Application of Multivariate Analysis Techniques for the identification of sulfates from Raman spectra: Implications for ExoMars

G. Lopez-Reyes¹, P. Sobron², C. Lefebvre², F. Rull¹

1. Unidad Asociada UVa-CSIC-Center of Astrobiology (Spain)
2. Space Science and Technology, Canadian Space Agency

ExoMars and the RLS instrument simulator
- The ExoMars rover includes a Raman Spectrometer (RLS instrument).
- The RLS instrument will acquire a set of ~30 spectra along a line from heterogeneous powdered samples.
- A simulator has been prepared to work in rover-like conditions
- Need for fast analytical tools which provide qualitative and quantitative information: MVT techniques

Multivariate Analysis Techniques (MVT)
They are powerful tools for data analysis, already applied in many fields
- Principal Component Analysis (PCA)
  - Computes new orthogonal variables by linear combinations of the original ones
- Partial Least Squares Regression (PLS)
  - Definition of output expected responses
  - Linear regression between input and output orthogonal variables
  - Over-fitting avoidance by Leave-One-Out Cross-Validation
- Artificial Neural Networks (ANN)
  - Non-linear
  - Several layers of interconnected neurons
  - Output is a function of neurons biases, transfer functions and weights

Principal Component Analysis (PCA)
Training with pure samples only
- 80% of variance with PC1+PC2
- Validation/test performed with pure samples and 1:1 mixtures
PCA differentiates hydrated from dehydrated sulfates

Partial Least Squares (PLS)
Training with pure samples only. Validation/test with mixtures.
- Model 1 (all inputs)
  - 7 components,
  - 93% correlation for pure sulfates
  - 89% for 1:1 mixtures
- Model 2 (selected inputs)
  - 4 components,
  - 99% correlation for pure sulfates
  - 95% for 1:1 mixtures

PLS predicts responses of pure sulfates and balanced mixtures of them

ExoMars rover

Spectral data sets
- 17 spectra of Fe-, Mg-, Ca- and Na-sulfates with different hydration states, and computed binary mixtures from linear combinations of them

Model input data
- PCA, and PLS (model 1): whole spectral range in which there are peaks (4004 inputs)
- PLS (model 2) and ANN: Sulfate main peak + non-overlapping secondary peaks (33 inputs)

Artificial Neural Network (ANN)
- Logsig transfer function
- Training and validation with pure samples + computed binary mixtures (25:75, 50:50, 75:25)
- Tested with pure samples + computed binary mixtures (5:95, 10:90, ..., 95:5)
- Verified with real spectra from the RLS simulator: mixture of Anhydrite (CaSO₄·0H₂O) + Thenardite (Na₂SO₄·H₂O)
- 33 input neurons corresponding to selected spectral positions
- 17 output neurons, each corresponding to one sulfate

Identification
- Outputs over 0.05 => Positive ID
- 100% accuracy for pure samples
- Fail ratio <3% for binary mixtures between 10:90 to 90:10

Quantification
- Raw output of the network provides a estimated proportion

ANN provides ID and quantification of pure and mixed sulfates with quite unbalanced proportions

Acknowledgements
This project was possible thanks to the support from R. Leveille. 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 (Spain) for providing funding for the project. PS and CL acknowledge support from NSERC.

References