InterIIT Tech Meet 13.0 ISRO Problem Statement Endterm Report

Team 24

ARTICLE INFO

Keywords:
Chandrayaan-2, X-ray fluorescence, Lunar elemental maps, Solar Flares, Catalogs

ABSTRACT

The Chandrayaan-2 Large Area Soft X-ray Spectrometer (CLASS) payload on-board Chandrayaan-2 provides time-integrated spectral data of the lunar surface in the soft X-ray region. Using the CLASS data allows for elemental abundances to be obtained, which, in turn, can be used for terrain mapping and acquiring variations in lunar composition. We propose an algorithm from which CLASS spectral data can be used to detect X-ray fluorescence and create elemental ratio maps, deriving key compositional groups.

1. Introduction

The compositional analysis of the lunar surface provides critical insights into the characteristics of the Moon's crust, impact-generated materials, and the effects of solar wind-induced space weathering. Systematic mapping of these elemental abundances is essential for advancing our understanding of the Moon's composition and identifying strategic sites for sample-return missions and in-situ resource utilization. One of the most direct ways to map the elemental composition is X-ray fluorescence spectrometry.

In this report, we describe our algorithm to detect solar flares, create a catalog of XRF line detections, produce elemental maps of the moon and analyze its compositional groups using spectra data from CLASS.

1.1. Solar Flares

Solar flares are stochastic, broadband emissions originating in the Sun's corona (Ackermann et al., 2014). These high-energy events emit X-rays, ultraviolet radiation, and charged particles (Zhang et al., 2021).

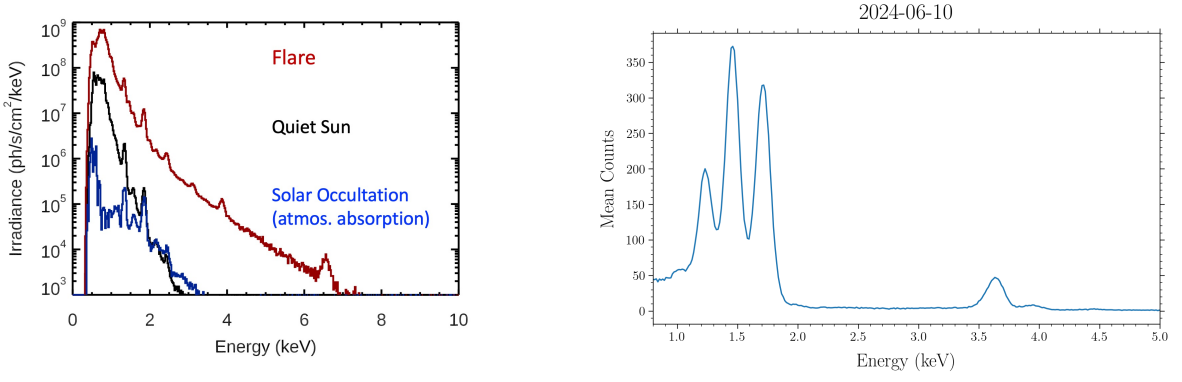

Figure 1: Left: Three example spectra from DAXSS (Mason, 2022) showing a solar flare in red (highest irradiance), quiet sun in black (middle irradiance), and a spectrum taken while the spacecraft was just moving behind the earth so that sunlight passes through Earth's atmosphere before reaching the spacecraft (blue; lowest irradiance). **Right:** CLASS X-Ray Fluorescence Spectra taken on 2024-06-10, showing Mg, Al, Si, and Ca lines respectively.

1.2. X-Ray Fluorescence Spectrometry

X-ray fluorescence (XRF) spectrometry is a non-destructive analytical technique employed to determine elemental composition by detecting X-ray emissions of specific elements. XRF is triggered when high-energy photons ionize inner-shell electrons, causing the emission of fluorescent radiation unique to each element (del Hoyo-Meléndez, 2018). The intensity of the element's spectral lines is proportional to its abundance. In space exploration, XRF is employed to analyze the surface composition of airless planetary bodies, using X-ray emissions from the solar corona during solar flares, which induce XRF on planetary surfaces, facilitating remote elemental mapping (Athiray et al., 2013).

1.3. Problem Understanding

The problem presented to us can be broken down into the following parts:

1. Detection of XRF spectral lines corresponding to solar flares using only CLASS spectral data.
2. Identifying the elements present and calculating corresponding line ratios along with deriving uncertainties.
3. Mapping uncertainty and area weighted ratios, achieving super resolution through overlapping tracks.
4. Deriving compositional groups and best ratios to take for data visualization.

Presently, there exist multiple missions (Yang et al., 2023; Wang et al., 2021; Zhang et al., 2023; Narendranath et al., 2024) aiming to map abundances (by weight percentage), but they depend on a model of the solar spectrum, which may not always be present. The XRF cross-sections are also heavily dependent on the excitation energy (Brunetti et al., 2004). We note that the method of ratios allows us to map more data, as it does not depend on the availability of a corresponding solar spectrum. We also achieve unbiased mapping of major elements on the lunar surface.

We consider that the best ratios to take are the combination of various ratios that best distinguish or identify already known selenographical features and trends obtained from ground truths (Turkevich, 1973; Yang et al., 2023). Additionally, with enough data, we can arrive at ratios that can predict composition of previously unmapped areas.

We build a dynamic detection and mapping pipeline that can process new data as and when it is available, and update our map on-the-fly.

2. Detection

In this section, we discuss the detection algorithm we employ for detecting solar flares, as outlined in figure 2.

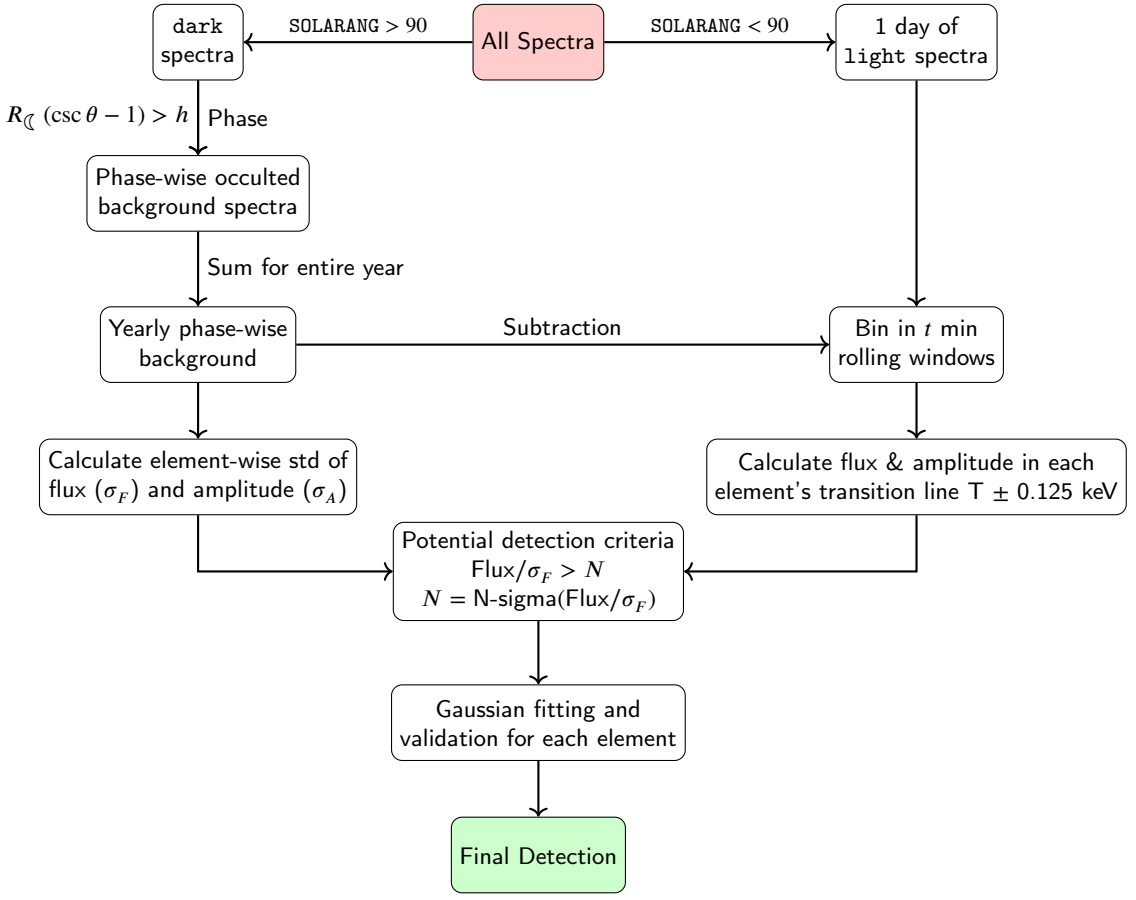

Figure 2: Flow chart of detection algorithm

2.1. Data Preprocessing

We perform some basic pre-processing steps on each day's data by first splitting the spectra into `light` (sun facing) and `dark` (away from sun) directories by using the `SOLARANG` $<$ 90 and `SOLARANG` $>$ 90 criteria respectively as mentioned in the CLASS Software Interface Specification (SIS). The spectra's FITS files are then updated to include a column for energy by simply multiplying the channel number by 13.5 eV, as described in the CLASS User Manual.

2.2. Background Estimation and Statistics

We encountered issues in modelling the background on a daily basis using all the files present in the `dark` folder of that day due to the presence of a very strong Aluminum line.

We note that this is likely due to the orbiter not being fully occulted by the moon even when the solar angle (θ) is greater than 90°. The altitude of the orbiter prevents it from being occulted as shown in figure 3.

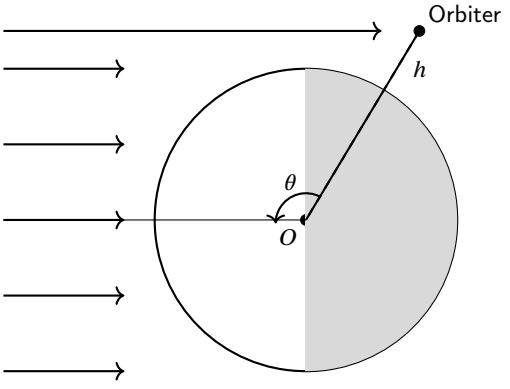

Figure 3: Chandrayaan-2 orbiter, on the `dark` side of the moon but still exposed to sunlight due to its altitude h

From some basic trigonometry, we get the condition for occultation as

$$R_{\leftmoon} \, (\csc \theta - 1) \geq h \tag{1}$$

where $R_{\leftmoon}$ is the radius of the moon and θ is the solar angle (`SOLARANG`).

Due to this exposure to the sun even on the `dark` side, the aluminum and copper on the orbiter gets excited by the incident solar flares, leading to a very strong Aluminum line in the background model as shown in figure 4 (left). This leads to an oversubtraction of the Aluminum line in the `light` data, and could possibly lead to incorrect results, as shown in figure 4 (right).

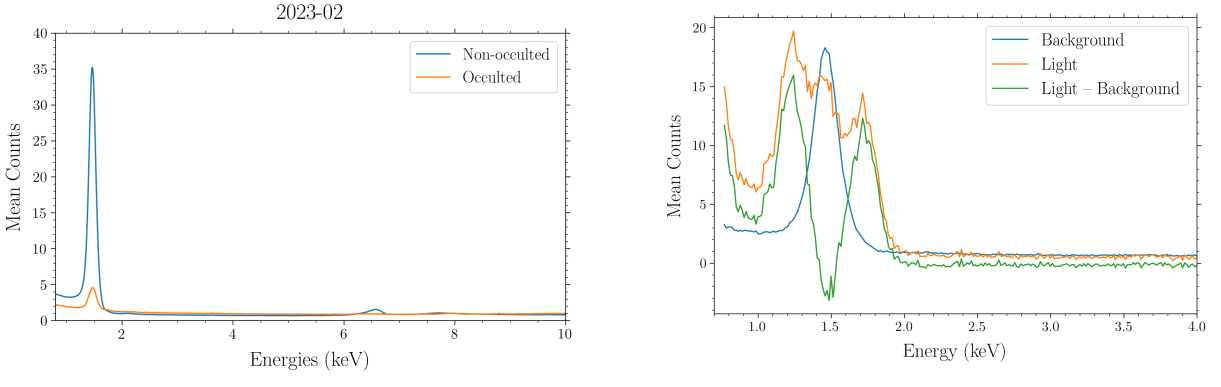

Figure 4: Left: The mean `dark` data when the orbiter is occulted vs non-occulted. **Right:** Negative Aluminum line in the background subtracted `light` data (green) due to a strong Aluminum line in the background

We also note that the amount of `dark` data per lunar day that is occulted is a very small fraction ($\sim$ 14%) of the total `dark` data, and often there is no occulted data. Due to this, we need to take a long term background model that also takes into account the geotail effects due to the phases of the moon.

We propose taking a *dynamic* yearlong phase-wise background, where we define the full moon phase as $\pm$3.5 days from the full moon, and similarly for the other phases (new moon, first quarter, third quarter). Through this, we can account for the phase-dependent background variations, as shown in figure 5. The background will be a year's worth of data prior to the time of detection, and capable of updating itself with new data as and when available.

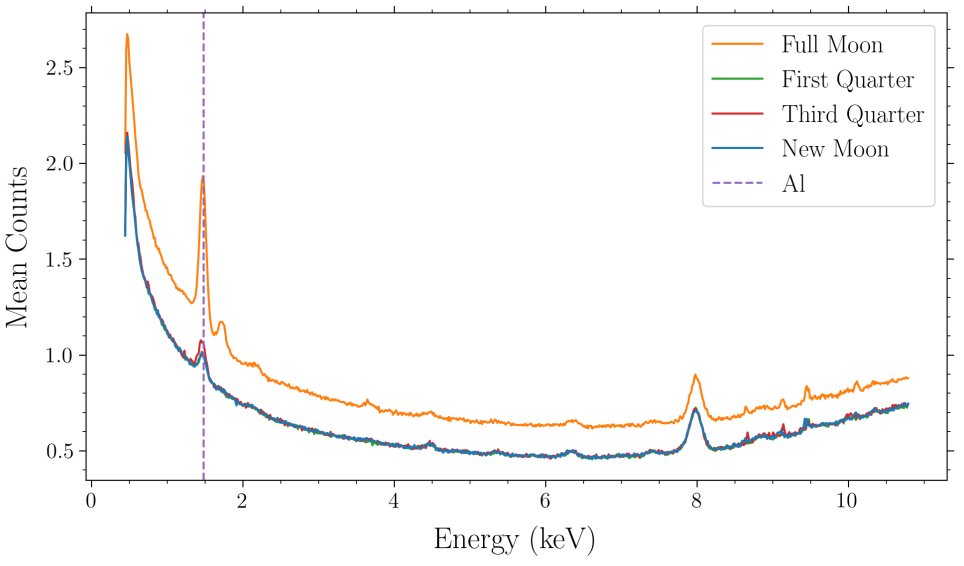

Figure 5: Phase-dependent background comparison. The full moon background is raised and has a stronger aluminum line compared to the new moon and first and third quarter phases.

The `dark` (background) spectra are first binned by taking the mean of counts in user specified time bin windows. Let $\mathcal{B}_P$ denote all the backgrounds throughout the year of phase P (NM: New Moon, FQ: First Quarter, FM: Full Moon, TQ: Third Quarter) and $\overline{\mathcal{B}}_P$ denote the mean background for phase P in a year. Although it is the standard, we do not take a sigma clipped mean, as there are large variations in the percentage of clipped values for both 5 and 3 sigma clipping, leading to a discontinuous and inaccurate background (figure 6).

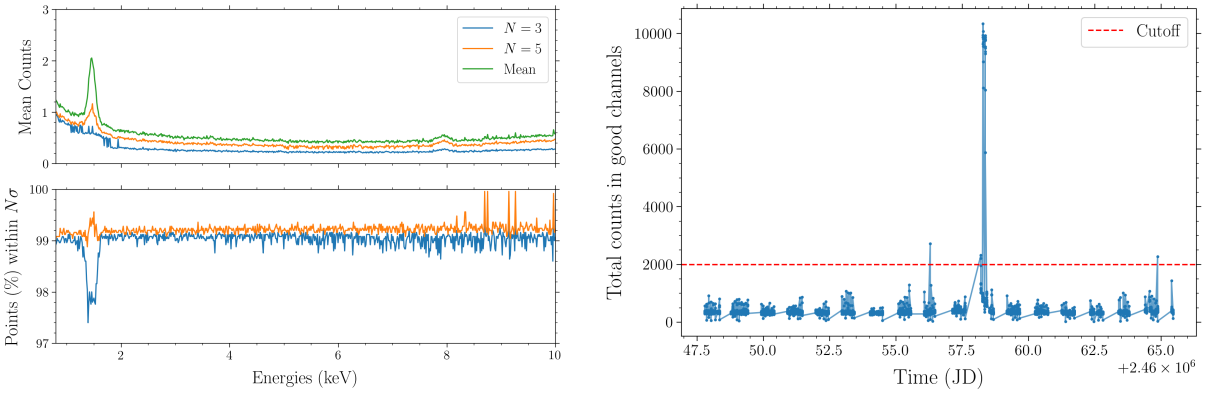

Figure 6: Left: The sigma clipped mean background vs the mean background. Note how the aluminum line is distorted even in 5-sigma clipping. **Right:** Light curve of sum of counts in 0.5 – 10.8 keV. The counts rise rapidly around the middle

We also notice certain random fluctuations in the occulted background data, likely caused by residual excitations from strong flares, as show in figure 6. Thus, we only consider those spectra which have total counts in the $37-800$ channel ($0.5-10.8$ keV) less than 2000.

We consider Oxygen, Magnesium, Aluminum, Silicon, Calcium (K-α and K-β), Titanium, Chromium, Manganese and Iron (K and L) lines for detection. A window of ± 0.125 keV is considered around each element L's transition line (T) energy $\left(E_T^L\right)$ for each of these time binned `dark` spectra, as 0.125 keV is half the typical energy difference between [Mg, Al] and [Al, Si] lines. Let the total counts (across energy) contained in this window be denoted as $\texttt{Flux}_T^L(\mathcal{B}_P)$, and the maximum counts as $\texttt{Amp}_T^L(\mathcal{B}_P)$, giving us an estimate of flux and amplitudes in the absence of spectral lines. This generates a time-series of flux and amplitudes of the background data in the window around each spectral line T.

We define the following terms:

- Standard deviation of the flux in each spectral line T as $\sigma\left(\texttt{Flux}_T^L(\mathcal{B}_P)\right)$
- Standard deviation of the amplitude of each spectral line T as $\sigma\left(\texttt{Amp}_T^L(\mathcal{B}_P)\right)$.

The trend $\sigma\left(\texttt{Flux}_T^L(\mathcal{B}_{\text{FM}})\right) \gg \sigma\left(\texttt{Flux}_T^L(\mathcal{B}_{\text{FQ}})\right) \approx \sigma\left(\texttt{Flux}_T^L(\mathcal{B}_{\text{TQ}})\right) > \sigma\left(\texttt{Flux}_T^L(\mathcal{B}_{\text{NM}})\right)$ is noticed, further reinforcing the choice of using phase dependent background.

2.3. Detection Criteria

We bin the `light` spectra in a user defined time bin rolling window, which moves forward by 8 s (1 spectra file) each time, allowing us to increase spatial resolution. We utilize rolling bins instead of disjoint bins to introduce a smoothing effect to the observed data, effectively performing a convolution with uniform weights across the binning window. This approach enhances the resolution of the resulting map by allowing each pixel within a track to be treated as distinct, rather than constraining it to a fixed track of pixels. Consequently, the method avoids sharp discontinuities between bins, yielding a continuous and more detailed representation of the data, which is particularly advantageous in scenarios requiring high-resolution mapping. Figure 7 demonstrates this increased resolution through smoothening.

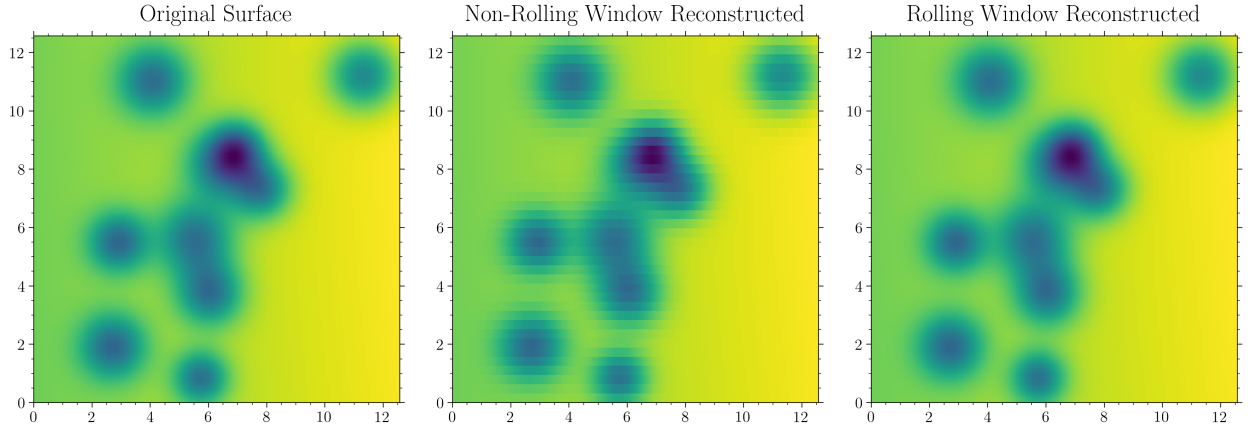

Figure 7: A smooth base image is sampled, and then reconstructed from the samples by non-rolling and rolling average of nearby samples. The rolling average creates a high resolution reconstruction. We also verify this by finding the root mean squared error to be 2 orders of magnitude lower for the rolling window reconstruction.

$\mathcal{L}_t$ is defined as the rolling window corresponding to time t. We obtain the flux values at each line T as $\texttt{Flux}_T^L(\mathcal{L}_t)$. We also define $\mathcal{L}_t^s = (\mathcal{L}_t - \overline{\mathcal{B}}_P) - med(\mathcal{L}_t - \overline{\mathcal{B}}_P)$, where med denotes the median counts of the spectra throughout the energy range of $2-7$ keV. The background subtraction is done to "detrend" the data, whereas we subtract $med(\mathcal{L}_t - \overline{\mathcal{B}}_P)$ from the detrended `light` spectra $(\mathcal{L}_t - \overline{\mathcal{B}}_P)$ to prevent offsets between the background and the `light` data that may arise due to solar activity or particle interactions.

We employed the following criteria for classifying the presence of a line T at time t in our midterm submission:

$$\texttt{Flux}_T^L(\mathcal{L}_t^s) \geq 5\sigma(\texttt{Flux}_T^L(\mathcal{B}_P)) \tag{2}$$

We now explore a different approach to create a variable detection threshold instead of keeping a constant 5 (see §2.3.1). This aims to increase the number of detections we get, and also to adapt to low solar activity days.

2.3.1. N-sigma

The N-sigma algorithm (Sharma et al., 2021) is a statistical technique employed to identify and reject outliers in light curve data. Utilizing the `sigma_clipped_stats` function from the `Astropy` module, the method iteratively refines the dataset. Starting with the light curve, it computes the median and standard deviation (σ), identifying outliers as data points that deviate by more than 3σ from the median. These outliers are removed, and the process is repeated using the updated light curve. The iteration continues until a maximum of five iterations is reached. The median and standard deviation derived from the final iteration are used as the parameters for outlier detection, further referred to as *med* and σ. Outliers are defined as data points with values $> med + N\sigma$, where we have taken the standard $N = 5$.

Now we apply the N-sigma algorithm for each day's observations, creating light curves by calculating the ratio $\frac{\texttt{Flux}_T^L(\mathcal{L}_t^s)}{\sigma(\texttt{Flux}_T^L(\mathcal{B}_P))}$. Dividing by $\sigma(\texttt{Flux}_T^L(\mathcal{B}_P))$ standardizes the source flux by the variations in the background flux, ensuring that the light curves reflect the intrinsic variations in the source flux, independent of the background fluctuations.

Instead of the constant coefficient of 5 that we used in equation 2, we replace it with $min\,(5, min(med + 5\sigma))$, where $(med + 5\sigma)$ is a list across all elements. This is done as high solar activity days creates large median and variance which causes a very high detection threshold to be set, thus we take the minimum threshold across elements and ensure this minimum threshold also isn't too large by taking its minimum with 5. We see in figure 8 the outlier threshold ($med_T^L + 5\sigma_T^L$) for *silicon* on that day is approximately 11, while the minimum threshold we obtain from the minimum across all elements is approximately 2.3. We also compare this with the constant threshold of 5.

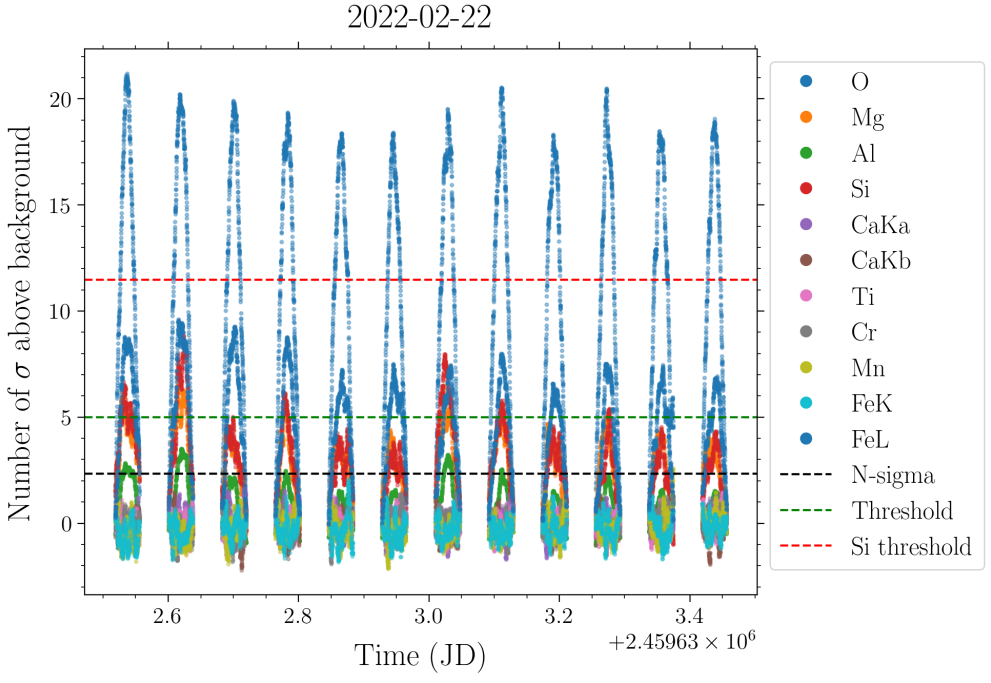

Figure 8: Light curve for a single day's data showing the outlier threshold for silicon (red), minimum across elements N-sigma threshold (black) and constant 5-sigma threshold (green). The gaps in the data is due to the orbiter going to the dark side of the moon. We note the number of data points above each threshold line is greatest for the black line, and the red line overestimates the threshold, leading to very few points.

Our new detection criteria becomes:

$$\texttt{Flux}_T^L(\mathcal{L}_t^s) \geq min\,(5, min(med + 5\sigma))\ \sigma\left(\texttt{Flux}_T^L(\mathcal{B}_P)\right) \tag{3}$$

The binned `light` spectra at time t ($\mathcal{L}_t$) that satisfy the criteria laid out in equation 3 are flagged as potential flares and move on to the next part of our algorithm. In this part, we fit a gaussian, $\mathcal{G}(\mathcal{L}_t^s)$, to each "detected" line as prescribed in §2.4 and apply cuts based on the criteria in equation 4 to eliminate false detections.

$$\begin{aligned} \mathrm{Amp}_T^L(\mathcal{G}(\mathcal{L}_t^s)) &\geq min\,(5, min(med + 5\sigma))\,\sigma(\mathrm{Amp}_T^L(\mathcal{B})) \\ \mu(\mathcal{G}(\mathcal{L}_t^s)) - E_T^L &\in [-0.05, 0.05] \\ \sigma(\mathcal{G}(\mathcal{L}_t^s)) &\in [0.05, 0.2] \end{aligned} \tag{4}$$

The values used for the mean and standard deviation limits are in units of keV and are further described in §2.4.

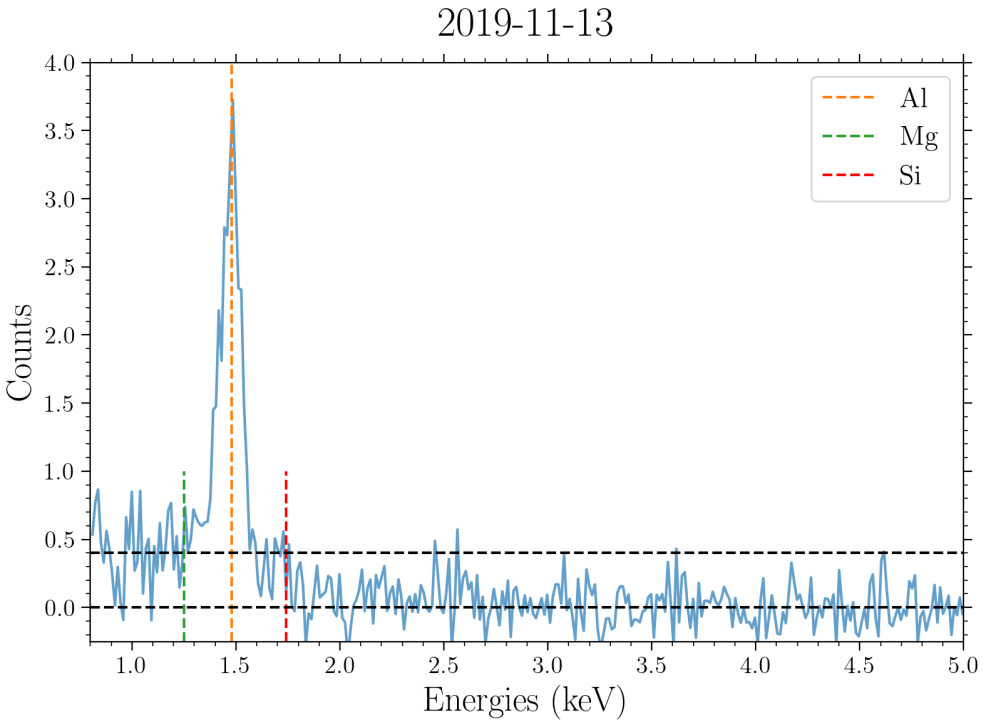

Figure 9: The left part of the spectra is rising due to the Al line, which can lead to a false detection of Mg and Si lines

Essentially, a detection is deemed as "confirmed" if there is a Gaussian profile at the corresponding element's spectral line. We can see in figure 9 how the left half of the spectra is raised above the mean level of 0 due to the strong Al line, which can lead to a false detection of Mg and Si based on equation 3 as their flux is artificially raised. Fitting gaussians to the lines helps prevent these false detections.

2.4. Gaussian Fitting

We use two gaussian fitting methods, namely `Scipy`'s `curve_fit` and `Astropy`'s `specutils` and validate them on synthetically generated spectra. We take randomly sampled 5 min binned `dark` (background) data and inject gaussians with randomly sampled standard deviations, amplitudes and means. The spectral resolution (FWHM) of $\sim$ 140 eV between -40 °C and -20 °C (Pillai et al., 2021) gives us a $\sigma \in (0.055, 0.065)$ keV. We choose a conservative range of $(0.05, 0.1)$ to uniformly sample the standard deviation.

The mean $\left(\mu = E_T^L + \varepsilon\right)$ is also sampled from a uniform distribution centered around each element's transition line, $T\ \left(E_T^L\right)$ taking $\varepsilon \sim U(-0.05, +0.05)$ keV to account for the error we obtain while fitting fixed parameter $\left(\mu = E_{\text{K-}\alpha}^{\text{Ca}}, \sigma = 0.07\text{ keV}, A = 20\text{ counts}\right)$ gaussians for 1000 randomly sampled 296 s backgrounds (figure 10).

This illustrates poissonian errors in the spectra's counts, as was told to be assumed in the FITS headers of each file. We define a conservative ± 0.05 keV window for ε to account for additional terms such as the spectra channel width of 13.5 eV as well as the spectral redistribution function (SRF), which we have not quantified.

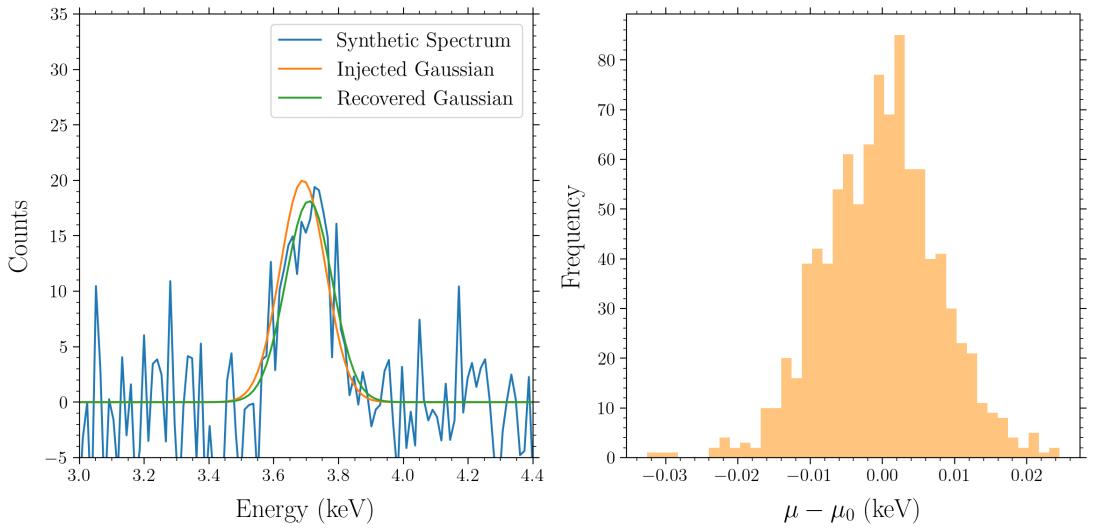

Figure 10: Left: Shift in μ for injected vs recovered gaussian after fitting on the synthetic spectrum. **Right**: Histogram of uncertainities in the recovered mean μ for 1000 injected gaussians at $\mu_0 = 3.69$ keV

2.4.1. Fitting Methods

1. `curve_fit`: This fits a model function to the data by minimizing the sum of squared residuals between the data points and the model using a non-linear least squares optimization approach. We set the initial guess (`p0`) parameters of the fit as $\mu = E_T^L$, $\sigma = 0.07$ keV, and $A = \texttt{Amp}_T^L(\mathcal{L}_i^s)$. Our fitting window is taken as a ± 0.125 keV window centered around E_T^L, as larger windows often led to the non convergence of the least squares optimizer.
2. `specutils`: This fits a one-dimensional Astropy gaussian model, `Gaussian1D`, to the spectral data by minimizing the residuals between the data points and the model using a Trust Region Reflective algorithm with least squares optimization. The Gaussian model is defined by parameters $\mu = E_T^L$, $\sigma = 0.07$ keV, and $A = \texttt{Amp}_T^L(\mathcal{L}_i^s)$. The `Spectrum1D` class is used as a container for the spectral data, representing flux (photon counts) as a function of energy (keV). The fit is performed using the `fit_lines` function from `specutils.fitting`. The fitting window is restricted to ± 0.125 keV for similar reasons as `curve_fit`.

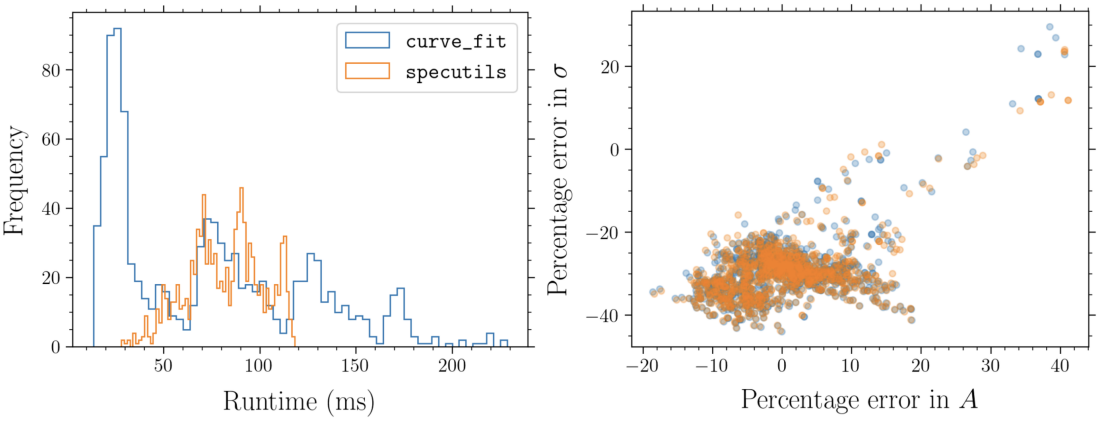

Figure 11: Comparison of `curve_fit` and `specutils`. Both functions were used to fit synthetically generated data.

The performance of `specutils` is comparable to `curve_fit` (figure 11) in terms of the runtime (`curve_fit` was set to `maxfev` = 250) and percentage errors between the fitted and injected values.

We choose `curve_fit` as our default fitting function as we found it to be faster while running the complete pipeline, but also keep the option to use `specutils` via a parser argument. All further results use the `curve_fit` fitting methodology.

From literature (Böhm-Vitense, 1992; Heiles & Troland, 2003), we know that a spectrum (S) is modelled as a continuum (C) and a combination of Gaussian, Lorentzian or Voigt absorption or emission lines. The model is as follows for a gaussian profile:

$$S(E) = C(E) + \sum_{i=1}^{N} \mathcal{G}(E_i, \sigma_i, A_i) \tag{5}$$

To obtain precise values of the spectral line parameters, we should ideally be fitting the superposition of all the possible spectral lines $\left(\sum_{i=1}^{N} \mathcal{G}(E_i, \sigma_i, A_i)\right)$. However, it is difficult to fit many parameters using the least squares optimization techniques used by `curve_fit` and `specutils`, as they are fairly sensitive to the initial condition provided (Juvela, Mika & Tharakkal, Devika, 2024), and often requires a bayesian Markov Chain Monte Carlo (MCMC) approach.

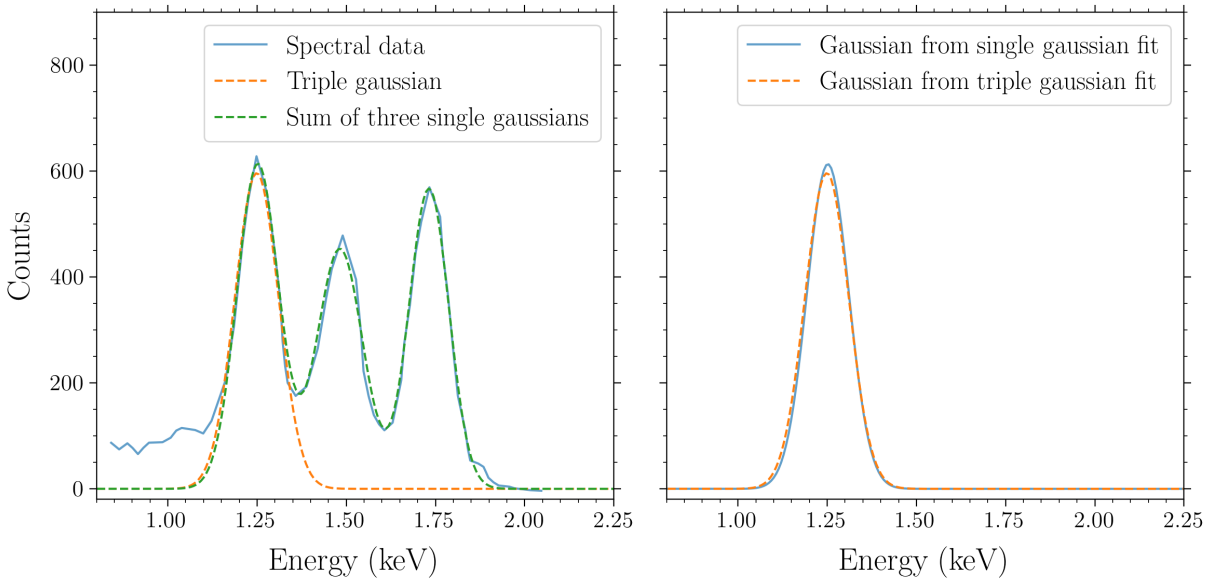

Figure 12: Left: Triple gaussian fit vs sum of single gaussian fits to the sample spectra presented in Narendranath et al. (2024). **Right**: Comparison of a single gaussian from the triple gaussian fit and the single gaussian fit for Mg. The difference in areas under the gaussians is typically $\sim 2\%$

We implemented a MCMC fitting approach, but found `curve_fit` and `specutils` to be faster by two orders of magnitude with similar accuracies. Thus, in the interest of computing time, we fit single gaussians to each spectral line and verify that this method is approximately equivalent to fitting the combination of lines as show in figure 12.

2.4.2. L-transition Lines

The maximum photons from a typical solar flare spectra, lie in the $1 - 2$ keV range as seen in figure 1. The XRF cross-section at this excitation energy for metals belonging to the first row of transition elements is much higher for the L-lines than the K-lines. As a result, the probability of K-line fluorescence is much lower than L-line fluorescence for these elements (Brunetti et al., 2004).

For the elements in the range $20 \leq Z \leq 26$ (Ca to Mn), K-line fluorescence only occurs at excitation energies of $4 - 6$ keV range, hence it is much less probable than L-line fluorescence. However, for the elements Ca, Sc, Ti the L-line energy is too low to be detected by the CLASS instrument. For the elements V, Cr and Mn, the L-line energy (0.521, 0.584, 0.64 keV respectively) is masked by the large flux of the $O_{\text{K-}\alpha}$ line at energy 0.525 keV and the detection

of these elements is made extremely unlikely through XRF methods. Further, these elements are only present in trace abundances, which makes abundance calculations difficult. However, Fe is present in much higher abundance on the lunar surface, and its L-line is sufficiently far away from the $O_{\text{K-}\alpha}$ line to be detectable.

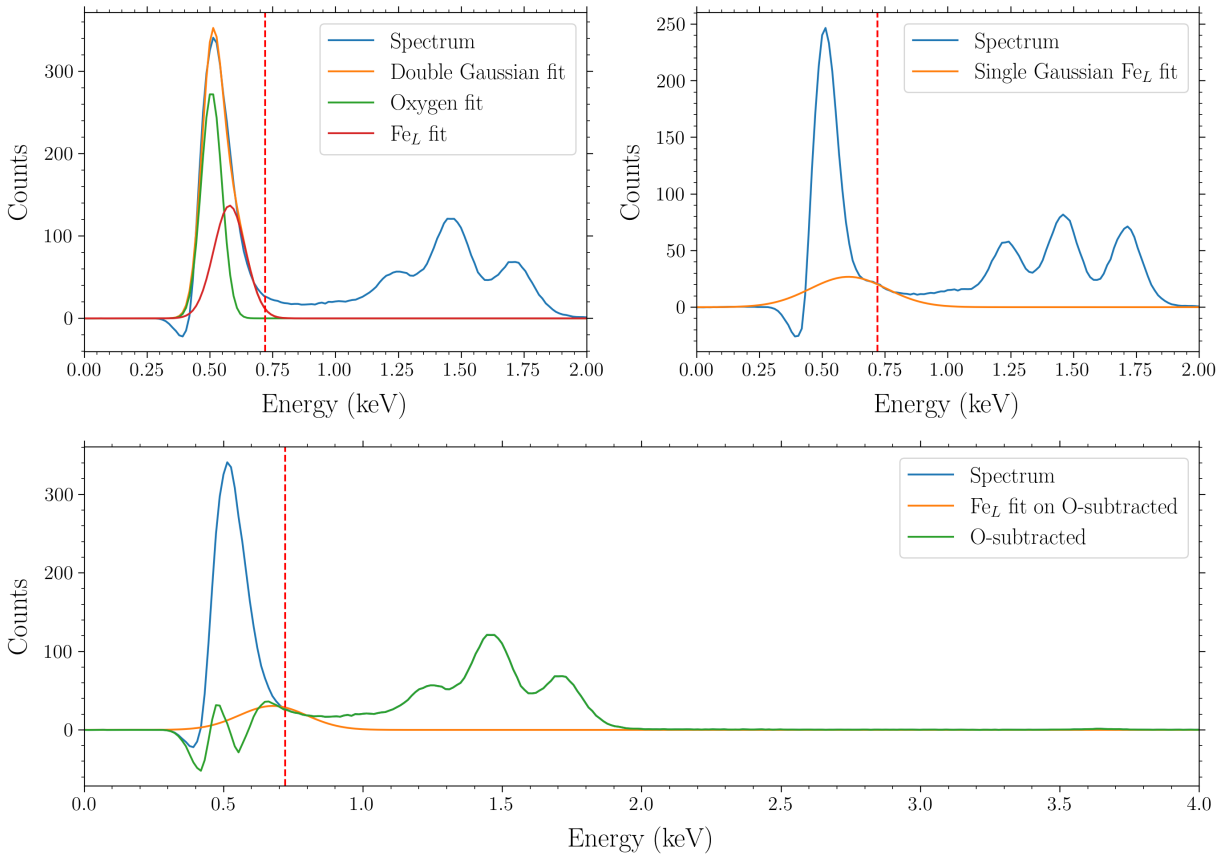

Figure 13: Top left: Double gaussian fitting leads to the two gaussians coming inside each other and does not fit the Fe_L line at its fluorescence energy. **Top right:** Single gaussian fit to the Fe_L line without subtracting the oxygen line. We see that the mean of the gaussian is shifted from the Fe_L and the standard deviation is large. **Bottom:** Subtracting the oxygen line and fitting Fe_L shows better results, with the mean matching the transition line and a lower standard deviation.

The Fe_L line at 0.72 keV is still quite close to the oxygen line at 0.52 keV, making it detectable but difficult to fit. We attempted to fit the line using a single gaussian model as discussed in §2.4.1, but find that the parameters of the fit vary drastically due to oxygen's high contamination. We also try fitting a double gaussian distribution, but find that the fit does not give satisfactory results, as shown in figure 13. In order to overcome this difficulty, instead of trying to fit both the lines, we fit the oxygen line separately, then subtract the gaussian fit from it and then fit the Fe_L line. We find this method gives much better results than the double gaussian or the single gaussian in the presence of oxygen, as shown in figure 13.

2.4.3. Uncertainty Propagation

We focus on errors in the fitted parameters of the gaussian model to the peaks. The estimated parameters are the mean (μ), standard deviation (σ) and amplitude (A) of the gaussian. We use two approaches to estimate the errors: the covariance matrix obtained by the optimization algorithm used by `curve_fit` and `specutils`, and simulations using randomly sampled backgrounds.

These incorporate the Poisson noise inherent in the incoming photon counts. In this section, we will also evaluate the accuracy of the methods used to estimate the errors in the fitted parameters.

The spectral and background data has inherent Poisson noise because the number of arrival photons follows the Poisson distribution. The distribution is characterized by a mean equal to the expected number of photons predicted from the model.

In the observed counts, we have contributions from the background and the source. When multiple sources contribute to observed counts, the variance in the observed counts is the sum of the variance in source counts and background counts. Hence, we can calculate the variance in the counts of the background subtracted spectra through the propagation of errors.

Assuming there are $N_{\mathcal{L}}$ spectra to be added to create the binned `light` spectra at time t, and $N_{\mathcal{B}}$ spectra that are added to make up the background spectrum:

$$\mathcal{L}_t = \frac{1}{N_{\mathcal{L}}} \sum_{i=1}^{N_{\mathcal{L}}} \mathcal{L}_i \quad \overline{\mathcal{B}} = \frac{1}{N_{\mathcal{B}}} \sum_{i=1}^{N_{\mathcal{B}}} \mathcal{B}_i$$

$$\mathcal{L}_t^s = \frac{1}{N_{\mathcal{L}}} \sum_{i=1}^{N_{\mathcal{L}}} \mathcal{L}_i - \frac{1}{N_{\mathcal{B}}} \sum_{i=1}^{N_{\mathcal{B}}} \mathcal{B}_i$$

$$\texttt{var}\left(\mathcal{L}_t^s\right) = \frac{1}{N_{\mathcal{L}}^2} \sum_{i=1}^{N_{\mathcal{L}}} \texttt{var}(\mathcal{L}_i) + \frac{1}{N_{\mathcal{B}}^2} \sum_{i=1}^{N_{\mathcal{B}}} \texttt{var}(\mathcal{B}_i)$$

Since we can assume Poisson errors for the counts as stated in the FITS headers, we have $\texttt{var}(\mathcal{L}_i) = \mathcal{L}_i$ and $\texttt{var}(\mathcal{B}_i) = \mathcal{B}_i$.

$$\texttt{var}\left(\mathcal{L}_t^s\right) = \frac{1}{N_{\mathcal{L}}^2} \sum_{i=1}^{N_{\mathcal{L}}} \mathcal{L}_i + \frac{1}{N_{\mathcal{B}}^2} \sum_{i=1}^{N_{\mathcal{B}}} \mathcal{B}_i = \frac{\mathcal{L}_t}{N_{\mathcal{L}}} + \frac{\mathcal{B}}{N_{\mathcal{B}}}$$

$$\texttt{std}\left(\mathcal{L}_t^s\right) = \sqrt{\frac{\mathcal{L}_t}{N_{\mathcal{L}}} + \frac{\mathcal{B}}{N_{\mathcal{B}}}} \tag{6}$$

We pass $\texttt{std}\left(\mathcal{L}_t^s\right)$ obtained from equation 6 to the fitting functions `curve_fit` and `specutils`, which then weigh each data point accordingly. These weights are used to decide the optimal fit and are propagated in the fitted parameters. In figure 14, we verify that the errors predicted by the covariance matrix are accurate estimates of the true errors.

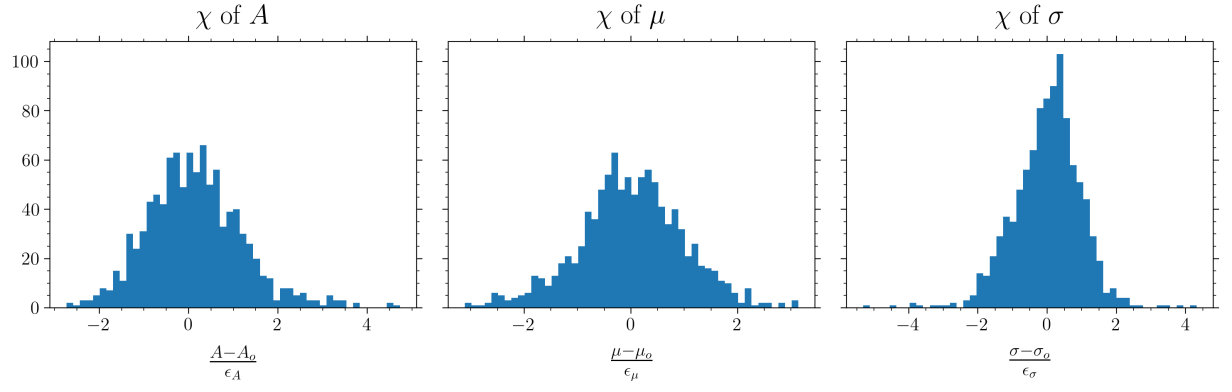

Figure 14: Histogram of χ of gaussian fits, obtained by random sampling of 300 s backgrounds, injecting gaussians to it and fitting them back. We can see that the distribution is centered around 0 with a standard deviation approximately equal to 1, which implies that the model (`curve_fit` or `specutils` gaussian model) correctly estimates the errors in the fit.

2.5. Results and Discussion

We compare the fractional increase in detections using the N-sigma algorithm compared to our previous constant threshold of 5 in table 1. These are the final detections that pass both detection criteria 3 and 4.

Element	2019	2020	2021	2022	2023	2024
Mg	0.0577	0.2321	0.3785	0.1119	0.0306	0.0224
Al	0.0549	0.0956	0.2989	0.1202	0.0685	0.0254
Si	0.1132	0.1476	0.3067	0.0745	0.0275	0.0150
$Ca_{K\text{-}\alpha}$	0.5172	1.2109	0.5685	0.4124	0.5025	0.6973
$Ca_{K\text{-}\beta}$	0.0227	5.3289	0.8476	0.7069	1.2057	1.6672
Ti	0.0	12.0435	6.8007	1.7739	2.1085	0.8336
Cr	*4	89.5344	16.6744	974.0	8.6667	0.3077
Mn	*0	*18886	*21382	*6552	*270	*0
Fe_K	*36	*15902	201.182	6.2008	8.6255	4.678
Fe_L	*0	0.2262	*354	2.4375	0.9986	*604

*have taken the N-sigma counts as 5-sigma was zero

Table 1
Year-wise relative counts $\left(\frac{N\text{-}sigma - 5\sigma}{5\sigma}\right)$ of element detections

The element-wise flux recorded shows a strong correlation with the `SOLARANG` parameter. It increases to a maximum when the solar angle is close to 0°, and a minimum at the times when the satellite transitions to night side, and the solar angle is ±90°. The flux towards the poles decreases significantly due to the lower intensity of the incident solar flares. The gaps represent night side where solar angle ≥ 90° in magnitude.

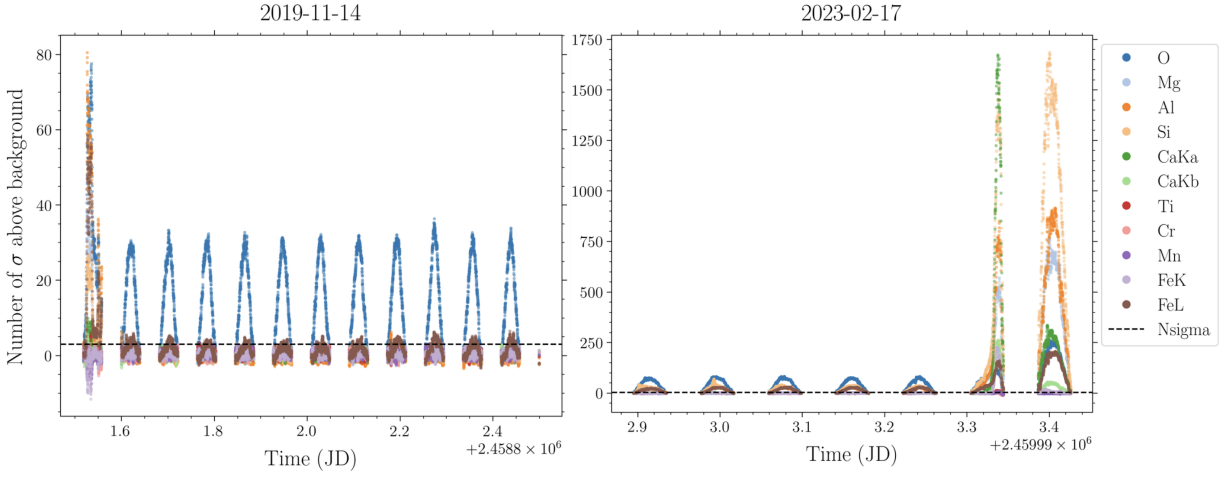

Figure 15: Left: Sample light curve (similar to figure 8) showing a strong flare on a typical day during solar minima. **Right:** The light curve for a day during solar maxima. We note the detection of flares continuously in every orbit during solar maxima as seen from the Silicon line excitations, which are not present in the solar minima light curve. Oxygen line is typically present throughout due to noise at the lower limits of the detector, which gets rejected by our cuts.

For each day, we store the gaussian fit parameters, element-wise, and lunar coordinates for each detection in a `JSON` file. This file is then parsed to filter out detections based on equation 4, then move on to mapping.

3. Mapping

The maps are produced as a `GeoTIFF` file using the standard coordinate reference system (CRS) `EPSG:4326`. We initially prepare a binary mask for each element based on the line detections at the specific locations.

For calculating the ratios, we filter out the detections based on the criteria described in equation 4, and calculate the area under the gaussians for various elements to calculate their ratios. The area under a gaussian defined as $\mathcal{G} = Ae^{\frac{-(x-x_0)^2}{2\sigma^2}}$ is equal to $\sqrt{2\pi}A\sigma$, so we define the ratio $\mathcal{R}_L^M$ of element M with element L as:

$$\mathcal{R}_L^M = \frac{A_M \sigma_M}{A_L \sigma_L} \tag{7}$$

We take most ratios with the base element L being Silicon, since its composition is assumed to be roughly uniform throughout the surface of the Moon (Yang et al., 2023). We also take ratios with other reference elements, namely oxygen and aluminum.

Across multiple passes over the same point, we average the ratio obtained for that point using an uncertainty weighted average as described in Kirchner (2016); Taylor (1997). Suppose there are n observations of the ratio $\mathcal{R}_L^M$ denoted by ${}^1\mathcal{R}_L^M, {}^2\mathcal{R}_L^M, {}^3\mathcal{R}_L^M, \cdots {}^n\mathcal{R}_L^M$. The following equation captures the uncertainty propagation for reading i:

$$\Delta_i = \Delta\left({}^i\mathcal{R}_L^M\right) = \Delta\left(\frac{A_M\sigma_M}{A_L\sigma_L}\right) = {}^i\mathcal{R}_L^M \sqrt{\left[\frac{\Delta\left(A_M\right)}{A_M}\right]^2 + \left[\frac{\Delta\left(\sigma_M\right)}{\sigma_M}\right]^2 + \left[\frac{\Delta\left(A_L\right)}{A_L}\right]^2 + \left[\frac{\Delta\left(\sigma_L\right)}{\sigma_L}\right]^2} \tag{8}$$

where $\Delta(A_M)$, $\Delta(\sigma_M)$, $\Delta(A_L)$, $\Delta(\sigma_L)$ are evaluated by the covariance matrix as explained in §2.4.3.

In addition to the above uncertainty in the fitting model, the height of the orbiter SAT_ALT from the surface also varies in the range $70 - 130$ km as mentioned in Pillai et al. (2021). This scales the surface area corresponding to each reading by $(\text{SAT\_ALT})^2$, since the 7° FWHM angle of the instrument is constant. Readings taken over a higher surface area correspond to proportionally higher variance. Thus, the overall variance is proportional to $\Delta_i^2 A_i = \Delta_i^2 \times (\text{SAT\_ALT}_i)^2$. Finally, the uncertainty weighted average and the propagated uncertainty are given by the following equations:

$${}^{\text{avg}}\mathcal{R}_L^M = \frac{\displaystyle\sum_{i=1}^{n} \frac{{}^i\mathcal{R}_L^M}{A_i\Delta_i^2}}{\displaystyle\sum_{i=1}^{n} \frac{1}{A_i\Delta_i^2}} \tag{9}$$

$$\Delta\left({}^{\text{avg}}\mathcal{R}_L^M\right) = \frac{1}{\sqrt{\displaystyle\sum_{i=1}^{n} \frac{1}{A_i\Delta_i^2}}} \tag{10}$$

We achieve a sub-pixel resolution map using the rolling window as explained in figure 7 along with taking the weighted average of the ratios over intersecting tracks.

3.1. Methodology

We obtain the coordinates of each spectra file from its FITS header, then take the central $(7° \times 7°)$ region, since that is the spatial FWHM of the CLASS instrument as mentioned in Pillai et al. (2021). This corresponds to a surface square of side length $(\texttt{SAT\_ALT}) \cdot \tan 7°$.

We use the `Rasterio` package to read and write geospatial raster data. The `rasterize` function of `Rasterio` is used to project the central square into the grid pixels. It takes as inputs coordinates and ratio values, packaged as `Polygon` objects using the `Polygon` method from the `Shapely` module.

We create the maps using the following procedure:

1. Every detection corresponds to multiple 8 s spectra based on the time-bin chosen for detection. Each of these 8 s spectra are first converted to individual `Polygon` objects, and then merged into a single `Polygon` object using the `unary_union` method in `shapely.ops`.

2. Certain shapes which cross the ante–meridian line are split into two polygons first and then dispatched as `MultiPolygons` to avoid wraparound (boundary) issues.
3. Finally, the `MultiPolygons` are merged again using `unary_union` into `Polygon` objects with the same ratio values.

Since a single grid pixel could be part of multiple tracks, we maintain the average value as per equation 9, of the pixel over all tracks it was a part of. This is done by having 2 bands in the `raster` data, one for line ratios, and the other for fit model uncertainties (Δ_i values). The area of the rasterized square is also used for averaging, as per equations 9 and 10.

The maps can be viewed using any GIS-application like `QGIS`[1] by adding it as an overlay as a Raster Layer and choosing the corresponding bands to see the required ratios of interest. To transform it to other projections, `gdaltranslate` with the relevant CRS can be used. It can also be uploaded as a layer on the `LROC Quick Map`[2].

3.2. Effect of time bin on mapping

We compare the maps generated by a 296 s and 96 s time binning. We notably see that lower time bin is better for ratios with higher signal, as it is able to resolve selenological features better, as seen in figure 16. The lower signal ratios resolve features better with a higher time bin as shown in figure 17.

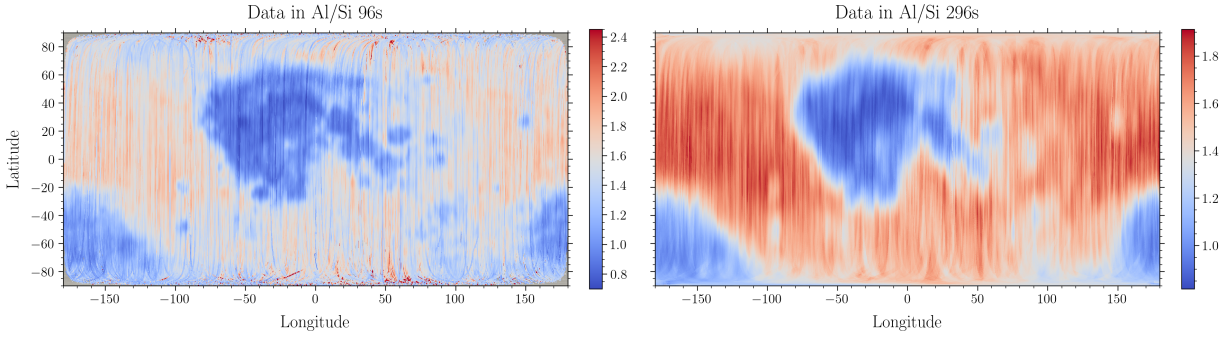

Figure 16: The Al maps demonstrate higher spatial resolution when the time bin is lowered, since smaller tracks are used for mapping, which increases contrast between nearby points

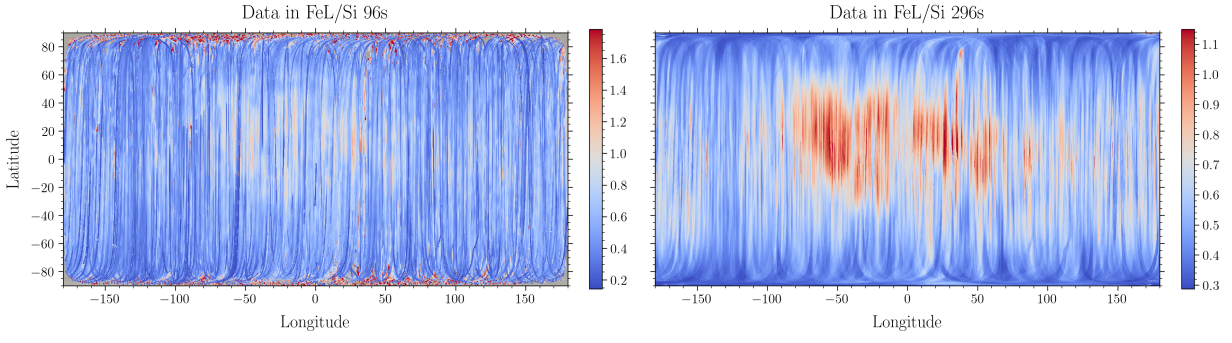

Figure 17: The Fe maps demonstrate data loss when the time bin is lowered, since less data is added, implying a low SNR.

[1] `https://www.qgis.org/`
[2] `https://quickmap.lroc.asu.edu/`

4. Lunar Compositional Groups

The online article by Korotev (2024a) provides analysis of almost all known lunar samples. Over 98 – 99% of the rocks and soil samples contain the following elements: Oxygen (41 – 45%), Silicon, Aluminum, Calcium, Iron, Magnesium and Titanium. Of the remaining amount, nearly all is Manganese, Potassium, Sodium and Calcium and these groupings together are called major and minor elements, respectively. The major cations are not present in their simple oxide forms and instead are present at the cation sites of compositionally complex silicate or oxide minerals or in glasses that have been produced by impact melting of rocks containing those elements.

These minerals are even better understood as solid-solutions of simpler silicate endmember compounds (Heiken et al., 1991). Still, lunar petrologists often quote compositions as though the oxide was present by converting the elemental abundance to the oxide considering the oxidation state of the cation, e.g. 10% Si corresponds to 21.4% SiO_2.

All rocks on the Moon were originally igneous and were formed by the cooling of molten material, after the separation of the Moon from the Earth. Cycles of cooling and melting occurred on different parts of the Moon at different times. The other main shaping force of the Moon's surface is meteoroid impacts, of which many tend to fuse rocks together and form glassy and highly variable rocks (breccia) or impact-melts, which are more homogenous solidification of rocks after heating by impact (Korotev, 2024b).

Ultimately, when classifying any rock or sample from the moon, the most useful distinction would be between highland anorthositic rocks that generally have higher aluminum weight percentage (corresponding to high Al/Si) and basaltic rocks that have higher iron percentage (corresponding to high Fe/Si) compositions, with more detailed subclassifications also possible. The reason for considering Si in this manner is that its presence on the Moon tends to be relatively constant throughout ($\approx 20\%$) (Heiken et al., 1991). Motivation for the classification we landed upon is found in the third edition of the Lunar Sourcebook and figure 18.

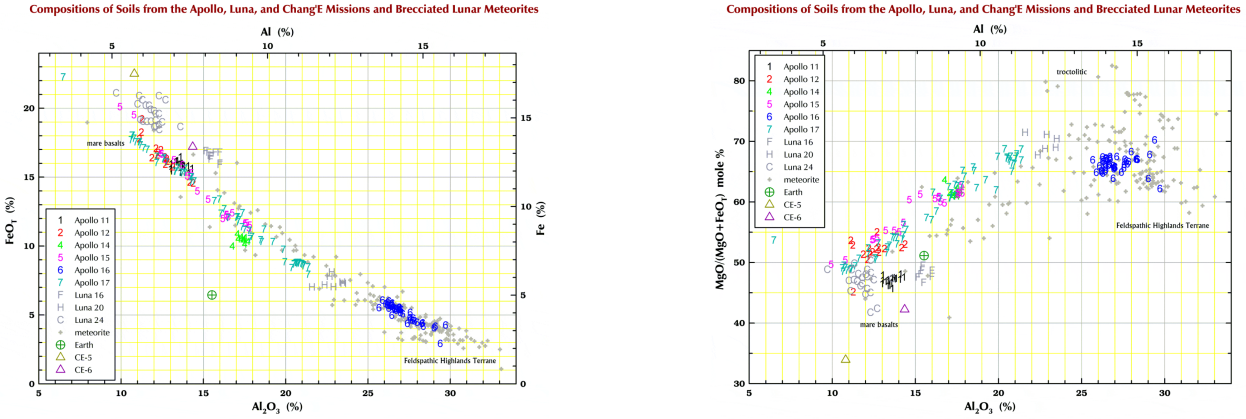

Figure 18: Left: Scatter plot of composition of returned Lunar Samples and known Lunar meteorite shows strong linear anti-correlation between iron and aluminum content in weight percentage. **Right:** Scatter plot of Mg# vs aluminium oxide content roughly, showing three kinds of clusters (Korotev, 2024a).

4.1. High Al/Si ratio

The earliest cycle of cooling and melting is thought to have produced the highland regions Heiken et al. (1991), which are dominant in aluminum. The first subclassification that is often used is that of *ferroan anorthosite* which are very rich in plagioclase feldspar and higher iron compared to magnesium. *Mg-suite* is the name that is given to highland regions which have lesser feldspar but more Mg-rich pyroxene and olivine rocks, which are distinguished into rock types such as Gabbros and Norites that form 20% of the highland crust and 2% of the Moon overall (Taylor et al., 1993).

4.2. High Fe/Si ratio

Later episodes of internal melting of the material at depths of 100 – 400 km produced volcanic eruptions that form the basaltic lava flows in the Moon's mare regions. Lunar lava tends to have higher Fe and lesser Al compared to terrestrial lava.

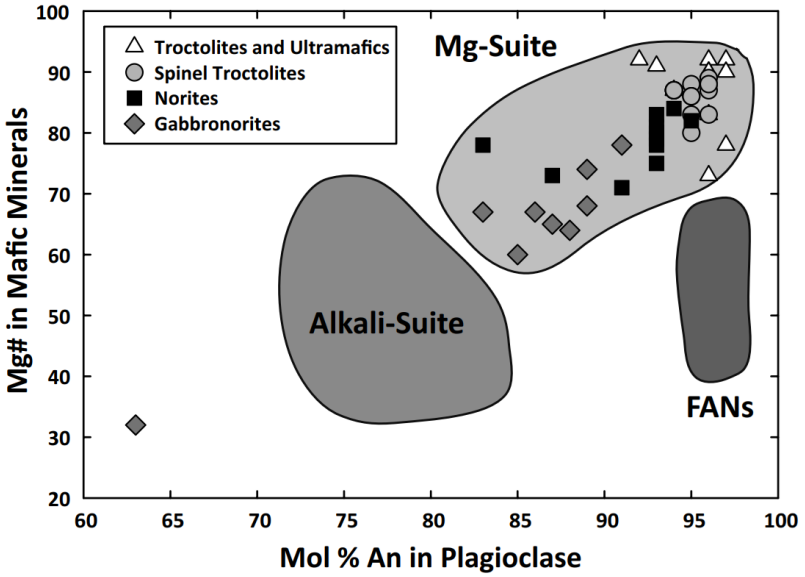

Figure 19: Grouping of the three types of Mg-rich rocks in the lunar highlands based off of Mg#. The Mg-suite stands out as these rocks have a high Mg to Fe ratio which is expected as with increasing Anorthosite percentage we expect less iron to be present (Shearer et al., 2015).

- **High Titanium** ($>$ 9%) basalts (HT) tend to have higher concentrations of ilmenites ($FeTiO_3$) which outcompete Mg-rich olivines and pyroxenes during crystallisation. The pyroxenes that do crystallise tend to have higher Calcium than usual.
- **Low Titanium** (LT) usually refers to basalts which are below 9% by weight which have elevated Mg/Si ratios compared to HT basalts due to the discussed anti-correlation.
- **Very Low Titanium** classification (VLT; $<$ 1.5%) in which the colour of the basalt lightens significantly due to the rapidly declining concentrations of ilmenite, which is quite dark in colour.

Overall Titanium is an important element as it helps distinguish regions within the Procellarum-KREEP Terrane (PKT). It is unique as even though it is a minor element compared to iron, magnesium or aluminum, its abundance spans a whole order of magnitude, going from around 1% to above 10% as well.

4.3. Need for More Intricate Ratios

We, with the methodology developed, shall often be combining together spectral files taken over a period of 96 seconds or more. This means that the instrument will receive emissions from not just a few discrete rock samples but from a large area on the surface of the moon, as discussed above in the mapping methodology.

This means that the emissions will often inevitably be a combination of the above kinds of compositional groups. which means that the sources contributing to a particular track or shape file could reasonably be of very different compositions. Thus, we will not be able to give a sure shot classification for the surface composition of the track with just the above ratios, if none of the above conditions defined for classification are unambiguously met, as defined by the ground truths from the previous Apollo, Luna or Chang'E missions given in figure 18. Thus, we expect scatter plots to not neatly separate into distinct populations and non-linear modelling to be required to capture the underlying variation.

Another reason for ambiguity could be difficulties with observing certain elemental lines frequently enough, with unambiguous fits and good selenographical coverage. Thus elements which are only weakly excited would be difficult to create maps for and instead histograms would be better ways to visualise the data. An example of one such quantity already used in the literature is the "Mg Number" which is the ratio of the molar percentage of magnesium oxide with the sum of that of magnesium and iron oxide.

5. XRF maps of the entire Lunar Surface

Given the criteria we have used to claim that an element has been detected in a particular light curve and the criteria used to ascertain whether the elements of interest are present or not, we arrive at the following maps of ${}^{\text{avg}}\mathcal{R}^{\text{Fe}}_{\text{Si}}$, ${}^{\text{avg}}\mathcal{R}^{\text{Al}}_{\text{Si}}$ and ${}^{\text{avg}}\mathcal{R}^{\text{Mg}}_{\text{Si}}$. These are unique from the previous full moon maps in the following significant ways:

1. Maps by (Zhang et al., 2023; Wang et al., 2021; Yang et al., 2023) are the most detailed maps available, however, they were obtained using the Kaguya Multi-Band Imager which had an inherent resolution of $\sim$ 100 meters per pixel as well as machine-learning based interpolation, using returned samples from the moon as the ground truth.
2. The previous Chandrayaan-2 maps (Narendranath et al., 2024) themselves do not cover the entirety of the moon, as explained in §1.3. Both the above types of maps were of the actual elemental molar compositions and not the line ratios as we have produced here.
3. Mapping elemental line ratios was undertaken by Gloudemans, A. J. et al. (2021), however, they were limited to only Apollo 15 and 16 data and could not produce a map of the whole moon. Due to this lack of data they could only produce maps of $\mathcal{R}^{\text{Al}}_{\text{Si}}$, $\mathcal{R}^{\text{Mg}}_{\text{Si}}$, and $\mathcal{R}^{\text{Mg}}_{\text{Al}}$, crucially, missing out mappings with iron.

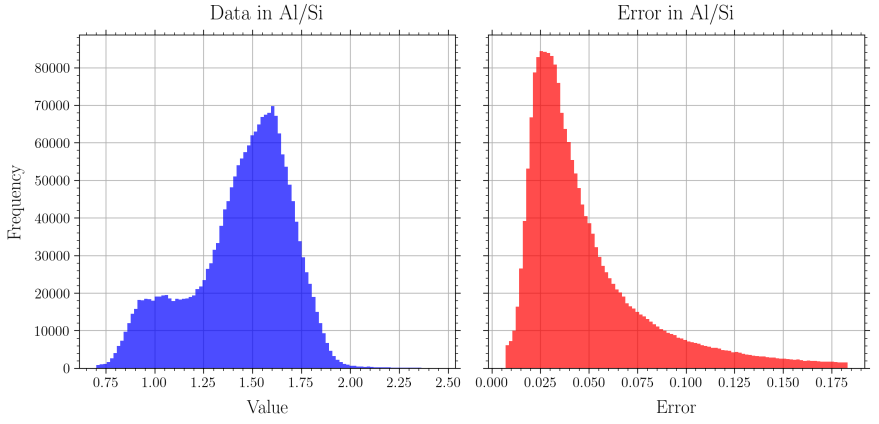

Figure 20: Left: Histogram of values of average aluminum line ratio over all the pixels in the image. This is close to the histogram of the same ratio found in Gloudemans, A. J. et al. (2021). Ours is similar to theirs insofar as we also obtain a bimodal distribution. Having mapped more of the lunar highland, our obtained second peak is understandably higher. **Right:** Histogram of uncertainties in average aluminum line ratio over all the pixels.

s

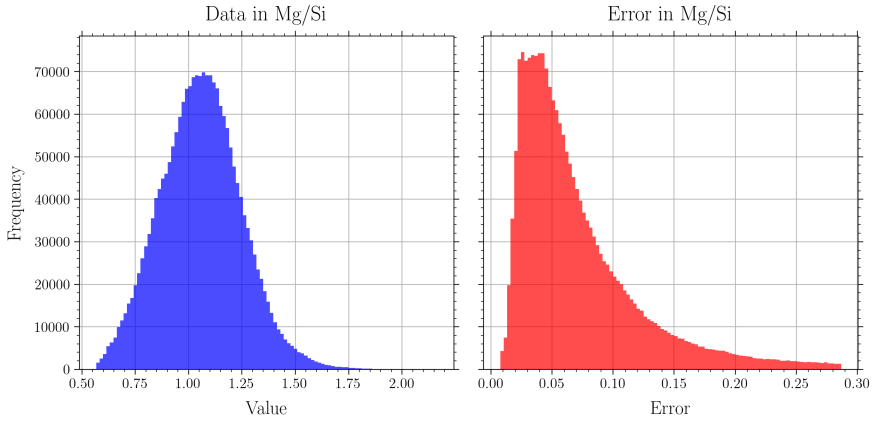

Figure 21: Left: Histogram of values of average magnesium line ratio over all the pixels in the image. This closely resembles the single modal distribution presented in Gloudemans, A. J. et al. (2021). **Right:** Histogram of uncertainities in average magnesium line ratio over all the pixels.

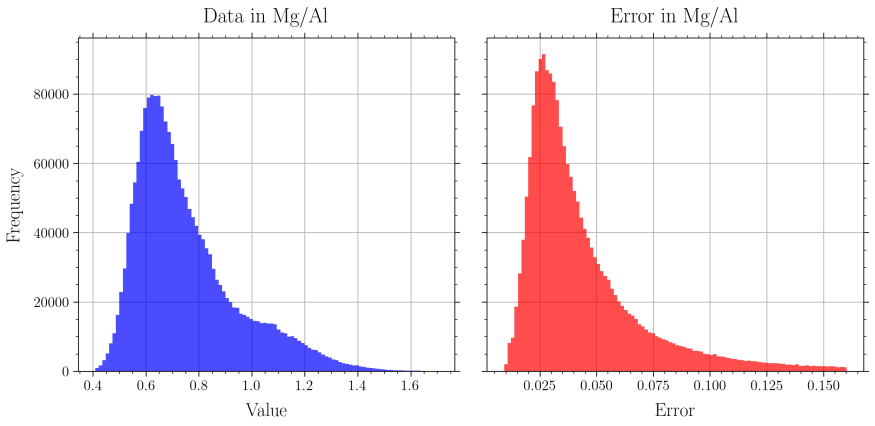

Figure 22: Left: Histogram of values of average magnesium by aluminum line ratio over all the pixels in the image. This also resembles the bi-modal distribution presented in Gloudemans, A. J. et al. (2021). **Right:** Histogram of uncertainities in average magnesium line ratio over all the pixels.

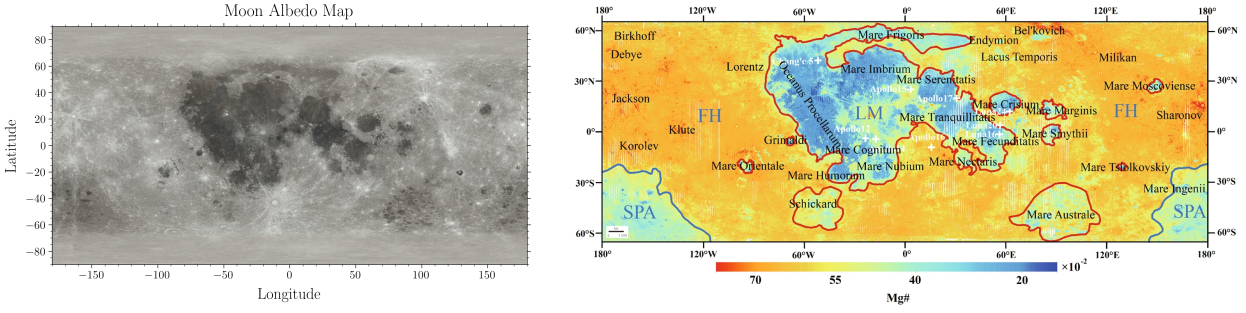

Figure 23: Left: Lunar Albedo map for reference throughout this report. **Right:** Map of the moon with detailed labels, from Yang et al. (2023). Colors represent Mg# intensities. We aim to recreate as many of these features as possible.

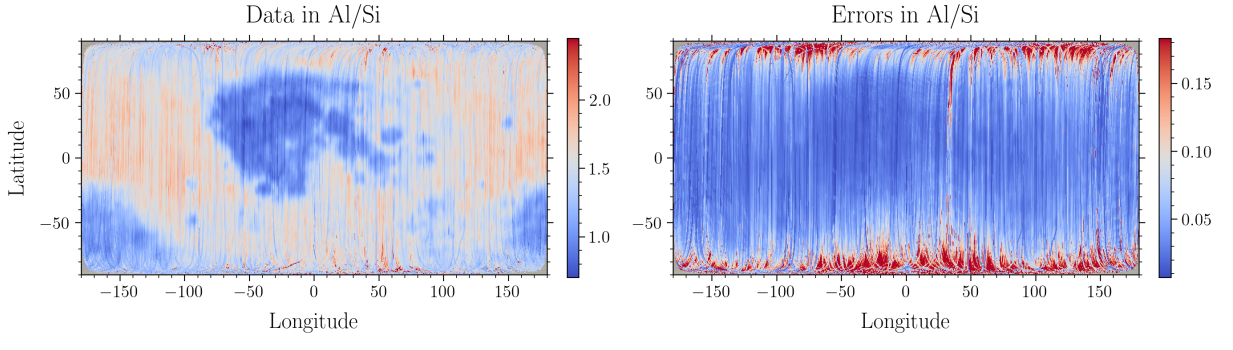

Figure 24: Left: Map of average aluminum line ratios for 96 s time bin. Separation between Maria, Highlands and South Pole Aitken (SPA) Basin is achieved to a great extent purely on the basis of this ratio alone. Smaller features such as various maria outside the central zone also appear as faint features. **Right:** Map of errors in the line ratios, we note higher errors in the $30° - 40°$ longitude region.

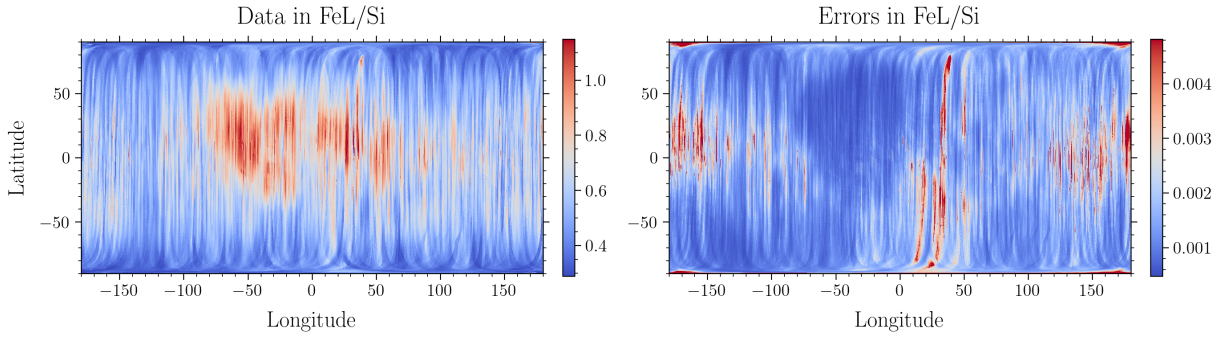

Figure 25: Left: Map of average iron line ratios for 296 s. Separation between Maria with higher Fe and Highlands with lower Fe is not achieved to the same extent as in figure 24. This is due to significantly lower detections of iron due to aforementioned difficulties with detecting the Fe_L line. **Right:** Map of errors in the line ratios, we note higher errors in the $30° - 40°$ longitude region as well as the sides.

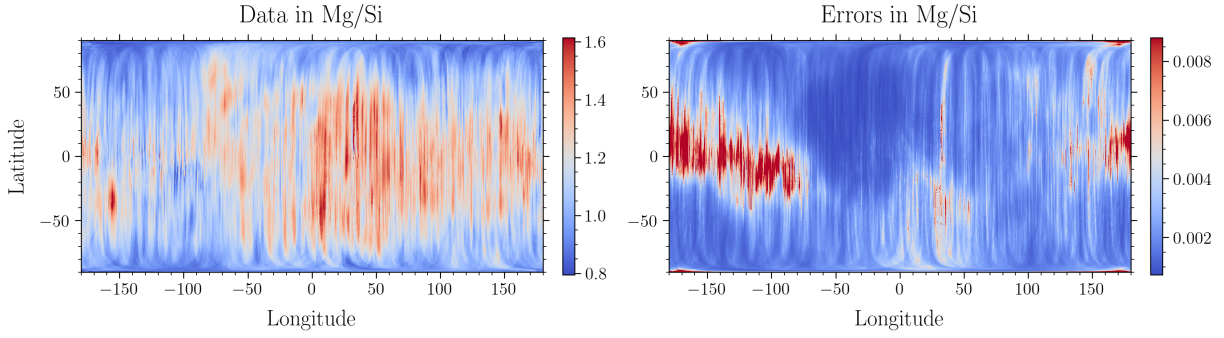

Figure 26: Left: Map of magnesium ratios for 296 s. This mapping is weakly similar to maps published in the literature, however, on its own isn't extremely useful to identify features on the moon. As we shall see, magnesium ratios in combination with others with reveal much more. **Right:** Map of errors in the line ratios, we note higher errors in the same region as figure 25.

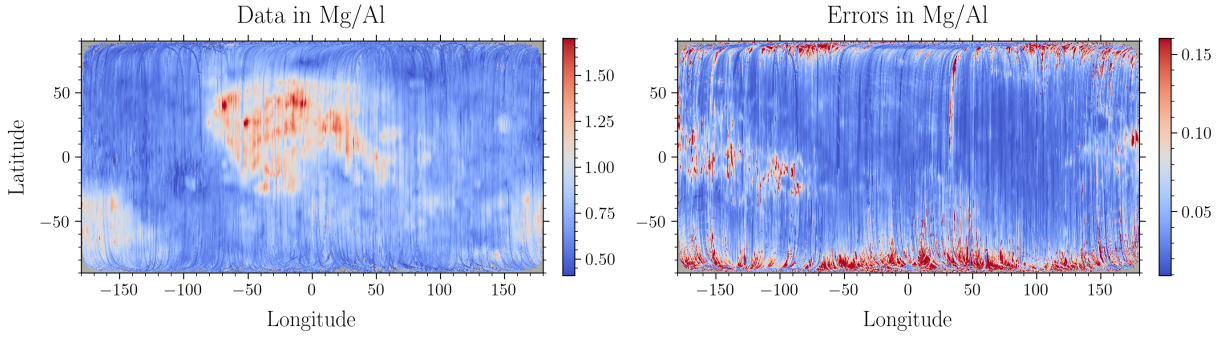

Figure 27: Map of average flux ratio of magnesium taken with respect to aluminum. This is a much more extensive map than found in Gloudemans, A. J. et al. (2021) which covered a smaller portion of the moon. We are able to recreate the gradients and relative values in the regions they mapped.

6. Best Ratios to Study Lunar Terrain Chemistry

The lunar surface chemical composition allows us insight into both the tectonic and volcanic processes that have shaped the Moon's crust and provides us a window to view the young Earth's composition, which has been papered over by more recent geological processes (Turkevich, 1973). Resolving questions around the composition of the Moon is thus important to gain selenological and geological insights. There are three sources of data we can use for this purpose, namely, meteorites, samples and radiation and have led us to characterize the elemental composition of over 99% of the mass of the lunar crust (Korotev, 2024a).

In this report, we focus on selecting the best combinations of elemental line ratios to visualize key compositional characteristics on the lunar surface, highlighting the most effective elemental combinations to distinguish between maria, highlands, and other prominent geological characteristics.

6.1. $\mathcal{R}^{\mathbf{Mg}}_{\mathbf{Si}}$ and $\mathcal{R}^{\mathbf{Al}}_{\mathbf{Si}}$

The most powerful combination of ratios, according to our analysis, comes from looking at the 2D histogram of $\mathcal{R}^{\mathrm{Mg}}_{\mathrm{Si}}$ and $\mathcal{R}^{\mathrm{Al}}_{\mathrm{Si}}$:

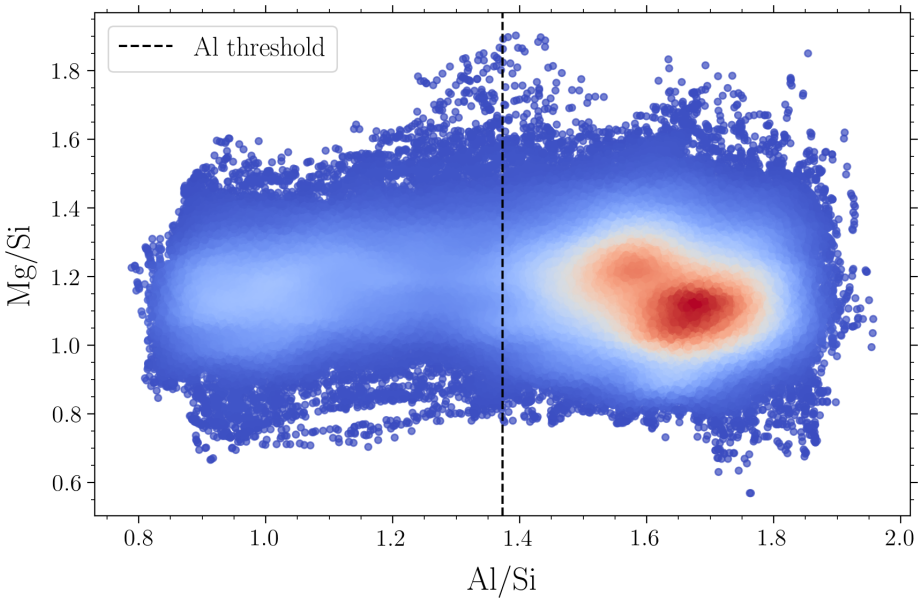

Figure 28: Scatter plot of average aluminium and magnesium line ratio. The plot visibly splits into two halves. We have plotted the individual histogram of aluminum ratios, to find the separating line between the right and left major clusters.

As can be seen, the plot (figure 28) easily bifurcates into a high Al and low Al region, which is to be expected given the already well known distribution of Al based on topography. On top of this, the high Al data points also visibly bifurcate into two clusters, which we will see correspond to selenographically distinct regions.

Continuing with our analysis (see figures 29, 30, 31, 32), we fit three 2D gaussians using a 3-component gaussian mixture model (GMM). The peak of each Gaussian is assigned a different color and the points in between are intermediately colored, representing their probability of belonging to any of the underlying populations assumed. Plotting on the lunar mercator projection with this same color scheme, we can see 3 regions distinctively. (1) The Basaltic Lunar Maria. (2) South Pole Aitken (SPA) Basin (3) the Feldspathic Highlands separate out pretty clearly, mostly on the basis of the distinctive variation in Al, as also was seen in the $\mathcal{R}^{\mathrm{Al}}_{\mathrm{Si}}$ map.

However, within the highland, we see a clear distinction between the higher and lower Mg clusters. Low Mg seems to pervade the northeast and wraps around to the northwest, which is in close agreement with the Chinese UV maps Yang et al. (2023); Wang et al. (2021). Alongside these, we can identify several secluded features on the moon's surface, such as the mare crisius and mare moscoviense. Resemblances of other mares can be found in figure 34, but nothing truly identifying.

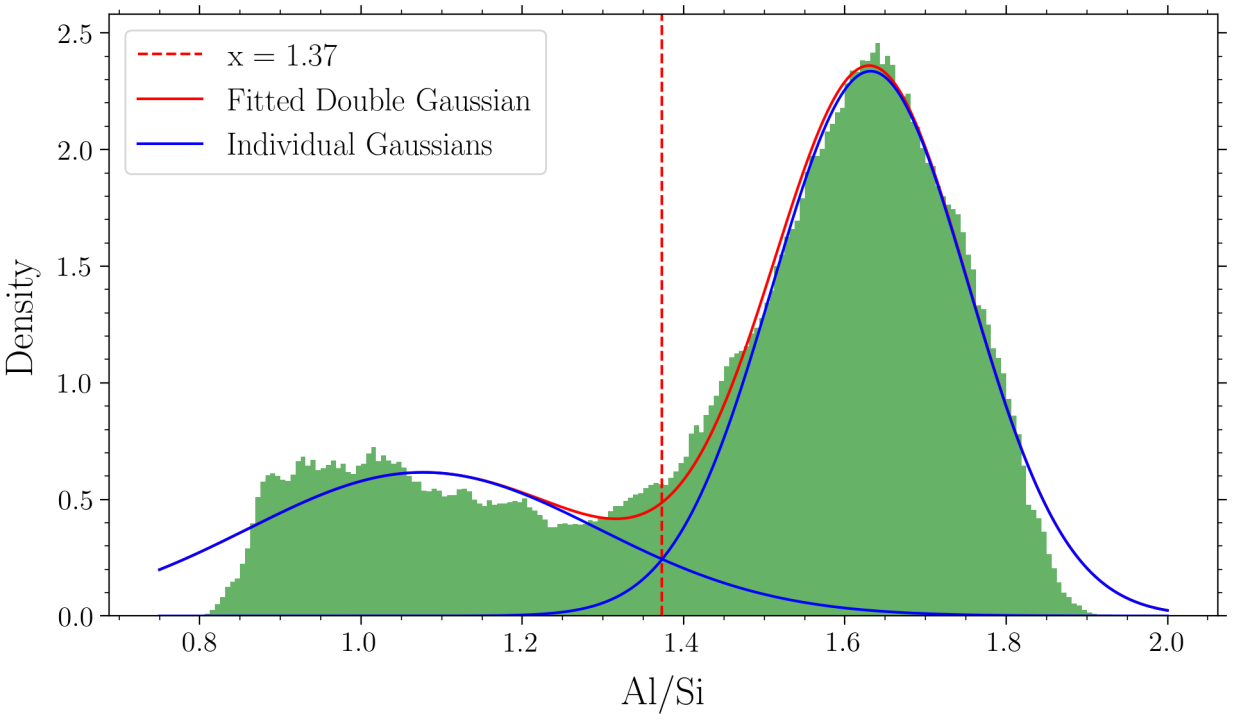

Figure 29: Double Gaussian fitting of the average aluminum ratio histogram over all points on the lunar map. This is a successful recreation of the result in Gloudemans, A. J. et al. (2021) where they also found a bimodal distribution for aluminum ratios.

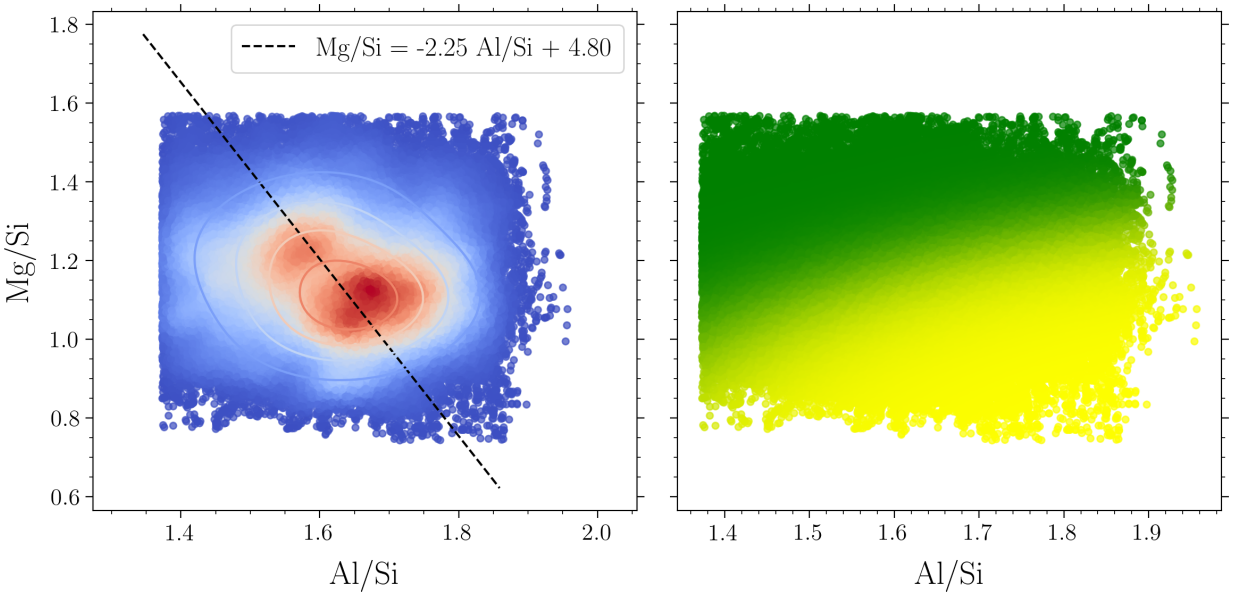

Figure 30: Left: Within the right cluster of the scatter plot figure, another splitting is observable, which we have modelled as two more 2D gaussians. A linear fit between the peaks of those gaussians has been provided. The fit is not good with a low R^2, which confirms non-linear modelling is the way to go. **Right:** Based on a gaussian mixture model, we have divided the sampled points into a probabilistic mixture of which population they belong to.

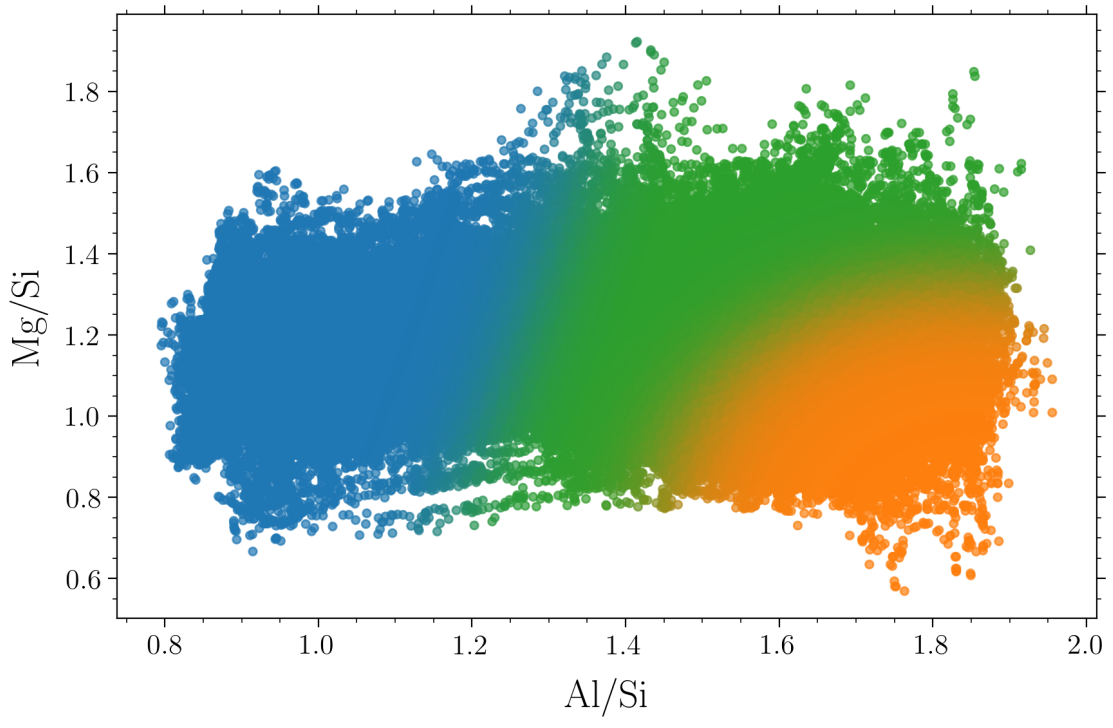

Figure 31: Left: Density heat map of full magnesium and aluminum scatter plot. **Right:** Based on a triple 2D gaussian mixture model, we have divided the sampled points into a probabilistic mixture of which population they correspond to selenologically continuous regions.

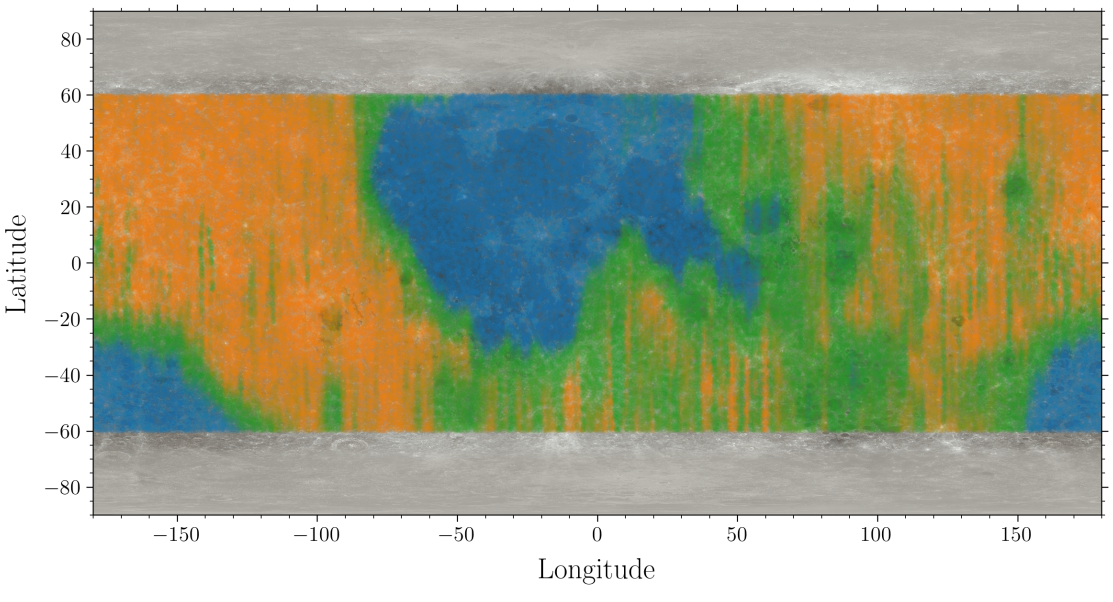

Figure 32: Lunar map based on the triple GMM model above. Maria and highland are well separated in this model. The highland itself is able to be divided into the relatively higher magnesium green region and the lower magnesium orange region. These regions map well to maps in Yang et al. (2023); Wang et al. (2021)

6.2. Mg#

A useful fraction that has been used in literature to characterize the variation of magnesium within lunar samples is the Magnesium Number (Mg#) which is the mole ratio of magnesium oxide in magnesium oxide and ferrous oxide mixture. We define an analogous quantity here which is defined in a similar vein with line ratios, i.e. $\mathcal{R}_{\mathrm{Si}}^{\mathrm{Mg}}/(\mathcal{R}_{\mathrm{Si}}^{\mathrm{Fe}}+\mathcal{R}_{\mathrm{Si}}^{\mathrm{Mg}})$

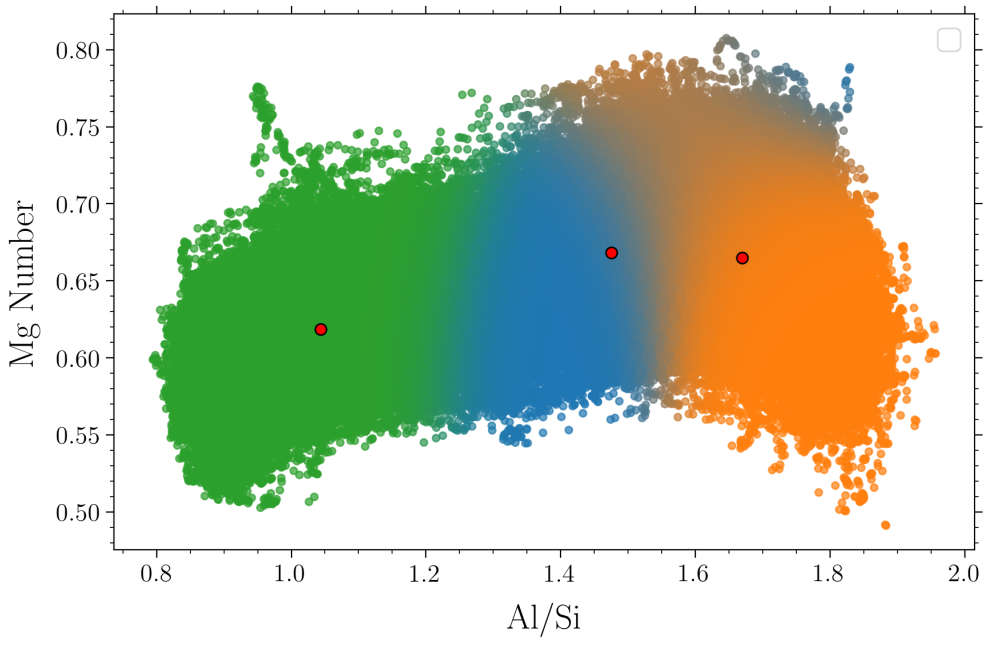

Figure 33: Magnesium number scattered against average aluminum ratios. A triple 2D gaussian mixture model has been applied and colors represent probabilities of belonging to certain populations.

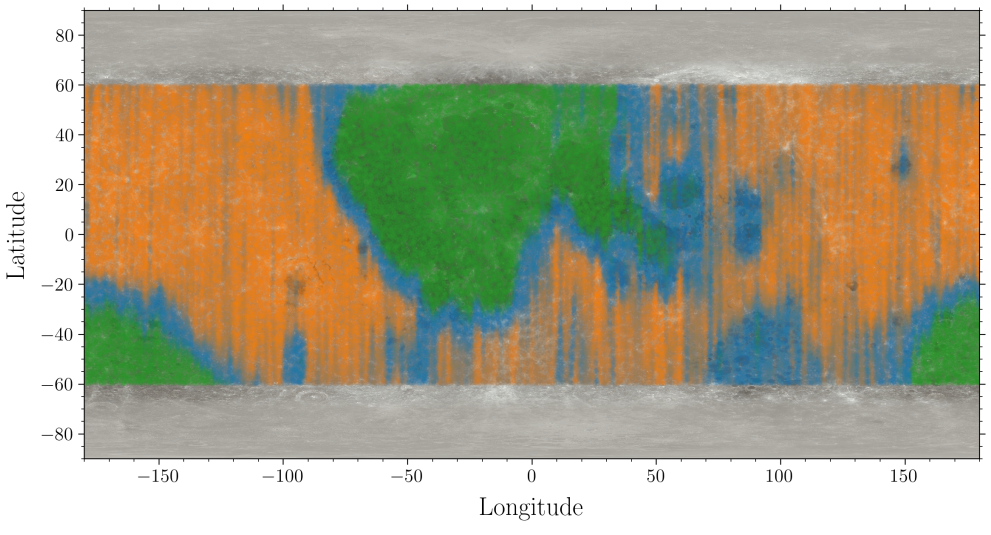

Figure 34: Lunar map of magnesium number. As can be seen, magnesium number is able to differentiate the surface into maria (green), highland (orange) and SPA basin (green) quite well. The blue zones seem to serve dual purposes, both representing the transition from highland to maria or SPA basin and spotting maria such as Mare Australe, Moscoviense and Orientale, separating them from the highland. The central maria region is not well differentiated in this classification.

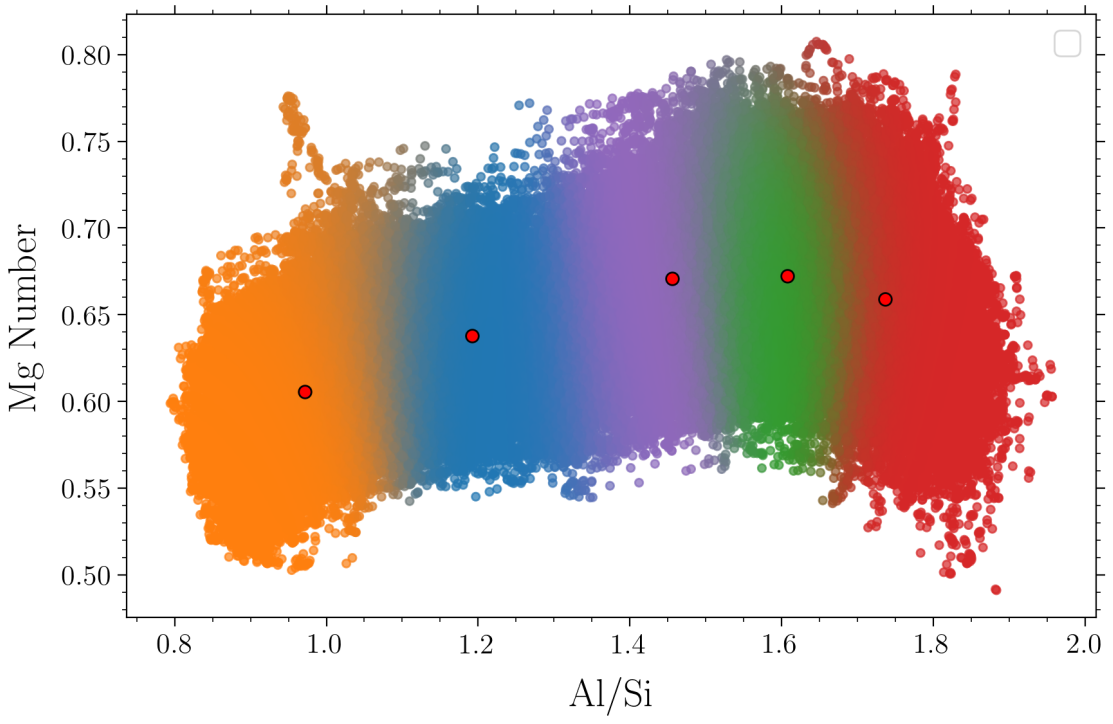

Figure 35: Going beyond figure 33, here we have applied a quintuple Gaussian mixture. Centroids of all the regions in composition space are given as red markers. Interpretation given in figure 36.

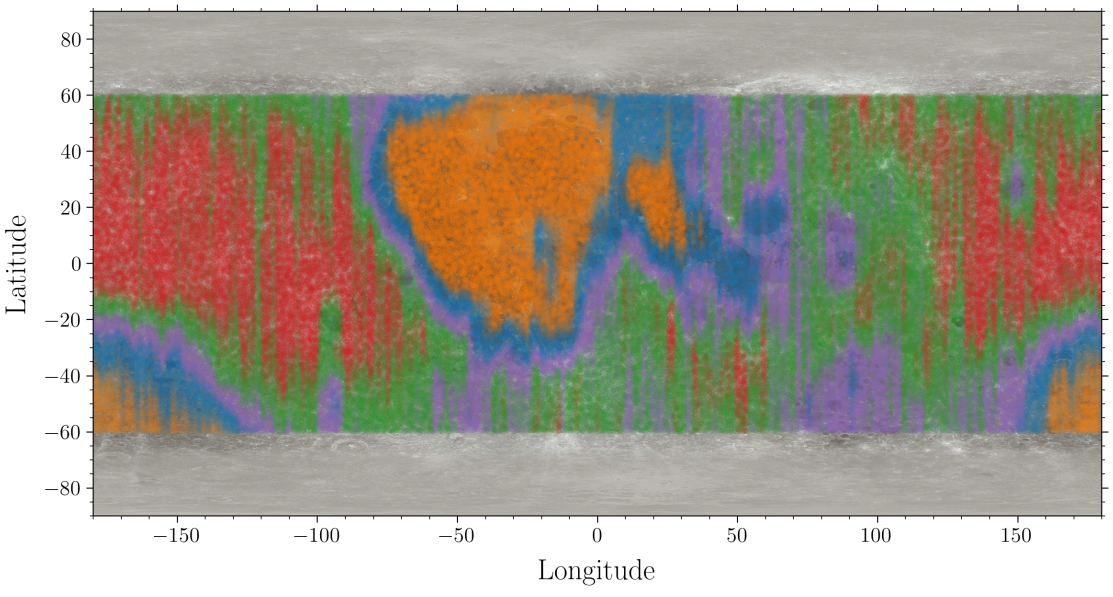

Figure 36: Quintuple Gaussian map, corresponding to figure 35. A better classification of the maria is enabled, with lower magnesium maria such as Tranquillatis and Oceanus Procellarum getting separated out. Within the highlands, the green region is sufficiently different from the red region, in the same way green was from orange in figure 32.

We note that using 5 gaussians to fit the 2D scatter plot as shown in figure 35 results in the gaussian components end up having decision boundaries nearly vertical, implying that the clustering is done purely on an Aluminum basis. We therefore conclude the triple clustering is superior, as adding more clusters washes out the variation in Mg#.

6.3. $\mathcal{R}^{\mathbf{Fe}}_{\mathbf{Si}}$ and $\mathcal{R}^{\mathbf{Al}}_{\mathbf{Si}}$

The inverse relationship between iron and aluminum is well known from the study of soil samples and lunar meteorites. Here we examine the scatter plots of average iron and aluminum ratios with respect to silicon and, as we shall show, we can differentiate regions within the Procellarum-Kreep-Terrane.

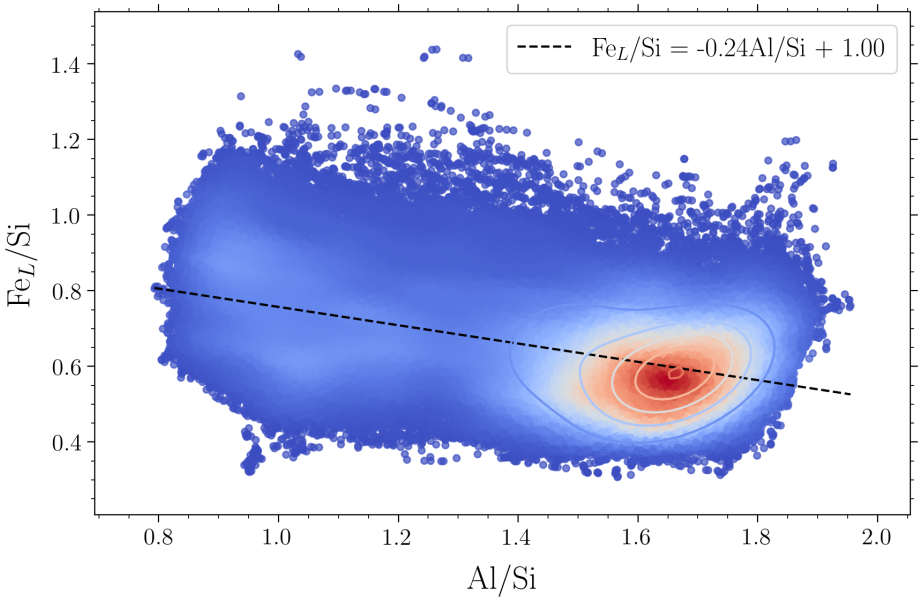

Figure 37: Scatter plot of average iron and aluminium line ratio. The plot visibly has a clear maxima with an overall linear trend. We have fit a straight line to the entire dataset and we get a weak linear fit with an R^2 of 0.05. Thus we prefer to go with a gaussian mixture model to model this histogram.

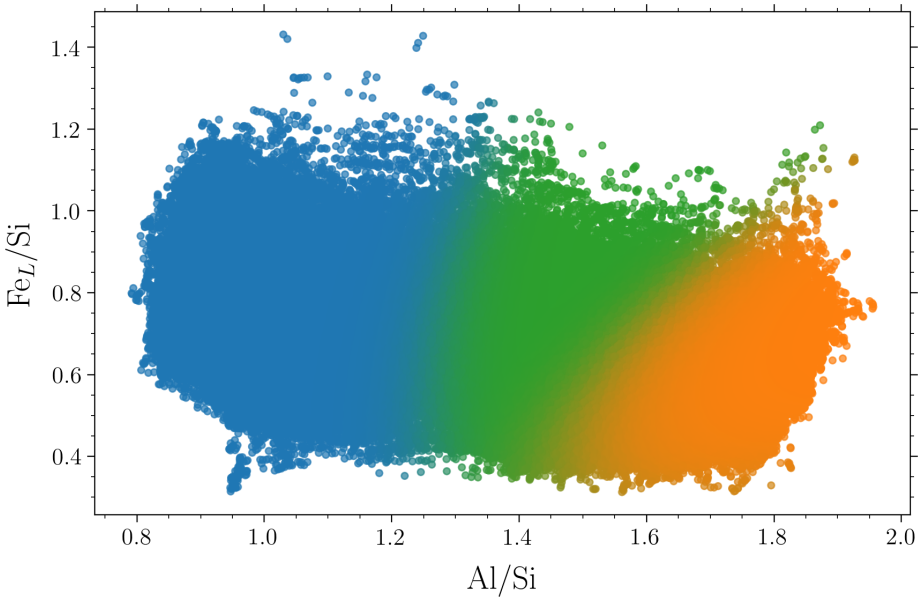

Figure 38: Scatter plot of average iron and aluminum ratios, with a triple gaussian mixture model. Higher aluminum zones once again split into two clusters. For interpretation, see figure 39.

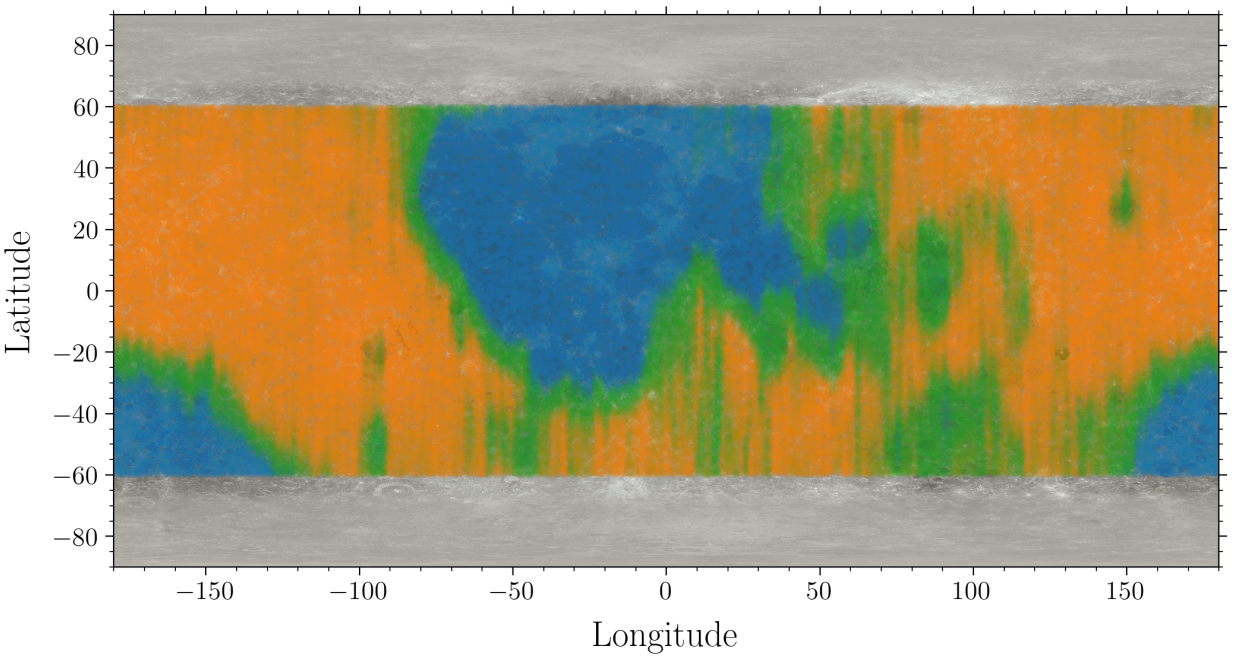

Figure 39: Lower aluminum regions cleanly end up representing the PKT and SPA basin. Smaller scale variation separating central maria is not captured by this three component model and as we shall see gets improved upon in the higher component model. Green zones seem to be capable of separating out smaller mare regions such as Mare Australe, Moscoviense, Orientale, Crisium, Marginis and Schickard.

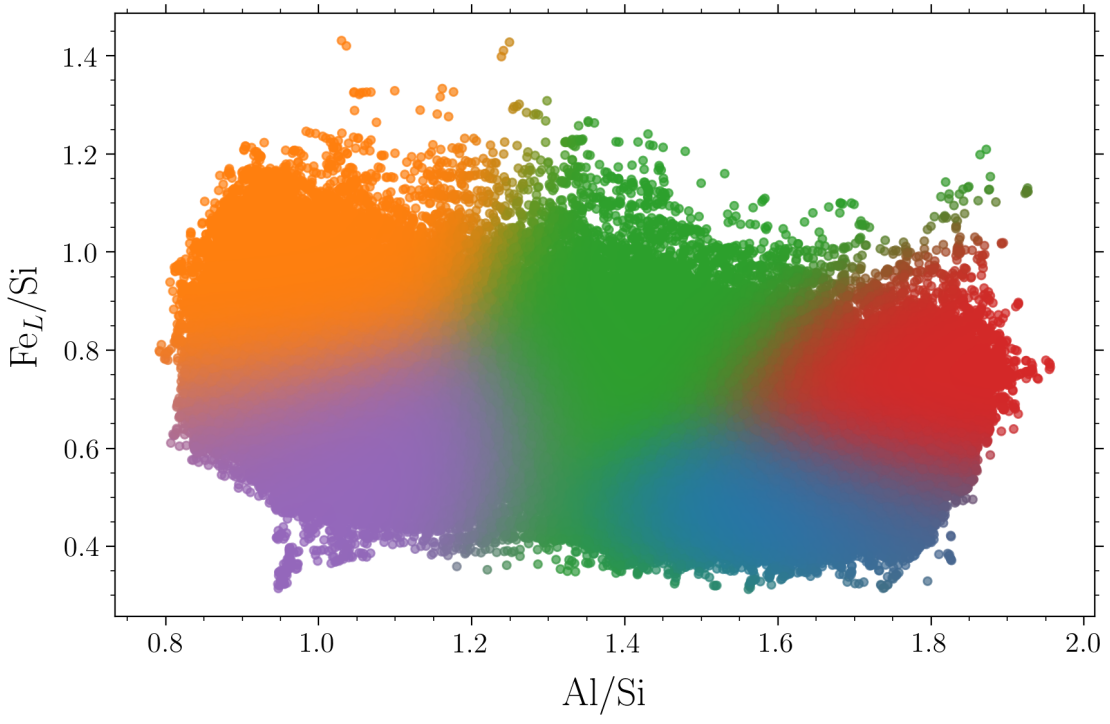

Figure 40: Scatter plot of average iron and aluminum ratios with a quintuple Gaussian mixture. The higher and lower aluminum parts of the figure get split into two more clusters, whilst middle aluminum remains mostly untouched.

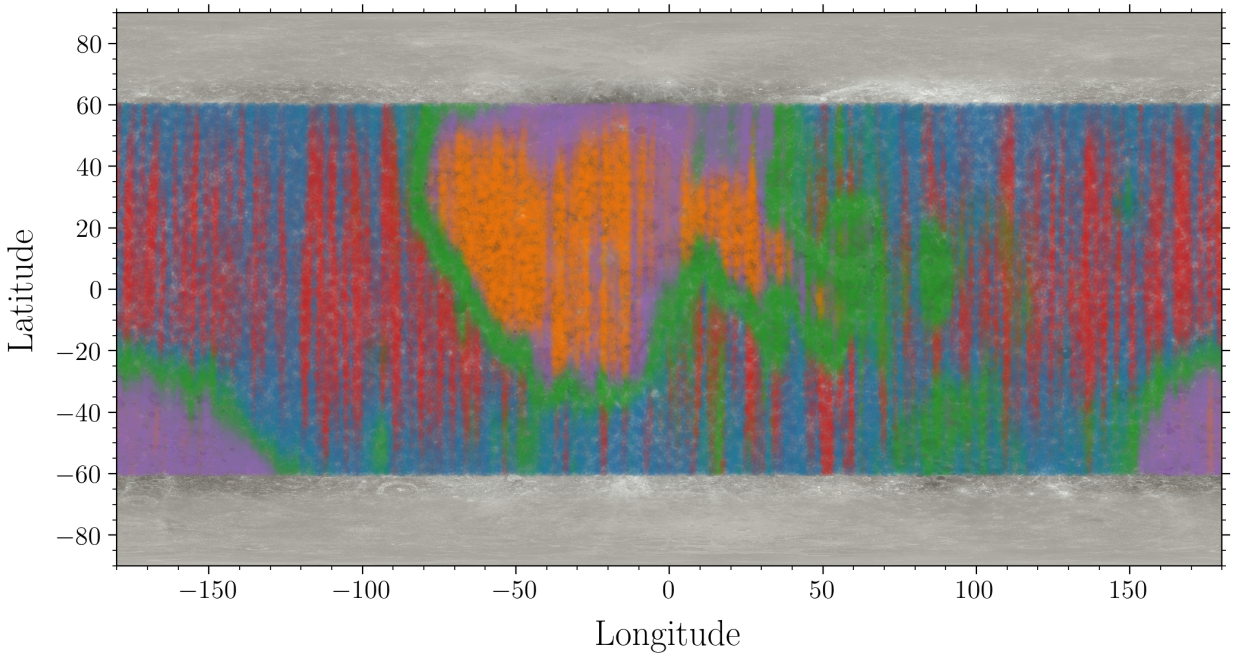

Figure 41: Quintuple Gaussian map, corresponding to figure 40. Similar to figure 36, the central mare region gets well distinguished, with spatial variation closely matching the previous literature on the weight distribution of iron on the moon (Yang et al., 2023; Wang et al., 2021). The blue and red classification doesn't have much geographical explanatory power and should be combined, leaving us effectively with a four component model.

7. Challenges and Lessons

7.1. Flare Detection Algorithm

In flare detection, we made initial attempts to classify flares by comparing the probability of observing a certain `light` spectrum given the distribution of background counts and applying a p-value hypothesis test. However, analysis of the p-value as a time series did not yield any significant trends and failed to correctly identify flare times, even with subtracted counts. One reason for the failure of this method was the discreteness in the cumulative distribution function constructed for background counts. To solve this, we developed an approach to quantify variations in subtracted counts by comparing it to the standard deviation of the background counts, which developed into the approach we have used currently.

7.2. Background Model Construction

While constructing a background model to subtract from `light` spectra, we discovered that a large chunk of the `dark` spectra had excessive excitations caused by the satellite not being fully occulted by the moon. These excessive excitations predominantly affected the Al line, causing it to go highly negative, leading to poor fits towards the poles. However, the occulted spectra are only a small fraction of the `dark` spectra, and requires us to consider a long-term background for enough statistics. Further, to adapt to the geotail effects as seen in figure 5, we group the spectra by phase of the moon.

7.3. Detection of Ca, Fe and other 3d-series elements

The 3d-series are elements of interest for abundance mapping, but only Fe is present in large abundances on the moon, the others are only expected in trace amounts. As discussed in §2.4.2, these elements emit L lines which are too low energy and are clouded by noise. The $\mathrm{Fe_L}$ line is just far enough from the $O_{\mathrm{K}-\alpha}$ line to be detectable, but presents challenges to fit gaussians. The approaches to resolve this are discussed in figure 13. The K-lines for these elements are unable to capture major selenographical features and thus are ill-suited for inference regarding these elements, as seen in figures 42, 43, and 44. Since these lines only appear weakly in the spectra, the amplitude cut for the gaussian fit described in equation 4 was removed to produce the maps.

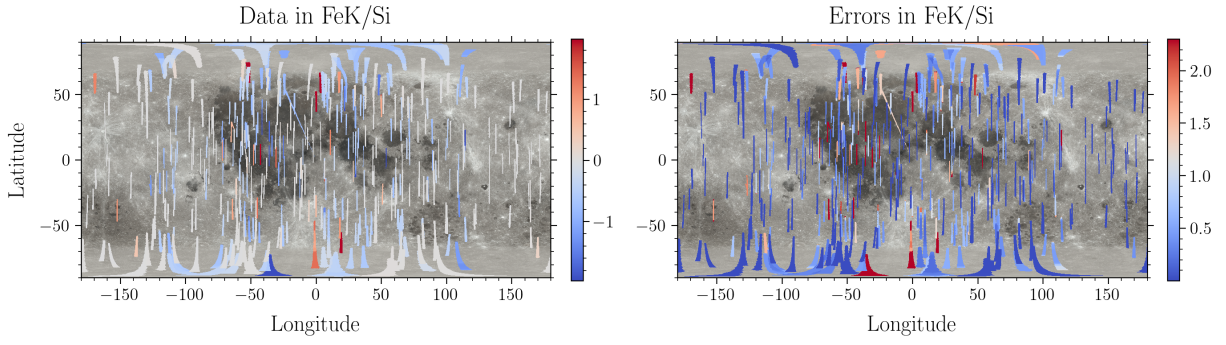

Figure 42: The coverage of Fe_K with 296s binning is fragmented and doesn't capture the features obtained through Fe_L

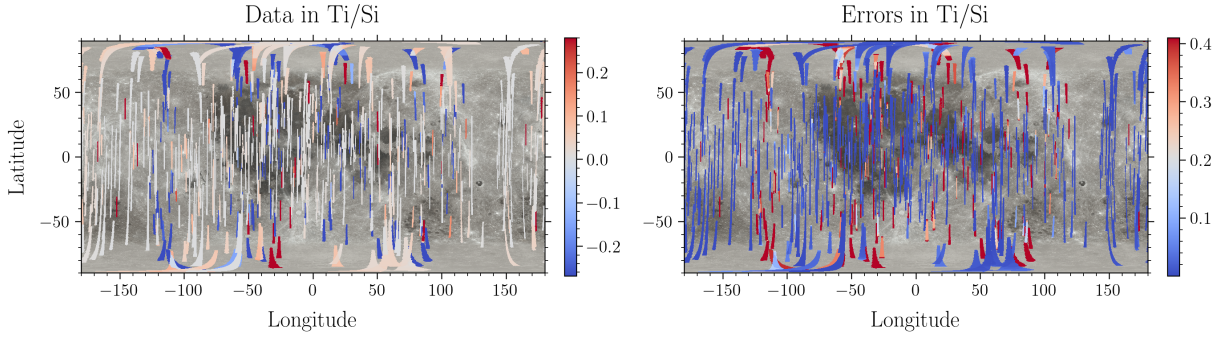

Figure 43: Coverage of Ti_K lines with 296s binning is larger than Fe_K but doesn't show significant selenographic contrast

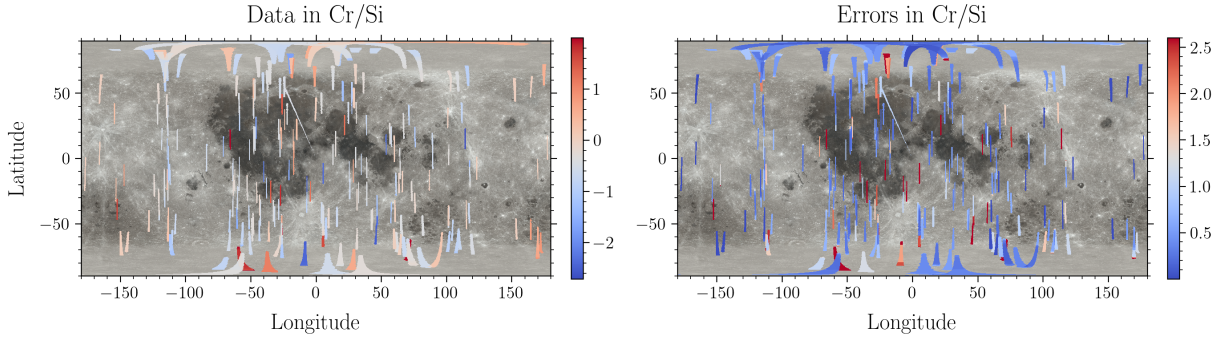

Figure 44: Coverage of Cr_K lines with 296s binning is minimal due to low abundance and XRF cross-section of the lines

Ca emits K-lines at excitation greater than 4 keV, and L-lines at $1-4$ keV (Brunetti et al., 2004), hence it requires strong flares to emit detectable XRF. However, since Ca is present in significant abundance, the $\text{Ca}_{\text{K-}\alpha}$ and $\text{Ca}_{\text{K-}\beta}$ lines still cover the entire lunar surface, as seen in figures 45, 46. The $\text{Ca}_{\text{K-}\alpha}$ data faintly distinguishes the maria from the highlands, and if added to the $\text{Ca}_{\text{K-}\beta}$ it may show further contrast. However, calcium based ratio groups were found to be ill-suited for inference based on clustering. Further, an analysis of the $\log \mathcal{R}_{\text{Si}}^{\text{Ca}_{\text{K-}\alpha}}$ shows a unimodal profile.

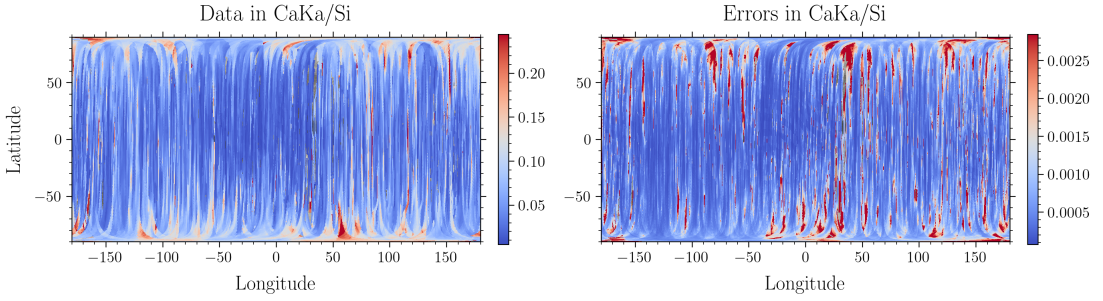

Figure 45: Faint separation of maria and highland observed through $\text{Ca}_{\text{K-}\alpha}$/Si data with 296s binning

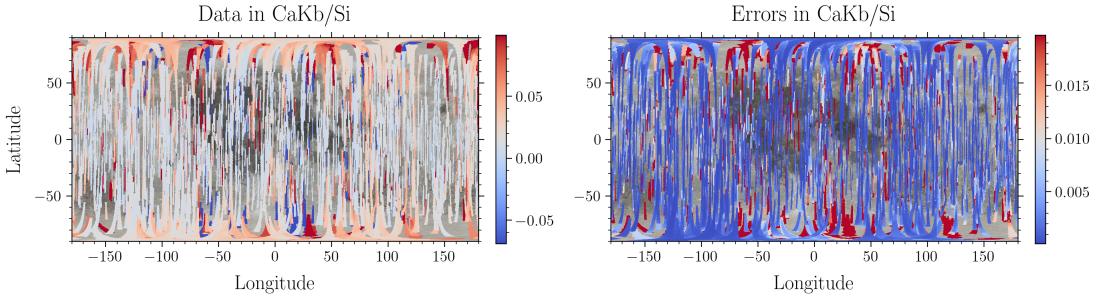

Figure 46: $\text{Ca}_{\text{K-}\beta}$ line detections with 296s binning may be used to add to the above $\text{Ca}_{\text{K-}\alpha}$ data to obtain better maps

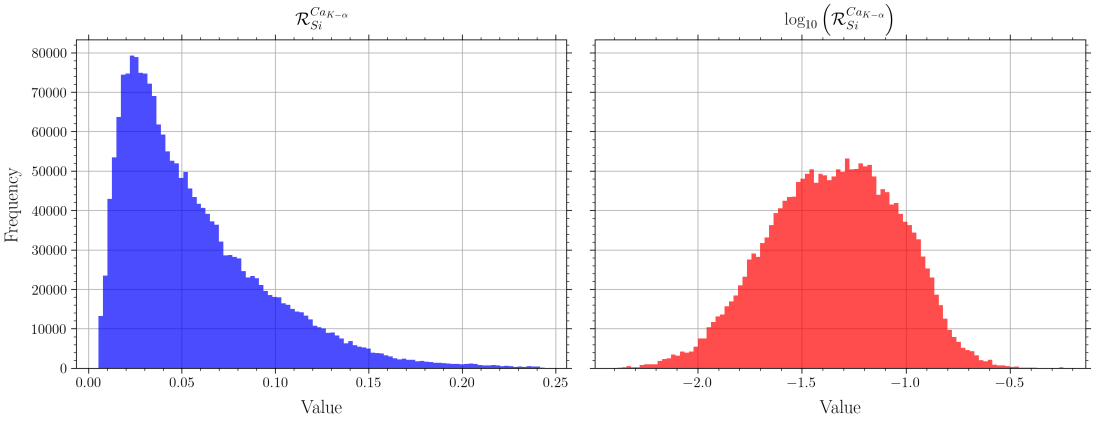

Figure 47: Left: Lognormal distribution shown by $\mathcal{R}_{\text{Si}}^{\text{Ca}_{\text{K-}\alpha}}$ **Right:** Histogram of $\log \mathcal{R}_{\text{Si}}^{\text{Ca}_{\text{K-}\alpha}}$ shows a likely gaussian distribution

7.4. Oxygen line

We attempted to fit the oxygen line with single gaussians as described in §2.4.2. The fit has greater uncertainty due to the instrumental noise present in that energy range. We also do not obtain enough coverage due to points being rejected by our gaussian fit criteria defined in equation 4.

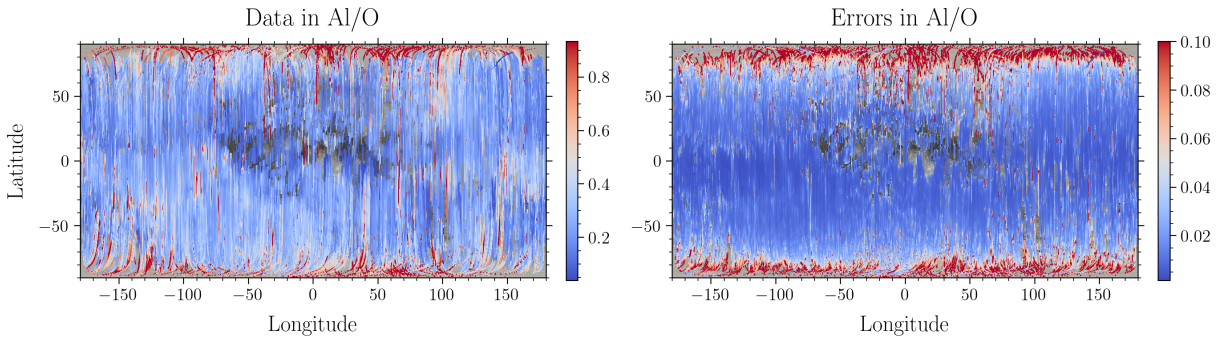

Figure 48: The map of $\mathcal{R}_{\mathrm{O}}^{\mathrm{Al}}$ shows sparser coverage compared to $\mathcal{R}_{\mathrm{Si}}^{\mathrm{Al}}$

7.5. Ratio Clustering

We attempted to cluster the derived ratio values along with spatial correlation using multiple approaches and models to do the clustering, including trying to cluster all ratio values at once, clustering 4-dimensionally with values and coordinates. However, 4-dimensional clustering requires carefully choosing an appropriately weighted distance metric, since it can affect the results significantly. Similarly, we are unable to infer new information by clustering all 3 ratios Fe/Si, Mg/Si and Al/Si as it is harder to analyze as compared to 2-at-a-time clustering. Thus, we choose to analyze ratio value clusters 2-at-a-time and use the spatial correlation to validate the clustering.

7.6. Computation

Since the data is quite large (500 GB), computational constraints such as runtime and memory are important considerations. We make use of multiprocessing as far as possible, such as separating dark and light spectra, constructing phase-wise background masterfiles, fitting gaussians, and mapping across different years. Due to the presence of so many files, a large number of `open` system calls are made, which can cause the file handlers of the OS to fail. So, small FITS files must be read with the `memmap=False` setting in `astropy.io`'s `fits.open` function. The entire pipeline running on all the data can take up to 5 hours to finish, with parallelization across years.

The detection pipeline requires reading a large number of files, hence it has the largest memory requirement. It consumes 6 GB of RAM per year of processing (30 GB if all years are done in parallel). The mapping program requires around 2.5 GB per year of data (12.5 GB if all years are done in parallel).

8. Future Work

The following ideas can be used to extend our analysis:

- The clustering model can be used to make a classifier which classifies samples and predicts the location on the moon. Stronger clustering models like supervised neural networks can be used with training from ground truth data, to extract non-linear trends in compositional ratios effectively.
- We may be able to analyze the maps using other representative ratios, like the Mg number. For example, we can analyze an analogous "Ca number" and possibly use $\log \mathcal{R}_{\mathrm{Si}}^{\mathrm{Ca_{K\text{-}\alpha}}}$ to perform clustering.
- We can try to combine the fluxes for different lines (K and L lines, or K-α and K-β lines) of a given element using cross-section information in order to get higher coverage and contrast.

- A possible method to improve the fit of the oxygen line is to fit a half gaussian in the $0.52 - 0.65$ keV range, as the lower window ($0.4 - 0.52$ keV) is near the limits of the detector's capabilites and would deviate from a gaussian profile. Thus it would be beneficial to fit the gaussian on the "good" side of the $O_{\text{K-}\alpha}$ line.
- As seen in figure 5, the background readings during full moon can reveal information about the geotail and its duration. We can consider finer binning on the phases of the moon to determine the exact duration of background rise due to geotail, and refine our background model.
- The background update is dynamic, but it can be further optimized for online updates. The update can be made efficient by storing and updating only mean backgrounds, and it can be made real-time by polling for new data continuously, and the memory usage of the program can also be reduced.

Acknowledgement

We acknowledge the use of data from the Chandrayaan–II, second lunar mission of the Indian Space Research Organisation (ISRO), archived at the Indian Space Science Data Centre (ISSDC)

Software: `Python v3.10.15` (Van Rossum & Drake, 2009), `NumPy` (Harris et al., 2020), `SciPy` (Virtanen et al., 2020), `Astropy` (Astropy Collaboration et al., 2013, 2018, 2022), `Pandas` (McKinney et al., 2010), `Matplotlib` (Hunter, 2007), `Specutils` (Earl et al., 2024), `Rasterio` (Gillies et al., 2013–), `Shapely` (Gillies et al., 2007–), `QGIS` (QGIS Development Team, 2024),

References

Ackermann, M., Ajello, M., Albert, A., et al. 2014, The Astrophysical Journal, 787, 15, doi: 10.1088/0004-637X/787/1/15
Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, , 558, A33, doi: 10.1051/0004-6361/201322068
Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, , 156, 123, doi: 10.3847/1538-3881/aabc4f
Astropy Collaboration, Price-Whelan, A. M., Lim, P. L., et al. 2022, , 935, 167, doi: 10.3847/1538-4357/ac7c74
Athiray, P., Narendranath, S., Sreekumar, P., Dash, S., & Babu, B. 2013, Planetary and Space Science, 75, 188, doi: https://doi.org/10.1016/j.pss.2012.10.003
Brunetti, A., Sanchez del Rio, M., Golosio, B., Simionovici, A., & Somogyi, A. 2004, Spectrochimica Acta Part B: Atomic Spectroscopy, 59, 1725, doi: https://doi.org/10.1016/j.sab.2004.03.014
Böhm-Vitense, E. 1992, Introduction to Stellar Astrophysics (Cambridge University Press)
del Hoyo-Meléndez, J. 2018, X-RAY FLUORESCENCE (XRF) SPECTROMETRY, 257–262, doi: 10.1515/9781942401353-028
Earl, N., Tollerud, E., O'Steen, R., et al. 2024, astropy/specutils: v1.18.0, v1.18.0, Zenodo, doi: 10.5281/zenodo.13942238
Gillies, S., et al. 2007–, Shapely: manipulation and analysis of geometric objects. https://github.com/Toblerity/Shapely
—. 2013–, Rasterio: geospatial raster I/O for Python programmers. https://github.com/mapbox/rasterio
Gloudemans, A. J., Kuulkers, E., Campana, R., et al. 2021, AA, 649, A174, doi: 10.1051/0004-6361/202140321
Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, Nature, 585, 357–362, doi: 10.1038/s41586-020-2649-2
Heiken, G. H., Vaniman, D. T., & French, B. M. 1991, Lunar Sourcebook, A User's Guide to the Moon
Heiles, C., & Troland, T. H. 2003, The Astrophysical Journal Supplement Series, 145, 329, doi: 10.1086/367785
Hunter, J. D. 2007, Computing in Science & Engineering, 9, 90, doi: 10.1109/MCSE.2007.55
Juvela, Mika, & Tharakkal, Devika. 2024, A&A, 685, A164, doi: 10.1051/0004-6361/202349044
Kirchner. 2016, Data Analysis Toolkit #5: Uncertainty Analysis and Error Propagation, https://seismo.berkeley.edu/~kirchner/Toolkits/Toolkit_05.pdf
Korotev, R. L. 2024a, The chemical composition of lunar soil | Some Meteorite Information | Washington University in St. Louis — sites.wustl.edu. https://sites.wustl.edu/meteoritesite/items/the-chemical-composition-of-lunar-soil/
—. 2024b, How do we know that it is a rock from the moon? | Some Meteorite Information | Washington University in St. Louis — sites.wustl.edu. https://sites.wustl.edu/meteoritesite/items/how-do-we-know-that-its-a-rock-from-the-moon/
Mason, J. 2022, INSPIRESat-1/DAXSS (AKA MinXSS-3) Level 1 data available — lasp.colorado.edu, https://lasp.colorado.edu/minxss/2022/06/06/inspiresat-1-daxss-aka-minxss-3-level-1-data-available/
McKinney, W., et al. 2010, in Proceedings of the 9th Python in Science Conference, Vol. 445, Austin, TX, 51–56
Narendranath, S., Pillai, N. S., Bhatt, M., et al. 2024, Icarus, 410, 115898, doi: https://doi.org/10.1016/j.icarus.2023.115898
Pillai, N. S., Narendranath, S., Vadodariya, K., et al. 2021, Icarus, 363, 114436, doi: https://doi.org/10.1016/j.icarus.2021.114436
QGIS Development Team. 2024, QGIS Geographic Information System, QGIS Association. https://www.qgis.org
Sharma, Y., Marathe, A., Bhalerao, V., et al. 2021, Journal of Astrophysics and Astronomy, 42, 73, doi: 10.1007/s12036-021-09714-6
Shearer, C. K., Elardo, S. M., Petro, N. E., Borg, L. E., & McCubbin, F. M. 2015, American Mineralogist, 100, 294, doi: doi:10.2138/am-2015-4817
Taylor, J. 1997, An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements, ASMSU/Spartans.4.Spartans Textbook (University Science Books). https://books.google.co.in/books?id=ypNnQgAACAAJ
Taylor, S. R., Norman, M. D., & Esat, T. M. 1993, in Lunar and Planetary Science Conference, Lunar and Planetary Science Conference, 1413
Turkevich, A. L. 1973, The moon, 8, 365, doi: 10.1007/BF00581730
Van Rossum, G., & Drake, F. L. 2009, Python 3 Reference Manual (Scotts Valley, CA: CreateSpace)
Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, Nature Methods, 17, 261, doi: 10.1038/s41592-019-0686-2
Wang, X., Zhang, J., & Ren, H. 2021, Planetary and Space Science, 209, 105360, doi: https://doi.org/10.1016/j.pss.2021.105360
Yang, C., Zhang, X., Bruzzone, L., et al. 2023, Nature Communications, 14, 7554, doi: 10.1038/s41467-023-43358-0
Zhang, J., Temmer, M., Gopalswamy, N., et al. 2021, Progress in Earth and Planetary Science, 8, 56, doi: 10.1186/s40645-021-00426-7
Zhang, L., Zhang, X., Yang, M., et al. 2023, Icarus, 397, 115505, doi: https://doi.org/10.1016/j.icarus.2023.115505