Aerosol and Air Quality Research

ORIGINAL RESEARCH
https://doi.org/10.4209/aaqr.220386

Machine Learning Classification Model to Label Sources Derived from Factor Analysis Receptor Models for Source Apportionment

Vikas Kumar[1], Vasudev Malyan[2], Manoranjan Sahu[2,1,3*], Basudev Biswal[4]

[1] Interdisciplinary Program in Climate Studies, Indian Institute of Technology Bombay, Mumbai 400076, India
[2] Aerosol and Nanoparticle Technology Laboratory, Environmental Science and Engineering Department, Indian Institute of Technology Bombay, Mumbai 400076, India
[3] Centre for Machine Intelligence and Data Science, Indian Institute of Technology Bombay, Mumbai 400076, India
[4] Department of Civil Engineering, Indian Institute of Technology Bombay, Mumbai 400076, India

Special Issue:
In honor of Prof. David Y.H. Pui for his "50 Years of Contribution in Aerosol Science and Technology" (VII)

ABSTRACT

Factor analysis (FA) receptor models are widely used for source apportionment (SA) due to their ability to extract the source contribution and profile from the data. However, there is subjectivity in the source identification and labelling due to manual interpretation, which is time-consuming. This raises a barrier to the development of the real-time SA process. In this study, a machine learning (ML) classification algorithm, k-nearest neighbour (kNN), is applied to the source profiles obtained from the United States Environmental Protection Agency's (U.S. EPA) SPECIATE database to develop a model that can automatically label the factors derived from FA receptor models. The train and test score of the model is 0.85 and 0.79, respectively. The overall weighted average precision, recall and F1 score is 0.79. The performance of the model during validation exhibits acceptable results. The application of ML models for source profile labelling will reduce the time taken and the subjectivity associated with results due to modeler bias. This process can act as another layer of the process for verification of the results of FA receptor models. The application of this methodology advances the process towards real-time SA.

Keywords: Particulate matter, Source apportionment, Receptor models, Machine learning, Classification

Received: November 7, 2022
Revised: March 26, 2023
Accepted: April 8, 2023

* **Corresponding Author:**
mrsahu@iitb.ac.in

Publisher:
Taiwan Association for Aerosol Research
ISSN: 1680-8584 print
ISSN: 2071-1409 online

1 INTRODUCTION

Source apportionment (SA) is the practice of identifying the emission sources and their contribution to develop control strategies for effective air quality management (Karagulian and Belis, 2012; Bove *et al.*, 2014; Hopke, 2016). Receptor models apportion the ambient concentrations to their respective sources. Chemical mass balance (CMB) is the most suitable receptor model if the source profiles are available. However, in the absence of source profiles, multivariate factor analysis (FA) models like positive matrix factorization (PMF), principal component analysis (PCA), and UNMIX are the alternatives (Viana *et al.*, 2008; Hopke and Cohen, 2011). FA models extract the source contribution and profile from the data itself (Viana *et al.*, 2008; Hopke and Cohen, 2011; Hopke, 2016) because of which it is the most widely used receptor model (Karagulian et al., 2015; Hopke et al., 2020). The profiles must be designated with source based on the literature or compared with measured source profiles, which is time-consuming and involves manual interpretation. It is the most subjective and least quantifiable step in applying FA receptor models (Reff *et al.*, 2007).

ORIGINAL RESEARCH
https://doi.org/10.4209/aaqr.220386

Hopke *et al.* (2020) reported 741 global particulate matter (PM) apportionments from 414 published papers during 2014–2019, of which only 89 applied CMB while others applied FA receptor models. 539 cases applied PMF, which is the most utilized receptor model. A review of these studies exhibits that the average (± SD) gap between the data collection and result publication is 4 (± 2) years, while the average delay in the publication of FA receptor models is 4 (± 3) years. Only 38 (6%) studies reported results for FA receptor models within a year. 166 (25%) and 149 (23%) studies reported results after 2 and 3 years, respectively. A few studies have taken almost 8–12 years to publish results by the time the results become irrelevant. This raises a barrier to developing the real-time SA process to track the source of pollution and their emission in real-time. Real-time SA is essential for managing air quality as a lot of information from dynamic source activity and various episodic events are suppressed when samples are collected for a long period during the study period. The improved time resolution of identification of pollution sources can help regulatory agencies take immediate action and reduce the emission at the source to improve air quality. Long-term online measurements coupled with the receptor model could be considered for real-time SA for managing air quality (Rai *et al.*, 2020; Lalchandani *et al.*, 2021; Prakash *et al.*, 2021; Yang *et al.*, 2022). In these studies, real-time chemical speciation instruments are used to obtain high-resolution (30 min or 1-h) chemical characterization of the aerosols. However, there is still a gap of almost 1–2 years in obtaining the results of these studies. This indicates that the process of applying the receptor model and result interpretation needs to be updated to achieve real-time SA.

Pernigotti and Belis (2018) developed DeltaSA, a tool to assign a factor to a source in FA receptor models based on the similarity between a given factor and source chemical profiles from public databases. Similarly, Liao *et al.* (2022) developed a numerical method to identify apportioned factor profiles by integrating distance and probability-based profile matching approaches. It is also essential to have models which can be integrated easily into the receptor models. The application of machine learning (ML) algorithms can be a potential solution to facilitate source identification. Automation of the process of labelling profiles derived from FA receptor models using ML can reduce the time taken for the process immensely and the uncertainty associated with results due to modeler bias. This will also streamline the process of identification and labelling of source profiles. Classification models are the most appropriate choice for the problem because of their ability to designate new samples with labels based on previous data. To our knowledge, no approaches have been developed to apply ML algorithms for linking factors to the appropriate sources.

The objective of this study are as follows: (a) investigate the application of ML classification algorithms to label profiles with appropriate sources based on the SPECIATE database; (b) validate the model on source profiles available in the literature.

2 MATERIALS AND METHODS

2.1 Data

The United States Environmental Protection Agency's (U.S. EPA) SPECIATE is a repository of organic gas and particulate matter (PM) category-specific emission speciation profiles of air pollution sources. SPECIATE 5.1, the latest version, includes 6,746 PM, gas, and other profiles. These emission source profiles are used to provide input to the Chemical Mass Balance (CMB) receptor models, verify profiles derived from ambient measurements by multivariate receptor models (e.g., factor analysis and positive matrix factorization) and interpret ambient measurement data (Simon *et al.*, 2010; U.S. EPA, 2015; Bray *et al.*, 2019; U.S. EPA, 2019). Data used to create these profiles come from various sources, including peer-reviewed journal articles and emissions testing conducted primarily by the EPA (Bray *et al.*, 2019). Further details about the SPECIATE database is available in Simon *et al.* (2010), Bray *et al.* (2019) and U.S. EPA (2019). In this study, only $PM_{2.5}$ source profiles are used for model development. 1731 $PM_{2.5}$ source profiles were collected from the SPECIATE repository (U.S. EPA, 2015) and were grouped into five major categories (count in parentheses) namely, biomass burning (325), coal combustion (108), dust (431), industrial (312) and traffic (555). The details of SPECIATE profiles used in this study and the respective assigned sources is provided in Table S1 of the Supporting Information. The uncertainties associated with the species in the source profiles, as provided by the SPECIATE is also considered in this study.

ORIGINAL RESEARCH
https://doi.org/10.4209/aaqr.220386

The total data was randomly split into a 70/30 ratio with train and test sizes of 1211 and 520, respectively. The source-wise train and test samples (count in parentheses) are biomass burning (219; 106), coal combustion (71; 37), dust (299; 132), industrial (224; 88) and traffic (398; 157). The split tries to maintain a 70/30 ratio for the whole dataset and each source. The list of elements and composite profiles of the five categories is provided in Table S2 of the Supporting Information.

2.2 Methodology

The $PM_{2.5}$ source profiles collected from SPECIATE database were fed into the ML model as input data, and the trained model was validated on the test dataset. The trained model can label the new factors derived from FA receptor models by assigning a source profile to the factors. The modeling framework implemented in this study is presented in Fig. 1. The k- Nearest Neighbour (kNN) classification ML algorithm is applied to develop the model. kNN is a memory-based algorithm and assumes that similar things exist nearby. kNN classifies the value for new samples using the majority vote of the k closest samples from the memorized data (Hastie *et al.*, 2009; Kuhn and Johnson, 2013; Sammut and Webb, 2011; Yang, 2019). kNN classification algorithm follows a five-step process: (a) select distance metric; (b) select number of nearest neighbours ($k < n$); (c) compute the distance from other data points to the desired point; (d) sort the points in increasing order of distance; (e) compute the average of k nearest neighbours' responses (Hastie *et al.*, 2009; Kuhn and Johnson, 2013; Sammut and Webb, 2011; Yang, 2019). In kNN, data is the model as it makes no assumptions about the relationship among the data and predicts output based on training data. kNN is simple, efficient and flexible in its speed and scalability, and since the training data in this problem will remain the same and predictions will be made based on that only, kNN is the most suitable algorithm for this problem (Sammut and Webb, 2011; Winters-Miner *et al.*, 2015). The methods for assignment of a factor to a source, as reported in Pernigotti and Belis (2018) and Liao *et al.* (2022), apply either distance-based indicators or a combined approach of distance-based proximity measures and probability-based matching algorithm (naïve Bayes classifier, NBC). kNN also applies distance metric to compute the distance from other data points to desired point. kNN has the option to utilize various distance metrics such as Euclidean, Manhattan and Minkowski. However, kNN has the option to either use uniform or distance-based weights. In the case of uniform weights, all the neighbouring points are weighted equally, while in distance-based weights, the closer neighbors will have a greater influence than points that are far away. kNN then classifies using a majority vote among the k neighbours (Hastie *et al.*, 2009; Kuhn and Johnson, 2013; Sammut and Webb, 2011; Yang, 2019). Another significant difference between kNN and NBC (applied by Liao *et al.* (2022)) is that NBC is a parametric model while kNN is a non-parametric model. NBC uses a fixed number of parameters for model building and considers strong assumptions about the data. Contrary to that, kNN is flexible with the number of parameters for building the model and considers fewer data assumptions. Also, NBC requires the underlying probability distributions for categories to obtain acceptable results and

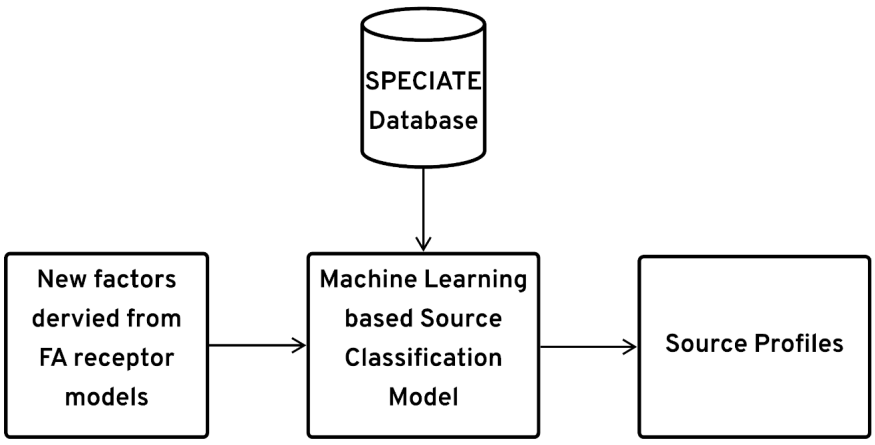

Fig. 1. Modeling framework for source classification model.

ORIGINAL RESEARCH
https://doi.org/10.4209/aaqr.220386

will only work if the decision boundary is linear, elliptic, or parabolic, while kNN does not require any such information and is often successful where the decision boundary is irregular (Hastie *et al.*, 2009; Kuhn and Johnson, 2013; Sammut and Webb, 2011; Yang, 2019). Further details of the kNN algorithm is available in the literature (Hastie *et al.*, 2009; Kuhn and Johnson, 2013; Sammut and Webb, 2011; Yang, 2019).

The KNeighborsClassifier module of the Sklearn Python library is used in this study for model development (Scikit-learn, 2011). The ML model should be evaluated on samples not used while building the model to assess the unbiased sense of model effectiveness (Kuhn and Johnson, 2013). Due to this, the data is divided into train and test data. The train data is used to develop the model, while the test data is used for evaluating the model's performance. However, if the test data is locked away, train data is further divided into the train and validation sets to measure performance on unseen data to select a good hypothesis (Kuhn and Johnson, 2013; Russell and Norvig, 2018). But, resampling methods such as cross-validation can also be used to assess the model performance using the training set (Kuhn and Johnson, 2013). Hyperparameter tuning using grid search with 10-fold cross-validation was conducted to determine the value of k in the kNN model. Based on the hyperparameter tuning results, the kNN model was trained for k = 5 (number of neighbours). The model takes around 6 seconds to train the model and provide the results on the test and validation datasets in approximately 3 seconds.

The accuracy of classification models can be quantified by precision, recall, F1 score and accuracy. Precision is the ability of the classifier not to label the sample as positive, a sample that is negative, while recall is the ability of the classifier to find all the positive samples. Precision and recall can be calculated by Eqs. (1) and (2),

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

where TP, FP, TN and FN are true positive (number of positive instances correctly classified), false positive (number of negative instances incorrectly classified), true negative (number of negative instances correctly classified) and false negative (number of positive instances misclassified) respectively. F1 score is the weighted harmonic mean of the precision and recall and can be calculated by Eq. (3).

$$F1\ score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3}$$

The F1 score varies from 0 (worst) to 1 (best) (Sammut and Webb, 2011). The accuracy of the model is calculated by Eq. (4).

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{4}$$

3 RESULTS AND DISCUSSION

3.1 Model Development

kNN is applied to the train data for model development and test data for validation. The train and test accuracy for the model is 0.85 and 0.79. The precision, recall and F1 score for the five major sources and the overall average on the test data is presented in Fig. 2. It is observed that the sample size significantly affects the model performance. The performance for the traffic source is the best as it has the highest number of samples, followed by dust and industrial sources. The precision of coal combustion is high, but the recall is low, reducing the F1 score.

ORIGINAL RESEARCH
https://doi.org/10.4209/aaqr.220386

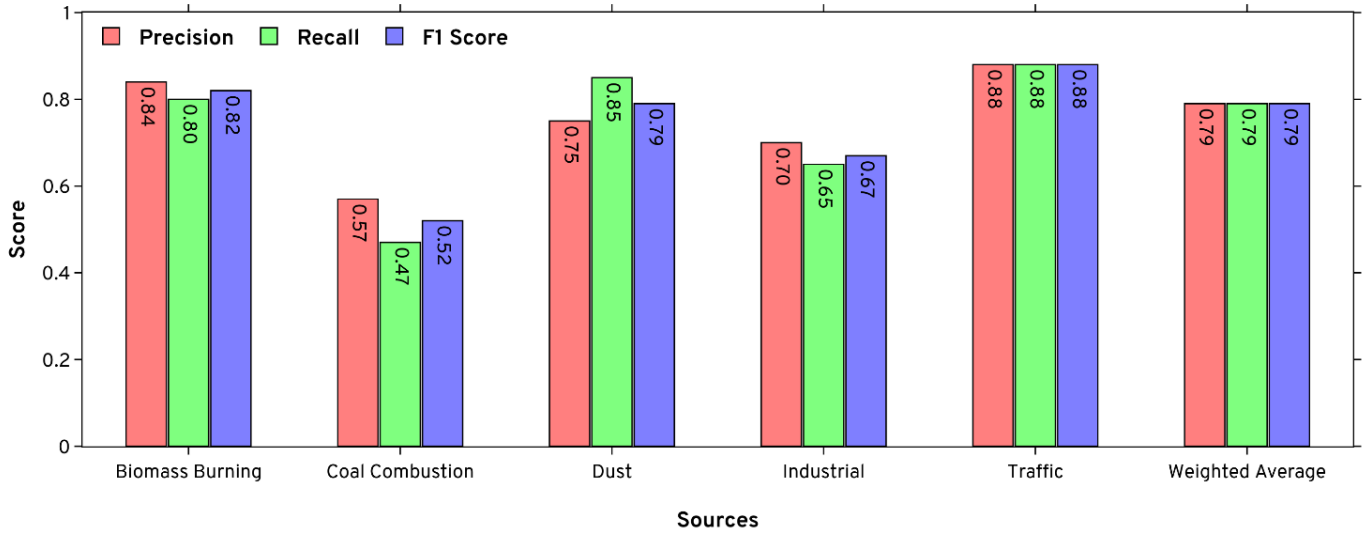

Fig. 2. Accuracy metrics for the sources and overall average.

Table 1. Confusion matrix.

		Predicted Label					
	Source	Biomass Burning	Coal Combustion	Dust	Industrial	Traffic	Count
True Label	Biomass Burning	86	1	5	5	9	106
	Coal Combustion	1	19	8	4	5	37
	Dust	2	6	111	12	1	132
	Industrial	3	5	17	59	4	88
	Traffic	9	0	6	4	138	157
	Count	101	31	147	84	157	520

The confusion matrix presented in Table 1 explains the model's performance for classifying source-wise samples. Out of 106 biomass burning samples in test data, only 101 are assigned as biomass burning by the model, out of which only 86 are correctly labelled. Similarly, only 19 and 111 out of 37 and 132 coal combustion and dust samples are assigned to the correct class. The low performance of the model for coal combustion is due to the lack of enough training samples to develop a robust theory about the source. For traffic source, 138 samples were assigned to the correct class by the model out of 157. Also, the true and predicted label count is equal for traffic source, which is not equivalent for other sources. This explains the effect of sufficient training data. The details of source-wise samples assigned to each class is available in Table 1.

3.2. Validation

The model developed above is applied to the source profiles obtained through measurement for Delhi, India (Prakash *et al.*, 2021) and the receptor model for Cincinnati, USA (Sahu *et al.*, 2011) that have been used for source apportionment of $PM_{2.5}$ in literature. The mass and the uncertainties associated with the species were used for prediction. The objective is to predict the label of the sources and compare them with those provided in the literature.

3.2.1 Measured source profiles

The measured source profiles for Delhi are taken from Prakash *et al.* (2021). Eight major sources are selected from this study for validation, viz., industrial (IND), coal combustion (CC), diesel vehicle (DV), gasoline vehicle (GV), construction dust (CD), soil dust (SD), road dust (RD) and biomass burning (BB). The elemental composition of the eight sources is presented in Fig. 3. The IND source is dominated by OC (50.31%) and a comparatively smaller quantity of EC (17.75%), Fe (10.65%)

ORIGINAL RESEARCH
https://doi.org/10.4209/aaqr.220386

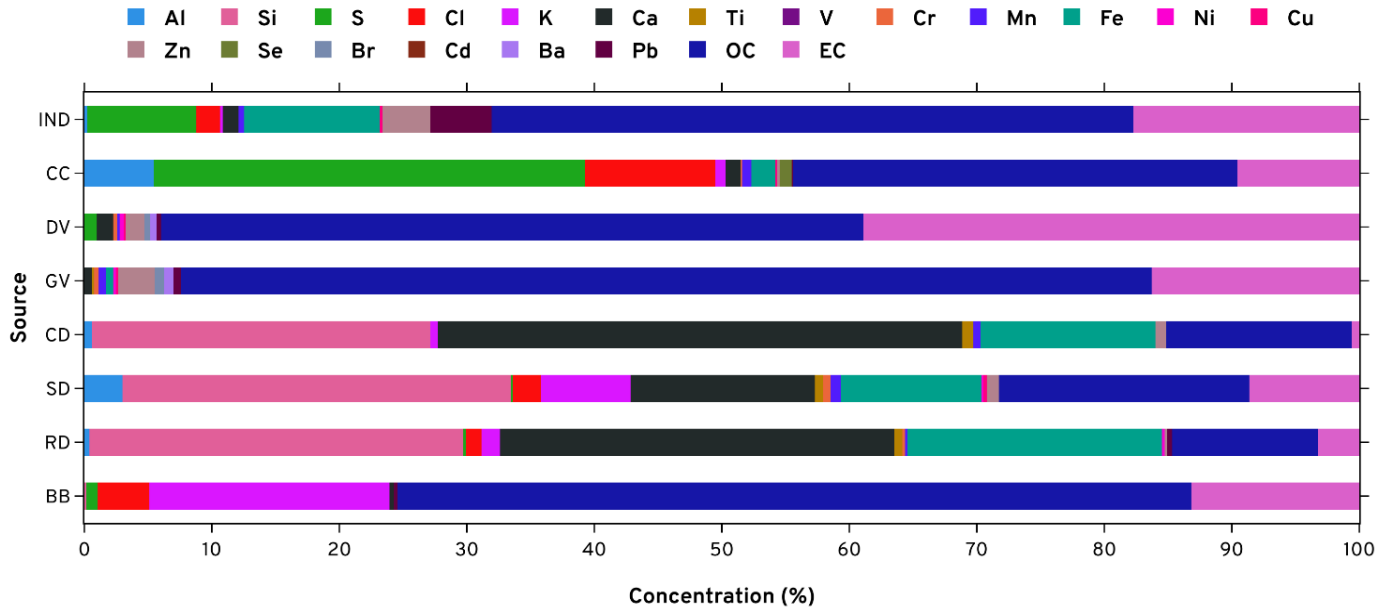

Fig. 3. Elemental composition of source profiles for Delhi.

and S (8.56%), while CC is dominated by S (34.84%) and OC (33.79%). The DV and GV exhibit a high concentration of OC and EC, with DV having higher EC than GV. DV has 55.11% and 38.90% OC and EC, respectively while GV has 76.14% and 16.29% OC and EC, respectively. The dust profiles CD, SD and RD, have high concentrations of Ca (41.14%, 14.44%, 30.96%) and Si (26.57%, 30.46%, 29.33%), followed by Fe (13.71%, 10.99%, 19.96%), OC and EC. BB has a high concentration of OC (62.28%), K (18.86%) and EC (13.17%).

The comparison of profiles for similarity indicates that IND is highly correlated with BB (0.91), GV (0.96), DV (0.92) and CC (0.76). Similarly, CC also has a high correlation with BB (0.96), GV (0.68) and DV (0.63). Other profiles also indicate similarities, as presented in Fig. 4. This demonstrates that even measured profiles have similarities, affecting the model performance as sources are not very well alienated from each other.

The probability matrix shows the possibility of the samples being assigned to a specific class. The probability matrix for the Delhi dataset is presented in Table 2. BB is assigned to the correct class of biomass burning with a probability of 0.6. RD, SD and CD were also allocated to the suitable class dust with probabilities of 0.4, 0.6, and 0.6, respectively. GV (0.8) and DV (0.8) are allotted to the traffic source. However, IND is labelled as a traffic source for two reasons. First, the number of traffic source samples is higher than any other source in training data, causing a bias in the model. Second, as discussed above also, IND, DV and GV are dominated by the same chemical species OC (50.31%, 55.11%, 76.14%) and EC (17.75%, 38.90%, 16.29%), and the IND source profile shows substantial similarity with DV (0.92) and GV (0.96), which can confuse the model.

3.2.2 Model derived source profiles

The PMF receptor model-derived source profiles for Cincinnati are taken from Sahu *et al.* (2011). Six major sources are selected from this study for validation, viz., soil dust (SD), metal processing (MP), biomass burning (BB), coal combustion (CC), gasoline vehicle (GV), and diesel vehicle (DV). Si (36.41%), Al (14.17%), Ca (6.32%) and Fe (10.83%) characterize SD, while MP has a relatively high concentration of Cu, Zn, Fe, and Pb. CC has the highest concentration of S (58%), while BB has an abundance of K. GV and DV have high OC and EC concentrations. The detailed elemental composition of the six sources is presented in Fig. 5.

The similarity between the source profiles for Cincinnati is presented in Fig. 6. DV and GV show a strong correlation (0.81) with each other and MP (0.65, 0.74). BB exhibits a good correlation with CC (0.43) and strong with MP (0.57). This indicates that model-derived profiles are similar to the measured profiles, affecting the model performance.

ORIGINAL RESEARCH
https://doi.org/10.4209/aaqr.220386

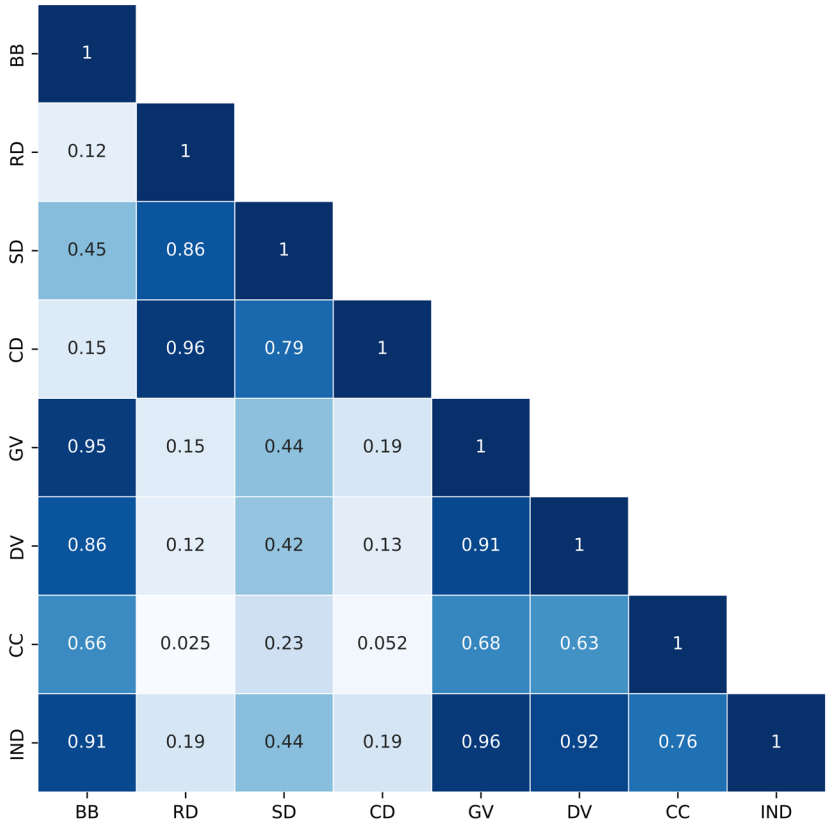

Fig. 4. Similarity between source profiles of Delhi.

Table 2. Probability matrix for the Delhi dataset.

		Probability of Predicted Source					Source Labelled
		BB	CC	Dust	Industrial	Traffic	
Actual	BB	0.6	0.2	0	0.2	0	Biomass Burning
Source	RD	0.2	0	0.4	0	0.4	Dust
	SD	0.2	0.2	0.6	0	0	Dust
	CD	0.2	0	0.6	0.2	0	Dust
	GV	0	0	0.2	0	0.8	Traffic
	DV	0.2	0	0	0	0.8	Traffic
	CC	0	0.4	0.2	0.4	0	Coal Combustion
	IND	0.2	0	0	0.2	0.6	Traffic

The probability matrix for the Cincinnati dataset presented in Table 3 assigns DV and GV to the correct class of traffic source. CC, BB and SD are also assigned to the suitable sources, but MP is appointed to the traffic source, possibly because of the same reasons discussed above for the measured profiles.

The ML approach presented in this study has its strengths and limitations. The strengths of this model are no human intervention and subjectivity, an efficient automated process, and the model can be coupled with receptor models easily for verification of results. However, the limitations are the requirement of huge datasets and model bias is there. However, this is different than human bias as it does not depend on the modeler and can be reduced over the period with more training.

This study attempted to identify the issues with the FA receptor models, which is a barrier to real-time SA. An ML-based classification model is developed to label the sources derived from FA receptor models automatically. The model's validation on measured and receptor model-derived source profiles provides evidence that the ML model performance is within the acceptable estimation range. The performance can be further increased by balancing the number of samples

ORIGINAL RESEARCH
https://doi.org/10.4209/aaqr.220386

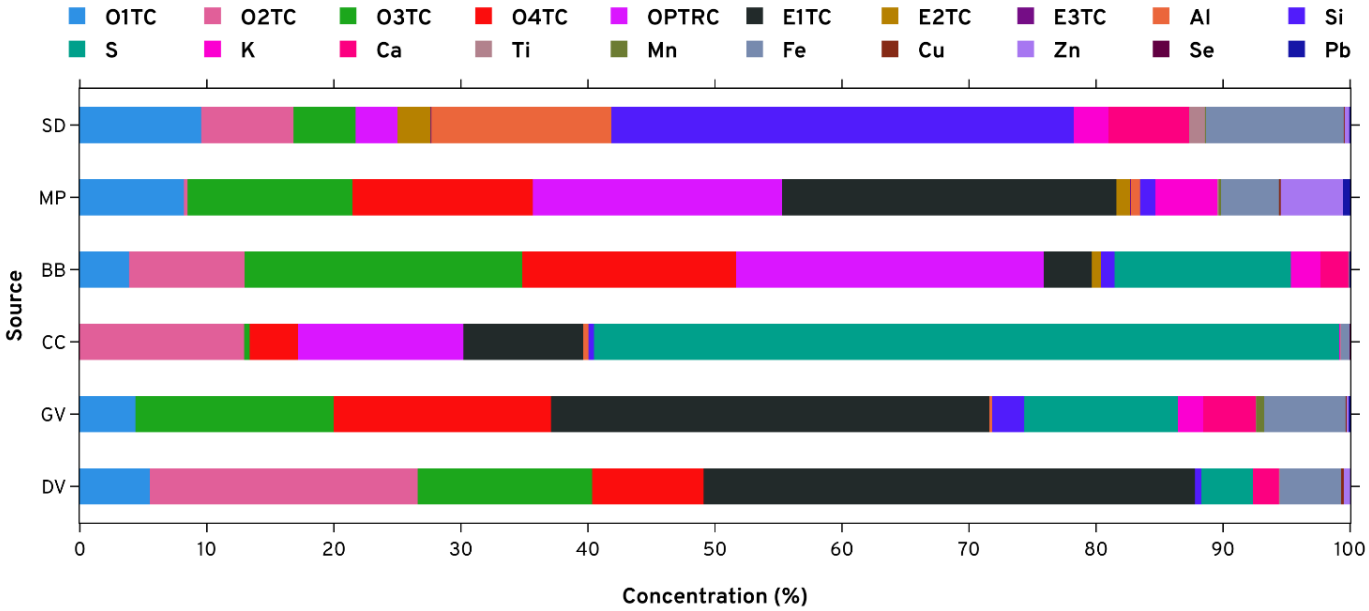

Fig. 5. Elemental composition of source profiles for Cincinnati.

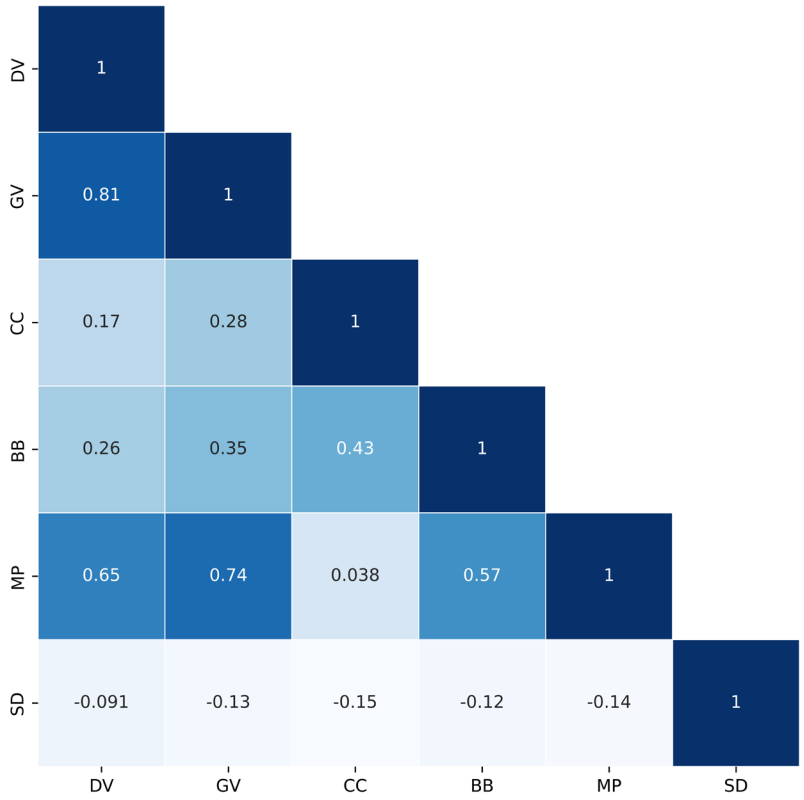

Fig. 6. Similarity between source profiles of Cincinnati.

for each source in the training data. This model can act as another layer of the process for verification of the results of FA receptor models and is a small step towards automation of the process for real-time SA. The aforementioned being said, we acknowledge that in this study, no secondary sources are used while prediction due to a lack of training data and only $PM_{2.5}$ source profiles are used. The same framework can be transferred to other pollutant sources with relevant data.

Table 3. Probability matrix for the Cincinnati dataset.

		Probability of Predicted Source					Source Labelled
		Biomass Burning	Coal Combustion	Dust	Industrial	Traffic	
Actual Source	DV	0.2	0	0	0	0.8	Traffic
	GV	0	0	0	0.4	0.6	Traffic
	CC	0	0.4	0	0.2	0.4	Coal Combustion
	BB	0.4	0.4	0	0.2	0	Biomass Burning
	MP	0	0	0	0.4	0.6	Traffic
	SD	0.2	0	0.8	0	0	Dust

4 CONCLUSION

This study implemented k- Nearest Neighbour (kNN), a classification machine learning (ML) algorithm for the automatic labelling of profiles derived from factor analysis (FA) receptor models based on the SPECIATE database. The train and test score of the model is 0.85 and 0.79, respectively. The overall weighted average precision, recall and F1 score is 0.79. The performance of the model during validation exhibits acceptable results. The application of ML models for source profile labelling will reduce the time taken and the subjectivity associated with results due to modeler bias. This process can act as another layer of the process for verification of the results of FA receptor models. The application of this methodology advances the process towards real-time source apportionment.

ACKNOWLEDGMENTS

The authors would like to thank and acknowledge the United States Environmental Protection Agency (US EPA) for making the SPECIATE data available for public use.

DATA AVAILABILITY

The data that support the findings of this study are available open and free to the public and can be accessed from the United States Environmental Protection Agency's (U.S. EPA) SPECIATE website (U.S. EPA, 2015).

SUPPLEMENTARY MATERIAL

Supplementary material for this article can be found in the online version at https://doi.org/10.4209/aaqr.220386

REFERENCES

Bove, M.C., Brotto, P., Cassola, F., Cuccia, E., Massabò, D., Mazzino, A., Piazzalunga, A., Prati, P. (2014). An integrated $PM_{2.5}$ source apportionment study: positive matrix factorisation vs. the chemical transport model CAMx. Atmos. Environ. 94, 274–286. https://doi.org/10.1016/j.atmosenv.2014.05.039

Bray, C.D., Strum, M., Simon, H., Riddick, L., Kosusko, M., Menetrez, M., Hays, M.D., Rao, V. (2019). An assessment of important SPECIATE profiles in the EPA emissions modeling platform and current data gaps. Atmos. Environ. 207, 93–104. https://doi.org/10.1016/j.atmosenv.2019.03.013

Hastie, T., Tibshirani, R., Friedman, J. (2009). The Elements of Statistical Learning. Springer New York, New York, NY. https://doi.org/10.1007/978-0-387-84858-7

Hopke, P.K., Cohen, D.D. (2011). Application of receptor modeling methods. Atmos. Pollut. Res.

2, 122–125. https://doi.org/10.5094/apr.2011.016

Hopke, P.K. (2016). Review of receptor modeling methods for source apportionment. J. Air Waste Manage. Assoc. 66, 237–259. https://doi.org/10.1080/10962247.2016.1140693

Hopke, P.K., Dai, Q., Li, L., Feng, Y. (2020). Global review of recent source apportionments for airborne particulate matter. Sci. Total Environ. 740, 140091. https://doi.org/10.1016/j.scitotenv.2020.140091

Karagulian, F., Belis, C.A. (2012). Enhancing source apportionment with receptor models to foster the air quality directive implementation. Int. J. Environ. Pollut. 50, 190. https://doi.org/10.1504/ijep.2012.051192

Karagulian, F., Belis, C.A., Dora, C.F.C., Prüss-Ustün, A.M., Bonjour, S., Adair-Rohani, H., Amann, M. (2015). Contributions to cities' ambient particulate matter (PM): A systematic review of local source contributions at global level. Atmos. Environ. 120, 475–483. https://doi.org/10.1016/j.atmosenv.2015.08.087

Kuhn, M., Johnson, K. (2013). Applied Predictive Modeling. Springer, New York. https://doi.org/10.1007/978-1-4614-6849-3

Lalchandani, V., Kumar, V., Tobler, A., Thamban, N.M., Mishra, S., Slowik, J.G., Bhattu, D., Rai, P., Satish, R., Ganguly, D., Tiwari, S., Rastogi, N., Tiwari, S., Močnik, G., Prévôt, A.S.H., Tripathi, S. N. (2021). Real-time characterization and source apportionment of fine particulate matter in the Delhi megacity area during late winter. Sci. Total Environ. 770, 145324. https://doi.org/10.1016/j.scitotenv.2021.145324

Liao, H.T., Hsieh, P.Y., Hopke, P.K., Wu, C.F. (2022). Development and evaluation of an integrated method using distance- and probability-based profile matching approaches in receptor modeling. Atmos. Pollut. Res. 13, 101423. https://doi.org/10.1016/j.apr.2022.101423

Pernigotti, D., Belis, C.A. (2018). DeltaSA tool for source apportionment benchmarking, description and sensitivity analysis. Atmos. Environ. 180, 138–148. https://doi.org/10.1016/j.atmosenv.2018.02.046

Prakash, J., Choudhary, S., Raliya, R., Chadha, T.S., Fang, J., Biswas, P. (2021). Real-time source apportionment of fine particle inorganic and organic constituents at an urban site in Delhi city: An IoT-based approach. Atmos. Pollut. Res. 12, 101206. https://doi.org/10.1016/j.apr.2021.101206

Rai, P., Furger, M., El Haddad, I., Kumar, V., Wang, L., Singh, A., Dixit, K., Bhattu, D., Petit, J.E., Ganguly, D., Rastogi, N., Baltensperger, U., Tripathi, S.N., Slowik, J.G., Prévôt, A.S.H. (2020). Real-time measurement and source apportionment of elements in Delhi's atmosphere. Sci. Total Environ. 742, 140332. https://doi.org/10.1016/j.scitotenv.2020.140332

Reff, A., Eberly, S.I., Bhave, P.V. (2007). Receptor modeling of ambient particulate matter data using positive matrix factorization: review of existing methods. J. Air Waste Manage. Assoc. 57, 146–154. https://doi.org/10.1080/10473289.2007.10465319

Russell, S.J., Norvig, P. (2018). Artificial intelligence: a modern approach. Pearson India Education Services Pvt. Ltd., Noida, India.

Sahu, M., Hu, S., Ryan, P.H., Le Masters, G., Grinshpun, S.A., Chow, J.C., Biswas, P. (2011). Chemical compositions and source identification of $PM_{2.5}$ aerosols for estimation of a diesel source surrogate. Sci. Total Environ. 409, 2642–2651. https://doi.org/10.1016/j.scitotenv.2011.03.032

Sammut, C., Webb, G.I. (2011). Encyclopedia of Machine Learning. Springer, New York. https://doi.org/10.1007/978-0-387-30164-8

Scikit-learn (2011). Nearest Neighbors Classification. https://scikit-learn.org/stable/modules/neighbors.html#nearest-neighbors-classification (accessed 23 March 2022).

Simon, H., Beck, L., Bhave, P.V., Divita, F., Hsu, Y., Luecken, D., Mobley, J.D., Pouliot, G.A., Reff, A., Sarwar, G., Strum, M. (2010). The development and uses of EPA's SPECIATE database. Atmos. Pollut. Res. 1, 196–206. https://doi.org/10.5094/apr.2010.026

U.S. Environmental Protection Agency (U.S. EPA) (2015). SPECIATE. https://www.epa.gov/air-emissions-modeling/speciate (accessed 12 December 2021).

U.S. Environmental Protection Agency (U.S. EPA) (2019). SPECIATE Version 5.0 Database Development Documentation Final Report. https://www.epa.gov/sites/default/files/2019-07/documents/speciate_5.0.pdf (accessed 05 December 2021).

Viana, M., Pandolfi, M., Minguillón, M.C., Querol, X., Alastuey, A., Monfort, E., Celades, I. (2008).

Inter-comparison of receptor models for PM source apportionment: case study in an industrial area. Atmos. Environ. 42, 3820–3832. https://doi.org/10.1016/j.atmosenv.2007.12.056

Winters-Miner, L.A., Bolding, P., Hill, T., Nisbet, B., Goldstein, M., Hilbe, J.M., Walton, N., Miner, G., Rastunkov, V., Stout, D. (2015). Chapter 15 - Prediction in Medicine – The Data Mining Algorithms of Predictive Analytics, in: Winters-Miner, L.A., Bolding, P.S., Hilbe, J.M., Goldstein, M., Hill, T., Nisbet, R., Walton, N., Miner, G.D. (Eds.), Practical Predictive Analytics and Decisioning Systems for Medicine, Academic Press, pp. 239–259. https://doi.org/10.1016/B978-0-12-411643-6.00015-6

Yang, X.S. (2019). Introduction to algorithms for data mining and machine learning. Academic Press. https://doi.org/10.1016/C2018-0-02034-4

Yang, X., Zheng, M., Liu, Y., Yan, C., Liu, J., Liu, J., Cheng, Y. (2022). Exploring sources and health risks of metals in Beijing $PM_{2.5}$: insights from long-term online measurements. Sci. Total Environ. 814, 151954. https://doi.org/10.1016/j.scitotenv.2021.151954