23rd International Conference on
Harmonisation within Atmospheric Dispersion Modelling
for Regulatory Purposes
15-19 September 2025, Hamburg, Germany

EXTENDED ABSTRACT

Machine learning-based aerosol classification and source apportionment using ground photometry data

Vikija Kupca, Faculty of Science and Technology, University of Latvia

kupcavikija@gmail.com

Iveta Steinberga, Faculty of Science and Technology, University of Latvia

Introduction

Aerosol classification is crucial for understanding and linking together atmospheric processes such as climate regulation and the interconnection with air quality and health effects. More knowledge of the complexity and the significance of aerosols and their sources is essential for tackling global climate and air pollution challenges. A key component of this understanding is source apportionment, the process of identifying and quantifying the origin of aerosols. This task remains a challenge due to the complexity of overlapping emission sources and their different chemical fingerprints, and the influence of varying atmospheric dynamics. Accurate source identification is crucial for effective mitigation strategies and environmental policies.

Aerosol monitoring, however, has its own limitations. Satellite observations, while sufficient in spatial coverage, have limited spatial resolution and are less reliable under cloudy conditions. Ground-based monitoring stations, such as sun photometers, offer more precise measurements but do not have spatial coverage. Data gaps further the challenge, for example, the single scattering albedo (SSA), a characteristic that determines the radiative effect, is found to be frequently missing. Such data gaps hinder effective classification methods and correct assessment of air quality.

To address this challenge, we employ machine learning algorithms to investigate whether relying primarily on photometry data can accurately classify aerosols in remote settings. This approach is contrasted with urban environments, where more comprehensive meteorological and trajectory datasets are often required. The models will be validated using a combination of chemical analysis and HYSPLIT back-trajectory models. This work not only addresses current limitations in aerosol monitoring but also contributes to standardized classification methodologies that could be extended to other remote sites.

This study aims to answer the following questions:

1. Can machine learning models effectively classify aerosol types and sources at a remote marine site using a combination of ground-based photometry, chemical, and trajectory data?
2. Are simplified classification algorithms sufficient for this task, or is a full suite of meteorological and trajectory data necessary?

Approach

The study is based on data collected at the AERONET station Lampedusa, Italy (35.5°N, 12.6°E) (figure 1), a baseline site located on the island of Lampedusa in the central Mediterranean Sea. This location is strategic for monitoring atmospheric composition, as it is frequently influenced by air masses originating from North Africa, Europe, and the surrounding marine environment. The data analyzed in this study covers the daily average measurements dated from January 1st, 2014, to December 31st, 2020.

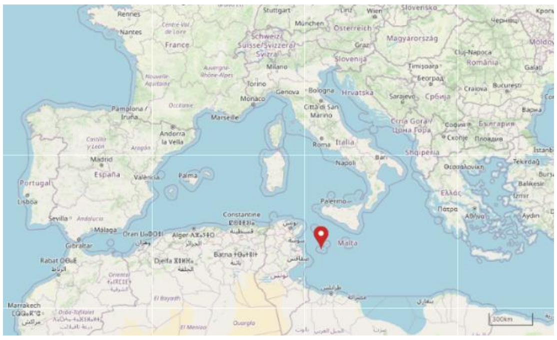


Figure 1. Location of the Lampedusa study site. Basemap data © OpenStreetMap contributors

Three primary types of data were used in this study:

1. AERONET optical properties: the core inputs are the Level 2.0 cloud-screened and quality assured optical properties from the AERONET sun photometer. The key parameters include Aerosol optical depth (AOD) at multiple wavelengths (440, 500, 675, 870, 1020 nm), Angstrom Exponent (AE) calculated over various wavelength pairs (440-870 nm), which serves as an indicator of particle sizes and Fine mode Fraction (FMF) at 500 nm, indicating the contribution of fine particles to the total AOD.
2. Aerosol type classifications: already existing classification methods (threshold methods) are derived from Annapurna et al. (2023), Stefan et al. (2020), Ozdemir et al. (2020). The primary aerosol classes to be identified are marine, desert dust, urban/industrial, and mixed.
3. Chemical data for validation: the dataset contains chemical speciation data, including concentrations of ions (e.g. Na, Cl, SO_4) and elements (e.g. Al, Fe, Si). This data is used as an independent dataset used for validation, for instance, high concentrations of Na and Cl are strong indicators of marine aerosols, elevated levels of elements like Al, Si and Fe will be used to validate mineral dust aerosols, high SO_4 and NO_3 concentrations will help to confirm anthropogenic pollution events.

Before model training, the raw data was processed by data cleaning, days with missing values for the core characteristics (AOD and AE) were removed from the dataset to ensure a complete set for the model.

Several classification algorithms are tested, with a primary focus on the random forest model, that is chosen for its robustness and high performance in other studies. Other models, such as gradient boosting, may also be evaluated to ensure a comprehensive comparison.

The model's ability to accurately classify aerosol types is assessed using standard performance metrics like accuracy, precision, recall and F1-score. A confusion matrix is generated to provide a detailed breakdown of the model's performance for each aerosol class. This helps to identify which aerosol type is easily distinguishable and where it faces difficulty.

Finally, the classification from the best performing model will be independently validated against the chemical profiles and a HYSPLIT back trajectory model to confirm the source origins. This multi-level validation approach ensures the robustness and reliability of our findings.

Results and discussion

A preliminary analysis was conducted on the 2014-2020 dataset to establish the viability of using machine learning for aerosol classification at Lampedusa. The frequency distribution of aerosol types, based on one of the used classification methods, is shown in Figure 2. Marine aerosols, as expected, make up the largest fraction of 44%, followed by 32% mixed aerosols, this highlights the complex atmospheric classification challenge. Dust aerosols make up 20% due to the proximity of North Africa and episodes of Urban/Industrial pollution constitute 4%.

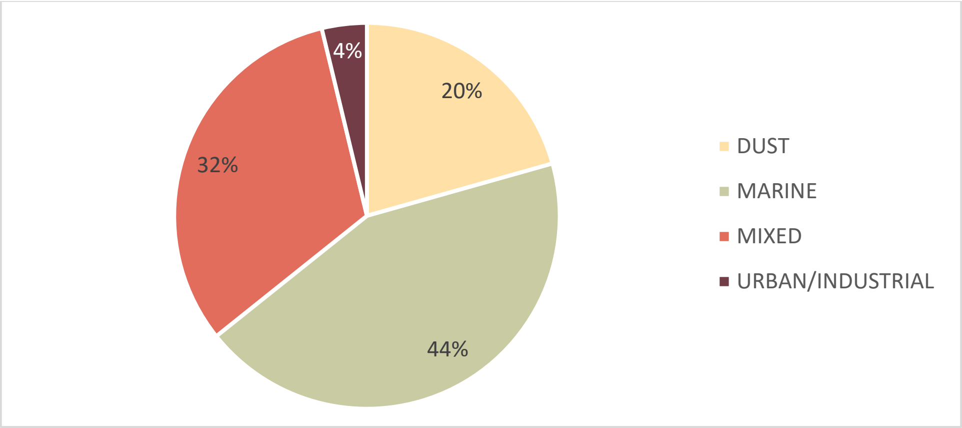


Figure 2. Classified aerosol types in Lampedusa from 2014-2020, based on the threshold method derived from Stefan et al. 2020

The fundamental premise of this study is that these different aerosol types have distinct optical signatures that can be learned by a classification algorithm. Figure 3 validates this premise by plotting AOD against AE. A clear separation between aerosol classes is visible. Coarse-mode particles, such as desert dust (low AE, high AOD) and marine aerosols (low AE, low AOD), occupy a distinct region of the plot from fine-mode particles associated with urban/industrial type or mixed type (high AE). This clear optical separability strongly suggests that supervised machine learning models can be effectively trained on this data.

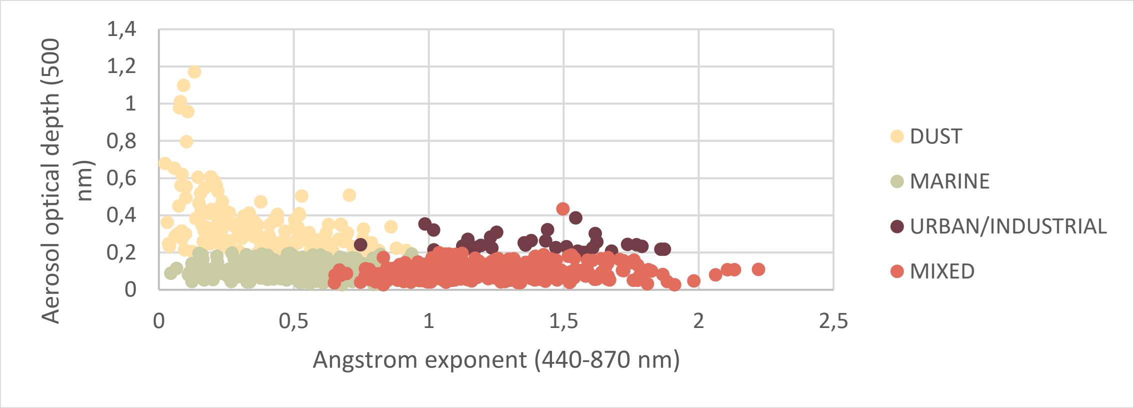
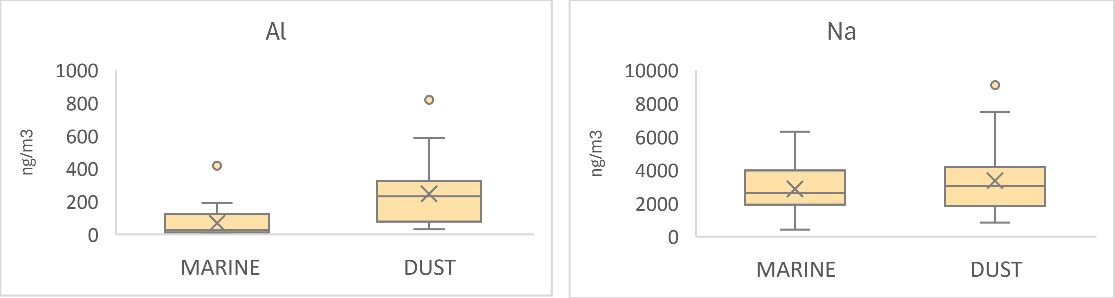


Figure 3. The optical properties of major aerosol types at Lampedusa

To confirm that these optical classifications correspond to chemically distinct aerosol compositions, we analyzed key chemical tracers. Figure 4a confirms that the chemical signature of dust. Days classified as DUST show significantly higher atmospheric concentrations of Aluminum (Al), a primary tracer for crustal material, when compared to days classified as MARINE. This result provides strong independent validation for the optical identification of mineral dust events. More interestingly, Figure 4b shows the mixed nature of aerosols in Lampedusa. While MARINE aerosol events are, as expected, rich in sodium (Na), so are the DUST events. This indicates that air masses from North Africa become heavily laden with sea salt during the transit over the Mediterranean Sea. Therefore, what is by optical properties identified as a dust event, is chemically mixed dust-marine aerosol.



(a) Al

(b) Na

Figure 4. Chemical validation of optical aerosol classifications using key tracer elements on days classified as Marine versus Dust

These preliminary results are compelling. Together they demonstrate that while the primary aerosol types have distinct optical and chemical properties, there is significant mixing between sources. This complexity underscores the limitations of threshold-based classification schemes and strongly motivates the application of more sophisticated methods like machine learning.

Conclusion and future work

Thus far we have successfully established a strong, data-driven foundation for the application of machine learning to aerosol classification at the remote marine site of Lampedusa. Through a preliminary analysis of optical and chemical data, we have demonstrated two key points.

1. Major aerosol types, particularly marine and desert dust, possess distinct and separable optical characteristics based on their AOD and AE.
2. These optical classifications are physically meaningful, as they correlate with expected chemical tracers - elevated aluminum for dust events and sodium for marine sources.

Building on this, our future work will proceed with:

1. Model development and evaluation - next step is to train and rigorously evaluate machine learning algorithms on AERONET optical data.
2. Source apportionment validation - to validate models' ability to identify aerosol origins, we will conduct back-trajectory analysis using HYSPLIT model.
3. Methodology scalability - the long-term objective is to refine this methodology so it can be applied to other AERONET sites around the world, particularly those lacking in chemical measurements. This will contribute to the development of standardized, scalable tools for improving global aerosol monitoring.

References

Annapurna, S., Anitha, M., & Kumar, L. S. (2023). Composition and source based aerosol classification using machine learning algorithms. *Advances in Space Research*, *73*(1), 474–497. https://doi.org/10.1016/j.asr.2023.09.068

OpenStreetMap contributors. (n.d.). OpenStreetMap. Retrieved August 15, 2025, from https://www.openstreetmap.org

Ozdemir, E., Tuygun, G. T., & Elbir, T. (2020). Application of aerosol classification methods based on AERONET version 3 product over eastern Mediterranean and Black Sea. Atmospheric Pollution Research, 11(12), 2226–2243. https://doi.org/10.1016/j.apr.2020.06.008

Stefan, S., Voinea, S., & Iorga, G. (2020). Study of the aerosol optical characteristics over the Romanian Black Sea Coast using AERONET data. Atmospheric Pollution Research, 11(7), 1165–1178. https://doi.org/10.1016/j.apr.2020.04.007