# Report on the outcomes of a Virtual Mobility[1]

**Action number: CA21119**

**Grantee name: Iveta Steinberga**

**Virtual Mobility Details**

Title: Machine Learning techniques applied to photometry

Start and end date: 20/06/2024 to 19/08/2024

Our research and analysis of scientific literature has revealed a significant trend. The use of machine learning algorithms in the classification of atmospheric aerosols, whether for determining potential climate impact or identifying the origin of potential aerosols, has seen a substantial increase in recent years. The variety of classification algorithms employed is extensive. However, a notable gap exists in the form of a lack of best practice recommendations for selecting the most suitable mode, depending on the nature of the data set or the objective pursued.

The classification of aerosols in publications could be divided in 2 main parts: (1) a classification based on the composition of the particles and (2) a classification to determine the origin of the primary source of aerosols. And then, particle composition groups (e.g. coarse non-absorbing, black carbon high etc.) or pollution source groups are often identified depending on the target source (urban, biomass, dessert etc.). In the form of a standardized approach, publications reveal 3-6 classifications of different particle compositions, while the number of pollution source classes in some cases reaches up to 20 classes.

In general, about 20 publications have been analysed for analysed. For data analysis measurements were used in Lampedusa (Italy), time period – 1/1/2014-4/21/2020, daily data, parameters used: AOD for different wavelengths, AE, FMF, chemical composition data (site measurements). Testing has been carried out for different classification algorithms, 4 different approaches analysed, and also mixed algorithm proposed, in order to avoid situations when classification couldn`t be performed or lot of outliers detected.

Validation of the results obtained has been performed using fixed aerosol chemistry measurements, identification of sources has been performed based on aerosol profiling results.

In addition to Lampedusa measurements, model testing has been carried out for AERONET-derived measurements in Rome (Italy), period – 1/1/2019-12/31/2022, daily data set, parameters – mainly the same.

[1] This report is submitted by the grantee to the Action MC for approval and for claiming payment of the awarded grant. The Grant Awarding Coordinator coordinates the evaluation of this report on behalf of the Action MC and instructs the GH for payment of the Grant.

Source apportionment classification was validated based on site chemical analysis (in case of Lampedusa), but in the case of Rome individual source profile data (from EU aerosol profiling database SPECIEUROPE, https://source-apportionment.jrc.ec.europa.eu/Specieurope/index.aspx) and air quality monitoring data obtained within national monitoring network were also used.

Data sets with inputs and results were prepared and after the Virtual Mobility extended report in the form of publication is planned. Preparation of publication is planned about end of the 2024 / beginning of 2025, and submission to the journal “Remote Sensing” or similar is intended.

The classification method, used for Lampedusa, based on only two parameters (AOD 550 nm; 440-870_Angstrom_Exponent), identified as DESERT aerosols in around 3% of cases, MARINE in 24%, urban in 2%, all other most of cases could not be classified. Sources such as BIOMASS, ARID, are also used, but none were found. The classification algorithm is described in Annapurna et al. (2024)[2]. The result shows that the established classification scheme is not robust and that the application options are relatively low, potentially for a specific location.

Another method that is widely used is described in Stefan et al. (2020)[3]. The technique is similar, resulting in aerosols being classified according to their origin - MARINE, DUST, MIXED, URBAN/INDUSTRIAL, BIOMASS BURNING. Also, only two parameters (AOD 550 nm; Angstrom_Exponent-Total_500nm) are used as classification criteria. The situation was analogous, but the results quite often were opposite (65% of cases). For example, if Annapurna et al. (2024), classification result was MARINE, then result according to Stefan et al. 2020 was DUST/SAND.

Besides above mentioned method described in Ozdemir et al. (2020)[4], where two parameters are used in the classification: AOD 550 nm and FineModeFraction_500nm gives only three classes: MARINE, DESERT, and CONTINENTAL. Although the established classifier works well and there are no cases where classification is not possible, the result is too homogeneous. Almost 90% of the cases correspond to the MARINE class. According to literature research, Ozdemir et al. (2020) modification analyzed as the last of the classification methods, results in 5 classes: MARINE, DESERT, CONTINENTAL, BIOMASS BURNING and MIXED, classifier criteria—AOD_870nm; Aod_440nm; 440-870_Angstrom_Exponent. As a result, it was not possible to classify aerosol sources in at least 50% of the cases, indicating deficiencies in the method.

Using ground-level aerosol chemistry measurements and SPECIEUROPE profiles, sea aerosols are strongly prevalent in Lampedusa for the whole observation period based on sodium and sulfate ion proportions, while specific proportions of sulfates, calcium, aluminum, and iron pointed to Saharan dust dominance. It allows to prepare an advanced aerosol classification algorithm was developed that considers the interdependence between photometric parameters (statistically relevant parameters have been identified) and chemical composition measurements with specific parameter limits. The algorithm developed includes the following parameters: AOD_500nm. 440-870_Angstrom_Exponent; Finemodefraction_500nm Extinction_angstrom_exponent_440-870 nm-Total; aerosol classes obtained — MARINE, VERY MARINE, DUST/DESERT, clear configurations/LOW AEROSOL, urban/INDUSTRIAL, BIOMASS, mixed.

Based on the established classification, the dataset uses a machine learning algorithm to train the classifier and use it for other datasets. The Random Forest algorithm was used in this case,

[2] https://doi.org/10.1016/j.asr.2023.09.068
[3] https://doi.org/10.1016/j.apr.2020.04.007
[4] https://doi.org/10.1016/j.apr.2020.06.008

and training was conducted in a JASP 0.18.3 environment (R-based). The learning accuracy of the resulting algorithms reaches 83.3%, a good enough indicator. A higher rate could be achieved by increasing the dataset.

In addition, in test mode, the algorithm was also used for the Roman dataset. The data set used was more complete, less data gaps, and the resulting algorithm machine learning performance rate was significantly higher. The test accuracy, an impressive 96%, showcases the algorithm's robust performance.