Source Apportionment of Particulate Matter by Application of Machine Learning Clustering Algorithms

Vikas Kumar¹, Manoranjan Sahu^{2,1,3*}, Pratim Biswas⁴

¹Interdisciplinary Program in Climate Studies, Indian Institute of Technology Bombay, Mumbai 400076, India

²Aerosol and Nanoparticle Technology Laboratory, Environmental Science and Engineering Department, Indian Institute of Technology Bombay, Mumbai 400076, India

³Centre for Machine Intelligence and Data Science, Indian Institute of Technology Bombay, Mumbai 400076, India

⁴Aerosol and Air Quality Research Laboratory, University of Miami, College of Engineering, Coral Gables, FL 33146, USA

ABSTRACT

A source apportionment (SA) study was conducted on two PM_{2.5} data sets, two carbon fractions and eight temperature-resolved carbon fractions collected during Cincinnati Childhood Allergy and Air Pollution Study (CCAAPS). This study aimed to evaluate two clustering algorithms: k-means clustering (kMC) and spectral clustering (SC) as potential receptor models for source apportionment. The application of kMC produced unsatisfactory results, but the results obtained from SC demonstrated a significant correlation with the results obtained using positive matrix factorization (PMF). The clustering results obtained were associated with practical evidence available in the literature. SC identified six source factors on analyzing two carbon fractions data set and seven factors from eight temperature-resolved carbon fractions data set. The sources (source contribution in parentheses) identified are: combustion (45.9 ± 3.66%) and secondary sulfate (11.4 ± 1.09%), vegetative/wood burning (17.5 ± 1.46%), diesel (10.6 ± 0.92%) and gasoline (3.6 ± 0.33%) vehicles, soil/crustal (2.07 ± 0.2%), traffic (9.3 ± 0.81%), and metal processing (8.8 ± 0.72%). The source profiles obtained using SC also show similarity with the profiles derived using PMF. In summary, this study presented a basic framework for applying Machine Learning algorithms for SA analysis. Also, it presents SC as a potential receptor model technique for SA.

Keywords: PM_{2.5}, Source apportionment, Receptor modeling, Positive matrix factorization, Machine learning, Clustering algorithms

1 INTRODUCTION

Receptor modeling for source apportionment (SA) is an integral element of contemporary aerosol research. Receptor models aim to identify emission sources and apportion their respective contributions to aerosol quantity and composition. The information obtained from SA is critical to evolving existing and developing novel control strategies or legislation for effective air quality management (Begum *et al.*, 2007; Hopke and Cohen, 2011; Karagulian; 2012; Bove *et al.*, 2014; Hopke, 2016). Viana *et al.* (2008) compared receptor models available in literature based on the extent of information required about emission sources and their profile before implementing the model to get credible results. In the comparison, chemical mass balance (CMB) and multivariate factor analysis (FA) are at the two extremes of the spectrum. CMB is the most appropriate model and provides the most objective result if the aerosol sources' details and profiles for a local region are available. In most cases, the local source profiles are not available and applying profiles that

OPEN ACCESS

Received: October 17, 2021
Revised: December 22, 2021
Accepted: January 27, 2022

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

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

Copyright: The Author(s).
This is an open access article
distributed under the terms of the
[Creative Commons Attribution
License \(CC BY 4.0\)](https://creativecommons.org/licenses/by/4.0/), which permits
unrestricted use, distribution, and
reproduction in any medium,
provided the original author and
source are cited.

do not adequately reflect the emission sources can provide inaccurate SA results (Hopke, 2016). So in the absence of source profiles, multivariate models such as Positive Matrix Factorization (PMF) play a pivotal role in SA (Ramadan *et al.*, 2003; Viana *et al.*, 2008; Hopke and Cohen, 2011; Hopke, 2016). The PM_{2.5} speciated physico-chemical properties and uncertainty associated with each species collected from *in-situ* measurements are used by PMF and other receptor models for SA. PMF decomposes the PM_{2.5} matrix as a product of source pattern and contribution matrices and searches for correlation between chemical species. PMF assumes that constituents emitted from the same source exhibit high correlation unique to that source to estimate the number, profile, and contribution of the sources from the ambient data itself (Paatero and Tapper, 1994; Paatero, 1997; Watson *et al.*, 2002; Ramadan *et al.*, 2003; Caselli *et al.*, 2006; Viana *et al.*, 2008; Hopke and Cohen, 2011; Karagulian, 2012; Hopke, 2016). PMF is the most widely used receptor model for SA (61.4%) during 1990–2019, as illustrated in Fig. S1 in the Supporting Information (Karagulian *et al.*, 2015; Hopke *et al.*, 2020). Other receptor models and their application is also mentioned in Fig. S1.

Although traditional receptor models such as PMF (Yue *et al.*, 2008; Harrison *et al.*, 2011; Dall'Osto *et al.*, 2012; Tan *et al.*, 2014; Liang *et al.*, 2020; Liang *et al.*, 2021) and k-means clustering (Charron *et al.*, 2008; Beddows *et al.*, 2009; Dall'Osto *et al.*, 2012; Wegner *et al.*, 2012; Salimi *et al.*, 2014; Liang *et al.*, 2020) have been applied for SA using particle number concentration, no attempts have been made previously to apply clustering algorithms on mass concentration data. In this study, we have implemented k-means clustering (kMC) as well as Spectral Clustering (SC) algorithms which are a part of Machine Learning (ML), on mass concentration data. ML executes mathematical models based on statistical theory in the form of computer programs. ML model optimizes the performance criterion or pre-defined parameters based on past data. The fundamental difference between traditional and ML programs is that the ML model can learn from the data and adapt to the changing environment, making it part of artificial intelligence. This is advantageous as the modeler does not need to provide the solution for all the possible scenarios (Alpaydin, 2020). ML algorithms' unique learning and adaptability feature can be critical for SA as the chemical composition of PM_{2.5} is not constant and varies from region to region. ML has a wide range of algorithms in its arsenal for various types of data and problems (Alpaydin, 2020).

Clustering, an unsupervised ML algorithm, is of interest for this study due to its ability to extract hidden patterns and information from the data itself, similar to FA models. Clustering is a significant element in machine learning and data mining processes and has been employed extensively to address issues in various disciplines, viz. psychology, biology, applied sciences, computer science, and so forth (Rokach and Maimon, 2005; Gan *et al.*, 2007; Higham *et al.*, 2007; Von Luxburg, 2007; Filippone *et al.*, 2008; Aggarwal and Reddy, 2014). PMF selects sources based on the trial-and-error method and depends on the researcher to choose the appropriate number of sources based on results obtained and practical evidence. The clustering process also falls in the same line. However, certain techniques are available to initialize the number of clusters (sources) from the data itself, which can help analyze new data sets.

The following objectives are addressed in this paper: (a) to investigate the feasibility of clustering as a potential receptor model for SA and (b) to compare the source contributions and profiles derived using clustering algorithms with the results obtained using PMF using two distinct datasets, viz. two carbon (2C) and eight temperature-resolved carbon (8C) fractions. In this study, two clustering algorithms were used namely k-Means Clustering (kMC) and Spectral Clustering (SC), which have been discussed in detail in the methodology section. The primary source contributions and profiles derived from both the clustering algorithms were compared for both datasets.

2 MATERIALS AND METHODS

2.1 Data Description

The PM_{2.5} and associated chemical composition data used in this study were collected at 11 sampling sites of the CCAAPS network. The exact sampling method and chemical composition analysis can be found in the previous research papers on this study (Hu *et al.*, 2006; Hu, 2007; Sahu *et al.*, 2011). The descriptive statistics of two distinct datasets used in this study viz. 2C and 8C are described in Table S1 in the Supporting Information. The EC and OC for the 2C dataset

were estimated by the National Institute for Occupational Safety and Health (NIOSH) method that uses the thermal optical transmittance (TOT) technique. The temperature resolved carbon's (O1TC, O2TC, O3TC, O4TC, OPTRC, E1TC, E2TC, and E3TC) for the 8C dataset were evaluated as per the Interagency Monitoring of Protected Visual Environments (IMPROVE) thermal optical reflectance (TOR) protocol. The OC fractions O1TC, O2TC, O3TC, and O4TC, were measured at 120°C, 250°C, 450°C, and 550°C, respectively in a 100% helium (He) atmosphere. The EC fractions were measured at 550°C, 700°C, and 800°C for E1TC, E2TC, and E3TC, respectively, in a 98% He and 2% oxygen mixture. The pyrolyzed organic carbon (OPTRC) was measured following laser response. The temperature resolved carbon fractions help differentiate diesel and gasoline vehicle sources based on their source signatures.

2.2 Receptor Modelling

2.2.1 Basic framework

Receptor models are based on the principle of mass conservation and decompose a matrix of PM_{2.5} speciation data with the purpose of (a) identifying the sources and their composition (source profile), (b) contribution of each source to ambient concentration (Watson *et al.*, 2002; Ramadan *et al.*, 2003; Viana *et al.*, 2008; Hopke and Cohen, 2011; Karagulian; 2012; Hopke, 2016). The receptor models must adhere to the natural physical constraints imposed on the model, which are: (a) neither a source contribution nor any element in source composition can be negative, (b) the model must reproduce the data, (c) the total predicted mass contribution of aerosol species for each source could not be higher than the measured mass for each species. For m elements in n samples for p independent sources, the mass balance equation can be written as:

$$x_{ij} = \sum_{k=1}^p g_{ik} f_{kj} + e_{ij} \quad (1)$$

where x_{ij} is the i^{th} concentration measured in the j^{th} sample, g_{ik} is the concentration of i^{th} parameter in the j^{th} sample, f_{kj} is the contribution from the k^{th} source to the j^{th} sample, and e_{ij} is the residue.

In matrix form,

$$X = FG + E \quad (2)$$

where X ($n \times m$) is the aerosol samples measured concentration matrix, F ($n \times p$) is the source contribution matrix, G ($p \times m$) is the source profile matrix, and E is the residue matrix.

2.2.2 Clustering

Clustering is a sequence of procedures for partitioning unlabelled data in groups (or clusters) with no prior information. In clustering, the objects in the same cluster exhibit similarity while objects in different clusters are quite contrasting (Gan *et al.*, 2007; Xu and Wunsch, 2008; Clarke *et al.*, 2009; Aggarwal and Reddy, 2014; Alpaydin, 2020). Clustering is different from classification and regression, which are supervised learning algorithms. Classification and regression require two variables, i.e., target/dependent and independent. In regression, the variables are real-valued continuous variables, while in classification, the variables are categorical (Sammur and Webb, 2011; Mendenhall and Sincich, 2014). Contrary to that, clustering assumes no output or target variable and only input data known as unlabelled data. Clustering relies primarily on the underlying or hidden structure of the data for exploration and investigation to discover and extract useful information. Due to its exploratory nature, clustering is suitable for this study as the source details are not available since they vary from widely and has to be extracted from the PM_{2.5} matrix, which is the only information available to the modelers, similar to FA models (Chen *et al.*, 1993; Gan *et al.*, 2007; Clarke *et al.*, 2009; Olivas *et al.*, 2009; Sammut and Webb, 2011). As illustrated in Fig. 1, clustering is an iterative process, including several steps, and often requires the repetition of certain steps to obtain satisfactory results. The major steps involved in the clustering process are briefly discussed in the Supporting Information.

Fig. 1. Overview of the steps in the clustering process (Fayyad *et al.*, 1996; Jain *et al.*, 1999; Halkidi *et al.*, 2001; Xu and Wunsch, 2008).

2.2.3 k-means clustering

kMC is one of the most popular, fundamental, and comprehensively used partitional clustering algorithms because of its simple implementation, speed, and memory efficiency (MacQueen, 1967; Lloyd, 1982; Gan *et al.*, 2007; Xu and Wunsch, 2008; Aggarwal and Reddy, 2014). Partitional clustering algorithms are iterative in nature and attempts to construct partitions (or clusters) in the data while minimizing the objective or cost function. kMC minimizes the sum of square error (SSE) for the given set of centroids through the minimum distance rule according to which entities are assigned to the cluster whose centroid is nearest to them (called error) calculated using Euclidean distance (Xu and Wunsch, 2005; Olivas *et al.*, 2009; Sammut and Webb, 2011; Morissette and Chartier, 2013). Further details about the algorithm are presented in Supporting Information.

2.2.4 Spectral clustering

Contrary to the conventional clustering algorithms, SC is based on graph theory, emphasizing the Laplacian matrix's eigenvalue decomposition to partition or cluster the data. The Laplacian matrix is generated from the weighted graph. (Bach and Jordan, 2004; McSherry, 2004; Gan *et al.*, 2007; Clarke *et al.*, 2009; Aggarwal and Reddy, 2014; Celebi and Aydin, 2016).

SC can outperform traditional clustering algorithms such as the kMC, expectation minimization (EM) which assumes the data to have a spherical or elliptical distribution, whereas SC makes no such assumptions (Von Luxburg, 2007; Aggarwal and Reddy, 2014). SC's basic framework applies eigenvalue decomposition, similar to UNMIX/PMF (Lewis *et al.*, 2003). SC shares the same objective function as PMF with slightly varying constraints and optimization methods. PMF expects F , a matrix constructed by the eigenvectors of the Laplacian matrix $L(G)$, to be non-negative while SC assumes F to be orthonormal (Merris, 1994; Liu and Han, 2013; Aggarwal and Reddy, 2014) which justifies the application of SC. Further details about the algorithm are presented in Supporting Information.

3 RESULTS AND DISCUSSION

3.1 Clustering

The results obtained from the kMC algorithm for 2C and 8C datasets are shown in Figs. S3(a) and S3(b) respectively, in the Supporting Information. It can be visualized from the kMC results that the clusters' data points encircle around the centroid of the clusters. The source contributions and profiles obtained from kMC were not promising compared with the results obtained using PMF which is the guiding principle for this study. The speculation for the inaccuracy in kMC results could be the assumption that the input data has a spherical distribution that might not necessarily be accurate for this kind of real-world data. The data's high dimension adds another layer of complexity to the problem that kMC cannot handle.

The results obtained from kMC are not explained further and only presented while comparing SC and PMF results. The other algorithm SC, produced results that display significant resemblance with the SA results obtained by Sahu *et al.* (2011) and are shown in Figs. S4(a) and S4(b) for 2C and 8C datasets respectively in the Supporting Information. Optical analysis of the SC results

indicates that the cluster's data points are much more scattered than the datapoints of kMC and do not exhibit any specific shape or distribution.

Using validity indices with contrasting scales is always good to effectively understand the clustering results. A higher CH index represents a good clustering while vice versa for the DB index. We can compare different clustering results and select those that provide a good value for both indexes. No specific CH index is defined to decide whether the clustering result is good or bad, especially for this kind of data. So, it becomes essential to compare the index results obtained from other clustering algorithms to get a comparative idea. The validity indices' values were calculated for this study and tabulated in Table S2 in the Supporting Information. The CH index is higher for both datasets when SC is applied compared to kMC. The CH index is highest (468.14) for 8C when SC is applied, while the DB index (0.19) is the lowest, making it the best result.

3.2 Sources Identified

The objective of clustering algorithms is to extract applicable information from the data to assist in solving problems. Thus, it is vital to associate the clustering results with the practical evidence of the domain available in the literature. The objective of this study is to extract information about emission sources, profiles, and their contribution. Among the sources identified, five sources are common for both datasets, i.e., combustion sulfate, secondary sulfate, vegetative/wood burning, soil/crustal, metal processing. 2C has a traffic source, while for 8C two additional sources namely diesel and gasoline vehicles were identified using eight temperature resolved carbon fractions which are thermal sensitive volatility fractions of carbon materials for differentiating diesel and gasoline vehicle sources. A summary of the sources and their respective contributions identified using SC is presented in Fig. 2 and explained briefly below.

3.2.1 Sulfate sources

One cluster for each dataset with the highest loadings of S, high OC, EC and traces of Se was assigned as the combustion sulfate sources. Another cluster for each dataset with the abundance of S and OC and traces of Si, K, and Ca was attributed to secondary sulfate sources (Kim *et al.*, 2003; Lewis *et al.*, 2003; Lee *et al.*, 2008; Chen *et al.*, 2011). PM_{2.5} contributions proportioned from combustion and secondary sulfate sources are 45.9% and 11.7% for 2C and 45.9% and 11.4% for 8C respectively. The combustion and secondary sulfate sources contribute 65.2% and 16.4% of the total sulfur in PM_{2.5} concentration. The sum of the contribution from both sources, 54.7% for 8C is consistent with the value (56.9%) reported by Sahu *et al.* (2011).

3.2.2 Metal processing

A cluster identified by SC with a relatively high concentration of Cu, Zn, Fe, and Pb was attributed as metal processing (Kim *et al.*, 2003, 2004, 2005). The source contributes 4.2% and 8.8% for 2C and 8C respectively.

3.2.3 Vegetative/wood-burning

The vegetative/wood-burning source was identified by an abundance of K, OC, and EC (Turn *et al.*, 1997; Song *et al.*, 2001; Pozza *et al.*, 2006; Belis *et al.*, 2013) and contributed 27.2% and 17.5% for 2C and 8C respectively.

Fig. 2. Source contributions obtained using SC for 2C and 8C data set.

to ambient PM_{2.5} concentration for 2C and 8C respectively. The source is a major producer of potassium and accounts for 45.9% of the total concentration.

3.2.4 Soil/crustal

The presence of Al, Si, Ca, Fe and K characterized soil/crustal source (Polissar *et al.*, 2001; Liu *et al.*, 2003; Kim *et al.*, 2004; Amato *et al.*, 2014) and arise from construction sites, unpaved roads, natural wind blow, resuspended traffic dust, etc. contributing 2.07% of total PM_{2.5}.

3.2.5 Traffic sources

A cluster with a high mass fraction of OC and EC and traces of Pb, Mn, Fe, and Zn was determined as traffic source (Ramadan *et al.*, 2000; Chow and Watson, 2002; Viana *et al.*, 2008; Amato *et al.*, 2009; Jeong *et al.*, 2011) for 2C dataset. The traffic source contributes 9.38% of total ambient PM_{2.5}. It includes the contribution from numerous activities and processes apart from exhaust emissions like the clutch, tyre, and brake linings wear, which impart trace elements like Mn, Fe, Zn, etc. For the 8C dataset, the traffic source was separated into diesel and gasoline, using temperature resolved carbon fractions. The diesel and gasoline vehicle sources were identified based on high EC and OC factors and contributed 10.6% and 3.6% respectively. The trace elements for these sources have the same origin as the traffic source. The EC/OC ratio for diesel and gasoline vehicles in this study are 0.65 and 0.37 respectively. The total traffic contribution for 8C (14.2%) is consistent with the value (16.5%) reported by Sahu *et al.* (2011).

It is important to outline that SC not only gives better results than kMC but also provides accurate results for the 8C dataset compared to 2C. The ability to perform better on datasets that have a large number of samples and features is an added advantage of SC as most models fail to accommodate higher dimensional data due to the computational complexity it brings along. This makes SC a suitable and effective receptor model for SA as the data has a high number of variables and are of different kinds with varying ranges and properties.

3.3 Comparison with the Previous PMF Modeling Study

It is always recommended to compare the clustering results with external results, if available, to understand the model's capabilities thoroughly. The source contribution results obtained from SC were compared with kMC and PMF as presented in Table 1. However, this study only compares the results of clustering algorithms with PMF and not their respective methodologies. The SA results provided by SC are close to the source contributions apportioned by PMF which is not the case with the kMC. The kMC results are quite equally distributed among certain sources, leading to inappropriate interpretations as it is known that certain sources produce more emissions than others. PMF is a standard receptor model for SA, and a good correlation with its results indicates the caliber of SC as a potential receptor model. The source profiles acquired on applying SC on 2C and 8C datasets were compared with the PMF profiles by Sahu *et al.* (2011). The comparison profiles are illustrated in Figs. 3 and 4 respectively and discussed below briefly. The percentage deviation of the source contribution by SC is calculated on the mass of PM_{2.5} contributed by the source. The absolute deviation is calculated to address the error, so the deviations are positive.

Table 1. Comparison of SC derived source contribution with the previous studies.

Source	2C			8C		
	Sahu <i>et al.</i> (2011)	This study		Sahu <i>et al.</i> (2011)	This study	
	PMF	kMC	SC	PMF	kMC	SC
Soil/crustal	0.6 ± 0.4	3.8 ± 0.34	1.5 ± 0.14	0.7 ± 0.4	5.07 ± 0.47	2.07 ± 0.2
Vegetative/wood burning	25.3 ± 1.7	22.9 ± 1.91	27.2 ± 2.28	15.7 ± 1.0	22.5 ± 1.88	17.5 ± 1.46
Combustion sulfate	46.1 ± 0.6	23.1 ± 1.89	45.9 ± 3.72	46.8 ± 0.7	26.2 ± 2.1	45.9 ± 3.66
Secondary sulfate	10.8 ± 1.3	22.9 ± 1.91	11.7 ± 1.04	12.9 ± 0.5	15.8 ± 1.35	11.4 ± 1.09
Metal processing	7.7 ± 0.4	16.4 ± 1.4	4.22 ± 0.38	7.5 ± 0.3	11.87 ± 1.0	8.8 ± 0.72
Diesel vehicle				10.9 ± 1.9	13.2 ± 1.1	10.6 ± 0.92
Gasoline vehicle				5.6 ± 1.4	5.3 ± 0.46	3.6 ± 0.33
Total Traffic	9.5 ± 2.4	10.1 ± 0.91	9.38 ± 0.81	16.5	18.5	14.2

Fig. 3. Comparison of source profiles obtained from PMF (green) and SC (white) for 2C data set.

For the 8C dataset, metal processing source contribution depicts a difference of 1.3% between PMF and SC results. The diesel and gasoline source contributions show a difference of 0.3% and 2% between PMF and SC results. Source contributions from PMF and SC for vegetative/wood burning sources show an error of 1.8%. The difference between contributions apportioned by PMF and SC for combustion and secondary sulfate sources is 0.9% and 1.5% respectively. Soil/crustal source shows a variation of 1.4% in source contribution. For the 2C dataset, vegetative/wood burning, soil/crustal, metal processing and traffic sources exhibit a deviation of 1.6%, 0.9%, 3.5% and 0.1% respectively. For the combustion and secondary sulfate sources, the variation between

Fig. 4. Comparison of source profiles obtained from PMF (green) and SC (white) for 8C data set.

PMF and SC for the 2C dataset is 0.2% and 0.9% respectively. A statistical t-test was conducted to compare the source contributions derived by PMF and SC. The results indicate a significant difference in the source contribution for soil/crustal ($p < 0.05$). There is no significant difference in the source contribution apportioned by PMF and SC for the 8C dataset for all the other sources. For the 2C dataset, soil/crustal and metal processing source contributions derived by PMF and SC exhibit a significant difference ($p < 0.05$). This indicates that for the 8C dataset with more features, there is an improvement in source contribution estimation. The elemental concentrations in the source profiles apportioned by SC and PMF are very similar. For the 8C dataset, the Pearson's correlation coefficient (R) between the source profiles obtained from PMF and SC for combustion (0.97) and secondary (0.85) sulfate, metal processing (0.73), vegetative/wood burning (0.74), diesel (0.81) and gasoline (0.66) vehicle are high while for soil/crustal source (0.09) however, the value is low. Similarly, for the 2C dataset, R is high for combustion (0.95) and secondary (0.68) sulfate, metal processing (0.88), vegetative/wood burning (0.99), traffic (0.93) while for soil/crustal (0.10) is low.

The statistical analysis for source contributions for 8C derived from KMC, soil/crustal, vegetative/wood burning, combustion and secondary sulfate and metal processing sources ($p < 0.05$) exhibit significant differences compared to PMF. Similarly, for the 2C dataset, soil/crustal, combustion and secondary sulfate and metal processing sources ($p < 0.05$) have significant differences in source contributions compared to PMF.

For multivariate receptor models, the source profiles should be comparable, but no profiles are considered standard results and only enough evidence needs to be collected to conclude convincing sources (Lewis *et al.*, 2003). The overall difference in contribution and chemical composition of the sources using PMF and SC is well within the acceptable estimation range. The difference between SC and PMF results can be reduced if we have large datasets as ML algorithms require extensive data for training and the accuracy improves with the size of the data. We are collecting large data sets from other sampling sites to apply our SC algorithm for further improvement. Overfitting is a scenario where the model performs better on training data than on testing/validation data. However, there are no objective criteria in clustering to declare that a certain output is correct as there are no labels or target variables in unsupervised ML algorithms. So, cross-validation on test data is not applicable in clustering as we need to assess the disparity in results from ground truth which is not available. However, the overfitting in clustering is expressed in terms of determining optimal cluster numbers. If n clusters are fitted on n samples, then the results are impractical and do not reflect the structure of the data (Schölkopf *et al.*, 2007; Gan *et al.*, 2007; Xu and Wunsch, 2008; Aggarwal and Reddy, 2014). So, choosing the optimal number of clusters and comparison with external results if available are techniques to avoid overfitting in clustering.

The results obtained in this study provides an optimistic outlook towards SC as a receptor model. The aforementioned being said, we acknowledge that this is just a comparative study whose objective was to introduce and estimate the feasibility of SC as a receptor model for SA. The results are encouraging and provide evidence for SC's feasibility as a potential mechanism for receptor modeling and SA. More sophisticated, accurate, and robust models can be developed using clustering algorithms as the framework's base with further research on larger datasets.

4 CONCLUSIONS

A new receptor modeling approach was examined for SA (source apportionment) using a machine learning (ML) methodology, the SC (spectral clustering) algorithm. This was demonstrated for two $PM_{2.5}$ datasets, 2C (two carbon fractions) and 8C (eight carbon fractions) which identified six and seven contributing sources for 2C and 8C datasets. The sources identified are combustion ($45.9 \pm 3.66\%$) and secondary sulfate ($11.4 \pm 1.09\%$), vegetative/wood burning ($17.5 \pm 1.46\%$), diesel ($10.6 \pm 0.92\%$) and gasoline ($3.6 \pm 0.33\%$) vehicles, soil/crustal ($2.07 \pm 0.2\%$), traffic ($9.3 \pm 0.81\%$), and metal processing ($8.8 \pm 0.72\%$) out of which combustion sulfate has the highest contribution followed by vegetative/wood burning. Collectively, the results obtained using SC were compared with PMF results and they appear consistent. Importantly, our results provide evidence for SC's feasibility as a potential mechanism for receptor modeling and SA. This study presented a basic framework for applying Machine Learning algorithms for source apportionment analysis and SC as a potential receptor model for SA providing a basis for further research.

ACKNOWLEDGMENTS

This work has been supported by the Central Pollution Control Board as a part of the study "Pilot Study for Assessment of Reducing Particulate Air Pollution in Urban Areas by Using Air Cleaning System (sometimes called as Smog Tower)" (Grant no: RD/0120-CPCB000-001). Partial support from the study "Application of Nanoparticles in ESP for Inactivation of Microorganisms and Degradation of VOCs for Air Purification" (Grant no: RD/0119-DST0000-048) is acknowledged.

SUPPLEMENTARY MATERIAL

Supplementary material for this article can be found in the online version at <https://doi.org/10.4209/aaqr.210240>

REFERENCES

- Aggarwal, C.C., Reddy, C.K. (2014). Data clustering: Algorithms and applications. Chapman And Hall/CRC.
- Alpaydin, E. (2020). Introduction to machine learning. MIT Press.
- Amato, F., Alastuey, A., de la Rosa, J., Gonzalez Castanedo, Y., Sánchez de la Campa, A.M., Pandolfi, M., Lozano, A., Contreras González, J., Querol, X. (2014). Trends of road dust emissions contributions on ambient air particulate levels at rural, urban and industrial sites in southern Spain. *Atmospheric Chem. Phys.* 14, 3533–3544. <https://doi.org/10.5194/acp-14-3533-2014>
- Amato, F., Pandolfi, M., Escrig, A., Querol, X., Alastuey, A., Pey, J., Perez, N., Hopke, P.K. (2009). Quantifying road dust resuspension in urban environment by multilinear engine: a comparison with PMF2. *Atmos. Environ.* 43, 2770–2780. <https://doi.org/10.1016/j.atmosenv.2009.02.039>
- Bach, F., Jordan, M. (2004). Learning spectral clustering. In *Adv. Neural Inf. Process. Syst.*
- Beddows, D.C.S., Dall'Osto, M., Harrison, R.M. (2009). Cluster analysis of rural, urban, and curbside atmospheric particle size data. *Environ. Sci. Technol.* 43, 4694–4700. <https://doi.org/10.1021/es803121t>
- Begum, B.A., Biswas, S.K., Hopke, P.K. (2007). Source apportionment of air particulate matter by chemical mass balance (CMB) and comparison with positive matrix factorization (PMF) Model. *Aerosol Air Qual. Res.* 7, 446–468. <https://doi.org/10.4209/aaqr.2006.10.0021>
- Belis, C.A., Karagulian, F., Larsen, B.R., Hopke, P.K. (2013). Critical review and meta-analysis of ambient particulate matter source apportionment using receptor models in Europe. *Atmos. Environ.* 69, 94–108. <https://doi.org/10.1016/j.atmosenv.2012.11.009>
- 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>
- Caselli, M., de Gennaro, G., Ielpo, P. (2006). A comparison between two receptor models to determine the source apportionment of atmospheric pollutants. *Environmetrics* 17, 507–516. <https://doi.org/10.1002/env.788>
- Celebi, M.E., Aydin, K. (2016). Unsupervised learning algorithms. Springer, Cham.
- Charron, A., Birmili, W., Harrison, R.M. (2008). Fingerprinting particle origins according to their size distribution at a UK rural site. *J. Geophys. Res.* 113, D07202. <https://doi.org/10.1029/2007jd008562>
- Chen, C.H., Pau, L.F., Wang, P.S.P. (1993). Handbook of pattern recognition and computer vision. World Scientific, Singapore.
- Chen, L.-W. A., Watson, J.G., Chow, J.C., DuBois, D.W., Herschberger, L. (2011). PM_{2.5} source apportionment: Reconciling receptor models for U.S. nonurban and urban long-term networks. *J. Air Waste Manage. Assoc.* 61, 1204–1217. <https://doi.org/10.1080/10473289.2011.619082>
- Chow, J.C., Watson, J.G. (2002). Review of PM_{2.5} and PM₁₀ apportionment for fossil fuel

- combustion and other sources by the chemical mass balance receptor model. *Energy Fuels* 16, 222–260. <https://doi.org/10.1021/ef0101715>
- Clarke, B., Fokoue, E., Zhang, H.H. (2009). *Principles and theory for data mining and machine learning*. New York, Ny Springer New York.
- Dall’Osto, M., Beddows, D.C.S., Pey, J., Rodriguez, S., Alastuey, A., Harrison, R.M., Querol, X. (2012). Urban aerosol size distributions over the Mediterranean city of Barcelona, NE Spain. *Atmos. Chem. Phys.* 12, 10693–10707. <https://doi.org/10.5194/acp-12-10693-2012>
- Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Mag.* 17, 37–54.
- Filippone, M., Camastra, F., Masulli, F., Rovetta, S. (2008). A survey of kernel and spectral methods for clustering. *Pattern Recognit.* 41, 176–190. <https://doi.org/10.1016/j.patcog.2007.05.018>
- Gan, G., Ma, C., Wu, J. (2007). *Data Clustering: Theory, Algorithms, and Applications*. Society for Industrial and Applied Mathematics. <https://doi.org/10.1137/1.9780898718348>
- Halkidi, M., Batistakis, Y., Vazirgiannis, M. (2001). On clustering validation techniques. *J. Intell. Inf. Syst.* 17, 107–145. <https://doi.org/10.1023/a:1012801612483>
- Harrison, R.M., Beddows, D.C.S., Dall’Osto, M. (2011). PMF analysis of wide-range particle size spectra collected on a major highway. *Environ. Sci. Technol.* 45, 5522–5528. <https://doi.org/10.1021/es2006622>
- 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>
- Hu, S., McDonald, R., Martuzevicius, D., Biswas, P., Grinshpun, S.A., Kelley, A., Reponen, T., Lockey, J., LeMasters, G. (2006). UNMIX modeling of ambient PM_{2.5} near an interstate highway in Cincinnati, OH, USA. *Atmos. Environ.* 40, 378–395. <https://doi.org/10.1016/j.atmosenv.2006.02.038>
- Hu, S. (2007). *Approaches to estimating exposure levels of diesel exhaust particles (DEP) in an urban airshed: Model development and applications* (Ph.D. Thesis). Washington University, Saint Louis, MO, USA.
- Jain, A.K., Murty, M.N., Flynn, P.J. (1999). Data clustering: A review. *ACM Comput. Surv.* 31, 264–323. <https://doi.org/10.1145/331499.331504>
- Jeong, C., McGuire, M.L., Herod, D., Dann, T., Dabek-Zlotorzynska, E., Wang, D., Ding, L., Celo, V., Mathieu, D., Evans, G. (2011). Receptor model based identification of PM_{2.5} sources in Canadian cities. *Atmos. Pollut. Res.* 2, 158–171. <https://doi.org/10.5094/apr.2011.021>
- 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>
- Kim, E., Larson, T.V., Hopke, P.K., Slaughter, C., Sheppard, L.E., Claiborn, C. (2003). Source identification of PM_{2.5} in an arid Northwest U.S. City by positive matrix factorization. *Atmos. Res.* 66, 291–305. [https://doi.org/10.1016/s0169-8095\(03\)00025-5](https://doi.org/10.1016/s0169-8095(03)00025-5)
- Kim, E., Hopke, P.K., Edgerton, E.S. (2004). Improving source identification of Atlanta aerosol using temperature resolved carbon fractions in positive matrix factorization. *Atmos. Environ.* 38, 3349–3362. <https://doi.org/10.1016/j.atmosenv.2004.03.012>
- Kim, E., Hopke, P.K., Pinto, J.P., Wilson, W.E. (2005). Spatial variability of fine particle mass, components, and source contributions during the regional air pollution study in St. Louis. *Environ. Sci. Technol.* 39, 4172–4179. <https://doi.org/10.1021/es049824x>
- Lee, S., Liu, W., Wang, Y., Russell, A.G., Edgerton, E.S. (2008). Source apportionment of PM_{2.5}: comparing PMF and CMB results for four ambient monitoring sites in the southeastern United States. *Atmos. Environ.* 42, 4126–4137. <https://doi.org/10.1016/j.atmosenv.2008.01.025>

- Lewis, C.W., Norris, G.A., Conner, T.L., Henry, R.C. (2003). Source apportionment of Phoenix PM_{2.5} aerosol with the UNMIX receptor model. *J. Air Waste Manage. Assoc.* 53, 325–338. <https://doi.org/10.1080/10473289.2003.10466155>
- Liang, C.S., Wu, H., Li, H.Y., Zhang, Q., Li, Z., He, K.B. (2020). Efficient data preprocessing, episode classification, and source apportionment of particle number concentrations. *Sci. Total Environ.* 744, 140923. <https://doi.org/10.1016/j.scitotenv.2020.140923>
- Liang, C.S., Yue, D., Wu, H., Shi, J.S., He, K.B. (2021). Source apportionment of atmospheric particle number concentrations with wide size range by nonnegative matrix factorization (NMF). *Environ. Pollut.* 289, 117846. <https://doi.org/10.1016/j.envpol.2021.117846>
- Liu, J., Han, J. (2013). Spectral Clustering. In *Data Clustering: Algorithms and Applications*.
- Liu, W., Hopke, P.K., Han, Y., Yi, S.M., Holsen, T.M., Cybart, S., Kozlowski, K., Milligan, M. (2003). Application of receptor modeling to atmospheric constituents at Potsdam and Stockton, NY. *Atmos. Environ.* 37, 4997–5007. <https://doi.org/10.1016/j.atmosenv.2003.08.036>
- Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Trans. Inf. Theory* 28, 129–137. <https://doi.org/10.1109/tit.1982.1056489>
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations, *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, pp. 281–297.
- McSherry, F. (2004). Spectral methods for data analysis (Ph.D. Thesis). University of Washington, Seattle, WA, USA.
- Mendenhall, W., Sincich, T. (2014). *A second course in statistics: Regression analysis*. Pearson, Harlow, Essex.
- Merris, R. (1994). Laplacian matrices of graphs: A survey. *Linear Algebra Its Appl.* 197–198, 143–176. [https://doi.org/10.1016/0024-3795\(94\)90486-3](https://doi.org/10.1016/0024-3795(94)90486-3)
- Morissette, L., Chartier, S. (2013). The k-means clustering technique: General considerations and implementation in Mathematica. *TQMP* 9, 15–24. <https://doi.org/10.20982/tqmp.09.1.p015>
- Olivas, E.S., Guerrero, J.D.M., Martinez-Sober, M., Magdalena-Benedito, J.R., Serrano, L. (2009). *Handbook of research on machine learning applications and trends: Algorithms, methods, and techniques*. IGI Global.
- Paatero, P., Tapper, U. (1994). Positive matrix factorization: A nonnegative factor model with optimal utilization of error estimates of data values. *Environmetrics* 5, 111–126. <https://doi.org/10.1002/env.3170050203>
- Paatero, P. (1997). Least squares formulation of robust nonnegative factor analysis. *Chemom. Intell. Lab. Syst.* 37, 23–35. [https://doi.org/10.1016/s0169-7439\(96\)00044-5](https://doi.org/10.1016/s0169-7439(96)00044-5)
- Polissar, A.V., Hopke, P.K., Poirot, R.L. (2001). Atmospheric aerosol over Vermont: Chemical composition and sources. *Environ. Sci. Technol.* 35, 4604–4621. <https://doi.org/10.1021/es0105865>
- Pozza, S.A., Bruno, R.L., Tazinassi, M.G., Goncalves, J.A.S., Filho, V.F.D.N., Barrozo, M.A.S., Coury, J.R. (2009). Sources of particulate matter: emission profile of biomass burning. *Int. J. Environ. Pollut.* 36, 276. <https://doi.org/10.1504/ijep.2009.021832>
- Ramadan, Z., Song, X.H., Hopke, P.K. (2000). Identification of sources of Phoenix aerosol by positive matrix factorization. *J. Air Waste Manage. Assoc.* 50, 1308–1320. <https://doi.org/10.1080/10473289.2000.10464173>
- Ramadan, Z., Eickhout, B., Song, X.H., Buydens, L.M.C., Hopke, P.K. (2003). Comparison of positive matrix factorization and multilinear engine for the source apportionment of particulate pollutants. *Chemom. Intell. Lab. Syst.* 66, 15–28. [https://doi.org/10.1016/s0169-7439\(02\)00160-0](https://doi.org/10.1016/s0169-7439(02)00160-0)
- Rokach L., Maimon O. (2005) Clustering methods. In: Maimon O., Rokach L. (Eds.), *Data Mining and Knowledge Discovery Handbook*. Springer, Boston, MA. https://doi.org/10.1007/0-387-25465-X_15
- 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>
- Salimi, F., Ristovski, Z., Mazaheri, M., Laiman, R., Crilley, L.R., He, C., Clifford, S., Morawska, L. (2014). Assessment and application of clustering techniques to atmospheric particle number size distribution for the purpose of source apportionment. *Atmos. Chem. Phys.* 14, 11883–11892. <https://doi.org/10.5194/acp-14-11883-2014>

- Sammur, C., Webb, G.I. (2011). Encyclopedia of machine learning. Springer, New York.
- Schölkopf, B., Platt, J., Hofmann, T. (Eds.) (2007). Advances in neural information processing systems 19: Proceedings of the 2006 Neural Information Processing System Conference. MIT Press, Cambridge, Massachusetts.
- Song, X.H., Polissar, A.V., Hopke, P.K. (2001). Sources of fine particle composition in the northeastern US. *Atmos. Environ.* 35, 5277–5286. [https://doi.org/10.1016/s1352-2310\(01\)00338-7](https://doi.org/10.1016/s1352-2310(01)00338-7)
- Tan, J., Duan, J., Chai, F., He, K., Hao, J.M. (2014). Source apportionment of size segregated fine/ultrafine particle by PMF in Beijing. *Atmos. Res.* 139, 90–100. <https://doi.org/10.1016/j.atmosres.2014.01.007>
- Turn, S.Q., Jenkins, B.M., Chow, J.C., Pritchett, L.C., Campbell, D., Cahill, T., Whalen, S.A. (1997). Elemental characterization of particulate matter emitted from biomass burning: Wind tunnel derived source profiles for herbaceous and wood fuels. *J. Geophys. Res.* 102, 3683–3699. <https://doi.org/10.1029/96JD02979>
- 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>
- Von Luxburg, U. (2007). A tutorial on spectral clustering. *Stat. Comput.* 17, 395–416. <https://doi.org/10.1007/s11222-007-9033-z>
- Watson, J.G., Zhu, T., Chow, J.C., Engelbrecht, J., Fujita, E.M., Wilson, W.E. (2002). Receptor modeling application framework for particle source apportionment. *Chemosphere* 49, 1093–1136. [https://doi.org/10.1016/s0045-6535\(02\)00243-6](https://doi.org/10.1016/s0045-6535(02)00243-6)
- Wegner, T., Hussein, T., Hämeri, K., Vesala, T., Kulmala, M., Weber, S. (2012). Properties of aerosol signature size distributions in the urban environment as derived by cluster analysis. *Atmos. Environ.* 61, 350–360. <https://doi.org/10.1016/j.atmosenv.2012.07.048>
- Xu, R., Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Trans. Neural Networks* 16, 645–678. <https://doi.org/10.1109/tnn.2005.845141>
- Xu, R., Wunsch, D.C. (2008). Clustering. John Wiley and Sons, Piscataway, NJ.
- Yue, W., Stölzel, M., Cyrys, J., Pitz, M., Heinrich, J., Kreyling, W.G., Wichmann, H.E., Peters, A., Wang, S., Hopke, P.K. (2008). Source apportionment of ambient fine particle size distribution using positive matrix factorization in Erfurt, Germany. *Sci. Total Environ.* 398, 133–144. <https://doi.org/10.1016/j.scitotenv.2008.02.049>