Using Multivariate Statistical Analysis to fit spectroscopy data from remote and \textit{in situ} analysis of planetary surfaces: A Proof-of-Concept Assessment

Principal Investigator: Monica M. Grady
Project Manager: Jamie Bantock
Raman Spectroscopy Specialist: Feargus Abernethy
Project Scientist: Lily Hills

Final Report
UK Space Agency NSTP Contract No. G21_0100
29\textsuperscript{th} April 2022
This page intentionally left blank
# Table of Contents

1 Front Matter ...................................................................................................................... 4  
1.1 Document Change Log: ............................................................................................... 4  
1.2 Abbreviations and Acronyms....................................................................................... 4  
1.3 Applicable Documents: ............................................................................................... 4  
1.4 Acknowledgements:.................................................................................................... 4  
2 Executive Summary ............................................................................................................ 5  
3 Introduction ....................................................................................................................... 6  
4 Modification of Project ...................................................................................................... 6  
5 Sample Analysis ................................................................................................................. 7  
6 Statistical Analysis ............................................................................................................ 8  
7 Methodology Development and Testing ............................................................................ 8  
8 Results ............................................................................................................................. 10  
9 Follow-on Plan .................................................................................................................. 12
1 Front Matter

1.1 Document Change Log:

<table>
<thead>
<tr>
<th>Date</th>
<th>Version</th>
<th>Comment</th>
<th>Prepared by</th>
</tr>
</thead>
<tbody>
<tr>
<td>06/04/2022</td>
<td>1.0</td>
<td>Input from TN1 and TN2 included</td>
<td>M. Grady</td>
</tr>
<tr>
<td>15/04/2022</td>
<td>2.0</td>
<td>Input from TN3 and TN4 included</td>
<td>M. Grady</td>
</tr>
<tr>
<td>29/04/2022</td>
<td>3.0</td>
<td>FR completed</td>
<td>M. Grady</td>
</tr>
</tbody>
</table>

1.2 Abbreviations and Acronyms

- FR: Final review
- K/O: Kick Off
- MTR: Mid-Term Review
- NSTP: National Space Technology Programme
- OP: Oxia Planum (simulant)
- OU: Open University
- PAH: Polycyclic Aromatic Hydrocarbon
- PH: Phobos (simulant)
- SP4PS: Spectroscopy for Planetary Sciences
- TN: Technical Note
- UKSA: UK Space Agency
- WP: Work Package

1.3 Applicable Documents:

- [AD1] NSTP Space Science Announcement of Opportunity
- [AD2] Grady Application Form: Template A
- [AD3] Grady Application Form: Template B
- [AD4] Grant Agreement Contract (January 2022)
- [AD5] Contract Change Notice (March 2022)

1.4 Acknowledgements:

We are indebted to Ms Amy Dugdale and Ms Zoe Morland for generous provision of their planetary simulants of Oxia Planum and Phobos, respectively.
2 Executive Summary

We have undertaken a ‘Proof of Concept’ study to determine whether multivariate statistical analysis can be used to fit spectra from planetary surfaces to library data from minerals. The study was reduced in scope because of unavoidable delays to the start of the project. Despite this, most of the required deliverables were completed. The outcomes of the study are as follows:

- We have a library of minerals, organic compounds, terrestrial analogues and simulants with accompanying Raman spectroscopic data (Technical Note 1);
- Raman spectra were acquired from a simple set of minerals (the silicate mineral plagioclase, the organic compound thymine plus a sample of plagioclase doped with a dilute concentration of thymine) to develop the statistical methodology (Technical Note 2);
- Results from the test dataset showed that the most useful characteristics of the spectra to use for statistical analysis was the intensity of each point in a spectrum \(I_p\) normalised to the total intensity of the spectrum \(I_T\) (Technical Note 3);
- For intercomparison of spectra from different species, the spectral range was divided into 10 equal wavebands from 200 – 1200 wavenumbers (cm\(^{-1}\)). The average value of \(I_p/I_T\) across each waveband was calculated and used as the basis for statistical analysis of the spectra (Technical Note 3);
- Using the Principal Component Analysis (PCA) feature of the commercial statistics package, SPSS, we were able to separate the spectra from plagioclase, thymine and plagioclase + thymine into 3 components. Linear discriminant analysis (LDA) showed how well the three groups were separated (Technical Note 3);
- The methodology was repeated successfully on a simulant of the Mars surface (Oxia Planum Simulant, OPS), showing that its four main components were pyroxene, plagioclase, olivine and a clay mineral (represented by vermiculite) (Technical Note 4);
- We conclude that multivariate statistical analysis shows promise as a useful tool that can be used to describe spectra from planetary surfaces, matching them to combinations of mineral spectra. There is, however, much additional work still required to ensure that the methodology is robust (Technical Note 4).
3 Introduction

The project is a feasibility study to assess whether multivariate statistical analysis is an appropriate tool to deconvolute spectra from planetary surfaces for accurate matching to laboratory analogues. The project aims to advance understanding of the data processing required to facilitate interpretation of science data acquired by instrumentation either in orbit around a planetary body or on a lander or rover. The main activity within the project is data processing and algorithm development, developing a tool to enable multivariate statistical modelling of spectral data. An important part of this activity is the laboratory analysis of simulants, providing an internally self-consistent set of materials to act as calibration standards for the spectral modelling.

The project has a management Work Package (WP) and four task-based WP; each of the latter resulted in a Technical Notes (TN). The TN form part of this Final Report (FR).

4 Modification of Project

The original project was designed to last for 5 months (21 weeks, assuming 1 week break over Christmas), kicking off at the beginning of November 2021, as shown in the following chart:

Following delays in award of funds (UK Space Agency side) and acceptance of contract conditions (Open University side), kick-off did not occur until 26th January, leaving just over 9 weeks to complete the project. Clearly, the scope of the project had to be reduced – this was effected by focussing our efforts on Raman spectroscopic analysis of the materials. The UV-Vis spectral analysis of the same materials is, therefore, postponed until additional resource becomes available. The abbreviated timeline for the project is as shown below:
An additional effect of the compressed timescale was that Technical Notes (TN) 1 and 2 were combined, as were TN 3 and 4; because of staff resource issues, these notes are now being submitted at the same time as this Final Report.

5 Sample Analysis

Two types of material were selected for analysis: silicates and organic compounds. The silicates were natural materials drawn from the School of Physical Sciences research collections. The organic compounds were synthetic and purchased from Sigma-Aldrich Ltd. The selected samples are listed in Table 1; justification for their selection is contained in TN1, as is how they were prepared.

<table>
<thead>
<tr>
<th>Table 1: Samples selected for spectroscopic analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sample Type</strong></td>
</tr>
<tr>
<td>Mineral</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Terrestrial Analogues</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Meteorite</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Planetary Simulants</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Organic Compounds</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

The materials were analysed by Raman Spectroscopy, a technique used to determine the presence of a particular molecule or mineral in a sample. It can also identify the composition of a component. Identification of molecules depends on comparison of the sample spectrum with spectra from known materials. The frequency of each peak in a spectrum can be related to specific functional groups or atomic bonds (e.g., an aldehyde group or an Si-O bond), but it is
the combination of peaks from different bonds that lead to final identification. TN2 is a full description of the analytical technique, the instrument used and the experiment parameters.

6 Statistical Analysis

Chemometric modelling is the method of using statistical techniques to recognize patterns in data that might not otherwise be apparent. The specific application to be appraised by this project is that of multivariate analysis. This is a well-known statistical technique that is used for pattern recognition in sets of data that contain more than two variables. It is appropriate here because a spectrum from a single species is a dataset of several variables as a function of wavelength: peak position, intensity, peak width, area, etc., all of which may (or may not) shift with changes in composition. The beauty of multivariate analysis is that it allows a spectrum to be broken down into its principal components, which can then be matched with the principal components of known materials.

Originally, four widely-available commercial specialist statistical software packages (Matlab, Nvivo, SigmaPlot and SPSS) plus an add-in for Microsoft Excel (XLStat) were to be evaluated. However, because of time constraints, it was not possible to undertake this evaluation. The statistical software package SPSS was used to test the methodology because it was the most familiar to the project team – and it also had a fairly intuitive interface for data import and manipulation.

7 Methodology Development and Testing

Methodology development and testing took place in parallel. A fairly simple test data set of a silicate (plagioclase, a common component of planetary surfaces) mixed with a single organic compound (thymine) was selected. Thymine has a very distinctive Raman spectrum with strong peaks, several of which coincide with those from plagioclase, hence it was an ideal choice for the second component in the mixture. Figure 1 shows the logic chain for the final methodology.

![Figure 1: Logic chain for taking data through Principal Component Analysis](image-url)
The goal was to discriminate between different features in a spectrum so that a spectrum of mixed components could be broken down into the individual components of the mixture. Several parameter sets were tested: peak position, peak height (intensity), peak area, peak distribution and relative peak intensities. The final choice of parameters was a trade-off between selection of sufficient data to represent a spectrum and the time taken to deconvolute the spectra. Iteration of statistical analysis of the test data eventually resulted in selection of the average count intensity within a specific waveband normalised to the total number of counts across the entire spectrum.

A grid of nine spectra was acquired from a 50 x 50 µm area in a sample from each of the three sets of material. It also shows the wavebands eventually selected for statistical analysis. The selected wavebands were (in cm\(^{-1}\)): 200 – 299; 300 – 399.....1000 – 1099; 1100 – 1199. The test data comprised 3 groups (plagioclase, thymine and plagioclase plus thymine) each with 9 cases (the array of 9 spectra taken on each sample) across 10 wavebands, a total of 270 data points. A full description of the methodology development is contained in TN3.

Figure 2: Left: Raman spectra of (a) plagioclase; (b) thymine and (c) plagioclase plus thymine with their most intense peaks labelled (in cm\(^{-1}\)). Also shown are the wavebands across which the intensities were averaged.
8 Results

The results from the test dataset are in Figure 3, from which it can be seen that multivariate statistics can separate spectra from the three groups of materials from each other. A more complete discussion of the results is in TN3.

(a)

(b)

Figure 3: (a) A matrix scatter plot showing relationships between the 3 Principal Components of the spectra from plagioclase, thymine and a mixture of plagioclase and thymine; (b) A stacked histogram showing that discriminant analysis can separate spectra from two end-member compounds.

Having shown that the methodology works, we tested it on a Mars surface simulant. The simulant was provided to us as a powder, without any compositional information – we know approximately what components went into its mix, but not their relative proportions or exact compositions. A spectrum of the simulant is given in Figure 4.

Figure 4: Spectrum from the Oxia Planum simulant showing the most intense features
The simulant was produced for testing instruments designed for the ExoMars 2022 rover, so was regarded as a simulant for Oxia Planum, the designated landing site of the rover. Knowledge of the landing site led us to test the four most common minerals likely to be present at Oxia Planum: olivine, pyroxene, plagioclase and clay (here represented by vermiculite). A full description of the test process for the OPS is contained in TN4. A summary of the results is in Figure 5.

(a)

<table>
<thead>
<tr>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Olivine</td>
</tr>
<tr>
<td>Pyroxene</td>
</tr>
<tr>
<td>Plagioclase</td>
</tr>
<tr>
<td>Vermiculite</td>
</tr>
<tr>
<td>OPS</td>
</tr>
</tbody>
</table>

(b)

Figure 5: (a) PCA of 4 individual minerals and the Oxia Planum Simulant (OPS); (b) Histograms showing how the 5 groups can be distinguished from each other using 4th order discriminant analysis
There is some separation between the mineral species on the basis of 4 extracted components (Figure 5a): PC1 is a factor that separates olivine and pyroxene from plagioclase and vermiculite – with the OPS lying across the two groupings. Similarly, PC2 distinguishes vermiculite from the other minerals and the simulant. The factors can then be interpreted in terms of the specific wavebands in which the biggest differences between the minerals occur – although it is not straightforward because there are many overlaps. Similarly, it is apparent that the minerals can be distinguished from each other, if not individually, then in pairs, by linear discriminant analysis (Figure 5b): olivine plus pyroxene are distinguished from plagioclase and vermiculite by Disc1, 2 and 3. Vermiculite is distinguished from all the other groups by Disc 3 and plagioclase is distinguished from OPS by Disc 4. The corollary of matching a mixture of materials (the simulant) to spectra from individual minerals, as achieved here, is that

The full statistical dataset for the simulant and minerals is in TN4.

9 Follow-on Plan

Although the deliverables for the project have been completed, there are still several developments required before the project should be considered finished and ready for publication. The website on which the data library will be hosted is not yet publicly accessible, mainly because the database itself is not complete – we have library of Raman spectra, but acquisition and reduction of UV-Vis and NIR spectra of the minerals, organics, etc, is not finished. If staff resource became available (subject to additional funding), then data acquisition would continue in late summer 2022, with a view to completion of the data library by September 2022.

We have shown that the spectra from different species can be resolved into components on the basis of intensity normalised to the total number of counts integrated across the entire spectrum. For ease of data processing, the spectra were (arbitrarily) divided into 10 wavebands and an average spectrum taken for each waveband. Refinement of the data processing algorithms, using perhaps a greater number of spectra for each species, spectra taken more closely together on individual grains or narrower wavebands should all allow better resolution of different components. Research in this area will continue.