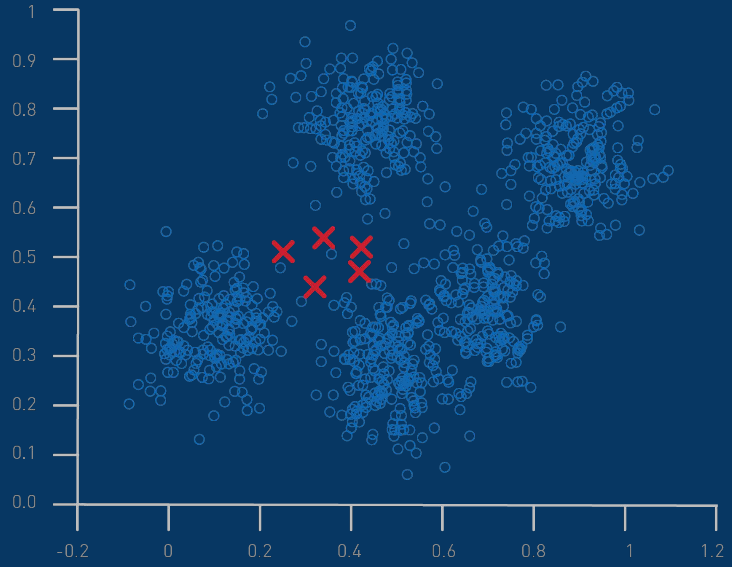
# N-Dimensional Auto-Clustering of Lidar Measurements to define Defacto Aerosol Types





by Zohaer Al Mahatab, *St. Edward's University*

Mentor: Connor Flynn


ARM

# Contents

# Problem

- Remote sensing instruments like HSRL provide rich information about aerosol and cloud properties throughout the atmospheric column.

- It does NOT provide good interpretation despite being sophisticated and depends a lot on the eye of beholder.

# Hypothesis

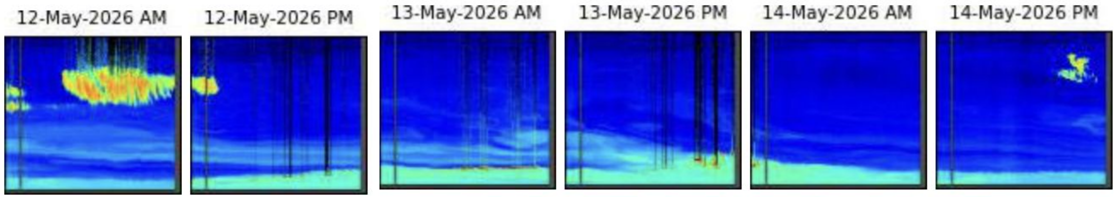
- Each observation is a 3D point in space. Do natural clusters emerge and if we color-code them by aerosol type, do they match what scientists already know?

- If their clusters can be reproduced with cheap uncalibrated LIDARs, research-grade data can be extended to sites that cannot afford HSRL.

# Instruments & Data

- High Spectral Resolution Lidar (HSRL) measurements from 2026-05-12 to 2026-05-14, (Holz, R., Garcia, J., Schuman, E., Bambha, R., Ermold, B., Eloranta, E., & Garcia, J.)

- Bankhead National Forest, AL, USA; Long-term Mobile Facility (BNF), Bankhead National Forest, AL, AMF3 (Main Site) (M1).




ARM

# Data Analysis

- Features: Color Ratio, Linear Depolarization 532 & 1064, No. of Aerosol Photons 532 & 1064, No. of Molecular Photons 532 & 1064, Attenuated Molecular Backscatter
- Plot each physical property and then adjust the plotting ranges to see the useful high contrast regions.
- Apply screens to filter only physically meaningful pixels enter the clustering. Otherwise they would distort the cluster centroids and produce meaningless groupings.

ARM

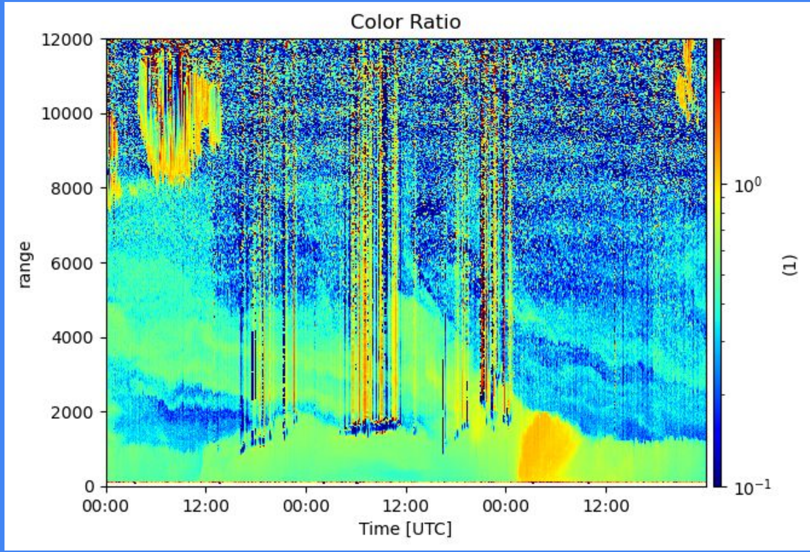



Fig 1: Color Ratio

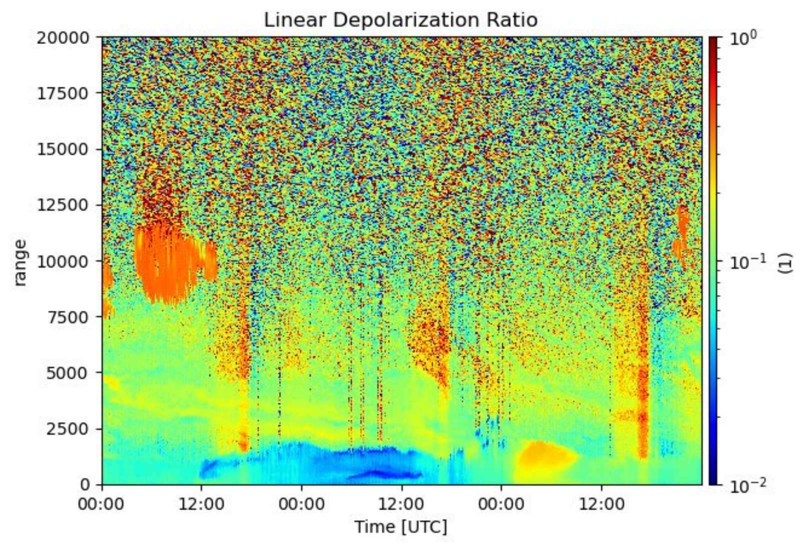



Fig 2: Linear Depolarization Ratio

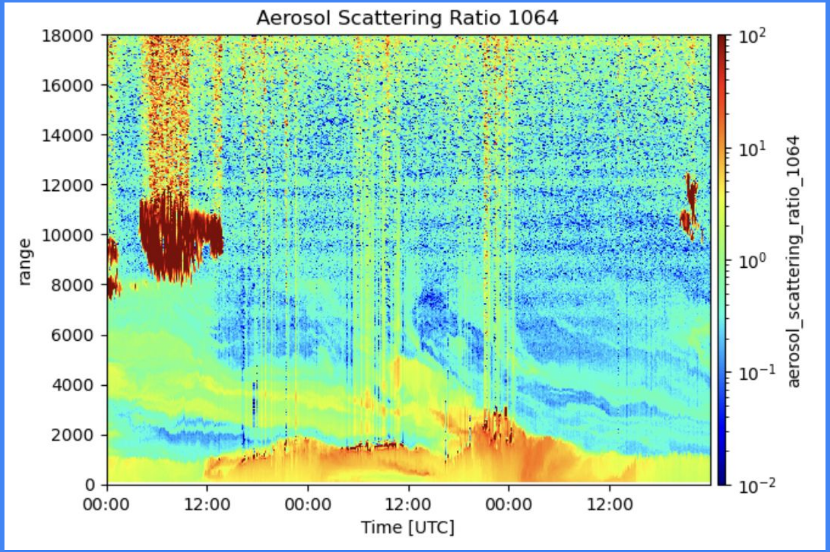



Fig 3: Aerosol Scattering Ratio

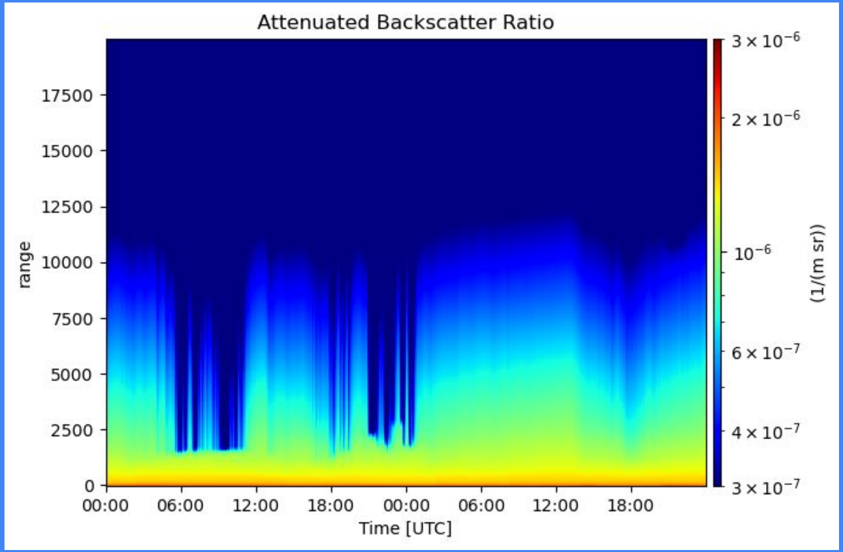



Fig 4: Attenuated Backscatter Ratio

# Feature Engineering

- Aerosol Scattering Ratio (ASR) is top-heavy. To normalize it, we did the total scattering and inverted it. We are calling it Inverse Scattering Ratio (ISR).

$$ISR = 1 + \frac{1}{ASR}$$

- Color Ratio and Linear Depolarization have a small range (~0-1) but are very top heavy. We converted them to their log-scale to create uniform distribution.

- Filtered out max values exceeding empirical limits.

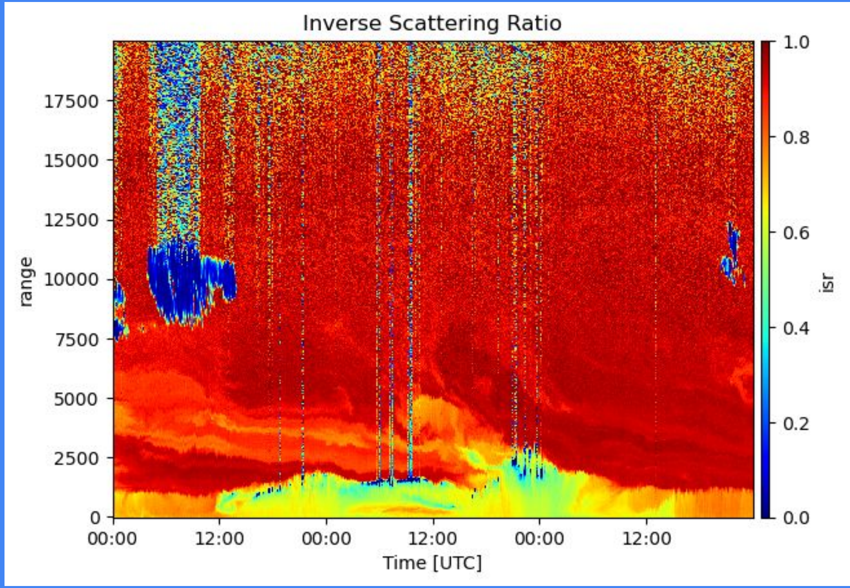



Fig 5: Inverse Scattering Ratio

# K-Means Clustering

- Selected Features for Clustering: Log Color Ratio, Log Linear Depolarization Ratio, Inverse Scattering Ratio.
- Removed negative LDR, CR, ISR. They arise from noise, dark current subtraction, or retrieval artifacts. Keeping them would skew K-means cluster centroids since it's a distance-based algorithm.
- Flattened each feature array and stacked them on top of each other to create a 2D features Array.
- Removed all rows with NaN values for any of the 3 selected features. Apply Standard Scaler to the features Array.

# K-Means Clustering

- Executed Elbow Method and Silhouette Scoring to find ideal k.

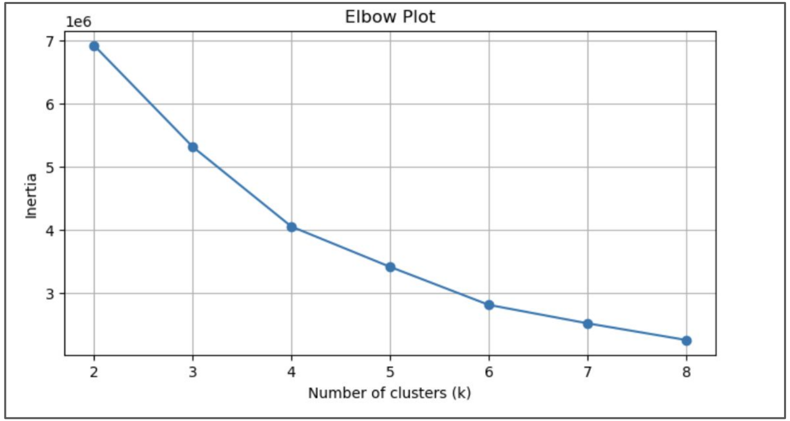



Fig 6: Elbow Method

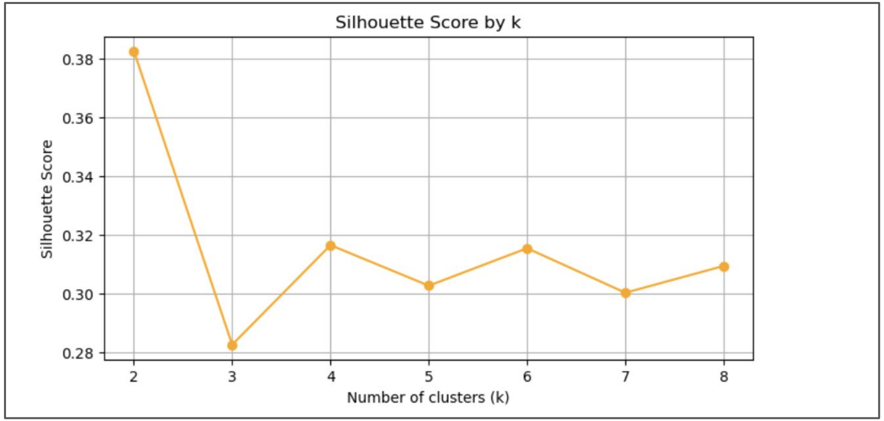



Fig 7: Silhouette Score

# Results

K = 4

Cluster 1: 535088 pixels (16.8%)

Cluster 2: 109023 pixels (3.4%)

Cluster 3: 1056790 pixels (33.1%)

Cluster 4: 1488052 pixels (46.7%)

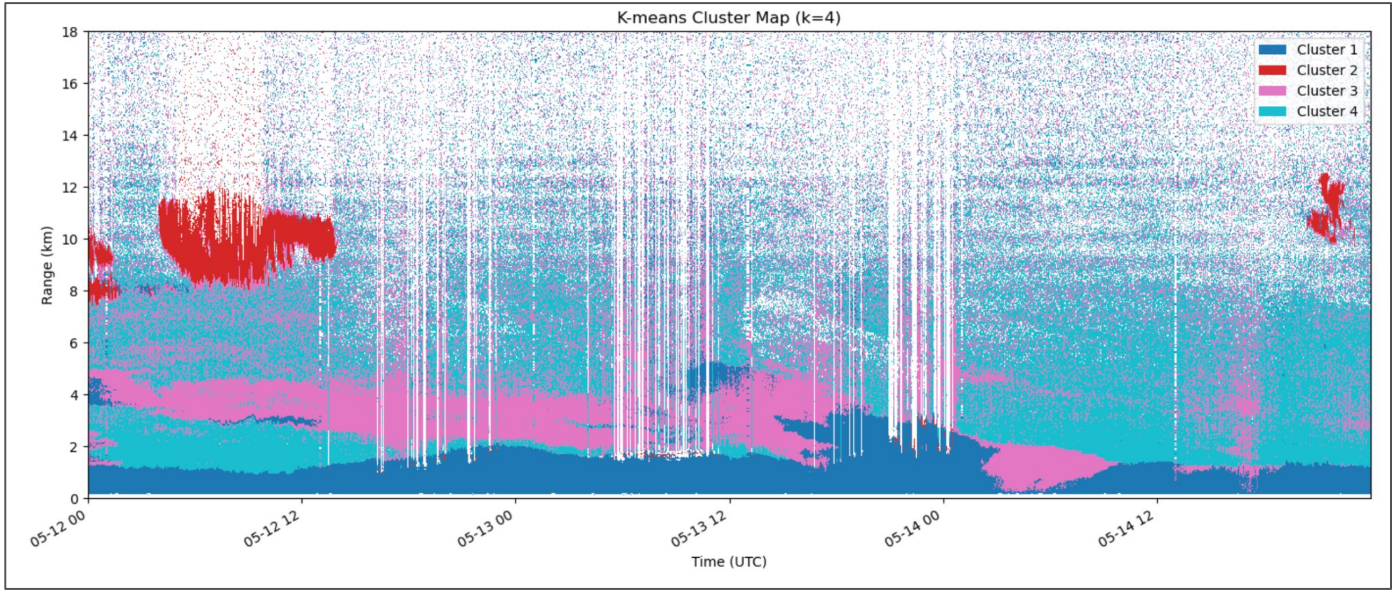



Fig 8: K-Means Clustering (K=4)

# Results

K = 5

Cluster 1: 526680 pixels (16.5%)

Cluster 2: 107528 pixels (3.4%)

Cluster 3: 365010 pixels (11.4%)

Cluster 4: 782503 pixels (24.5%)

Cluster 5: 1407232 pixels (44.1%)

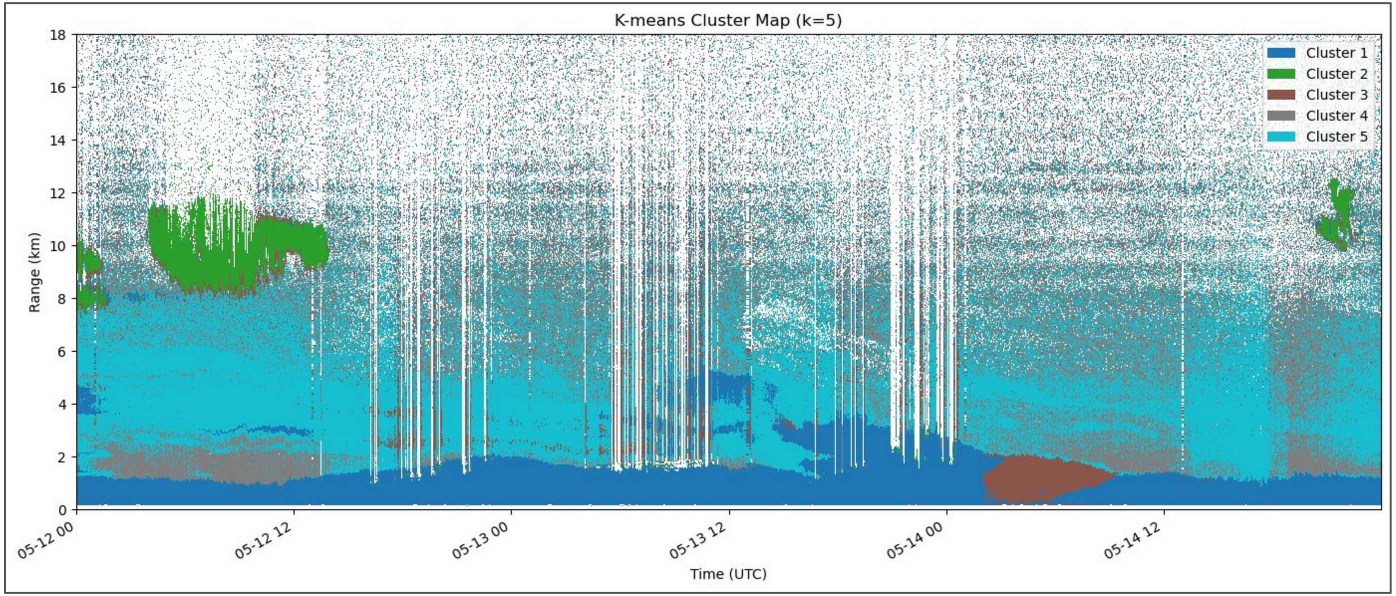



Fig 9: K-Means Clustering (K=5)

# Future Work

- Trying DBSCAN to handle outliers better and find ideal k.

- Looking more into the impact of the features of used for clustering and fine-tune them.

- **Reproduce clusters with cheap uncalibrated LIDARs.**

# Thank You!