Estimation of microphysical parameters of atmospheric pollution using Machine Learning

C. Llerena[1]**, D. Müller[2], R. Adams[3], N. Davey[3], and Y. Sun[3]

[1]Polytechnic school, University of Alcalá, Spain
[2] School of Physics, Astronomy & Mathematics, University of Hertfordshire
[3]Centre for Computer Science and Informatics Research, University of Hertforshire, United Kingdom, AL10 9AB
cosme.llerag@gmail.com
{d.mueller,r.g.adams,n.davey,y.2.sun}@herts.ac.uk

Abstract. The estimation of microphysical parameters of pollution (effective radius and complex refractive index) from optical aerosol parameters entails a complex problem. In previous work based on machine learning techniques, Artificial Neural Networks have been used to solve this problem. In this paper, the use of a classification and regression solution based on the k-Nearest Neighbour algorithm is proposed. Results show that this contribution achieves better results in terms of accuracy than the previous work.

Keywords: LIDAR, Particle Extinction Coefficient, Particle Backscatter, Effective Radius, Complex Refractive Index, k-Nearest Neighbour

1 Introduction

One of the main important factors that drive climate change is particulate pollution [1]. To understand atmospheric temperatures changes that cause climate change, it is necessary to study and characterise the optical, chemical and microphysical properties of these particles in the atmosphere. Some technologies like radiometers or Light Detection and Ranging (LIDAR) make possible the observation of aerosols. LIDAR has become a key tool for the characterization of atmospheric pollution in the atmosphere. LIDAR is the only remote sensing technique used in research on atmospheric pollution that allows for vertically-resolved observations of particulate pollution, for example, [2]. Using LIDAR, optical aerosol parameters can be extracted [3] but more information about particles is required to understand the impact of pollution on climate change.

Microphysical particle parameters are also of key interest to determine pollution effects. In [4–8], inversion algorithms are used to estimate microphysical information (particle size or complex refractive index) from optical data. Their estimation is a very complex task because many factors such as ambient atmospheric humidity, the condensation of gases on existing particles or the mixing

** Corresponding author.

of particles of different chemical properties modify the values of the optical data [9]. Due to these difficulties, inversion algorithms are very complex and require an extensive mathematical background as we are dealing with ill-posed inverse problems [10].

Therefore, less complex solutions must be proposed as we need techniques that a) allow for fast data processing in view of current and up-coming LIDAR space missions; b) offer autonomous data retrieval in view of serious lack of experts in this research field; and c) provide us with ways of exploiting the information content of these highly complex data sets in an optimum way. Using synthetic optical data, authors in [9] have developed a computational model using Artificial Neural Networks (ANNs) [11] to estimate the effective radius of particles (r_{eff}) and the complex refractive index from combinations of extinction (α) and backscatter (β) coefficients. Specifically, these authors use values of α and β at different wavelengths ($\lambda = 355, 532$ and 1064 nm). These wavelengths are currently used by most of the LIDAR system in the world for the investigation of particualate pollution in the atmosphere. Most notably and in view of advantages not further detailed in this contribution, there has been a push for emitting all three wavelengths simultaneously in the past 20 years. Five combinations of α and β were tested, resulting in finding the most suitable one which uses the values of α at $\lambda = 355$ and 532 nm (α_{355} and α_{532}) and the values of β at $\lambda = 355, 532$ and 1064 nm (β_{355}, $\beta_{532}$and β_{1064}). For technical reasons the measurement of α at 1064 nm has become possible just recently. The quality of these data, however, still needs to be improved before tests with this extended set of $\beta + \alpha$ data can be carried out. Moreover, ANNs were evaluated for three different size ranges of effective radii, that is, particles with r_{eff} between $10-100$ nm, $110-250$ nm and $260-500$ nm, respectively. This separation was performed by hand due to two reasons: a) to limit the computation time and b) to separate particles according to their nature. Without going into further details particles in these three different size ranges have different effect on climate change and human health.

The aim of this work is to investigate whether we can develop a computational method based on ML techniques which can first classify particles in to the three categories, then, can estimate particle properties within each category, or not. In addition, we look for a model with less computational cost than the one proposed in [9].

2 Data Description

The dataset is the synthetic one used in [9], which was generated using a Mie scattering algorithm [12]. It contains 1,665,343 particles. According to the three ranges of r_{eff}, there are 330,480 particles with a radius between 10 nm and 100 nm, 503,155 samples with a radius between 110 nm and 250 nm and 831,708 particles with a radius between 260 nm and 500 nm.

The following information for each *particle size distribution* can be found:

- Extinction and backscatter coefficients at different wavelengths (355, 532 and 1064 nm). As in [9], the best combination of α and β will be used (α_{355}, α_{532}, β_{355}, β_{532} and β_{1064}).
- Mode width, from 1.4 to 2.5 in step of 0.1. The mode width is the geometrical standard deviation of the theoretical model that describes in an approximate manner the shape (number concentration versus particle size) of naturally occurring atmospheric size distributions. This shape, referred to as *logarithmic-normal* can be described as a Gauss distribution if particle radius is plotted on a logarithmic scale. More details can be found in e.g. [13]. The total number of particles in the atmosphere can be modelled by a sum of sets (distributions) according to the radius. Particle size distributions in the atmosphere can be described by 5-6 different modes. Each mode has its own mean radius (or alternatively we can also use mode radius) which is the value where the size distribution reaches its maximum value) and the mode width. In the present case we simplified our simulations in the sense that we did not use combinations of these modes in order to cover the vast size range of particles from a few nanometers to several tens of micrometers. In this first set of studies we mimicked these naturally-occurring multimodal size distributions by the use of single-mode logarithmic-normal (log-normal) distributions which not only cover the relevant particle radius range but result in sufficiently realistic optical properties as well. Furthermore, effective radius is a commonly used number in climate modelling. It reduced the complexity of size distribution information from *mean radius and mode width* (in each mode) to a single number. Alternatively effective radius can also be used for each individual mode. Optical properties of particle size distributions described in terms of effective radius are sufficiently close to optical properties of the underlying size distribution when used in modelling.
- Mean radius (nm), from 10 nm to 500 nm in step size of 10.
- Real (from 1.2 to 2 in step size of 0.025) and imaginary part (from 0 to -0.1 in step size of $9.99 \cdot 10^{-6}$) of the complex refractive index. This index indicates the attenuation suffered by light when passing through a particle.

Figure 1 shows how α and β vary with respect to λ for a mode width equal to 1.4. Looking at this figure, the reader can note that α has quite similar values at the different wavelengths, while β decreases as the wavelength increases. It must be said that similar behaviour is observed in the rest of the mode widths (from 1.5 to 2.5). It can be seen that the backscatter coefficient increases as the effective radius increases. Those variations are larger for higher mode widths. Similar observations can be found in the variations of α.

Furthermore, the variation of α and β across the particle effective radius in different width modes are also investigated. Figures 2 and 3 show the variation of β in two different width modes, 2.0 and 2.5, namely. It can be seen that the backscatter coefficient increases as the effective radius increases. These variations are larger when the value of the mode width is higher. Similar observations can also be found in the variations of α.

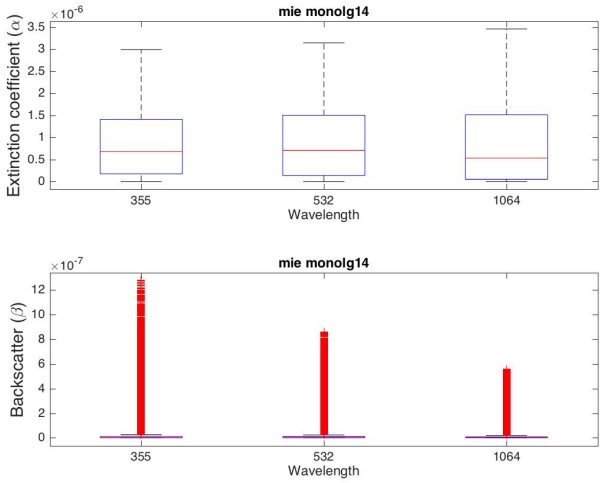

Fig. 1. Variation of α and β in mode width 1.4 at the different values of λ.

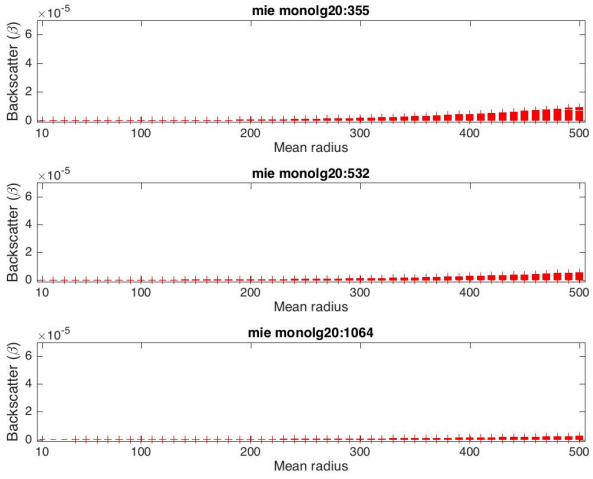

Fig. 2. Variation of β with respect to effective radius in mode width equal to 2.0.

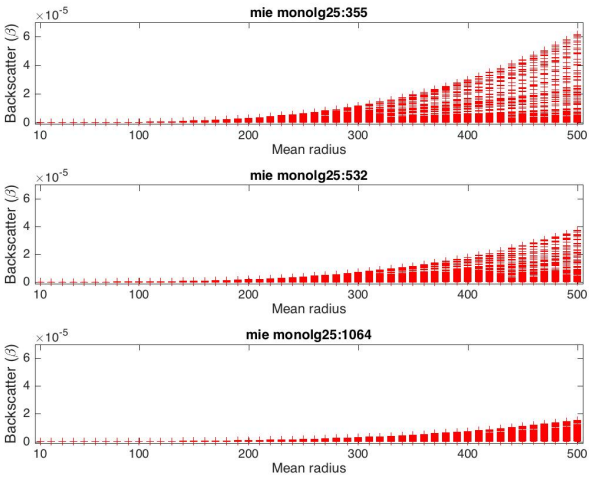

Fig. 3. Variation of β with respect to effective radius in mode width equal to 2.5.

To determine which ML techniques can be applied to the classification and estimation stages, first Principal Component Analysis (PCA) has been applied. Figure 4 shows a PCA plot of the original synthetic data. It can be seen that the class of $110 - 250$ nm overlaps with the class of $10 - 100$ nm and the class of $260 - 500$ nm, but the class of $10 - 100$ nm and the class of $260 - 500$ nm do not overlap between them.

According to Figure 4, k-NN [14]) can be considered as a solution because the three classes are not strongly overlapped. In addition, Extreme Learning Machine (ELM) [15] has been chosen to be a potential solution since this deep learning solution uses a similar ANNs architecture to the one used in [9] and it can have a lower computational cost.

3 Retrieval of Microphysical Parameters

The estimation of microphysical parameters using ANNs was addressed in [9]. Due to the computational cost and the nature of particles, this estimation was performed in three different ranges of particle sizes separately. Taking this into account, we propose two solutions in this paper. The first one is a single regression solution, which uses an ML technique to estimate microphysical parameters for all the particles together. This technique must outperform ANNs in terms of accuracy. The second one includes two steps: 1) a classification that will separate particles into the three classes and then 2) a regression that will estimate microphysical parameters.

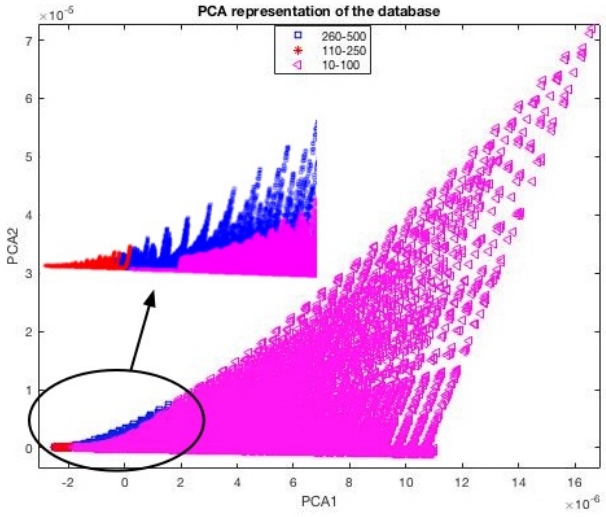

Fig. 4. PCA2 versus PCA1.

3.1 Single regression solution

In work [9], a Multi-Layer Perceptron (MLP) with one hidden layer was used to estimate microphysical parameters. Specifically, different configurations of MLPs (training algorithms, number of neurons in the hidden layer, activation functions) were tested. They concluded that the most suitable MLP contains five neurons in the hidden layer and uses the Levenberg Marquardt training algorithm. In this sense, we have implemented the same MLP to be compared with ELM and k-NN both in terms of accuracy and computational cost.

3.2 Combined solution

Figure 5 shows a flow diagram of this combined solution. It can be seen that five-feature vectors are classified into three classes according to r_{eff} first. Then one single regression estimation model is trained within each class. In both classification and estimation stages, the suitability of using ELM and K-NN will be evaluated. The detailed solution is given as follows:

1. The whole dataset has been split into training (75%) and test (25%) sets.
2. Training a classification model using the training set.
3. The test set has been split into the three classes.
4. Within each class, a regression model is trained.
5. Finally, the regression models in step 4 were used to estimate microphysical parameters on the test set.

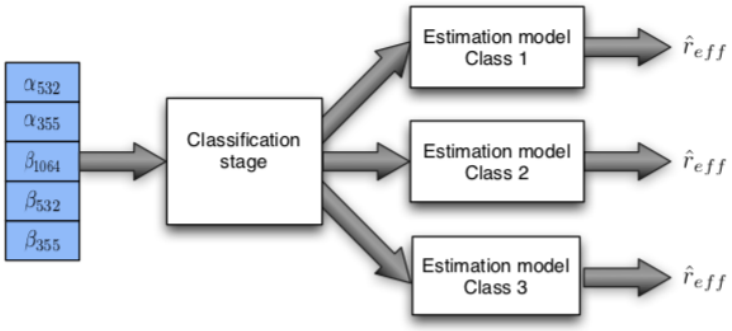

Fig. 5. Scheme of the full solution combining classification and regression estimation.

4 Experiments

In this section the results of both solutions are provided. To evaluate the classification solution, precision, recall and F1-score [16] are calculated for each class since the three classes are imbalanced. To test the suitability of each microphysical parameter estimation solution, the Root Mean Square Error (RMSE) has been used. The computational cost of each technique is evaluated with CPU time.

4.1 Single regression solution

In Table 1, ELM and k-NN are compared with the MLP configuration in the previous work, when all the particles are analysed together. For the MLP, the whole dataset has been split into training (65%), validation (15%) (where the validation set is used to control overfitting) and test (25%) sets while for k-NN and ELM, the dataset has been split into training (75%) and test (25%). Note that the results for the different classes are also extracted from the total results and shown in the table.

Due to the space limit, only the results for the best configuration of each method are presented in the table. For instance, the methodology based on k-NN has been tested for different number of neighbours ($1, 3, 4, 5, 7, 11, 15, 19$, 29 and 39), resulting in $k = 1$ being the best choice. In the case of ELM, different activation functions (sigmoid, sine or hard limit) and number of neurons in the hidden layer ($N = 2, 3, 4, 5, 7, 10, 20, 30$, $50, 75, 100, 150, 200$ and 300) have been tested. Sigmoid function and $N = 300$ achieve the best results.

Looking at Table 1, it is clear that k-NN produces the best results (the lowest values of RMSE) for all the parameters and across all class of particles.

Table 1. RMSE obtained by MLP [9], ELM and k-NN based solutions when r_{eff}, real and imaginary part of the complex index are estimated.

Param.	Method	Whole data	10 - 100 nm	110 - 250 nm	260 - 500 nm
r_{eff}	*MLP* [9]	43.40	14.51	37.97	53.03
	kNN (k=1)	**31.18**	**6.17**	**27.80**	**38.24**
	ELM (N=300)	44.37	15.29	39.82	53.74
Real Part	*MLP* [9]	0.15	0.19	0.14	0.14
	kNN (k=1)	**0.08**	**0.13**	**0.07**	**0.07**
	ELM (N=300)	0.19	0.23	0.17	0.18
Imag. Part	*MLP* [9]	0.18	0.25	0.16	0.15
	kNN (k=1)	**0.08**	**0.10**	**0.09**	**0.06**
	ELM (N=300)	0.26	0.29	0.27	0.26

Another aspect to be considered is the computational cost associated with each solution. In some applications, study of the atmospheric pollution must be in real-time and so, it is important to have fast algorithms. Bearing this in mind, Table 2 shows the mean values of CPU times of each solution for training and test. These experiments have been carried out on a computer with a $2,8$ GHz Intel Core i7 processor and 8 Gb RAM.

Table 2. Mean values of CPU time (s) associated with MLP [9], ELM and k-NN based solutions.

	MLP [9]	**k-NN**	**ELM**
Training	753.86	7.25	181.79
Test	0.44	3.00	5.29

It is clear that the solution based on k-NNs is less expensive in terms of CPU time when training. In test, the MLP needs less time but it is less accurate obtaining parameters. In the case of ELM, larger values of CPU time are required and it achieves worst RMSE values. For these reasons, larger number of neurons have not been studied with ELM.

4.2 Combined solution

As it is shown in Figure 5, the combined solution allows us to split particles into the three classes (10 - 100 nm, 110 - 250 nm, 260 - 500 nm) before estimating parameters. At the classification stage, MLP, k-NN and ELM based classifiers have been studied with k-NN giving the highest accuracy (96.09%) when $k = 3$. Since classes are unbalanced, precision, recall and F1- scores are also provided for each class in Table 3.

Table 3. Precision, recall and F1- scores obtained by the solution based on k-NNs ($k = 3$) for each particle class.

Class	**Precision (%)**	**Recall** (%)	**F1-score**
Class 1 (10 - 100 nm)	98.76	97.69	0.98
Class 2 (110 - 250 nm)	93.63	93.43	0.94
Class 3 (260 - 500 nm)	96.52	97.07	0.97

Very good results are achieved, the worst being for class 2, what makes sense, since it overlaps with the rest of the classes (Figure 5).

Once particles in the test set have been separated using k-NN, the estimation models must be applied. Obtained RMSE scores are presented in Table 4.

Table 4. RMSE obtained by MLP [9], ELM and k-NN based solutions when r_{eff}, real and imaginary part of the complex index are estimated for Combined solution.

Parameter	**Method**	**10 - 100 nm**	**110 - 250 nm**	**260 - 500 nm**
r_{eff}	*MLP* [9]	10.56	30.29	51.17
	kNN (k=1)	**5.86**	**27.83**	**37.82**
	ELM (N=300)	11.01	32.38	50.79
Real Part	*MLP* [9]	0.22	0.16	0.17
	kNN (k=1)	**0.09**	**0.07**	**0.07**
	ELM (N=300)	0.21	0.15	0.17
Imaginary Part	*MLP* [9]	0.27	0.27	0.15
	kNN (k=1)	**0.06**	**0.08**	**0.07**
	ELM (N=300)	0.24	0.25	0.23

It can be seen from this table, the combined solution based on k-NN achieves the best results for all the classes and microphysical parameters. If we compare these results with those in Table 1, we can see that in general the combined method performs better or similar to the single regression solution when the classification stage has been applied previously. Moreover, it shows ELM performs much better after classification is done first, that is, it performs better within each class. As for the computational cost, similar conclusions as those from Table 2 can be made.

5 Discussion & Conclusions

Estimating microphysical parameters from optical data can be done using ML techniques. [9] is a very interesting paper and from it, two objectives have been met in our work. Firstly, we provide a solution that produces lower RMSE and computational cost when estimating microphysical parameters. Secondly, a new

combined solution, which produces high accuracy at the classification stage, has been implemented.

References

1. Stocker, T. F., Qin, D., Plattner, G. K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Midgley, P. M.: Climate change 2013: the physical science basis. Intergovernmental panel on climate change, working group I contribution to the IPCC fifth assessment report (AR5). New York. Cambridge, United Kingdom and New York, NY, USA, pp. 1535 (2013)
2. Ansmann, A., Müller, D.: Lidar and Atmospheric Aerosol Particles, in: Weitkamp, C. (Ed.), Lidar. Springer New York, pp. 105–141 (2005).
3. Ansmann, A., Riebesell, M., Weitkamp, C.: Measurement of atmospheric aerosol extinction profiles with a Raman lidar. Optics Letters, vol. 15, pp. 746–748 (1990)
4. Veselovskii, I., Kolgotin, A., Griaznov, V., Müller, D, Wandinger, U., Whiteman, N. D.: Inversion with regularization for the retrieval of tropospheric aerosol parameters from multiwavelength lidar sounding. Applied Optics. Vol. 41, No. 18, pp. 3685-3699 (2002)
5. Müller, D., Wandinger, U., Ansmann, A.: Microphysical Particle Parameters from Extinction and Backscatter Lidar Data by Inversion With Regularization: Simulation. Applied Optics vol. 38, pp. 2358–2368 (1999)
6. Böckmann, C., Mironova, I., Müller, D., Schneidenbach, L., Nessler, R.: Microphysical aerosol parameters from multiwavelength lidar. Journal of the Optical Society of America A vol. 22, pp. 518–528 (2005)
7. Kolgotin, A., Müller, D.: Theory of inversion with two-dimensional regularization: profiles of microphysical particle properties derived from multiwavelength lidar measurements. Applied Optics vol. 47, pp. 4472–4490 (2008)
8. Müller, D., Kolgotin, A., Mattis, I., Petzold, A., Stohl, A.: Vertical profiles of microphysical particle properties derived from inversion with two-dimensional regularization of multiwavelength Raman lidar data: experiment. Applied Optics 50, 2069–2079 (2011)
9. Mamun, M. M., Müller, D.: Retrieval of Intensive Aerosol Microphysical Parameters from Multiwavelength Raman/HSRL Lidar: Feasibility Study with Artificial Neural Networks. Neural Networks. Atmos. Meas. Tech. Discuss, 7 (2016)
10. Hadamard, J.: Bull. Univ. of Princeton 13, 49 (1902).
11. Schalkoff, R. J.: Artificial neural networks. Vol. 1. New York: McGraw-Hill (1997)
12. Bohren, C., Huffman, D.: Absorption and Scattering of Light by Small Particles. Wiley science paperback series. (1998)
13. Hinds, C.W.: Aerosol Technology: Properties, Behavior, and Measurement of Airborne Particles, 2nd Edition, ISBN: 978-0-471-19410-1, Jan 1999, 504 pages.
14. Peterson, L. E.: K-nearest neighbor. Scholarpedia, vol. 4(2), pp. 1883 (2009)
15. Huang, G-B., Zhu, Q-Y., Siew, C-K.: Extreme learning machine: theory and applications. Neurocomputing, vol. 70, no 1–3, p. 489–501 (2006)
16. Goutte, C., Gaussier, E.: A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In European Conference on Information Retrieval pp. 345–359, Springer, Berlin, Heidelberg (2005)