A Framework for Dynamic Hyper-local Source Apportionment using Low-cost Sensors for Real-time Policy Action

Shoubhik Chakraborty[1,2,4], Sachchida Nand Tripathi[2,3], Davender Sethi[2], Akanksha Lakra[2], Ambasht Kumar[2], Pranjal Kumar Srivastava[1], Nihal Thukarama Rao[1], Avnish Tripathi[1], and Purushottam Kar[1]

[1]Department of Computer Science and Engineering, Indian Institute of Technology Kanpur, Kanpur, 208016, India
[2]Department of Civil Engineering, Indian Institute of Technology Kanpur, Kanpur, 208016, India
[3]Department of Sustainable Energy Engineering, Indian Institute of Technology Kanpur, Kanpur, 208016, India
[4]Department of Computer Science and Engineering, Shiv Nadar University Chennai, Chennai, 603110, India

Correspondence: Shoubhik Chakraborty (shbhk1908@gmail.com), Sachchida Nand Tripathi (snt@iitk.ac.in), and Purushottam Kar (purushot@cse.iitk.ac.in)

Abstract. The presence of particulate matter, toxic gases and other pollutants in the air pose significant risk to human health and the environment. Identifying the different sources of air pollution which is termed as Source Apportionment (SA), needs to be done in real-time in order to understand the dynamics of the contributing sources and also to enable the policy makers frame effective regulatory measures to curb air pollution. The unit deployed for implementing the SA framework at a particular location must also be cost-effective, so that it becomes feasible to create a dense network with such units and thus cover a wide geographical area. The use of low-cost air quality monitoring sensors have become popular in this regard. In our proposed framework we use low-cost air quality sensor units in conjunction with machine learning models to develop a low-cost real-time solution for SA. Multi output regression models, which are supervised machine learning models are used for this purpose. Reference Grade Instruments are used for learning calibration models for the low-cost sensors as well as the multi output regression models for SA. Once the calibration and multi output regression models are learnt during training, the proposed framework allows the low-cost sensors to be deployed on the field as a standalone device, where it collects on-field data and stores it in a remote server through a wireless network. This data can be pulled at the user end, calibrated and then fed to the trained model to obtain the SA results in terms of the relative abundance of the different sources in ambient air. Mean Absolute Error (MAE) has been used as the metric to measure the accuracy in predicting the relative abundance of different sources, while Spearman's Rank Order Correlation Coefficient (SROCC) and Normalized Discounted Cumulative Gain (NDCG) are the metrics that have been used to get an estimate of how well the proposed approach performs in predicting the relative abundance of the different sources in the correct order. Extensive experimentation done using data gathered from two different environments in the city of Lucknow, India shows the robustness of the proposed approach in doing real-time SA. MAE of less than 5% have been obtained in predicting the relative abundance of most of the organic as well as elemental sources, while values of SROCC greater than 0.75 and NDCG greater than 0.85 obtained for all the sources shows that the proposed framework also performs very well in predicting most of the sources in correct order of their actual contribution to air pollution.

Keywords Source Apportionment, Low-cost Sensors, Real-time, Machine Learning, Multi output regression Model.

https://doi.org/10.5194/egusphere-2025-5677
Preprint. Discussion started: 9 December 2025

Abbreviations

PM: Particulate Matter

VOC: Volatile Organic Compounds

SA: Source Apportionment

RGI: Reference Grade Instrument

LCAQ: Low Cost Air Quality (sensor)

PMF: Positive Matrix Factorization

BC: Black Carbon

1 Introduction

The rising concentrations of PM and toxic gases such as carbon monoxide (CO) and sulphur dioxide (SO_2), originating from both natural and anthropogenic activities, pose serious environmental and public health risks (Manisalidis et al., 2020). Among these, PM and BC are of particular concern, as they can penetrate deep into the lungs and enter the bloodstream, adversely affecting respiratory and cardiovascular health (Cohen et al., 2017). Long-term exposure to PM, VOCs and other toxic gases makes an individual prone to lung cancer, asthma and heart disease (Sharma et al., 2013; Zhou et al., 2023). They also contribute to ground-level ozone (O_3) formation (Ying and Krishnan, 2010; Li et al., 2015; Zhang et al., 2020), which harms ecosystems by damaging plant life. Pollutants like NO_2 and SO_2 are precursors to acid rain (Mehta, 2010), affecting biodiversity and soil quality, while O_3 is harmful to both plant and human health (Nuvolone et al., 2018; Karmakar et al., 2022), exacerbating asthma and reducing lung function. Regulatory measures need to be adopted and implemented by policy makers in order to curb air pollution and this needs detailed knowledge of the sources contributing to air pollution. This process of identifying the sources of air pollution is known as *Source Apportionment* (SA) (Coelho et al., 2022). The implementation of SA techniques can aid policy makers to target the most harmful emission contributors. By understanding whether traffic, industry, or biomass burning is the primary culprit in specific regions, governments can enact measures like stricter vehicular emission standards (Singh et al., 2023) to reduce traffic emissions, industrial restrictions by setting maximum allowable emissions for specific pollutants or by providing incentives for the use of cleaner technologies and renewable energies (Munsif et al., 2021). Without precise source attribution, efforts to reduce air pollution may be ineffective or misdirected. It is also advantageous to have a SA technique which produces output in real-time. Any system or process which responds to its external stimuli within a definite amount of time is said to be a real-time system or process (Liu et al., 2006). SA done in real-time will keep the end-users updated about the dynamics of the various pollution sources and enable them to act quickly in case of any emergency (Yoo, 2021). Furthermore, in order to collect SA information of a large area with high spatial resolution, it is necessary to create a dense network of devices or instruments responsible for capturing and analysing the sampled air (Ajnoti et al., 2024). This in turn requires that the device or instrument deployed at each node of such a network be cost-effective. One of the traditional SA modalities, defined as filter based techniques (Cohen et al., 2010; Jaiprakash et al., 2017; Jain et al., 2018; Sharma et al., 2016), involves collecting air samples on filters and analysing the composition of PM with techniques such as X-ray fluorescence, ion chromatography etc. Finally, to trace the sources of PM, statistical or machine learning methods like Chemical Mass Balance (CMB) (Watson et al., 1991), Positive Matrix Factorization (PMF) (Paatero and Tapper, 1994) or clustering (Kumar et al., 2022) are used. Although, the method provides accurate and detailed chemical composition of PM in the sampled air, however, such methods are very time as well as resource consuming. Thus, they are not suitable for obtaining SA results with high spatio-temporal resolution. The use of expensive RGI like Aerosol Mass Spectrometer, metal monitoring using Xact-625i, Environmental Beta Attenuation Mass Monitor (EBAM) and others coupled with computationally intensive receptor models like PMF (Bhandari et al., 2020; Lakra et al., 2024; Lalchandani et al., 2021; Shukla et al., 2021) has made possible the implementation of SA much faster as compared to the filter based techniques. However, their cost and the requirement of skilled researchers to operate these instruments makes it infeasible for large scale deployment. Recently people have started exploring the potential of LCAQ sensors in performing SA with high spatio-temporal resolution (Bousiotis et al., 2021, 2022; Hagan et al., 2019; Owoade et al., 2021). LCAQ sensors have become increasingly popular for monitoring PM due to their affordability, portability, and ease of deployment. The effectiveness of LCAQ sensors in PM monitoring has been demonstrated through several studies (Liu et al., 2020; Pope et al., 2018; Shen et al., 2021; Kulkarni et al., 2022; Sahu et al., 2020). Feenstra et al. (2019) demonstrated that LCAQ sensors can closely match the PM concentration obtained using a co-located RGI. Similarly, Snyder et al. (2013) showed that LCAQ sensors could be used to estimate near-roadway PM concentrations. A study by Steinle et al. (2015) showed that LCAQ sensors could be used in participatory sensing projects to involve citizens in air quality monitoring and raise awareness about air pollution. However, LCAQ sensors have less measurement accuracy as compared to RGI. Hence, they require calibration before deployment in field and machine learning techniques are generally used for calibration of such sensors (Sahu et al., 2020, 2021).

1.1 Related Works and State of the Art

Until recently, receptor modelling techniques such as PMF and the Chemical Mass Balance (CMB) model have been the de facto standard approaches for source apportionment studies (Reff et al., 2007; Sun et al., 2020; Hopke, 2016). Traditionally, these techniques were applied on data captured using RGI. However, recently researchers have explored the efficacy of these techniques on data captured using LCAQ sensors as well. Bousiotis et al. (2025) highlights the rapid evolution of source apportionment techniques from traditional receptor-based approaches toward data-driven and hybrid machine learning frameworks. Most recent studies employed unsupervised learning algorithms such as k-Means clustering, Principal Component Analysis (PCA), and Non-Negative Matrix Factorization (NMF) to extract latent temporal or spatial patterns from sensor data. These methods are particularly suitable when chemical speciation data are unavailable, which is an inherent limitation of most low-cost sensor (LCS) networks. Studies like Hagan et al. (2019), Yang et al. (2022), and Kumar et al. (2024a) illustrate the effectiveness of NMF in decomposing mixed signals into interpretable sources, while Bousiotis et al. (2022) used two-step PMF model for source apportionment using data obtained from LCAQ sensors. However, due to the high computational complexity associated with these methods, researchers have resorted to the use of more sophisticated machine learning algorithms. Kumar et al. (2022) demonstrated the potential of spectral clustering for source apportionment using $PM_{2.5}$ datasets, while Bousiotis et al. (2021, 2023) and Dimitriou et al. (2023) demonstrate the use of k-Means to identify recurring source-like behaviours in PM and gaseous pollutant data.

In terms of datasets, nearly all studies rely on PM measurements obtained from low-cost optical particle counters (e.g., Alphasense OPC-N3) combined with limited gaseous pollutant sensors (NO_2, O_3, CO) and occasional black carbon (BC) monitors. Co-location with RGI remains a critical step for calibration and bias correction. Some studies, such as Liang et al. (2025), explore mobile networks of low-cost sensor units (e.g., sensors mounted on taxis) to map spatial emission patterns dynamically using clustering algorithms. While ML methods significantly improve scalability and automation, they are constrained by sensor drift, limited chemical specificity, and environmental interference, emphasizing the need for robust cross-validation against traditional receptor models.

A few recent studies by Mills et al. (2023) and Kumar et al. (2024b), have integrated supervised machine learning techniques with receptor-based approaches like Positive Matrix Factorization (PMF) and Chemical Mass Balance (CMB), thereby enhancing both the predictive accuracy and interpretability of SA results. In particular, Kumar et al. (2024b) proposes a practical supervised ML framework capable of generating real-time SA estimates by leveraging time-series data from low-cost particle sensors in conjunction with reference SA outputs derived from the CMB model applied to data collected from RGI. The framework employs multi-output regression models to learn the mapping between the low-cost sensor measurements and the ground-truth source contribution profiles obtained from RGI. Once trained, these regression models are deployed for rapid inference on streaming sensor data, enabling near real-time source attribution. The study is strongly motivated by the operational demand for fast and scalable SA estimates across dense networks of low-cost air-quality sensors.

Motivated by the methodology proposed by Kumar et al. (2024b), the present study seeks to develop a novel receptor-modeling paradigm that jointly leverages data from LCAQ sensors and RGI to achieve high-accuracy SA while enabling real-time operational capability. The proposed framework integrates pollutant gas concentrations, such as CO, NO_2, SO_2 and O_3 along with $PM_{2.5}$, VOCs, BC and environmental parameters including temperature and relative humidity for source apportionment. In addition, the study attempts to identify emission sources based on their associated non-refractive organic components and characteristic trace elements. Whereas most previous studies have relied exclusively on LCAQ data for training receptor models, our approach investigates the potential advantages of integrating RGI measurements into the modeling process, with the objective of enhancing model robustness, calibration fidelity, and generalizability across diverse sensing environments.

1.2 Summary of Contributions

This paper contributes towards the development of a machine learning based framework for doing SA in real-time using data obtained from LCAQ sensor units. The use of the term *real-time SA* in the context of this paper implies that the proposed framework can provide continuous updated information regarding the contribution or the relative abundance of the different sources contributing to air pollution at fixed regular intervals of time. The LCAQ sensor unit houses several sensors for measurement of some common gas species, $PM_{2.5}$, VOCs, Temperature and Relative Humidity. RGI like Gas Analysers, EBAM, High Resolution Time of Flight Aerosol Mass Spectrometer (HR-ToF-AMS) and instruments for monitoring the concentration

Table 1. Make and model number of all the instruments used in this experiment.

Instruments	**Make/Model number**
High Resolution Time of Flight Aerosol Mass Spectrometer (HR-ToF-AMS)	Aerodyne/HR-ToF-AMS 215-127
Xact-625i	SailBri Cooper/Xact-625i
Gas Analysers for CO, NO_2, O_3 and SO_2	Thermo-Scientific Gas Analyser/ Model $43i - SO_2$, Model $48i - CO$, Model $42i - NO_X$, Model $49i - O_3$, Model $146i - DGC$
Micro-Aethlometer	AethLabs/AE-51
E-BAM	MetOne/8037 Mass FOIL SET Serial C21637
Low-cost sensors for CO, NO_2, O_3 and SO_2	Alphasense/B4
Low-cost sensors for $PM_{2.5}$and VOC	Sensiron/SEN54
Low-cost Temperature and Relative Humidity sensors	Bosch/BME-280

of different trace elements in ambient air are used for training the machine learning models for SA and also for calibrating the LCAQ sensor units. The proposed machine learning framework for SA is a multi output regression model (Borchani et al., 2015). Such a learning model falls under the category of supervised machine learning (Sen et al., 2020), wherein the model is fed with ground truth data for learning during the training phase. Once the model has been trained using the requisite amount of data from both LCAQ sensor units as well as ground truth data from RGI, it can be deployed on the field without any further requirement of data being fed from expensive RGI. The current work is a novel technique, which contributes to the development of SA as a real-time task, which can be done on a hyper-local scale by using a dense network of these LCAQ sensor units. During the testing phase, the LCAQ sensor units are deployed as standalone devices in field, which capture the on-field data and uploads them onto a remote server. This uploaded data can be pulled at the user-end and fed to the already trained machine learning model to compute the SA results. Extensive experimentation has been carried out to validate the robustness of the proposed framework on two deployments with contrasting environments within Lucknow city of India during the month of October 2023.

The rest of the paper is organized as follows. Section 2 discusses in detail the experimental set up and the deployment sites used for these experiments. Section 3 explains in detail about the data used for these experiments along with the methodology. Section 4 shows the comparison of the SA results predicted using the proposed machine learning framework with the ground truth SA results obtained using RGI like HR-ToF-AMS and Xact-625i. Finally, Section 5 concludes the paper.

https://doi.org/10.5194/egusphere-2025-5677
Preprint. Discussion started: 9 December 2025

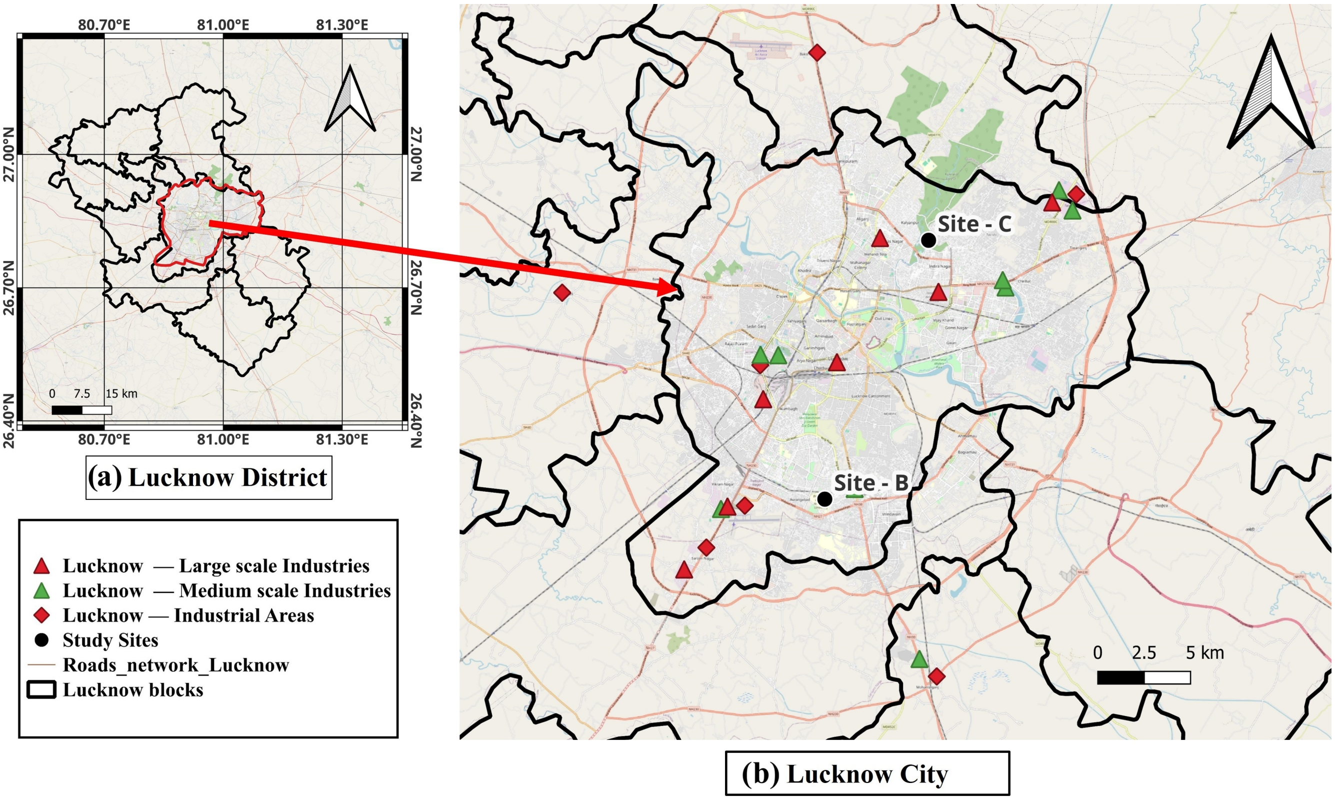


Figure 1. The locations of the deployment sites: **Site-B** and **Site-C** are shown in panel (b). The district and city boundaries of Lucknow in panel (a) and (b) respectively were taken from QGIS 3.4 Madeira.

2 Deployment Setup

Data used in our study were gathered from two deployment sites in Lucknow city: (a) **Site-B**: Babasaheb Bhimrao Ambedkar University (BBAU), Lucknow $(26.89° N, 80.98° E)$ from 13^{th} October 2023 to 22^{nd} October 2023, (b) **Site-C**: Central Institute of Medicinal and Aromatic Plants (CIMAP), which is a plant research laboratory and part of the Council of Scientific and Industrial Research (CSIR) $(26.83° N, 80.89° E)$ from 2^{nd} October 2023 to 11^{th} October 2023. **Site-B** is majorly a traffic site located near the Amar Shaheed Path highway of Lucknow city. It has some residential and commercial areas in its vicinity. **Site-C** is a research institute, also located in Lucknow and dedicated to medicinal and aromatic plants research. The surrounding areas of the institute is majorly covered by forests with some residential and commercial areas as well. It has a nearby tourist spot, which is the Kukrail Picnic spot. The site has lesser influence of anthropogenic activities such as industrial and traffic emissions. Figure 1 shows the location of these two sampling sites. As can be observed from this figure, there are large and medium scale industries located surrounding **Site-C** as well as **Site-B**. The presence of these industries can cause the pollutants emitted by them to be transported to these sites. In order to facilitate data collection at different sites without the need to set up an entire new laboratory for each site, a mobile air quality monitoring platform was set up. This system consisted of a mobile van equipped with state of the art instruments for sampling of the ambient air followed by subsequent detailed analysis of its constituents. The most notable instruments deployed in the mobile van include the High-Resolution Time of Flight Aerosol Mass Spectrometer (HR-ToF-AMS) which details the composition of the non-refractive organic and inorganic (sulphates, nitrates, ammonium, chloride) aerosols present in ambient air in terms of their mass to charge ratio (DeCarlo et al., 2006), the Xact-625i for detecting the concentration of the different trace elements (U.S. Environmental Protection Agency), Gas Analysers for measuring concentration of gases like CO, NO_2, SO_2, and O_3 (Oil and Gas Online), Micro-Aethlometer for measurement of BC concentration (South Coast Air Quality Management District (SCAQMD)), the EBAM for measurement of total $PM_{2.5}$ concentration. All these instruments provide highly accurate measurements and hence they are also termed as Regulatory Grade Instruments or Reference Grade Instruments (RGI) in literature (Bousiotis et al., 2022). The detail of these instruments and their working can be found in our previous works by Bhowmik et al. (2022) and Shukla et al. (2021). Most of these instruments are very costly and hence it becomes infeasible to deploy them on a large scale to create a dense network for measuring air quality on a hyper-local scale. Apart from these costly instruments, the mobile van also houses LCAQ sensor units manufactured by Respirer Living Sciences (Respirer Living Sciences). Each such unit contains separate electrochemical sensors for measuring the concentration of four pollutant gases: CO, NO_2, SO_2, and O_3, optical sensors for measuring the concentration of $PM_{2.5}$ and sensors for estimating the quantity of VOC in the sampled air. These LCAQ units are also equipped with Temperature and Relative Humidity sensors. The make and model number of all the instruments used in this experiment are mentioned in Table 1.

Figure 2 shows the road map for the proposed framework. The entire process of developing a low-cost sensor based SA framework can be divided into two phases: the training phase followed by the testing phase. Under the training phase, we categorize the instruments sampling the ambient air into three groups as shown in Fig. 2. The first group consists of the LCAQ sensor unit and a Micro-Aethalometer. The measurement accuracy of most of the sensors in the LCAQ sensor unit are low and hence the data obtained from these sensors need calibration before they can be used for further processing. The Micro-Aethalometer used for measuring BC concentration is a low-cost device and has accuracy comparable to that of any RGI (Alas et al., 2020). Hence, it comes lab calibrated. The working principle of the different sensors used in the LCAQ sensor unit and the Micro-Aethalometer are detailed in Appendix B. The second group of instruments consists of a set of RGI like the Gas Analysers and EBAM. The main purpose of these instruments is to measure the concentration of the gases CO, NO_2, SO_2, and O_3 along with $PM_{2.5}$ (Barbiere et al., 2019; Huang et al., 2007), which are then used as ground truth for learning linear regression models to calibrate the low-cost sensors for each of these gases and $PM_{2.5}$. The third set of instruments are those which are used to perform SA directly from the sampled air with very good accuracy. These constitute the HR-ToF-AMS and the Xact-625i. The data captured by the Xact-625i is used for performing elemental SA, while the data captured by the HR-ToF-AMS corresponding to the non-refractive organic components is used for performing organic SA (Shukla et al., 2021). The data matrix constructed from the time series data captured with these instruments is then split into two factors: one being the source profile matrix and the other being the coefficient matrix with the help of Positive Matrix Factorization (PMF). PMF is performed using Source Finder interface (SoFi version 9.4.1, Datalystica Ltd., Villigen, Switzerland) implemented with the

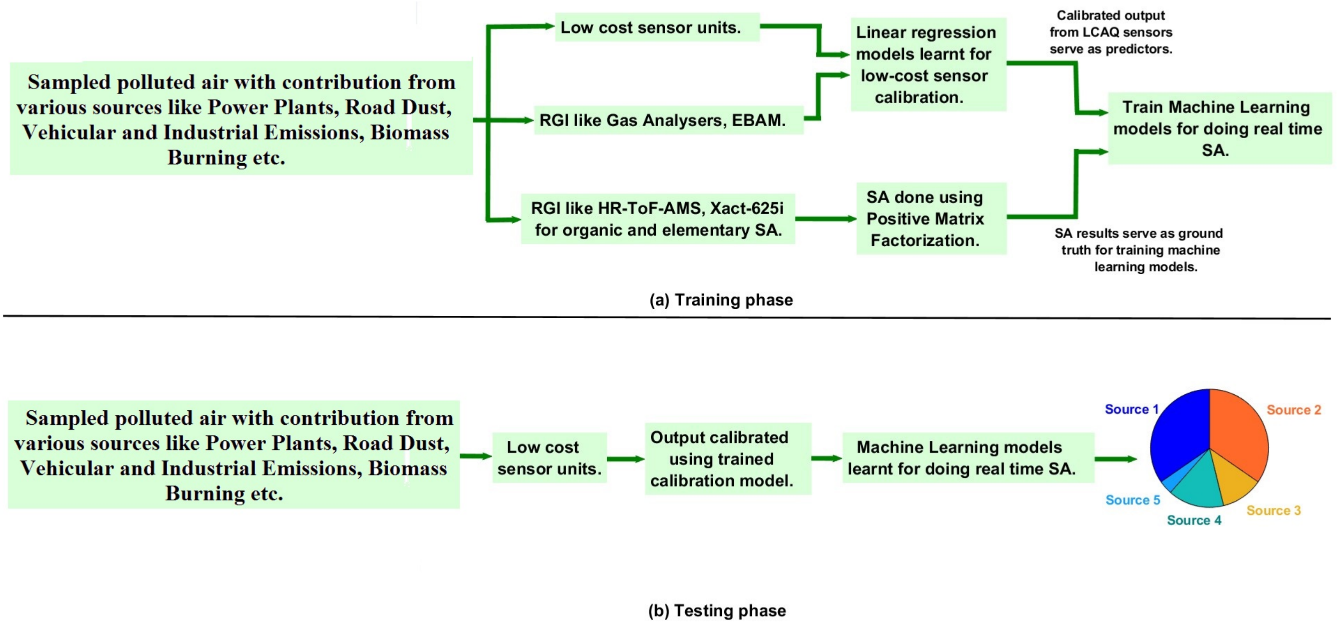

Sampled polluted air with contribution from various sources like Power Plants, Road Dust, Vehicular and Industrial Emissions, Biomass Burning etc.
Low cost sensor units.
RGI like Gas Analysers, EBAM.
RGI like HR-ToF-AMS, Xact-625i for organic and elementary SA.
Linear regression models learnt for low-cost sensor calibration.
SA done using Positive Matrix Factorization.
Calibrated output from LCAQ sensors serve as predictors.
SA results serve as ground truth for training machine learning models.
Train Machine Learning models for doing real time SA.
(a) Training phase
Sampled polluted air with contribution from various sources like Power Plants, Road Dust, Vehicular and Industrial Emissions, Biomass Burning etc.
Low cost sensor units.
Output calibrated using trained calibration model.
Machine Learning models learnt for doing real time SA.
Source 1
Source 2
Source 3
Source 4
Source 5
(b) Testing phase

Figure 2. Block diagram showing the road map of the proposed approach.

IGOR Pro 9.0.2.4 software (Wavemetrics, Inc., Portland, OR, USA) for SA analysis (Canonaco et al., 2013; Crippa et al., 2014). The coefficient matrix contains information regarding the contribution of the different sources of pollution as a function of time. The PMF algorithm is run separately on data captured from HR-ToF-AMS and Xact-625i for organic and elemental SA respectively. The calibrated data from the low-cost sensors serve as the set of predictors, while the source apportionment results obtained from the coefficient matrix serve as the ground truth for training machine learning models. The trained machine learning models are expected to replicate the source apportionment results obtained using RGI by feeding on data exclusively from low-cost sensor units. In the testing phase, the performance of these machine learning models are assessed by feeding them with data from LCAQ sensor unit after calibration and comparing the output of these models with ground truth results obtained using RGI.

3 Methodology

All experiments were conducted on a Desktop with an Intel (R) Core (TM) i7-9700 CPU @ 3.00 GHz with 32 GB RAM and Windows 11 operating system. Python 3.9.12 was used in conjunction with Jupyter Notebook as editor for implementing all the data processing algorithms. Standard machine-learning packages like Scikit-learn, Pandas NumPy and SciPy were used in all implementations. In this section, the details of the data captured by the LCAQ sensor unit, which serve as the predictors for

https://doi.org/10.5194/egusphere-2025-5677
Preprint. Discussion started: 9 December 2025

source apportionment are discussed. The data obtained after performing PMF on the HR-ToF-AMS/ Xact-625i output, which serve as the ground truth for training the machine learning model are also discussed in this section.

3.1 Datasets

For each of the two sites: **Site-B** and **Site-C**, two separate datasets are constructed, one for organic source apportionment and the other for elemental source apportionment. Each dataset consists of a time series of observations, where each observation has its associated time-stamp, predictor/ feature vector and target vector. The i^{th} observation in a dataset can be denoted by the tuple $(\mathbf{T_i}, \mathbf{x_i}, \mathbf{y_i})$, where T_i, x_i and y_i refers to the time-stamp, feature vector and target vector respectively. Furthermore, each component in a tuple is also provided with a superscript which contains information regarding the nature of the dataset. Thus, the tuple $(\mathbf{T_i^{BO}}, \mathbf{x_i^{BO}}, \mathbf{y_i^{BO}})$ contains observation from **Site-B** for organic source apportionment (**O**) and tuple $(\mathbf{T_i^{CE}}, \mathbf{x_i^{CE}}, \mathbf{y_i^{CE}})$ contains observation from **Site-C** for elemental source apportionment (**E**) etc. In any tuple $(\mathbf{T_i}, \mathbf{x_i}, \mathbf{y_i})$, the feature vector $x_i \in \mathbb{R}^D$ and the target vector $y_i \in \mathbb{R}^L$, where $\{D, L \in \mathbb{N}\}$, $\mathbb{N}$ being the set of natural numbers. The feature vector is constructed from the data obtained from LCAQ sensor unit after calibration. The different features constituting this feature vector are the concentration of four gases, $PM_{2.5}$, VOC, Temperature (T) and Relative Humidity (RH) as measured by the LCAQ sensor unit and BC as measured by Micro-Aethlometer. Thus, we have a total of nine different features and hence $D = 9$. The target vector y_i is the relative abundance of the different organic/ elemental sources polluting the sampled air. For organic source apportionment, five different sources have been identified using PMF, and hence $y_i \in \mathbb{R}^5$ for the organic case; whereas for elemental source apportionment, seven different sources have been identified and hence $y_i \in \mathbb{R}^7$ for the elemental case. Henceforth, we refer to the feature vector x_i as *LCAQ feature* vector and y_i vector as *SA target* vector. The entire dataset was split in an 80:20 ratio for training and testing purposes. The split was performed in a temporal manner, with the first 80% of observations used for training and the remaining 20% reserved for testing. To prevent artificial temporal correlation between the training and test sets (which can result from randomly mixing time-ordered observations), no shuffling was applied prior to the split. We first discuss about the LCAQ sensor unit data in next subsection, followed by the RTSA ground truth data for organic and elemental SA in subsequent subsection in detail.

3.2 LCAQ Sensor Unit Data

The LCAQ sensor units provide time-series data corresponding to the concentration of four gases CO, NO_2, SO_2, and O_3 along with $PM_{2.5}$, VOC and meteorological parameters like T and RH. This data which serves as the LCAQ feature vector (along with BC data obtained from Micro-Aethlometer) is generated with a time resolution of 1 minute. The data samples collected within each non-overlapping 30-minute interval are summarized by a single representative value, computed as their median. There are two reasons for doing this. Firstly, the median, being less sensitive to outliers, reduces the influence of measurement noise. Secondly, the RTSA ground truth data (as discussed in Section 3.3), which is used to generate the SA target vector, is available at a 30-minute resolution. Hence, for time alignment of the LCAQ feature vectors with their associated SA target vectors, the data generated from both these sources must be made available at the same time resolution.

The data obtained from the gas and $PM_{2.5}$ sensors (housed within the LCAQ sensor units) need to be calibrated before they can be processed further. For each of the gas and $PM_{2.5}$ sensors, separate linear regression models (Montgomery et al., 2021) are learnt for calibration. These linear regression models are learnt using data obtained from the sensors and their corresponding co-located RGI (Gas Analysers for gases and EBAM for $PM_{2.5}$). The gas sensors (which are of Alphasense make, model B4) give an indirect measure of actual gas concentration in the form of voltage developed at the two electrodes viz. the working electrode (V_{op1}) and the reference electrode (V_{op2}). For the $PM_{2.5}$ sensor, we obtain a single electrode output (V_{op}). Furthermore, these measurements are dependent on meteorological parameters T and RH. The output voltage developed at the electrodes of gas ($PM_{2.5}$) sensors along with T and RH form the feature vector for the linear regression model, while the data from Gas Analysers (EBAM) serve as the target. Thus, the set of linear equations describing the calibration model for the four gases and $PM_{2.5}$ are as follows:

$$\begin{aligned}
\alpha_0^{CO} + \alpha_1^{CO} V_{op1}^{CO} + \alpha_2^{CO} V_{op2}^{CO} + \alpha_3^{CO} T + \alpha_4^{CO} RH &= y_{CO} \\
\alpha_0^{NO_2} + \alpha_1^{NO_2} V_{op1}^{NO_2} + \alpha_2^{NO_2} V_{op2}^{NO_2} + \alpha_3^{NO_2} T + \alpha_4^{NO_2} RH &= y_{NO_2} \\
\alpha_0^{O_3} + \alpha_1^{O_3} V_{op1}^{O_3} + \alpha_2^{O_3} V_{op2}^{O_3} + \alpha_3^{O_3} T + \alpha_4^{O_3} RH &= y_{O_3} \\
\alpha_0^{SO_2} + \alpha_1^{SO_2} V_{op1}^{SO_2} + \alpha_2^{SO_2} V_{op2}^{SO_2} + \alpha_3^{SO_2} T + \alpha_4^{SO_2} RH &= y_{SO_2} \\
\alpha_0^{PM_{2.5}} + \alpha_1^{PM_{2.5}} V_{op}^{PM_{2.5}} + \alpha_2^{PM_{2.5}} T + \alpha_3^{PM_{2.5}} RH &= y_{PM_{2.5}}
\end{aligned} \tag{1}$$

where, $\alpha_i^{CO}, \alpha_i^{NO_2}, \alpha_i^{O_3}, \alpha_i^{SO_2}$ and $\alpha_i^{PM_{2.5}}, i = 0, 1, 2..$ etc. are the coefficients of linear regression model for CO, NO_2, O_3, SO_2 and $PM_{2.5}$ respectively, which are to be learnt in the training phase. The dataset comprising feature vectors and their corresponding targets, is split into training and test subsets in the ratio of $80\% : 20\%$. Shuffling is deliberately avoided before splitting to prevent temporal correlations between the training and test data. This ensures a fair evaluation of the model's generalization capability on test (unseen) data. In order to assess the performance of the linear models used in calibration, the coefficient of determination also known as the R^2 score is used as the performance metric, which is defined as:

$$R^2 = 1 - \frac{\sum_i (w_i - \hat{w}_i)^2}{\sum_i (w_i - \bar{w})^2} \tag{2}$$

where, w_i denotes the reference or target variable, $\hat{w}_i$ denotes the predicted variable and $\bar{w}$ denotes the mean of reference variable. The R^2 score takes a value of unity for perfect prediction i.e. when $w_i = \hat{w}_i$, $\forall i$ and this score decreases from its maximum value of unity with degradation in prediction performance. Fig. 3 and Fig. 4 show the calibration performance of the linear regression models for the four gases and $PM_{2.5}$ at **Site-B** and **Site-C** respectively. It was observed that the predicted time series closely follows the variations in the reference time series for a majority of the pollutants at both the sites with reasonably good R^2 scores. This tracking of the variations in the reference signal is better for CO, O_3 and $PM_{2.5}$ at both the sites because of the relatively higher concentration of these pollutants. Particularly for SO_2, the R^2 scores are close to zero at both the sites because the ambient concentration of SO_2 lies in the Below Detection Limit (~5 ppb for Alphasense/B4 sensors) (BDL) range and hence measurements are not very reliable (Ltd., 2023).

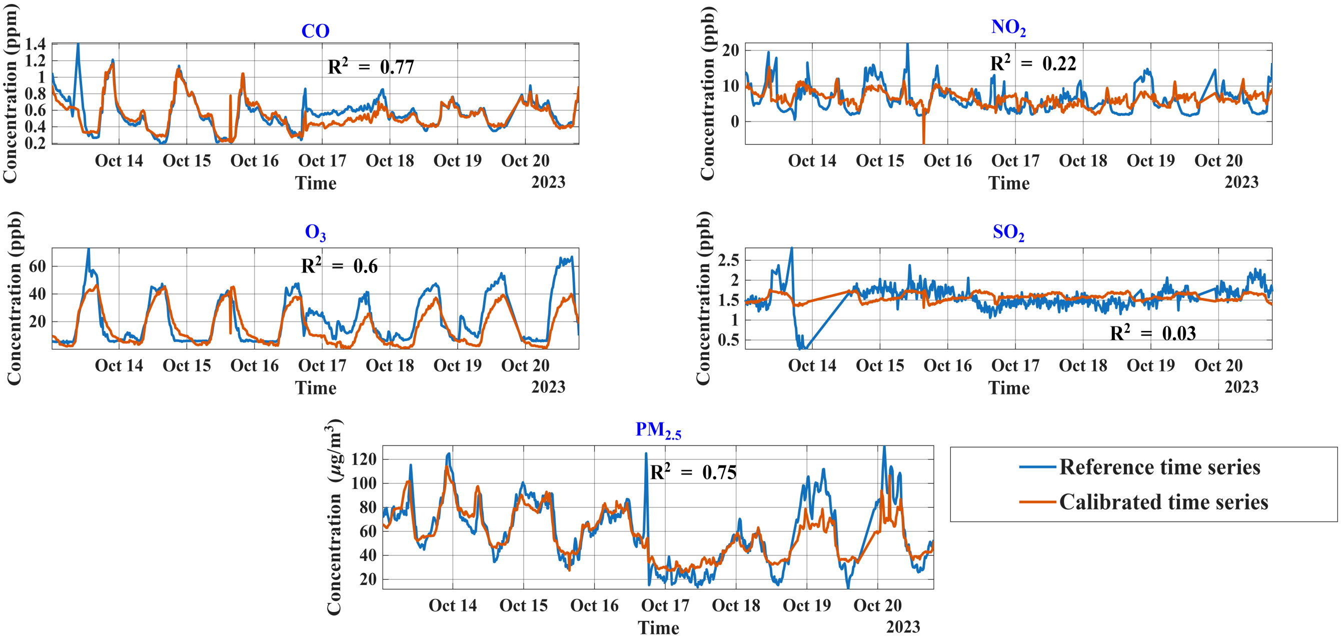


Figure 3. Time series plots of the concentration of various gases and $PM_{2.5}$ at **Site-B** as obtained after calibration along with their concentration as obtained from the Gas Analysers and EBAM (which serve as the reference).

The box and whisker plots of the different features in the training data are shown in Fig. 5. From this figure, it can be observed that the ranges of the feature values vary significantly. For instance, among the four gases, the range of values of CO concentration is much higher (median value being 0.57 ppm at **Site-B** and 0.45 ppm at **Site-C**) compared to the other three gases, whose median values are on the order of a few tens of ppb at both sites. Similarly, the numerical values of BC concentration are considerably higher than those of the other features. Hence, it is necessary to scale the data so that all features lie on a comparable scale. We adopt standardization (or Z-score scaling) as the preferred scaling method, since it does not require prior knowledge of the maximum possible feature values (scikit-learn developers, 2025). The standardization parameters (i.e., the mean and standard deviation) for each feature and each site are estimated from the training data and stored for use during the testing phase.

It is always desirable to have low correlation between the different features in the data because correlated features introduce redundancy in the data. The correlation between the different features in the training data are examined with the help of a heat map as shown in Fig. 6. From this heat map it can be observed that none of the features except $PM_{2.5}$ and BC are consistently correlated with each other across the two sites, which is desirable. The strong correlation between $PM_{2.5}$ and BC is expected because BC is a component of $PM_{2.5}$, and their co-emission from combustion sources like vehicles and biomass burning often means they are found together (Taheri et al., 2019; Choomanee et al., 2024; Tiwari et al., 2013). Furthermore, Gong et al. (2015) demonstrated in their paper that this correlation varies across seasons and they observed it to be higher during the autumn and summer season. However, despite this high correlation observed at these two sites, we need to use both BC

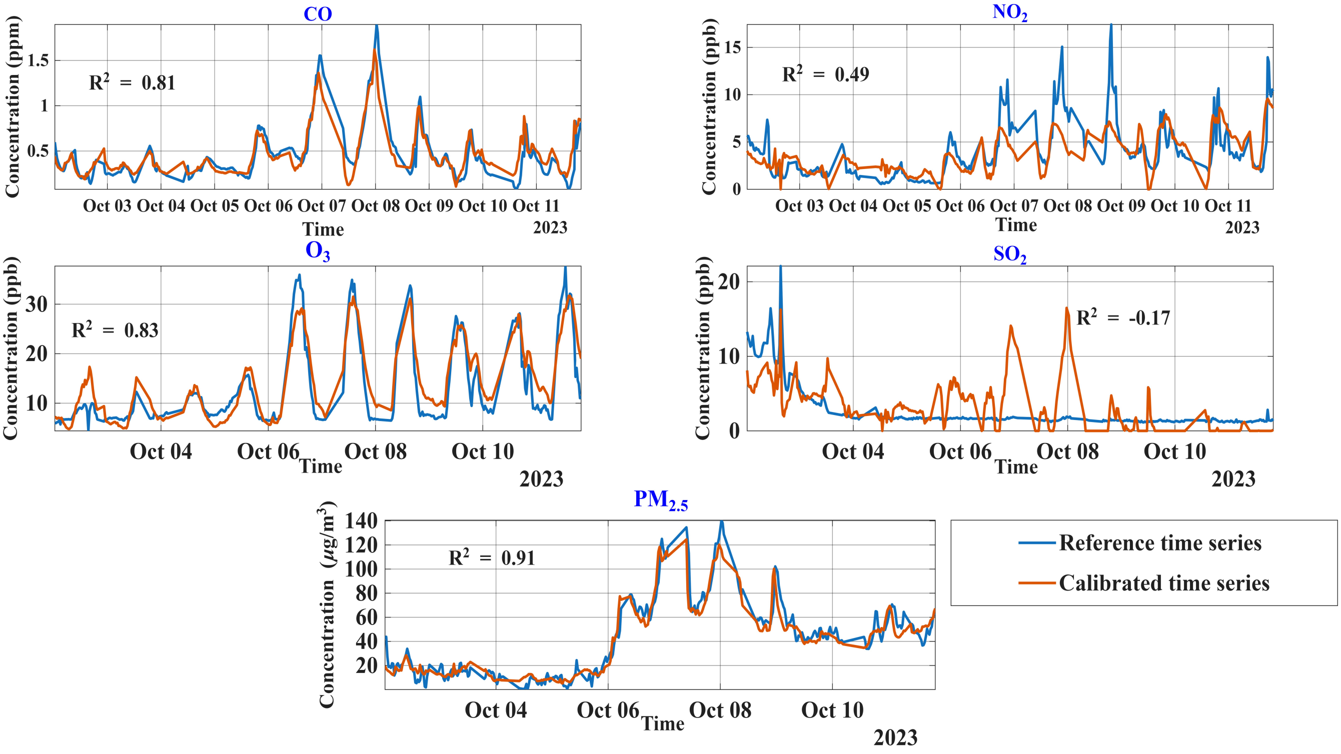


Figure 4. Time series plots of the concentration of various gases and $PM_{2.5}$ at **Site-C** as obtained after calibration along with their concentration as obtained from the Gas Analysers and EBAM (which serve as the reference).

as well as $PM_{2.5}$ as predictors for SA because BC is a specific indicator of primary combustion sources, whereas Secondary Organic Aerosols (SOA) often constitute a significant fraction of $PM_{2.5}$(Lalchandani et al., 2021; Shukla et al., 2021). This happens because abundant gaseous precursors like SO_2, NO_x, NH_3 and VOCs are oxidized and partitioned into the particle phase, forming sulphates, nitrates, ammonium and SOA. The relatively long atmospheric lifetimes and regional transport of these precursors further enable the accumulation of secondary mass in the fine mode. The need to incorporate both BC and $PM_{2.5}$ as LCAQ features is further underscored by the correlation heat map in Fig. 7, as their correlation with some organic and elemental sources differ appreciably.

3.3 RTSA Ground Truth

The details of operation of HR-ToF-AMS and Xact-625i along with PMF-based Source Apportionment methodology are presented in Appendix A as well as in Lalchandani et al. (2021) and Shukla et al. (2021). Both organic as well as elemental source apportionment were carried out on the combined data from the two sites discussed earlier in Section 2 i.e. **Site-B** and **Site-C**. In organic source apportionment, the sources of organic pollutants are identified along with their contribution in polluting the ambient air. Organic aerosols primarily generated from traffic, especially diesel exhaust are termed as Hydrocarbon

https://doi.org/10.5194/egusphere-2025-5677
Preprint. Discussion started: 9 December 2025

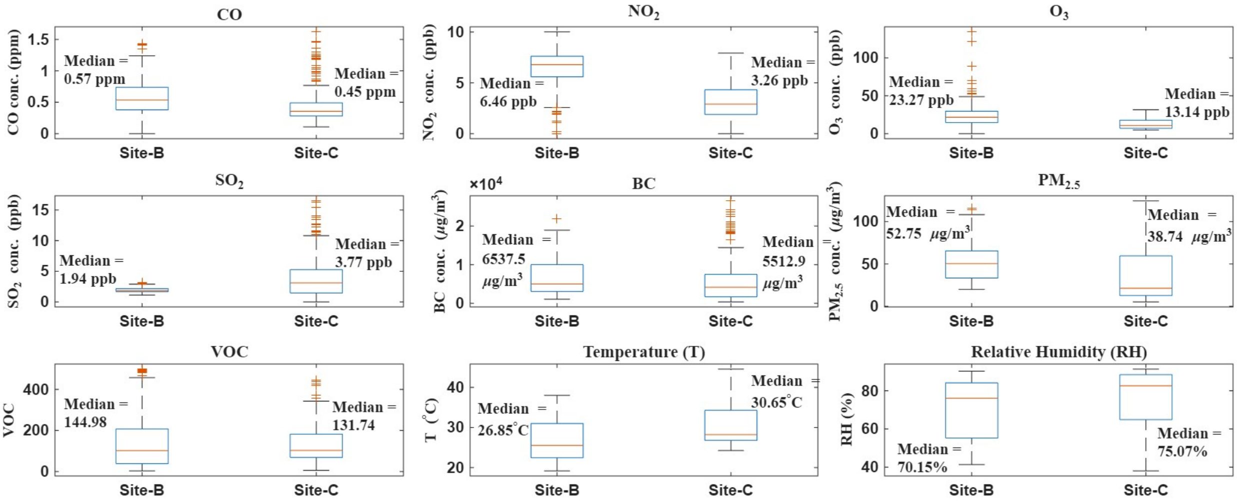
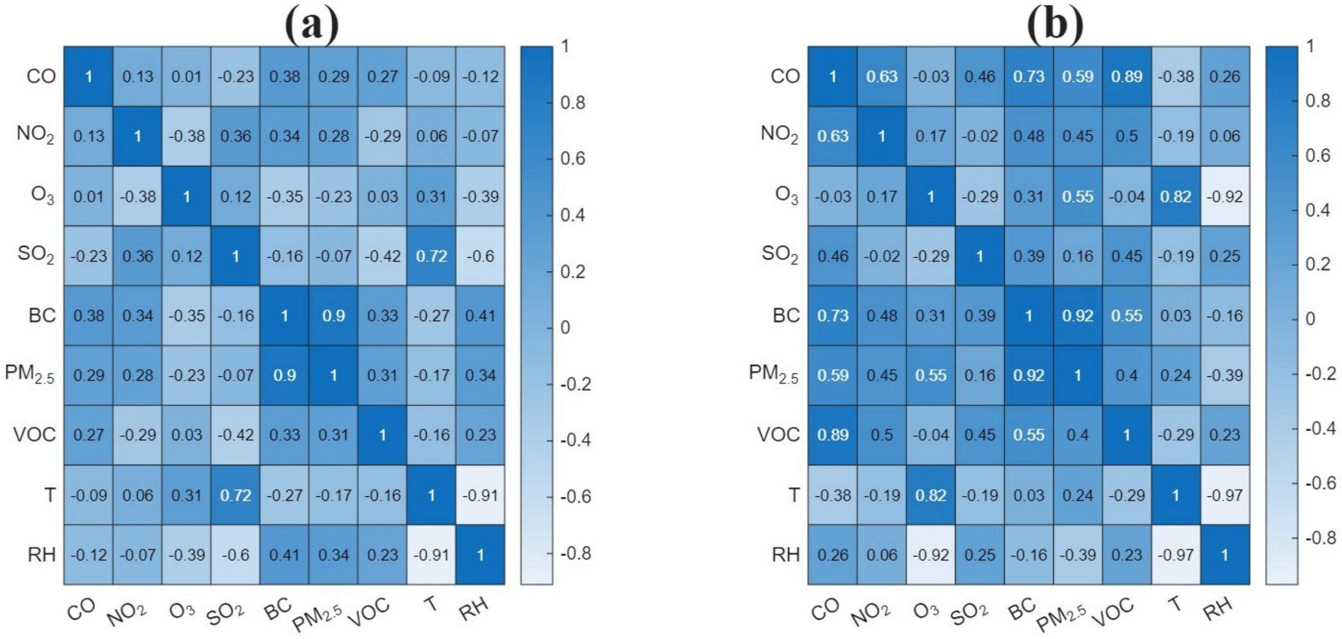


Figure 5. Box and whisker plots showing the distribution of the different features from the training data at **Site-B** and **Site-C**. The box represents the interquartile range (IQR), with the lower and upper whiskers extending to $1.5 \times IQR$ below the first quartile (Q_1) and above the third quartile (Q_3), respectively. Data points beyond the whiskers are plotted individually as outliers.



(a)

	CO	NO_2	O_3	SO_2	BC	$PM_{2.5}$	VOC	T	RH
CO	1	0.13	0.01	-0.23	0.38	0.29	0.27	-0.09	-0.12
NO_2	0.13	1	-0.38	0.36	0.34	0.28	-0.29	0.06	-0.07
O_3	0.01	-0.38	1	0.12	-0.35	-0.23	0.03	0.31	-0.39
SO_2	-0.23	0.36	0.12	1	-0.16	-0.07	-0.42	0.72	-0.6
BC	0.38	0.34	-0.35	-0.16	1	0.9	0.33	-0.27	0.41
$PM_{2.5}$	0.29	0.28	-0.23	-0.07	0.9	1	0.31	-0.17	0.34
VOC	0.27	-0.29	0.03	-0.42	0.33	0.31	1	-0.16	0.23
T	-0.09	0.06	0.31	0.72	-0.27	-0.17	-0.16	1	-0.91
RH	-0.12	-0.07	-0.39	-0.6	0.41	0.34	0.23	-0.91	1

(b)

	CO	NO_2	O_3	SO_2	BC	$PM_{2.5}$	VOC	T	RH
CO	1	0.63	-0.03	0.46	0.73	0.59	0.89	-0.38	0.26
NO_2	0.63	1	0.17	-0.02	0.48	0.45	0.5	-0.19	0.06
O_3	-0.03	0.17	1	-0.29	0.31	0.55	-0.04	0.82	-0.92
SO_2	0.46	-0.02	-0.29	1	0.39	0.16	0.45	-0.19	0.25
BC	0.73	0.48	0.31	0.39	1	0.92	0.55	0.03	-0.16
$PM_{2.5}$	0.59	0.45	0.55	0.16	0.92	1	0.4	0.24	-0.39
VOC	0.89	0.5	-0.04	0.45	0.55	0.4	1	-0.29	0.23
T	-0.38	-0.19	0.82	-0.19	0.03	0.24	-0.29	1	-0.97
RH	0.26	0.06	-0.92	0.25	-0.16	-0.39	0.23	-0.97	1

Figure 6. Heat maps obtained from the correlation coefficient between the various predictors at (a) **Site-B** and (b) **Site-C**

like Organic Aerosol (**HOA**) (Zhang et al., 2005). Organic aerosols originated from both natural as well as anthropogenic
295 burning such as wood, dry leaves, cow-dung burning, stubble burning etc. are termed as Biomass Burning Organic Aerosol

(**BBOA**) (Smith et al., 2009). In our work, we further sub-categorize BBOA as primary (**BBOA-1**) and secondary (**BBOA-2**) depending on whether it is less oxidised or more oxidised respectively (Lakra et al., 2024). SA of organics also resolved **SOA** like Low Volatile Oxygenated Organic Aerosol (**LVOOA**) and Semi Volatile Oxygenated Organic Aerosol (**SVOOA**). In general LVOOA is oxygenated to a much higher degree as compared to SVOOA, and hence there is a higher chance that organic aerosols under LVOOA may have been generated at a distant place and transported to the site under observation by the influence of wind and other factors (Saarikoski et al., 2023).

In elemental source apportionment, the PMF resolved the sources under the following categories viz. Dust, Fireworks, Solid Fuel Combustion (**SFC**), Lead-rich (**Pb-rich**), Sulphur-rich (**S-rich**), Ferrous smelting and Chlorine-rich (**Cl-rich**). SFC is further categorized as **SFC-I** and **SFC-II**, where SFC-I is generated from biomass/wood burning or different kinds of waste incineration and SFC-II is generated from electronic waste incineration of Zinc rich products. Further details about these factors are described in Appendix A.

We are interested in learning a mapping function $f : \mathbb{R}^D \to \mathbb{R}^L$ which can predict the SA target vector y_i (representing the relative abundance of the different sources) using the LCAQ feature vector x_i. The different elements in the SA target vector y_i sum to unity, since they represent the relative contribution of different organic (elemental) sources. Let $\tilde{y}_i$ denote the vector constructed using the SA results obtained from PMF at the i^{th} time instant. The elements of $\tilde{y}_i$ do not sum to unity, as the SA results derived from PMF are not constrained to sum to one. Thus, in order to obtain y_i from $\tilde{y}_i$, we normalize $\tilde{y}_i$ by its L_1 norm to obtain the normalized vector y_i.

$$y_i = \frac{\tilde{y}_i}{||\tilde{y}_i||_1} \tag{3}$$

The task of learning a multi output mapping function $f : \mathbb{R}^D \to \mathbb{R}^L$ is equivalent to the task of learning L separate mapping function of the form $f_l : \mathbb{R}^D \to \mathbb{R}$, and then combining the model outputs. Let f_l denote the regression model learnt using the training data $(x_i, y_i[l]), i \in \{1, 2, ...T\}, l \in \{1, 2, ...L\}$. During model evaluation, let x_{test} denote the test LCAQ feature vector and $\hat{y}_{test}$ denote the final predicted SA target vector. Then $\hat{y}_{test}$ can be written as follows

$$\tilde{y}_{test} = [f_1(x_{test}), f_2(x_{test}), f_3(x_{test}), ..., f_L(x_{test})]$$
$$\hat{y}_{test} = \frac{\tilde{y}_{test}}{||\tilde{y}_{test}||_1} \tag{4}$$

An Internet of Things (IoT) network is used to extract data from the LCAQ sensor units. An Application Programming Interface (API) is set up to store the data from the LCAQ sensor units into a server. This data is pulled at the user end with the help of the API and fed to the data processing and machine learning algorithms. The low cost sensors produce null values at certain times during their operation, due to reasons such as loss of data over the IoT network or sensor malfunctioning. The observations corresponding to such timestamps are discarded in the first step of the data processing stage.

The LCAQ feature vectors were time-aligned with their corresponding SA target vectors based on the recorded timestamps. The HR-ToF-AMS and Xact-625i instruments produced data with a 30-minute time resolution, the EBAM and Micro-Aethalometer provided data at 5-minute intervals, and the Gas Analyzers generated data with a time resolution of 15-minutes. As discussed in Section 3.2, the LCAQ sensor unit data, originally recorded at a 1-minute resolution, were converted to a

https://doi.org/10.5194/egusphere-2025-5677
Preprint. Discussion started: 9 December 2025

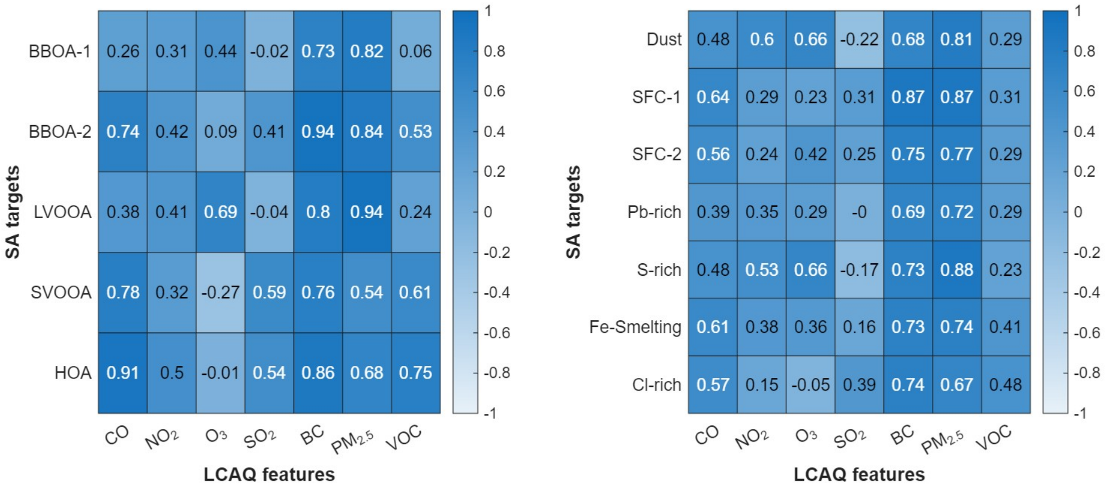


Figure 7. Heatmaps depicting correlations between *LCAQ features* and *SA targets*. It is noticeable that certain attributes like $PM_{2.5}, BC$ etc. are well correlated with the various organic as well as elemental sources. These latent correlations are exploited by the regression-based receptor model to predict relative abundance of the various sources.

30-minute resolution. Since the lowest available time resolution (30 minutes) corresponds to the HR-ToF-AMS, Xact-625i, and LCAQ sensor units, the data from the EBAM, Micro-Aethalometer, and Gas Analyzers were also converted to a 30-minute resolution. This conversion was performed by computing the median of all samples within each 30-minute interval. After time-alignment of the data, we had 364 observations for organic SA and 369 observations for elemental SA at **Site-B**. For **Site-C**, we had 229 observations for organic SA and 351 observations for elemental SA.

The correlation between the different attributes of the predictor vector x_i^{BO} and their corresponding target vector y_i^{BO} are shown in Fig. 7 (a) in the form of heat map, while Fig. 7 (b) shows the heat map for the same set of correlation values between x_i^{BI} and y_i^{BI}. It can be observed that there exist strong correlation between several attributes of x_i and y_i. The attributes $PM_{2.5}$, BC and CO exhibit good correlation with almost all the organic as well as elemental sources and the correlations are in particular better for organic sources as compared to the elemental ones. NO_2 shows moderate correlation with both organic and elemental sources. VOC shows much better correlation with organic sources as compared to the elemental ones, while O_3 shows more consistent and better correlation with most of the elemental sources. These correlations between the *LCAQ features* and *SA targets* provides the motivation to learn regression-based receptor models that translates to learning a vector valued mapping function $f : \mathbb{R}^D \rightarrow \mathbb{R}^L$. Such a mapping function can learn the relation between the LCAQ feature vectors and SA target vectors.

3.4 Metrics used for evaluation

In order to make a quantitative assessment of the performance of the proposed framework, the following metrics are used:

- Mean Absolute Error (MAE) and Standard Deviation of Absolute error (SAE): Let $y_{test}[i][l]$, constitute the ground truth relative abundance in the test set and $\hat{y}_{test}[i][l], i = 1,2,...,N_{test}, l = 1,2,....,L$ denote the corresponding predictions. The variable i refers to the time index of the sample under consideration, while the variable l refers to the index of the source. If $e[i][l]$ denotes the absolute prediction error for the i^{th} sample of the l^{th} source, then the MAE and SAE for the l^{th} source are defined as follows:

$$e[i][l] = |y_{test}[i][l] - \hat{y}_{test}[i][l]|\,,\ i = 1,2,...,N_{test},\ l = 1,2,....,L$$
$$MAE = \frac{\sum_{i=1}^{N_{test}} e[i][l]}{N_{test}}$$
$$SAE = \frac{\sum_{i=1}^{N_{test}} (e[i][l] - MAE)^2}{N_{test}} \tag{5}$$

The MAE and SAE provides an estimation of the accuracy of prediction. The coefficient of determination or R^2 score is another metric which can be used in this scenario to measure the quality of prediction for time series data. However, we prefer not to use this metric because it is affected by the variance of the ground truth time series. Smaller values of variance in the ground truth data tend to amplify even small errors in prediction. For example, if the relative abundance of a particular source tends to remain constant with time in the ground truth data, then even a small error of say 1% in prediction will make $R^2 \rightarrow -\infty$, which is a misleading assessment of the prediction error.

- Spearman's Rank Order Correlation Coefficient ($SROCC$) and Normalized Discounted Cumulative Gain ($NDCG$): The performance of the proposed framework in terms of correctly predicting the order of relative abundance of the various sources is very important and measured with the help of the SROCC metric (Myers et al., 2013). The SROCC metric assigns equal importance to all the sources. However, it is more important that the the sources having the highest relative abundance be recognized in the correct order. Ordering error can be tolerated for the sources which have very less contribution. The NDCG score is a more appropriate metric in this context (Järvelin and Kekäläinen, 2002), which assigns more weightage to the sources having the highest contribution and vice versa. Hence, these two metrics are also used for the performance evaluation of our proposed framework.

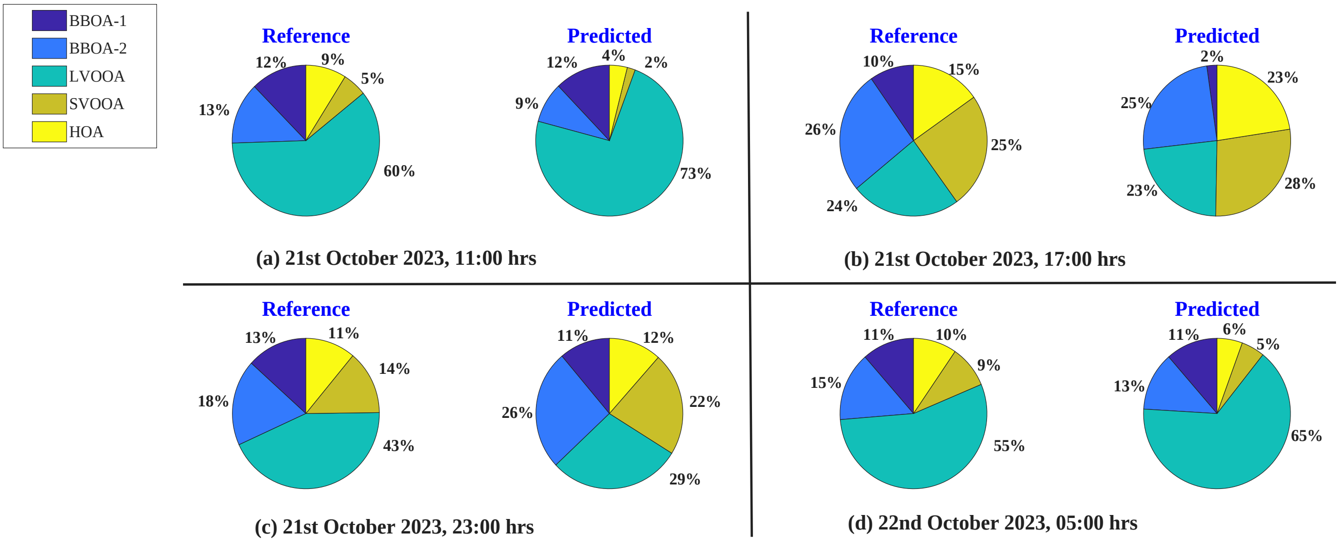
4 Results and Discussion

The performance assessment of the proposed framework is based on how closely the SA results obtained from data captured using RGI like HR-ToF-AMS and Xact-625i can be replicated by the data obtained from LCAQ sensor units. As mentioned earlier, the HR-ToF-AMS performs chemical speciation of the non-refractive organic and inorganic aerosols present in the sampled air in terms of their mass to charge (or m/z) ratio. The chemical speciation of the non-refractive organic aerosols having m/z ratio upto 300, obtained from HR-ToF-AMS is used in our experiment (DeCarlo et al., 2006). Hence, the SA results generated by doing PMF of the time series data captured using the HR-ToF-AMS is considered quite accurate and used as the ground truth for training our supervised machine learning algorithms. Similarly, for elemental source apportionment, the results obtained from PMF of the time series data captured using the Xact-625i serves as the ground truth. The Xact-625i captures the absolute concentration of several trace elements, out of which twenty-two common elements which contaminate the air, and are above the minimum detection limit like *Al, Si, S, Cl, K, Ca, Pb* etc., are used for elemental SA (Shukla et al., 2021). The SA results generated from data captured using the HR-ToF-AMS and Xact-625i serve as the *SA targets* for our proposed framework; while the pollutant gases CO, NO_2, O_3, SO_2 along with $PM_{2.5}$, VOC, T and RH measured using the LCAQ sensor units and BC measured using Micro-Aethlometer serve as the LCAQ features. The problem of learning a mapping function $f : \mathbb{R}^D \rightarrow \mathbb{R}^L$ for mapping the *LCAQ feature* vector x_i to the *SA target* vector y_i is equivalent to learning L separate functions $f_l, l = 1,2,...L$ from $\mathbb{R}^D \rightarrow \mathbb{R}$ as shown in Eqn (4), where $D = 9$. The performance of the learning algorithm in learning the individual functions f_l is assessed based on two different criteria: the first criteria being how closely the predicted relative abundance follows the reference or ground truth relative abundance for the various sources, while the second criteria tries to evaluate whether the predicted relative abundance of the different sources are in the same order as their corresponding ground truth values. The performance based on the first criteria can be evaluated with the help of time-series plots, pie-charts and the MAE/ SAE metric described in Eqn 5, while information regarding efficacy of the proposed method in preserving the correct order of relative abundance of the different sources in the predictions can be obtained from the NDCG and SROCC metric described in Section 3. In the following, the performance of the proposed framework in predicting the organic SA results are discussed first followed by their elemental counterpart. The SA results discussed in the following subsections are obtained by choosing Linear regression models as individual mapping functions $f_l, l = 1,2,....,L$ (Seber and Lee, 2012). The performance of some other popular regression models like k-Nearest Neighbour Regression (Song et al., 2017), Gradient Boosting Regression (Friedman, 2001), Ridge Regression (McDonald, 2009), Random Forest Regression (Segal, 2004), Support vector regression (Awad et al., 2015) are presented in Appendix C. The performance of the proposed framework is less sensitive with respect to the various regression models used for the mapping function as can be observed from the results presented in Appendix C.

4.1 Organic source apportionment

The organic source apportionment results for **Site-B** are discussed first. Figure 8 shows sample pie-charts obtained using the proposed method. Each section of Fig. 8 consists of two pie-charts, one constructed from *SA target* using Eqn (3) and the other obtained from the corresponding predictions using Eqn (4). It can be observed from this figure that the ordering of the top three sources with highest contribution is preserved in two out of the four cases, i.e. in Fig. 8 (c) and (d). In order to get a quantitative estimate of the overall prediction performance in this case, the MAE and SAE as described in Eqn (5) is computed for each of the five components and the results are tabulated in Table 2. Apart from LVOOA, the MAE corresponding to other sources are either less than or around 5%. Since LVOOA has a mean relative abundance of 48.61%, which is much higher compared to the source having the second largest mean relative abundance (i.e. BBOA-2 with 17.09%), thus there is less probability of the order of the sources getting perturbed during prediction. This observation is supported by the results in Table 3, where we get very good SROCC and NDCG scores related to organic source apportionment at **Site-B**.

https://doi.org/10.5194/egusphere-2025-5677
Preprint. Discussion started: 9 December 2025



(a) 21st October 2023, 11:00 hrs

(b) 21st October 2023, 17:00 hrs

(c) 21st October 2023, 23:00 hrs

(d) 22nd October 2023, 05:00 hrs

Figure 8. Sample pie-charts showing organic source apportionment results predicted using the proposed method at different times during the day at **Site-B**.

Table 2. Component wise MAE and SAE computed for BBOA-1, BBOA-2, LVOOA, SVOOA and HOA according to Eqn (5) for deployment at **Site-B**.

		BBOA-1	**BBOA-2**	**LVOOA**	**SVOOA**	**HOA**
Site-B	**MAE (in %)**	2.89	2.87	10.78	5.20	4.39
	SAE (in %)	2.63	2.25	7.96	4.93	3.83
	Mean of reference Fractions (in %)	9.91	17.09	48.61	12.68	11.72

Table 3 shows quantification of the efficacy of the proposed method in predicting the relative abundance of the different organic sources at **Site-B** in the correct order based on the mean SROCC and mean NDCG scores. The NDCG scores have been evaluated for different values of K, where K is the number of sources under consideration. An NDCG score of 0.88 for $K = 1$ indicates that the proposed framework detects the source with the highest contribution with very good efficiency. The efficiency improves for $K = 2, 3$ and for all the sources. The SROCC score is less as compared to the NDCG scores for all the sources, because the SROCC formulation assigns equal importance to all the sources irrespective of their relative abundance. A mean SROCC value of 0.77 also indicates that the method performs quite well in predicting the relative abundance of the

https://doi.org/10.5194/egusphere-2025-5677
Preprint. Discussion started: 9 December 2025

Table 3. Efficacy of the proposed method in predicting the relative abundance of the different components in the correct order for organic source apportionment at **Site-B**

	Mean Normalized Discounted Cumulative Gain (NDCG) in predicting the top K contributing sources				**Mean SROCC**
	K=1	***K=2***	***K=3***	**For all sources**	
Site-B	0.88	0.92	0.95	0.97	0.77

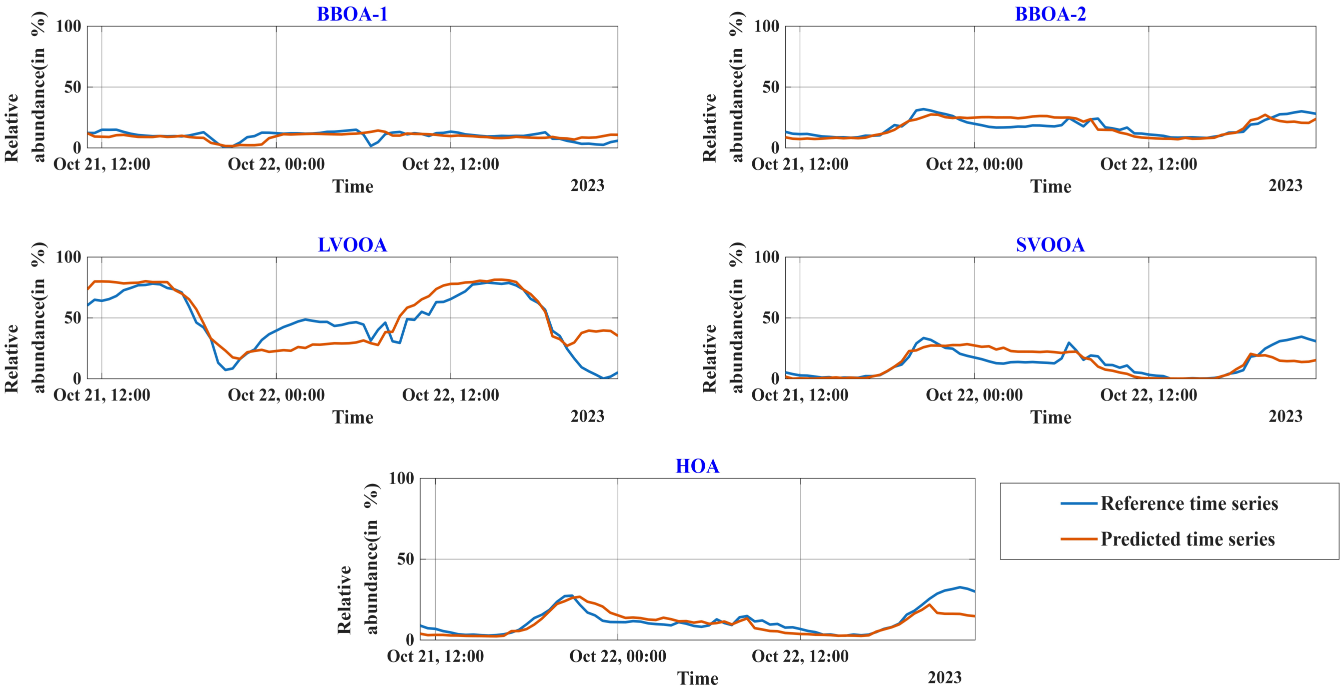
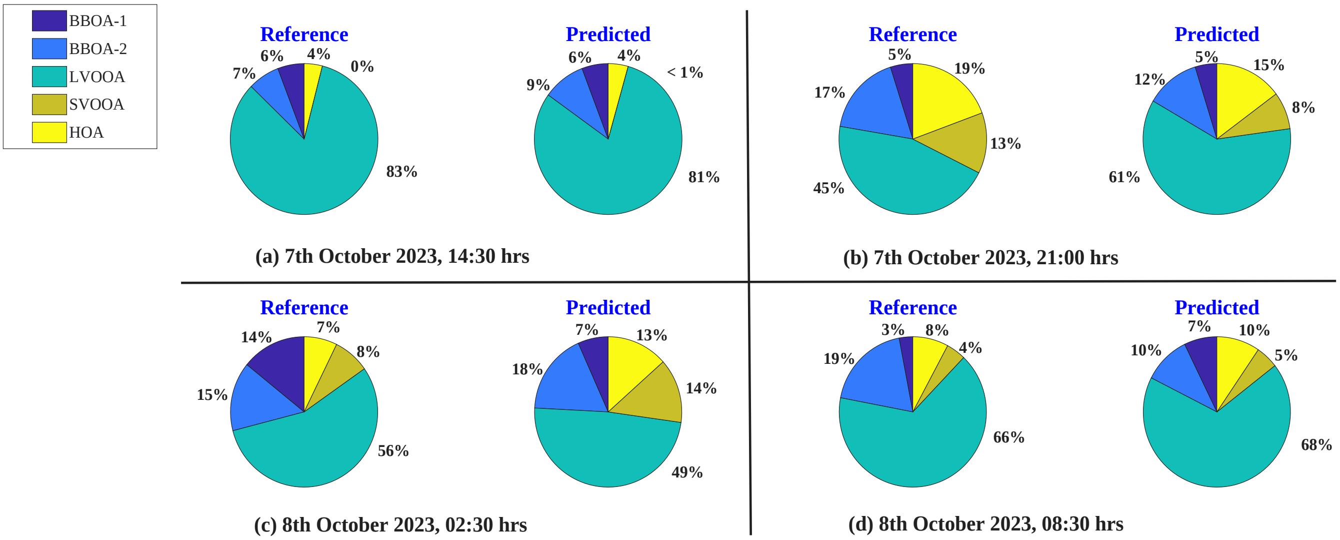


Figure 9. Reference time series plots showing the relative abundance of organic sources along with their corresponding predictions obtained using the proposed method at **Site-B**.

different sources in the correct order. Figure 9 shows the reference time series plots for the relative abundance of organic sources along with their corresponding predictions obtained using the proposed method at **Site-B**. The predicted time series plots for the relative abundance of all the organic sources at **Site-B** are in good agreement with their corresponding reference time series plots, which shows the good performance of the proposed approach in predicting the relative abundance of the different sources as a function of time.

Figure 10 shows sample pie-charts obtained using the proposed method at **Site-C**. It can be observed from these plots that the predicted pie-charts closely match the reference pie-charts at various time during the day. Table 4 shows the MAE and



(a) 7th October 2023, 14:30 hrs

(b) 7th October 2023, 21:00 hrs

(c) 8th October 2023, 02:30 hrs

(d) 8th October 2023, 08:30 hrs

Figure 10. Sample pie-charts showing organic source apportionment results predicted using the proposed method at different times during the day at **Site-C**.

SAE computed according to Eqn (5). Maximum MAE and SAE of 7.41% and 6.34% respectively are obtained for LVOOA, while these metrics are much lower for the other sources. Since LVOOA has a relative abundance which is much higher than the other sources for a major part of the day, hence error of around $10-15\%$ is not sufficient to cause an error in detecting the source with highest contribution.

Table 5 tabulates the mean SROCC and the mean NDCG scores for $K = 1,2,3$ and for all the sources. Mean NDCG score of unity for $K = 1$ indicates that in this case, the source with the highest relative abundance is always predicted correctly. Mean SROCC of 0.8 and mean NDCG scores close to unity for other values of K implies that the sources with the highest contribution are detected correctly for most of the times. The erroneous cases arises for the sources with the lowest contribution. Figure 11 shows the reference time series plots for the relative abundance of different sources along with their corresponding predictions obtained using the proposed method for organic source apportionment at **Site-C**. The predicted time series plots showing the relative abundance of all the organic sources at **Site-C** are in good agreement with their corresponding reference time series plots, which shows the good performance of the proposed approach in predicting the relative abundance of the different sources as a function of time.

4.2 Elemental source apportionment

For elemental source apportionment, we have $L = 7$ sources as mentioned in Section 3. Proceeding in the same manner as for the case of organic source apportionment in Subsection 4.1, we first discuss the results of **Site-B** followed by the results of **Site-C**. Figure 12 shows sample pie-charts corresponding to different times during the day comparing the reference source apportionment results with the predicted ones at **Site-B**. The three sources with the highest contribution have been predicted

Table 4. Component wise MAE and SAE computed for BBOA-1, BBOA-2, LVOOA, SVOOA and HOA according to Eqn (5) for deployment at **Site-C**.

		BBOA-I	**BBOA-2**	**LVOOA**	**SVOOA**	**HOA**
Site-C	**MAE (in %)**	3.44	3.96	7.41	4.03	2.72
	SAE (in %)	2.42	2.61	6.34	3.22	2.37
	Mean of reference Fractions (in %)	8.43	13.22	58.30	8.63	11.42

Table 5. Efficacy of the proposed method in predicting the relative abundance of the different components in the correct order for organic source apportionment at **Site-C**

	Mean Normalized Discounted Cumulative Gain (NDCG) in predicting the top K contributing sources				**Mean SROCC**
	K = 1	**K = 2**	**K = 3**	**For all sources**	
Site-C	1	0.99	0.98	0.99	0.8

with good accuracy in Fig. 12 (b), (c) and (d) with a maximum error of less than 10%. Table 6 shows the MAE and SAE for each of the sources. The three major sources as observed from Fig. 12, i.e. Dust, SFC-1 and S-rich show MAE and SAE less than 10%, which is quite good for all practical purposes.

Table 7 shows the mean SROCC and mean NDCG scores for various values of K. Mean SROCC of 0.84 along with mean NDCG scores larger than 0.9 for all the cases shows that the proposed method has performed well in predicting the sources in the correct order of their relative abundance. The time series plots for the various sources are shown in Fig. 13. From the plots, it is evident that Dust, SFC-1, and S-rich are the dominant contributing sources, and the predicted time-series closely replicate both the temporal variations and the mean level observed in the reference data for all these three sources.

Figure 14 shows reference and predicted pie-charts for elemental source apportionment at **Site-C**. It can be observed that the top three contributing sources are the same in all the four cases and the order of each these three sources in terms of relative abundance are also the same in the reference and predicted pie-charts. The observations in these pie-charts are a testament to the results tabulated in Table 9. A value of 0.87 for SROCC and NDCG values close to unity for $K = 1,2,3$ shows that the proposed method has very accurately predicted the different elemental sources in the correct order of their relative abundance

https://doi.org/10.5194/egusphere-2025-5677
Preprint. Discussion started: 9 December 2025

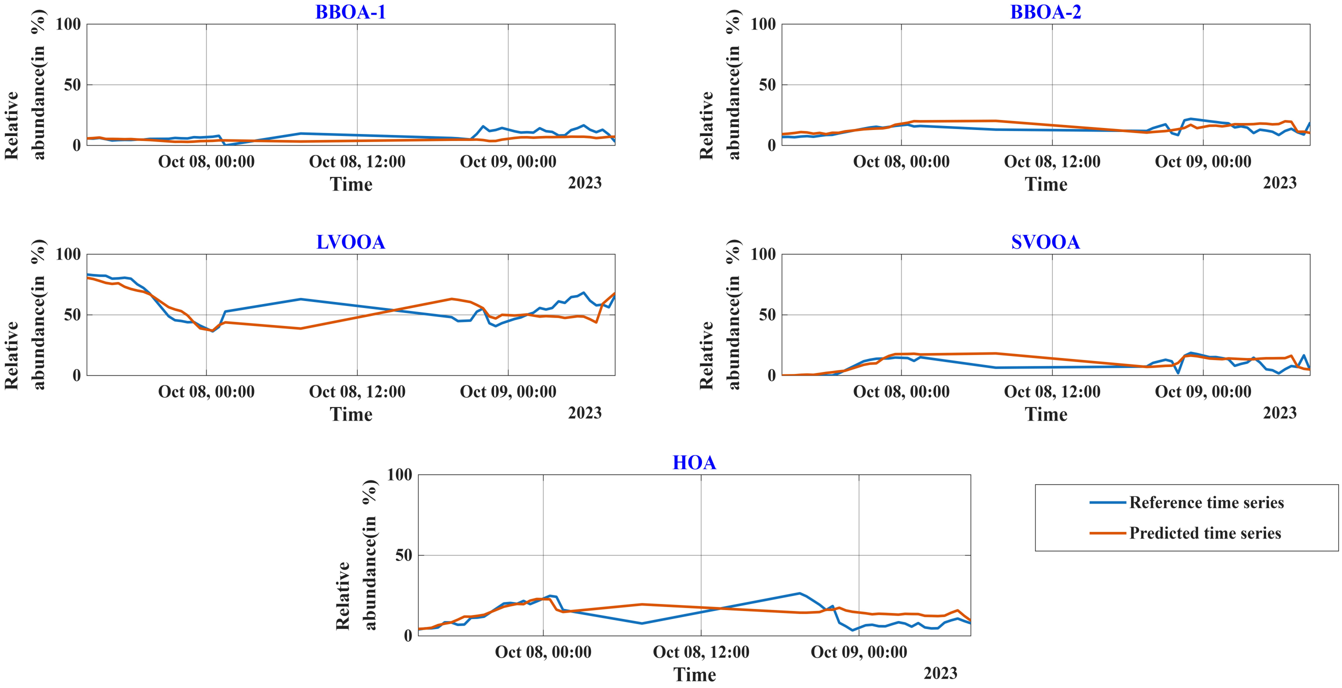
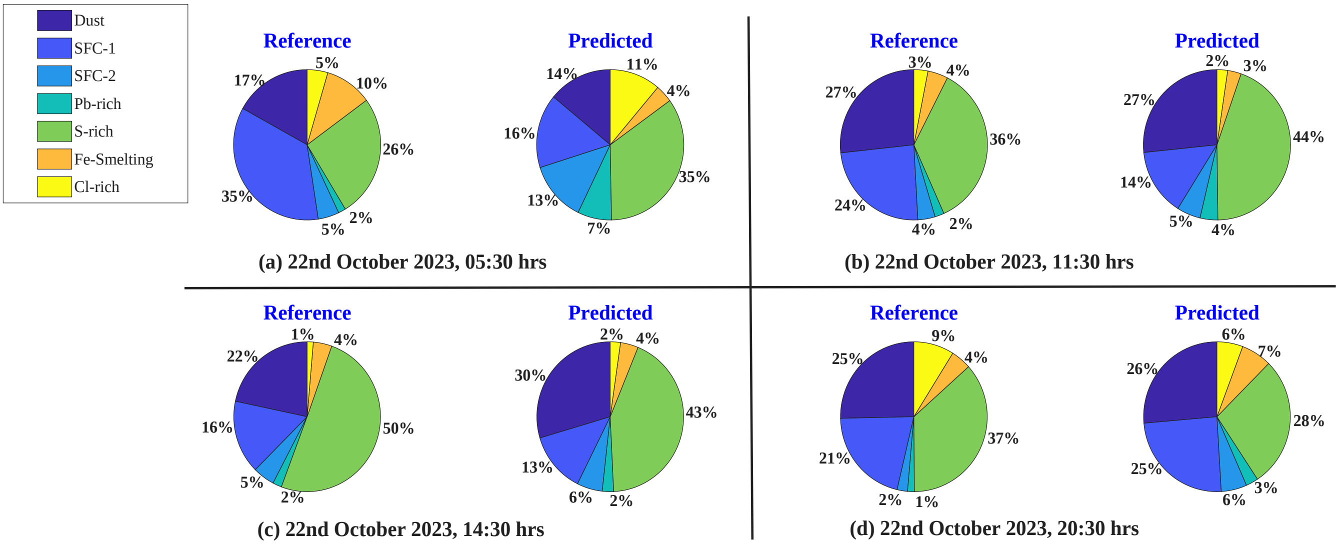


Figure 11. Reference time series plots showing the relative abundance of organic sources along with their corresponding predictions obtained using the proposed method at **Site-C**.



(a) 22nd October 2023, 05:30 hrs

(b) 22nd October 2023, 11:30 hrs

(c) 22nd October 2023, 14:30 hrs

(d) 22nd October 2023, 20:30 hrs

Figure 12. Sample pie-charts showing elemental source apportionment results predicted using the proposed method at different times during the day at **Site-B**.

Table 6. Component wise MAE and SAE computed for Dust, SFC-1, SFC-2, Pb-rich, S-rich, Ferrous Smelting and Cl-rich according to Eqn (5) for deployment at **Site-B**.

		Dust	**SFC-I**	**SFC-II**	**Pb-rich**	**S-rich**	**Ferrous Smelting**	**Cl-rich**
Site-B	**MAE (in %)**	4.91	6.86	2.48	2.12	8.54	1.67	2.62
	SAE (in %)	4.78	5.04	2.00	1.65	5.70	1.53	1.91
	Mean of reference Fractions (in %)	26.44	23.75	3.84	1.70	35.17	5.00	4.10

Table 7. Efficacy of the proposed method in predicting the relative abundance of the different components in the correct order for elemental source apportionment at **Site-B**

	Mean Normalized Discounted Cumulative Gain (NDCG) in predicting the top K contributing sources				**Mean SROCC**
	K=1	**K=2**	**K=3**	**For all sources**	
Site-B	0.91	0.94	0.97	0.97	0.84

at **Site-C**. Table 8 shows the MAE and SAE obtained at **Site-C**. A maximum MAE of 6.67% is also a testament to the good performance of the proposed method for elemental source apportionment at **Site-C**. Figure 15 shows the corresponding reference and predicted time series plots for each of the elemental sources at **Site-C**. Similar to **Site-B**, Dust, SFC-1 and S-rich turn out to be the highest contributing sources for most of the time at **Site-C** too and the temporal variations as well as the mean level in the reference time-series plots are well replicated by the predicted ones for all these three sources.

In linear regression, we attempt to approximate the target value as a linear combination of the predictors. The coefficients in this linear combination are estimated during the training phase by minimizing an objective function such as squared error loss or Huber loss. The magnitude of the coefficient obtained for a particular predictor in predicting the relative abundance of a source is an indicator of the weight assigned to it for the prediction. In our case, since we are using the same set of predictors for predicting all the sources, hence sources which have a higher contribution or higher relative abundance will result in coefficients which have higher magnitude. Our objective is to understand the relative contribution of each predictor in predicting a particular source. In order to achieve this objective, we normalize the coefficient magnitude corresponding to

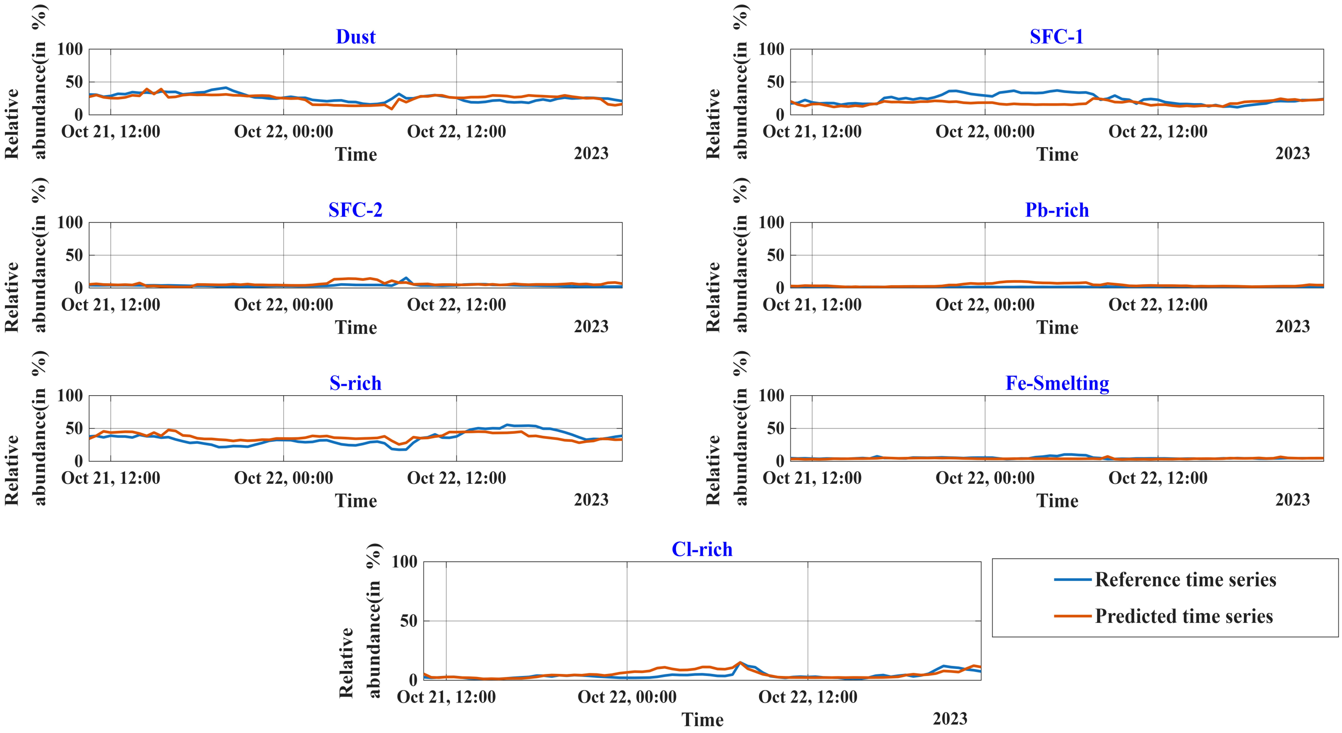


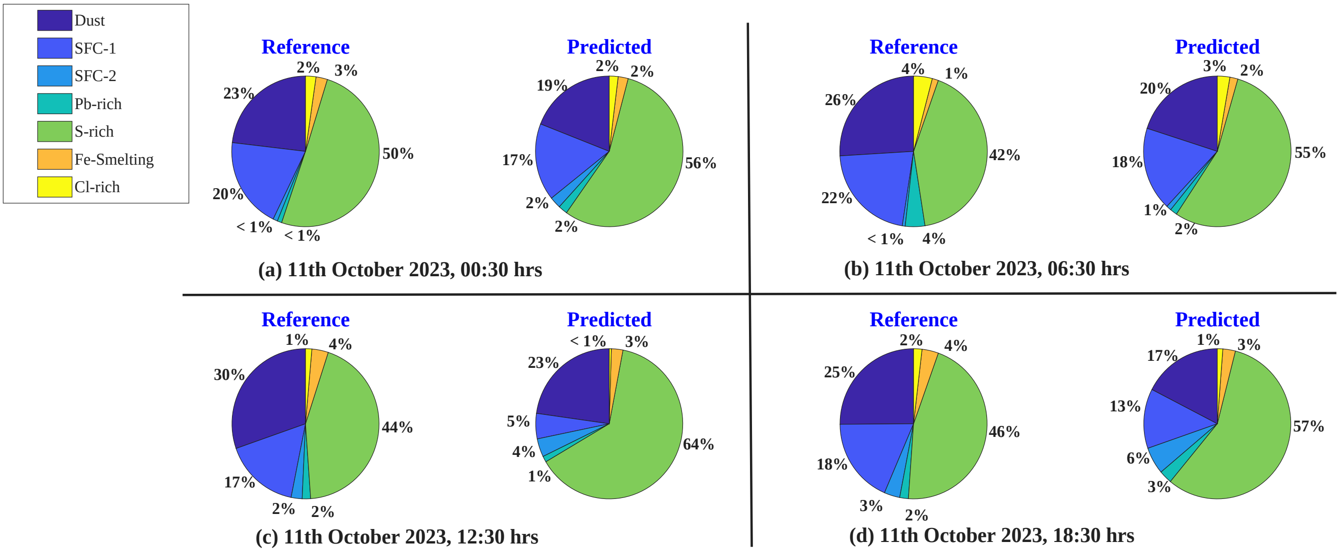
Figure 13. Reference time series plots showing the relative abundance of elemental sources along with their corresponding predictions obtained using the proposed method at **Site-B**.

Table 8. Component wise MAE and SAE computed for Dust, SFC-1, SFC-2, Pb-rich, S-rich, Ferrous Smelting and Cl-rich according to Eqn (5) for deployment at **Site-C**.

		Dust	SFC-I	SFC-II	Pb-rich	S-rich	Ferrous Smelting	Cl-rich
Site-C	**MAE (in %)**	4.48	6.67	1.25	1.01	6.49	1.25	2.41
	SAE (in %)	3.42	4.98	0.69	1.01	5.19	0.81	1.90
	Mean of reference Fractions (in %)	26.38	13.70	1.54	2.10	52.3	2.69	1.30

each predictor with respect to the sum of the coefficient magnitudes of all the predictors for that particular source. Figure 16 shows the plot of the magnitude of normalized coefficients obtained after learning the linear regression model as the mapping

https://doi.org/10.5194/egusphere-2025-5677
Preprint. Discussion started: 9 December 2025



(a) 11th October 2023, 00:30 hrs

(b) 11th October 2023, 06:30 hrs

(c) 11th October 2023, 12:30 hrs

(d) 11th October 2023, 18:30 hrs

Figure 14. Sample pie-charts showing elemental source apportionment results predicted using the proposed method at different times during the day at **Site-C**.

Table 9. Efficacy of the proposed method in predicting the relative abundance of the different components in the correct order for elemental source apportionment at **Site-C**

	Mean Normalized Discounted Cumulative Gain (NDCG) in predicting the top K contributing sources				**Mean SROCC**
	K=1	**K=2**	**K=3**	**For all sources**	
Site-C	1	0.99	0.99	0.99	0.87

function for the various organic as well as elemental sources at **Site-B** and **Site-C**. In the case of organic SA, it can be observed that the coefficients show relatively higher magnitudes for CO, $PM_{2.5}$, BC, T and RH for almost all the sources. The high latent correlation between CO, $PM_{2.5}$ and BC with most of the organic sources as observed from Fig. 7 could serve as the possible reason for the higher magnitude of the corresponding coefficients. The operation of the different sensors in the LCAQ sensor units are sensitive to T and RH and hence the coefficients related to these predictors also have higher value. In the case of elemental SA also, $PM_{2.5}$ and BC were found to be the predictors which had higher correlation with most of the sources as observed from Fig. 7. Among the various elemental sources, S-rich was observed as the most significant contributor from Fig. 12 and Fig. 14. The coefficients associated with $PM_{2.5}$, BC, SO_2, T and RH were found to be the most significant ones in terms of their magnitudes in predicting S-rich sources. It can be observed from Fig. 16 (c) and (d) that, the magnitudes of the normalized coefficients are more uniformly distributed in the elemental case as compared to the organic case. It is expected,

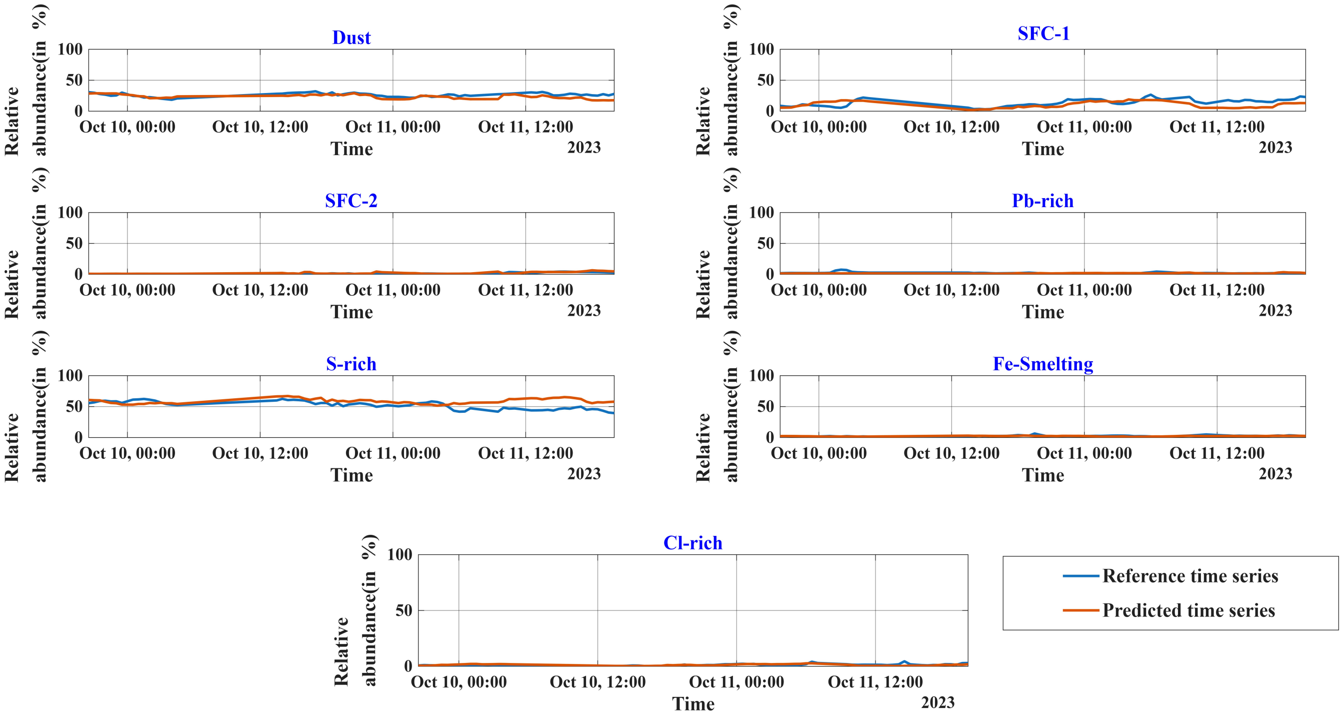


Figure 15. Time series plots showing the reference relative abundance of the elemental sources along with their corresponding predictions obtained using the proposed method at **Site-C**.

since in the case of elemental SA the different sources are correlated to any predictor to a much lesser extent as compared to the organic SA case. Hence, the linear regression model in the case of elemental SA cannot give too much importance to one or a small group of predictors leaving out the rest.

4.3 Discussion

As observed from Subsection 4.1 and 4.2, the SA results predicted by the proposed machine learning based framework show close resemblance to that obtained from data captured using RGI like HR-ToF-AMS and Xact-625i. In case of organic SA, all sources except LVOOA show MAE and SAE less than or around 5% for both the sites. As mentioned in Section 3, LVOOA is a source which is oxygenated to a higher degree as compared to SVOOA and others and hence there is a higher chance that the point of origin of such sources is at a large distance from the site under observation(Lakra et al., 2024; Tobler et al., 2020). It was observed from Table 4 that LVOOA has the highest mean relative abundance of 58.30% at **Site-C**. Since **Site-C** is a background site and there are a number of large and small scale industries (of steel, cement etc.) at a distance of few kilometres, these industries seem to be the major source of air pollution in the area surrounding **Site-C**. Even though **Site-B** also has small and large scale industries surrounding it, the mean relative abundance of LVOOA at **Site-B** is less as compared to **Site-C**, because it is a traffic related site and hence the contribution of other sources also becomes significant. Since, the

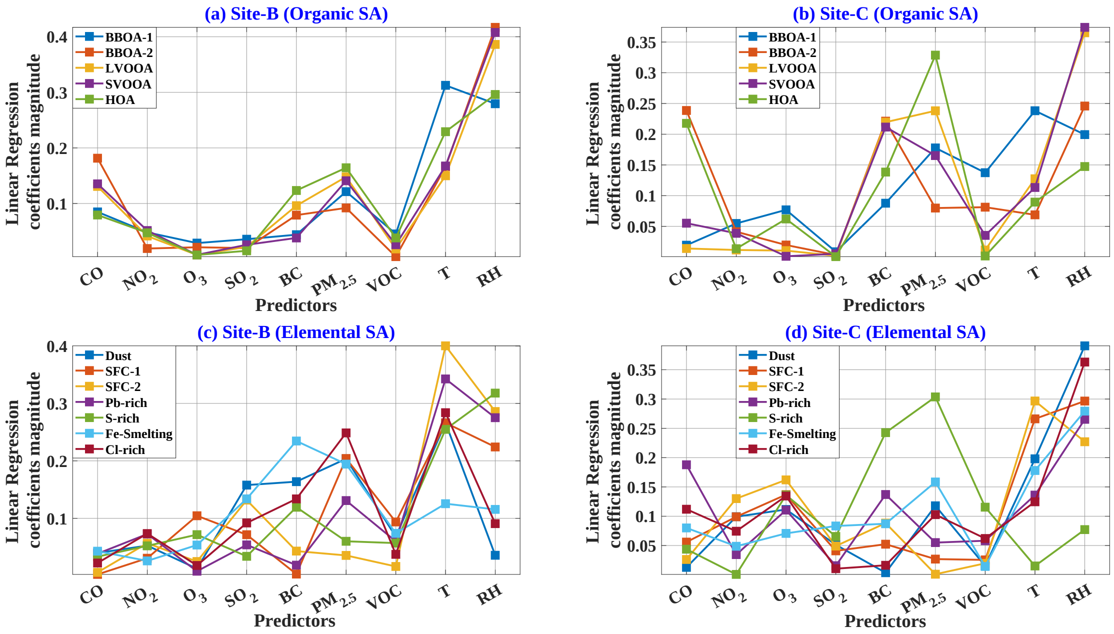


Figure 16. Magnitude of the coefficients obtained for the various predictors after linear regression for the different organic and elemental sources at **Site-B** and **Site-C**.

relative abundance of LVOOA dominates that of other sources by a large margin in case of **Site-C**, hence the NDCG score for $K = 1$ is unity in this case. However, the same metric has a value of 0.88 for **Site-B**, since the dominance of LVOOA as the source with the highest relative abundance is less. In the case of elemental SA, it turns out that the sources S-rich, Dust 495 and SFC-1 have the highest relative contribution. Here also the dominance of S-rich source can be attributed to the presence of industries manufacturing steel, cement etc. at a few kilometres distance from both the sites (Nidheesh and Kumar, 2019). Apart from the NDCG scores, the SROCC scores are also higher for both the sites as compared to their organic counterpart. This implies that in most of the cases the sources are predicted in the correct order of their relative abundance. In order for the policy makers to enact regulatory measures to curb air pollution, it is very important that they have the knowledge about 500 the dynamics of the main contributing sources. The results presented in this section clearly demonstrate the efficacy of the proposed framework in identifying the main sources contributing to air pollution with good precision at a particular location and as a function of time.

5 Conclusion

In this paper, a machine learning based framework for real-time SA is proposed. The proposed framework is based on using 505 multi output regression models on data obtained from LCAQ sensor units to predict the relative abundance of the different organic as well as elemental sources of air pollution. RGI like the HR-ToF-AMS and Xact-625i are used for training the multi output regression models for SA. Other RGI like Gas Analysers and EBAM were used for calibrating the LCAQ sensor data. The set up for the experiments consisted of the entire group of RGI and LCAQ sensor units mounted on a mobile van, so that the same set of experiments can be conducted at any location over a wide geographical area (such as a district) and at any time of the year without the need to dismantle and reassemble the set up time and again. The current work reports the performance of the proposed machine learning framework on two sites, one of which is a background site, while the other is a traffic related site. Different multi output regression models were used to assess the performance of the proposed framework. The performance of most of these models were close to each other with linear regression models providing the best performance in most of the cases. The performance assessment was done based on mainly two sets of metrics: the first set of metrics which include the MAE and SAE evaluate the accuracy of the predicted relative abundances with respect to the reference ones; while the second set of metrics which include the NDCG and SROCC scores provide a quantification of the ability of the proposed framework to predict the different sources in the correct order of their relative abundance. Very good MAE and SAE values of less than 5% have been obtained for most of the sources; while mean NDCG scores greater than 0.9 and mean SROCC greater than 0.75 have been obtained for both organic as well as elemental source apportionment at both the sites. In the current set of experiments, separate machine learning models have been learnt for the two sites. In the future, it is planned to develop transferable machine learning models, so that the same model is generalized across different locations as well as times.

https://doi.org/10.5194/egusphere-2025-5677
Preprint. Discussion started: 9 December 2025

Appendix A: Details of PMF-based RTSA using RGI Data

In this section we will discuss about the SA results using real-time measured data from RGI like HR-ToF-AMS and Xact-625i. As mentioned in Section 4 in the main paper, the HR-ToF-AMS measures the m/z spectra of the non-refractive organic and inorganic components present in the sampled air, while the Xact-625i measures the concentration of trace elements present in sampled air. The time series data obtained from each of these two instruments is then separately fed into PMF algorithm, which expresses the time series matrix as a product of two separate matrices: one matrix contains the profile of the different sources, while the other matrix contains the strength or concentration of the different sources as a function of time. PMF is a bilinear unmixing receptor model with non-negative constraints, widely employed for SA of ambient measurements (Paatero and Tapper, 1994). More details can be found in Lakra et al. (2024), Lalchandani et al. (2021), and Shukla et al. (2021). The output of the PMF algorithm in the context of our experiment is presented below both for the organic as well as elemental case. It is to be noted that in the context of the discussion in the next two subsections, the term *factor* is used to represent a "source", since it is the standard followed in PMF literature.

A1 Organic SA

The PMF model was applied on the Unit Mass Resolution (UMR) data obtained as output from the SQUIRREL (SeQUential Igor data RetRiEvaL, version 1.66B) v1.61 in Igor Pro 9.0.2.4 (Wave-metrics, Portland, USA) (Lakra et al., 2024). The 30 min average data of the two sites obtained form the SQUIRREL with m/z values in the range $1-300$ was used as input to the PMF. The $3-8$ factor solution was tested based on the Q/Q_{exp} ratio and scaled residuals. Based on this identification of clear factors was done. The final selected model consisted of five organic factors, with three secondary aerosol factors and two primary aerosol factors. The secondary organic aerosol factors, also known as Oxygenated Organic Aerosols (OOAs) include highly oxygenated factors known as Low Volatile OOA (LVOOA), freshly oxidized factor known as Semi Volatile OOA (SVOOA) and highly oxidized Biomass Burning Organic Aerosol (BBOA-2). Among the primary factors are Hydrocarbon like Organic Aerosols (HOA) related to the traffic and less oxidized Biomass Burning Organic Aerosol (BBOA-1). The m/z spectra of these different factors are shown in Fig. A1, while the time series plots of the different factors for both the sites are shown in Fig. A2. The m/z spectra of LVOOA is characterized by a sharp peak at $m/z = 44$ and negligible content beyond $m/z = 44$ (Ng et al., 2010). $m/z = 44$ represents the aged species in factor. The SVOOA is characterized by peaks at $m/z = 44$, 53, 67, 81 and 109 along with the aromatic fragments at $m/z = 63$. BBOA is characterized by the presence of signals at the $m/z44$, 60 and 73 which are identified as fragments from the anhydrous sugar present in the biomass smoke (Zhu et al., 2018). BBOA is further categorized as BBOA-1 which is less oxidized and BBOA-2 which is more oxidized. The HOA factor is characterized by the presence of the alkyl fragment signature with predominated signals at $m/z = 55$ and $m/z = 57$ with the absence of $m/z = 60$ (Ulbrich et al., 2009). SA carried out for the two sites results in the sources such as traffic (HOA), biomass burning (BBOA) and two secondary organic aerosol factors (LVOOA and SVOOA). Their diurnal variation reveals that the secondary aerosols such as LVOOA and SVOOA are higher during the daytime due to dominant photochemical processes and at night the primary sources such as HOA and BBOA-1 are dominating (Kumar et al., 2016). The comparison between the two sites show

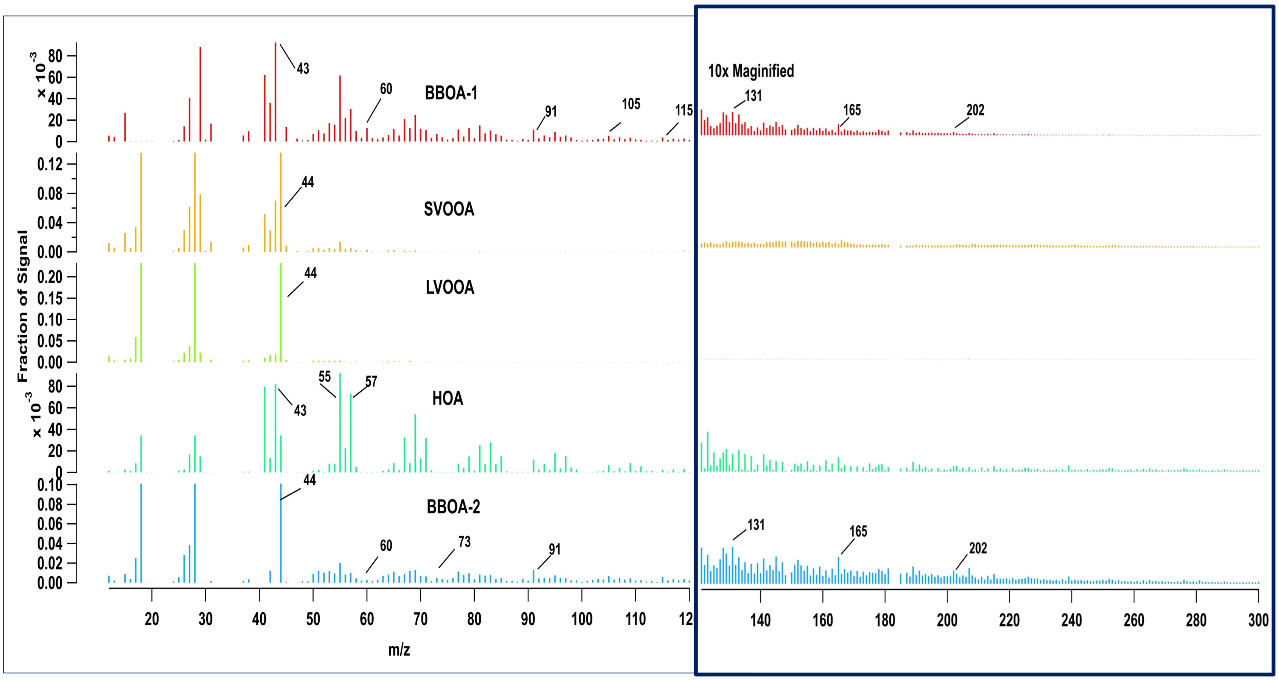


Figure A1. m/z spectra for the different organic factors over **Site-B** and **Site-C** during the measurement period.

that secondary organic aerosol factors concentration over **Site-C** is higher as compared to **Site-B** as **Site-C** is away from the anthropogenic activities of the city, where there are chances that particles from long-range transport are measured in this site which are changed to secondary form due to various meteorological factors and presence of other radicals in air (Lakra et al., 2024). However, the primary sources contribution is dominant at the site B due to near proximity to the local anthropogenic activities such as industrial and traffic.

A2 Elemental SA

Combined input data matrix obtained from **Site-B** and **Site-C** of the measured elements with 30-minute time resolution was prepared. Elements were first filtered based on the percentage of data points below their Minimum Detection Limit (MDL) (Cooper Environmental Services). PMF was applied to this elemental dataset, and a range of $3-10$ factor solutions were tested. A seven-factor solution was selected based on the Q/Q_{exp} ratio and scaled residuals, which resulted in identification of clear source factors: Dust, SFC-I, Pb-Rich, S-Rich, Ferrous Smelting, Cl-Rich, and SFC-II. Among these, more than 50% was contributed by S-rich, Dust and SFC-1 factors, whereas Cl-rich, Ferrous Smelting, Pb-rich, and SFC-2 have lower contribution at both sites. The profile of these different factors are shown in Fig. A3, while the time series plots of the different factors for both the sites are shown in Fig. A4.

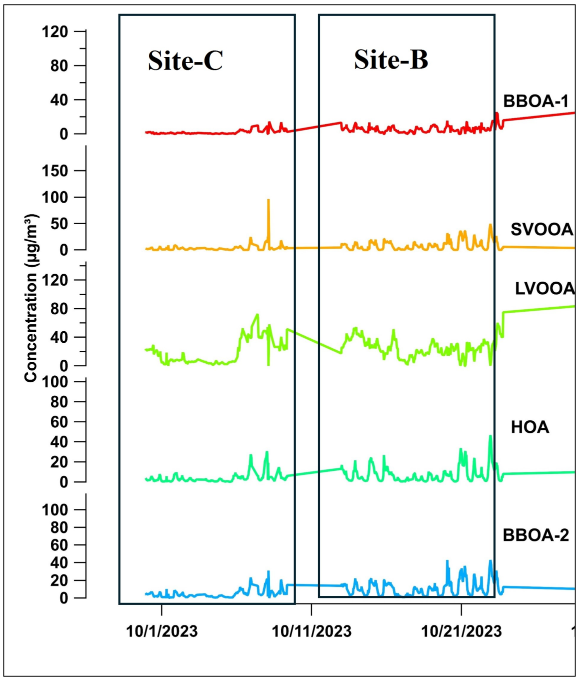


Figure A2. Time series plots for the different factors of organic aerosol obtained after PMF over **Site-B** and **Site-C**.

- **Dust Factor:** The relative contributions of elements to this factor are as follows: Silicon (Si) (80%), K (28%), Calcium (Ca) (83%), Titanium (Ti) (50%), Vanadium (V) (45%), Iron (Fe) (45%), Nickel (Ni) (35%), Rb (20%) and Sb (15%). However, the factor's mass is mainly dominated by Si (43%), Ca (25%), and Fe (20%). Si, Ca, and Ti are commonly used as tracers for dust-related sources (Pant et al., 2015; Rai et al., 2020a). The dust factor accounts for approximately 25% of contributions at both sites.

- **SFC-I Factor:** The total mass fraction of this factor is dominated by Potassium (K) at 65%, with additional contributions from As, Se, and Rubidium (Rb), which are widely recognized as markers for biomass burning (Rai et al., 2020b, a). Other potential sources of K, As, and Se include coal and wood combustion. The SFC-1 factor has a higher contribution at **Site-B** as compared to **Site-C**.

- **SFC-II Factor:** The relative percentage contribution and total mass fraction of this factor is dominated by Zinc (Zn) (80%). Zn is the tracer for coal combustion and can also be emitted from traffic related, and waste incineration (Pant et al., 2015). It has lowest contribution at both sites due to restriction in waste burning in the surrounding location.

- **Pb-Rich Factor:** This factor is predominantly characterized by Pb (80%), commonly used as a tracer for coal combustion and steel production (Li et al., 2012). Contributions made4 by this factor at both the sites were similar.

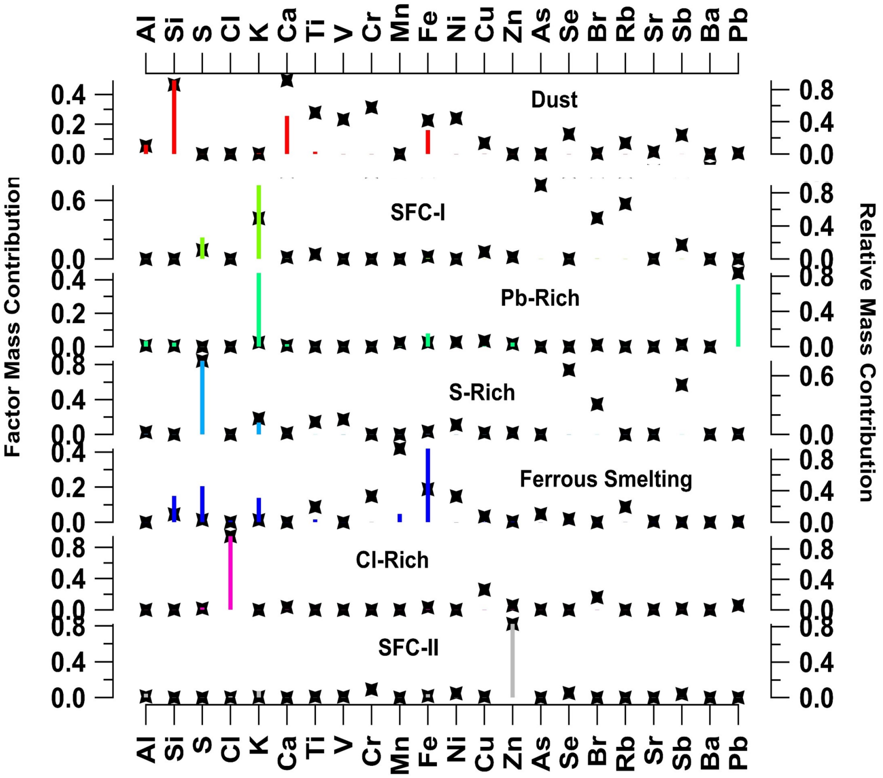


Figure A3. Factor profile for the different elemental sources obtained after PMF over **Site-B** and **Site-C**.

- **S-Rich Factor:** The elements which dominate the relative mass contribution in this factor are Sulphur (S) (80%), Selenium (Se) (60%), Antimony (Sb) (50%) and Bromine (Br) (25%). The total mass fraction of the factor is primarly driven by S. Elements such as S, Arsenic (As), Se and Sb are well-established markers for coal combustion (Rai et al., 2020b). This factor contributes the highest to the total elemental mass at both sites.

- **Ferrous Smelting Factor:** This factor is primarily dominated by Fe (40%) with relative contributions from Manganese (Mn), Chromium (Cr) and Nickel (Ni) which has the likelihood of originating from industrial sources or non-exhaust traffic emissions, such as brake or tire wear. Its overall contribution was minimal at both sites.

- **Cl-Rich Factor:** This factor exhibits the highest mass contribution of chloride (85%) with additional contributions from Bromine (Br) and copper (Cu) (approximately 20%). The Cl-Rich factor shows a higher contribution at **Site-B** compared to **Site-C**. This difference may be attributed to the lower metal concentrations at **Site-C**, which is a background

https://doi.org/10.5194/egusphere-2025-5677
Preprint. Discussion started: 9 December 2025

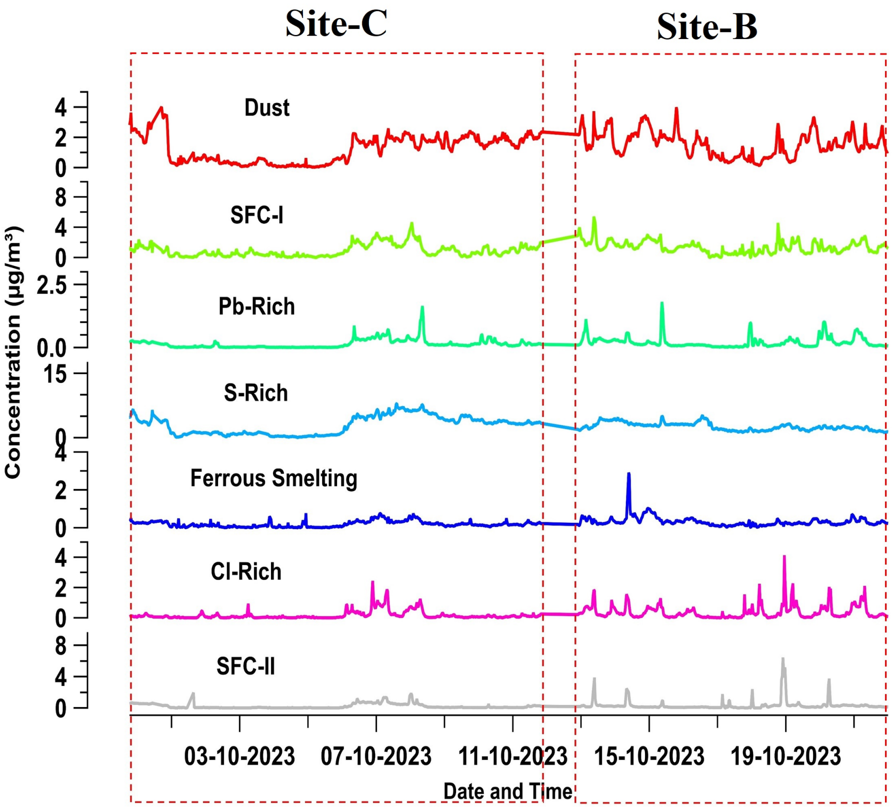


Figure A4. Time series plots for the different elemental factors over **Site-B** and **Site-C** during the measurement period.

site, whereas **Site-B** experiences nearby biomass burning or garbage burning which likely releases Cl and Cu into the atmosphere (Gani et al., 2019; Shukla et al., 2021).

Appendix B: Low-cost sensors

In this section, we discuss in brief about the working principle of the gas, PM and VOC sensors installed in the LCAQ sensor unit as well as the Micro-Aethalometer.

B1 Electrochemical sensors for measurement of Pollutant Gases

The sensors used for the measurement of gaseous pollutants such as carbon monoxide (CO), nitrogen dioxide (NO_2), sulphur dioxide (SO_2), and ozone (O_3) in this experiment are electrochemical sensors of Alphasense make with model number B4. Electrochemical sensors are advanced instruments designed for detecting specific gaseous pollutants such as carbon monoxide (CO), nitrogen dioxide (NO_2), sulphur dioxide (SO_2), and ozone (O_3). These sensors operate based on the principle of electrochemical reactions occurring at the electrodes within the sensor. Each sensor consists of a working electrode, an auxiliary electrode, a reference electrode, and an electrolyte. When the target gas diffuses through a selective, permeable membrane, it reacts electrochemically at the working electrode, generating an electrical current. The magnitude of this current is directly proportional to the concentration of the target gas.

To ensure stability and accuracy, the reference electrode maintains a constant potential, while the auxiliary electrode compensates for potential cross-sensitivities from interfering gases. The sensors are highly sensitive, capable of detecting gases at very low concentrations (parts per billion), and are designed to be compact, energy-efficient, and easily integrable into portable or stationary air quality monitoring systems. They are widely used in applications ranging from industrial emission monitoring to personal exposure assessment and urban air quality studies.

B2 Low-cost sensors for $PM_{2.5}$ and VOC measurement

For measurement of $PM_{2.5}$ and VOC, a sensor model of make Sensirion with model number SEN54 is used. This sensor module is a highly integrated device designed for measuring $PM_{2.5}$, VOC, and environmental parameters like humidity and temperature. For PM measurement, the SEN54 uses a laser-based light scattering technique. Inside the sensor, a laser beam illuminates airborne particles as they pass through a detection chamber. A photodetector captures the scattered light, and the intensity and angle of the scattering are analyzed to determine the size and concentration of particles in the air. This method provides high accuracy and resolution for detecting fine particles such as $PM_{2.5}$.

For VOC measurement, the sensor employs a metal-oxide semiconductor (MOX) gas sensor. The surface of the MOX sensor interacts with VOC molecules, altering the conductivity of the material in a way proportional to the VOC concentration. The SEN54 also includes advanced signal processing algorithms to translate raw data into reliable and calibrated outputs for VOC levels. Additionally, the sensor integrates temperature and humidity sensing elements to provide environmental context for $PM_{2.5}$ and VOC data, improving the accuracy of measurements under varying conditions.

The SEN54 is compact, robust, and energy-efficient and is ideal for indoor and outdoor air quality monitoring applications, providing comprehensive insights into particulate and gas-phase pollution in real time.

https://doi.org/10.5194/egusphere-2025-5677
Preprint. Discussion started: 9 December 2025

B3 Micro Aethalometer

Apart from the gas, PM and VOC sensors installed in the LCAQ sensor unit, the data captured using a Micro-Aethalometer is also used for predicting the different sources. The Micro-Aethalometer is a compact, portable device designed for real-time monitoring of black carbon (BC) concentrations in the air. It is widely used in environmental and occupational health studies to assess air quality and exposure to fine PM, particularly those originating from combustion sources like vehicle emissions and biomass burning. The working principle of the Micro-Aethalometer can be categorized under four sections as:

- **Air Sampling:** Ambient air is drawn through the device by an internal pump, and PM is collected on a quartz filter tape.
- **Optical Analysis:** A light-emitting diode (LED) generates a steady beam of near-infrared light that passes through the filter. Black carbon, being highly absorptive in this wavelength, reduces the intensity of the transmitted light.
- **Signal Processing:** The reduction in light intensity (attenuation) is measured by a photodetector. The attenuation signal is proportional to the BC concentration on the filter, allowing for continuous monitoring.
- **Data Output:** The signal is processed to compute the equivalent black carbon concentration in micrograms per cubic meter (μg/m^3). The data is logged internally and can be transmitted to external systems via USB or wireless communication.

For our experiments, we have used Micro-Aethalometer manufactured by AethLabs with model number AE-51. The AE-51 operates based on the principle of optical absorption. It uses a small pump to draw an air sample through a quartz filter tape, where airborne PM is deposited. A light beam, typically in the near-infrared (880 nm), is directed through the filter, and the device measures the attenuation of light caused by the black carbon particles. The attenuation is then converted into a BC concentration using calibration factors derived from laboratory and field studies.

The device is equipped with a sensitive photodetector and an internal pump to ensure a consistent sample flow rate, usually around 50-150 ml/min. It also features real-time data logging with high temporal resolution, which makes it suitable for both stationary and mobile measurements. Data can be exported for further analysis, enabling detailed spatial and temporal assessments of BC levels.

The Micro-Aethalometer AE-51 is lightweight and battery-operated, making it ideal for personal exposure studies, indoor air quality monitoring, and on-road measurements of vehicular emissions. Its compact design ensures portability while maintaining sensitivity and accuracy for black carbon detection in diverse environments.

Appendix C: Additional Results

In this appendix we tabulate the results obtained using multi output regression models with some of the most popular individual mapping functions apart from **Linear Regression (LR)**, which has already been presented in Section 4. The following regression models have been used as individual mapping functions:

- **K-Nearest Neighbours Regressor (K-NN):** The hyper-parameters used in this case include $K = 5$ as the number of nearest neighbours considered. Uniform weights are used for each of the K nearest neighbour and the distance parameter used is Minkowski distance with power parameter $p = 2$.
- **Gradient Boosting Regressor (GB):** The principal hyper-parameters used in this case include choosing a learning rate of 0.1, with the number of estimators equal to 100. Squared error as the loss function, while Friedman-mean square error was used as the function to measure the quality of a split.
- **Ridge Regressor (RR):** The hyper-parameter α responsible for controlling the regularization strength is set equal to unity.
- **Random Forest Regressor (RF):** In this case, the number of estimators or the number of trees in the forest is set equal to 100 and the function used to measure the quality of a split is set as the squared error loss function.
- **Support Vector Regressor (SV):** In this case an rbf kernel is chosen for implementing the Support Vector regressor. The smoothing parameter *gamma* for the rbf kernel is set to be adaptively chosen based on the nature of predictor matrix. It is equal to the reciprocal of the product of the dimension of the predictor and the variance of the predictor matrix.

The hyper-parameters mentioned above have been selected on a trial and error basis. It has been observed that the results are not sensitive to change in the hyper-parameters over a large region in the hyper-parameter space. Outside this region, the performance of the algorithm deteriorates further. The organic source apportionment results followed by their elemental counterpart are discussed next for **Site-B** and **Site-C**.

Table C1. Component wise MAE and SAE computed for BBOA-1, BBOA-2, LVOOA, SVOOA and HOA according to Eqn (5) for deployment at **Site-B** for different regression models discussed above.

			BBOA-1	BBOA-2	LVOOA	SVOOA	HOA
Site-B (Organic SA)	**K-NN**	**MAE (in %)**	2.84	3.38	9.85	4.91	3.34
		SAE (in %)	2.51	3.13	9.73	5.40	3.79
	GB	**MAE (in %)**	3.06	3.76	11.05	4.92	3.92
		SAE (in %)	2.78	3.58	11.34	5.55	4.96
	RR	**MAE (in %)**	2.85	2.79	10.47	5.02	4.23
		SAE (in %)	2.50	2.19	7.62	4.73	3.67
	RF	**MAE (in %)**	2.46	3.94	10.84	4.84	3.35
		SAE (in %)	2.41	2.94	9.72	5.01	4.05
	SV	**MAE (in %)**	3.57	4.68	14.51	7.15	5.81
		SAE (in %)	1.74	3.60	11.83	4.93	4.41
	LR	**MAE (in %)**	2.89	2.87	10.78	5.20	4.39
		SAE (in %)	2.63	2.25	7.96	4.93	3.83
	Mean of Actual Fractions (in %)		9.91	17.09	48.61	12.68	11.72

Table C2. Efficacy of the proposed method in predicting the relative abundance of the different components in the correct order for organic source apportionment at **Site-B** for different regression models discussed above.

		Mean Normalized Discounted Cumulative Gain (NDCG) in predicting the top K contributing sources				**Mean SROCC**
		K=1	**K=2**	**K=3**	**For all sources**	
Site-B (Organic SA)	**K-NN**	0.91	0.94	0.95	0.98	0.81
	GB	0.90	0.93	0.94	0.97	0.76
	RR	0.88	0.92	0.95	0.97	0.78
	RF	0.87	0.91	0.94	0.96	0.79
	SV	0.88	0.92	0.93	0.96	0.57
	LR	0.88	0.92	0.95	0.97	0.77

https://doi.org/10.5194/egusphere-2025-5677
Preprint. Discussion started: 9 December 2025

Table C3. Component wise MAE and SAE computed for BBOA-1, BBOA-2, LVOOA, SVOOA and HOA according to Eqn (5) for deployment at **Site-C** for different regression models discussed above.

			BBOA-1	**BBOA-2**	**LVOOA**	**SVOOA**	**HOA**
Site-C (Organic SA)	**K-NN**	**MAE (in %)**	3.59	3.57	9.71	4.57	3.97
		SAE (in %)	2.75	2.83	8.37	4.28	3.56
	GB	**MAE (in %)**	3.55	3.09	7.25	3.87	4.21
		SAE (in %)	2.91	2.16	5.47	3.23	3.37
	RR	**MAE (in %)**	3.34	3.79	7.17	3.99	2.50
		SAE (in %)	2.44	2.61	5.70	3.05	2.00
	RF	**MAE (in %)**	3.74	3.43	7.49	3.34	4.99
		SAE (in %)	2.94	2.28	5.98	3.34	3.68
	SV	**MAE (in %)**	3.65	3.82	11.00	5.17	7.07
		SAE (in %)	3.22	2.44	7.28	2.85	2.68
	LR	**MAE (in %)**	3.44	3.96	7.41	4.03	2.72
		SAE (in %)	2.42	2.61	6.34	3.22	2.37
	Mean of Actual Fractions (in %)		8.43	13.22	58.30	8.63	11.42

Table C4. Efficacy of the proposed method in predicting the relative abundance of the different components in the correct order for organic source apportionment at **Site-C** for different regression models discussed above.

		Mean Normalized Discounted Cumulative Gain (NDCG) in predicting the top K contributing sources				**Mean SROCC**
		K=1	**K=2**	**K=3**	**For all sources**	
Site-C (Organic SA)	**K-NN**	1	0.99	0.97	0.99	0.78
	GB	1	0.99	0.97	0.99	0.77
	RR	1	0.99	0.98	0.99	0.81
	RF	1	0.99	0.97	0.99	0.78
	SV	1	0.95	0.97	0.99	0.67
	LR	1	0.99	0.98	0.99	0.8

Table C5. Component wise MAE and SAE computed for Dust, SFC-1, SFC-2, Pb-rich, S-rich, Ferrous Smelting and Cl-rich according to Eqn (5) for deployment at **Site-B** for different regression models discussed above.

			Dust	SFC-I	SFC-II	Pb-rich	S-rich	Ferrous Smelting	Cl-rich
Site-B (Elemental SA)	**K-NN**	**MAE (in %)**	5.36	7.34	1.29	2.34	8.10	1.19	1.76
		SAE (in %)	4.62	5.92	1.86	2.61	4.34	1.29	1.67
	GB	**MAE (in %)**	5.12	7.71	3.50	2.91	6.94	1.25	2.02
		SAE (in %)	4.14	6.36	4.19	3.58	4.26	1.43	1.96
	RR	**MAE (in %)**	4.86	6.89	2.38	2.04	8.65	1.65	2.60
		SAE (in %)	4.74	5.07	1.91	1.59	5.38	1.53	1.87
	RF	**MAE (in %)**	4.77	7.51	2.93	2.73	7.13	1.15	2.01
		SAE (in %)	3.27	6.74	2.76	2.78	3.82	1.52	2.35
	SV	**MAE (in %)**	7.69	8.80	5.04	6.45	7.90	3.59	5.14
		SAE (in %)	4.80	6.62	1.20	0.55	6.30	1.06	1.59
	LR	**MAE (in %)**	4.91	6.86	2.48	2.12	8.54	1.67	2.62
		SAE (in %)	4.78	5.04	2.00	1.65	5.70	1.53	1.91
	Mean of Actual Fractions (in %)		26.44	23.75	3.84	1.70	35.17	5.00	4.10

Table C6. Efficacy of the proposed method in predicting the relative abundance of the different components in the correct order for elemental source apportionment at **Site-B** for different regression models discussed above.

		Mean Normalized Discounted Cumulative Gain (NDCG) in predicting the top K contributing sources				**Mean SROCC**
		K=1	**K=2**	**K=3**	**For all sources**	
Site-B (Elemental SA)	**K-NN**	0.89	0.93	0.97	0.97	0.84
	GB	0.91	0.91	0.94	0.97	0.8
	RR	0.91	0.94	0.97	0.97	0.84
	RF	0.91	0.93	0.96	0.97	0.82
	SV	0.91	0.95	0.98	0.98	0.88
	LR	0.91	0.94	0.97	0.97	0.84

https://doi.org/10.5194/egusphere-2025-5677
Preprint. Discussion started: 9 December 2025

Table C7. Component wise MAE and SAE computed for Dust, SFC-1, SFC-2, Pb-rich, S-rich, Ferrous Smelting and Cl-rich according to Eqn (5) for deployment at **Site-C** for different regression models discussed above.

			Dust	SFC-I	SFC-II	Pb-rich	S-rich	Ferrous Smelting	Cl-rich
Site-C (Elemental SA)	**K-NN**	**MAE (in %)**	3.84	6.14	0.61	0.78	8.84	0.68	0.65
		SAE (in %)	2.50	3.78	0.64	0.99	6.19	0.66	0.56
	GB	**MAE (in %)**	3.49	5.14	0.77	0.96	7.99	0.60	0.63
		SAE (in %)	2.46	3.60	0.90	1.05	5.55	0.67	0.59
	RR	**MAE (in %)**	3.53	5.05	0.77	0.96	7.93	0.57	0.62
		SAE (in %)	2.39	3.54	0.92	1.06	5.49	0.66	0.57
	RF	**MAE (in %)**	3.53	5.14	0.76	0.95	8.14	0.58	0.64
		SAE (in %)	2.44	3.71	0.84	1.03	5.61	0.64	0.60
	SV	**MAE (in %)**	3.55	5.08	0.74	0.96	7.93	0.55	0.61
		SAE (in %)	2.41	3.51	0.90	1.05	5.55	0.66	0.58
	LR	**MAE (in %)**	4.48	6.67	1.25	1.01	6.49	1.25	2.41
		SAE (in %)	3.42	4.98	0.69	1.01	5.19	0.81	1.90
	Mean of Actual Fractions (in %)		26.38	13.70	1.54	2.10	52.3	2.69	1.30

Table C8. Efficacy of the proposed method in predicting the relative abundance of the different components in the correct order for elemental source apportionment at **Site-C** for different regression models discussed above.

		Mean Normalized Discounted Cumulative Gain (NDCG) in predicting the top K contributing sources				**Mean SROCC**
		K=1	**K=2**	**K=3**	**For all sources**	
Site-C (Elemental SA)	**K-NN**	1	0.99	0.99	0.99	0.93
	GB	1	0.99	0.99	0.99	0.90
	RR	1	0.99	0.99	0.99	0.90
	RF	1	0.99	0.99	0.99	0.91
	SV	1	0.99	0.99	0.99	0.90
	LR	1	0.99	0.99	0.99	0.87

Financial support. We thank the Bloomberg Philanthropies, the Open Philanthropy, and the Clean Air Fund for their financial support.

Code and data availability. Sample data and code used for the experiments described in this manuscript will be made available online upon acceptance of the manuscript for publication.

Author contributions. The study was conceived and planned by SNT who also contributed to the final article. SC carried out the implementation of the machine learning algorithm and analysis and prepared the first draft. PKS, NTR, AT and PK also contributed to the design and implementation of the machine learning algorithm. DS, AL and AK contributed to data collection in the field and also did the analysis for source apportionment using data from RGI. All the authors contributed to the final shaping of the article.

Competing interests. The authors declare that they have no competing interests to disclose.

References

Ajnoti, N., Gehlot, H., and Tripathi, S. N.: Hybrid instrument network optimization for air quality monitoring, Atmospheric Measurement Techniques, 17, 1651–1664, 2024.

Alas, H. D. C., Müller, T., Weinhold, K., Pfeifer, S., Glojek, K., Gregorič, A., Močnik, G., Drinovec, L., Costabile, F., Ristorini, M., et al.: Performance of microAethalometers: real-world field intercomparisons from multiple mobile measurement campaigns in different atmospheric environments, Aerosol and Air Quality Research, 20, 2640–2653, 2020.

Awad, M., Khanna, R., Awad, M., and Khanna, R.: Support vector regression, Efficient learning machines: Theories, concepts, and applications for engineers and system designers, pp. 67–80, 2015.

Barbiere, M., Lagler, F., Borowiak, A., et al.: Evaluation of the Inter-Laboratory Comparison exercise for SO2, CO, O3, NO and NO2 (13-16 May 2019, Ispra), 2019.

Bhandari, S., Gani, S., Patel, K., Wang, D. S., Soni, P., Arub, Z., Habib, G., Apte, J. S., and Hildebrandt Ruiz, L.: Sources and atmospheric dynamics of organic aerosol in New Delhi, India: insights from receptor modeling, Atmospheric Chemistry and Physics, 20, 735–752, 2020.

Bhowmik, H. S., Shukla, A., Lalchandani, V., Dave, J., Rastogi, N., Kumar, M., Singh, V., and Tripathi, S. N.: Inter-comparison of online and offline methods for measuring ambient heavy and trace elements and water-soluble inorganic ions (NO 3-, SO 4 2-, NH 4+, and Cl-) in PM 2.5 over a heavily polluted megacity, Delhi, Atmospheric Measurement Techniques, 15, 2667–2684, 2022.

Borchani, H., Varando, G., Bielza, C., and Larranaga, P.: A survey on multi-output regression, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 5, 216–233, 2015.

Bousiotis, D., Singh, A., Haugen, M., Beddows, D., Diez, S., Murphy, K. L., Edwards, P. M., Boies, A., Harrison, R. M., and Pope, F. D.: Assessing the sources of particles at an urban background site using both regulatory instruments and low-cost sensors–a comparative study, Atmospheric Measurement Techniques, 14, 4139–4155, 2021.

Bousiotis, D., Beddows, D., Singh, A., Haugen, M., Diez, S., Edwards, P. M., Boies, A., Harrison, R. M., and Pope, F. D.: A study on the performance of low-cost sensors for source apportionment at an urban background site, Atmospheric Measurement Techniques, 15, 4047–4061, 2022.

Bousiotis, D., Allison, G., Beddows, D. C., Harrison, R. M., and Pope, F. D.: Towards comprehensive air quality management using low-cost sensors for pollution source apportionment, NPJ Climate and Atmospheric Science, 6, 122, 2023.

Bousiotis, D., Shaqiri, L. A., Sanghera, D. S., Tinker, D., and Pope, F. D.: Low-Cost Source Apportionment (LoCoSA) of air pollution-literature review of the state of the art, Science of The Total Environment, 998, 180 257, 2025.

Canonaco, F., Crippa, M., Slowik, J. G., Baltensperger, U., and Prévôt, A. S.: SoFi, an IGOR-based interface for the efficient use of the generalized multilinear engine (ME-2) for the source apportionment: ME-2 application to aerosol mass spectrometer data, Atmospheric Measurement Techniques, 6, 3649–3661, 2013.

Choomanee, P., Bualert, S., Thongyen, T., Rungratanaubon, T., Rattanapotanan, T., and Szymanski, W. W.: Beyond common urban air quality assessment: Relationship between PM2. 5 and black carbon during haze and non-haze periods in Bangkok, Atmospheric Pollution Research, 15, 101 992, 2024.

Coelho, S., Ferreira, J., Rodrigues, V., and Lopes, M.: Source apportionment of air pollution in European urban areas: Lessons from the ClairCity project, Journal of Environmental Management, 320, 115 899, 2022.

Cohen, A. J., Brauer, M., Burnett, R., Anderson, H. R., Frostad, J., Estep, K., Balakrishnan, K., Brunekreef, B., Dandona, L., Dandona, R., et al.: Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015, The lancet, 389, 1907–1918, 2017.

Cohen, D. D., Crawford, J., Stelcer, E., and Bac, V. T.: Characterisation and source apportionment of fine particulate sources at Hanoi from 2001 to 2008, Atmospheric Environment, 44, 320–328, 2010.

Crippa, M., Canonaco, F., Lanz, V., Äijälä, M., Allan, J., Carbone, S., Capes, G., Ceburnis, D., Dall'Osto, M., Day, D., et al.: Organic aerosol components derived from 25 AMS data sets across Europe using a consistent ME-2 based source apportionment approach, Atmospheric chemistry and physics, 14, 6159–6176, 2014.

DeCarlo, P. F., Kimmel, J. R., Trimborn, A., Northway, M. J., Jayne, J. T., Aiken, A. C., Gonin, M., Fuhrer, K., Horvath, T., Docherty, K. S., et al.: Field-deployable, high-resolution, time-of-flight aerosol mass spectrometer, Analytical chemistry, 78, 8281–8289, 2006.

Dimitriou, K., Stavroulas, I., Grivas, G., Chatzidiakos, C., Kosmopoulos, G., Kazantzidis, A., Kourtidis, K., Karagioras, A., Hatzianastassiou, N., Pandis, S. N., et al.: Intra-and inter-city variability of PM2. 5 concentrations in Greece as determined with a low-cost sensor network, Atmospheric environment, 301, 119 713, 2023.

Feenstra, B., Papapostolou, V., Hasheminassab, S., Zhang, H., Der Boghossian, B., Cocker, D., and Polidori, A.: Performance evaluation of twelve low-cost PM2. 5 sensors at an ambient air monitoring site, Atmospheric Environment, 216, 116 946, 2019.

Friedman, J. H.: Greedy function approximation: a gradient boosting machine, Annals of statistics, pp. 1189–1232, 2001.

Gani, S., Bhandari, S., Seraj, S., Wang, D. S., Patel, K., Soni, P., Arub, Z., Habib, G., Hildebrandt Ruiz, L., and Apte, J. S.: Submicron aerosol composition in the world's most polluted megacity: the Delhi Aerosol Supersite study, Atmospheric Chemistry and Physics, 19, 6843–6859, 2019.

Gong, W., Zhang, T., Zhu, Z., Ma, Y., Ma, X., and Wang, W.: Characteristics of PM1. 0, PM2. 5, and PM10, and their relation to black carbon in Wuhan, Central China, Atmosphere, 6, 1377–1387, 2015.

Hagan, D. H., Gani, S., Bhandari, S., Patel, K., Habib, G., Apte, J. S., Hildebrandt Ruiz, L., and Kroll, J. H.: Inferring aerosol sources from low-cost air quality sensor measurements: a case study in Delhi, India, Environmental Science & Technology Letters, 6, 467–472, 2019.

Hopke, P. K.: Case studies of source apportionment from North America, 2016.

Huang, C.-H. et al.: Field comparison of real-time PM2. 5 readings from a beta gauge monitor and a light scattering method, Aerosol and Air Quality Research, 7, 239–250, 2007.

Jain, S., Sharma, S., Mandal, T., and Saxena, M.: Source apportionment of PM10 in Delhi, India using PCA/APCS, UNMIX and PMF, Particuology, 37, 107–118, 2018.

Jaiprakash, Singhai, A., Habib, G., Raman, R. S., and Gupta, T.: Chemical characterization of PM 1.0 aerosol in Delhi and source apportionment using positive matrix factorization, Environmental Science and Pollution Research, 24, 445–462, 2017.

Järvelin, K. and Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques, ACM Transactions on Information Systems (TOIS), 20, 422–446, 2002.

Karmakar, S. P., Das, A. B., Gurung, C., and Ghosh, C.: Effects of ozone on plant health and environment: A mini review, Res. Jr. Agril. Sci, 13, 612–619, 2022.

Kulkarni, P., Puttaswamy, N., Prabhu, V., Agrawal, P., Upadhya, A. R., Rao, S., Sutaria, R., Mor, S., Dey, S., Khaiwal, R., et al.: Inter-versus Intracity Variations in the Performance and Calibration of Low-Cost PM2. 5 Sensors: A Multicity Assessment in India, ACS Earth and Space Chemistry, 6, 3007–3016, 2022.

Kumar, B., Chakraborty, A., Tripathi, S., and Bhattu, D.: Highly time resolved chemical characterization of submicron organic aerosols at a polluted urban location, Environmental Science: Processes & Impacts, 18, 1285–1296, 2016.

Kumar, V., Sahu, M., Biswas, P., et al.: Source Apportionment of Particulate Matter by Application of Machine Learning Clustering Algorithms, Aerosol and Air Quality Research, 22, 210 240, 2022.

Kumar, V., Malyan, V., Sahu, M., and Biswal, B.: Aerosol sources characterization and apportionment from low-cost particle sensors in an urban environment, Atmospheric Environment: X, 22, 100 271, 2024a.

Kumar, V., Sahu, M., Biswal, B., Prakash, J., Choudhary, S., Raliya, R., Chadha, T. S., Fang, J., and Biswas, P.: Real-Time Source Apportionment of Particulate Matter from Low-Cost Particle Sensors Using Machine Learning, Aerosol science and engineering, pp. 1–11, 2024b.

Lakra, A., Shukla, A. K., Bhowmik, H. S., Yadav, A. K., Jain, V., Murari, V., Gaddamidi, S., Lalchandani, V., and Tripathi, S. N.: Comparative analysis of winter composite-PM2. 5 in Central Indo Gangetic Plain cities: Combined organic and inorganic source apportionment and characterization, with a focus on the photochemical age effect on secondary organic aerosol formation, Atmospheric Environment, 338, 120 827, 2024.

Lalchandani, V., Kumar, V., Tobler, A., Thamban, N. M., Mishra, S., Slowik, J. G., Bhattu, D., Rai, P., Satish, R., Ganguly, D., et al.: Real-time characterization and source apportionment of fine particulate matter in the Delhi megacity area during late winter, Science of the total environment, 770, 145 324, 2021.

Li, L., Xie, S., Zeng, L., Wu, R., and Li, J.: Characteristics of volatile organic compounds and their role in ground-level ozone formation in the Beijing-Tianjin-Hebei region, China, Atmospheric Environment, 113, 247–254, 2015.

Li, Q., Cheng, H., Zhou, T., Lin, C., and Guo, S.: The estimated atmospheric lead emissions in China, 1990–2009, Atmospheric Environment, 60, 1–8, 2012.

Liang, Y., Wang, X., Dong, Z., Wang, X., Wang, S., Si, S., Wang, J., Liu, H.-Y., Zhang, Q., and Wang, Q.: Understanding the origins of urban particulate matter pollution based on high-density vehicle-based sensor monitoring and big data analysis, Urban Climate, 59, 102 241, 2025.

Liu, J. W. et al.: Real-time systems, Pearson Education India, 2006.

Liu, X., Jayaratne, R., Thai, P., Kuhn, T., Zing, I., Christensen, B., Lamont, R., Dunbabin, M., Zhu, S., Gao, J., et al.: Low-cost sensors as an alternative for long-term air quality monitoring, Environmental research, 185, 109 438, 2020.

Ltd., A.: SO2-B4 Sulfur Dioxide Sensor: Technical Datasheet, Alphasense Ltd., Great Notley, Essex, UK, https://ametekcdn.azureedge.net/mediafiles/project/oneweb/oneweb/alphasense/products/datasheets/alphasense_so2-b4_datasheet_en_3.pdf, minimum detection limit: 5 ppb (noise ±2 standard deviations). Accessed: 4 November 2025., 2023.

Manisalidis, I., Stavropoulou, E., Stavropoulos, A., and Bezirtzoglou, E.: Environmental and health impacts of air pollution: a review, Frontiers in public health, 8, 14, 2020.

McDonald, G. C.: Ridge regression, Wiley Interdisciplinary Reviews: Computational Statistics, 1, 93–100, 2009.

Mehta, P.: Science behind acid rain: analysis of its impacts and advantages on life and heritage structures, south Asian journal of tourism and heritage, 3, 123–132, 2010.

Mills, S. A., Bousiotis, D., Maya-Manzano, J. M., Tummon, F., MacKenzie, A. R., and Pope, F. D.: Constructing a pollen proxy from low-cost Optical Particle Counter (OPC) data processed with Neural Networks and Random Forests, Science of the Total Environment, 871, 161 969, 2023.

Montgomery, D. C., Peck, E. A., and Vining, G. G.: Introduction to linear regression analysis, John Wiley & Sons, 2021.

Munsif, R., Zubair, M., Aziz, A., and Zafar, M. N.: Industrial air emission pollution: potential sources and sustainable mitigation, in: Environmental Emissions, IntechOpen, 2021.

Myers, J. L., Well, A. D., and Lorch Jr, R. F.: Research design and statistical analysis, Routledge, 2013.

Ng, N., Canagaratna, M., Zhang, Q., Jimenez, J., Tian, J., Ulbrich, I., Kroll, J., Docherty, K., Chhabra, P., Bahreini, R., et al.: Organic aerosol components observed in Northern Hemispheric datasets from Aerosol Mass Spectrometry, Atmospheric Chemistry and Physics, 10, 4625–4641, 2010.

Nidheesh, P. and Kumar, M. S.: An overview of environmental sustainability in cement and steel production, Journal of cleaner production, 231, 856–871, 2019.

Nuvolone, D., Petri, D., and Voller, F.: The effects of ozone on human health, Environmental Science and Pollution Research, 25, 8074–8088, 2018.

Oil and Gas Online: B-Series Platform Overview, https://www.oilandgasonline.com/doc/b-iseries-platform-b-0001.

Owoade, O. K., Abiodun, P. O., Omokungbe, O. R., Fawole, O. G., Olise, F. S., Popoola, O. O., Jones, R. L., and Hopke, P. K.: Spatial-temporal Variation and Local Source Identification of Air Pollutants in a Semi-urban Settlement in Nigeria Using Low-cost Sensors, Aerosol and Air Quality Research, 21, 200 598, https://doi.org/10.4209/aaqr.200598, 2021.

Paatero, P. and Tapper, U.: Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values, Environmetrics, 5, 111–126, 1994.

Pant, P., Shukla, A., Kohl, S. D., Chow, J. C., Watson, J. G., and Harrison, R. M.: Characterization of ambient PM2. 5 at a pollution hotspot in New Delhi, India and inference of sources, Atmospheric environment, 109, 178–189, 2015.

Pope, F. D., Gatari, M., Ng'ang'a, D., Poynter, A., and Blake, R.: Airborne particulate matter monitoring in Kenya using calibrated low-cost sensors, Atmospheric Chemistry and Physics, 18, 15 403–15 418, 2018.

Rai, P., Furger, M., El Haddad, I., Kumar, V., Wang, L., Singh, A., Dixit, K., Bhattu, D., Petit, J.-E., Ganguly, D., et al.: Real-time measurement and source apportionment of elements in Delhi's atmosphere, Science of the Total Environment, 742, 140 332, 2020a.

Rai, P., Slowik, J. G., Furger, M., El Haddad, I., Visser, S., Tong, Y., Singh, A., Wehrle, G., Kumar, V., Tobler, A. K., et al.: Highly time-resolved measurements of element concentrations in PM 10 and PM 2.5: comparison of Delhi, Beijing, London, and Krakow, Atmospheric Chemistry and Physics Discussions, 2020, 1–18, 2020b.

Reff, A., Eberly, S. I., and Bhave, P. V.: Receptor modeling of ambient particulate matter data using positive matrix factorization: review of existing methods, Journal of the Air & Waste Management Association, 57, 146–154, 2007.

Respirer Living Sciences: Respirer Website, https://www.respirer.in/.

Saarikoski, S., Hellén, H., Praplan, A. P., Schallhart, S., Clusius, P., Niemi, J. V., Kousa, A., Tykkä, T., Kouznetsov, R., Aurela, M., et al.: Characterization of volatile organic compounds and submicron organic aerosol in a traffic environment, Atmospheric Chemistry and Physics, 23, 2963–2982, 2023.

Sahu, R., Dixit, K. K., Mishra, S., Kumar, P., Shukla, A. K., Sutaria, R., Tiwari, S., and Tripathi, S. N.: Validation of low-cost sensors in measuring real-time PM10 concentrations at two sites in Delhi national capital region, Sensors, 20, 1347, 2020.

Sahu, R., Nagal, A., Dixit, K. K., Unnibhavi, H., Mantravadi, S., Nair, S., Simmhan, Y., Mishra, B., Zele, R., Sutaria, R., et al.: Robust statistical calibration and characterization of portable low-cost air quality monitoring sensors to quantify real-time O 3 and NO 2 concentrations in diverse environments, Atmospheric Measurement Techniques, 14, 37–52, 2021.

scikit-learn developers: scikit-learn: Preprocessing data, https://scikit-learn.org/stable/modules/preprocessing.html, accessed: 2025-09-29, 2025.

Seber, G. A. and Lee, A. J.: Linear regression analysis, John Wiley & Sons, 2012.

Segal, M. R.: Machine learning benchmarks and random forest regression, 2004.

Sen, P. C., Hajra, M., and Ghosh, M.: Supervised classification algorithms in machine learning: A survey and review, in: Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018, pp. 99–111, Springer, 2020.

Sharma, S. B., Jain, S., Khirwadkar, P., and Kulkarni, S.: The effects of air pollution on the environment and human health, Indian Journal of Research in Pharmacy and Biotechnology, 1, 391–396, 2013.

Sharma, S. K., Sharma, A., Saxena, M., Choudhary, N., Masiwal, R., Mandal, T. K., and Sharma, C.: Chemical characterization and source apportionment of aerosol at an urban area of Central Delhi, India, Atmospheric Pollution Research, 7, 110–121, 2016.

Shen, H., Hou, W., Zhu, Y., Zheng, S., Ainiwaer, S., Shen, G., Chen, Y., Cheng, H., Hu, J., Wan, Y., et al.: Temporal and spatial variation of PM2. 5 in indoor air monitored by low-cost sensors, Science of The Total Environment, 770, 145 304, 2021.

Shukla, A. K., Lalchandani, V., Bhattu, D., Dave, J. S., Rai, P., Thamban, N. M., Mishra, S., Gaddamidi, S., Tripathi, N., Vats, P., et al.: Real-time quantification and source apportionment of fine particulate matter including organics and elements in Delhi during summertime, Atmospheric Environment, 261, 118 598, 2021.

Singh, S., Kulshrestha, M. J., Rani, N., Kumar, K., Sharma, C., and Aswal, D.: An overview of vehicular emission standards, Mapan, 38, 241–263, 2023.

Smith, J. S., Laskin, A., and Laskin, J.: Molecular characterization of biomass burning aerosols using high-resolution mass spectrometry, Analytical chemistry, 81, 1512–1521, 2009.

Snyder, E. G., Watkins, T. H., Solomon, P. A., Thoma, E. D., Williams, R. W., Hagler, G. S., Shelow, D., Hindin, D. A., Kilaru, V. J., and Preuss, P. W.: The changing paradigm of air pollution monitoring, Environmental science & technology, 47, 11 369–11 377, 2013.

Song, Y., Liang, J., Lu, J., and Zhao, X.: An efficient instance selection algorithm for k nearest neighbor regression, Neurocomputing, 251, 26–34, 2017.

South Coast Air Quality Management District (SCAQMD): AethLabs microAeth Model AE51, https://www.aqmd.gov/aq-spec/product/aethlabs-microaeth-model-ae51.

Steinle, S., Reis, S., Sabel, C. E., Semple, S., Twigg, M. M., Braban, C. F., Leeson, S. R., Heal, M. R., Harrison, D., Lin, C., et al.: Personal exposure monitoring of PM2. 5 in indoor and outdoor microenvironments, Science of the Total Environment, 508, 383–394, 2015.

Sun, X., Wang, H., Guo, Z., Lu, P., Song, F., Liu, L., Liu, J., Rose, N. L., and Wang, F.: Positive matrix factorization on source apportionment for typical pollutants in different environmental media: a review, Environmental Science: Processes & Impacts, 22, 239–255, 2020.

Taheri, A., Aliasghari, P., and Hosseini, V.: Black carbon and PM2. 5 monitoring campaign on the roadside and residential urban background sites in the city of Tehran, Atmospheric environment, 218, 116 928, 2019.

Tiwari, S., Srivastava, A. K., Bisht, D. S., Parmita, P., Srivastava, M. K., and Attri, S.: Diurnal and seasonal variations of black carbon and PM2. 5 over New Delhi, India: Influence of meteorology, Atmospheric Research, 125, 50–62, 2013.

Tobler, A., Bhattu, D., Canonaco, F., Lalchandani, V., Shukla, A., Thamban, N. M., Mishra, S., Srivastava, A. K., Bisht, D. S., Tiwari, S., et al.: Chemical characterization of PM2. 5 and source apportionment of organic aerosol in New Delhi, India, Science of The Total Environment, 745, 140 924, 2020.

Ulbrich, I., Canagaratna, M., Zhang, Q., Worsnop, D., and Jimenez, J.: Interpretation of organic components from Positive Matrix Factorization of aerosol mass spectrometric data, Atmospheric Chemistry and Physics, 9, 2891–2918, 2009.

U.S. Environmental Protection Agency: Environmental Technology Verification Report, https://archive.epa.gov/nrmrl/archive-etv/web/pdf/p100fk6b.pdf.

Watson, J. G., Chow, J. C., and Pace, T. G.: Chemical mass balance, in: Data Handling in Science and Technology, vol. 7, pp. 83–116, Elsevier, 1991.

Yang, L. H., Hagan, D. H., Rivera-Rios, J. C., Kelp, M. M., Cross, E. S., Peng, Y., Kaiser, J., Williams, L. R., Croteau, P. L., Jayne, J. T., et al.: Investigating the sources of urban air pollution using low-cost air quality sensors at an urban Atlanta site, Environmental Science & Technology, 56, 7063–7073, 2022.

Ying, Q. and Krishnan, A.: Source contributions of volatile organic compounds to ozone formation in southeast Texas, Journal of Geophysical Research: Atmospheres, 115, 2010.

Yoo, G.: Real-time information on air pollution and avoidance behavior: evidence from South Korea, Population and Environment, 42, 406–424, 2021.

Zhang, K., Li, L., Huang, L., Wang, Y., Huo, J., Duan, Y., Wang, Y., and Fu, Q.: The impact of volatile organic compounds on ozone formation in the suburban area of Shanghai, Atmospheric environment, 232, 117 511, 2020.

Zhang, Q., Alfarra, M. R., Worsnop, D. R., Allan, J. D., Coe, H., Canagaratna, M. R., and Jimenez, J. L.: Deconvolution and quantification of hydrocarbon-like and oxygenated organic aerosols based on aerosol mass spectrometry, Environmental science & technology, 39, 4938–4952, 2005.

Zhou, X., Zhou, X., Wang, C., and Zhou, H.: Environmental and human health impacts of volatile organic compounds: A perspective review, Chemosphere, 313, 137 489, 2023.

Zhu, Q., Huang, X.-F., Cao, L.-M., Wei, L.-T., Zhang, B., He, L.-Y., Elser, M., Canonaco, F., Slowik, J. G., Bozzetti, C., et al.: Improved source apportionment of organic aerosols in complex urban air pollution using the multilinear engine (ME-2), Atmospheric Measurement Techniques, 11, 1049–1060, 2018.