibration curves that could be used to compute these values.

ANN

The training process of our ANN consisted on providing outputs proportional to the abundance of the materials. As ANN can model any non-linear function, this seemed to be the most convenient way to do it, contrary to PLS, where the underlying physical principles tried to be modeled.

In order to evaluate the identification accuracy of the ANN, we established the criteria to consider that a sulfate is detected when the corresponding output is higher than a determined threshold. For this network, this threshold was set to 0.06. As the output of the network ranges from 0 to 1, this value will be the theoretically lowest detection threshold for this model (i.e. no sulfates will be detected below concentrations of 6%).

Under these premises, the training set spectra were detected with 100% accuracy with this ANN. Furthermore, the major phase present in all the samples of validation and test sets was also detected in 100% of the cases. Both phases of the mixtures were detected with 100% accuracy only when the minor phase was present with at least 10% abundance. In other words, the ANN detects 100% of the minerals present in binary mixtures with proportions ranging from 10:90 to 90:10. This implies that the ANN model provides a robust system for the qualitative detection of sulfates in this kind of mixtures, with a

384 detection threshold for minor phases of around 10%, even for spectra acquired with a
385 different hardware setup than the training spectra.

386 The representation of the ANN outputs for spectra of mixtures with different
387 proportions (validation and test sets) shows that the network outputs also provide
388 information on the relative abundance of the materials. As part of the validation set was
389 used to train the model, the results for this set proved very good also in terms of
390 quantification of mineral abundance, as expected. More interesting are the results for the
391 test spectra, which are shown in Figures 10 and 11, where the estimated concentration
392 values from the ANN with respect to the expected ones are represented. These results show
393 that the model accuracy might somehow depend on the samples (e.g., the results for the
394 mixture in Fig. 11 show that the Mg-sulfate concentration tends to be underestimated for
395 mixtures where it is the major component). However, a certain degree of correlation
396 between the modeled values and the actual abundances of the mixtures exists. This is an
397 interesting result, especially bearing in mind that this is true for the test set spectra, while
398 the model was trained with pure spectra obtained with a different spectrometer (training
399 set) and computed spectra of mixtures calculated from those samples.

400 The representation of the maximum and minimum values of the outputs of the ANN
401 which do not correspond to the minerals present in the spectra can be seen in Figures 10
402 and 11 – b. The representation of these values, which should always be 0, is of relevance to
403 show that, with a threshold of 6%, no false identifications are obtained. The conclusion is
404 that the use of ANN looks promising for providing robust and reliable results under the
405 described premises, not only for identifying the phases present in binary mixtures of
406 sulfates, but also to provide some kind of quantification of their abundance.

MVAT comparative

PCA, PLS and ANN models have been trained based-only in 17 spectra of pure sulfates and mixed spectra computed as linear combinations of those. This procedure has allowed evaluating these analytical techniques without the need to actually prepare all possible combinations of samples for the calibration of the models.

The analysis of selected regions from the Raman spectrum implies the analysis of several thousands of variables at the same time. To deal with this amount of information, PCA and PLS calculate new sets of orthogonal variables as linear combinations of the original ones. PLS then regresses these variables to expected responses, while PCA doesn't. These new variables correspond to latent variables which ideally should be directly related to physical properties of the system (e.g., the degree of hydration of a sulfate). However, this is not always the case. This is probably due to the non-linear nature of the Raman emission, which PCA and PLS try to model with linear processing. ANNs, on the other hand, can provide non-linear transfer functions. Thus, they might be more adequate for the modeling of non-linear effects (as the Raman emission). However, ANNs require a relatively low number of input variables to provide any relevant results. This technique does not perform a reduction to latent variables on its own, so it has been made by only inputting the most relevant spectral positions (corresponding to the wavenumbers of the most representative peaks of the training spectra).

We have shown the ability of these MVAT models of providing useful qualitative and even quantitative information for simple binary mixtures, even when the training and testing were performed with spectra recorded with different hardware setups. As discussed in the previous sections, PCA separated low from high hydrated sulfates, and also somewhat classified samples of mixtures depending on their relative abundance. PLS

model outputs presented good correlations to the expected responses. However, though well correlated, the responses in some cases were relatively far from the expected values. The conclusion is that PCA and PLS provided a classification method which needs a previous calibration for the sample under analysis, and a classification method. On the other hand, the ANN model outputs directly provided the abundance of the corresponding salt, in addition to a 100% qualitative detection for mixtures with abundances as low as 10%.

To overcome the various limitations of these MVATs, a synergy between them might be interesting to improve the overall performance of the models. Some classifiers for the qualitative analysis of minerals have been proposed based on integrated PCA and ANN models (Dorfer et al., 2010; Ishikawa and Gulick, 2013). Future research will thus focus in developing models for the quantification of mineral abundances which integrate different MVATs.

IMPLICATIONS FOR EXOMARS

ExoMars' RLS instrument will determine the structural and compositional features of materials in rocks and soils at the surface and subsurface of Mars. The ExoMars samples will be collected by a drill, then crushed and delivered to a suite of instruments located in the rover's analytical laboratory, where the RLS instrument sits. A crushing station will provide homogenized powdered samples that will likely feature complex mixtures of mineral phases.

Fast, robust, unsupervised RLS data processing tools able to interpret the intricate sprectra that will be obtained from Martian samples would benefit ExoMars mission operations in that they may directly support the daily tactical operations of the rover. The different multivariate analysis techniques methodologies – PCA, PLS, and ANN – we have

455 discussed here promise to provide an efficient way to process RLS data during the ExoMars
456 mission.

457 Future work in this direction will focus on exploring the capabilites of these
458 methodologies to evaluate more complex mixtures; these will include additional synthetic
459 sulfates, oxides, clays, phyllosilicates, carbonates, perchlorates, as well as natural samples.

460 We will carry out this research using the RLS ExoMars Simulator we have developed
461 at the University of Valladolid-Centro de Astrobiologia and the different RLS models and
462 prototypes that we are developing with the Spanish National Institute for Aerospace
463 Technology (INTA).

464 **ACKNOWLEDGMENTS**

465 We kindly thank A. Wang for providing the Raman spectra of the pure sulfate
466 samples used in this work and A. Koujelev for the helpful insights on the implementation of
467 ANN. R. Leveille and the Canadian Space Agency facilitated this project. Preliminary
468 results of this work have been reported at the 15th Annual Conference of the International
469 Association for Mathematical Geosciences (Lopez-Reyes, 2013b). GLR acknowledges the
470 University of Valladolid (Spain) for providing funding for the project. PS and CL
471 acknowledge support from the Natural Sciences and Engineering Research Council of
472 Canada (NSERC) and the Canadian Space Agency.

473 **REFERENCES**

474 Angel, S.M., Gomer, N.R., Sharma, S.K., and McKay, C. (2012) Remote Raman
475 spectroscopy for planetary exploration: A review. Applied Spectroscopy, 66(2), 137-150.

476 Bonny, S., and Jones, B. (2003) Microbes and mineral precipitation, Miette Hot Springs,
477 Jasper National Park, Alberta, Canada. Canadian Journal of Earth Sciences, 40(11), 1483-
478 1500.

479 Bower, D.M., Steele, A., Fries, M.D., and Kater, L. (2013) Micro Raman spectroscopy of
480 carbonaceous material in microfossils and meteorites: improving a method for life
481 detection. Astrobiology, 13(1), 103-113.

482 Chou, I.M., Seal, R.R., and Wang, A. (2013) The stability of sulfate and hydrated sulfate
483 minerals near ambient conditions and their significance in environmental and planetary
484 sciences. Journal of Asian Earth Sciences, 62, 734-758.

485 de Jong, S. (1993) SIMPLS: An alternative approach to partial least squares regression.
486 Chemometrics and Intelligent Laboratory Systems, 18(3), 251-263.

487 Dickensheets, D.L., Wynn-Williams, D.D., Edwards, H.G.M., Schoen, C., Crowder, C.,
488 and Newton, E.M. (2000) A novel miniature confocal microscope/Raman spectrometer
489 system for biomolecular analysis on future Mars missions after Antarctic trials. Journal of
490 Raman Spectroscopy, 31(7), 633-635.

491 Dolenko, T.A., Burikov, S.A., and Sugonjaev, A.V. (2005) Neural network technologies in
492 Raman spectroscopy of water solutions of inorganic salts. In H.J. Byrne, E. Lewis, B.D.
493 MacCraith, E. McGlynn, J.A. McLaughlin, G.D. Osullivan, A.G. Ryder, and J.E. Walsh,
494 Eds. Opto-Ireland 2005: Optical Sensing and Spectroscopy, 5826, p. 298-305.

Dorfer, T., Schumacher, W., Tarcea, N., Schmitt, M., and Popp, J. (2010) Quantitative mineral analysis using Raman spectroscopy and chemometric techniques. Journal of Raman Spectroscopy, 41(6), 684-689.

Downs, R.T. (2006) The RRUFF Project: an integrated study of the chemistry, crystallography, Raman and infrared spectroscopy of minerals. Program and Abstracts of the 19th General Meeting of the International Mineralogical Association in Kobe, Japan., O03-13.

Edwards, H.G.M., Hutchinson, I., and Ingley, R. (2012) The ExoMars Raman spectrometer and the identification of biogeological spectroscopic signatures using a flight-like prototype. Analytical and Bioanalytical Chemistry, 404(6-7), 1723-1731.

Edwards, H.G.M., Hutchinson, I.B., Ingley, R., Waltham, N.R., Beardsley, S., Dowson, S., and Woodward, S. (2011) The search for signatures of early life on Mars: Raman spectroscopy and the Exomars mission. Spectroscopy Europe, 23(1).

Edwards, H.G.M., Newton, E.M., Dickensheets, D.L., and Wynn-Williams, D.D. (2003) Raman spectroscopic detection of biomolecular markers from Antarctic materials: evaluation for putative Martian habitats. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 59(10), 2277-2290.

Ellery, A., and Wynn-Williams, D. (2003) Why Raman spectroscopy on Mars?--a case of the right tool for the right job. Astrobiology, 3(3), 565-79.

Fernandez-Remolar, D.C., Preston, L.J., Sanchez-Roman, M., Izawa, M.R.M., Huang, L., Southam, G., Banerjee, N.R., Osinski, G.R., Flemming, R., Gomez-Ortiz, D., Prieto

Ballesteros, O., Rodriguez, N., Amils, R., and Dyar, M.D. (2012) Carbonate precipitation under bulk acidic conditions as a potential biosignature for searching life on Mars. Earth and Planetary Science Letters, 351, 13-26.

Foucher, F., Lopez-Reyes, G., Bost, N., Rull, F., Rüßmann, P. and Westall, F. (2012) Effect of grain size distribution on Raman analyses and the consequences for in situ planetary missions. Journal of Raman Spectroscopy.

Ghesti, G.F., de Macedo, J.L., Resck, I.S., Dias, J.A., and Dias, S.C.L. (2007) FT-Raman spectroscopy quantification of biodiesel in a progressive soybean oil transesterification reaction and its correlation with H-1 NMR spectroscopy methods. Energy & Fuels, 21(5), 2475-2480.

Hassoun, M. H., (1995) Fundamentals of Artificial Neural Networks. MIT Press.

Hermosilla Rodriguez, I., Lopez-Reyes, G., Llanos, D.R., and Rull Perez, F. (2014) Automatic raman spectra processing for Exomars. In E. Pardo-Igúzquiza, C. Guardiola-Albert, J. Heredia, L. Moreno-Merino, J.J. Durán, and J.A. Vargas-Guzmán, Eds. Mathematics of Planet Earth, p. 127-130. Springer Berlin Heidelberg.

Ishikawa, S.T., and Gulick, V.C. (2013) An automated mineral classifier using raman spectra. Computers & Geosciences, 54(0), 259-268.

Johnson, R.A., and Wichern, D.W. (2002) Applied multivariate statistical analysis. Prentice hall Upper Saddle River, NJ.

Josset, J.-L., Westall, F., Hofmann, B.A., Spray, J.G., Cockell, C., Kempe, S., Griffiths, A.D., De Sanctis, M.C., Colangeli, L., Koschny, D., Pullan, D., Föllmi, K., Diamond, L.,

Josset, M., Javaux, E., Esposito, F., and Barnes, D. (2012) CLUPI, a high-performance imaging system on the ESA-NASA rover of the 2018 ExoMars mission to discover biofabrics on Mars. EGU General Assembly Conference Abstracts, 14, p. 13616.

King, P.L., and McLennan, S.M. (2010) Sulfur on Mars. Elements, 6(2), 107-112.

Knoll, A.H., Carr, M., Clark, B., Farmer, J.D., Fischer, W.W., Grotzinger, J.P., McLennan, S.M., Malin, M., Schroder, C., Squyres, S., Tosca, N.J., and Wdowiak, T. (2005) An astrobiological perspective on Meridiani Planum. Earth and Planetary Science Letters, 240(1), 11-11.

Kong, W.G., and Wang, A. (2010) Planetary Laser Raman Spectroscopy for surface exploration on C/D-type asteroids: A case study. 41st Lunar and Planetary Science Conference, p. Abstract #2730. Lunar and Planetary Institute, Houston.

Koujelev, A., Sabsabi, M., Motto-Ros, V., Laville, S., and Lui, S.L. (2010) Laser-induced breakdown spectroscopy with artificial neural network processing for material identification. Planetary and Space Science, 58(4), 682-690.

Kriesten, E., Alsmeyer, F., Bardow, A., and Marquardt, W. (2008) Fully automated indirect hard modeling of mixture spectra. Chemometrics and Intelligent Laboratory Systems, 91(2), 181-193.

Lafuente, B., Sansano, A., Navarro, R., Rull, F., Martínez Frías, J., Medina, J., Lopez, G., Sobron, P. (2012) Multivariate analysis of Raman spectra for geological classification and identification: Application to Exomars mission. GeoRaman 2012, Nancy.

557 Lambert, J.L., Morookian, J., Roberts, T., Polk, J., Smrekar, S., Clegg, S.M., Weins, R.C.,
558 Dyar, M.D., and Treiman, A. (2010) Standoff LIBS and Raman spectroscopy under venus
559 conditions. 41st Lunar and Planetary Science Conference, p. Abstract #2608. Lunar and
560 Planetary Institute, Houston.

561 Levenberg, K. (1944) A method for the solution of certain problems in least squares.
562 Quarterly of Applied Mathematics, 2, 164-168.

563 Ling, Z.C., Wang, A., Jollif, B.L., Arvidson, R.E., and Xia, H.R. (2008) A systematic
564 Raman, Mid-IR, and Vis-NIR spectroscopic study of ferric sulfates and implications for
565 sulfates on Mars. 39th Lunar and Planetary Science Conference, p. Abstract #1463. Lunar
566 and Planetary Institute, Houston.

567 Ling, Z.C., Wang, A., Jolliff, B.L., Li, C., Liu, J., Bian, W., Ren, X., Mu, L.L., and Su, Y.
568 (2009) Raman spectroscopic study of quartz in lunar soils from Apollo 14 and 15 missions.
569 40th Lunar and Planetary Science Conference, p. Abstract #1823. Lunar and Planetary
570 Institute, Houston.

571 Lopez-Reyes, G., Rull, F., Venegas, G., Westall, F., Foucher, F., Bost, N., Sanz, A., Catalá-
572 Espí, A., Vegas, A., Hermosilla, I., Sansano, A., and Medina, J. (2013a) Analysis of the
573 scientific capabilities of the ExoMars Raman Laser Spectrometer instrument. European
574 Journal of Mineralogy, 25(5), 721-733.

575 Lopez-Reyes, G., Sobron, P., Lefevbre, C. and Rull, F. (2013b) Application of multivariate
576 analysis techniques for the identification of sulfates from Raman spectra. 15th International
577 Association for Mathematical Geosciences Conference, Madrid.Marquardt, D.W. (1963)

578 An algorithm for least-squares estimation of nonlinear parameters. Journal of the Society
579 for Industrial & Applied Mathematics, 11(2), 431-441.

580 Martens, H., and Naes, T. (1992) Multivariate Calibration. Wiley.

581 Nixon, S.L., Cockell, C.S., and Cousins, C.R. (2013) Plausible microbial metabolisms on
582 Mars. Astronomy & Geophysics, 54(1), 13-16.

583 Özbalci, B., Boyaci, İ.H., Topcu, A., Kadılar, C., and Tamer, U. (2013) Rapid analysis of
584 sugars in honey by processing Raman spectrum using chemometric methods and artificial
585 neural networks. Food Chemistry, 136(3–4), 1444-1452.

586 Perez-Pueyo, R., Soneira, M.J., and Ruiz-Moreno, S. (2004) A fuzzy logic system for band
587 detection in Raman spectroscopy. Journal of Raman Spectroscopy, 35(8-9), 808-812.

588 Rull, F., Lopez, G., Catala, A., Medina, J., Sansano, A., Sanz, A., Sobron, F. (2011) Raman
589 spectroscopy analysis on powdered samples inside the Exomars mission. ESA Congrex
590 Lisbon 2011.

591 Rull, F., Maurice, S., Diaz, E., Tato, C., Pacros, A., and Team., t.R. (2011a) The Raman
592 Laser Spectrometer (RLS) on the ExoMars 2018 Rover Mission. 42nd Lunar and Planetary
593 Science Conference LPSC, p. Abstract #2400. Lunar and Planetary Institute, Houston.

594 Rull, F., Sansano, A., Díaz, E., Canora, C.P., Moral, A.G., Tato, C., Colombo, M.,
595 Belenguer, T., Fernández, M., Manfredi, J.A.R., Canchal, R., Dávila, B., Jiménez, A.,
596 Gallego, P., Ibarmia, S., Prieto, J.A.R., Santiago, A., Pla, J., Ramos, G., Díaz, C., and
597 González, C. (2011b) ExoMars Raman laser spectrometer for Exomars. Society of Photo-
598 Optical Instrumentation Engineers (SPIE), Conference Series, 8152, 12.

599 Sanchez-Andrea, I., Rojas-Ojeda, P., Amils, R., and Luis Sanz, J. (2012) Screening of
600 anaerobic activities in sediments of an acidic environment: Tinto River. Extremophiles,
601 16(6), 829-839.

602 Schumacher, W., Kuehnert, M., Roesch, P., and Popp, J. (2011) Identification and
603 classification of organic and inorganic components of particulate matter via Raman
604 spectroscopy and chemometric approaches. Journal of Raman Spectroscopy, 42(3), 383-
605 392.

606 Sharma, S.K., Lucey, P.G., Ghosh, M., Hubble, H.W., and Horton, K.A. (2003) Stand-off
607 Raman spectroscopic detection of minerals on planetary surfaces. Spectrochimica Acta Part
608 A: Molecular and Biomolecular Spectroscopy, 59(10), 2391-2407.

609 Sobron, P., Sobron, F., Sanz, A., and Rull, F. (2008) Raman signal processing software for
610 automated identification of mineral phases and biosignatures on Mars. Applied
611 Spectroscopy, 62(4), 364-370.

612 Sobron, P., Sanz, A., Acosta, T., and Rull, F. (2009) A Raman spectral study of stream
613 waters and efflorescent salts in Rio Tinto, Spain. Spectrochimica Acta part A: Molecular
614 and Biomolecular Spectroscopy, 71(5), 1678-82.

615 Sobron, P., and Alpers, C.N. (2013a) Raman spectroscopy of efflorescent sulfate salts from
616 iron mountain mine superfund site, California. Astrobiology, 13(3), 270-8.

617 Sobron, P., Bishop, J., Blake, D., Chen, B. and Rull, F. (2014) Natural Fe-bearing oxides
618 and sulfates from the Rio Tinto Mars analogue – Critical assessment of VNIR reflectance

spectroscopy, laser Raman spectroscopy, and XRD as mineral identification tools. American Mineralogist, This volume.

Sobron, P., Lefebvre, C., Koujelev, A., and Wang, A. (2013b) Why Raman and LIBS for exploring icy moons? 44th Lunar and Planetary Science Conference, p. Abstract #2381. Lunar and Planetary Institute, Houston.

Steele, A., McCubbin, F.M., Fries, M., Glamoclija, M., Kater, L., and Nekvasil, H. (2010) Graphite in an Apollo 17 impact melt breccia. Science, 329(5987), 51-51.

Vagenas, N.V., Kontoyannis, C. G. (2003) A methodology for quantitative determination of minor components in minerals based on FT-Raman spectroscopy - The case of calcite in dolomitic marble. Vibrational Spectroscopy, 32(2), 261-264.

Wang, A., Freeman, J.J., Jolliff, B.L., and Chou, I.M. (2006) Sulfates on Mars: A systematic Raman spectroscopic study of hydration states of magnesium sulfates. Geochimica Et Cosmochimica Acta, 70(24), 6118-6135.

Wang, A., Haskin, L.A., Lane, A.L., Wdowiak, T.J., Squyres, S.W., Wilson, R.J., Hovland, L.E., Manatt, K.S., Raouf, N., and Smith, C.D. (2003) Development of the Mars microbeam Raman spectrometer (MMRS). Journal of Geophysical Research, 108(E1), 5005.

Wang, A., Zheng, M.P., Kong, F.J., Ling, Z.C., Kong, W.G., Sobron, P., and Jolliff, B.L. (2011) A Low T, High RH, and potentially life-friendly environment within the martian salt-rich subsurface in equatorial regions. 42nd Lunar and Planetary Science Conference, p. Abstract #2049. Lunar and Planetary Institute, Houston.

Wiens, R.C., Sharma, S.K., Clegg, S.M., Misra, A.K., and Lucey, P.G. (2007) Combined remote Raman spectroscopy and LIBS instrumentation for Mars astrobiology exploration. Seventh International Conference on Mars, p. Abstract #3092. Lunar and Planetary Institute, Houston.

Wold, S., Martens, H., and Wold, H. (1983) The multivariate calibration problem in chemistry solved by the PLS method. Lecture Notes in Mathematics, 973, 286-293.

Wynn-Williams, D.D., and Edwards, H.G.M. (2000) Proximal analysis of regolith habitats and protective biomolecules in situ by Laser Raman Spectroscopy: Overview of terrestrial antarctic habitats and Mars analogs. Icarus, 144(2), 486-503.

650

651

Hydration state	Mg	Ca	Fe	Na
Anhydrous	Anhydrous Mg-Sulfate	Anhydrite	--	Thenardite
1/2 H_2O	--	Bassanite	--	--
1 H_2O	Kieserite	--	Szomolnokite	--
2 H_2O	Sanderite	Gypsum	--	--
3 H_2O	Mg-sulfate tri-hydrate	--	--	--
4 H_2O	Starkeyite	--	Rozenite	--
5 H_2O	Pentahydrate	--	--	--
6 H_2O	Hexahydrate	--	--	--
7 H_2O	Epsomite	--	Melanterite	--
10 H_2O	--	--	--	Glauber's salt
11 H_2O	Meridianiite	--	--	--

652 Table 1. Sulfates used for the analysis with multivariate techniques

653

Response	Training set (pure sulfates)	Validation set (Mixed spectra)	Test set (Anhydrite + Thenardite)	Test set (Thenardite + $MgSO_4$)
Hydration ratio	-4.5^{-17}	-0.0087	0.0814	0.0509
Ca- ratio	3.3^{-17}	0.0027	-0.0249	0.0046
Mg- ratio	-1.9^{-17}	-0.0061	-0.0288	-0.0625
Fe- ratio	-1.3^{-17}	0.0021	-0.0737	-0.0211
Na- ratio	-8.1^{-18}	0.0075	0.1123	0.0962
Average	-1.1^{-17}	-0.0004	0.0132	0.0136

654 Table 2. Average prediction error values with the 12 component PLS model

655

656

657

658

659

Response	Training set (pure sulfates)	Validation set (Mixed spectra)	Test set (Anhydrite + Thenardite)	Test set (Thenardite + $MgSO_4$)
Hydration ratio	0.0788	0.0821	0.0901	0.0838
Ca- ratio	0.0072	0.0212	0.0327	0.0178
Mg- ratio	0.0109	0.0193	0.0356	0.0639
Fe- ratio	0.0134	0.0227	0.0897	0.0925
Na- ratio	0.0269	0.0236	0.1368	0.1280
Average	0.0274	0.0338	0.0770	0.0772

660 Table 3. RMSEP values (error variance) with the 12 component PLS model

661

662

663

Response	Training set (pure sulfates)	Validation set (Mixed spectra)	Test set (Anhydrite + Thenardite)	Test set (Thenardite + $MgSO_4$)
Hydration ratio	92.9%	89.5%	--	--
Ca- ratio	99.7%	97.9%	98.1%	--
Mg- ratio	98.7%	96.2%	--	98.4%
Fe- ratio	98.9%	97.1%	--	--
Na- ratio	93.9%	93.3%	98.9%	98.8%
Average	96.8%	94.8%	98.5%	98.6%

664 Table 4. Correlation values (in %) for the Calculated vs. Expected responses with the

665 12 component PLS model

666

667

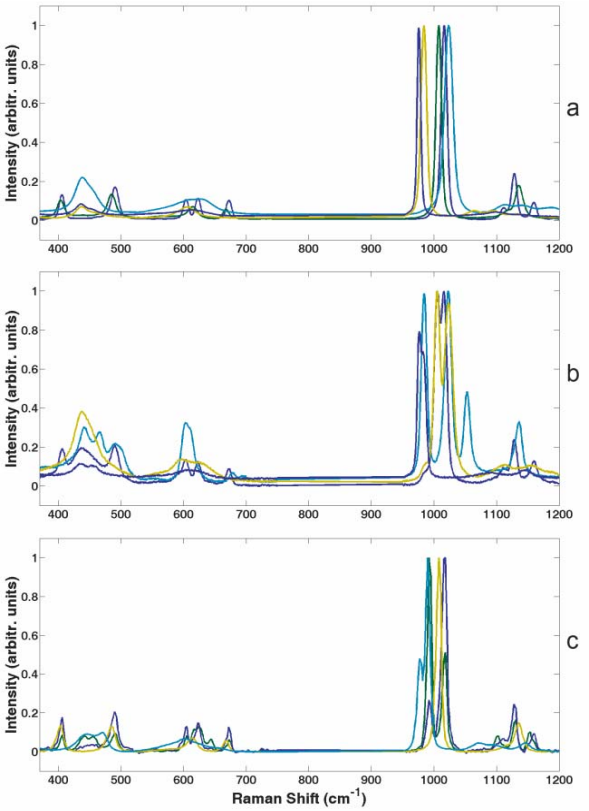


FIGURE 1

Figure 1. Example spectra from the training (a), validation (b) and test sets (c)

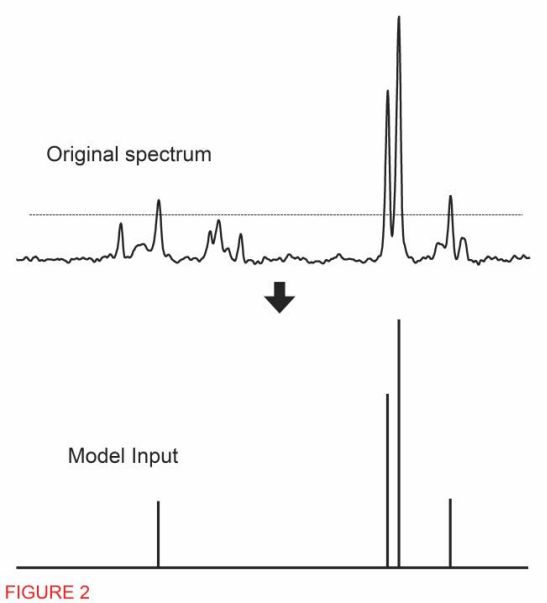


FIGURE 2

Figure 2. Spectrum pre-processing example for the ANN model

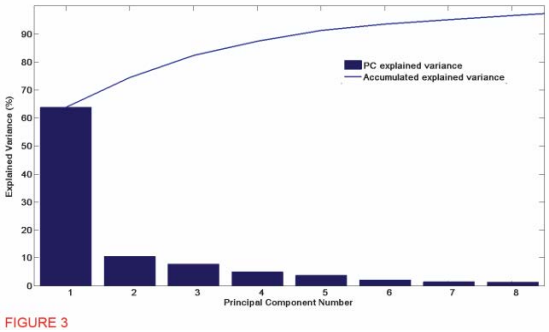


FIGURE 3

Figure 3. Explained variance of the data set (pure spectra of sulfates) after PCA analysis

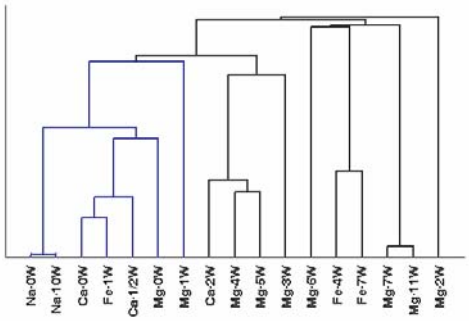


FIGURE 4

Figure 4. Dendrogram showing the classification of the PCA model with 3 PCs. As it can be seen, low-hydrated sulfates are contained in the blue branch of the dendrogram.

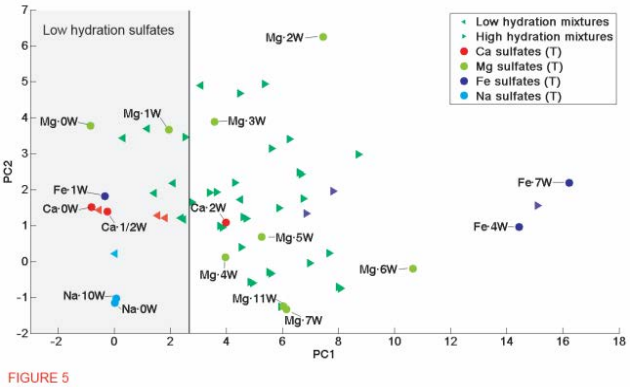
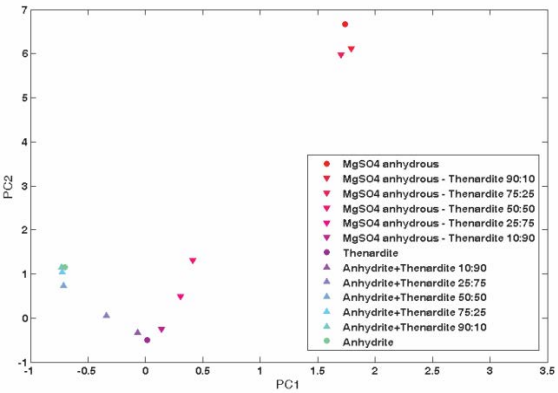


FIGURE 5

Figure 5. Representation of the scores for the training set (pure sulfates, circle) and 50:50 mixtures of the sulfates of the same cation from the validation set (triangles).

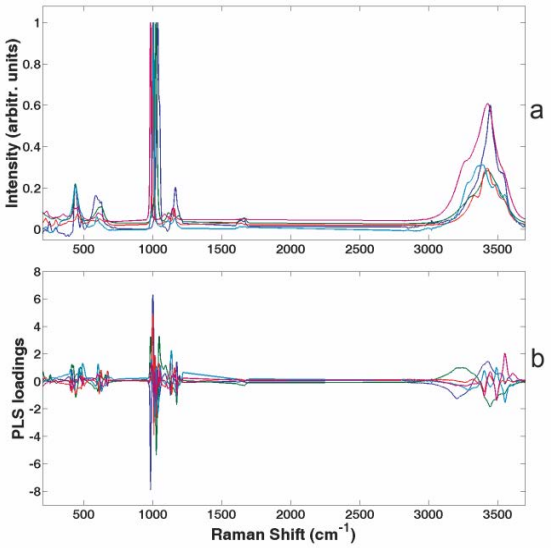
683 Different colors represent different cations. It can be seen how the low-hydration samples
684 and mixtures are found with lower values of PC1.



685 FIGURE 6

686 Figure 6. Scores for the averaged spectra of the different mixtures acquired with the
687 ExoMars Raman instrument. The scores are placed along lines between the endmembers at
688 a distance dependent on the mixture concentration.

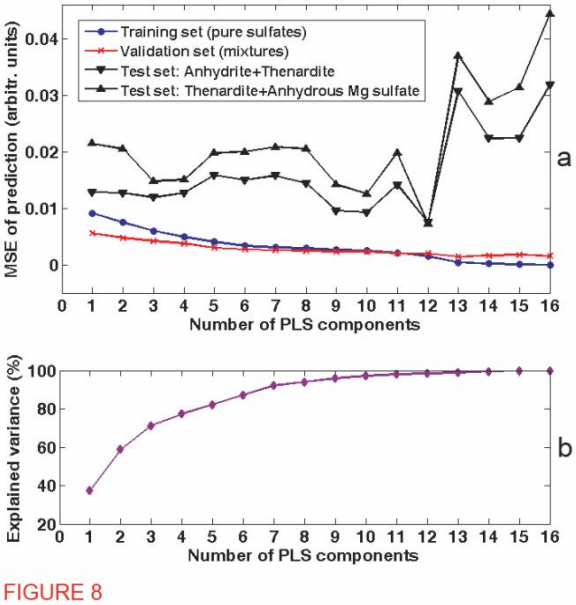
689



690 FIGURE 7

691 Figure 7. Some spectra from the training set (a) and corresponding PLS loadings (b)

692

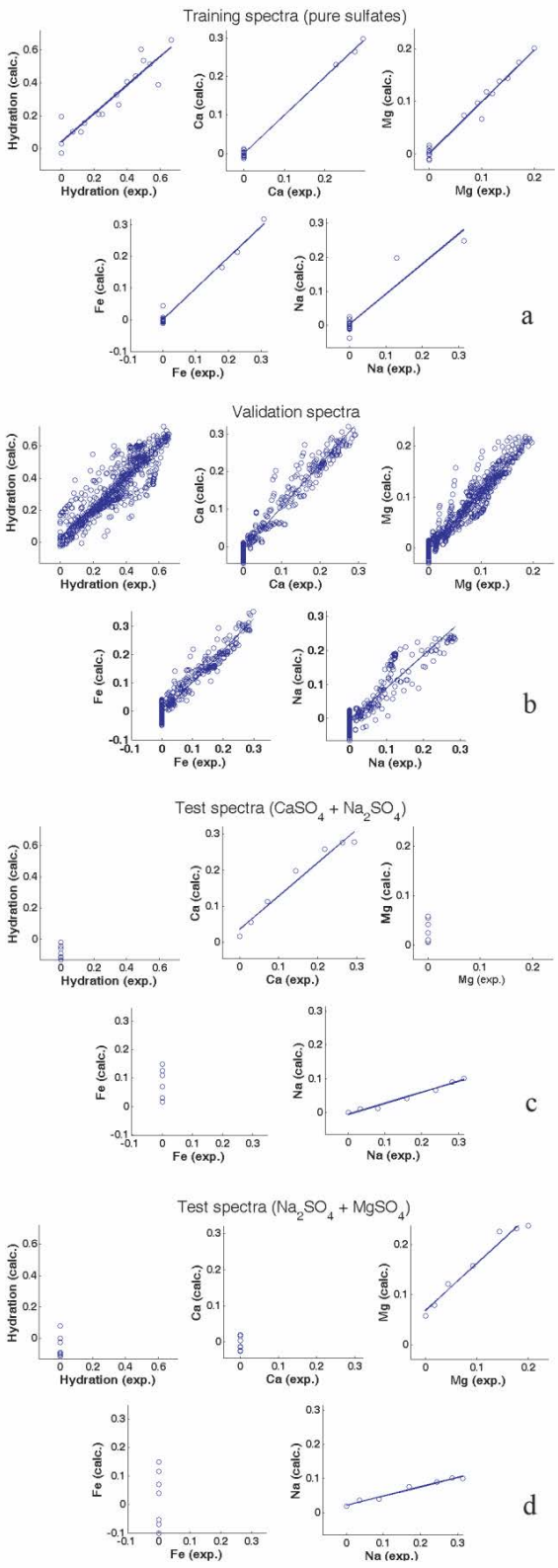


693 FIGURE 8

694 Figure 8. MSE of prediction (a), and accumulated explained variance (b) obtained by

695 the models with different numbers of components.

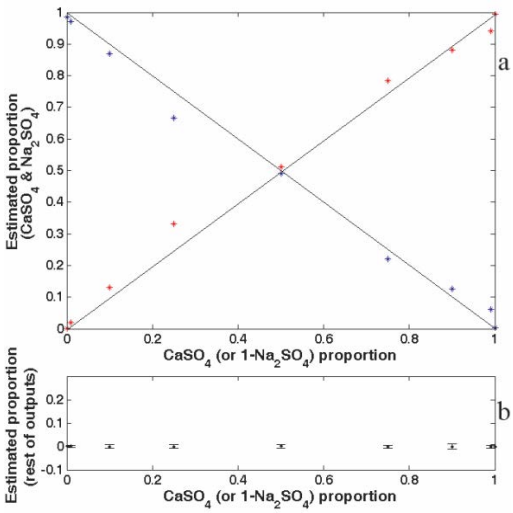
696



697 FIGURE 9

698 Figure 9. Calculated vs. Expected responses (cation weight ratio) for the training (a),
699 validation (b) and test (c,d) spectra for the 12 component PLS model

700



701 FIGURE 10

702 Figure 10. ANN outputs for the test set. Outputs corresponding to Anhydrite ($CaSO_4$
703 – red) and Thenardite (Na_2SO_4 – blue) (a). Outputs corresponding to the rest of sulfates (b).
704 Vertical axis represents the proportion estimated by the ANN, while horizontal axis
705 represents the expected (known) proportion.

706

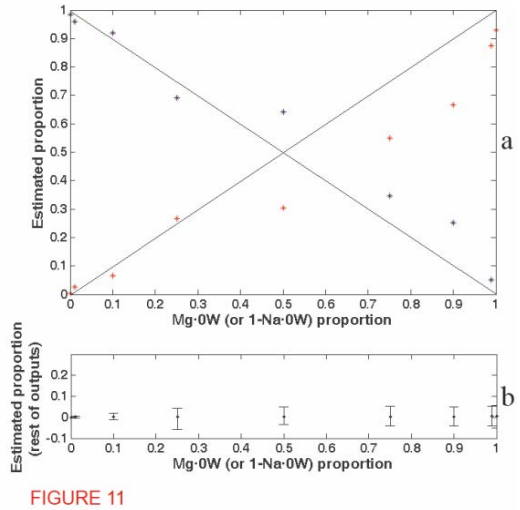


FIGURE 11

Figure 11. ANN outputs for the test set (mixtures in several proportions from RLS). Outputs corresponding to Thenardite (Na_2SO_4 – blue) and Mg-sulfate ($MgSO_4$ – red) (a). Outputs corresponding to the rest of sulfates (b). Vertical axis represents the proportion estimated by the ANN, while horizontal axis represents the expected (known) proportion.