Indian Journal of Radio & Space Physics
Vol 50, December 2021, pp. 217-223

Aerosol classification using machine learning algorithms

Annapurna Sheela Mohan*, Anitha Manisekaran & Lakshmi Sutha Kumar
Department of ECE, NIT Puducherry, Karaikal, 609609, India

Received:12 February 2021 Accepted:5 March 2021

Aerosols are particles that are omnipresent in the atmosphere. They vary in size, shape and composition. They can be naturally occurring or might be produced artificially. However, proper classification and characterization of aerosols have been still in progress and this creates uncertainty in climatological studies. In this paper, an aerosol classification scheme has been presented based on the measurements done using a CIMEL sunphotometer in Thessaloniki, Greece from 1998 to 2017. The study has been mainly upheld by the direct measurements of Single Scattering Albedo (SSA) at 440nm and Fine Mode Fraction (FMF) at 500nm. These parameters have been used to establish testing and training datasets. Machine learning algorithms have been used to validate the classified data. Various performance metrics have been evaluated. Also, the best-fit algorithm for classifying aerosol data has been found out.

Keywords: AERONET, Fine Mode Fraction, Machine learning, Single Scattering Albedo, Sunphotometer

1 Introduction

Atmospheric debris is omnipresent within the atmosphere of the earth. These particles interact with ecosystems and solar radiation. While the debris is regularly cited using the general term called aerosols, they shape a combination of more than one version of additives. These additives can be of various phases like liquid or solid and can range drastically in shape and length. They are also originating from distinct sources[1]. Identification of the primary aerosol characteristics is significant because individual species of aerosols interact with solar radiation in distinct ways. The capability of human beings to inhale these particles is dependent on the particle size and once inhaled, their toxicity can range considerably. These particles also influence the vegetation index, can cause visibility degradation and can debase monuments and properties. While in situ instruments such as mass spectrometers, particle counters, and nephelometers are great in figuring out the aerosol particles that are close to the earth's surface, it's complex and expensive to carry out measurements using these instruments in the elevated layers of the atmosphere. Remote sensing instruments such as LIDARs (Light Detection and Ranging), spectrophotometers, and sun photometers, permit aerosol tracking in remote atmospheric regions. Therefore, measurements using remote sensing instruments are recently engaged in the identification of aerosols andthe development of numerous aerosol classification methods[2].

A classification procedure for aerosols dependent on the estimations from a double monochromator Brewer spectrophotometer uses data from 1998 to 2017 in Thessaloniki region, Greece[3]. In this paper, a machine learning clustering method (decision tree) is used on the data, and the metric applied is the Mahalanobis distance metric. The output class distribution is as follows: Dust Mixtures (DUST): 8.1%, UV Single Absorbing Mixtures (FNA):64.7%, Mixed: 9.8%, and Black Carbon Mixtures (BC): 17.4%. The measurements for the work are done using a CIMEL sunphotometer. While comparing the clustering potential of the algorithm with manually classified cases, the Mahalanobis algorithm shows a high typing score for all predominant clusters with DUST: 83.3%, BC: 66.7%, and FNA: 100.0%.

An aerosol classification technique based on Mahalanobis distance calculation with data from 190 AERONET (Aerosol RObotic NETwork) sites during the period 1993 to 2012 showcases the use of Microphysical and optical aerosol properties such as complex refractive index, Absorption Angstrom Exponents (AAE), SSA, and Extinction Angstrom Exponent (EAE) present in the visible region of Electro-Magnetic spectrum to classify atmospheric aerosols on a global scale. The classified aerosol types are Mixed Aerosol, Biomass Burning, Dust, Urban-Industrial, and Maritime[4].

*Corresponding author (E mail: smannapurna@gmail.com)

Radiation absorptivity and dominant size mode, which are determined by SSA and FMF respectively can be used for aerosol classification[5]. SSA isindicative of the absorptivity of the aerosol particle. Hence, it classifies absorbing and non-absorbing aerosols. FMF at 550 nm is used to determine the dominant size mode particles. Four reference AERONET locations are chosen & aerosol types are studied. The analysis reveals that aerosol types are partly affected by relative humidity and strongly affected by their sources and also shows that the absorptivity of industrial/urban aerosol in Europe and North America is lower than the aerosol present in Central America and Asia.

2 Materials and methods

2.1 AERONET-sunphotometer data

AERONET, a passive remote sensing network was developed by NASA for aerosol monitoring. It consists of a spectral sun-sky radiometer which measures the sun radiances directly at 8 spectral channels centered at 340, 380, 440, 500, 670, 870, 940, and 1020 nm. It employs approximately 400 AERONET sites in 50 countries on seven continents and India have about 24 AERONET sites.

The CIMEL sunphotometer data for LAP in Thessaloniki, Greece (40.630ºN, 22.960ºE) during the period 1998–2017 isgiven by the AERONET website (https://aeronet.gsfc. nasa.gov/cgi-bin/ webtool _aod_v3) and version 3 level 2 data were considered for analysis. The main reason for choosing this site for this study is that it provides all the high-quality necessary data for continuous ten years. Furthermore, the location has relatively high AOD values and varying aerosol concentrations throughout the year[5]. So, it can identify the dominant aerosol type present in the earth's atmospheric layers by employing various machine learning approaches.

The measured radiance from the sun photometer isn't equal to the radiance emitted by the sun (extra-terrestrial radiance) because the solar flux is reduced by atmospheric absorption and scattering. This relation is given by Beer's law. To get rid of the atmospheric effect, Langley extrapolation is completed. Once this is done, the extra-terrestrial radiance will be accustomed to find the AOT/AOD (Aerosol Optical Thickness /Aerosol Optical Depth). This measurement represents the degradation of the sunlight beam by haze and dust. The other parameters used in this analysis from CIMEL sun photometers are Fine Mode Fraction at 500 nm, Absorption Angstrom Exponent at 340–380 nm, Single Scattering Albedo measured at 440 nm, Extinction Angstrom Exponent calculated at 320–360 nm and Refractive index (real and imaginary) at 440 nm. A detailed explanation of all these parameters is given in the next section.

2.2 Aerosol optical properties

AOD is a fundamental observation in a sun photometer and is used to measure atmospheric aerosols such as smoke particles, desert dust, sea salt, and urban haze, which emerged within the vertical column of the atmosphere. The other aerosol parameters are explained below.

2.2.1 Single Scattering Albedo[440nm]

The SSA is a fundamental factor used as a measure of the contribution of relative scattering to total extinction and also act as an important variable for assessing the various climatological effects of aerosols[4]. It indicates the absorptivity of an aerosol particle and is given by:

$$SSA = \frac{Scattering}{Scattering + Absorption} \quad \ldots (1)$$

An SSA value equal to 1 indicates the presence of non-absorbing (scattering) particles, and a value equal to zero indicates absorbing particles. In this analysis, the SSA data ranges from 0.8304 - 0.9953 are used. Both the values correspond to higher values of SSA (since it is closer to 1). This hints at the dominance of non-absorptive aerosol particles in the study region.

2.2.2 Refractive Index-Real Part[440nm]

Refractive Index in aerosol optical properties is a complex quantity. It has both real and imaginary parts and also exhibits scattering and absorption properties. The real part (IR-Real) has elevated values when the scattering capability of the aerosol is high[4]. Therefore, if the real part of the refractive index increases, the proportion of scattering particles increases. In this analysis, the Real index of refraction data ranges from 1.33 - 1.57305 are used.

2.2.3 Refractive Index-Imaginary Part [440nm]

The imaginary part (IR-Img) is dependent on the real part and increases with the absorbing property of the aerosol[4]. In this analysis, the imaginary index of refraction data ranges from 0.000502 - 0.020874 was

used. These are low values, which hints at the possibility of non-absorbing particles. This conclusion is having a good agreement with the SSA data analysis.

2.2.4 Fine Mode Fraction [500nm]

It is defined as the ratio of total AOD to the fine AOD measured directly at 550 nm[4]. Generally, these two aerosol products are not directly provided by AERONET so they need to be interpolated to get the FMF at 500 nm. A second-order polynomial fit is applied to the logarithmic scale of the total and fine mode AOD values measured at 440 nm, 675 nm, 870 nm, and 1020 nm for calculating it at 500 nm. These wavelengths are preferred because they are available for a longer period[6]. The FMF is a quantitative indicator. FMF value is an indicator of aerosol size. An FMF value equal to 1 indicates the presence of dominant fine aerosol mixtures, and a value equal to zero indicates dominant coarse aerosol mixtures. In this analysis, the FMF data range from 0.186585 - 0.998475 was used. Hence, the FMF values are distributed in all ranges and a wide variety of classification flags can be deduced.

2.2.5 Absorption Angstrom Exponent [440-870nm]

It is measured from the direct slope of absorbing aerosol optical thickness to the function of wavelength[4]. It can be deliberated using the aerosol absorption optical thickness (AAOT) at 870 and 440 nm.

$$AAE - \frac{log\big(AAOT(870nm)\big) - log\big(AAOT(440nm)\big)}{log\big((870nm)\big) - log\big((440nm)\big)} \quad \ldots (2)$$

Where AAOT can be calculated using:

$$AAOT(\lambda) = \big(1 - SSA(\lambda)\big) * AOD(\lambda) \quad \ldots (3)$$

The AAE data range from 0.403025 - 3.859151 are used in this study.

2.2.6 Extinction Angstrom Exponent [440-870nm]

It represents the slope between extinction optical thickness and measured wavelength (Extinction = absorption + scattering)[6]. EAE is calculated based on aerosol extinction optical thickness (EOT) at 440nm, and 870nm.

$$EAE = -\frac{log\big(EOT(870nm)\big) - log\big(EOT(440nm)\big)}{log\big((870nm)\big) - log\big((440nm)\big)} \quad \ldots (4)$$

In this analysis, the EAE data ranges from 0.150052 -2.121572 are used. Both the angstrom exponents are qualitative properties.

2.3 Machine-learning algorithms

Six machine learning algorithms were used in this paper to validate the classified data. The accuracy, precision, and recall of these algorithms are compared in section V.

2.3.1 Logistic Regression algorithm

In Regression, the output variable value is determined by considering the input variable values in the labelled datasets, and hence, it is a supervised learning approach. Logistic Regression (LR) gives the multivariate outcome. It predicts whether an event will occur or not from the set of classified outputs based on values of input variables[7].

2.3.2 Random Forest algorithm

Random Forest (RF) algorithm proceeds by creating decision trees on data samples. Each of the decision trees makes a prediction. The best solution is selected finally through voting[7]. Over-fitting is reduced by averaging the result. Hence, it is better than a single decision tree.

2.3.3 K Nearest Neighbor algorithm

K Nearest Neighbor (KNN) proceeds by finding the distance between all the examples and a query in the data. These distances are sorted in ascending order and the specified number of examples (K) that are closest to the query are chosen. Each sample votes for the most frequent label (in the case of classification) and the decision are made in favor of the most voted label. The Euclidean distance[7]between the new instance and the existing instances are calculated using the formula:

$$Euclidean\ distance(x, x_i) = \sqrt{\sum \big(x_j - x_{ij}\big)^2} \quad \ldots (5)$$

Then, the distances are sorted in ascending order. A suitable value of K is chosen and the first K number of points from the sorted list is taken. The response from these points is considered and the majority among the responses is accepted as the predicted output.

2.3.4 Decision Tree algorithm

In the Decision Tree (DT) algorithm, classification problems are solved by the continuous splitting of

data based on a certain parameter. The choices are in the leaves and the information is part of the hub. The decision variable is clear cut (result as Yes/No) in the classification tree[7].

2.3.5 Naïve Bayes algorithm

Naïve Bayes (NB) set of rules is primarily based totally on conditional probability. The probability table present in the algorithm is the model that is used to update the training data. The "probability table" is based on its feature values wherein one desires to look up the class probabilities for predicting a new observation[7]. The fundamental assumption is of conditional independence and that is why it is called "naïve".

2.3.6 Support Vector Machines

In, Support Vector Machines (SVM) a defined hyperplane is used as the decision boundary. A set of objects belonging to various classes are isolated by a decision plane. The objects may or may not be linearly separable. The complex mathematical functions called kernels are expected to isolate the objects which are members of various classes[7]. SVM expects to accurately characterize the objects based on examples in the training data set.

3 Results and Discussions

The comparison of various machine learning algorithms requires an input case of the pre-classified CIMEL training dataset. An automated threshold-based methodology utilizing measurements from a single instrument is used as it permits more proficient, less abstract, and quicker grouping of an enormous number of cases[8]. For that feature selection of various aerosol parameters are important[9].

3.1 Reference Clusters

Fig. 1 illustrates the plot of different aerosol properties &their feature importance in %. After analyzing the feature importance, it is found that SSA at 440 nm and FMF at 500 nm are significant features. Features like AAE, EAE, RI (real and imaginary) are not significant when compared to the former two. Therefore, a threshold-based classification scheme as shown in Fig. 2 is used to assign classification flags to the data.

An automated aerosol classification method is proposed based on thresholds of the SSA at 440nm

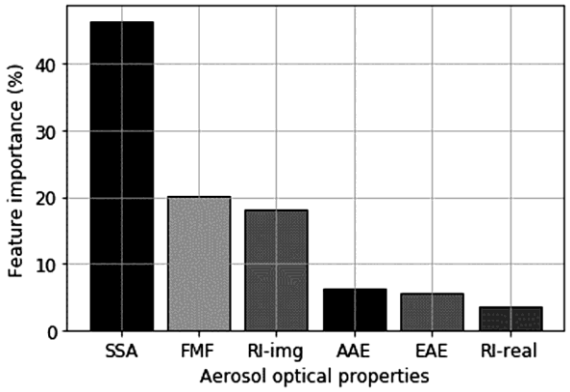

Fig. 1 — Feature importance of aerosol optical properties[4]

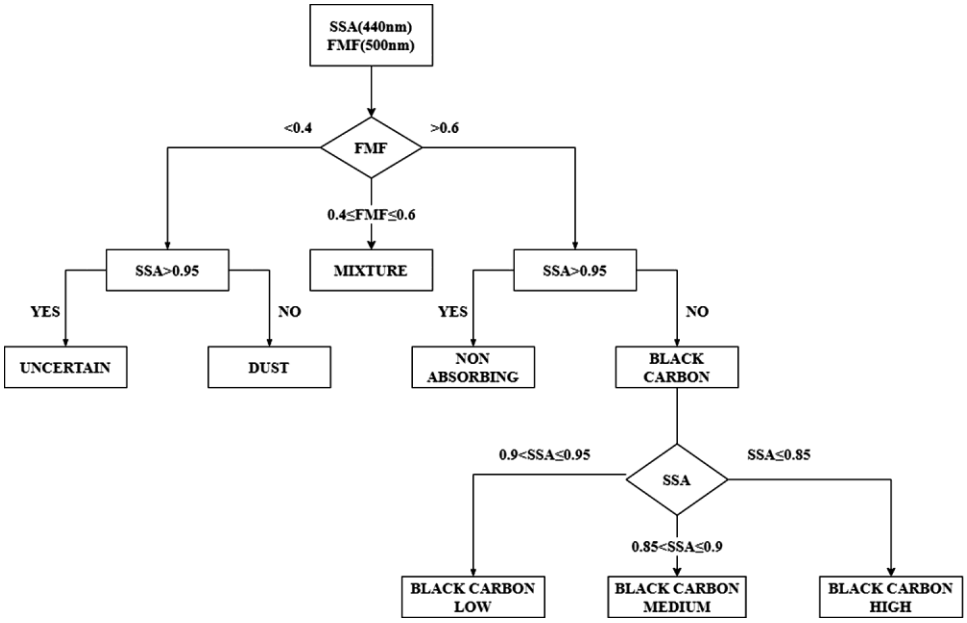

Fig. 2 — Aerosol classification algorithm process flow based on FMF and SSA[6]

and the FMF at 500 nm. The categories include Fine Non-Absorbing Mixtures (FNA), Dust Mixtures (DUST), Mixed, High Black Carbon Mixtures (BC High), Medium Black Carbon Mixtures (BC Med), and Low Black Carbon Mixtures (BC Low). Thresholds are applied to FMF at 500nm and SSA at 440nm to assign each type in one of the clusters. The SSA thresholds are 0.85 to 0.90 for BC Med, 0.90 to 0.95 for BC Low, and below 0.85, for BC High. For this examination, classification flags of the DUST group and FNA group are taken and BC High and BC Med clusters are converted into a solitary cluster named Black carbon Medium and it is preferred to utilize just the BC High cluster as it ought to contain combinations of more grounded Black Carbon component. Accessibility and the availability of BC High data is minor to shape a single cluster with numerous data points. It tends to be seen that the cluster optical properties are not unique concerning that the FNA cluster causing frequent misclassification between these two clusters. Merging of BC Low and FNA clusters also not ideal since it contrarily influences the BC classification scores. The mixed cluster is not considered in the reference dataset as it contains impacts of numerous aerosol types and the SALT cluster is barred as it contains just 3 cases.

The classification flags are obtained by applying the process of Fig. 2. On sunphotometer data (1760 cases) was shown in Fig. 3. Thresholds are applied to the SSA at 440nm and FMF at 500nm to assign each case in one of the clusters.

- FMF<0.4: Dominant coarse aerosol
- FMF>0.6: Dominant fine aerosol
- 0.4<FMF<0.6: safety margin

The FMF region between 0.4 and 0.6 is considered as a safe range and is applied here as opposed to utilizing the constraint of 0.6 to limit the immediate impact of different errors in the FMF count[5]. The SSA edge at 0.95 isolates ocean salt from dust for FMF at 500nm and values beneath 0.4. Fine particles are separated dependent on the FNA above 0.95 that incorporates ammonia, sulfate, nitrate, and organic aerosol, particles that predominantly disperse sunlight-based radiation, and three classes depend on the level of dark carbon particles it contains.

The output class distribution of the clusters in Fig. 4 indicates that black carbon low is the maximum found aerosol with 192 cases and non-absorbing is the second major component with 189 cases in the aerosol in Thessaloniki, Greece. Dust component is a minor one according to the composition having just 5 cases.

After assigning classification flags to each set of FMF and SSA, the next step is to apply various machine learning algorithms to the data and determine the accuracy of these algorithms for classifying the data. The machine learning algorithms used for the comparison are Support Vector Machine (SVM), Decision tree, Logistic Regression, Random Forest, K nearest neighbor, and Naïve Bayes

3.2 Performance metrics of machine learning algorithms

Various performance metrics are used to evaluate and compare machine learning algorithms.

3.2.1 Accuracy

It is the ratio of accurately anticipated perception of all the perceptions[10].

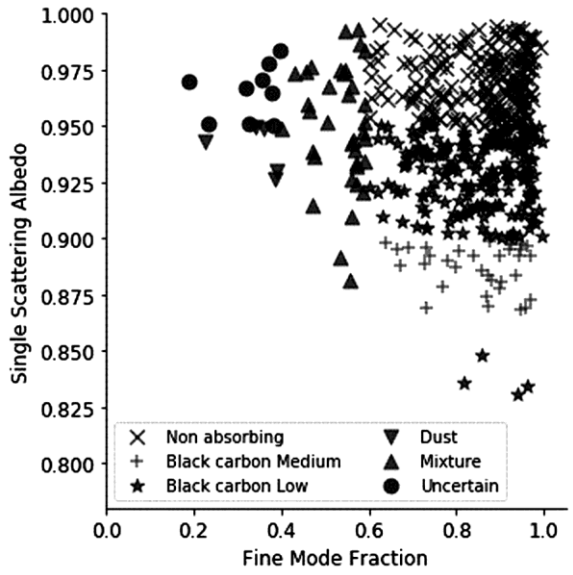

Fig. 3 — Classification flags on different aerosol types using FMF measured at 500nm & SSA measured at 440nm[6]

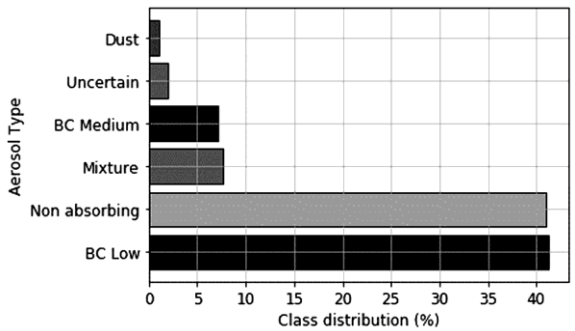

Fig. 4 — Output class distribution of Aerosol types[6]

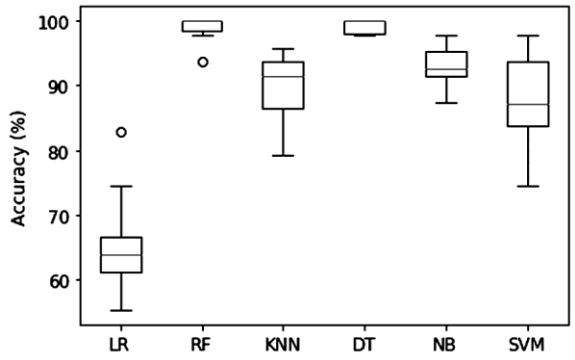

Fig. 5 — Comparison of accuracies of Machine Learning Algorithms[7]

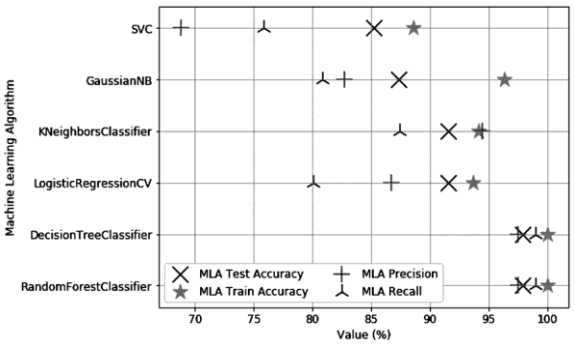

Fig. 6 — Comparison of performance metrics of Machine Learning algorithms[10] .

Table 1 — Accuracy of Machine Learning Algorithms[7]

Algorithm	Accuracy (%)
Logistic Regression	65.75
Random Forest	98.95
K Nearest Neighbor	89.66
Decision Tree	99.15
Naïve Bayes	93.03
SVM	86.92

Table 2 — Comparison of performance metrics of machine learning algorithms[10]

Algorithm	Testaccuracy (%)	Train accuracy (%)	Precision (%)	Recall (%)
LR	92.1	93.65	86.70	80.08
RF	98.95	100.00	97.47	98.97
KNN	91.66	94.17	94.46	87.39
DT	99.15	100.00	97.47	98.97
NB	87.03	96.29	82.71	81.86
SVM	85.02	88.62	68.84	75.86

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \quad \ldots (6)$$

3.2.2 Precision

Precision is the ratio of correctly anticipated positive perceptions of the absolute anticipated positive perceptions. High precision relates to the low false-positive rate[10].

$$Precision = \frac{TP}{(TP + FP)} \quad \ldots (7)$$

3.2.3 Recall

The recall is the ratio of effectively predicted positive observations to all observations in the actual class[10].

$$Recall = \frac{TP}{(TP + FN)} \quad \ldots (8)$$

- TP: True Positive
- TN: True Negative
- FP: False Positive
- FN: False Negative

Performance metrics such as accuracy (test and train), Precision, and Recall are found out and plotted in Fig. (5 and 6). From Table 1, it can be observed that the Decision Tree algorithm shows the highest accuracy, followed by Random Forest, Naïve Bayes, K Nearest Neighbor, SVM, and Logistic regression. The accuracy of the random forest algorithm is 98.95 and that of the decision tree is 99.15. The lowest accuracy is encountered in Logistic regression (65.75). The remaining algorithms work in between these two and the trends are plotted in Fig. 5.

The comparative analysis of Fig. 6 and Table 2 reveals that the accuracy (test and train), precision, and recall are found to be highest for the decision tree algorithm, followed by the random forest algorithm. While analyzing other algorithms MLA train accuracy shows high values followed by MLA test accuracy, MLA recall, and MLA precision. For all the machine learning algorithms the different performance metrics lie in the ranges from 0.5 to 1.

4 Conclusion

An automated machine learning approach has been applied to the CIMEL sunphotometer measurements of SSA and FMF collected from the AERONET site operated by NASA. The technique currently assigns the classification
flags to the daily data of Thessaloniki, Greece, measured using a CIMEL sunphotometer to build a reference dataset. Once the reference dataset is formed with proper classification flags, machine learning algorithms has been used to validate the model so formed. Various aerosol types have been

obtained and the classification flags have been labelled. The labels obtained are Fine Non-Absorbing Mixtures (FNA), Dust Mixtures (DUST), High Black Carbon Mixtures (BC High), Low Black Carbon Mixtures (BC Low), Medium Black Carbon Mixtures (BC Med), and Mixed. A total of 1760 cases have been taken during the measurement period of 1998–2017 and have been assigned automatically to one of these reference clusters based on a decision tree and threshold approach. FNA and BC low have been encountered more often in Thessaloniki. Mixtures and BC medium are less common and DUST is rarer. Because of the event of sea salt combinations in uncommon conditions, they have been eliminated from the aerosol classification algorithm. FNA mixtures have been commonly predominant from 1998 to 2017. DUST and BC mixtures have shown up just infrequently in the period 1998–2017. Further, this classified data has beenused as input to various machine learning algorithms. Decision Tree has exhibited maximum accuracy (test and train), precision, and recall. Hence, it has been recommended as the best-fit algorithm for classifying such type of data. In the future, the model will be evaluated for denser datasets of different regions with different geographical conditions.

References

1 Gao Song, Zhou Zhicheng, Tang Zhiwei & Bi Xiaotian, *IEEE International Conference on Power System Technology*, Guangzhou, China, 1 (2018) 3491.

2 Pramod Kulkarni, Paul A. Baron& Klaus Willeke, Aerosol Measurement: Principles, Techniques, and Applications (John Wiley &Sons,New Jersey), 3rd Edn, ISBN: 9780470387412, 1 (2011)1.

3 Nikolaos Siomos, Ilias Fountoulakis, Athanasios Natsis, Theano Drosoglou & Alkiviadis Bais, *Remote Sens*, 12 (2020) 965.

4 Patrick Hamill, Marco Giordano, Carolyne Ward, David Giles & Brent Holbein, Atmos Environ, 140 (2016) 213.

5 Sunita Verma, Divya Prakash, Philippe Ricaud, Swagata Payra, Jean-Luc Attié & Manish Soni1, Aerosol Air Qual Res, 15 (2015) 985.

6 J Lee, J Kim, Song C H, Kim S B, Chun Y, Sohn B J& Holben B N, Atmos Environ, 44 (2010) 3110.

7 Susmita Ray, *A Quick Review of Machine Learning Algorithms*, 1 (2019) 35.

8 Nikolaos Papagiannopoulos, Lucia Mona, Aldo A modeo, Gluseppe D Amico, Pilar Guma Claramunt, Gelsomina Pappalardo, Lucas A lados A rboledas, Juan Luis Guerrero-Rascado, Vassilis A miridis, Panagiotis Kokkalis, Arnoud Apituley, Holger Baars, Anja Schwarz, Ulla Wandinger, Ioannis Binietoglou, Doina Nicolae, Daniele Bortoli, Adolfo Comeron, Alejandro Rodriguez-Gomez, Michael Sicard, Alex Papayannis & Matthia Wiegner, Atmospheric Chem Phys, 18 (2018) 15879.

9 Burton S P, Ferrare R A, Hostetler C A, Hair J W, Rogers R R, Obland M D, Butler C F, Cook A L, Harper D B& Froyd K D, Atmos Meas Tech, 5 (2012) 73.

10 Mikhail D. Molovtsev & Irina S. Sineva, *Classification Algorithms Analysis in the Forest Fire Detection Problem*, 2019 International Conference "Quality Management, Transport and Information Security, Information Technologies", 1 (2019) 548.