Computational Approach for Annotation of IR Spectral Features

Patrick Ayres, Gattlin Walker, and Eathan Hickey
Missouri State University
Faculty advisors: Dr. Razib Iqbal, Dr. Keiichi Yoshimatsu

Abstract

Infrared is the section of electromagnetic radiation, or EMR, with wavelengths longer than visible light, ranging from 0.7 micrometers (μm) to 1000 μm. If the wavelengths within the near-IR and mid-IR range (0.7 μm - 25 μm) are shot at an object and then plotted against the reflectance at those wavelengths, it may be possible to extract information about the chemical composition of the object or substance in question. The purpose of this study is to explore computational approaches to annotate infrared spectrum to find similarities among spectra and classify those spectra based on the spectral features. We utilize the NASA ECOSTRESS dataset in our exploration. This multifaceted research includes identifying the effectiveness of a novel preprocessing algorithm called Normalized Local Change (NLC), in conjunction with traditional spectral similarity algorithms, such as Euclidean Distance and Pearson Correlation Coefficient, to increase the accuracy of spectral similarity measures. Initial investigations show that using NLC increases the accuracy when comparing spectra by an average of over 10% for the ECOSTRESS dataset. In addition, spectral features that are commonly observed and that are only observed in specific types of samples were sought by comparing several selected sets of IR spectra. Our research endeavor will also include experimental machine learning approaches, such as various clustering algorithms and convolutional neural networks.

Keywords: infrared spectrum, spectral similarity, clustering, convolutional neural network, dimensionality reduction.

Objective

Infrared spectroscopy is a powerful tool in research for its capability to remotely probe the information of the Earth as well as astronomical bodies. By taking an infrared scan of a new planet, they can then compare the unknown spectrum to known spectra and see if the planet has any known compounds. Towards this, our ongoing research project has the following objectives: A) Analyze the effectiveness of our Normalized Local Change (NLC) algorithm[1] on the ECOSTRESS dataset by comparing the accuracy of spectral similarity algorithms with and without the NLC preprocessing, B) Analyzing the effectiveness of clustering algorithms to classify unknown spectra, C) Analyzing the effectiveness of convolutional neural networks to find similar spectra to a given spectrum, and D) Identification of IR spectral features that are commonly observed and that are only observed in specific types of samples in ECOSTRESS database.

Section A: Analyzing the effectiveness of NLC

Background

NLC is a preprocessing algorithm that is used to translate values before they are used as inputs for the equations listed below [1]. It works by looking at the data to the left and right of a given data point and using that to nudge the current value to limit the amplitude of peaks and valleys. The behavior in which NLC uses to do this is governed by two parameters – the range and the floor multiplier. The range dictates how the width of the adjacent data points we look at, and the floor multiplier is used to calculate the floor value by multiplying the mean of the spectrum by the floor multiplier. If there is a data point that has a lower value, it is automatically brought up to this floor. The equation for NLC is shown below:

$$NLC_k = \frac{R}{L+R} \; where \; L = \sum_{i=k-r}^{k-1} A_i \; and \; R = \sum_{i=k+1}^{k+r} A_i$$

There are a few other existing methods to measure spectral similarity. Table-1 below shows the existing methods that were used in our study to test the effectiveness of NLC.

Table 1. Commonly used approaches for spectra comparison

Manhattan Distance (MAD)	$\sum_{1}^{n} \lvert x_i - y_i \rvert$
Euclidean Distance (MSD)	$\sum_{1}^{n} \sqrt{(x_i - y_i)^2}$
Cosine Similarity (DPN)	$\frac{A \cdot B}{\lVert A \rVert \lVert B \rVert}$
Pearson Correlation Coefficient (COR)	$\frac{\sum_{1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{1}^{n}\sqrt{(x_i - \bar{x})^2}\ \sum_{1}^{n}\sqrt{(y_i - \bar{y})^2}}$

Approach

The first step in being able to test the effectiveness of NLC was to organize the ECOSTRESS dataset. This was done by separating each data file according to different aspects of the spectrum it contains. There are two types of files in the dataset: materials and plants. The file itself lists which one of these two options it is. However, the main path for categorizing an item in the dataset stays the same and is shown in Figure 1.

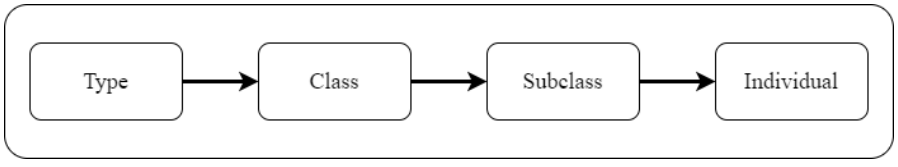

Figure 1 – Level Structure for a data file in the ECOSTRESS dataset

Now that the dataset is organized, the next step is to generate a "hitlist", which is a way to take a data file (which has the data for the spectrum of a single file) and compare it to all the other files in the dataset, and then sort them by their similarity using one of the equations above to figure out which spectra are similar. The next consideration to make is how to measure the accuracy of a given algorithm/configuration. We observed that items that are in more of the same classifications as referenced in Figure-1 are more likely to have similar spectra. With that in mind, we created the idea of measuring the accuracy at different "levels", which correspond to how closely the two compounds are classified. For example, if the top result of the hitlist was of the same "type", then it was a success under level 1, as that is the broadest thing to get correct. Extrapolating this same logic, class corresponds to level 2, subclass corresponds to level 3, and the unique individual name is tied to level 4.

As explained in the background, NLC has two parameters – range and floor multiplier. By changing these two parameters and generating hitlists for all the files, we can generate a heatmap of what works best for maximizing accuracy of NLC on the ECOSTRESS dataset. These heatmaps can be found in Figures 2-5. Once we have these values, we compared them to the accuracies obtained without NLC and compare them to see if NLC is effective for increasing accuracy in comparing spectral similarity.

Current Progress/Results

The following figures detail the accuracy of NLC using different values for the range and floor multipliers and at different levels of classification in the form of heatmaps.

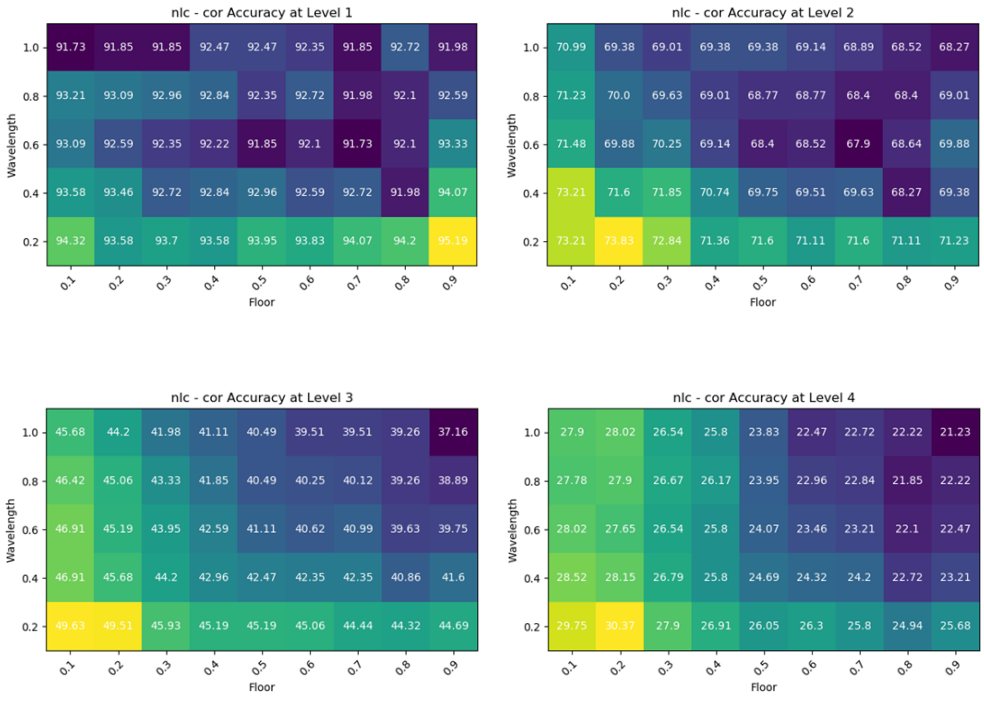

Figure 2 - Accuracy using Pearson Correlation Coefficient

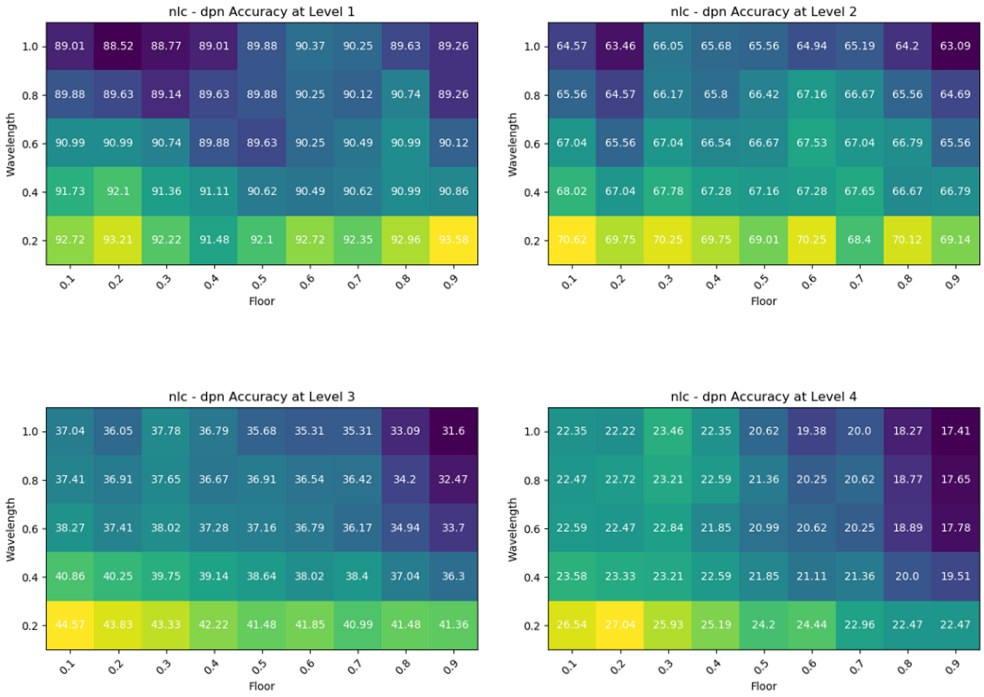

Figure 3 - Accuracy using Cosine Similarity

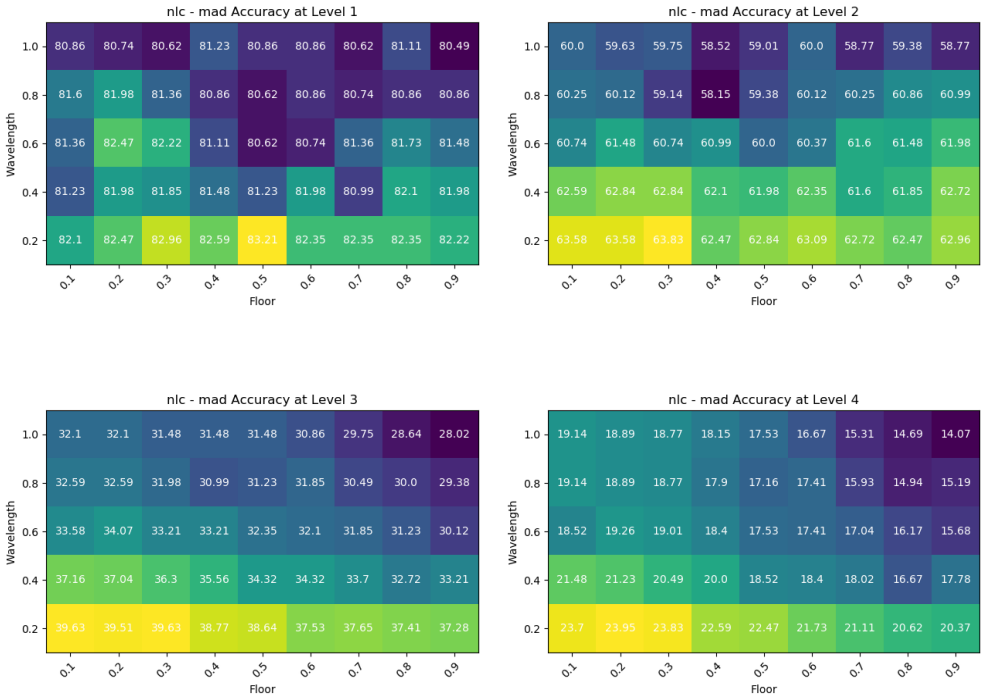

Figure 4 - Accuracy using Manhattan Distance

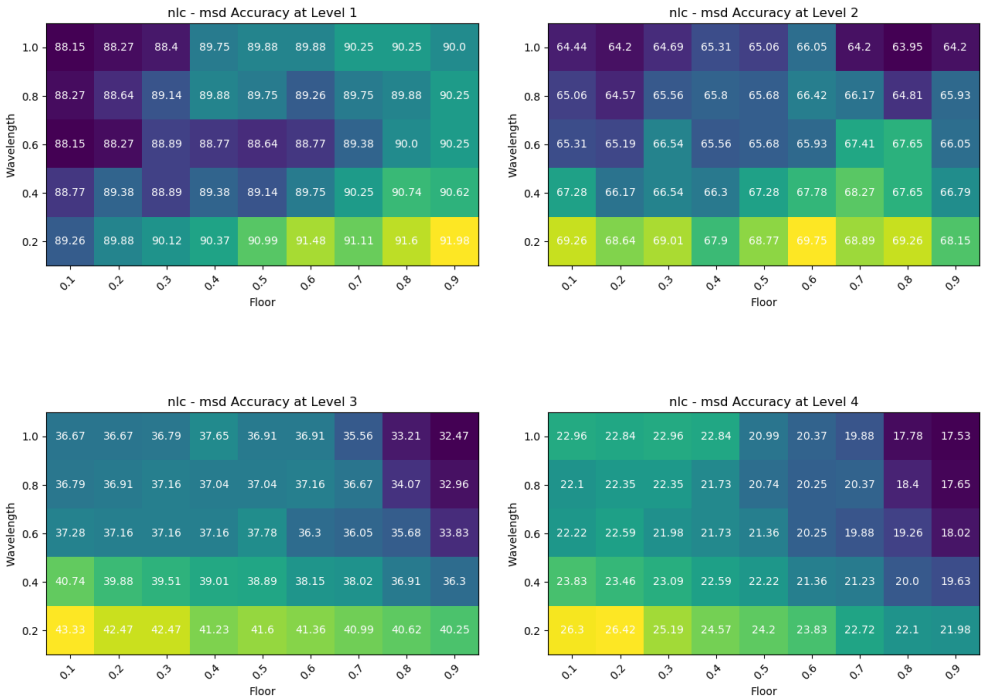

Figure 5 - Accuracy using Euclidean Distance

It is important to note that the different levels correlate to the specificity of the classification. Correctly guessing a value at level 1 means that the type was correct, while at level 2 a correct guess means the class was correct, so on and so forth. The different figures display the different algorithms we use. Figure-2 shows the results for Pearson correlation coefficient, Fgure-3 shows cosine similarity, Figure-4 shows Manhattan distance, and Figure-5 shows Euclidean distance. On the x-axis for these heatmaps is the floor multiplier while on the y-axis it shows the different range values. The accuracies are shown in the middle of the squares. As an example, referring to the level 2 heatmap in Figure-2, we used the Pearson correlation coefficient for these values, and we found that the best accuracy for comparing at level 2 was to use a floor multiplier of 0.2 and a range of 0.2 as well. The floor multiplier of 0.2 means that any values that are less than 0.2 times the mean of the graph were brought up to that value, and the range of 0.2 means that for any given wavelength, we would look to values in the range ±2 μm to determine where the given data point should be placed. This led to an accuracy of 73.83%.

Looking at the rest of the values in the heatmaps, it is easy to see that smaller ranges are much more effective, as 0.2 was the best value in nearly all test runs. The floor, on the other hand, varies. For level 1, the floor value was best at 0.9. For most of the other levels, it was best at 0.1-0.2. We can compare these accuracies with the accuracies gathered when not using NLC. All the values were compiled and shown in Table-2.

Table 2. Performance analysis with respect to NLC

	Manhattan Distance		Euclidean Distance		Cosine Similarity		Pearson Correlation	
Classification Level	w/o NLC	w/ NLC	w/o NLC	w/ NLC	w/o NLC	w/ NLC	w/o NLC	w/ NLC
Level 1	82.8%	83.2%	82.5%	92.0%	80.0%	93.6%	81.3%	95.2%
Level 2	53.6%	63.8%	55.7%	69.8%	55.6%	70.6%	60.3%	73.8%
Level 3	27.3%	39.6%	28.9%	43.3%	32.0%	44.6%	38.9%	49.6%
Level 4	16.3%	24.0%	17.5%	26.4%	20.9%	27.0%	24.7%	30.4%

In almost every situation outlined above, NLC increased the accuracy by a significant amount. Across all levels and algorithms, NLC increased the accuracy by an average of 10.53%.

Section B: Analyzing the effectiveness of clustering

Approach

Before analyzing the effectiveness of using common clustering algorithms, we want to first use different data visualization techniques to be able to look at the data and see if any conclusions can be drawn about the grouping of data that might lend itself to certain algorithms.

In order to be able to visualize the data, we need a way to reduce the dimensionality of the dataset. If we consider each row of values to be a new dimension, then some of the data files have ~2000 dimensions. We need a way to get this down to 3 or 2 dimensions. Therefore, the two main algorithms we explored used for this were Principle Component Analysis (PCA)[2] and T Distributed Stochastic Neighbor Embedding (t-SNE). PCA works by looking at each

dimension in the data and seeing which dimension is least spanned by the data. This dimension that is least spanned would be the one that can be taken out with minimal information loss. A visualization of this can be seen in the figure below, where the PCA 2nd dimension has a much shorter span, so it would be the one that can be removed.

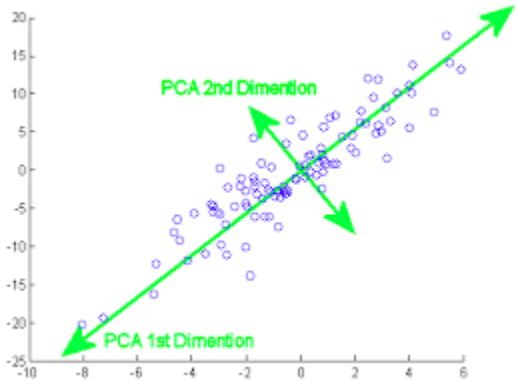

Figure 6 - A graph showing 2 PCA dimensions with the 2nd one being shorter

On the other hand, t-SNE is a probabilistic algorithm that realizes on minimizing the Kullback-Leibler divergence. While t-SNE generally works better at showing clean division in high-dimensional data, it requires more computational power. Because of this, it is common practice to first use PCA to get the dimensionality down to something more manageable for t-SNE, around 40-50.

Plan for Remaining Work

The work regarding clustering in relation to ECOSTRESS dataset and classifying spectrums is still ongoing. Once the data has been analyzed, we will be able to decide what route to take for deciding on the clustering algorithms for the dataset. If the dataset is grouped nicely, we will start with something like DBScan which works well on uniform density data, while if it is not, the first step will most likely be some sort of hierarchical algorithm.

Section C: Analyzing the effectiveness of Convolutional Neural Network (CNN)

Approach

The research for the neural network is split up into three different aspects: classifying spectra by type, classifying spectra by class, and classifying a spectra pair as matching or non-matching. The point of these three aspects is to compare the effectiveness of convolutional neural networks to the similarity algorithms used in Section A.

The CNN implementation for each aspect of the research is a sequential model from the Keras and Tensorflow APIs. The basic structure of the model is a variable number of convolutional and max-pooling layers followed by a variable number of dense layers. The initial convolutional layer will have the input shape of the samples in the dataset. The final dense layer will be the number of possible classes for the output. The variable number of layers for the convolutional, max-pooling, and dense layers is for experimental purposes in search of an optimal model.

Setting up the dataset for spectrum type classification involves taking each entry in the ECOSTRESS dataset and assigning it to its corresponding spectrum type. The data is then inputted into the CNN to create a model for classifying the data by type. Classifying by spectrum class is very similar to classifying by spectrum type. The only difference is every sample in the dataset now is classified by its spectrum class instead of type. Spectra pair classification for matching and non-matching involves a slightly different setup than the previous two aspects. Each sample in the dataset is comprised of two spectra that form a spectra pair. If the two spectra are different samples of the same compound, then they are considered matching. If the two spectra samples are not of the same compound, then they are considered non-matching. The training of the CNN should however be very similar to the previous two aspects.

Plan for Remaining Work

Moving forward, the ECOSTRESS dataset will be used for the three parts listed in the approach. The dataset will be split up into an 80/20 split. 80% of the data will be for training the models, and the 20% will be to validate the model. The validation set is used to ensure overfitting is not occurring with the training set. Each aspect will have its own CNN model created. To find the optimal model, a series of tests will be conducted that manipulates the number of convolutional layers and dense layers, number of nodes at each convolutional layer and dense layer, and the amount of noise added to the dataset.

Section D: Identification of IR spectral features that are commonly observed and that are only observed in specific types of samples in the ECOSTRESS database

Approach

The energy of IR light is close to the vibrational energy of chemical bonds. This makes IR spectroscopy useful in analysis and identification of chemical compounds. Therefore, in parallel to the spectra matching and classification approaches that encompasses to matching, classification, and clustering of IR spectra, we are examining the FTIR spectra of various class of compounds and materials in ECOSTRESS database to identify the unique spectrum characteristics that can be utilized as signature of each compounds. In addition, we plan to identify the regions of spectra on which higher degree of noises appears and/or on which peaks that tend to reflect characteristics that are not unique to each compound or material, such as water content.

Current Progress/Results

Our initial examination indicated that water samples show relatively small but some degree of reflectance around 2-3 μm and 10-12 μm. This suggest that the reflectance around these regions may come from variations in water content of samples (see Figure-7).

Figures 8-9 are some of the selected examples of the spectra for photosynthetic vegetation samples and non-photosynthetic vegetations. In general, these vegetation samples exhibited more peaks within 2-14 μm region. The spectra of many vegetation samples are resembling to each other even though each doe show some features of differences. In general, the spectra for the identical samples share more common spectral features in comparison to the spectra from other

species. As it can be seen in Figure-9, greater variations were observed in the spectra for non-photosynthetic vegetation samples. One of the interesting finding was that one or two peaks were observed around 10-12 µm in some of the samples whereas not in the other samples. At this point, it is uncertain that the differences like this comes from the fact that the differences on the parts of samples or those are more unique characteristics of certain plant species.

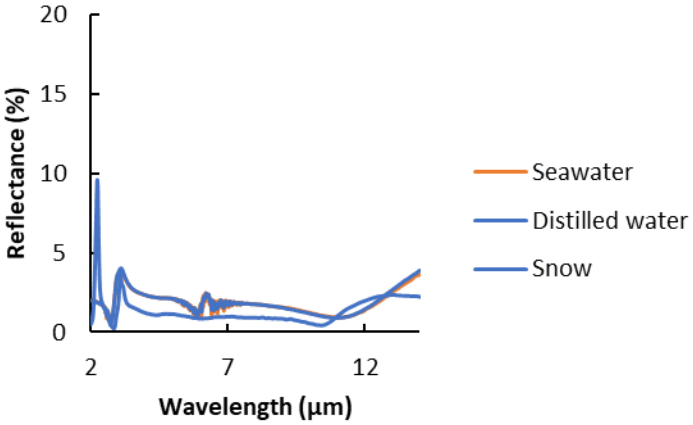

Figure 7 – IR reflectance spectra for water samples

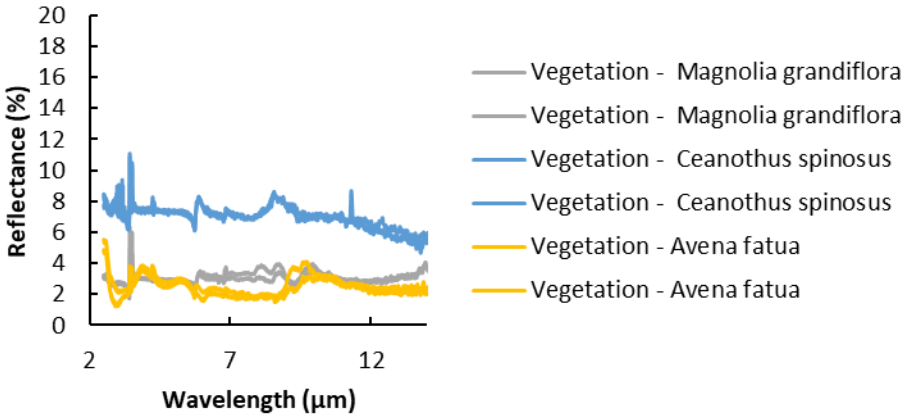

Figure 8 – IR reflectance spectra for selected examples of photosynthetic vegetation samples

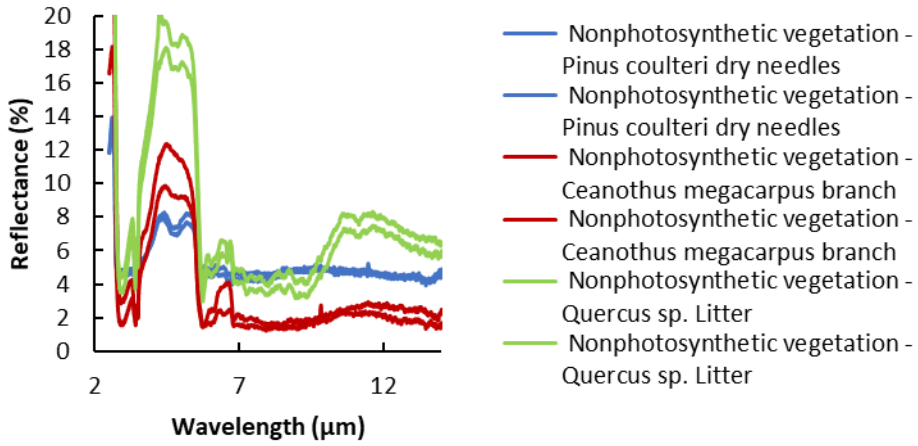

Figure 9 – IR reflectance spectra for selected examples of non-photosynthetic vegetation samples

Plan for Remaining Work

Based on the preliminary observations, we formulated a few hypotheses: the peaks that were observed on non-photosynthetic vegetation samples derives from either differences in the parts of the vegetations or unique characteristics of the plants; reflectance below 3 µm and above 10 µm tends to contain more noise due to the water content; generally, non-synthetic vegetations samples absorbs more IR light absorption 3-14 µm. In order to examine these hypotheses, we are currently working to compare the class-to-class in a systematic manner. In addition, we will continue working on examining the spectral characteristics on the spectra in the classes of soil, minerals, and rock. We hope to summarize the observations and determining the regions of spectra on which higher degree of noises appears and/or on which peaks that tend to reflect characteristics that are not unique to each compound or material, such as water content. Our plan is to incorporate these finding to assign proper weighing on computational methods for matching, classification, and clustering of IR spectra.

Acknowledgments

We would like to give a big thank you to Dr. Razib Iqbal & Dr. Keiichi Yoshimatsu, who guided us through the project and always were there to offer guidance and support. We would also like to thank Joshua Ellis, who was a fountain of ideas and always willing to let us bounce our ideas off him. Lastly, we would like to thank the NASA Missouri Space Grant Consortium for sponsoring our work, without which none of this would have been possible.

Biography

Patrick Ayres was born in Springfield, Missouri, and has lived there his whole life. He is a senior in Computer Science at Missouri State University and graduates in Spring 2020 with his B.S. He is currently a junior software developer at O'Reilly Auto Parts. He is a lifelong learner with a passion for knowledge from any source. He spends his free time watching math videos, playing video games, and going for runs/bike rides.

Gattlin Walker is originally from Bolivar, Missouri. He is a senior at Missouri State University and will graduate in the Spring of 2020 with a B.S. degree in Computer. After graduation he plans to work as a software developer in the Springfield area. In his free time, he enjoys reading, working on side projects, and spending time with friends and family.

Eathan Hickey is a Chemistry major at Missouri State University. He has been participating in undergraduate research projects in the laboratory of Dr. Yoshimatsu since Spring 2019. He is currently seeking for the career opportunities after the graduation. At home, he enjoys spending his free time on playing with his dog.

References

1. Joshua D. Ellis, Razib Iqbal, KeiichiYoshimatsu, *Anal. Chim. Acta* **2020**, *1103*, 49-57.2. Rasmus Bro a and Age K. Smilde, *Anal. Methods* **2014**, *6*, 2812-2831.