ANALYSIS OF THE EFFECTS OF COMPOSITIONAL ELEMENTS, DETECTOR TEMPERATURE, AND GEOGRAPHY ON THE VARIATION IN AMOUNT OF WATER FOUND ON SURFACE OF THE MOON

An Honors Thesis Presented

By

RITIK SHAH

Approved as to style and content by:

**** Mario Parente 05/25/23 17:37 ****
Chair

**** Marco F Duarte 05/26/23 19:49 ****
Committee Member

**** Dennis L Goeckel 05/29/23 18:24 ****
Honors Program Director

ABSTRACT

Discovering water on the moon was a major milestone in the frontier of space exploration, upon which a lot of research has been done to investigate the presence of water in the form of ice on the lunar surface. However, water is not only present on the moon in the form of ice, in fact there is water embedded into the soil in various regions scattered throughout the surface of the moon. My project in conjunction with NASA's Lunar Data Analysis Program (LDAP) involves analyzing the composition of lunar terrains and variables including but not limited to detector temperature, maturity, and geography to see which are best predictors for the variation in the abundance of water found. By developing a regression model using spectral parameters which indicate composition as independent variables and a spectral parameter which indicates the presence of water as the dependent variable, I was able to find that detector temperature, which varies according to the time of day on the moon, had the most significant effect on the results of this regression.

Table of Contents :

Acknowledgements:

I wish to express my sincere gratitude to Dr. Christian Wohler from Dortmund University of Technology in Germany who generously contributed additional information related to particular characteristics of the reflectance signals. I would also like to thank NASA's Lunar Data Analysis Program (LDAP) for allowing me to be a part of their groundbreaking research. Finally I would like to extend my heartfelt appreciation to my advisor Professor Mario Parente, without whom this research experience would not have been possible, for his immeasurable support, and invaluable feedback through the whole process.

1. INTRODUCTION

The discovery of water on the moon represented a landmark accomplishment in the field of space exploration and sparked in-depth research on the possibility of ice as a water source on the lunar surface. On the moon, however, water is not only present as ice. Surprisingly, water is also woven into the soil of several areas dispersed around the lunar surface.

My thesis is a part of a broad investigation carried out by NASA's Lunar Data Analysis Program (LDAP). The main goal of the LDAP team's study is to carefully analyze how the strength of the spectral OH/H_2O band, which ranges from 2.7 to 3 microns, changes throughout the day on the Moon. To do this, researchers make use of the invaluable data gathered by the Moon Mineralogy Mapper (M3) in order to understand the type and quantity of hydroxyl and/or water-bearing compounds. In this approach, near-infrared hyperspectral images are carefully examined in search of a specific spectral signature that can be used as a reliable indication of the presence of water or hydroxyl. Given that my thesis is a part of this vast LDAP project, it requires a thorough research of lunar terrains that takes into account a variety of factors like temperature, maturity, and geography. Finding the most effective forecasters of changes in water abundance is the main goal. I will use supervised machine learning to create a regression model to achieve this. The model will make use of spectral parameters as independent variables that provide information about composition and a

spectral parameter that indicates the presence of water as the dependent variable. By using this comprehensive strategy, I hope to improve the knowledge of lunar water distribution and further the overall goals of the LDAP mission.

The data I am analyzing is obtained from an instrument called the Moon Mineralogy Mapper (M3). M3, which was created by NASA and launched by ISRO's Chandrayaan 2 mission, has been instrumental in gathering reflectance values from various areas of the lunar surface. However, it can be difficult to determine the meaning of this raw data without additional information. I overcame this problem by using a set of equations provided by the NASA M3 team, which allowed the development of numerous parameters corresponding to various compositions found on the moon's surface, such as Olivine, Pyroxene, and others.

Additionally, Dr. Christian Wohler from Dortmund University of Technology in Germany generously contributed additional information related to particular characteristics of the reflectance signals, such as band depth and continuum slope, to further enhance my research. The dataset's depth and richness were increased by the availability of Geotiff photos of lunar craters. While the main goal of my thesis was to use this extensive data to investigate whether characteristics may be used to accurately anticipate changes in water distribution, there was a considerable degree of complexity. It was critical to distinguish between the two since the reflectance signal that shows the presence of water (H2O) also signals the presence of the hydroxyl

radical (OH). To obtain accurate and precise results, it was crucial to address this obstacle.

I am confident that undertaking a thorough analysis into the crucial factors determining the presence of water on the lunar surface and successfully differentiating between water and hydroxyl offers enormous importance for future research initiatives in this sector. Such findings represent a significant step forward, bringing scientists closer to solving many riddles that are currently cloaked in doubt or ignorance. The goal of this research project is to produce profound insights that will make it easier to identify areas on the moon that may contain significant water reservoirs. To do this, a detailed analysis of the elemental composition in and around these areas will be used, ideally with high accuracy.

It is impossible to overestimate the importance of such a finding, especially in the context of space exploration. Future expeditions might be able to concentrate their efforts more effectively with a better understanding of the amount of water present on the moon, easing the harvest of this priceless resource. There are practically endless applications for this recovered water, which emphasizes how important and practical this discovery is. The ramifications go well beyond what we can imagine, from maintaining human presence throughout protracted lunar trips to acting as a crucial resource for several scientific researches.

Additionally, this research paves the way for developments in allied fields of study. It establishes the groundwork for future investigations into the moon's

geological history, the possibility that it could support life, and the larger implications for planetary science by improving our understanding of the distribution of lunar water and polishing our capacity to distinguish between water and hydroxyl. In the end, this project offers a critical contribution to the rapidly increasing field of space research and creates a plethora of possibilities for future advancements in science.

2. REVIEW OF LITERATURE

The primary texts I concentrated on included important elements that complemented the research that I carried out. Each reading gave thorough explanations of the several variables that can affect the results of my regression models. Remarkably, these sections closely mirrored the thinking process I embarked upon right from the beginning of this research endeavor. It is essential to delve into the intricate details pertaining to the geological properties of the moon, the intricate composition of minerals found on its surface, the data acquisition procedures used by the moon mineralogical mapper, the statistical exploration of the acquired data, the regression process itself, and the diverse array of regression models in order to gain a thorough understanding of the methodology used in my research and my subsequent findings.

1. **Areas of the moon :**

The highlands and mares, two separate geological landforms that embellish the lunar surface, each have distinctive qualities and captivating insights. The brighter areas that embellish the lunar expanse are the highlands, which are radiant and severely scarred. The oldest regions of the moon's surface are these ancient terrains, which are weighed down by innumerable impact craters. These heavenly landscapes, which are predominantly made up of anorthosite and other siliceous rocks, have an abundance of aluminum and calcium, which makes them extraordinarily bright and luminous in the blackness of the night sky (1). The highlands rise above the lunar terrain and display magnificent formations that extend many kilometers into the depths.

In contrast, the darker, smoother craters that beautify the lunar landscape are the mare, with their obscure allure. These zones, made of basaltic and olivine-rich rocks (2), are the result of volcanic activity that formerly raged on the moon's surface. The lunar depths boldly let off molten lava, which bravely flowed before finally cooling and giving way to solidification. The mare were born during this change, carving their presence in sharp contrast to their more colorful highland counterparts. These areas are younger than the highlands because their formation is linked to the lunar volcanic eruptions. Their surfaces exhibit a tranquil flatness and silkiness that catches the observer's attention.

This distinction between the highlands and mare is crucial because it reveals the subtle differences of their reflecting signals, which are anchored in their fundamental composition. It was discovered that the solution to interpreting the many reflecting characteristics that result from these two lunar landscapes' dissimilar compositions is by realizing how drastically different they are from one another.

2. **The Moon Mineralogy Mapper (M3) :**

The Moon Mineralogy Mapper (M3) is a highly advanced spectrometer instrument that has been designed specifically to precisely map the surface composition of the moon. This allows scientists and researchers to better understand the geology and mineralogy of the moon. This cutting-edge apparatus meticulously examines the intensity of light that is reflected off the surface of the moon, encompassing a broad range of wavelengths ranging from 0 to 2900 nanometers, using a ground-breaking method called reflectance spectroscopy.

The M3 technology has the incredible ability to deliver priceless reflectance data throughout an outstanding spectrum of 84 different channels, each of which corresponds to particular wavelengths found between 0 and 2700 nanometers. Two separate absorption bands can be made of these wavelengths:

- the 1-micron band (800–1100 nm)
- the 2-micron band (1750–2000 nm)

Unfortunately, the most interesting band, the 3-micron band (2700–3000 nm), which has attracted the most attention due to the possibility of water or hydroxyl being present, is not present in the M3 data. Researchers have found fascinating signs of water or hydroxyl molecules in this 3-micron band.

Critical information regarding the composition of the moon's surface has been disclosed by the M3 instrument's priceless dataset. Plagioclase and pyroxene, two essential minerals, have been found to make up the majority of the lunar surface. However, olivine, ilmenite, and anorthite have also been found in trace amounts. These incredible findings were made possible by a careful examination of the reflectance signals given off by various minerals, which allowed researchers to establish links between the presence of particular minerals and particular wavelength bands(3).

The cornerstone of the data used in this study project is the reflectance signals collected from the M3 sensor. With the use of a thorough study of the 3000 nm (3-micron) band, I was able to investigate and determine whether water or hydroxyl compounds might be present. It is important to note that the M3 device was used to measure reflectance signals throughout the day at various times. The temperature of the detector fluctuated due to the sun's

location and the amount of light that struck it, which is an important issue that must be carefully taken into account while analyzing the data that the M3 instrument collected.

3. Understanding the reflectance signals :

A reflectance signal acts as a quantitative measure of how much light is reflected from a specific surface or substance. A portion of incident light that strikes a surface is absorbed by the surface, while the remaining portion is reflected back into the surroundings. The amount of light that is reflected is essentially represented by the reflectance signal. With the use of this useful metric, we can learn more about the characteristics of surfaces or materials, including characteristics like color, texture, and composition.

A continuum line plays a crucial function in the field of data visualization because it successfully illustrates the continuous and unbroken relationship between two or more variables. The goal of this graphical display is to make hidden patterns and trends in the data being displayed visible. We may determine the cohesive and interrelated character of the variables under study by using a continuous line, which helps with the interpretation and understanding of the data.

Contrarily, band depth acts as a measurable indicator of the degree of absorption or reflection that a particular feature or band within a spectrum

experiences. It is expressed as a percentage and obtained by dividing the maximum reflectance value in a specific band by the difference between its minimum and maximum reflectance values. We can calculate the amount of absorption at various light wavelengths using this formula. Band depth can also be used to locate and describe particular spectral features like absorption or reflection peaks.

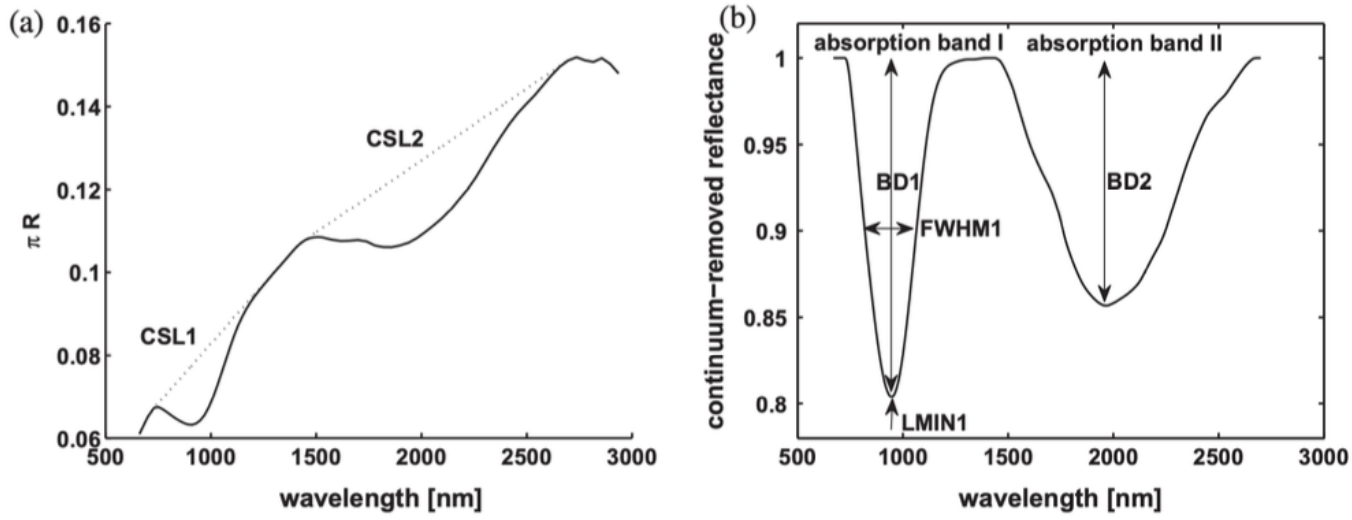


Fig.1. (a) The spectrum of reflected light from the surface on the inside of a small crater. (b) Continuum-removed spectrum. [4]

Fig.1 shows the reflectance signals and continuum removed spectra that is obtained from M3. The figure details the absorption band 1, which is around the 1 micrometer wavelength, and absorption band 2, which is around the 2 micrometer wavelength.

4. Parameters :

Creating a model that accurately illustrates the complex relationship between a dependent variable and one or more independent variables is a key goal in the field of regression analysis. The parameters, also referred to as variables or features, have the authority to decide the strength and direction of the link between the dependent variable and its independent counterparts, and they are responsible for establishing this relationship. The independent variables are the elements that are believed to have an impact on the dependent variable, whereas the dependent variable itself is the primary component being measured and is assumed to be influenced by the independent variables.

The choice of appropriate parameters during regression model development requires careful study. It is crucial to include particular independent variables in the model because doing so directly influences the model's ability to predict the dependent variable with accuracy. Incorrect parameter selections may result in an inaccurate representation of the connection between the dependent and independent variables, which will then produce inaccurate predictions.

The M3 data parameters play a crucial part in this study's overall scope. These variables represent the reflectance values at various wavelengths, which are used to identify specific minerals by their existence. BD1, FWHM1, LMIN1, CSL1, BD2, FWHM2, and CSL2, as shown in Fig.1 are the spectral

parameters used in this research. They represent how much absorption takes place, the wavelengths the absorption takes place over, the continuum slope of the absorption band, over absorption bands 1 and 2. Additionally, these variables can shed light on the structural traits of the reflectance signals at the chosen wavelengths. The temperature of the detector varies during different observations, as was previously explained. Therefore, it was essential to include a further parameter that represented the detector's temperature at the time of each observation. This research project also includes another parameter that provides information on the presence of water or hydroxyl. These parameters collectively make up the group of independent and dependent variables in my regression model.

5. Exploratory Data Analysis :

Exploratory data analysis (EDA) is a fundamental method for data analysis that makes it easier to find patterns and connections in a dataset. Its importance rests in its capacity to reveal the intricate foundations of complicated data, enabling the discovery of relationships and trends that might otherwise remain hidden in the raw data. A thorough analysis of Tukey et al.'s work (5) yields priceless insights into the EDA procedure. With the help of EDA, we are able to identify hidden structures in the data, create and test hypotheses, and unearth fascinating links. Data visualization is a key element

of EDA because it enables us to visually depict the data and identify subtle trends and associations that might otherwise escape our notice. Additionally, it helps in the identification of abnormalities and outliers that may contain significant data. Statistical methods like regression analysis can then be used to evaluate the hypotheses produced by EDA. Exploratory Data Analysis is, in essence, a crucial technique for understanding complex datasets. It not only makes it easier to look at relationships, trends, and anomalies, but it also gives us the ability to come up with hypotheses and learn things that we might not have learned from just looking at raw data.

6. Regression analysis :

A sophisticated statistical method used to evaluate and quantify the relationship between a dependent variable and one or more independent variables is known as regression analysis. Due to its widespread use in statistical modeling, it is possible to determine the strength and direction of the link between these variables, allowing for more accurate predictions of the value of the dependent variable based on the values of the independent variables. When dealing with variables that are difficult to measure or when attempting to understand complex interrelationships between numerous variables, this methodology is especially helpful. Regression analysis also helps in identifying the essential independent variables that are critical in explaining

the fluctuations in the dependent variable, providing insights into the primary elements influencing a certain phenomenon.

a. Linear regression -

The statistical method of linear regression is used to build a model that represents the relationship between a dependent variable and one or more independent variables. The assumption that there is a linear relationship between the dependent variable and the independent factors forms the basis of this hypothesis. In essence, this means that as the value of an independent variable rises or falls, the corresponding value of the dependent variable also rises or falls at the same rate. Because it offers a simple and effective way to predict the value of a dependent variable based on one or more independent variables, linear regression is important because it can make regression analysis easier provided that there is a linear relationship between the independent variables and the dependent variable. Additionally, due to its broad use and deep understanding within the statistical community, it is a preferred option for a variety of domains and applications. An important aspect of linear regression's adaptability is that it may be used to model complex and diverse interactions in situations where there are several independent variables. By embracing this addition, analysts can more sophisticatedly investigate and comprehend complicated processes within their data.

b. Decision Trees -

A well-known type of supervised learning algorithms that is useful for both classification and regression applications are decision trees. Decision trees are useful tools for making predictions about continuous target variables, specifically in the context of regression. According to the explanation given by James et al. (6), the main goal of a regression decision tree is to build an accurate model that can predict the values of the target variable by utilizing the unique features available in the dataset.

Powerful algorithms called decision trees divide data into smaller and smaller groupings based on the values of certain features. This iterative process keeps going until the data is split up into groups that have a high degree of homogeneity with regard to the desired variable or a specified maximum length is reached. Each leaf node in the tree structure denotes a predicted value for the target variable, whereas each interior node reflects a choice based on the value of a particular attribute. The intrinsic interpretability and comprehensibility of decision trees is one of the factors contributing to their widespread use in regression analysis. They can handle both continuous and categorical data, as well as many target variables, demonstrating their versatility. Decision trees also excel at capturing nonlinear relationships between the characteristics and the target variable, which expands the range of datasets to which they can be applied.

Beyond its interpretability, decision trees have further uses in regression analysis. They help to comprehend links and trends by providing insights into the

underlying structure of the data. Decision trees are well known as a trustworthy modeling approach due to their propensity for delivering precise forecasts.

Even though decision trees are capable of a lot on their own, their full potential is only reached when they are merged into ensembles. An ensemble, or group of decision trees, makes use of the combined knowledge of several different trees, greatly improving their value. Random forests and boosted decision trees are two frequently used methods to build ensembles. Random forests combine the predictions of many decision trees that have each been trained on a random portion of the data. Boosted decision trees, on the other hand, build decision trees repeatedly, with each new tree attempting to fix the mistakes produced by the preceding one. Ensembles are able to produce reliable and incredibly precise outcomes by combining the predictions of these separate trees.

i. Random Forests :

A highly effective ensemble machine learning technique, random forests are useful for both classification and regression tasks. Their primary mode of operation is the parallel training of many decision trees, each on a distinct subset of the data. The random forest's individual trees each create a class prediction, and the ensemble of trees' most commonly produced prediction serves as the model's final output.

Random forests show better performance than individual trees because they act as a committee made up of many reasonably uncorrelated trees. The fact that the trees

in the ensemble protect one another from their individual mistakes is thought to be the cause of this phenomenon. While some trees may make mistakes, many others will produce accurate forecasts, allowing the entire group of trees to move in the right direction. Multiple trees' predictions are averaged in order to reduce overfitting problems and improve the model's overall accuracy.

Because of its ability to efficiently handle enormous volumes of data, random forests are used extensively in regression analysis. The benefit of providing a measure of the relative relevance of each feature included in the dataset is another benefit of random forests. This attribute helps in the selection of features and the comprehension of the underlying relationships in the data by identifying the influence and significance of various features on the outcome.

ii. Boosted decision trees :

Due to its outstanding qualities, boosted decision trees, an ensemble model used in regression analysis, have become increasingly popular. These models' technique of "boosting" entails instructing a series of weak learners. Weak learners are trees whose predictions are only slightly better than random guessing. Multiple trees applied sequentially, aggregate result in the ensemble being capable of producing accurate predictions. Each succeeding learner is made to fix the errors committed by the one before it, creating a cumulative learning effect. These learners relate to individual trees specifically in the domain of regression trees, and the boosting procedure maintains their dependency. By adjusting the residual mistakes left by the

previous trees, the algorithm effectively learns. Therefore, using boosting in a decision tree ensemble leads to a reduction in the errors. Due to their amazing ability to detect complex and nonlinear correlations within the dataset, boosted decision trees are an effective tool for regression analysis.

c. Support Vector Regression

Multivariate support vector regression (MSVR) is a powerful machine learning technique used for solving regression problems involving multiple input variables. It is an extension of the traditional support vector regression (SVR) method, which is designed for univariate regression tasks. In MSVR, the goal is to find a hyperplane that best fits the data points in a high-dimensional space. This hyperplane is determined by a subset of support vectors, which are the data points that lie closest to the hyperplane. The distance between the hyperplane and the support vectors is maximized, while also minimizing the prediction error on the training data. The main idea behind MSVR is to transform the input variables into a higher-dimensional feature space using a kernel function. This transformation allows MSVR to capture complex relationships between the input variables and the target variable. The choice of kernel function depends on the specific problem and can include linear, polynomial, or radial basis functions (rbf). In MSVR, there are two key parameters that need to be tuned: C and epsilon. The parameter C controls the trade-off between achieving a small prediction error and maximizing the margin between the hyperplane and the support vectors. A larger value of C allows for a smaller margin but reduces

the training error. On the other hand, a smaller value of C encourages a larger margin but may increase the training error. The parameter epsilon represents the width of the epsilon-tube around the predicted value where no penalty is applied to errors. Data points within this tube are considered to have accurate predictions. By adjusting epsilon, the model's sensitivity to errors can be controlled. During the training phase, MSVR optimizes a cost function that balances the margin maximization and the error minimization, taking into account the values of C and epsilon. This optimization is typically achieved using techniques such as quadratic programming. The resulting MSVR model can then be used to predict the target variable for new input data by mapping the inputs to the higher-dimensional feature space using the chosen kernel function and determining the position of the data point with respect to the hyperplane.

d. Residuals :

When performing a regression analysis, residuals—differences between the observed value of the dependent variable (y) and the anticipated value of y produced from the regression model ($\bar{y}$)—are examined while taking the provided independent variables into account. The differences between the dependent variable's actual value and the value predicted by the regression model are essentially represented by residuals. The importance of residuals comes from their capacity to measure how closely the regression model fits the data. Examining the residuals makes it possible to spot any trends that can point to the model's inadequacy in capturing the underlying patterns in the data. Thus, by carefully examining the residuals, we can learn a great deal about

the efficacy and precision of the regression model in capturing the real relationship between the variables under study.

7. Cross Validation

Cross-validation is a widely used technique in machine learning and statistical modeling to assess the performance and generalizability of a predictive model. It is particularly useful when the dataset is limited or when there is a need to estimate how well the model will perform on unseen data. The basic idea behind cross-validation is to divide the available data into multiple subsets or folds. One of the folds is kept aside as the validation set, while the model is trained on the remaining folds. The validation set is then used to evaluate the model's performance. This process is repeated several times, each time with a different fold serving as the validation set, and the performance results are averaged. The most commonly used cross-validation technique is k-fold cross-validation. In k-fold cross-validation, the data is divided into k equal-sized folds. The model is trained on k-1 folds and evaluated on the remaining fold. This process is repeated k times, with each fold serving as the validation set once. The performance results from each fold are then averaged to obtain an overall assessment of the model's performance. Cross-validation helps in estimating how well a model will generalize to unseen data by providing an unbiased evaluation of its performance. It helps to mitigate the risk of overfitting, where a model performs well on the training data but fails to generalize to new data. Cross-validation is also useful

for hyperparameter tuning, where different combinations of model parameters are evaluated to find the optimal configuration. By performing cross-validation on each parameter combination, it is possible to select the best set of hyperparameters that yield the highest performance.

3. METHODS

In order to analyze the data, I use the Python programming language as well as softwares such as MATLAB. Since the nature of my thesis is focused on leveraging Machine Learning, I heavily rely on the statistical models of regression which I implement by using the Python scikit-learn library as well as the MATLAB statistics and machine learning toolbox. A very important aspect of my thesis is to make various statistical plots such as scatter plots of each individual parameter, histograms of the parameters, plots of the regression residuals, etc. for which I use Python's matplotlib library as well as MATLAB inbuilt functions.

My research experience started off with analyzing data of the Bullialdus crater, which is in the Mare region of the moon, that was given to me as 13 geotiff images where each image was 1800 by 1800 pixels over 84 channels and each pixel had one reflectance value attached to it. The detector temperature when each image was taken by M3 was recorded and can be seen below. Every pixel (or datapoint) was then

associated with a temperature variable that represented the detector temperature during the observation.

Bullialdus

Image	*Detector Temperature (* celsius)*
1	160.63
2	150.35
3	150.29
4	150.35
5	150.16
6	146.3
7	146.61
8	146.42
9	159.23
10	159.54
11	164.12
12	164.04
13	164.45

Since one of the important factors that was to be investigated in my research was the effect of temperature on the variation in the 3 micron water or hydroxyl band, the data of temperatures was crucial.

I then analyzed the data of Clavius, which is a crater in the Highland region of the moon, and was given to me as 8 geotiff images of 1800 by 2100 pixels over 84 channels and each pixel had one reflectance value attached to it. For Clavius the detector temperature at the time of observation was not recorded like in Bullialdus, but since detector temperature is a proxy variable for the time of day of observation, I directly used the time of day. Similar to Bullaldus, every pixel was associated with a time of day variable that represented the time of day on the moon at the time of the observation.

Clavius

Image	***Time of Day***
1	8.44
2	8.42
3	8.47
4	16.03
5	16.05
6	14.26
7	12.38
8	12.53

All of the reflection data from Bullaldus and Clavius had to be made sense of, and hence the parameters mentioned in the literature review had to be calculated for

each data point. This resulted in each pixel having a value for each parameter associated with it.

The next step was to understand the data in a more detailed way, for which Exploratory Data Analysis was needed. The first part in EDA was plotting normalized histograms of each parameter, from which I could gain important insights about the data. Normalization is the process of scaling the data such that the mean of the data is 0 and the standard deviation is 1.

Bullialdus

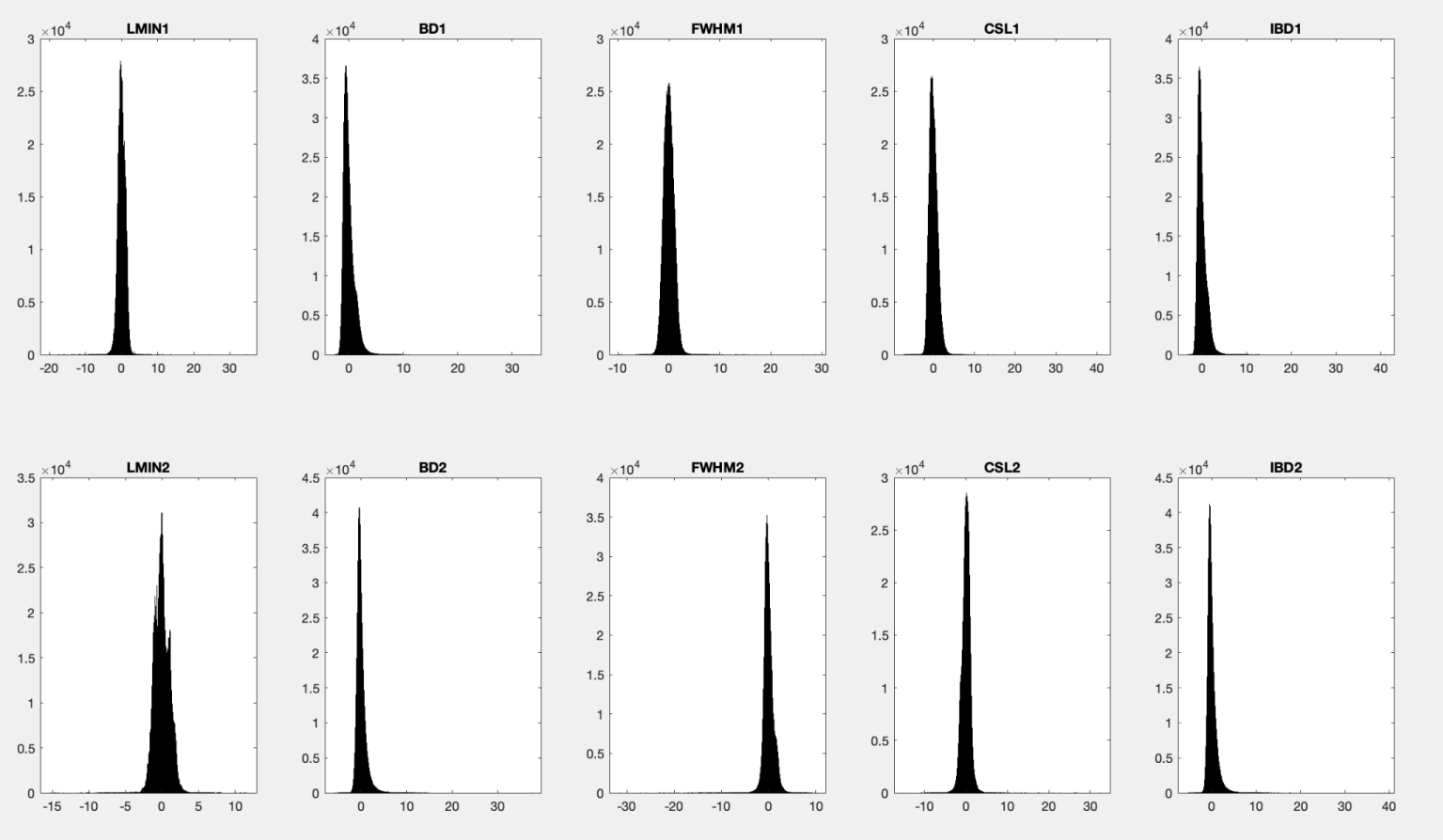


Fig. 2. Normalized histograms of all parameters for Bullialdus data

Clavius

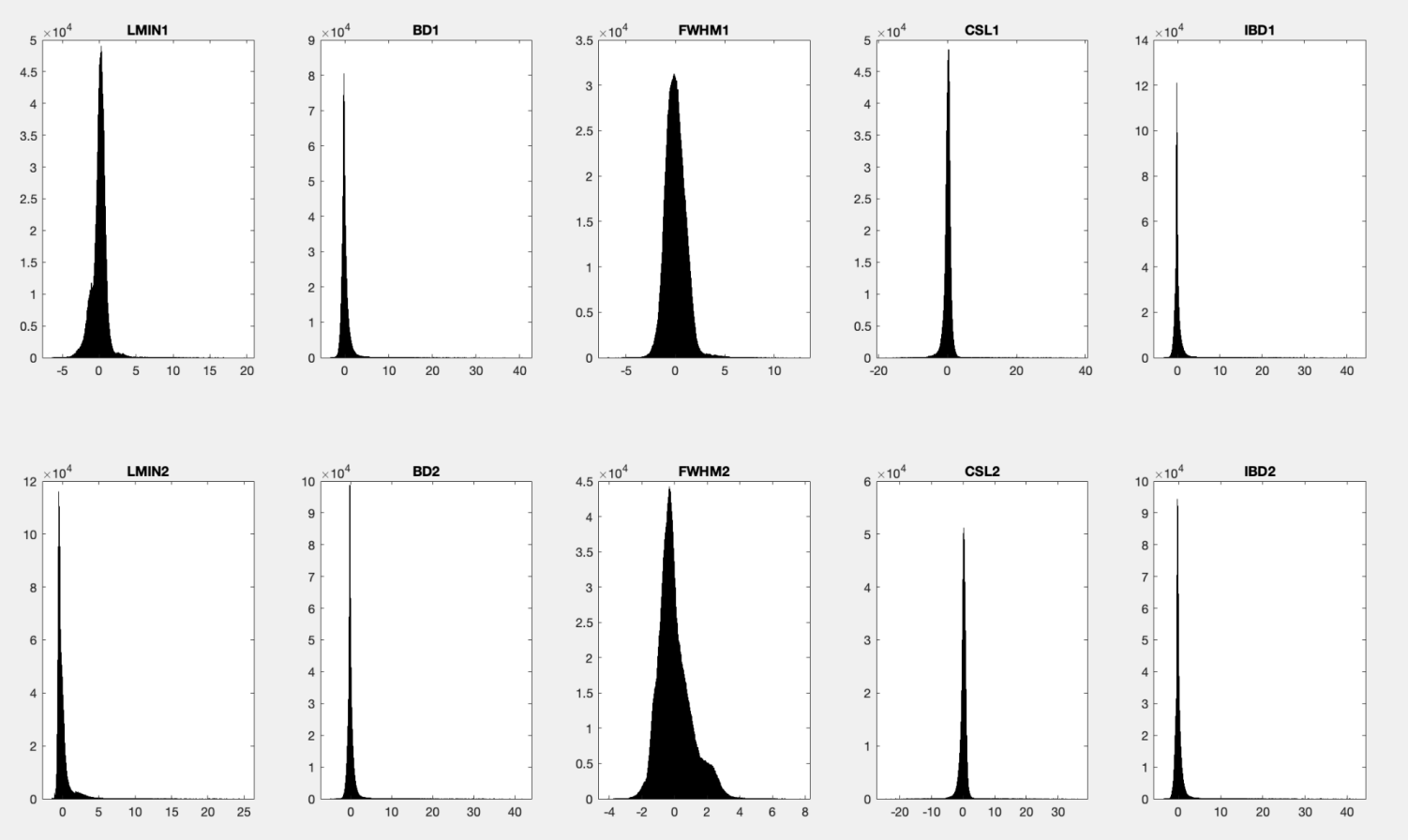


Fig. 3. Normalized histograms of all parameters for Clavius data

The most important insight was to discover that the data had a significant amount of outliers which had to be dealt with which can be inferred from the long tails on either side of the plots. Upon further investigation it was found that a large portion of these outliers were due to the presence of noisy data which had to be removed, but some of the data points which we had initially thought of as outliers were actually indicative of the presence of fresh craters as a result of the constant bombardment of the lunar surface by small meteors and these points should not be removed.

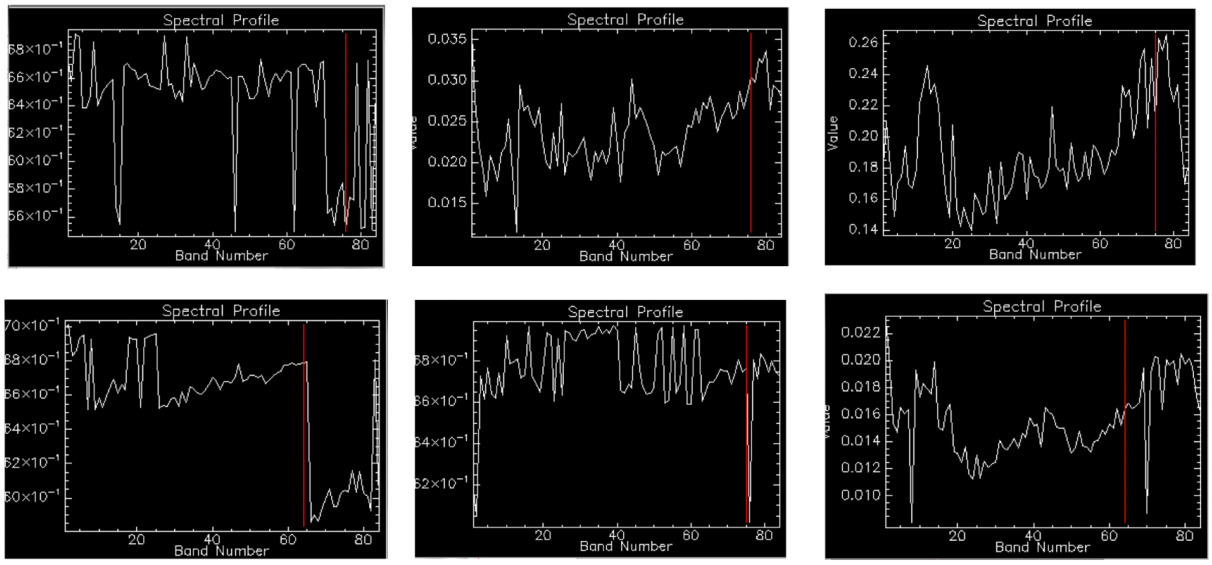


Fig. 4. Spectrums of most outliers - noisy data

After removing the outliers, it was time to understand exactly how important a role temperature plays, and hence overlaid histograms for each parameter using pixels grouped by temperature were plotted. The plots indicated that there is in fact a reasonable amount of variation with temperature and this fact needs to be accounted for in the regression. It was incorporated by assigning each data point a temperature parameter that represents the detector temperature at the time the data was captured. Since only the Bullialdus data had values for the detector temperature, Clavius data was not used for this purpose.

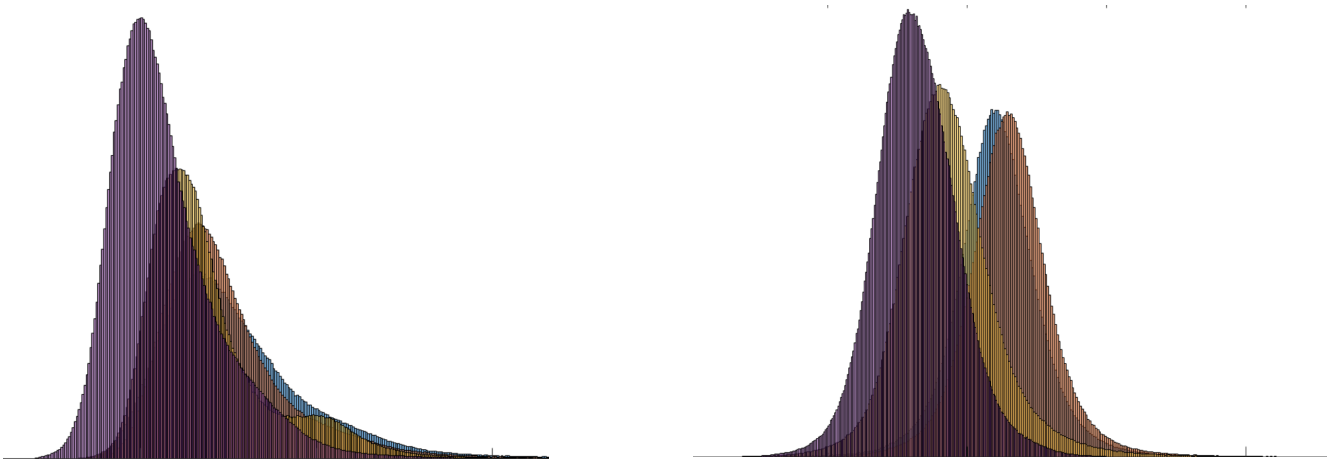


Fig. 5. Histogram of data points grouped by temperature (a) BD1, (b) ohibd

The final part of EDA involved measuring how correlated each of the independent variables were with one another. To calculate the correlation, Spearman's rank correlation coefficient was used, which can measure the statistical dependence of variables whether this dependence is linear or nonlinear. The correlation values showed that various parameters are highly correlated with each other, and thus some of the parameters can be eliminated since we do not want redundant data which will only increase the processing time as well as produce inferior results. The parameters eliminated were: LMIN2, IBD1, and IBD2. Heatmaps, that provide a visual representation of the correlation matrix, were then plotted to show the correlation between parameters that were not eliminated. When some parameters are eliminated, the resulting heatmap focuses solely on the relationships between the non-eliminated parameters. This can help uncover hidden relationships, identify strong positive or

negative correlations, and provide a clearer understanding of how the remaining variables interact with each other.

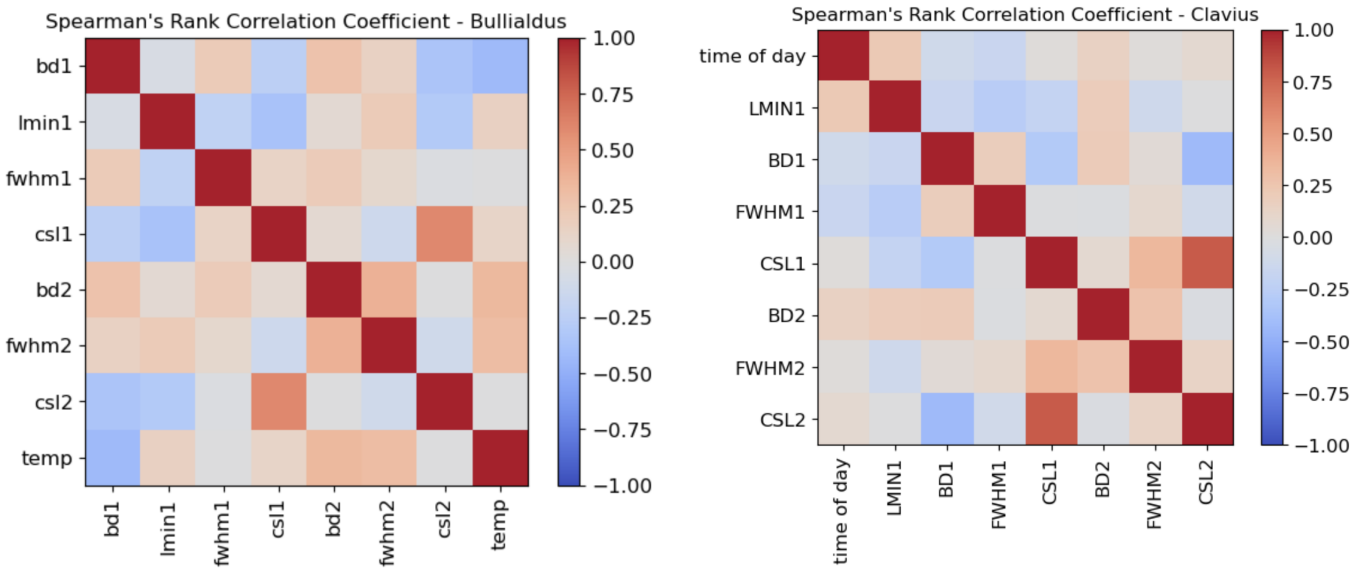


Fig. 6. (a) Spearman's rank correlation of non-eliminated parameters for Clavius

b) Spearman's rank correlation of non-eliminated parameters for Bullialdus

Once the process of exploratory data analysis was completed, regression analysis was performed on the data. To perform regression, the data was divided into two groups: 80% of the data was used for training, and 20% for testing. Cross validation was used while performing hyperparameter tuning for the models. The training data consists of a set of features and the corresponding values of the dependent variable, and is used to enable the model to learn relationships and patterns between the input features and dependent variables. During the training process, the model analyzes the training data and adjusts its internal parameters or weights based

on the patterns it discovers. The goal is to minimize the difference between predicted output of the model and the true output provided in the training data.

Validation data is a separate dataset that is used to assess the performance of a trained model during the training process. It is different from the training data and serves as a metric to understand how well the model generalizes to unseen data. The model is given data of the independent variables but not the corresponding values of the dependent variable. The validation data is used for hyperparameter tuning, model selection, and monitoring the model's performance. It helps in detecting overfitting or underfitting of the model and gives insights about the generalization capability of the model. The goal is to find the set of hyperparameters that performs well on the validation data, since it indicates how the model might perform on new, unseen data.

Testing data is yet another distinct dataset that is used to evaluate the final performance and generalization ability of the trained model. It serves as an independent benchmark to assess how well the model performs on unseen data. The testing data is not shown to the model until it is fully trained and ready for evaluation. The main purpose of the testing data is to understand if the model has learned the underlying patterns and relationships in the data, and can make accurate predictions.

The independent variables used were: LMIN1, BD1, FWHM1, CSL1, BD2, FWHM2, CSL2, and Temperature (for Bullialdus)/Time of Day (for Clavius) and the

dependent variable was OHIBD which is a parameter that is indicative of the presence of water.

To understand whether the independent variables had a linear relationship with the dependent variable, I initially started the regression analysis by using multivariate linear regression, which is the simplest regression model. This model assumes that the independent variables and the dependent variable have a linear relationship, and is effectively a line of best fit. It is calculated as follows:

$$y_i = b_0 + \sum_{j=1}^{p} b_j x_{ij} + e_i$$

for i $\in$ {1, . . . , n} where:

- $y_i \in R$ is the real-valued response for the i-th observation
- $b_0 \in R$ is the regression intercept
- $b_j \in R$ is the j-th predictor's regression slope
- $x_{ij} \in R$ is the j-th predictor for the i-th observation
- e_i is the Gaussian error term

However, I did not have much success doing so. When looking at the scatter plots of the residuals, it is clear that the residuals have a clear pattern of skewness formed by the points. This was because of the fact that the data probably does not have a strong linear relationship between the independent and dependent variables.

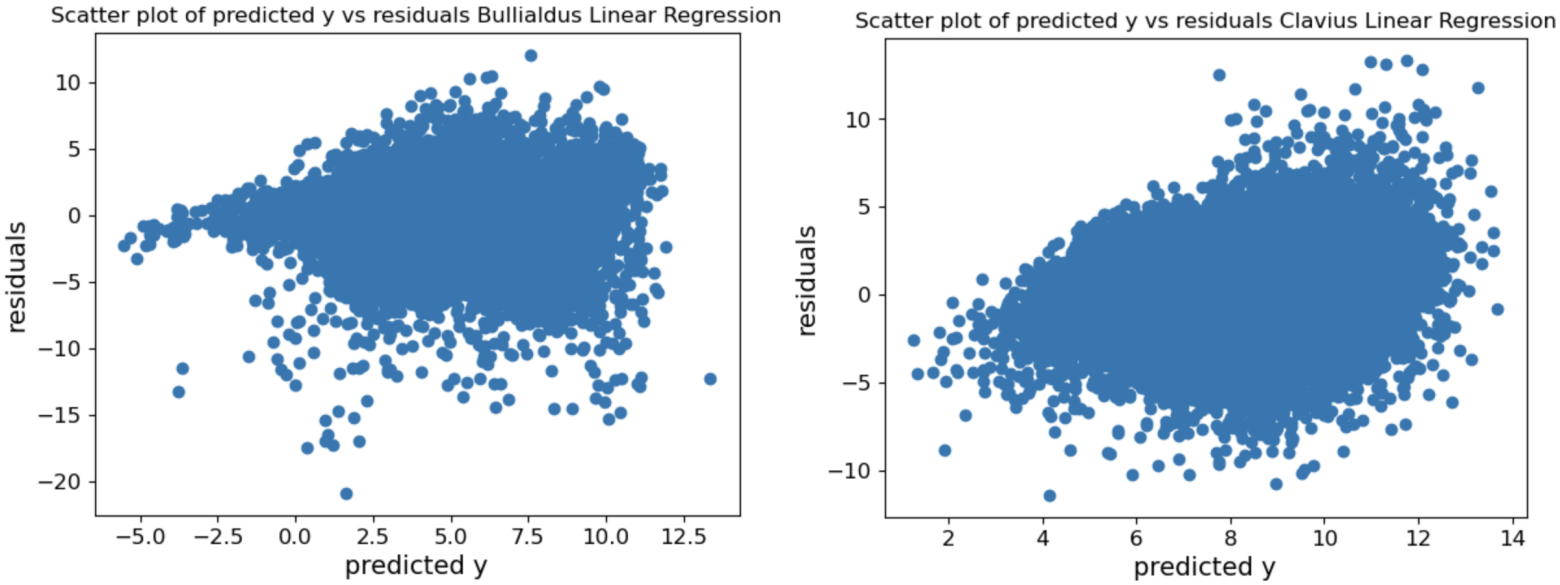


Fig. 7. Residual plots of linear regression model for (a) Bullialdus, (b) Clavius

Eventually, I moved on to using boosted regression trees to develop further insights on the effect that each parameter has on the dependent variable. This method was chosen because of the fact that the data is probably non linear and regression trees are an excellent way to understand how much of a role each parameter plays in the variation in amount of water found. I used a specific algorithm for boosted regression trees known as Gradient Boosting. The term "gradient" refers to the optimization process used to minimize the loss function. The algorithm calculates the gradient of the loss function with respect to the predictions made by the model and tries to find the optimal direction to update the model's parameters in order to minimize the loss. This process is performed using gradient descent optimization to minimize the Mean

Squared Error, which is a commonly used metric to measure the average squared difference between the predicted values and the actual values.

$$MSE = (1/n) * \Sigma(y_i - \hat{y}_i)^2$$

Since gradient boosted trees are a complicated model, it is extremely important to correctly select the model's hyperparameters so as to minimize underfitting, which occurs when a model fails to capture the relationships in the data, or overfitting, which occurs when the model captures the relationships in the training data extremely well, but fails to generalize on unseen data. To build an optimal model, the number of trees in the model, the depth of each tree, and the learning rate of the gradient descent must be tuned to ensure there is neither underfitting nor overfitting affecting the model's performance.

A Grid Search was used to find the optimal values for each parameter. The algorithm works by taking names of the hyperparameter to tune and multiple values of each hyperparameter. An exhaustive search over all possible combinations of hyperparameter values is performed, and for each combination, 5-fold cross validation is conducted to evaluate the model's performance.

Data	Model	Number of Trees	Maximum Depth	Learning Rate
Bullialdus	Gradient Boosting	25	4	0.3
Clavius	Gradient Boosting	30	3	0.15

Final values of hyperparameters as found by Grid Search method for Gradient Boosting models

The gradient boosted regression tree model gave very important insights into the data. The optimal gradient boosted model was used to plot the importance of each feature in the model's prediction, and to plot the residuals of the model to be used in the analysis. During the construction of each tree in the Gradient Boosted Regression ensemble, features are evaluated to determine the best split points. The reduction in mean squared error achieved by a particular feature when making a split is recorded. Features that lead to a larger decrease in the mean squared error are considered more important. Feature importance of the model is calculated by averaging the reduction in mean squared error by each feature across all trees in the ensemble, providing a relative ranking of feature importance, indicating which features had the most significant impact on the predictive performance of the model. Feature importances provide insights into the relative influence of different features on the target variable.

To verify the results of the gradient boosted model, a Random Forest regression model and a Support Vector Machine regression model was created. The best parameters for the random forest model and support vector machine model were also found using the grid search method, so as to minimize the effects of underfitting and overfitting on the model. The rationale behind creating different models to verify the results was that each of these models has different approaches for the regression process and doing so can provide reliability in the results achieved:

1) By comparing the results of different models, the consistency and robustness of the model's findings can be maintained. If multiple models provide similar results, the confidence in the accuracy and reliability of the analysis is increased.
2) Different models have different assumptions and characteristics. By comparing the performance of multiple models, potential biases and overfitting issues that may arise from solely relying on one model can be identified. If the results of all models are similar, the observed patterns are more likely to be genuine rather than artifacts specific to a particular model.
3) Different models may assign different levels of importance to various features, and by examining the consistency in feature importance rankings across models, a deeper understanding of the variables that have the most significant impact on the dependent variable can be gained.

Data	Model	Number of Trees	Maximum Depth
Bullialdus	Random Forest	50	9
Clavius	Random Forest	45	8

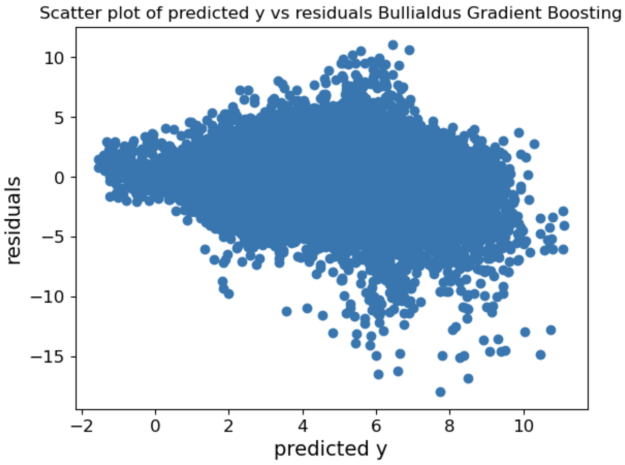
Data	Model	C value	Epsilon value	Kernel
Bullialdus	Support Vector Regression	3	0.15	Radial basis function
Clavius	Support Vector Regression	5	0.15	Radial basis function

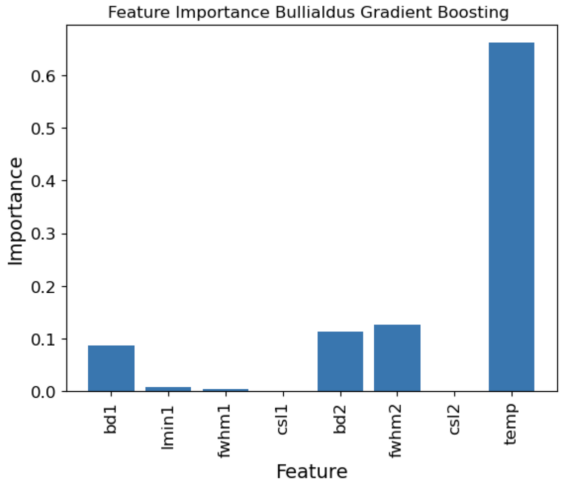
Final values of hyperparameters as found by Grid Search method

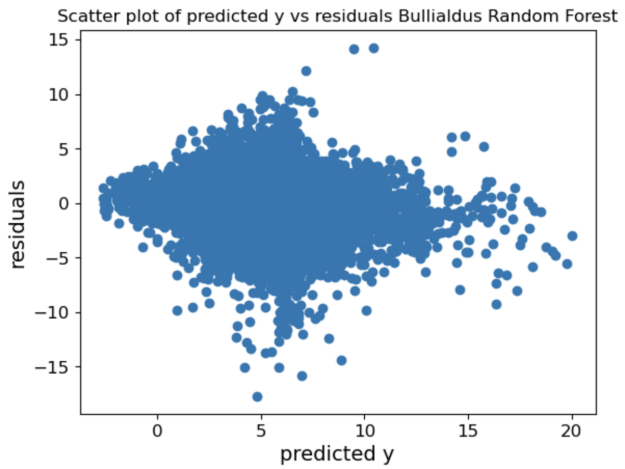
4. Results and discussion :

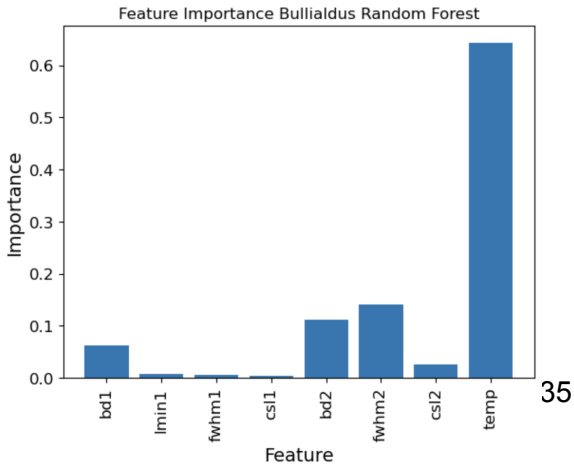
The following section presents a comprehensive analysis and interpretation of the results obtained from the research conducted in this study. This empirical investigation aimed to address the research questions and hypotheses formulated in the earlier sections, providing valuable insights into the phenomenon under investigation. By analyzing the gathered data and employing appropriate statistical techniques, this section presents a detailed account of the key findings, trends, and patterns observed throughout the study.

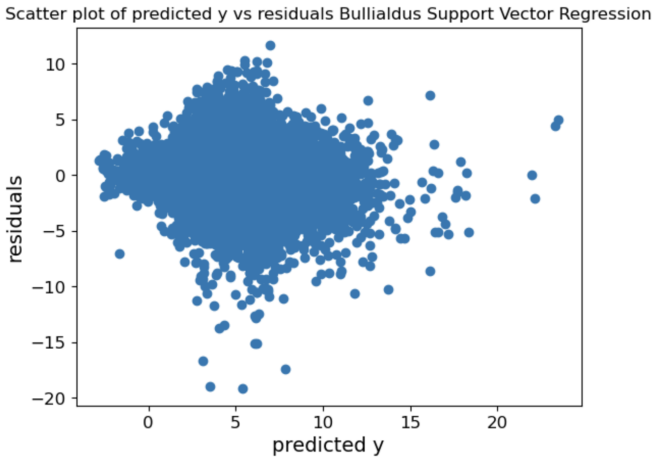
Regression results for Bullialdus data :

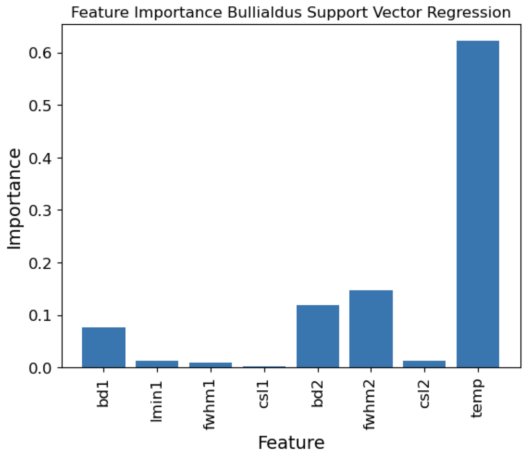








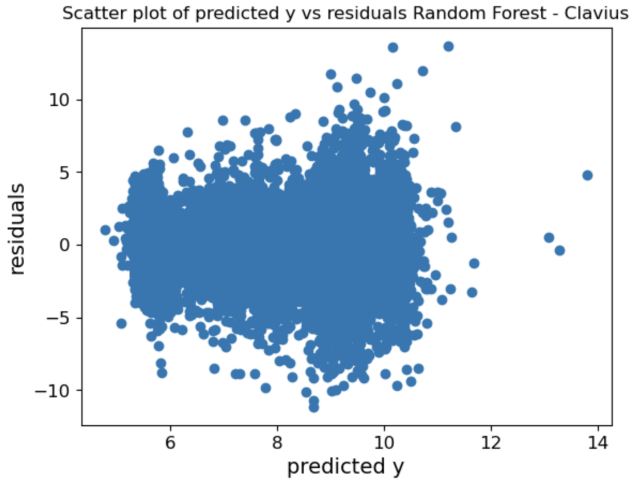


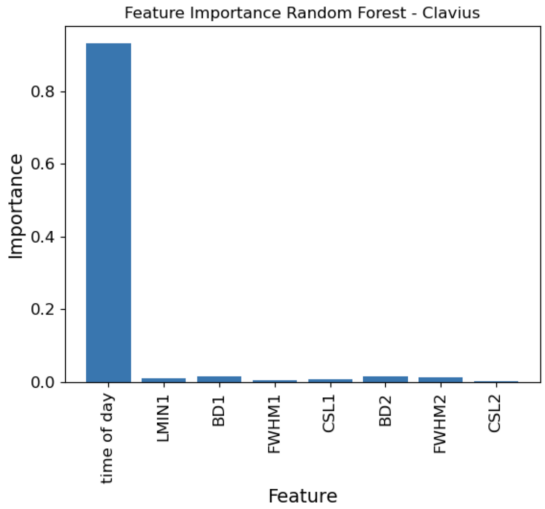


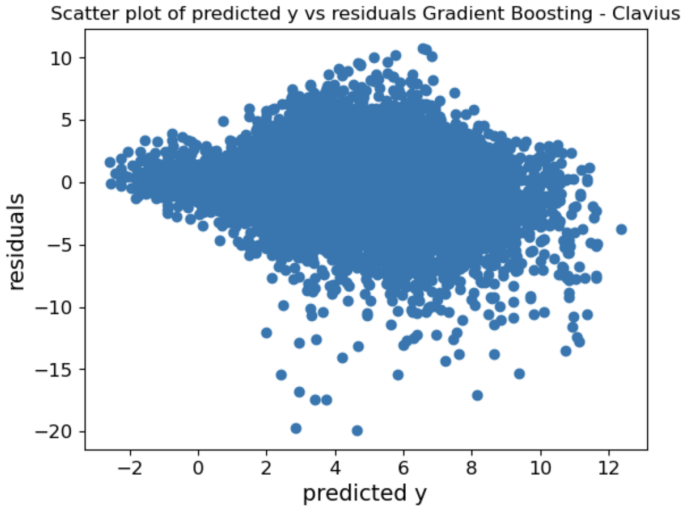
Model	**Mean Squared Error - Training data**	**Mean Squared Error - Testing data**
Gradient Boosting	1.65	1.64
Random Forest	1.53	1.57
Support Vector Regression	1.58	1.61

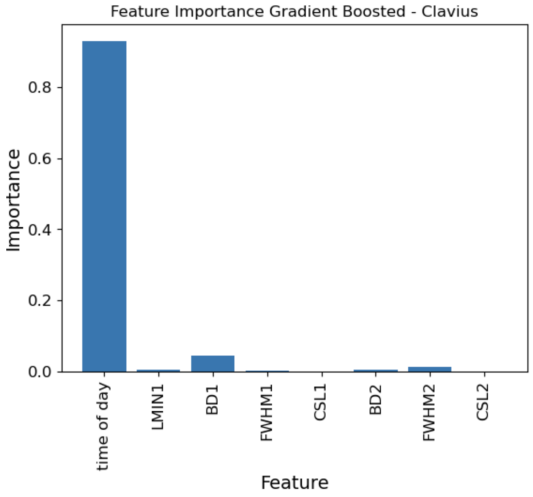
It can be inferred from the feature importance plots that all models found Temperature followed by FWHM2, BD2, and BD1 to be the most important features in the model's prediction of the dependent variable. Temperature has a much higher feature importance suggesting that changes in temperature had the most significant influence on the dependent variable. From the mean squared errors on the training and testing data of all the models, one can conclude that the models are well tuned, and the effects of overfitting and underfitting are negligible. The fact that different models each using a different approach for prediction, all telling the same story provides further confidence in the results.

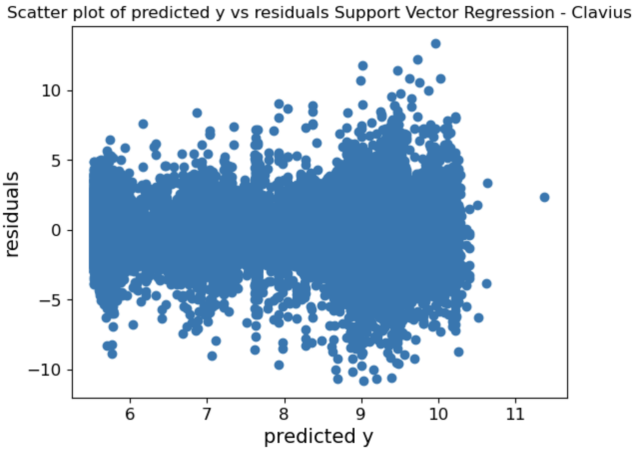
Regression results for Clavius data:

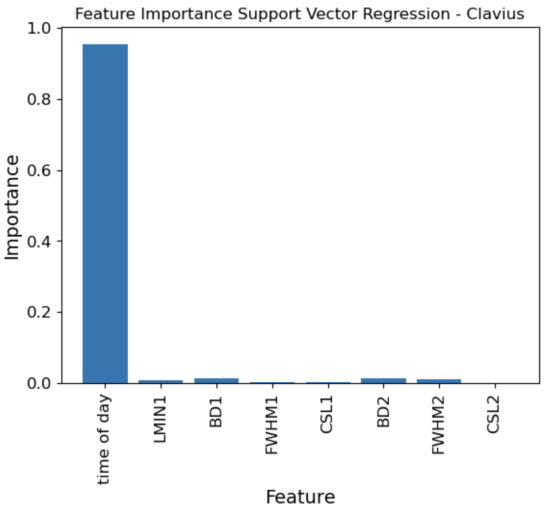












Model	Mean Squared Error - Training data	Mean Squared Error - Testing data
Gradient Boosting	1.80	1.81
Random Forest	1.69	1.72
Support Vector Regression	1.74	1.76

From the feature importance plots, one can tell that all models found Time of Day to be the most important feature in the model's prediction of the dependent variable. Time of Day has the majority of the feature importance suggesting that changes in Time of Day had the most significant influence on the dependent variable. Time of Day has a higher importance on the Clavius data as compared to Temperature on Bullialdus data because the Highlands have less variation in composition as compared to the Mare, so the other features are not changing as much. From the mean squared errors on the training and testing data of all the models, one can conclude that the models are well tuned, and the effects of overfitting and underfitting are minimized. The fact that once again different models each using a different approach for prediction, all giving similar results proves that the results of the regression are statistically valid.

The range of the dependent variable, ohibd, in the data was between 0 and 30. This means that the model could predict values of the dependent variable between these values. The mean squared testing error of the worst performing model in this

research, was 1.81, indicating that on average the model was 1.81 units away from perfectly predicting the value of the dependent variable. When looking at this mean squared error in comparison with the range of values that the dependent variable can take, it can be understood that this error is not very big, indicating that the models are performing well. When dividing the highest mean squared testing error by the range of the dependent variable (1.81/30), the mean squared testing errors can be quantified against the range, showing that the worst performing model only has an error of about 6%.

The reason that Temperature/Time of Day dominates the feature importance can be attributed to a variety of factors. First and foremost is that the values of other features vary in only a small portion of the data, and are stagnant throughout the majority of the data. On the other hand Temperature/Time of Day varies in a larger portion of the data and thus, the models have attributed this to Temperature/Time of Day having a more significant impact on the dependent variable. In addition to this, it is also possible that the moon mineralogical mapper's data collection or thermal calibration process is not perfectly done and can be affected by changes in the detector temperature. If this is the case, it would explain why Temperature/Time of Day has the highest feature importance since the dependent variable would vary more based on Temperature/Time of Day as compared to other features.

When scientifically understanding these results, it is clear that they do not suggest that the amount of water in a particular part of the lunar surface changes with

the time of day (and subsequently the detector temperature), this is not physically possible, but instead that this is all that can be analyzed from this data. For a more in-depth analysis to be carried out and to find with high confidence a set of features, that are representative of various aspects of the absorption bands, whose variations can accurately account for the variation in the dependent variable, better data is necessary.

5. Summary and conclusion :

The complexity of how detector temperature affects the regression problem in question has been clarified by this study, and it has also given the possible reasons behind the phenomenon observed:

1) The values of other features vary in only a small portion of the data, and are stagnant throughout the majority of the data, whereas, Temperature/Time of Day varies in a larger portion of the data and thus, the models have attributed this to Temperature/Time of Day having a more significant impact on the dependent variable.
2) The moon mineralogical mapper's data collection or thermal calibration process might not be perfectly done and can be affected by changes in the detector temperature. If this is the case, it would explain why Temperature/Time of Day has the highest feature importance since the

dependent variable would vary more based on Temperature/Time of Day as compared to other features.

During my research I have improved my grasp of the methodology used and its anticipated results by rigorous research approaches and in-depth data analysis.

The results of this investigation have significant implications for the future. They add to the body of knowledge by demonstrating how important a role detector temperature plays in the regression problem. These findings open up fresh perspectives and directions for investigation, as well as opportunities for possible applications in different study areas.

It is critical to recognize this study's constraints, most important of which is the data obtained. The reason for data being the study's main constraint is detailed above. Due to these restrictions, the field can continue to develop and improve by addressing and building upon the current findings in subsequent studies once better data has been obtained.

In conclusion, this work is an advancement in our efforts to understand the questions surrounding the parameters which affect the presence of water on the moon.

The results obtained from this study opens up fascinating new directions for further study and investigation in the area. As I draw to a close, it is clear that there are a number of intriguing topics worth looking into further.

References :

1. “Chemical composition of lunar meteorites and the lunar crust” [Online]. Available: https://www.researchgate.net/publication/226104384_Chemical_Composition_of_Lunar_Meteorites_and_the_Lunar_Crust. [Accessed: 16-Dec-2022].
2. “Petrology and chemical composition of lunar mare Diabase ... - USRA.” [Online]. Available: https://www.hou.usra.edu/meetings/lpsc2020/pdf/2613.pdf. [Accessed: 16-Dec-2022].
3. J. F. Mustard, C. M. Pieters, P. J. Isaacson, J. W. Head, S. Besse, R. N. Clark, R. L. Klima, N. E. Petro, M. I. Staid, J. M. Sunshine, C. J. Runyon, and S. Tompkins, “Compositional diversity and geologic insights of the Aristarchus Crater from Moon Mineralogy Mapper Data,” *NASA/ADS*, 10-May-2011. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2011JGRE..116.0G12M/abstract.
4. C. Wöhler, A. Grumpe, A. Berezhnoy, M. U. Bhatt, and U. Mall, “Integrated Topographic, photometric and spectral analysis of the lunar surface: Application to impact melt flows and ponds,” *Icarus*, 20-Mar-2014. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0019103514001328.
5. J. W. Tukey, *Exploratory Data Analysis*. Hoboken, New Jersey: Pearson, 2020.
6. G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning: With applications in R*.