

Gaussian Model on Emission Spectroscopy Data with Cloud Computing

Amita Behuria Pathak, Joanna Duran, Sheri Loftin, Anjali Solsi

Abstract— The analysis of light from emission spectra is one of the ways we know the composition of the atmospheres of the various moons and planets in our solar system. Recent advances in the field of cloud computing make it possible to perform more complex functions and in a shorter amount of time. Using synthetically created spectra dataset, we show that computing methods no longer have to be time consuming and can be advanced with the use of cloud computing. This demonstrates that cloud computing can improve the identification of elements present in spectrographic signal.

Index Terms—Cloud Computing, NASA, Gaussian Model, Emission Spectroscopy, Google Cloud Platform

I. INTRODUCTION

ONE of the more time consuming and work intensive tasks of an astronomer is identifying the elements present in a spectrographic signal. The analysis of light from emission spectra is one of the ways we know the composition of the atmospheres of the various moons and planets in our solar system. With the explosion of new exoplanets discovered in the last decade, analyzing atmospheric spectroscopy is becoming even more important. “About 20 years after the discovery of the first extrasolar planet, the number of planets known has grown by three orders of magnitude, and continues to increase at neck breaking pace. For most of these planets we have little information, except for the fact that they exist and possess an address in our Galaxy. For about one third of them, we know how much they weigh, their size and their orbital parameters. For less than 20, we start to have some clues about their atmospheric temperature and composition.”[1] With the increase in targets, a more precise and time saving method of analyzing the spectra from these targets is needed.

The Gaussian Mixture Model (GMM) is a very precise and exact method for the task of examining emission spectra that have thermal doppler spreading. Pairing this method, which can be time consuming in itself, with cloud computing further streamlines the process of analysis. The need for tools to examine these spectra is real. “Understanding a planet’s atmosphere is a necessary condition for understanding not only the planet itself, but also its formation, structure, evolution, and habitability. This requirement puts a premium on obtaining spectra and developing credible interpretative tools with which

to retrieve vital planetary information. However, for exoplanets, these twin goals are far from being realized.” [2]

II. EMISSION SPECTROSCOPY

When elements are heated, they emit energy in the form of light. With the help of a spectroscope, a discontinuous spectrum can be viewed. A spectroscope or a spectrometer is an instrument which is used for separating the components of light, which have different wavelengths. The atomic spectrum appears in a series of lines called the line spectrum (Fig 1). Each element emits a characteristic set of discrete wavelengths according to its electronic structure, and by observing these wavelengths the elemental composition of the sample can be determined. Each element has a different atomic spectrum and therefore have their own line spectrum. The emission spectrum can be used to determine the composition of a material, since each element of the periodic table has its own line spectrum.

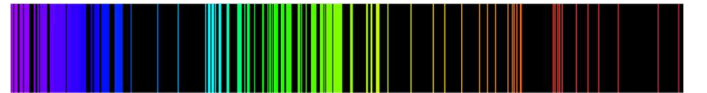


Fig. 1 Emission spectra of iron [3]

Emission spectroscopy is a spectroscopic technique which examines the wavelengths of photons emitted by atoms or molecules during their transition from an excited state to a lower energy state. Emission spectroscopy developed in the late 19th century and efforts in theoretical explanation of atomic emission spectra eventually led to quantum mechanics [3]. The dataset used in this paper are astronomical emission spectra, whose purpose is to identify the composition of stars by analyzing the received light. Data is captured through spectroscope (Fig 2)

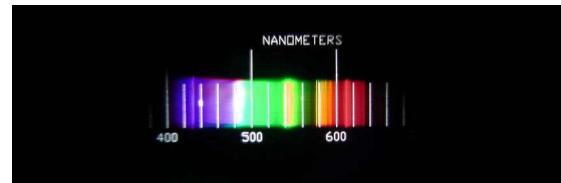


Fig. 2 An example of spectrum captured by a spectroscope¹

This project involves identifying the peaks in an emission spectrum. The spectra for this data set is noted in Fig 3. The spectra behave this way because the particles in the

¹<https://www.nasa.gov/image-feature/spectroscopy>

atmosphere of the planet being observed are moving relative to one another. This causes a shift in the light emission due to doppler shifting. Doppler shift creates a gaussian curve.

Each peak represents a wavelength that is being produced by an element in the atmosphere. By identifying the peaks, the element can be identified and therefore it can be deduced what the atmosphere is made of. This process can be used on planets, stars and nebulae. This method is how it is determined what the universe is made of.

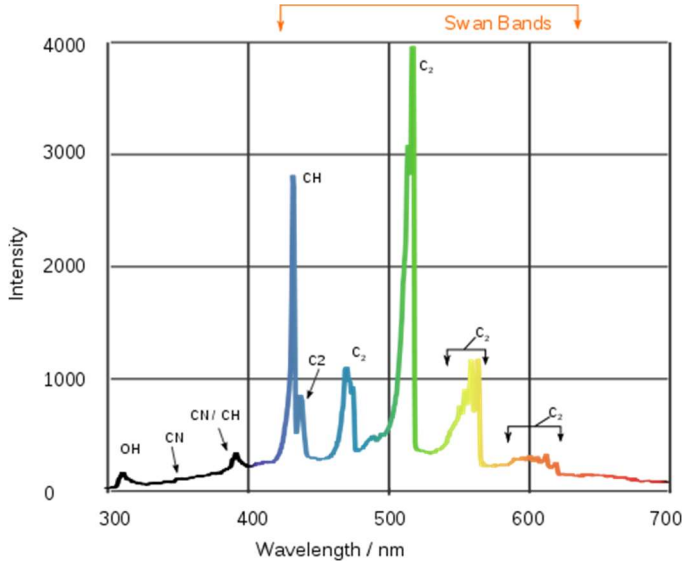


Fig. 3 Spectra Data²

III. DATA

A. Acquire data

The data was acquired by Sheri Loftin with the permission of NASA. The data is synthetically created spectra using only 20 elements (instead of the entire 118 on the periodic table). The spectral curve from the dataset is pictured in Fig. 4. This dataset creation was part of a team lead by Dr. Tim McClanahan, Evana Gizzi, and Sheri Loftin. The project is not going forward at this time due to Dr. McClanahan having other commitments.

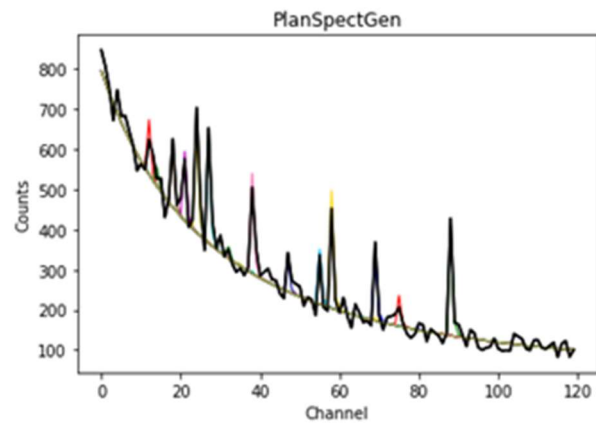
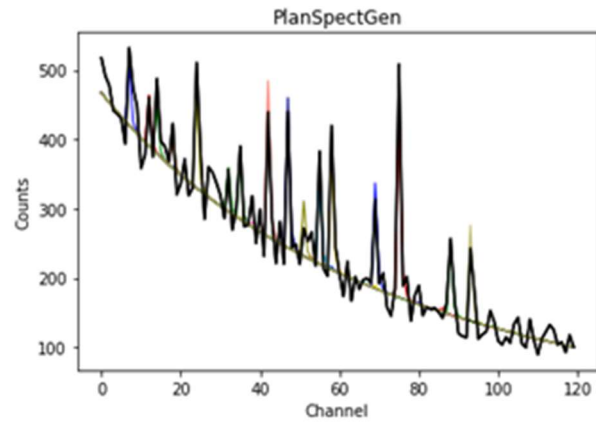


Fig. 4 Spectral curve

B. Data Description

The data is comprised of five very large files (Fig. 4). One file "Backgrd.csv" which is comprised of some background noise level that has been incorporated into the spectra files. There are four files "CCSpectra.csv" that each contain 50,000 spectra emission signals. Each signal contains 120 wavelength measurements at various strengths, emulating a real signal.

Buckets / ccproject-spectra							
<input type="checkbox"/> Name	Size	Type	Storage class	Last modified	Public access	Encryption	
<input type="checkbox"/> Backgrd.csv	51.16 MB	application/vnd.ms-excel	Standard	11/19/19, 3:02:45 PM UTC-6	Not public	Google-managed key	
<input type="checkbox"/> CCSpectra.csv	59.37 MB	application/vnd.ms-excel	Standard	11/7/19, 9:59:39 AM UTC-6	Not public	Google-managed key	
<input type="checkbox"/> CCSpectra2.csv	51.16 MB	application/vnd.ms-excel	Standard	11/7/19, 12:12:54 PM UTC-6	Not public	Google-managed key	
<input type="checkbox"/> CCSpectra3.csv	51.16 MB	application/vnd.ms-excel	Standard	11/7/19, 12:20:27 PM UTC-6	Not public	Google-managed key	
<input type="checkbox"/> CCSpectra4.csv	51.16 MB	application/vnd.ms-excel	Standard	11/7/19, 12:19:28 PM UTC-6	Not public	Google-managed key	

Fig. 5 Datasets

IV. ADVANTAGES OF CLOUD

There are many advantages to moving data computing to cloud based computing. Advantages that have been realized are as follows.

² https://en.wikipedia.org/wiki/File:Spectrum_of_blue_flame.svg

A. Flexibility

Moving the data to cloud helps to scale up or down as per demands, it's ideal for growing or fluctuating bandwidth demands.

B. Disaster Recovery

Disaster recovery and backups are offered by public and private cloud providers as a managed service. Users have the option to build their own disaster recovery solution using Google's regular cloud services. Google's disaster recovery cookbook explains how to setup disaster recovery.

C. Capital -expenditure Free

Cloud computing cuts out the high cost of hardware. Ease of setup and management.

D. Increased collaboration

When your teams can access, edit and share documents anytime, from anywhere, they're able to do more together, and do it better. Cloud-based workflow and file sharing apps help them make updates in real time and gives them full visibility of their collaborations.

E. Document Control

The more employees and partners collaborate on documents, the greater the need for watertight document control. All files are stored centrally and everyone sees one version of the truth. Greater visibility means improved collaboration, which ultimately means better work

F. Security

Because your data is stored in the cloud, you can access it no matter what happens to your machine. And you can even remotely wipe data from lost laptops so it doesn't get into the wrong hands. Risk of loss of sensitive data is reduced.

G. Competitiveness

Moving to the cloud gives access to enterprise-class technology, for everyone.

H. Environmentally friendly

When cloud needs fluctuate, server capacity can scale up and down to fit the requirements needed. Only the energy needed is used and reduces carbon footprints.

V. MIGRATE DATA TO CLOUD

The provider chosen for this project was Google Cloud Platform. It was chosen because of the user-friendly nature of the services needed for this analysis in the cloud. From uploading and storing the data to performing the analysis through the Jupyter notebook in the AI Platform, this service provided all necessary functionality.

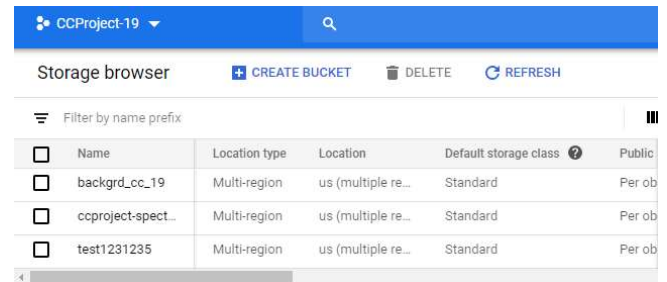
A. Google Cloud Platform

Google Cloud Platform is a suite of cloud computing services that is run on the same infrastructure used for Google Search. Some of the services provided include computing, networking, data storage, data analytics, and machine learning. It provides

infrastructure as a service, platform as a service, and serverless computing environments. It is available in approximately 20 regions, each with three zones where users can deploy cloud resources.

B. Creating Bucket and Uploading Data

In the Storage section of Google Cloud Platform there is a sub-section named Browser. In Storage browser a bucket was created called ccproject-spectra, as seen in Fig. 6, to house the various data files. The bucket created stored the data in multi-region location, which allows for the highest availability across the largest area. A standard storage class was used as it is the best for short-term storage and frequently accessed data. Once the bucket was created, the four spectra csv files were uploaded.



<input type="checkbox"/>	Name	Location type	Location	Default storage class	Public
<input type="checkbox"/>	backgrd_cc_19	Multi-region	us (multiple re...	Standard	Per ob
<input type="checkbox"/>	ccproject-spect...	Multi-region	us (multiple re...	Standard	Per ob
<input type="checkbox"/>	test1231235	Multi-region	us (multiple re...	Standard	Per ob

Fig. 6 Bucket created

C. Creating Notebook and Compute

In the AI Platform section of Google Cloud Platform there is a section names Notebooks. In Notebook instances, a new instance was created with Python called ccproject and configured with the default python environment and packages as well as default subnetwork. From the instance, the JupyterLab and notebook created can be directly accessed. This analysis focused on creating Gaussian mixture models to identify the elements from emission spectra. The Python notebook, as seen in Fig. 7, contained our code for specifying the required libraries and functions, reading in the data from the bucket, modeling the data, and plotting the models. The models will be detailed in the next section.

```
spectra_analysis.ipynb
[2]: from __future__ import absolute_import, division, print_function
    from __future__ import print_function

import os
import seaborn as sns
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

from sklearn import mixture, svm
from sklearn.decomposition import PCA
from sklearn.cluster import KMeans #Code source: https://jakevdp.github.io/P
from scipy.spatial.distance import cdist
from sklearn.mixture import GaussianMixture as GMM
from sklearn.mixture import BayesianGaussianMixture as BGM

...
Plan:
1. Read in data (Spectra.csv and Backgrd.csv)
2. Label the backgrd values and see if we can pick out the remaining peak
3. Run a Gaussian Mixture model on the peaks information to see if we can sep

[2]: '\nPlan: \n      1. Read in data (Spectra.csv and Backgrd.csv)\n      2. Label t
3. Run a Gaussian Mixture model on the peaks information to see if we can sep

[5]: #FOR SPECTRA DATA
def plot_kmeans(kmeans, X, n_clusters, rseed=0, ax=None):
    labels = kmeans.fit_predict(X)

    # plot the input data
```

Fig. 7 Jupyter notebook

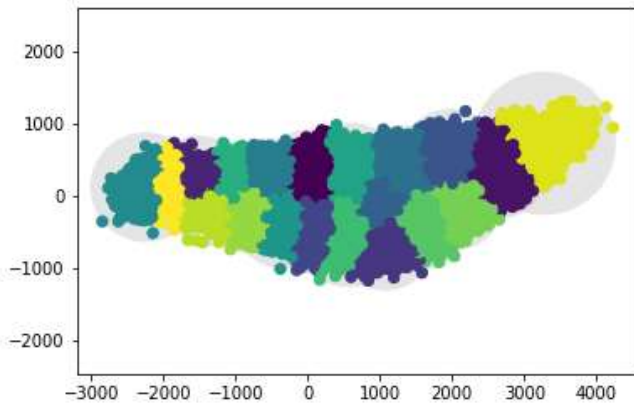


Fig 8. Plot of K-Means model

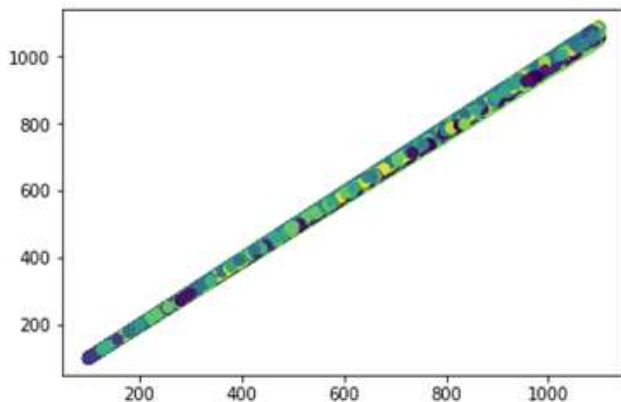


Fig 9. Plot of GMM model

VI. ANALYSIS

The analysis focused on running a Gaussian mixture model on the peaks information from the spectra data to determine whether the element signatures can be separated and identified.

When a set of data points is grouped together in some form based on similarities, that is referred to as clustering in terms of machine learning. Clustering is a type of unsupervised learning, where input data does not have labeled responses. Two types of clustering used in this analysis are K-means clustering and Gaussian Mixture Models. The k-means algorithm starts by randomly initializing the cluster centers. Then, each point in the dataset is assigned to the closet cluster using the Euclidean distance, and the position of the point is updated. [5]

The central limit theorem states that as more samples are collected from a dataset, it tends to resemble a Gaussian distribution, which makes this type of model very useful. The probability density function is used to tell the probability of observing an input x , given that specific normal distribution. The foundation of Gaussian Mixture Models (GMM) is to find the parameters of the Gaussians that best explain the data, which is referred to as generative modeling. The first step is to compute the probability that each data point was generated by each of the k (number of clusters) Gaussians. The next step is to update the weights, means, and covariances. [6]

The spectrum of all planetary objects has two peaks. One is in the visible region due to reflected sunlight and the second is in the infrared due to thermal emission of the planet. An example of the emission spectrum can be seen in Fig. 10. This is the emission spectrum from Titan's stratosphere, which is the largest moon of Saturn.

Before the data is read in, there is a function defined for plotting the representation of the k-means model and initializing parameters. After the data is read in, both the K-means and Gaussian mixture model are performed. The output plots of the respective models are shown in Fig. 8 and Fig. 9. Since there are twenty elements in the data, the number of clusters/components is specified as 20 in the models.

For this Gaussian mixture model we have used unsupervised learning. There are some advantages with unsupervised learning:

- “Unsupervised machine learning finds all kind of unknown patterns in data.
- Unsupervised methods help you to find features which can be useful for categorization.
- It is taken place in real time, so all the input data to be analyzed and labeled in the presence of learners.
- It is easier to get unlabeled data from a computer than labeled data, which needs manual intervention.” [7]

There are also some difficulties:

- “You cannot get precise information regarding data sorting, and the output as data used in unsupervised learning is labeled and not known

- Less accuracy of the results is because the input data is not known and not labeled by people in advance. This means that the machine requires to do this itself.
- The spectral classes do not always correspond to informational classes.
- The user needs to spend time interpreting and label the classes which follow that classification.
- Spectral properties of classes can also change over time so you can't have the same class information while moving from one image to another.” [7]

Our project has encountered some of these difficulties, primarily in having the necessary time to interpret and label the classes that follow the classification. The output from our GMM, seen in Figure 9, lacks a discernable structure. In addition, there are no labels on the axes to help us determine what we are seeing. The timeframe of this project has not allowed us enough time to perform an adequate analysis.

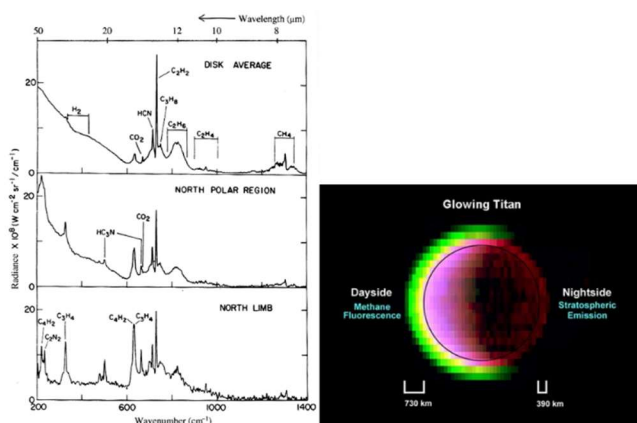


Fig 10. Sample emission spectrum [4]

VII. FURTHER RECOMMENDATIONS

This project started as an effort to streamline and improve the methods used to analyze emission spectra. This process is complicated and will require further work to fully understand the results of the analysis.

One of the primary difficulties is the difficulty that frequently presents itself when using unsupervised learning: what structure has the model found.

We provided our algorithm with minimal parameters: that we expect the signal to contain the summation of 20 gaussian curves. We are expecting the system to find that each curve is the summation of the gaussian curves related to the peak wavelengths, at various signal strengths, of each of the 20 elements. Until we do further analysis, we will not be able to confirm this expectation exists in the results.

The next steps include:

- Adjusting the visualization to enhance the results of the model.
- Once understanding of what structure has been found in the data, we will need to confirm whether it is the

structure we desired to see or if further algorithm adjustment is needed.

- If the model is shown to support our results, further testing is needed to ensure the results are real and not just a quirk of this dataset.

There is future work potential beyond this as well:

- Expansion of the dataset components to include more elements.
- Addition of molecular data to enhance the model. See Figure 10.

The importance of this project is clear and this work will be done whether by our team or by another. Whether GMM ends up being the most useful method remains to be seen.

VIII. CONCLUSION

It is evident throughout our assignment that valuable data available on-premise can be used to gain new insights. Analysis of the data helped the team to get visibility into the elements present in a spectrographic signal that help understand the composition of the planet's atmosphere. If the team has insight they are able to understand complex situations.

The foundational approach is to leverage the benefits of Cloud computing, use advanced techniques to analyze the available data to get insights. This analysis can be extended to large amount of data and all the elements that are available. Additional techniques of data analysis can be used in future to arrive at enhanced results.

APPENDIX

See Jupyter Notebook for Analysis

ACKNOWLEDGMENT

The authors would like to thank Dr. Sohail Rafiqi for his guidance in the development of this paper. The authors would also like to thank NASA for the opportunity to use the data.

REFERENCES AND FOOTNOTES

REFERENCES

- [1] Tinetti, G., Encrenaz, T. & Coustenis, A. *Astron Astrophys Rev* (2013) 21: 63. <https://doi.org/10.1007/s00159-013-0063-6>
- [2] A.S. Burrows, "Spectra as windows into exoplanet atmospheres", *PNAS* September 2, 2014 111 (35) 12601-12609; first published January 13, 2014 <https://doi.org/10.1073/pnas.1304208111>
- [3] "Astronomical Spectroscopy." Wikipedia, Wikimedia Foundation, 7 Nov. 2019, https://en.wikipedia.org/wiki/Astronomical_spectroscopy
- [4] Bagenal, Fran. "Spectroscopy of Atmospheres." University of Colorado, Boulder. <http://lasp.colorado.edu/~bagenal/3720/CLASS5/5Spectroscopy.html>

- [5] Dabbura, Imad (2018, September 17). Gaussian Mixture Models Explained. K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks. Retrieved November 20, 2019, from <https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a>.
- [6] Carrasco, O. C. (2019, August 21). Gaussian Mixture Models Explained. Retrieved November 21, 2019, from <https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95>.
- [7] “Unsupervised Machine Learning : What it, Algorithms, Example. ” <https://www.guru99.com/unsupervised-machine-learning.html>.