Contents lists available at ScienceDirect

Environmental Research

journal homepage: www.elsevier.com/locate/envres

Source apportionment of PM_{10} particles in the urban atmosphere using PMF and LPO-XGBoost

Ying Liu [a,b], Bowen Jin [a,b], Xun Zhang [a,b,c], Xiansheng Liu [d,e,f,*], Tao Wang [g], Vy Ngoc Thuy Dinh [h], Jean-Luc Jaffrezo [h], Gaëlle Uzu [h], Pamela Dominutti [h], Sophie Darfeuil [h], Olivier Favez [i,j], Sébastien Conil [k], Nicolas Marchand [l], Sonia Castillo [m,n], Jesús D. de la Rosa [o], Stuart Grange [p], Christoph Hueglin [p], Konstantinos Eleftheriadis [q], Evangelia Diapouli [q], Manousos-Ioannis Manousakas [q], Maria Gini [q], Giulia Calzolai [r], Célia Alves [s], Marta Monge [f], Cristina Reche [f], Roy M. Harrison [t], Philip K. Hopke [u,v], Andrés Alastuey [f], Xavier Querol [f]

[a] *School of Computer and Artificial Intelligence, Beijing Technology and Business University, Beijing, 100048, China*
[b] *Beijing Laboratory for System Engineering of Carbon Neutrality, Beijing Municipal Education Commission, Beijing, 100048, China*
[c] *School of Computer Science and Artificial Intelligence, Xinjiang HeTian College, Hotan 848000, China*
[d] *Guangdong Key Laboratory of Environmental Catalysis and Health Risk Control, Guangdong-Hong Kong-Macao Joint Laboratory for Contaminants Exposure and Health, Institute of Environmental Health and Pollution Control, Guangdong University of Technology, Guangzhou, 510006, China*
[e] *Guangzhou Key Laboratory of Environmental Catalysis and Pollution Control, Guangdong Technology Research Center for Photocatalytic Technology Integration and Equipment Engineering, School of Environmental Science and Engineering, Guangdong University of Technology, Guangzhou, 510006, China*
[f] *Institute of Environmental Assessment and Water Research (IDAEA-CSIC), 08034, Barcelona, Spain*
[g] *Shanghai Key Laboratory of Atmospheric Particle Pollution and Prevention, Department of Environmental Science & Engineering, Fudan University, Shanghai, 200433, China*
[h] *Univ. Grenoble Alpes, IRD, CNRS, INRAE, Grenoble INP, IGE, UMR 5001, 38000, Grenoble, France*
[i] *INERIS, Parc Technologique Alata, BP 2, Verneuil-en-Halatte, 60550, France*
[j] *Laboratoire central de surveillance de la qualité de l'air (LCSQA), Verneuil-en-Halatte, 60550, France*
[k] *ANDRA DISTEC/EES Observatoire Pérenne de l'Environnement, F–55290, Bure, France*
[l] *Aix Marseille Univ, CNRS, LCE, Marseille, France*
[m] *Department of Applied Physics, University of Granada, 18011, Granada, Spain*
[n] *Andalusian Institute of Earth System Research, IISTA-CEAMA, University of Granada, 18006, Granada, Spain*
[o] *Associate Unit CSIC-UHU "Atmospheric Pollution", CIQSO, University of Huelva, 21071, Huelva, Spain*
[p] *Swiss Federal Laboratories for Materials Science and Technology (Empa), 8600, Dübendorf, Switzerland*
[q] *ENRACT Lab, National Centre for Scientific Research "Demokritos", Athens, 15341, Greece*
[r] *INFN Division of Florence and Department of Physics and Astronomy, University of Florence, via G.Sansone 1, 50019, Sesto Fiorentino, Italy*
[s] *Department of Environment and Planning, Centre for Environmental and Marine Studies (CESAM), University of Aveiro, 3810-193, Aveiro, Portugal*
[t] *School of Geography Earth and Environmental Sciences, University of Birmingham, B15 2TT, Birmingham, United Kingdom*
[u] *Departments of Public Health Sciences and Environmental Medicine, University of Rochester School of Medicine and Dentistry, Rochester, NY, 14642, USA*
[v] *Institute for a Sustainable Environment, Clarkson University, Potsdam, NY, 13699, USA*

ARTICLE INFO

Keywords:
Source apportionment
PMF
Machine learning
LPO-XGBoost
Air pollution

ABSTRACT

Atmospheric particulate matter (PM), as a leading part of air pollution, affects health in many ways. Thus, identifying and quantifying the contribution of atmospheric particulate matter sources of PM is vital for developing effective air quality management strategies. Positive Matrix Factorization (PMF) is one of the most common methods for source apportionment. However, PMF has some limitations, particularly its assumption that each source contributes linearly. In reality, some sources may exhibit nonlinear behaviors, which can compromise the accuracy of source apportionment. This study introduces a Lung Performance Optimization-based XGBoost (LPO-XGBoost) model, which leverages adaptive optimization principles inspired by lung function to enhance classic PM source apportionment. We demonstrate the potential for efficient, real-time

[*] Corresponding author. Guangdong Key Laboratory of Environmental Catalysis and Health Risk Control, Guangdong-Hong Kong-Macao Joint Laboratory for Contaminants Exposure and Health, Institute of Environmental Health and Pollution Control, Guangdong University of Technology, Guangzhou, 510006, China.
E-mail addresses: xiansheng.liu@idaea.csic.es, xianshengliu@gdut.edu.cn (X. Liu).

https://doi.org/10.1016/j.envres.2025.121659
Received 21 March 2025; Received in revised form 12 April 2025; Accepted 19 April 2025
Available online 22 April 2025

application of the LPO-XGBoost model across 21 monitoring sites in 6 European countries. Trained and validated on extensive environmental datasets, the model is capable of predicting major pollution sources, including road traffic, biomass burning, crustal, industrial, nitrate-rich particles, sulfate-rich particles, heavy fuel oil, and sea salt. It outperforms other machine learning models with an overall predictive coefficient of determination ($r^2 = 0.88$). Notably, the model performs exceptionally well in predicting sources such as sea salt ($r^2 = 0.97$) and biomass burning ($r^2 = 0.89$), but shows lower accuracy for the sulfate-rich particles source ($r^2 = 0.75$). Comparative analyses with models including Random Forest (RF), Support Vector Machine (SVM), and their LPO-enhanced variants confirm that LPO-XGBoost provides the most reliable performance in estimating pollution source contributions, offering scalability and robustness ideal for high-time-resolution observational data. This model has significant potential to support targeted air quality management strategies. Future research should focus on expanding key species measurements at monitoring sites, ensuring consistent temporal coverage, and optimizing the model for improved mixed-source predictions to strengthen its applicability in comprehensive urban air quality assessments.

1. Introduction

In recent years, air pollution has become a major environmental risk factor for human health and ranks as the fourth deadliest health risk worldwide, significantly contributing to the global burden of disease (Cohen et al., 2017; Wang et al., 2017, 2023c; Dominguez et al., 2024; Liu et al., 2025a,b). Among them, atmospheric particulate matter (PM), which is a major component of air pollution, makes a significant contribution through various pathways, via (a) combustion of fossil fuels for domestic heating, power generation, and transportation; (b) waste incineration in residential and municipal facilities; (c) industrial processes; and (d) natural sources, such as sea salt, volcanic eruptions, windblown dust, and pollen (Kleinman et al., 1976; Nagar et al., 2014; Nieder et al., 2018; Tong et al., 2021; Liu et al., 2024b). Moreover, it can also remain suspended for longer periods in the air, which enables it to get involved in secondary pollution and spreads to other areas through horizontal transport (Chen et al., 2018). At the same time, atmospheric PM from specific sources have a significant effect on human health, especially in causing respiratory diseases (such as asthma and chronic obstructive pulmonary disease) and heart diseases (such as high blood pressure and coronary artery disease) (Croft et al., 2019; Rich et al., 2019; Hopke et al., 2020a; Wang et al., 2023a,b). Therefore, identifying and quantifying pollution sources for PM are essential for developing effective air quality management strategies. Moreover, understanding the specific contributions of different sources not only facilitates the design of targeted interventions but also plays a crucial role in safeguarding the health of inhabited environments (Jafari et al., 2021).

Identifying multiple potential contributors to atmospheric particle pollution is a critical step toward achieving comprehensive source apportionment, with source apportionment methods playing an essential role in this process (Balachandran et al., 2013; Alias et al., 2020). Receptor models, which are well-established tools for source apportionment, are extensively employed to identify source categories and quantify contributions based on the chemical composition of pollutants collected at receptor sites (Hopke, 2016; Hopke et al., 2020a). These models rely exclusively on observational data for source apportionment, eliminating the need for emission inventories or complex transport models, thus making them highly valued in atmospheric research (Belis et al., 2013). Frequently applied receptor models include Positive Matrix Factorization (PMF) (Paatero et al., 1994), and Chemical Mass Balance (CMB) (Miller et al., 1972). Among these receptor models, PMF is one of the most widely adopted methods (Hopke, 2016; Wen et al., 2016; Dai et al., 2020; De Angelis et al., 2020; Gao et al., 2021), which has been successfully applied to many areas with different characteristics (Querol et al., 2001; Kim et al., 2003; Moon et al., 2008; Cohen et al., 2009; Amato et al., 2015; Liang et al., 2016; Manousakas et al., 2017; Borlaza et al., 2020; Mardoñez et al., 2023). However, the current PMF software (EPA PMF 5.0) has several limitations, such as the need for manual parameter settings. Users with different levels of experience may introduce subjectivity, leading to discrepancies in the results of PMF source apportionment. Therefore, after obtaining accurate pollution source contributions for a specific region using PMF, how to better apply these results for long-term pollution source prediction without manual intervention is an urgent issue that needs to be addressed (Zhang et al., 2019; Wang et al., 2022; Xu et al., 2023).

Recently, machine learning techniques have gained increasing interest in atmospheric science due to their strong performance in several key areas, including automatic parameter selection, model optimization, large-scale data handling, superior predictive accuracy, and robust generalization (Liang et al., 2020; Yang et al., 2020; Peng et al., 2024). For instance, the Extreme Gradient Boosting (XGBoost) model demonstrated high predictive accuracy in estimating $PM_{2.5}$ and PM_{10} concentrations, achieving r^2 values exceeding 0.9 (Pan, 2018; Zhong et al., 2022). This model outperformed other approaches, such as Random Forest (RF), Support Vector Machine (SVM), and Decision Tree Regression (DTR), and effectively mitigated overfitting. Additionally, XGBoost has demonstrated exceptional performance on large-scale population distribution mapping datasets, achieving the highest r^2 value ($r^2 = 0.8$) along with the lowest RMSE and MAE, showcasing its superior predictive capability. This dataset includes multiple variables such as geographic, geospatial big data, remote sensing data, and building data. Compared to other models, XGBoost outperforms in terms of accuracy and reliability (Zhao et al., 2021). Therefore, integrating machine learning methods (such as XGBoost), known for their robust performance in predicting large-scale datasets, with PMF may offer a potential solution to address PMF's limitations in processing large-scale atmospheric outdoor data.

This study intend to demonstrated the capability of coupling XGBoost and PMF in order to predicts large-scale outdoor pollution source apportionment through the following steps: (i) conducting detailed source apportionment using offline speciation data of PM_{10} from 21 monitoring stations across 6 European countries, applying the PMF method with repeated validations to ensure the reliability of the results; (ii) integrating XGBoost with the Lung Performance Optimization (LPO) algorithm (Ghasemi et al., 2024) to predict results from PMF source apportionment and using SHapley Additive exPlanations (SHAP) values to assess the contributions of the most relevant species for each pollution source; (iii) analyzing the performance of this model and comparing it with other models, such as SVM and RF, to evaluate its effectiveness. This analysis aims to identify the driving factors behind pollution sources, providing critical insights to develop more targeted and effective air quality management strategies.

2. Methodology

2.1. *Offline chemical speciation data in PM_{10}*

This study used a large dataset of concentration data from 21 air quality super monitoring sites. The data span from 2013 to 2021, with each site having at least one year of data to ensure the representativeness of the sites. The time resolution of PM sample collection is an average of one sample every three days. In detail, a total of 3112 daily samples are obtained from a distribution of seven countries: France (5 sites), Greece (1 site), Italy (2 sites), Portugal (3 sites), Spain (5 sites), and Switzerland (5 sites) (Fig. S1 & Table S1).

A total of thirty-two parameters were selected from the PM_{10} measurements at most sites, which including organic carbon (OC) and elemental carbon (EC), 21 metal elements (Al, Ca, Fe, Ti, K, Mg, As, Se, Cd, Mn, V, Ni, Cr, Pb, Ba, Cu, Sb, Sn, Zn, Rb, Na), 8 water-soluble ions (Ca^{2+}, K^{+}, Mg^{2+}, SO_4^{2-}, NO_3^{-}, NH_4^{+}, Na^{+}, Cl^{-}), and levoglucosan (LGA). When the monitoring metrics at the site simultaneously include Ca or Ca^{2+}, K or K^{+},and Mg or Mg^{2+}, only the elements Ca, K, and Mg were used as inputs to the source apportionment model. The OC and EC were measured by a thermal-optical method using the EUSAAR_2 protocol (Cavalli et al., 2010). Metal elements were analyzed via inductively coupled plasma-mass spectrometry (ICP-MS), Proton induced X-ray emission (PIXE), and X-ray fluorescence (XRF) techniques. Water-soluble ions were quantified via ion chromatography (IC), while levoglucosan was measured using gas chromatography-mass spectrometry (GC-MS). The levoglucosan data from France were obtained by High Performance Liquid Chromatography with Pulsed Amperometric Detection (HPLC-PAD) (Glojek et al., 2024; Liu et al., 2024c).

2.2. PMF

The PMF receptor model, used for source apportionment, is a factor analysis-based approach that decomposes the original data matrix X into two separate matrices: the source contribution matrix G and the source profile matrix F. This can be mathematically expressed as follows (Paatero et al., 1994; Paatero et al., 2014):

$$x_{ij} = \sum_{K=1}^{P} g_{ik} f_{kj} + e_{ij}$$

where x_i is the concentration of *jth* species measured in the *ith* sample (time), p stands for the number of factors, the factor profile f_{kj} is the concentration of *jth* species from the *kth* source, and the factor time series g_{ik} is the contribution of the *kth* source to the *ith* sample (time), while the residual matrix e_{ij} represents the error of *jth* species measured in *ith* sample (time).

The optimal solution for source apportionment was based on the following diagnostic parameters: S/N, Q_{robust}, Q_{true}/Q_{exp}, residual distribution at scaled residuals, G-space plots, interpretability of the factor profiles, and the seasonality of source contributions, and Pearson's correlation coefficients between input variables and reconstructed ones. For S/N, thresholds were categorized as “strong” ($S/N > 1$), “weak” ($0.5 \leq S/N \leq 1$), and “bad” ($S/N < 0.5$). The r^2 values were classified as “strong” ($r^2 > 0.6$) or “weak” ($0.2 \leq r^2 \leq 0.6$). When a species was critical for identifying a specific source, S/N thresholds were reevaluated (e.g., reclassifying “weak” or “bad” as “strong”). Conversely, species with significant relative contributions but lower importance to source identification were adjusted to “weak” or “bad.” The ratio of Q_{true}/Q_{robust} was analyzed for trends, with plateaus indicating the optimal range for the number of factors. Within this range, additional evaluation of source profiles and matrices determined the final solution.

Other diagnostic parameters, including the distribution of scaled residuals, G-space plots, the interpretability of factor profiles, and the reasonableness of temporal variations in source contributions, were also used to identify the optimal solution. The final solutions integrated these diagnostic parameters with DISP and bootstrap analysis (Paatero et al., 2013) to verify their robustness. No factor swaps were observed within the allowed dQmax range, and bootstrap simulations consistently demonstrated strong correlations with the baseline factors. The interpretation of potential sources was informed by both seasonal patterns and geochemical factors (Brown et al., 2015). This methodology ensured that the PMF analyses remained reliable, providing solutions that accurately reflected source contributions across the various monitoring sites. Based on these diagnostic parameters, PMF was applied individually at each site to identify PM sources, as it is capable of handling complex datasets and resolving multiple sources without requiring prior information about the sources (Querol et al., 2001; Kim et al., 2004; Amato et al., 2015; Liang et al., 2016; Weber et al., 2019; Hopke et al., 2020b).

2.3. Machine learning model: Extreme Gradient Boosting

Supervised learning is a machine learning approach where models learn from labelled data to predict unseen examples (Cunningham et al., 2008). XGBoost, as a popular machine learning library, demonstrates superior performance in supervised learning tasks through its gradient boosting algorithms (Chen et al., 2016). It is recognized for its scalability and efficiency. The gradient boosting method iteratively builds decision trees, where each new tree corrects the residual errors made by the previous trees, thereby minimizing the loss function and improving predictive accuracy. The gradient boosting method iteratively trains decision trees using input data, reducing the loss function and improving predictive accuracy. Initially, the mean of the target values is determined, followed by the calculation of residuals for each training data point as the preliminary predictions (Zhu et al., 2021). To prevent overfitting, an "early stopping round" parameter is applied to ensure that the model's iterations cease when no further improvements, such as reduced error, are achieved (Karimi et al., 2023).

In model construction and training, XGBoost enhances performance and processing capabilities through efficient feature layout and optimization techniques (Chen et al., 2016). During the feature layout transformation process, XGBoost organizes the feature values along with their gradient statistics. These statistics are used to efficiently identify the optimal split points during training (Fig. 1). The gradient statistics correspond to the sum of gradients for potential splits, which are crucial for determining the split that minimizes the loss function. In this study, each "feature value" represents the concentration of a specific observational species at a given time point. In the feature layout transformation process, each feature value is stored along with its corresponding gradient statistics, significantly accelerating the identification of split points during training (Fig. 1). Notably, it does not directly store missing values; instead, it maps feature values to instance indices via pointers, enabling the efficient handling of sparse data. Since feature columns are pre-sorted, it performs linear scans through cumulative gradient statistics to identify the optimal split point. These gradient statistics represent the sums of gradients for the left and right subtrees, which are critical for determining the split point that minimizes the loss function. By pre-sorting feature columns, it efficiently executes linear scans to quickly identify the optimal split point.

Hyperparameter tuning is crucial for optimizing model performance, as efficiently exploring hyperparameter configurations can significantly impact prediction accuracy. This is particularly effective in tuning the hyperparameters of complex models like XGBoost (Chen et al., 2016; Satpathy et al., 2024). Furthermore, in this study, the LPO algorithm is adopted to optimize the hyperparameters of the XGBoost model. LPO is an innovative approach that efficiently explores various hyperparameter combinations to determine the optimal configuration. For further details on the LPO algorithm, please refer to the supplementary materials.

2.4. Evaluation of model performance

In this study, the XGBoost model was employed to predict PM source contribution, with various evaluation metrics used to validate its performance. Our dataset uses the concentrations of 32 observational species as independent variables, with the concentration contributions of each pollution source for each sample derived from PMF as the dependent variables. Each pollution source is assigned a label, and the samples are categorized based on these labels. This allows the model to train and evaluate on data from different pollution sources, enabling it to learn the relationship between each observational species and each pollution

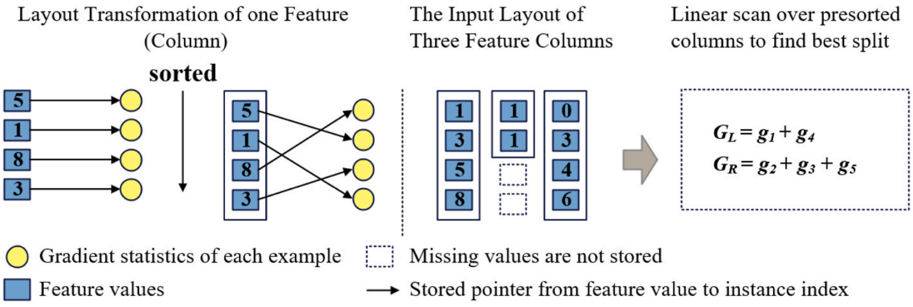
Layout Transformation of one Feature (Column)
sorted
5
1
8
3
The Input Layout of Three Feature Columns
1
3
5
8
1
1
0
3
4
6
Linear scan over presorted columns to find best split
$G_L = g_1 + g_4$
$G_R = g_2 + g_3 + g_5$
Gradient statistics of each example
Feature values
Missing values are not stored
Stored pointer from feature value to instance index

Fig. 1. Efficient feature layout transformation and split point calculation in XGBoost.

source through these samples. Our dataset were divided into training and test sets with three different ratios, namely 6:4, 7:3, and 8:2, to compare the performance of the models trained under each ratio. The results indicated that the optimal ratio was 8:2 (Table S2). Model performance was assessed through ten-fold cross-validation. In each iteration of the cross-validation process, the dataset was randomly divided into 10 folds. The model was trained on 9 of the folds (training set) and tested on the remaining 1-fold (test set). This process was repeated 10 times, with each subset serving as the test set once. Evaluation metrics, including the coefficient of determination (r^2), normalized root mean square error (RMSE), and normalized mean absolute error (MAE), were used to comprehensively assess the model's predictive accuracy and robustness. Normalized RMSE and normalized MAE are commonly used metrics for assessing model prediction performance. By comparing the errors to the data range, they eliminate the scale differences caused by units. These normalized metrics provide a unified evaluation approach, making comparisons between different models more intuitive (Willmott, 1982; Hyndman et al., 2006).

Furthermore, SVM have demonstrated strong generalization capabilities in numerous studies, making them particularly suitable for short-term air quality forecasting (Lei et al., 2023). RF models utilizing multi-source remote sensing data have also achieved high accuracy in predicting $PM_{2.5}$ concentrations (Zhong et al., 2022). Therefore, this study additionally employs SVM and RF for comparison with XGBoost to validate the performance of the XGBoost model. The techniques and parameters used for the SVM and RF models, along with their specific settings and configurations, are provided in the Supporting Information.

2.5. *Interpreting LPO-XGBoost model with SHAP method*

This study employs SHAP values to analyze the model. SHAP values, based on the concept of Shapley values, assign importance by quantifying the marginal contribution of each feature (Lundberg et al., 2017). In this study, SHAP values are used to quantify the predictive contribution of each observational species to each pollution source, rather than directly reflecting the mass contribution of the sources to the total PM mass. In this way, SHAP values help us understand the importance of each feature in the model's predictions and its impact. These values are used to explain the output of the LPO-XGBoost model and identify the relationships between features and pollutant concentrations, thereby providing deeper insights into the model's behavior and predictive logic (Lundberg et al., 2018). This approach enhances the interpretability of the model and offers essential data support and theoretical insights for developing more precise air quality management strategies.

3. Results and discussion

3.1. *Source apportionment of PM_{10} in Europe*

PMF was initially used to profile the PM_{10} sources at 21 monitoring sites across 6 countries (Table 1, Figs. S2 and S3). A total of 8 sources were identified, including road traffic (21 sites), biomass burning (20 sites), crustal (21 sites), industrial (8 sites), nitrate-rich particles (15 sites), sulfate-rich particles (16 sites), heavy oil combustion (10 sites), and sea salt (6 sites). For a detailed analysis of the source contributions obtained from PMF at these sites, please refer to the article Source

Table 1
Source contributions of different pollution types across monitoring sites.

Site	Road traffic	Biomass burning	Crustal	Industrial	Nitrate-rich particles	Sulfate-rich particles	Heavy oil combustion	Sea salt
BAS_UB	13 %	12 %	12 %	–	31 %	32 %	–	–
BCN_UB	21 %	–	13 %	11 %	–	24 %	10 %	21 %
CA_UB	28 %	40 %	6 %	–	10 %	16 %	–	–
COIM_UB	29 %	34 %	4 %	–	19 %	–	14 %	–
FLO_UB	20 %	33 %	7 %	–	15 %	18 %	–	7 %
GRA_UB	22 %	25 %	14 %	3 %	–	21 %	15 %	–
GRE-fr_UB	25 %	18 %	8 %	5 %	20 %	24 %	–	–
LEN_UB	8 %	16 %	5 %	7 %	41 %	20 %	3 %	–
MAG_UB	5 %	32 %	15 %	–	22 %	26 %	–	–
MRS-LCP_UB	33 %	20 %	16 %	–	–	21 %	10 %	–
MRS-AIX_UB	32 %	21 %	7 %	3 %	–	25 %	12 %	–
PAY_UB	41 %	8 %	15 %	–	31 %	5 %	–	–
ZUR_UB	16 %	10 %	6 %	–	32 %	36 %	–	–
BER_TR	13 %	29 %	9 %	–	24 %	25 %	–	–
COIM_TR	27 %	33 %	9 %	14 %	–	17 %	–	–
MAD-EA_TR	9 %	30 %	17 %	–	11 %	13 %	15 %	5 %
PORT_TR	33 %	21 %	16 %	–	15 %	–	–	15 %
BAI_UI	27 %	35 %	25 %	8 %	–	–	5 %	–
GIJ_UI	18 %	17 %	33 %	–	16 %	–	3 %	13 %
DEM_SUB	5 %	23 %	12 %	–	35 %	–	8 %	17 %
GRE-vif_SUB	20 %	25 %	6 %	2 %	22 %	25 %	–	–

Note:'-' represents the absence of the certain source.

apportionment of atmospheric pollutants based on the offline data in pan-European urban atmosphere (Liu et al., 2025b).

Road traffic sources contributed to all monitoring sites, having a relatively strong impact on PM_{10} concentrations at most sites. However, contributions at PAY_UB, MRS-LCP_UB, and PORT_TR exceeded 30 %, showing that high traffic density could significantly raise PM_{10} concentrations in these areas. In contrast, at sites like MAG_UB, LEN_UB, and DEM_SUB, traffic-related contributions were below 10 %, possibly due to fewer vehicles or effective pollution control measures.

Biomass burning is an important source of air pollution at all monitoring sites except for BCN_UB, with its contribution higher than 30 % for certain stations like CA_UB, BAI_UI, and COIM_UB, showing that biomass combustion plays a big role in the air quality in those regions (Table 1). On the other hand, ZUR_UB and PAY_UB show little biomass burning contribution, at 10 % and 8 %, respectively, or only a small part of the area was affected by this combustion source, reflecting fewer occurrences or better control in these areas.

Crustal sources caused pollution at all monitoring sites, with contribution rates ranging from 4 % to 33 % (Table 1). Notably, the GIJ_UI site showed a large 33 % contribution, highlighting the need for targeted strategies to reduce crustal source pollution.

Industrial sources contributed relatively little overall, with contributions exceeding 10 % only at COIM_TR and BCN_UB, showing a direct link between these areas and industrial activities. At sites with lower industrial emissions, such as GRE-fr_UB and GRE-vif_SUB, the contribution from industrial sources is less than 5 % (Table 1).

Nitrate-rich particles sources show varying contributions across different sites, including GRE-fr_UB, GRE-vif_SUB, LEN_UB, DEM_SUB, FLO_UB, COIM_UB, MAD-EA_TR, MAD-EV_UB, BAS_UB, BER_TR, MAG_UB, PAY_UB, and ZUR_UB, with contributions ranging from 11 % to 41 % (Table 1). Notably, nitrate-rich particles contributions at LEN_UB, DEM_SUB, BAS_UB, PAY_UB, and ZUR_UB exceed 30 %.

Sulfate-rich particles sources were observed at sites such as GRE-fr_UB, LEN_UB, MRS-AIX_UB, CA_UB, COIM_TR, PORT_TR, BCN_UB, GIJ_UI, BAS_UB, BER_TR, MAG_UB, and PAY_UB, with their contributions ranging from 5 % to 32 % (Table 1). Most of these sites have contributions in the range of 15 %–26 %, showing moderate pollution levels.

Heavy oil combustion sources were more important in GRA_UB and COIM_UB than other materials at port-adjacent sites, with contributions of 15 % and 14 %, respectively (Table 1). This shows that shipping emissions play a significant role in the level of air pollution in these coastal cities (Toscano, 2023).

Sea salt was another important source of PM_{10}, found at locations like BCN_UB, DEM_SUB, FLO_UB, GJ_UI, MAD-EA_TR, and PORT_TR, with contributions ranging from 5 % to 21 % (Table 1). Among these coastal sites, BCN_UB, PORT_TR, DEM_SUB, and GIJ_UI show particularly high contributions, each above 10 %.

Therefore, different pollution sources affect air quality in various ways, and their impact varies across regions, environmental conditions, and human activities. In some areas, biomass burning and surface dust are the primary sources of pollution, while in other regions, road traffic, industrial activities, and shipping have a greater impact on air quality.

3.2. *Performance comparison among models*

In this section, we compare the predictive performance of the SVM, RF, and XGBoost models, along with their versions optimized using the LPO parameter adjustment algorithm (LPO-SVM, LPO-RF, and LPO-XGBoost), across different pollution sources through 10-fold cross-validation.

For example, in the case of crustal source prediction, the standard RF model achieved an r^2 of 0.78, normalized RMSE of 0.061 and normalized MAE of 0.031. The XGBoost model showed better performance with an r^2 of 0.84, normalized RMSE of 0.052, and normalized MAE of 0.025. In contrast, the SVM model only achieved an r^2 of 0.56, with normalized RMSE of 0.087 and normalized MAE of 0.045 (Fig. 2). After applying LPO optimization, the RF model's performance improved significantly, with r^2 increasing to 0.83, normalized RMSE decreasing to 0.054, and normalized MAE dropping to 0.028. The XGBoost model also showed improvement, with r^2 rising to 0.87, normalized RMSE dropping to 0.047, and normalized MAE remaining at 0.026. Only the SVM model exhibited a slight decline, with r^2 dropping to 0.30, normalized RMSE at 0.111, and normalized MAE at 0.054 (Fig. 2). Additional performance statistics for the models are provided in Table S3. Overall, LPO-XGBoost demonstrated excellent performance in predicting emissions from biomass burning, crustal, heavy oil combustion, industrial, road traffic, nitrate-rich particles, sulfate-rich particles, and sea salt. Although LPO-XGBoost slightly underperformed compared to RF and XGBoost in predicting heavy oil combustion sources, it achieved the highest r^2 of 0.88 in total prediction performance, surpassing all other models (Fig. 2). The overall predictive performance is calculated by combining the predicted values and true values for all pollution sources. This method provides a more comprehensive assessment of the model's overall performance in predicting multiple pollution sources, thereby avoiding bias from the results of a single pollution source. The performance of each machine learning model varies across different pollution sources because, in this study, a separate machine learning model was trained for each pollution source. The relationships between different pollution sources and various species differ, and the concentration contributions of each pollution source, as determined by PMF analysis, also vary. These factors ultimately contribute to the differences observed in the model's training performance.

The exceptional performance of LPO-XGBoost can be attributed to XGBoost's inherent strength in handling large datasets (Chen et al., 2016). When combined with the LPO parameter optimization algorithm, the model benefits from local parameter adjustments that further improve its fitting ability and predictive efficiency. In contrast, both SVM and LPO-SVM exhibited poorer performance (Fig. 2 & Table S3), likely due to the inherent limitations of SVM in managing high-dimensional data, especially in the presence of complex nonlinear relationships (Cervantes et al., 2020).

Therefore, based on the above comparisons, LPO-XGBoost is established as an innovative model that has been demonstrated to be an outstanding solution for addressing the challenges associated with complex pollution source predictions, solidifying its position as the optimal model for such tasks. And the analysis of each pollution source

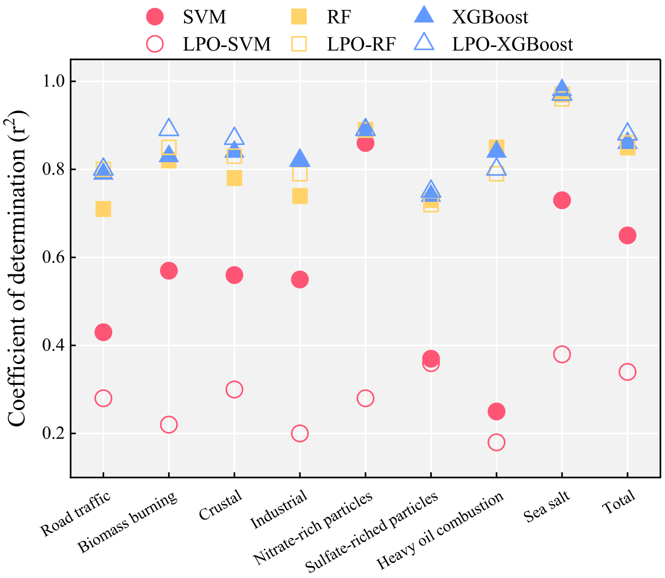

Fig. 2. The r^2 results of 10-fold cross-validation for different models under various pollution sources.

in the following sections will be based on the experimental results of LPO-XGBoost.

3.3. Performance analysis of LPO-XGBoost model

Fig. 3 shows the predictive performance of the LPO-XGBoost model for each pollution source, where scatter plots compare the predicted values with the actual values obtained by PMF. Each subplot corresponds to a specific pollution source, clearly demonstrating the model's ability to capture the complexity of various sources and its predictive accuracy. The majority of predicted values align closely with the 1:1 line (red line in the figure), indicating the model's ability to accurately predict the PMF results on the test set.

To further evaluate the performance of the LPO-XGBoost model, SHAP values were applied to explain the characteristic species of each pollution source. This approach helps identify the contribution of individual chemical species to the model's predictions. The results are shown in Fig. 4, where red dots represent samples with higher feature values, and blue dots represent samples with lower feature values. The gradient between these colors illustrates the varying impacts of feature values, where higher feature values (red) have a larger influence, and lower feature values (blue) have a smaller impact. The vertical axis is ordered by descending importance of the related species or elements. Features closer to the top generally have a greater contribution to the model's output, emphasizing their significance in shaping the prediction results. The X-axis represents SHAP values, which quantify the impact of individual attributes on the model's predictions. Positive values indicate that the attribute's value increases the prediction, while negative values indicate that the attribute's value decreases the prediction. SHAP analysis assigns a SHAP value to each sample. Negative values may arise because certain attributes are negatively correlated with the target variable or serve to reduce the prediction for specific samples, leading to negative adjustments.

The following discussion of the SHAP values obtained for each factor is compared to the results on the composition of the chemical profiles obtained by the PMF, in order to evaluate the consistency of the two approaches.

The r^2 value for road traffic sources is 0.80 (Fig. 3a), with key characteristic species including EC, Ca^{2+}, Fe, Cr, Sn, and Cu (Fig. 4a), which closely align with the results obtained from PMF analysis (Fig. S3). For biomass burning, the r^2 reaches 0.89, indicating that the model performs well in predicting this pollution source, with predicted values closely matching the true values (Fig. 3b). The primary characteristic species identified through SHAP analysis include OC, LGA, K, EC, and K^+, with OC and LGA having the highest SHAP values (Fig. 4b), indicating that these two substances contribute the most to the prediction of biomass burning sources (Liu et al., 2023). Similarly, the source apportionment results obtained through PMF primarily identify OC, EC, K^+/K, and LGA as key markers (Fig. S3). For crustal sources, the r^2 on the test set reaches 0.87 (Fig. 3c), with SHAP analysis identifying the primary characteristic species as Al, Ti, Fe, and Ca^{2+}, with Al and Ti having the highest SHAP values (Fig. 4a). PMF source apportionment results also show that Al, Ti, V, Ca^{2+}/Ca, Fe, and Mg^{2+}/Mg contribute most significantly to crustal sources (Fig. S3). For industrial sources, the r^2 value is 0.82 (Fig. 3d), with the primary characteristic species identified through SHAP analysis including NH_4^+, Cr, NO_3^-, Zn, Rb, Ba, Mg, and Pb (Fig. 4d). In contrast, PMF analysis identifies Ni, Cr, Zn, Pb, and Mn as the key species (Fig. S3). The presence of numerous marker elements is attributed to the diversity of industrial sectors across different monitoring sites, leading to variations in the markers (Rodríguez et al., 2004; Amatoa et al., 2014; Minguillón et al., 2014).

The LPO-XGBoost model achieved an r^2 value of 0.89 for predicting the nitrate-rich particle source (Fig. 3e), with NO_3- and NH_4^+ having the highest SHAP values (Fig. 4g), which is consistent with the results obtained from PMF analysis (Fig. S3). However, for the sulfate-rich particle source, the r^2 is the lowest among all sources at 0.75 (Fig. 3f). The sulfate-rich source includes both primary and secondary sulfates (Fig. S3). Secondary sulfates are typically associated with NH_4^+ and SO_4^{2-} (Cheng et al., 2016; Wang et al., 2016, 2018, 2023b; Liu et al., 2024a), which is reflected in their higher SHAP values (Fig. 4h). Primary sulfates may originate from coal and oil combustion, which generates elements such as Se, As, Pb, V, Ni, and OC (Jafarinejad, 2016; Pokorná et al., 2018), or from dust and mineral particles (Tang et al., 2019; Wu et al., 2020). However, SHAP analysis shows that elements from mixed sources such as Se, As, Pb, V, Ni, and OC contribute minimally (Fig. 4h). This result suggests that the LPO-XGBoost model struggles to capture the complex impacts of mixed and aged sources, such as those from coal and oil combustion, leading to reduced prediction accuracy.

The r^2 value for the heavy oil combustion source is 0.80 (Fig. 3g), with the primary characteristic species identified through SHAP analysis being V and Ni (Fig. 4c), which are typical markers for this source (Fig. S3). The r^2 value for sea salt sources is 0.97 (Fig. 3h), showing the highest predictive accuracy among all pollution sources. The SHAP values for Na and Na^+ are the highest, followed by Mg and Cl^- (Fig. 4f). In the PMF analysis, the primary characteristic species for sea salt sources are Na^+/Na and Cl^-/Cl, consistent with existing literature

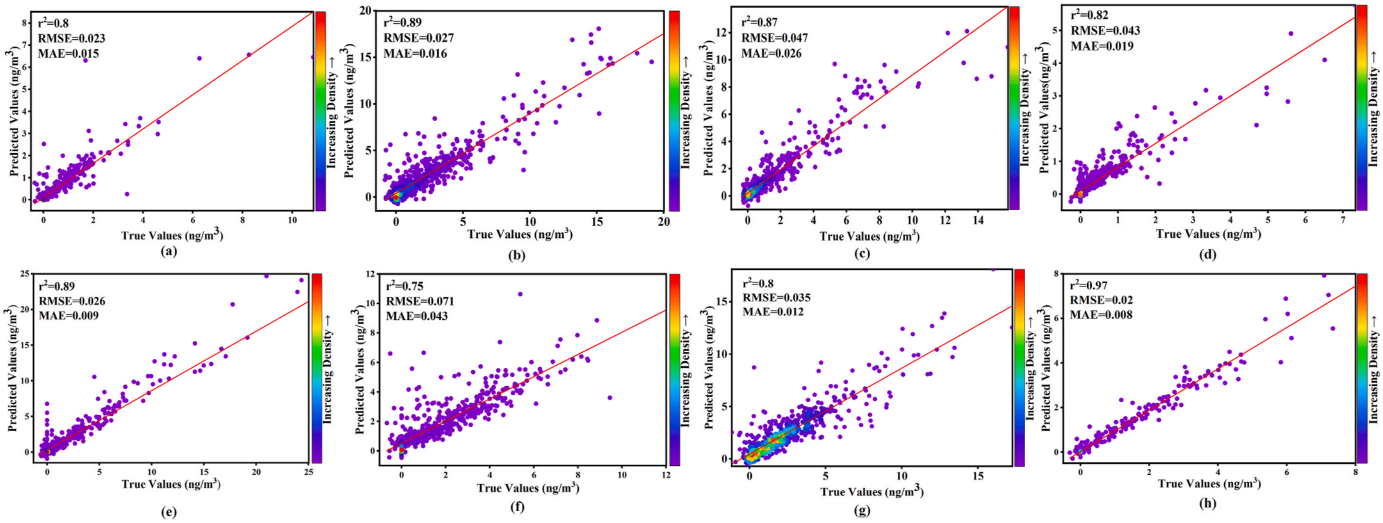

Fig. 3. Scatter plot comparing predicted and true values for each pollution source using the LPO-XGBoost model on the test set: (a) Road traffic, (b) Biomass burning, (c) Crustal, (d) Industrial, (e) nitrate-rich particles, (f) Sulfate-rich particles, (g) Heavy oil combustion and (h) Sea salt.

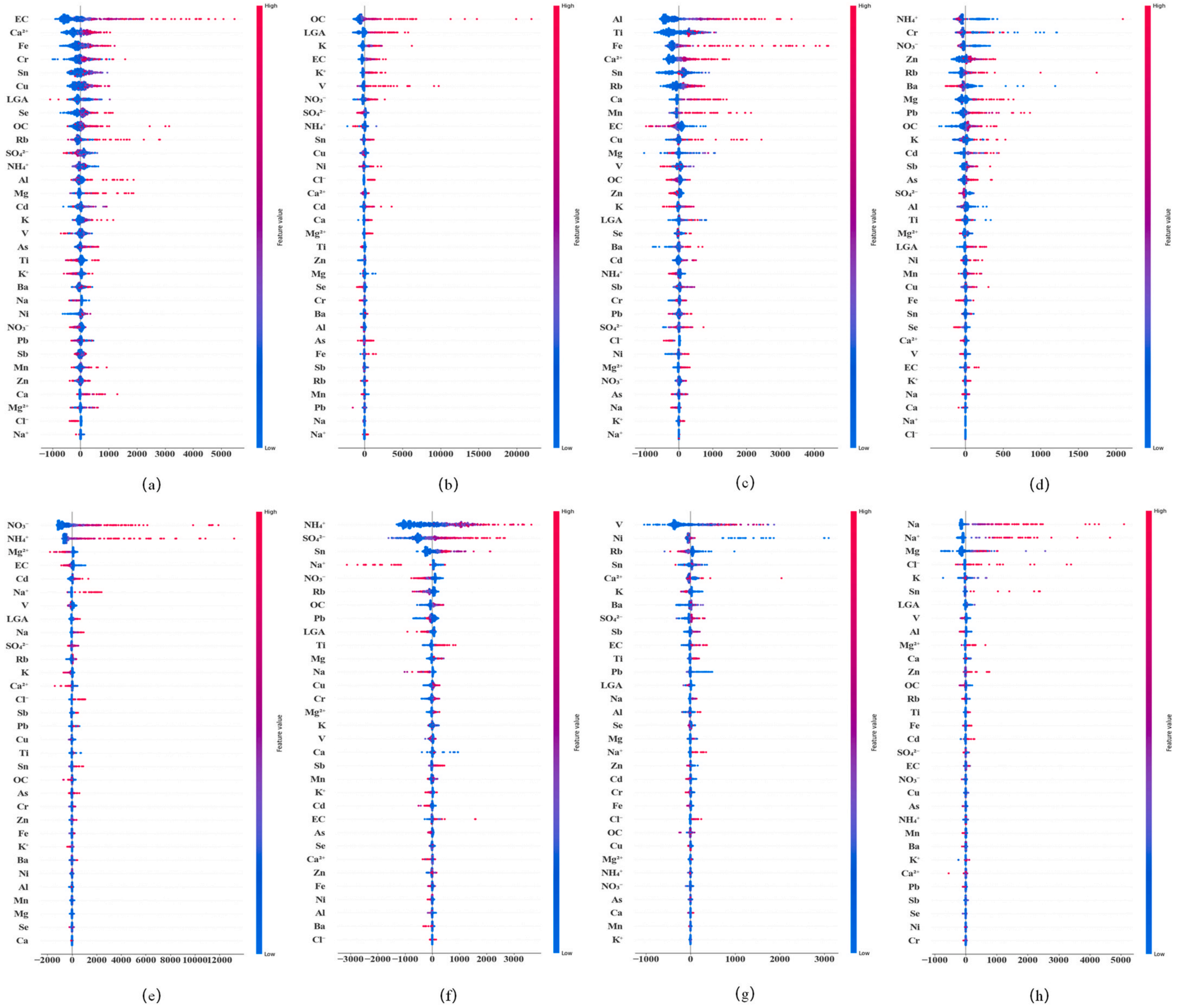

Fig. 4. SHAP value distribution for different pollution sources using the LPO-XGBoost model: (a) Road traffic, (b) Biomass burning, (c) Crustal, (d) Industrial, (e) Nitrate-rich particles, (f) Sulfate-rich particles (g) Heavy oil combustion, and (h) Sea salt.

(Aldabe et al., 2011). Therefore, the LPO-XGBoost model accurately captures the characteristic species of these sources.

3.4. Limitations

Although the dataset spans 2013 to 2021, each monitoring site lacks data for every year within this period, with combined data from all sites covering the entire timeframe (Table S4). This lack of data continuity may introduce a degree of bias when predicting trends over extended time scales. Despite the lack of continuity in the temporal scale of the data, LPO-XGBoost achieved an r^2 of 0.88 for overall prediction performance. With improved temporal continuity in the data, the model's performance is expected to further improve. To minimize the impact of limited data coverage, the XGBoost model used demonstrates robust predictive accuracy even under conditions of limited data coverage. However, we still need to measure the concentration data for each indicator. However, if the pollution sources in the region remain stable over the long term, PMF analysis may no longer be necessary, as our model can be directly applied for prediction, enabling faster source apportionment results. This capability is attributed to its tree-based ensemble learning structure, which effectively captures complex relationships within sparse data (Zhu et al., 2021).

Additionally, the monitoring data in this study were collected from multiple sites across different countries. While most monitoring sites measured the same species, some sites may lack data for certain key species (Table S5). This inconsistency could introduce significant errors in the model's predictions at sites where critical species data are missing. For instance, at the coastal site MRS-LCP_UB, the sea salt source was not identified due to the absence of key species associated with sea salt, such as Na/Na^+ and Cl/Cl^- ions, in the source apportionment analysis. The PMF indicates, in the process of source apportionment, that different pollution sources have varying dependencies on specific species, and the absence of these key species may impact the LPO-XGBoost's accuracy in identifying certain pollution sources. This issue is especially pronounced at sites with sparse species measurements, where the model may exhibit prediction bias. This limitation suggests that future work could focus on standardizing or supplementing monitoring data to improve the coverage of key species, thereby enhancing the model's applicability across all monitoring sites. It is also worth noting that in this study, the concentration contribution data for each pollution source were derived by integrating the individual PMF analysis results from each site. This approach was chosen because the PMF results from a single site provide more precise estimates compared to a combined analysis of all sites. Furthermore, a separate LPO-XGBoost model was trained for each pollution source to ensure that the model could more accurately learn the specific characteristics and differences of the PMF results from each site, thus improving prediction accuracy and enhancing the model's generalization capability.

Furthermore, according to the results presented in Section 3.3, although the LPO-XGBoost model performs strongly in predicting various pollution sources, its accuracy is notably lower for specific types of sources, such as industrial sources and sulfate-enriched sources. SHAP analysis reveals that some feature species contribute minimally to these pollution sources, which may stem from the model's limited ability to analyze complex mixed pollution from multiple sources. However, the current selection of chemical species used as input data species may not effectively capture the characteristics of these mixed sources. This issue is also related to the fact that PMF, in some cases, may produce mixed factors, a phenomenon that can arise due to various reasons, such as the number of samples and species used in the analysis. This could explain why sulfate-enriched sources include components from secondary organic aerosols and other sources emissions, which in turn results in slightly lower performance of the LPO-XGBoost model when predicting these sources. This limitation suggests that future research should focus on incorporating additional key species and optimizing the model's ability to identify mixed sources, thereby improving its overall predictive performance and ability to differentiate between pollution sources. Additionally, apart from the existing pollution sources, local activities such as off-road traffic, generators, and construction machinery also play a significant role in PM_{10} emissions. These activities are particularly prevalent in urban development areas but have not been adequately considered in the current study. Future research should incorporate these activities into the analysis.

4. Conclusions and outlooks

This study uses results from a large PMF analysis of the composition and contributions of pollution sources at 21 monitoring sites across 6 European countries, including road traffic, biomass burning, crustal, industrial, nitrate-rich particles, sulfate-rich particles, heavy oil combustion, and sea salt. The contributions of these sources varied due to differences in geographic and environmental conditions and human activities.

In order to develop near-real-time source apportionment results quickly and accurately without manual intervention, we explored machine learning methods based on a large set of prior results to predict pollution source contributions, including XGBoost, RF, and SVM, along with their LPO-enhanced variants. We ultimately selected the LPO-XGBoost model, which, after training with pollution source concentration contribution samples derived from PMF analysis, achieved an overall averaged r^2 of 0.88 for the contribution of all sources across all sites, based on a comparison between PMF results and the test set samples. A separate model was trained for each pollution source. The r^2 values for all sources were greater than or equal to 0.75. Compared with other machine learning models, including RF, SVM, and their LPO-enhanced variants, the LPO-XGBoost model demonstrates the best overall predictive performance, indicating its particular suitability for complex data and large-scale pollution source apportionment.

Future research is expected in order to enhance and test the model capabilities, for example withcollecting data with broader temporal coverage and ensuring consistent measurement of key speciesacross all monitoring sites. For instance, the inclusion of key species data for sulfate-rich particle sources.particularly those associated with other pollutants such as coal and oil combustion, should be consideredAddressing issues related to site data continuity and species variability will also further enhance themodel's prediction accuracy. Additionally, future studies will integrate the contribution of local activities to PM_{10} into the existing model, thereby enhancing the comprehensiveness of the source apportionment analysis.

CRediT authorship contribution statement

Ying Liu: Writing – original draft. **Bowen Jin:** Writing – original draft, Formal analysis. **Xun Zhang:** Writing – review & editing. **Xiansheng Liu:** Supervision, Data curation, Conceptualization. **Tao Wang:** Writing – review & editing. **Vy Ngoc Thuy Dinh:** Supervision, Investigation, Data curation. **Jean-Luc Jaffrezo:** Supervision, Investigation, Data curation. **Gaëlle Uzu:** Supervision, Investigation, Data curation. **Pamela Dominutti:** Supervision, Investigation, Data curation. **Sophie Darfeuil:** Supervision, Investigation, Data curation. **Olivier Favez:** Supervision, Investigation, Conceptualization. **Sébastien Conil:** Supervision, Investigation, Data curation. **Nicolas Marchand:** Supervision, Investigation, Data curation. **Sonia Castillo:** Supervision, Investigation, Data curation. **Jesús D. de la Rosa:** Supervision, Investigation, Data curation. **Stuart Grange:** Supervision, Investigation, Data curation. **Christoph Hueglin:** Supervision, Investigation, Data curation. **Konstantinos Eleftheriadis:** Supervision, Investigation, Data curation. **Evangelia Diapouli:** Supervision, Investigation, Data curation. **Manousos-Ioannis Manousakas:** Supervision, Investigation, Data curation. **Maria Gini:** Supervision, Investigation, Data curation. **Giulia Calzolai:** Supervision, Investigation, Data curation. **Célia Alves:** Supervision, Investigation, Data curation. **Marta Monge:** Project administration. **Cristina Reche:** Supervision, Investigation, Data curation. **Roy M. Harrison:** Writing – review & editing. **Philip K. Hopke:** Writing – review & editing. **Andrés Alastuey:** Writing – review & editing. **Xavier Querol:** Writing – review & editing, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

5. Acknowledgements

This study has been supported and been done by RI-URBANS project (Research Infrastructures Services Reinforcing Air Quality Monitoring Capacities in European Urban & Industrial Areas, European Union's Horizon 2020 research and innovation program, Green Deal, European Commission, contract 101036245). Furthermore, additional support has been received from State Key Laboratory of Resources and Environmental Information System, the National Natural Science Foundation of China (42407566, 42205099, 72441001), and the Chunhui Project Foundation of the Education Department of China (HZKY20220053), the Science Fund for Creative Research Groups of the National Natural Science Foundation of China (72221002) that carried out the ML implementation. Meanwhile, samples in France were collected within many research and Air Quality assessment programs, including the programs CARA (funded by the Ministry of Environment within the LCSQA), DECOMBIO, CAMERA, and QAMECS (all funded by Ademe), ACME and MIAI-Airquality (funded by University Grenoble Alpes), OPE – Andra (funded by Andra), and multiple fundings by Atmo AuRA, Atmo Sud, Atmo Grand Est, Atmo Haut de France, Atmo Normandie, for the sampling and analyses. We would like to express our deep thanks to many people in the AASQA France for the sampling of all these samples, and to people in several laboratories in France, including IGE, for the analyses of these samples. The University of Aveiro thanks the Foundation for Science and Technology (FCT) for funding CESAM (UID Centro de Estudos do Ambiente e do Mar + LAP/0094/2020). The University of Granada acknowledges the financial support of the Spanish

Ministry of Science and Innovation through the project ELPIS PID2020-120015RB-I00. Samples in Switzerland were collected by the Swiss National Air Pollution Monitoring Network NABEL (BAFU/Empa).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envres.2025.121659.

Data availability

Data will be made available on request.

References

Aldabe, J., Elustondo, D., Santamaría, C., Lasheras, E., Pandolfi, M., Alastuey, A., Querol, X., Santamaría, J.M., 2011. Chemical characterisation and source apportionment of PM2.5 and PM10 at rural, urban and traffic sites in navarra (North of Spain). Atmos. Res. 102 (1–2), 191–205.

Alias, N.F., Khan, F., Khan, F., Sairi, N.A., Zain, S.M., Suradi, H., Rahim, H.A., Banerjee, T., Bari, M.A., Othman, M., Latif, M.T., 2020. Characteristics, Emission Sources, and Risk Factors of Heavy Metals in PM2.5 from Southern Malaysia.

Amato, F., Alastuey, A., Karanasiou, A., Lucarelli, F., Nava, S., Calzolai, G., Severi, M., Becagli, S., Gianelle, V.L., Colombi, C., Alves, C.A., Custódio, D., Nunes, T., Cerqueira, M., Pio, C., Eleftheriadis, K., Diapouli, E., Reche, C., María, Minguillón, C., Manousakas, M.I., Maggos, T., Vratolis, S., Harrison, R.M., Querol, X., 2015. AIRUSE-LIFE+: a harmonized PM speciation and source apportionment in five southern European cities. Atmos. Chem. Phys. 16, 3289–3309.

Amatoa, u., Casseeb, F.R., Gonc, H.A. C.D.v. d., Gehrigd, R., Gustafssone, a., Hafner, W., Harrisong, R.M., Jozwickac, M., Kellyh, J., Morenoa, T., Prévôt, A.S.H., Schaapc, M., Sunyer, J., Querola, A., 2014. Eview Rban Air Quality : the Challenge of Traffic Non-exhaust Emissions.

Balachandran, S., Chang, H.H., Pachon, J.E., Holmes, H.A., Mulholland, J.A., Russell, A. G., 2013. Bayesian-based ensemble source apportionment of PM2.5. Environ. Sci. Technol. 47 23, 13511–13518.

Belis, C.A., Larsen, B.R., Amato, F., Haddad, I.E., Viana, M., 2013. *European Guide on Air Pollution Source Apportionment with Receptor Models*: European Guide on Air Pollution Source Apportionment with Receptor Models.

Borlaza, L.J.S., Weber, S., Uzu, G., Jacob, V., Cañete, T., Micallef, S., Trébuchon, C., Slama, R., Favez, O., Jaffrezo, J.L., 2020. Disparities in particulate matter (PM10) origins and oxidative potential at a city scale (Grenoble, France) – part 1: source apportionment at three neighbouring sites. Atmos. Chem. Phys. 21, 5415–5437.

Brown, S.G., Eberly, S.I., Paatero, P., Norris, G.A., 2015. Methods for estimating uncertainty in PMF solutions: examples with ambient air and water quality data and guidance on reporting PMF results. Sci. Total Environ. 518–519, 626–635.

Cavalli, F., Viana, M., Yttri, K.E., Genberg, J., Putaud, J.P., 2010. Toward a standardised thermal-optical protocol for measuring atmospheric organic and elemental carbon: the EUSAAR protocol. Atmos. Meas. Tech. 3 (1).

Cervantes, J., García, F., Rodríguez-Mazahua, L., Chau, A.L., 2020. A comprehensive survey on support vector machine classification: applications, challenges and trends. Neurocomputing 408, 189–215.

Chen, Q., Wang, M., Sun, H., Wang, X., Wang, Y., Li, Y., Zhang, L., Mu, Z., 2018. Enhanced health risks from exposure to environmentally persistent free radicals and the oxidative stress of PM2.5 from Asian dust storms in erenhot, Zhangbei and Jinan, China. Environ. Int. 121, 260–268. https://doi.org/10.1016/j.envint.2018.09.012.

Chen, T., Guestrin, C., Assoc Comp, M., 2016. XGBoost: a scalable tree boosting system. Paper Presented at the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (KDD), San Francisco, CA, pp. 785–794, 2016 Aug 13-17.

Cheng, Y., Zheng, G., Wei, C., Mu, Q., Zheng, B., Wang, Z., Gao, M., Zhang, Q., He, K., Carmichael, G.R., Pöschl, U., Su, H., 2016. Reactive nitrogen chemistry in aerosol water as a source of sulfate during haze events in China. Sci. Adv. 2.

Cohen, A.J., Brauer, M., Burnett, R., Anderson, H.R., Frostad, J., Estep, K., Balakrishnan, K., Brunekreef, B., Dandona, L., Dandona, R., Feigin, V., Freedman, G., Hubbell, B., Jobling, A., Kan, H., Knibbs, L., Liu, Y., Martin, R., Morawska, L., Pope III, C.A., Shin, H., Straif, K., Shaddick, G., Thomas, M., van Dingenen, R., van Donkelaar, A., Vos, T., Murray, C.J.L., Forouzanfar, M.H., 2017. 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. Lancet 389 (10082), 1907–1918. https://doi.org/10.1016/s0140-6736(17)30505-6.

Cohen, D.D., Stelcer, E., Santos, F.L., Prior, M., Thompson, C.A., Pabroa, P.C.B., 2009. Fingerprinting and source apportionment of fine particle pollution in manila by IBA and PMF techniques : a 7-year study. X Ray Spectrom. 38, 18–25.

Croft, D.P., Zhang, W., Lin, S., Thurston, S.W., Hopke, P.K., van Wijngaarden, E., Squizzato, S., Masiol, M., Utell, M.J., Rich, D.Q., 2019. Associations between source-specific particulate matter and respiratory infections in New York state adults. Environ. Sci. Technol. 54 (2), 975–984. https://doi.org/10.1021/acs.est.9b04295.

Cunningham, P., Cord, M., Delany, S.J., 2008. Supervised learning. In: Cord, M., Cunningham, P. (Eds.), Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 21–49.

Dai, Q., Hopke, P.K., Bi, X., Feng, Y., 2020. Improving apportionment of PM2.5 using multisite PMF by constraining G -values with a priori information. Sci. Total Environ. 736. https://doi.org/10.1016/j.scitotenv.2020.139657.

De Angelis, E., Carnevale, C., Turrini, E., Volta, M., 2020. Source apportionment and integrated assessment modelling for air quality planning. Electronics 9 (7). https://doi.org/10.3390/electronics9071098.

Dominguez, A., Dadvand, P., Cirach, M., Arevalo, G., Barril, L., Foraster, M., Gascon, M., Raimbault, B., Galmes, T., Gomez-Herrera, L., Persavento, C., Samuelsson, K., Lao, J., Moreno, T., Querol, X., Jerrett, M., Schwartz, J., Tonne, C., Nieuwenhuijsen, M.J., Sunyer, J., Basagana, X., Rivas, I., 2024. Development of land use regression, dispersion, and hybrid models for prediction of outdoor air pollution exposure in Barcelona. Sci. Total Environ. 954, 176632. https://doi.org/10.1016/j.scitotenv.2024.176632, 176632.

Gao, A., Wang, J., Luo, J., Li, A., Chen, K., Wang, P., Wang, Y., Li, J., Hu, J., Zhang, H., 2021. Temporal variation of PM2.5 -associated health effects in Shijiazhuang, Hebei. Front. Environ. Sci. Eng. 15 (5). https://doi.org/10.1007/s11783-020-1376-0.

Ghasemi, M., Zare, M., Zahedi, A., Trojovsky, P., Abualigah, L., Trojovska, E., 2024. Optimization based on performance of lungs in body: lungs performance-based optimization (LPO). Comput. Methods Appl. Mech. Eng. 419. https://doi.org/10.1016/j.cma.2023.116582.

Glojek, K., Thuy, V.D.N., Weber, S., Uzu, G., Manousakas, M., Elazzouzi, R., Dzepina, K., Darfeuil, S., Ginot, P., Jaffrezo, J.L., Zabkar, R., Tursic, J., Podkoritnik, A., Mocnik, G., 2024. Annual variation of source contributions to PM 10 and oxidative potential in a mountainous area with traffic, biomass burning, cement-plant and biogenic influences. Environ. Int. 189. https://doi.org/10.1016/j.envint.2024.108787.

Hopke, P.K., 2016. Review of receptor modeling methods for source apportionment. J. Air Waste Manag. Assoc. 66 (3), 237.

Hopke, P.K., Croft, D.P., Zhang, W., Lin, S., Masiol, M., Squizzato, S., Thurston, S.W., van Wijngaarden, E., Utell, M.J., Rich, D.Q., 2020a. Changes in the hospitalization and ED visit rates for respiratory diseases associated with source-specific PM2.5 in New York state from 2005 to 2016. Environ. Res. 181. https://doi.org/10.1016/j.envres.2019.108912.

Hopke, P.K., Dai, Q., Li, L., Feng, Y., 2020b. Global review of recent source apportionments for airborne particulate matter. Sci. Total Environ. 740.

Hyndman, R.J., Koehler, A.B., 2006. Another look at measures of forecast accuracy. Int. J. Forecast. 22, 679–688.

Jafari, A.J., Charkhloo, E., Pasalari, H., 2021. Urban air pollution control policies and strategies: a systematic review. Journal of environmental health science & engineering 19 (2), 1911–1940.

Jafarinejad, S., 2016. Control and Treatment of Sulfur Compounds Specially Sulfur Oxides (Sox) Emissions from the Petroleum Industry: a Review.

Karimi, S., Asghari, M., Rabie, R., Niri, M.E., 2023. Machine learning-based white-box prediction and correlation analysis of air pollutants in proximity to industrial zones. Process Saf. Environ. Prot. 178, 1009–1025. https://doi.org/10.1016/j.psep.2023.08.096.

Kim, E., Hopke, P.K., 2004. Source apportionment of fine particles in Washington, DC, utilizing temperature-resolved carbon fractions. J. Air Waste Manag. Assoc. 54, 773–785.

Kim, E., Larson, T.V., Hopke, P.K., Slaughter, C., Sheppard, L., Claiborn, C.S., 2003. Source identification of PM2.5 in an arid Northwest U.S. City by positive matrix factorization. Atmos. Res. 66, 291–305.

Kleinman, M.T., Kneip, T.J., Eisenbud, M., 1976. Seasonal patterns of airborne particulate concentrations in New York city. Atmos. Environ. 10 (1), 9–11.

Lei, T.M.T., Ng, S.C.W., Siu, S.W.I., 2023. Application of ANN, XGBoost, and other ML methods to forecast air quality in Macau. Sustainability 15 (6). https://doi.org/10.3390/su15065341.

Liang, C.-S., Duan, F., He, K., Ma, Y.-l., 2016. Review on recent progress in observations, source identifications and countermeasures of PM2.5. Environ. Int. 86, 150–170.

Liang, Y.-C., Maimury, Y., Chen, A.H.-L., Juarez, J.R.C., 2020. Machine learning-based prediction of air quality. Applied Sciences-Basel 10 (24). https://doi.org/10.3390/app10249151.

Liu, X., Lara, R., Dufresne, M., Wu, L., Zhang, X., Wang, T., Monge, M., Reche, C., Di Leo, A., Lanzani, G., Colombi, C., Font, A., Sheehan, A., Green, D.C., Makkonen, U., Sauvage, S., Salameh, T., Petit, J.-E., Chatain, M., Coe, H., Hou, S., Harrison, R.M., Hopke, P., Petäjä, T., Alastuey, A., Querol, X., 2024a. Variability of ambient air ammonia in urban Europe (Finland, France, Italy, Spain, and the UK). Environ. Int. 185, 108519.

Liu, X., Zhang, X., Dufresne, M., Wang, T., Wu, L., Lara, R., Seco, R., Monge, M., Yáñez-Serrano, A.M., Gohy, M., Petit, P., Chevalier, A., Vagnot, M.P., Fortier, Y., Baudic, A., Ghersi, V., Gille, G., Lanzi, L., Gros, V., Simon, L., Héllen, H., Reimann, S., Le Bras, Z., Müller, M.J., Beddows, D., Hou, S., Shi, Z., Harrison, R.M., Bloss, W., Dernie, J., Sauvage, S., Hopke, P.K., Duan, X., An, T., Lewis, A.C., Hopkins, J.R., Liakakou, E., Mihalopoulos, N., Zhang, X., Alastuey, A., Querol, X., Salameh, T., 2025a. Measurement report: exploring the variations in ambient BTEX in urban Europe and their environmental health implications. Atmos. Chem. Phys. 25 (1), 625–638. https://doi.org/10.5194/acp-25-625-2025.

Liu, X., Zhang, X., Jin, B., Hadiatullah, H., Zhang, L., Zhang, P., Wang, T., Deng, Q., Querol, X., 2023. Online monitoring of carbonaceous aerosols in a northern Chinese city: temporal variations, main drivers, and health risks. Atmos. Environ. 316, 120169.

Liu, X., Zhang, X., Wang, R., Liu, Y., Hadiatullah, H., Xu, Y., Wang, T., Bendl, J., Adam, T., Schnelle-Kreis, J., Querol, X., 2024b. High-precision microscale particulate matter prediction in diverse environments using a long short-term memory neural network and street view imagery. Environ. Sci. Technol. 58, 3869–3882.

Liu, X., Zhang, X., Wang, T., Jin, B., Wu, L., Lara, R., Monge, M., Reche, C., Jaffrezo, J.-L., Uzu, G., Dominutti, P., Darfeuil, S., Favez, O., Conil, S., Marchand, N., Castillo, S., de la Rosa, J.D., Stuart, G., Eleftheriadis, K., Diapouli, E., Gini, M.I., Nava, S., Alves, C., Wang, X., Xu, Y., Green, D.C., Beddows, D.C.S., Harrison, R.M., Alastuey, A., Querol, X., 2024c. PM10-bound trace elements in Pan-European urban atmosphere. Environ. Res. 260. https://doi.org/10.1016/j.envres.2024.119630.

Liu, X., Zhang, X., Jin, B., Wang, T., Dinh, V.N.T., Jaffrezo, J.-L., Uzu, G., Dominutti, P., Darfeuil, S., Favez, O., Conil, S., Marchand, N., Castillo, S., de la Rosa, J.D., Stuart, G., Eleftheriadis, K., Diapouli, E., Manousakas, M.-I., Nava, S., Alves, C., Monge, M., Reche, C., Harrison, R.M., Hopke, P.K., Alastuey, A., Querola, X., 2025b. Source apportionment of PM10 based on offline chemical speciation data in urban Europe. submit to Npj Climate and Atmospheric Science for Peer Review.

Lundberg, S., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Arxiv arXiv:1705.07874, 4768–4777.

Lundberg, S.M., Erion, G.G., Lee, S.I., 2018. Consistent individualized feature attribution for tree ensembles. Arxiv, arXiv:1802.03888. 9 (12), 1–8.

Manousakas, M.I., Papaefthymiou, H., Diapouli, E., Migliori, A., Karydas, A.-G., Bogdanović-Radović, I., Eleftheriadis, K., 2017. Assessment of PM2.5 sources and their corresponding level of uncertainty in a coastal urban area using EPA PMF 5.0 enhanced diagnostics. Sci. Total Environ. 574, 155–164.

Mardoñez, V., Pandolfi, M., Borlaza, L.J.S., Jaffrezo, J.L., Alastuey, A., Besombes, J.L., Moreno, R.I., Pérez, N., Močnik, G., Ginot, P., Krejci, R., Chrastný, V., Wiedensohler, A., Laj, P., Andrade, M., Uzu, G., 2023. Source apportionment study on particulate air pollution in two high-altitude Bolivian cities: La Paz and El Alto. Atmos. Chem. Phys. 23, 10325–10347.

Miller, M.S., Friedlander, S.K., Hidy, G.M., 1972. A chemical element balance for the Pasadena aerosol. J. Colloid Interface Sci. 39 (1), 165–176.

Minguillón, M.C., Cirach, M., Hoek, G., Brunekreef, B., Tsai, M.-Y., Hoogh, K.d., Jedynska, A.D., Kooter, I.M., Nieuwenhuijsen, M.J., Querol, X., 2014. Spatial variability of trace elements and sources for improved exposure assessment in Barcelona. Atmos. Environ. 89, 268–281.

Moon, K.J., Han, J.S., Ghim, Y.S., Kim, Y.J., 2008. Source apportionment of fine carbonaceous particles by positive matrix factorization at Gosan background site in East Asia. Environ. Int. 34 (5), 654–664.

Nagar, J.K., Akolkar, A.B., Kumar, R., 2014. A review on airborne particulate matter and its sources, chemical composition and impact on human respiratory system. Int. J. Environ. 5 (2), 447–463.

Nieder, R., Benbi, D.K., Reichl, F.-X., 2018. Soil-Borne Particles and Their Impact on Environment and Human Health.

Paatero, P., Eberly, S., Brown, S.G., Norris, G.A., 2014. Methods for estimating uncertainty in factor analytic solutions. Atmos. Meas. Tech. 7 (3), 781–797. https://doi.org/10.5194/amt-7-781-2014.

Paatero, P., Eberly, S.I., Brown, S., Norris, G.A., 2013. Methods for estimating uncertainty in factor analytic solutions. Atmos. Meas. Tech. 7, 781–797.

Paatero, P., Tapper, U., 1994. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. Environmetrics 5 (2).

Pan, B., 2018. Application of XGBoost algorithm in hourly PM2.5 concentration prediction. IOP Conf. Ser. Earth Environ. Sci. 113.

Peng, Z., Zhang, B., Wang, D., Niu, X., Sun, J., Xu, H., Cao, J., Shen, Z., 2024. Application of machine learning in atmospheric pollution research: a state-of-art review. Sci. Total Environ. 910. https://doi.org/10.1016/j.scitotenv.2023.168588.

Pokorná, P., Schwarz, J., Krejci, R., Swietlicki, E., Havránek, V., Ždímal, V., 2018. Comparison of PM2.5 chemical composition and sources at a rural background site in central Europe between 1993/1994/1995 and 2009/2010: effect of legislative regulations and economic transformation on the air quality. Environ. Pollut. 241, 841–851.

Querol, X., Alastuey, A., Rodríguez, S., Plana, F., Ruiz, C.R., Cots, N., Massagué, G., Puig, O., 2001. PM10 and PM2.5 source apportionment in the Barcelona metropolitan area, Catalonia, Spain. Atmos. Environ. 35, 6407–6419.

Rich, D.Q., Zhang, W., Lin, S., Squizzato, S., Thurston, S.W., van Wijngaarden, E., Croft, D., Masiol, M., Hopke, P.K., 2019. Triggering of cardiovascular hospital admissions by source specific fine particle concentrations in urban centers of New York state. Environ. Int. 126, 387–394. https://doi.org/10.1016/j.envint.2019.02.018.

Rodríguez, S., Querol, X., Alastuey, A., Viana, M.-M., Alarcón, M., Mantilla, E., Ruiz, C. R., 2004. Comparative PM10-PM2.5 source contribution study at rural, urban and industrial sites during PM episodes in Eastern Spain. Sci. Total Environ. 328 (1–3), 95–113. https://doi.org/10.1016/s0048-9697(03)00411-x.

Satpathy, P., Boopathy, R., Gogoi, M.M., Babu, S.S., Das, T., 2024. Machine learning techniques to predict atmospheric black carbon in a tropical coastal environment. Remote Sensing Applications-Society and Environment 34. https://doi.org/10.1016/j.rsase.2024.101154.

Tang, M., Zhang, H., Gu, W., Gao, J., Jian, X., Shi, G., Zhu, B., Xie, L., Guo, L., Gao, X., Wang, Z., Zhang, G., Wang, X., 2019. Hygroscopic properties of saline mineral dust from different regions in China: geographical variations, compositional dependence, and atmospheric implications. J. Geophys. Res. Atmos. 124, 10844–10857.

Tong, D.Q., Baklanov, A.A., Barker, B.M., Castillo-Lugo, J.J., Gassó, S., Gaston, C.J., Gill, T.E., Griffin, D.W., Huneeus, N., Kahn, R.A., Kuciauskas, A.P., Ladino, L.A., Li, J., Mayol-Bracero, O.L., McCotter, O.Z., Méndez-Lázaro, P.A., Mudu, P., Nickovic, S., Oyarzún, D., Prospero, J.M., Raga, G.B., Raysoni, A.U., Ren, L., Sarafoglou, N., Sealy, A.M., Sprigg, W.A., Sun, Z., Van Pelt, R.S., Vimić, A.V., 2021. Health and safety effects of airborne soil dust in the americas and beyond. Rev. Geophys. 61.

Toscano, D., 2023. The impact of shipping on air quality in the port cities of the Mediterranean Area: a review. Atmosphere 14 (7), 1180.

Wang, G., Zhang, R., Gómez, M.E., Yang, L., Levy Zamora, M., Hu, M., Lin, Y., Peng, J., Guo, S., Meng, J., Li, J., Cheng, C., Hu, T., Ren, Y., Wang, Y., Gao, J., Cao, J., An, Z., Zhou, W., Li, G., Wang, J.-y., Tian, P., Marrero-Ortiz, W., Secrest, J., Du, Z., Zheng, J., Shang, D., Zeng, L., Shao, M., Wang, W., Huang, Y., Wang, Y., Zhu, Y., Li, Y., Hu, J., Pan, B., Cai, L., Cheng, Y., Ji, Y., Zhang, F., Rosenfeld, D., Liss, P.S., Duce, R.A., Kolb, C.E., Molina, M.J., 2016. Persistent sulfate formation from London fog to Chinese haze. Proc. Natl. Acad. Sci. 113, 13630–13635.

Wang, J., Wu, H., Wei, W., Xu, C., Tan, X., Wen, Y., Lin, A., 2022. Health risk assessment of heavy metal(loid)s in the farmland of megalopolis in China by using APCS-MLR and PMF receptor models: taking huairou district of beijing as an example. Sci. Total Environ. 835. https://doi.org/10.1016/j.scitotenv.2022.155313.

Wang, T., Liu, Y., Deng, Y., Fu, H., Zhang, L., Chen, J., 2018. The influence of temperature on the heterogeneous uptake of SO2 on hematite particles. Sci. Total Environ. 644, 1493–1502.

Wang, T., Liu, Y., Zhou, S., Wang, G., Liu, X., Wang, L., Fu, H., Chen, J., Zhang, L., 2023a. Key factors determining the formation of sulfate aerosols through multiphase chemistry—A kinetic modeling study based on beijing conditions. J. Geophys. Res. Atmos. 128 (20), e2022JD038382.

Wang, T., Xia, Z., Wu, M., Zhang, Q., Sun, S., Yin, J., Zhou, Y., Yang, H., 2017. Pollution characteristics, sources and lung cancer risk of atmospheric polycyclic aromatic hydrocarbons in a new urban district of nanjing, China. J. Environ. Sci. 55, 118–128.

Wang, T., Zhang, L., Zhang, P., Yu, G., Chen, C., Qin, X., Wang, G., Liu, X., Li, R., Zhang, L., Xia, Z., 2023b. Unveiling the pollution and risk of atmospheric (gaseous and particulate) polycyclic aromatic hydrocarbons (PAHs) in a heavily polluted Chinese city: a multi-site observation research. J. Clean. Prod. 428, 139454.

Weber, S., Salameh, D., Albinet, A., Alleman, L.Y., Waked, A., Besombes, J.L., Jacob, V., Guillaud, G., Meshbah, B., Rocq, B., Hulin, A., Dominik-Sègue, M., Chrétien, E., Jaffrezo, J.L., Favez, O., 2019. Comparison of PM10 sources profiles at 15 French sites using a harmonized constrained positive matrix factorization approach. Atmosphere 10, 310.

Wen, W., Cheng, S., Liu, L., Wang, G., Wang, X., 2016. Source apportionment of PM2.5 in Tangshan, China-Hybrid approaches for primary and secondary species apportionment. Front. Environ. Sci. Eng. 10 (5). https://doi.org/10.1007/s11783-016-0839-9.

Willmott, C.J., 1982. Some comments on the evaluation of model performance. Bull. Am. Meteorol. Soc. 63, 1309–1313.

Wu, C., Zhang, S., Wang, G., Lv, S., Li, D., Liu, L., Li, J., Liu, S., Du, W., Meng, J., Qiao, L., Zhou, M., Huang, C., Wang, H., 2020. Efficient heterogeneous formation of ammonium nitrate on the saline mineral particle surface in the atmosphere of East Asia during dust storm periods. Environ. Sci. Technol. 54 (24), 15622–15630.

Xu, B., Xu, H., Zhao, H., Gao, J., Liang, D., Li, Y., Wang, W., Feng, Y., Shi, G., 2023. Source apportionment of fine particulate matter at a megacity in China, using an improved regularization supervised PMF model. Sci. Total Environ. 879. https://doi.org/10.1016/j.scitotenv.2023.163198.

Yang, L., Shami, A., 2020. On hyperparameter optimization of machine learning algorithms: theory and practice. Arxiv 415, 295–316 doi:arXiv:2007.15745.

Zhang, K., Shang, X., Herrmann, H., Meng, F., Mo, Z., Chen, J., Lv, W., 2019. Approaches for identifying $PM_{2.5}$ source types and source areas at a remote background site of south China in spring. Sci. Total Environ. 691, 1320–1327. https://doi.org/10.1016/j.scitotenv.2019.07.178.

Zhao, X., Xia, N., Xu, Y., Huang, X., Li, M., 2021. Mapping population distribution based on XGBoost using multisource data. IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. 14, 11567–11580.

Zhong, W., Lian, X., Gao, C., Chen, X., Tan, H., 2022. PM2.5 concentration prediction based onmRMR-XGBoost model. Paper Presented at the International Conference on Machine Learning and Intelligent Communications.

Zhu, X., Chu, J., Wang, K., Wu, S., Yan, W., Chiam, K., 2021. Prediction of rockhead using a hybrid N-XGBoost machine learning framework. J. Rock Mech. Geotech. Eng. 13 (6), 1231–1245. https://doi.org/10.1016/j.jrmge.2021.06.012.