

A machine learning approach for aerosol classification using sun photometer and lidar data

Joelle Buxmann^(a), Martin Osborne^(a), Mike Protts^(a), Panuganti C.S. Devara^(b), Pawan Gupta^(c),
Elena S. Lind^(c), Govindan Pandithurai^(d), Debbie O'Sullivan^(a),

^(a)*the Met Office, Exeter, UK*

^(b)*Amity Centre for Ocean-Atmospheric Science and Technology, Amity University Haryana, India*

^(c)*Biospheric Sciences Laboratory, NASA Goddard Space Flight Center, Greenbelt, Maryland, USA*

^(d)*Indian Institute of Tropical. Meteorology, Pune, India*

Lead Author e-mail address: joelle.c.buxmann@metoffice.gov.uk

Abstract: The Met Office operates a ground based operational network of nine polarisation Raman lidars (aerosol profiling instruments) and sun photometers (column integrated information) across the United Kingdom (UK). An aerosol classification scheme using supervised machine learning has been developed. The concept of Mahalanobis (~normalized) distance to identify the aerosol type from individual Aerosol Robotic Network (AERONET) measurements including Extinction Angstrom Exponent, Absorption Angstrom Exponent, Single Scattering Albedo and Index of refraction is used for a subset of AERONET stations around the globe of known main aerosol types (training set). The aerosol types so far include marine, urban industrial, biomass burning and dust. The relation of particle linear depolarisation ratio (PLDR) and lidar ratio (LR) from the Raman lidar is used in synergy to validate the particle type.

1. Introduction

Aerosols play a significant role in Earth's climate system and air quality. The composition, size, and distribution of aerosols influence atmospheric processes, such as radiative forcing and aerosol cloud interaction [1]. Additionally, aerosols negatively impact human health in form of air pollution [2]. Thus, accurate aerosol classification and characterization are essential for understanding their impacts on climate, health, and ecosystems. The Aerosol Robotic Network (AERONET) is a globally distributed network of ground-based sun photometers, designed to measure aerosol optical properties with high precision and temporal resolution [3]. AERONET aerosol physical and optical properties have been used for aerosol classification before [e.g. 4-7]. Additionally, lidar observations show well established aerosol classification schemes based on lidar ratio (LR) and particle linear depolarisation ratio (PLDR) [8,9]. Recent advancements in machine learning techniques offer promising avenues automating aerosol classification processes and can be used for both lidar [10] and sun photometer data [11]. This paper presents an integrated approach for aerosol

classification utilizing AERONET sun photometer data and the K nearest neighbor (KNN) machine learning (ML) algorithm. By harnessing the rich dataset provided by AERONET, including an approach to utilize cloud screened level 1.5 data, we aim to enhance the efficiency and timeliness of aerosol classification, making it potentially available for nowcasting and/or model assimilation.

2. Methodology

Our aerosol clusters are based on geographical location [6] with predominant aerosol types rather than by composition or clustering analysis [4]. To build a training data set we assign a specific type of aerosol with a given location (like in [6,12]). Locations with predominantly urban industrial aerosols include major cities like Hamburg, Moscow Stennis, and GSFC (Goddard Space Flight Center NASA in Greenbelt, Maryland), Aerosols from biomass burning are expected at sites in the Amazon and Southern Africa: Mongu, Cuiaba, Etosha Pan, Alta Floresta. Dust is prevalent in Saudi Arabia, e.g. Solar Village, Dakar, Djou Gou, and Blida. Marine aerosols are found on islands and coastal locations, e.g. Lanai. For the mixed aerosol definition,

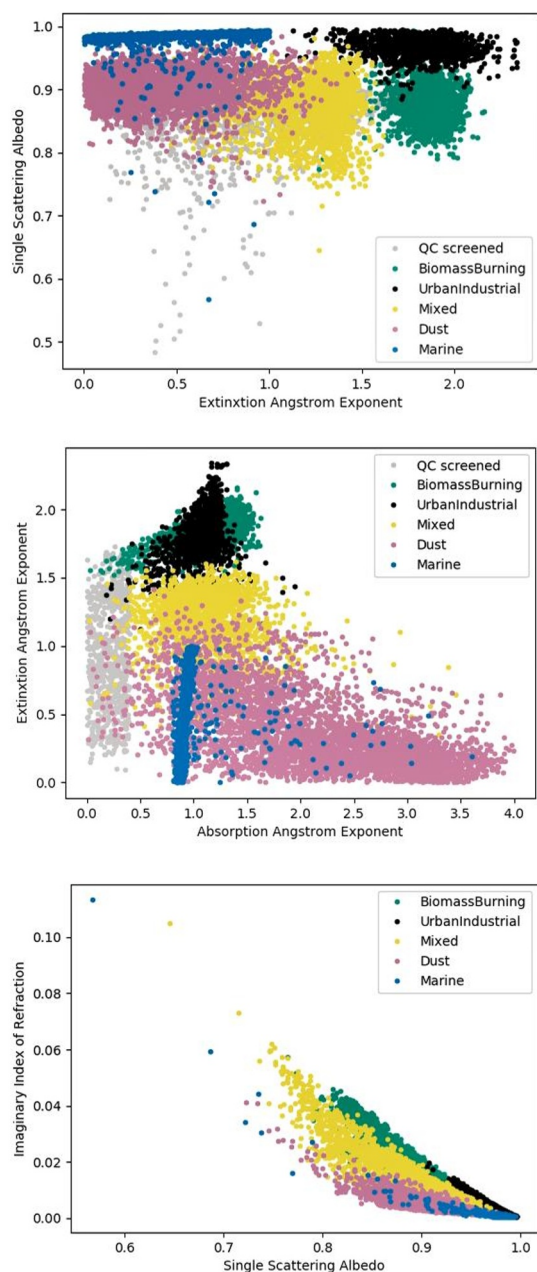


Figure 1: Examples of the aerosol clusters in 2-dimensional representation.

including but not limited to Asian cities (e.g. Pune, Osaka, Pokhara, Beijing), we follow the nomenclature of [13]. The KNN technique is then applied using the Mahalanobis distance [14], as it is independent of dimension and there is no limit to the number of variables that can be used to evaluate it [6]. We build a predictive model from AERONET data using a training data set according to the geographical aerosol classification mentioned above. As basis for our classification we use the relationship of Imaginary and Real Refractive Index, Absorption Angstrom Exponents (AAE), Single Scattering Albedo (SSA), and

Extinction Angstrom Exponent(EAE) based on [6]. Examples of the clusters are shown in figure 1. The aim is to build a real time aerosol classification application that can be used in Nowcasting. As level 2.0 data is usually only available after 1-2 years (after a new calibration has been performed), it is important to analyze the properties as compared to more readily available level 1.5 (cloud screened) data. We found that the difference between level 1.5 and level 2.0 AERONET data was significant especially in SSA and AAE. Therefore, we defined an additional cluster, referred to here as ‘QC screened’ from level 1.5 data with $SSA < 0.9$ and $AAE < 0.4$. As seen in figure 1 those properties were not/rarely found in the chosen reference clusters level 2.0 data. This is not a replacement of AERONET level 2.0 screening but used as practical parametrisation. An additional criterion was included for marine aerosols using level 1.5 data, filtered $EAE < 1$, as suggested in [15].

In a sensitivity study we split the training data set into 30% test data and 70% training data. We found that including 5 sites (with 5 different aerosol types) and $K=13$ showed the highest accuracy of 92% (figure 2), where accuracy is defined as fraction of correctly predicted aerosol type according to the initial assignment in the test dataset.

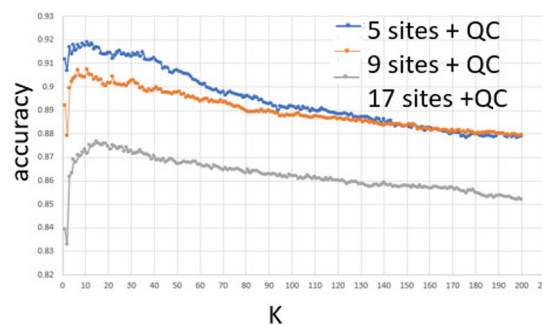


Figure 2: Sensitivity study including various numbers of aerosol type specific sites and Quality Control (QC) screened cluster.

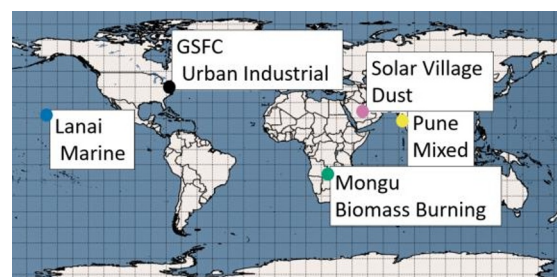


Figure 3: Locations with specific dominant aerosol types chosen for the ML algorithm.

Therefore, the core of our model includes five AERONET sites with predominant aerosol types (figure 3): mixed aerosol (Pune), biomass burning (Mongu), dust (Solar Village), urban industrial (GSFC), and marine (Lanai).

3. Results

Here, we present more detailed data from October 2017 ([16] hurricane Ophelia). We selected two examples where both lidar and sun photometer data is available.

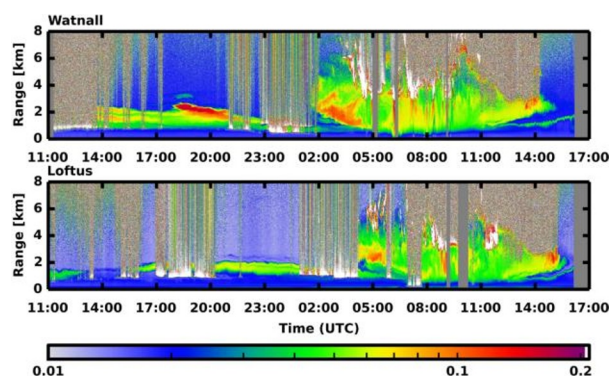


Figure 4: Lidar volume linear depolarisation ratios for 15 and 16 October 2017 [16].

Table 1. Example Results of lidar (L) and sun photometer(SP) ML aerosol classification from two locations in the UK: Watnall (WA) and Loftus (LO)(Ma=marine, Mi=mixed, D=dust)

Place Date Time	Layer height (km)	PL DR %	LR (sr)	Aerosol type	
				L	SP ML
LO 15/10/ 2017	0.3 to 1.1 km	2	27	Ma	Mi
3-3:30 pm	1.4 to 2.45 km	26	27	D	
WA 16/10/ 2017	0.3 to 1 km	2	19.5	Ma	Ma
2:30- 3pm	1 to 1.5 km	6	19.5	Ma	
	1.95 to 2.9 km	16	19.5	D	

The first location in Loftus on 15th Oct 2017 shows an equally balanced contribution of marine aerosols (PLDR=2%) and dust

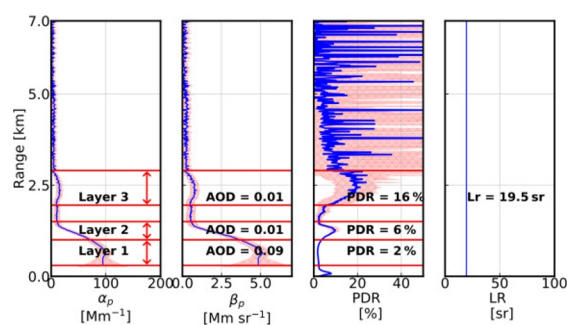


Figure 5: Optical properties and mass concentration estimates calculated from averaged profiled from Watnall on 16 October from 2:30 pm to 3:00 pm UTC.

(PLDR=26 %) in two layers with an overall LR of 27 sr (using Klett-Fernald [17]). This is interpreted as mixed aerosol by the ML algorithm from the column integrated sun photometer data (Table1). Based on the data from 16th October 2017 (Figure 4 and 5) from the sun photometer in Watnall (Nottingham, UK), the KNN retrieval predicted the aerosol to be of marine origine (table 1). As shown in figure 5 the lidar ratio (LR) is set to 19.5 sr throughout the whole profile using the Fernald–Klett method [17]. The PLDRs are 2 % and 6 % between 0.3-1 km and 1-15 km, respectively. However, some aerosol is present at 1.95-2.9km with a PLDR of 16%. These values suggest the lower layers are of marine origine with some dust above 1.95 km [8,9]. The findings of the lidar are in good agreement with the ML prediction of marine aerosol. The sun photometer is a column integrated measurement, as most of the profile shows as marine aerosol from the lidar measurements with only a small dust layer in comparison.

4. Summary and Discussion

The KNN technique has been applied to build an aerosol Classification Scheme using AERONET data. Preliminary results show high accuracy (92%) on a test/train data set including level 1.5 data, which is suitable for Nowcasting. A few case studies so far show good agreement with lidar prediction of aerosol types. One issue of the column integrated sun photometer data is that mixtures of aerosols cannot be distinguished easily and may be overestimated. While there are certainly some shortcomings in our approach as discussed, we are confident that generally our algorithm leads to plausible aerosol identifications.

5. Outlook

We plan to increase the test and train data set to improve the accuracy of our model. Further we will add a volcanic ash cluster to the scheme based on volcanic eruptions over the last 8-10 years using global data. We will add more validation utilizing more case studies of lidar and sun photometer synergetic measurements as well as satellite data comparison.

Acknowledgements

We are grateful to the Civil Aviation Authority and Department for Transport for funding the lidar project. This abstract is part of a project that is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-INFRADEV-2019-2, Grant Agreement number: 871115. We thank the network managers, PI's, CoPI's and engineers for their hard work managing and maintaining the sites!

References

- [1] Masson-Delmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (eds.)], *IPCC, 2021: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2391 pp, 2021.
- [2] [https://www.who.int/news-room/fact-sheets/detail/ambient-\(outdoor\)-air-quality-and-health](https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health), last access March 2024.
- [3] B.N. Holben, T.F. Eck, I. Slutsker, A. Smirnov, A. Sinyuk, J. Schafer, D. Giles, O. Dubovik, *AERONET's version 2.0 quality assurance criteria, Remote Sensing of the Atmosphere and Clouds*, Proc. SPIE, 6408, p. 64080Q, 2006.
- [4] A.H. Omar, J.-G. Won, D.M. Winker, S.-C. Yoon, O. Dubovick, M.P. McCormick, *Development of global aerosol models using cluster analysis of Aeronet Robotic Network (AERONET) measurements*, J. Geophys. Res., 110, 2005.
- [5] C. Toledano, M. Wiegner, S. Gross, V. Freudenthaler, J. Gasteiger, D. Muller, T. Muller, A. Schladitz, B. Weinzierl, B. Torres, N.T. O'Neill, *Optical properties of aerosol mixtures derived from sun-sky radiometry during SAMUM-2*, Tellus, Ser. B, 63, pp. 635-648, 2011.
- [6] Patrick Hamill, Marco Giordano, Carolyne Ward, David Giles, Brent Holben, *An AERONET-based aerosol classification using the Mahalanobis distance*, *Atmospheric Environment*, Volume 140, Pages 213-233, ISSN 1352-2310, 2016.
- [7] Vijayakumar, K., P.C.S. Devara, S.M. Sowbane, D.M. Giles, B.N. Holben, S.V.B. Rao and C.J. Shankar, *Solar radiometer sensing of multi-year aerosol features over a tropical urban station: direct-Sun and inversion products*, Atmos. Meas. Tech., 13, 5569–5593, 2020.
- [8] Groß, S., Freudenthaler, V., Wirth, M., and Weinzierl, B., *Towards an aerosol classification scheme for future EarthCARE lidar observations and implications for research needs*, Atmos. Sci. Lett., 16, 77–82, <https://doi.org/10.1002/asl2.524>, 2015.
- [9] Mona, L., Liu, Z., Müller, D., Omar, A., Papayannis, A., Pappalardo, G., Sugimoto, N., and Vaughan, M., *Lidar Measurements for Desert Dust Characterization: An Overview*, Adv. Meteorol., 2012, 356265, 2012.
- [10] Nicolae, D., Vasilescu, J., Talianu, C., Binietoglou, I., Nicolae, V., Andrei, S., and Antonescu, B., *A neural network aerosol-typing algorithm based on lidar data*, Atmos. Chem. Phys., 18, 14511–14537, 2018.
- [11] Mohan, A. S., Manisekaran, A., & Kumar, L. S., *Aerosol classification using machine learning algorithms. Indian Journal of Radio & Space Physics*, 50(4), 217–223, 2021.
- [12] C. Cattrall, J. Reagan, K. Thome, O. Dubovik, *Variability of aerosol and spectral lidar and backscatter and extinction ratios of key aerosol types derived from selected aerosol robotic network locations*, J. Geophys. Res., 110, 2005.
- [13] D.M. Giles, B.N. Holben, T.F. Eck, A. Sinyuk, A. Smirnov, I. Slutsker, R.R. Dickenson, A.M. Thompson, J.S. Schafer *An analysis of AERONET aerosol absorption properties and classifications representative of aerosol source regions* J. Geophys. Res., 117, 2012.
- [14] P.C. Mahalanobis, *On the generalized distance in statistics*, Proc. Natl. Inst. Sci. India, 2, pp. 49-55, 1936.
- [15] A.M. Sayer, A. Smirnov, N.C. Hsu, B.N. Holben, *A pure marine aerosol model for use in remote sensing applications*, J. Geophys. Res., 117, 2012.
- [16] Osborne, M., Malavelle, F. F., Adam, M., Buxmann, J., Sugier, J., Marengo, F., and Haywood, J.: *Saharan dust and biomass burning aerosols during ex-hurricane Ophelia: observations from the new UK lidar and sun-photometer network*, Atmos. Chem. Phys., 19, 3557–3578, 2019.
- [17] Johannes Speidel and Hannes Vogelmann, *Correct(ed) Klett–Fernald algorithm for elastic aerosol backscatter retrievals: a sensitivity analysis*, Appl. Opt. 62, 861-868, 2023.