AI201 Mini-Project
Aerosol Classification using AERONET Optical Properties: A Case Study at the Manila Observatory

Hans Jarett Ong and Rossjyn Fallorina

Date of Submission: January 6, 2023

1. INTRODUCTION

Aerosol particles play a crucial role in the study of earth and climate systems as they influence the radiative and thermodynamic properties of the atmosphere [1]. This is particularly relevant in the warm tropical environment of Southeast Asia (SEA), where there are still many gaps in knowledge and large uncertainties regarding the relationship between aerosol radiative properties and the atmosphere's thermodynamic properties. Moreover, the SEA region is also known to be vulnerable to climate impacts [2]. In order to better understand the SEA climate system, the 7-Southeast Asian Studies (7SEAS) mission was established with the objective of facilitating interdisciplinary research into the integrated SEA aerosol environment via grass roots style collaboration [3]. In particular, 7-SEAS uses ground-based, remotely-sensed, and modeled data sets to study the aerosol-environment interaction in the region of Java through the Malay Peninsula and SEA to Taiwan.

One of the key challenges in the study of aerosols is determining their types, which is usually done using chemical sampling and analysis. However, these methods can be expensive and time-consuming, and there are many places where chemical sampling data are limited or unavailable. In such cases, remote sensing provides an advantage as it allows for the continuous gathering of large amounts of data with low maintenance. One widely used remote sensing system in the 7-SEAS mission is NASA's Aerosol Robotic Network (AERONET) [4]. AERONET is a global network of ground-based sun photometers that can continuously measure aerosol optical properties such as absorption, scattering, optical depth, and aerosol size distributions.

In this study, we will focus on AERONET data from the Manila Observatory which has available data starting from January 2009. This long-term data will allow us to examine the temporal variability and seasonality of aerosol. Choosing Manila Observatory is meant to be a proof of concept, and the analysis presented here can be extended to other sites as well.

2. OBJECTIVES

The main objective of this study is to create a model that can classify aerosols using their optical properties. Using this classifier, we aim to accurately identify the types of aerosols present at the Manila Observatory and understand their temporal variability. In addition, we aim to compare our results with known events and trends, such as examining how the distribution of aerosol types changed during the COVID-19 pandemic which might show the relationship between aerosol properties (e.g. pollution levels) and human activities.

3. METHODOLOGY

For this project, we will focus on supervised classification, relying on domain knowledge to label some of the training data. For example, we may label aerosols from highly-urbanized cities like Beijing as "urban dust" (a mixture of dust and smoke particles) or label aerosols from areas with agricultural biomass burning processes, such as Alta Floresta, Brazil during harvest season, as "biomass burning white smoke." Once we have labeled the training data, we can explore different classification techniques.

3.1 Data Description

3.1.1 AERONET

NASA's AERONET is a global network of sun photometers that continuously measure aerosol optical properties. These photometers, known as the CIMEL Electronique CE-318 sun-sky radiometers, are installed at various locations around the world and have a 1.2° field of view and 2 detectors for measuring direct sun and sky radiance. The sun photometers can operate in two modes: direct sun and sky radiance. Direct sun measurements are performed every 15 minutes at multiple wavelengths, while sky radiance measurements are taken at various scattering angles to deduce particle size distribution and phase functions.

AERONET data is available in three quality levels: 1.0, 1.5, and 2.0. Level 1.0 data is raw and unprocessed, level 1.5 data is cloud-screened using a specific algorithm [5], and level 2.0 data is both cloud-screened and quality-assured through instrument tests and manual inspection [6]. Level 2.0 data is considered to have the best quality and will thus be used for this project.

3.1.2 Aerosol Optical Properties

Aerosol Optical Thickness (AOT) is a measure of the extinction of light due to aerosol. AERONET sun photometers calculate AOT using the spectral extinction of the direct beam radiation according to the Beer-Lambert-Bouguer Law:

$$V_\lambda = V_{0\lambda} d^2 e^{-\tau_\lambda m} \cdot t_y \tag{1}$$

where V is the digital voltage, V_0 is the extraterrestrial voltage, m is the optical air mass (which is approximately the secant of the zenith angle), τ is the total optical depth, λ is the wavelength, d is the ratio of the average to the actual Earth-Sun distance, and t_y is the transmission of the absorbing gasses. The AOT (τ_a) is the τ minus the absorption by atmospheric gasses, water vapor, and the effects of Rayleigh scattering [4]:

$$\tau_a = \tau - \tau_{H_2O} - \tau_{Rayleigh} - \tau_{O_3} - \tau_{NO_2} - \tau_{CO_2} - \tau_{CH_4} \quad (2)$$

The Angstrom Exponent (AE) can be derived from the AOT per wavelength. AE is defined to be the slope of the AOT with respect to the wavelength in a logarithmic scale:

$$\alpha = -\frac{d \ln \tau_a}{d \ln \lambda} \quad (3)$$

where α is the Angstrom Exponent, τ_a is the AOT, and λ is the wavelength [7]. AE is a particle size indicator where $AE < 1$ suggests the dominance of coarse aerosols while $AE \geq 2$ suggests the dominance of fine aerosols. The derivative of AE with the wavelength, α', is also a good indicator for particle size where $\alpha' > 0$ suggests the dominance of fine aerosols and $\alpha' < 0$ suggests the dominance of coarse aerosols. α' is obtained from the second order polynomial fit of AOT vs wavelength in log-log space [8].

More parameters that describe particle size and absorption can be derived using the inversion algorithm. The inversion algorithm was developed using almucantar and principal plane measurements as inputs in a radiative transfer model [9], and was further developed in subsequent works [10, 11, 12, 13, 14]. The inversion algorithm assumes that aerosol particles are partitioned into spherical and non-spherical components, and the percentage of spherical particles is denoted by the asymmetry parameter ($g(\lambda)$). In addition to sphericity, the algorithm also retrieves the volume concentration (C_V), volume radius (r_V), and effective radius (r_{eff}), along with their corresponding standard deviations (σ). The volume particle size distribution ($dV(r)/d\ln r$) is also retrieved for 22 logarithmically equidistant points (r_i) in the range $0.05\mu m \leq r \leq 15\mu m$. The single scattering albedo ($\omega(\lambda)$) retrieval, which assumes that a sunbeam is only reflected off a single particle, is the ratio of the scattering efficiency to the extinction efficiency. The real ($n(\lambda)$), imaginary ($k(\lambda)$ refractive indices, and the single scattering albedo describe the scattering and absorbing properties of aerosols. It should be noted that retrievals of the complex refractive index ($n + ik$) require $AOT_{440} \geq 0.4$.

3.1.3 Aerosol Reference Clusters

The reference clusters used in this study are based on previous research [12, 15, 16, 17]. There are a total of 6 reference clusters, each of which represents a distinct class of aerosol as shown in Table 1.

3.2 Classification

3.2.1 Data Preparation

Since the goal is to classify aerosol types, it was necessary to consider only features that do not depend on quantity. Therefore, features such as AOT were excluded. The selected features included various optical properties of the aerosols at different wavelengths, such as the single scattering albedo, refractive index, asymmetry factor, and Angstrom exponent.

There are many null values for inversion products in the data due to the requirement that $AOT_{440} > 0.4$ (as mentioned earlier). Additionally, cloudy and rainy weather can make it difficult for the instrument to make accurate measurements, leading to further missing values. To address this issue, the IterativeImputer from the scikit-learn library was used with a Bayesian Ridge estimator to impute the missing values. This method uses an iterative approach to estimate the missing values by fitting a model to the observed data and then using the model to predict the missing values. The imputed values are then used to fit the model in the next iteration, until the imputed values converge. This can be an effective way to fill in missing values in the data while maintaining the underlying relationships between the features [18].

Finally, feature selection was done by removing features with high multicollinearity using the variance inflation factor (VIF). The VIF is a measure of how much the variance of a given model coefficient is inflated due to collinearity with other variables [19]. The built-in VIF method, called variance_inflation_factor, from the statsmodels library was used to identify and remove highly correlated features. The process of removing high-VIF features was done iteratively, by first computing the VIFs for all variables, then removing the feature with the highest VIF, and recomputing the VIFs until all VIFs were less than 10.

The resulting preprocessed dataset was then split into training and validation sets, with the Manila Observatory data reserved as the test set. It is important to note that the IterativeImputer and VIF computation were both trained on and applied to the training set first. After these steps were completed on the training set, the trained IterativeImputer and the selected features from VIF were then applied to the test and validation sets as well. The resulting processed datasets ended up with 12 features. After preparing the data, three classifier models were trained, and the best performing model will later be used for the test set. These models are the Mahalanobis Classifier, k-Nearest Neighbors, and Naive Bayes, which are discussed in more detail in the following sections.

3.2.2 Mahalanobis Classifier

The Mahalanobis classifier (MC) is a particular kind of clustering method that uses the Mahalanobis distance to measure the distance of a point to each cluster. The Mahalanobis distance is defined as:

$$d_M(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)} \quad (4)$$

where x is the point being considered, and μ and Σ are the mean and covariance matrix of the cluster. A test sample is

Class Name	Abbreviation	Site	Period
Mineral Dust	MD	Solar Village, Saudi Arabia	Mar-Jul (1999-2015)
Polluted Dust	PD	Beijing, China	Whole Year (2001-2013)
Biomass Burning, Dark Smoke	BB-D	Mongu, Nigeria	Aug-Nov (1995-2009)
Biomass Burning, White Smoke	BB-W	Alta Floresta, Brazil	Aug-Oct (1995-2013)
Urban/Industrial (Developed Economy)	UI	GSFC, Maryland, USA	Jun-Sept (1993-2013)
Urban/Industrial (Developing Economy)	UI-D	Chen Kung Univ., Tainan, Taiwan	Whole Year (2002-2014)

Table 1: Aerosol classes used as reference clusters in this study and the AERONET site they were taken from.

then classified into the cluster from which it has the smallest Mahalanobis distance. The included mahalanobis.py script contains the implementation used in this project.

3.2.3 *k-Nearest Neighbors*

The k-nearest neighbor classifier (KNN) works by finding the k points in the training set that are closest (using some specified distance metric) to the point being classified and then assigning the point to the majority class of those k points. In this project, the KNeighborsClassifier from the scikit-learn library was used [18].

In addition, to obtain an improved KNN model, the number of neighbors was subject to hyperparameter tuning along with grid search in order to maximize model performance. The built-in GridSearchCV from the scikit-learn library was used to evaluate through several choices of number of neighbors, which resulted in improved overall model performance.

3.2.4 *Naive Bayes Classification*

Finally, the Naive Bayes classifier (NB) is a simple classifier based on Bayes theorem:

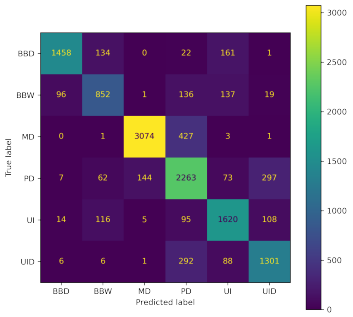
$$P(y|x_1, x_2, ..., x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{\prod_{i=1}^{n} P(x_i)} \qquad (5)$$

where $x_1, x_2, ..., x_n$ are the features of some sample point x and y is the class label. It is called "naive" because of its assumption that features are independent from one another. In this project, the GaussianNB implementation from the scikit-learn libary was used [18]. And similar to the KNN model, GridSearchCV was also used to fine tune a hyperparameter, namely the variance smoothing factor used to ensure calculation stability.

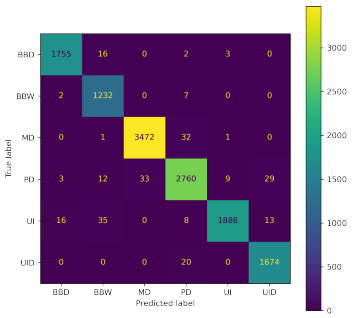
4. EXPERIMENTAL RESULTS

Table 2 shows the performance of the three classifiers on the validation set. It can be seen that the k-nearest neighbor classifier (KNN) performed the best with an accuracy 98.1%, while the naive Bayes classifier (NB), having a 72.8% accuracy was the worst model.

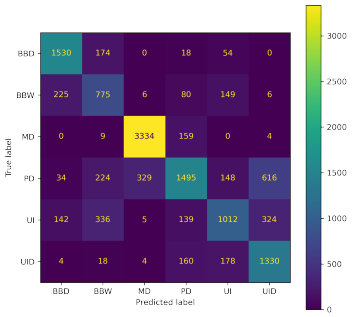
The confusion matrices in Figure (1) show the performance breakdown of each of the three models by how accurately they were able to classify the different types of aerosols. The x-axis represents the predicted labels while the y-axis represents the true label. Hence, a high off-diagonal value indicates that the model tends to "confuse" together certain classes of aerosols. For example, NB (fig. 1c) predicted 616 instances of PD as UID. In line with the results in Table 2, the best performer, KNN, has the fewest off-diagonal counts, while the worst model, NB, has the most.



(a) Mahalanobis Classifier (MC)



(b) k-Nearest Neighbors (KNN)



(c) Naive Bayes (NB)

Figure 1: Confusion matrices for the three classifier models.

	Accuracy	Precision (Macro)	Precision (Weighted)	Recall (Macro)	Recall (Weighted)	F1 Score (Macro)	F1 Score (Weighted)
MC	0.812	0.806	0.820	0.796	0.812	0.799	0.814
KNN	0.981	0.979	0.982	0.982	0.981	0.980	0.981
NB	0.727	0.695	0.734	0.711	0.728	0.695	0.723

Table 2: A comparison of the performance metrics for the three classifier models.

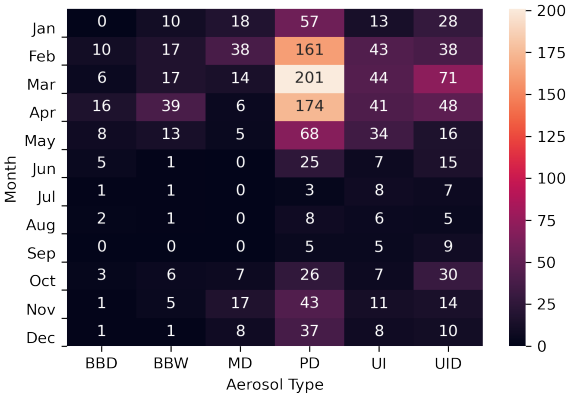


Figure 2: Heatmap of predicted aerosol types in the Manila Observatory data using KNN.

Aerosol Type	Pre-COVID-19 (%)	COVID-19 (%)
BBD	3.378	0.000
BBW	6.947	5.882
MD	7.075	5.882
PD	50.988	23.529
UI	13.384	50.000
UID	18.228	14.706

Table 3: Aerosol Types Before and During COVID-19.

The KNN model was used to predict the aerosol types in the test set (i.e. data from Manila Observatory) since it was the best-performing model out of the three. The predicted aerosol types are displayed as a heatmap in Figure 2. The heatmap shows the counts of the predicted aerosol types observed in each month, with the x-axis representing the actual count of aerosol types and the y-axis representing the months. Finally, we also present table 3 which compares the aerosol composition before and during the COVID-19 pandemic.

5. ANALYSIS AND DISCUSSION OF RESULTS

As shown earlier, the k-nearest neighbor (KNN) classifier was found to be the best performing model out of the three models tested in this study. This is likely due to the fact that KNN is a non-parametric method, meaning it does not make assumptions about the underlying data distribution. In comparison, both the Mahalanobis classifier (MC) and Naive Bayes (NB) are parametric models, which may have led to poorer performance. The KNN model was able to accurately classify the different types of aerosols present at the Manila Observatory, as demonstrated by its high overall accuracy and low off-diagonal counts in the confusion matrix (Figure 1). In contrast, the MC and NB models had lower overall accuracy and higher off-diagonal counts, indicating a greater tendency to confuse different aerosol classes. Overall, the results of this study suggest that KNN is a promising model for classifying aerosols based on their optical properties.

The results of the pivot table, shown in Figure 2, reveal the monthly distribution of aerosol types as measured by the AERONET instrument at the Manila Observatory. It can be seen that the majority of aerosols present are of the PD, UI, and UID types, with relatively fewer occurrences of BBD, BBW, and MD. These results are consistent with the fact that the site is located in an urban area since PD, UI, and UID are typically associated with vehicular exhaust or industrial activities. Interestingly, the distribution of aerosol types exhibits some seasonal variation, with the highest counts of PD, UI, and UID aerosols observed during the dry season (around January to May). However, it must be noted that the low observation counts during the wet season is due to the instrument not being able to collect data when it is raining or when it is too cloudy.

Finally, table 3 shows the percentage of aerosol types before and during COVID-19. It can be seen that, during the COVID-19 period, there was a decrease in the percentage of PD aerosols and an increase in the percentage of UI aerosols. This shift in trend may be due to the changes in human activity during COVID-19. In terms of optical properties, PD aerosols are typically coarser and less reflective compared to UI aerosols. Thus, the decrease in the proportion of PD may indicate a decrease in sources of course aerosols such as construction or transportation during the pandemic. On the other hand, the increase in UI aerosols may suggest some increase in sources of fine aerosol such as fires or industrial activities (it must also be noted that fine aerosols can be transported over longer distances compared to coarse aerosols). Further research is needed to confirm these hypotheses.

6. CONCLUSION

In this paper, we have demonstrated the predictive capabilities of classifying an aerosol's type through data from its optical properties. Through NASA's AERONET network, data on an aerosol's optical properties, such as its refractive index and single scattering albedo, were taken and matched with labels of aerosol types generated from geographic domain knowledge. This study limits its scope of study to data taken from the Manila Observatory in the Philippines.

Three supervised learning algorithms were then employed to classify six (6) aerosol types: the Mahalanobis (MC), k-nearest neighbors (KNN), and Naive Bayes (NB) classifiers. With an accuracy of 98.1%, the KNN classifier was shown to perform best in classifying different aerosol types, and is consistent with six other performance metrics. The

Mahalanobis classifier performed best after KNN (accuracy: 81.2%), and NB yielded the worst performance scores among the three (accuracy: 72.7%). With this, the KNN classifier was chosen as the best classifier over MC and NB, and was finally used to predict aerosol types of test data taken from the Manila Observatory. The aerosol types predicted were used to create aerosol distributions around the Manila Observatory: PD, UI, and UID aerosols were found to be the most abundant. A comparison between aerosol compositions pre-COVID 19 and during COVID-19 was also conducted. It was shown that PD levels were significantly reduced during the onset of the COVID-19 period.

For future work, alternative techniques on multiclass classification may be explored, such as tree-based algorithms and artificial neural networks. Additionally, to utilize the full capability of AERONET's vast range of data, the scope of the study may be extended to other regions of the world.

7. REFERENCES

[1] S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. Averyt, M. Tingor, and H. Miller, *Climate Change 2007: The Physical Science Basis*. Cambridge University Press, 2007.

[2] M. Parry, O. Canziani, J. Palutikof, P. van der Linden, and C. Hanson, *Impacts, Adaptation, and Vulnerability: Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, 2007.

[3] J. S. Reid, E. J. Hyer, R. S. Johnson, B. N. Holben, R. J. Yokelson, J. Zhang, J. R. Campbell, S. A. Christopher, L. Di Girolamo, L. Giglio, R. E. Holz, C. Kearney, J. Miettinen, and E. A. Reid, "Observing and understanding the southeast asian aerosol system by remote sensing: An initial review and analysis for the seven southeast asian studies (7seas) program," *Atmospheric Research*, vol. 122, pp. 403–468, 2013.

[4] B. Holben, T. Eck, I. Slutsker, D. Tanre, J. Buis, A. Setzer, E. Vermote, J. Reagan, Y. Kaufman, T. Nakajima, F. Lavenu, I. Jankowiak, and A. Smirnov, "Aeronet - a federated instrument network and data archive for aerosol characterization," *Remote Sensing of Environment*, vol. 66, pp. 1–16, 1998.

[5] A. Smirnov, B. Holben, T. Eck, O. Dubovik, and I. Slutsker, "Cloud-screening and quality control algorithms for the aeronet database," *Remote Sensing of Environment*, no. 73, pp. 337–349, 2000.

[6] B. N. Holben, T. F. Eck, I. Slutsker, A. Smirnov, A. Sinyuk, J. Schafer, D. Giles, and O. Dubovik, "Aeronet's version 2.0 quality assurance criteria," *Remote Sensing of the Atmosphere and Clouds*, 2006.

[7] A. Ångström, "The parameters of atmospheric turbidity," *Tellus A*, no. 16, pp. 64–75, 1964.

[8] T. Eck, B. Holben, J. Reid, O. Dubovik, A. Smirnov, N. O'Neill, I. Slutsker, and S. Kinne, "Wavelength dependence of the optical depth of biomass burning, urban and desert dust aerosols," *Journal of Geophysical Research*, no. 104, pp. 31 333–31 350, 1999.

[9] O. Dubovik and M. King, "A flexible inversion algorithm for retrieval of aerosol," *Journal of Geophysical Research*, vol. 20, no. 105, pp. 673–696, 2000.

[10] O. Dubovik, A. Smirnov, B. Holben, M. King, Y. Kaufman, T. Eck, and I. Slutsker, "Accuracy assessment of aerosol optical properties retrieval from aeronet sun and sky radiance measurements," *Journal of Geophysical Research*, no. 105, pp. 9791–9806, 2000.

[11] O. Dubovik, B. Holben, T. Lapyonok, A. Sinyuk, M. Mishchenko, P. Yang, and I. Slutsker, "Non-spherical aerosol retrieval method employing light scattering by spheroids," *Journal of Geophysical Research*, no. 107, p. 4739, 2002.

[12] O. Dubovik, B. Holben, T. Eck, A. Smirnov, Y. Kaufman, M. King, D. Tanr'e, and I. Slutsker, "Variability of absorption and optical properties of key aerosol types observed in worldwide locations," *Journal of Atmospheric Science*, vol. 59, pp. 590–608, 2002b.

[13] O. Dubovik, A. Sinyuk, T. Lapyonok, B. Holben, M. Mishchenko, P. Yang, T. Eck, H. Volten, O. Munoz, B. Veihelmann *et al.*, "Application of light scattering by spheroids for accounting for particle nonsphericity in remote sensing of desert dust," *Journal of Geophysical Research*, vol. 111, 2006.

[14] A. Sinyuk, O. Dubovik, B. Holben, T. F. Eck, F.-M. Breon, J. Martonchik, R. Kahn, D. J. Diner, E. F. Vermote, J.-C. Roger, T. Lapyonok, and I. Slutsker, "Simultaneous retrieval of aerosol and surface properties from a combination of AERONET and satellite," *Remote Sensing of the Environment*, vol. 107, 2007.

[15] C. Cattrall, J. Reagan, K. Thome, and O. Dubovik, "Variability of aerosol and spectral lidar and backscatter and extinction ratios of key aerosol types derived from selected aerosol robotic network locations," *Journal of Geophysical Research*, vol. 110, 2005.

[16] D. Giles, B. Holben, S. Tripathi, T. Eck, W. Newcomb, I. Slutsker, R. Dickerson, A. Thompson, S. Mattoo, S. Wang, R. Singh, A. Sinyuk, and J. Scafer, "Aerosol properties over the indo-gangetic plain: mesoscale perspective from the tigerz experiment," *Journal of Geophysical Research*, vol. 116, 2011.

[17] P. B. Russell, M. Kacenelenbogen, J. M. Livingston, and O. P. Hasekamp, "A multiparameter aerosol classification method and its application to retrievals from spaceborne polarimetry," *Journal of Geophysical Research: Atmospheres*, vol. 119, no. 19, pp. 11,171–11,183, 2014.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer,

R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," 2011–. [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html

[19] D. A. Belsley, E. Kuh, and R. E. Welsch, *Regression diagnostics: Identifying influential data and sources of collinearity.* New York: Wiley, 1980.