



Unveiling the optimal regression model for source apportionment of the oxidative potential of PM₁₀

Vy Dinh Ngoc Thuy¹, Jean-Luc Jaffrezo¹, Ian Hough¹, Pamela A. Dominutti¹,
Guillaume Salque Moreton², Grégory Gille³, Florie Francony⁴, Arabelle Patron-Anquez⁵,
Olivier Favez^{6,7}, and Gaëlle Uzu¹

¹Université Grenoble Alpes, CNRS, IRD, INP-G, INRAE, IGE (UMR 5001), 38000 Grenoble, France

²Atmo AuRA, 69500 Bron, France

³Atmo Sud, 13006 Marseille, France

⁴Atmo Nouvelle Aquitaine, 33692 Merignac, France

⁵Atmo Hauts de France, 59044 Lille, France

⁶INERIS, Parc Technologique Alata, BP 2, 60550 Verneuil-en-Halatte, France

⁷Laboratoire central de surveillance de la qualité de l'air (LCSQA), 60550 Verneuil-en-Halatte, France

Correspondence: Gaëlle Uzu (gaelle.uzu@ird.fr)

Received: 6 February 2024 – Discussion started: 19 February 2024

Revised: 26 April 2024 – Accepted: 14 May 2024 – Published: 26 June 2024

Abstract. The capacity of particulate matter (PM) to generate reactive oxygen species (ROS) in vivo leading to oxidative stress is thought to be a main pathway in the health effects of PM inhalation. Exogenous ROS from PM can be assessed by acellular oxidative potential (OP) measurements as a proxy of the induction of oxidative stress in the lungs. Here, we investigate the importance of OP apportionment methods for OP distribution by PM₁₀ sources in different types of environments. PM₁₀ sources derived from receptor models (e.g., EPA positive matrix factorization (EPA PMF)) are coupled with regression models expressing the associations between PM₁₀ sources and PM₁₀ OP measured by ascorbic acid (OP_{AA}) and dithiothreitol assay (OP_{DTT}). These relationships are compared for eight regression techniques: ordinary least squares, weighted least squares, positive least squares, Ridge, Lasso, generalized linear model, random forest, and multilayer perceptron. The models are evaluated on 1 year of PM₁₀ samples and chemical analyses at each of six sites of different typologies in France to assess the possible impact of PM source variability on PM₁₀ OP apportionment. PM₁₀ source-specific OP_{DTT} and OP_{AA} and out-of-sample apportionment accuracy vary substantially by model, highlighting the importance of model selection according to the datasets. Recommendations for the selection of the most accurate model are provided, encompassing considerations such as multicollinearity and homoscedasticity.

1 Introduction

Ambient particulate matter (PM) is one of the key contributors to atmospheric pollution and is responsible for approximately 7 million premature deaths worldwide yearly (WHO, 2021). Many epidemiological studies have linked PM exposure to adverse health effects including (i) acute effects studies using time series and related studies to evaluate the immediate impact of PM exposure (Bell et al., 2004; Dominici, 2004; Pope and Dockery, 2006; Peng et al., 2009) and (ii) cohort studies aiming to evaluate the long-term effects of

chronic PM exposure (Pelucchi et al., 2009; Crouse et al., 2012, 2015; Beelen et al., 2014; Ayres et al., 2008; Yu et al., 2021). These studies mainly focused on the association with PM mass concentrations. However, various research shows that the impacts of PM also depend on other factors such as chemical composition, size distribution, particle morphology, and biological mechanisms (Brook et al., 2010). The capacity of PM to generate reactive oxygen species (ROS) in vivo has recently been introduced as a pivotal indicator of PM biological mechanism with direct implications for oxida-

tive stress and cellular damage (Li et al., 2008; Lodovici and Bigagli, 2011; Mudway et al., 2020; Nelin et al., 2012; Rao et al., 2018; Ayres et al., 2008; Akhtar et al., 2010; Leni et al., 2020). The quantification of the PM capacity to oxidize biological media is called “oxidative potential” (OP) (Bates et al., 2019; Daellenbach et al., 2020; Dominutti et al., 2023). Various acellular assays of OP have been introduced, differentiating ROS generation mechanisms of PM (Dominutti et al., 2023; Calas et al., 2018). Dithiothreitol (DTT) and ascorbic acid (AA) assays are two of the commonly used procedures in the literature (Liu and Ng, 2023).

The relationship between PM chemical components and OP activities may indicate which components are the most prone to generating ROS (Calas et al., 2019; Godri et al., 2011; Yang et al., 2014; Janssen et al., 2014; Crobeddu et al., 2017; Szigeti et al., 2015, 2016; Calas et al., 2018). However, this research pathway struggles with the co-variation between measured and unmeasured PM components (Calas et al., 2018; Weber et al., 2018). An alternative approach is to examine the association between OP and sources of PM obtained using receptor models such as chemical mass balance, positive matrix factorization (PMF), or principal component analysis. PMF is the most popular method for its ability to quantify PM source contributions without extensive prior information on specific sources at the site studied (Belis et al., 2013; Viana et al., 2008; Paatero and Tappert, 1994; Brown et al., 2015; Paatero and Hopke, 2009).

Regression analysis is the most common and effective way to estimate the redox activity of receptor-model-derived PM sources (Bates et al., 2015; Deng et al., 2022; Li et al., 2023; Liu et al., 2018; Shangguan et al., 2022; Verma et al., 2014; J. Wang et al., 2020; Yu et al., 2019). Generally, this is achieved by regression analyses to characterize the relationship between OP activities ($\text{nmol min}^{-1} \text{m}^{-3}$) and PM source contributions ($\mu\text{g m}^{-3}$). This approach provides the OP activities attributed to each microgram of each source ($\text{nmol min}^{-1} \mu\text{g}^{-1}$), denoted as “intrinsic OP”, which can be used to calculate the contribution of each source for each observation day. Numerous regression models can be used for such OP source apportionment (SA), with multiple linear regression fitted by ordinary least squares (OLS) being the most common regression technique (Bates et al., 2015; Deng et al., 2022; Li et al., 2023; Liu et al., 2018; Shangguan et al., 2022; Verma et al., 2014; Y. Wang et al., 2020; Yu et al., 2019). Further, some studies exclude sources with negative intrinsic OP, assuming that negative OP activities are geochemically nonsensical (Bates et al., 2018; Weber et al., 2018). Additionally, weighted least squares can be used to introduce a weighting term, generally using the OP analysis uncertainties to take into account the measurement uncertainties of the OP assays (Borlaza et al., 2021; Daellenbach et al., 2020; Dominutti et al., 2023; Fadel et al., 2023; in’t Veld et al., 2023; Weber et al., 2021). Finally, non-linear models, such as multilayer perceptron, have been used to try to capture possible non-linearities between OP activities and

PM sources (Borlaza et al., 2021; Elangasinghe et al., 2014; D. Wang et al., 2023). However, no study to date has compared the performance and applicability of these various regression models. Each model entails different assumptions which should be carefully considered when selecting a given model.

This study aims to evaluate the variability in PM_{10} OP SA techniques by comparing eight regression techniques: multiple linear regression fitted by OLS, weighted least squares (WLS), positive least squares (PLS), ridge regression (Ridge), least absolute shrinkage and selection operator (Lasso), generalized linear model (GLM), random forest (RF), and multilayer perceptron (MLP). These techniques are applied to apportion PM_{10} OP_{AA} and PM_{10} OP_{DTT} to PM_{10} sources at six sites in France. The PM_{10} SA outputs have been published by Weber et al. (2021), using a harmonized PMF methodology based on 1 year of sampling with similar chemical analyses for a large set of chemical tracers. The results of the PM_{10} OP SA models are compared with regard to the estimated intrinsic PM_{10} OP of each source, the out-of-sample accuracy of the apportionment, and the assumptions inherent in each model. The most appropriate model at each site is compared with OLS to quantify the difference between choosing a model based on data characteristics vs. using the most common approach. Finally, this study provides guidelines for selecting the most suitable model in the strategy for OP contribution regarding sources of PM_{10} . This holds particular significance in the context of the implementation of OP monitoring as a novel air quality metric as foreseen in research programs (such as RI-Urbans) and in the process of the revision of European Directive 2008/50/CE.

2 Methodology

2.1 General organization of the study

Figure 1 illustrates the general workflow of this study. Sections 2.2, 2.3, and 2.4 describe the methods used to analyze the temporal evolution of PM_{10} sources and PM_{10} OP, identify collinearity among PM_{10} sources, and examine homoscedasticity in the relationship between PM_{10} OP and PM_{10} sources. Section 2.5 describes the eight regression techniques (OLS, WLS, PLS, Ridge, Lasso, GLM, RF, and MLP), used for PM_{10} OP SA. Each technique is applied to each site separately using PM_{10} OP_v ($\text{nmol min}^{-1} \text{m}^{-3}$) as the dependent variable and PM_{10} sources ($\mu\text{g m}^{-3}$) as independent variables. The coefficient of the regression, called the “intrinsic PM_{10} OP” of the source ($\text{nmol min}^{-1} \mu\text{g}^{-1}$), represents the capacity of each μg of PM_{10} from the given source to generate oxidative stress; the higher the intrinsic PM_{10} OP of a source, the more redox-active. Each model is trained on a randomly selected (without replacement) 80 % subsample of the dataset and validated on the remaining 20 %. This process is repeated 500 times to estimate uncertainty, a method particularly needed for sources with strong

seasonality. For WLS, PLS, Ridge, and Lasso models, PM₁₀ OP analytical errors were used as a weighting, implying that the PM₁₀ OP with the high analysis uncertainties has less influence on the model. These eight regression techniques were applied to find the relationship between PM₁₀ OP and PM₁₀ sources; however, PLS, Ridge, and Lasso were performed twice, with and without weighting, and consequently there are 11 results of regression techniques that will be presented. Section 2.6 describes the statistical validation of the models using root mean square error (RMSE), mean absolute error (MAE), and R-squared (R^2). The geochemical validation is based on the regression coefficient (the intrinsic PM₁₀ OP) of each source. These are calculated separately for the training and testing data and averaged across the 500 sampling iterations.

2.2 Study sites and PM₁₀ sources

Six French sites are selected in this work for their different typologies: Roubaix and Nice (traffic sites within urban areas), Port-de-Bouc (industrial hotspot), Talence (urban background site), Grenoble and Chamonix (urban background sites in an alpine valley). At each site, sampling was conducted over at least 1 year to capture the complete annual evolution of PM₁₀ and its components. These sites and sampling series have been used and described by Weber et al. (2019).

In brief, daily filter samples were collected on pre-heated Pallflex quartz fiber filters every third day through high-volume sampling (DA80, Digitel). These filters were analyzed to determine the PM chemical species and OP activities. Further details regarding the chemical species and PM₁₀ OP analysis methodology can be found in Weber et al. (2019, 2021). Briefly, elemental carbon (EC) and organic carbon (OC) were analyzed using the EUSAAR2 thermo-optical protocol with a Sunset Laboratory analyzer. Major ionic components (Cl^- , NO_3^- , SO_4^{2-} , NH_4^+ , Na^+ , K^+ , Mg^{2+} , Ca^{2+}) and methanesulfonic acid (MSA) were measured by ion chromatography (IC). Anhydro sugars and saccharides (including levoglucosan, mannosan, arabinol, sorbitol, and mannitol) were analyzed by high-performance liquid chromatography with pulsed amperometric detection (HPLC-PAD). Major and trace elements (Al, Ca, Fe, K, As, Ba, Cd, Co, Cu, La, Mn, Mo, Ni, Pb, Rb, Sb, Sr, V, and Zn) were determined by inductively coupled plasma atomic emission spectroscopy or mass spectrometry (ICP-AES or ICP-MS). Furthermore, colocated PM₁₀ measurements were conducted automatically at each site using the Tapered Element Oscillating Microbalance equipped with a Filter Dynamics Measurement System (TEOM-FDMS).

We used the PM₁₀ sources identified by Weber et al. (2019), who performed a separate PMF for each site using a harmonized approach for all sites (the same chemical species and measurement methods, the same procedure to estimate uncertainties, and the same constraints on the prelim-

inary solutions). Table 1 provides a data description, including the sampling duration, the number of samples collected, and the PM₁₀ sources identified at each site, while Fig. 2 presents the location of the sites in France together with the respective proportion of each PM₁₀ source at each site.

2.3 OP analysis

PM₁₀ OP assays were performed on PM₁₀ extracted from the filters using simulated lung fluid, as detailed in Calas et al. (2017, 2018). The AA assay involved ascorbic acid, a natural antioxidant in the lungs inhibiting lipid and protein oxidation in the lining fluid, using the method presented by Kelly and Mudway (2003) and further described by Calas et al. (2018). Conversely, the DTT assay used dithiothreitol (DTT) as a chemical surrogate for cellular reducing agents, specifically nicotinamide adenine dinucleotide and nicotinamide adenine dinucleotide phosphate oxidase, thereby replicating in vivo interactions between PM₁₀ and biological oxidants (Cho et al., 2005; Calas et al., 2018). Both assays measured the consumption of AA or DTT during the assay, i.e., the rate of the transfer of electrons from AA or DTT to oxygen. The assays were conducted with 96-well plates of UV-transparent quality (CELLSTAR, Greiner Bio-One), and absorption measurements were acquired using a TECAN spectrophotometer (Infinite M200 Pro) at the wavelengths of 265 nm for the AA assay and 412 nm for the DTT assay (Calas et al., 2017, 2018, 2019). Each sample extraction was subjected to four analyses; the PM₁₀ OP in this study represents the mean and the analysis uncertainty is the standard deviation of these four PM₁₀ OP analyses. After analysis, the PM₁₀ OP activities of each sample were blank-subtracted using laboratory and field blanks, and normalized using the air sampling volumes and the mass concentration. The resulting OP_V represents the PM₁₀ OP due to PM₁₀ per cubic meter of air ($\text{nmol min}^{-1} \text{m}^{-3}$). To simplify the denotation of PM₁₀ OP, OP is used to represent PM₁₀ OP throughout this article.

2.4 Collinearity and heteroscedasticity tests

The result of a regression model strongly depends on the characteristics of the dataset because each model makes assumptions about the data. Two critical assumptions in OLS regression analysis are that (1) there is little collinearity between independent variables (the PM₁₀ sources in this study) and (2) the variance of the regression residuals is constant (called “homoscedasticity”). These assumptions should be tested in different ways.

2.4.1 Collinearity

Collinearity occurs when one or more of the independent variables is close to a linear combination of the other independent variables. When collinearity is present, small

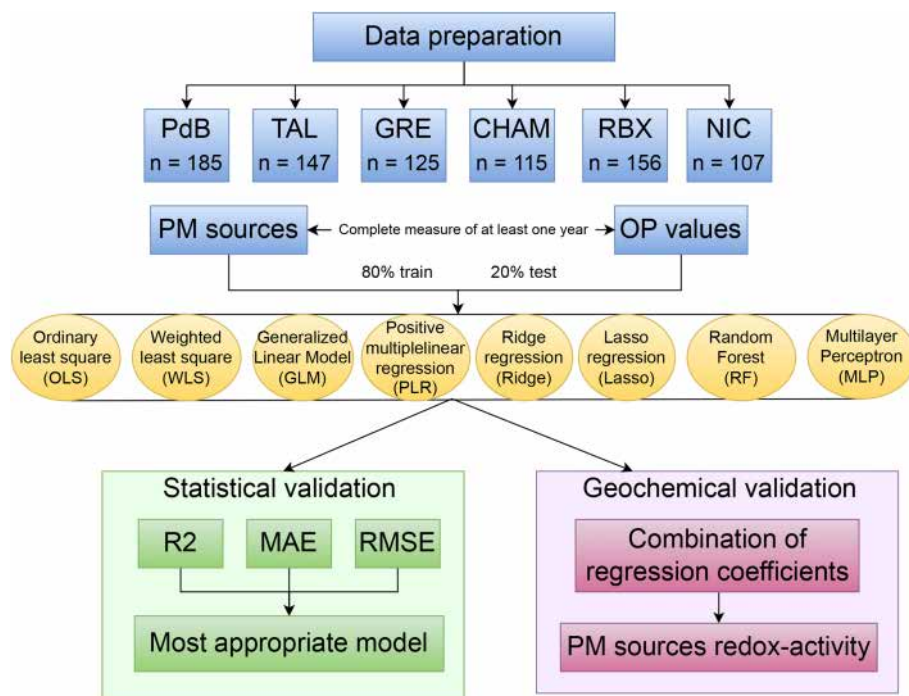


Figure 1. Workflow comparison of the methodology for PM₁₀ OP source apportionment.

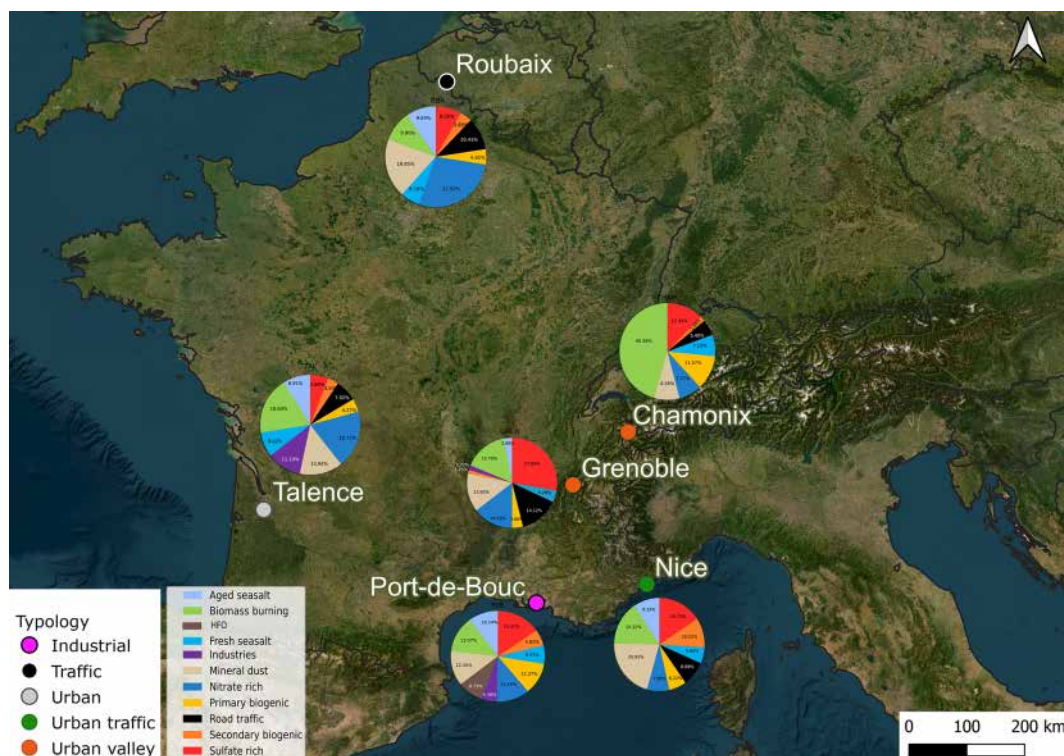


Figure 2. The location of the sites selected for this study. The small colored dots represent the typology of the sites. The pie charts are the PM₁₀ source apportionment for each site with the colors identifying the PM₁₀ sources. Background photography from ESRI satellite imagery.

Table 1. Data description.

	PdB	TAL	GRE-fr	CHAM	RBX	NIC
Name	Port-de-Bouc	Talence	Grenoble	Chamonix	Roubaix	Nice
No. of samples	185	147	125	115	156	107
Sampling dates	June 2014 to June 2016	February 2012 to April 2013	February 2017 to March 2018	November 2013 to October 2014	January 2013 to May 2014	July 2014 to May 2015
No. of sources	10	10	10	8	9	9

changes in the data can cause large changes in estimated coefficients, and the estimated standard errors of the coefficients are large. The variance inflation factor (VIF) is an indicator of the collinearity between independent variables (Crane and Surles, 2002; O'Brien, 2007; Rosenblad, 2011). The VIF of a specific source is calculated as

$$\text{VIF}_i = \frac{1}{1 - R_i^2}, i = 1, \dots, p - 1, \quad (1)$$

where p is the number of PM_{10} sources and R^2 is the coefficient of determination of a multiple linear regression model between the i th source and the other sources. The VIF values of a PM_{10} source present a range between 1 and ∞ . The higher the VIF values, the greater the collinearity between this PM_{10} source and the other sources. A VIF value between 5 and 10 is commonly interpreted as moderate collinearity, while values greater than 10 indicate high collinearity (Crane and Surles, 2002).

2.4.2 Heteroscedasticity

Heteroscedasticity occurs when the variance of regression residuals is not constant but varies for different values of the dependent variable. In this case, the estimated standard errors of the regression coefficients are not reliable. The Goldfeld–Quandt test was developed by Goldfeld and Quandt (1965) to evaluate residual variance in a regression model. To implement the Goldfeld–Quandt test, an OLS regression was performed between OP and PM_{10} sources to identify the residual of OP prediction. Next, the PM_{10} sources and corresponding residual are divided into three segments: the upper segment is the group with higher PM_{10} source concentration, the lower segment is the group with lower PM_{10} source concentration, and the middle segment, constituting 10 % of the moderate PM_{10} concentration, is excluded. A subsequent regression analysis is then conducted on the two remaining subgroups to determine the ratio of residual sums of squares. Finally, an F test is conducted on this ratio to assess whether the variances are the same, with a p value below 0.05 interpreted as evidence of heteroscedasticity.

The variance inflation factor (VIF) and the Goldfeld–Quandt test were performed in Python 3.9, using the statsmodels 0.14.0 package (Seabold and Perktold, 2010).

2.5 Regression models

The fundamental principle of regression models in this study is to use the PM_{10} sources to predict OP activities by identifying the parameters (coefficients and residuals) that minimize an error term (Hastie, 2009). A simple regression model can be represented by Eq. (2), which defines the estimated function of the regression model, and by Eq. (3), which estimates the residuals:

$$\hat{y} = f(X) + e, \quad (2)$$

$$e = y - \hat{y}, \quad (3)$$

where \hat{y} is the estimated OP ($\text{nmol min}^{-1} \text{m}^{-3}$), X is the PM_{10} source contribution ($\mu\text{g m}^{-3}$), y is the observed OP ($\text{nmol min}^{-1} \text{m}^{-3}$), and e denotes the residuals ($\text{nmol min}^{-1} \text{m}^{-3}$). Each model has certain assumptions and a minimization term, as presented in the next section.

2.5.1 Ordinary least squares (OLS)

OLS is a linear regression technique that minimizes the residual sum of squares. This model is based on several assumptions: (1) linearity – the relationship between OP and PM_{10} sources is linear; (2) independence – the PM_{10} sources must be independent, with no collinearity; (3) homoscedasticity – the variance of residuals is constant across all values of PM_{10} sources; and (4) normality – the residuals are normally distributed. In the OLS model, the estimated equation and objective to minimize are defined as follows:

$$\hat{y} = \beta_0 + \sum_{i=1}^p \beta_i \cdot x_i, \quad (4)$$

$$\text{Minimize } \sum_{i=1}^m (y_i - \hat{y}_i)^2 \quad (5)$$

where β_0 denotes the intercept ($\text{nmol min}^{-1} \text{m}^{-3}$), β_i represents the regression coefficient (intrinsic OP,

$\text{nmol min}^{-1} \mu\text{g}^{-1}$) of source i , x_i is the concentration of source i ($\mu\text{g m}^{-3}$), p is the number of PM_{10} sources, and m is the number of observations.

2.5.2 Weighted least square (WLS)

The assumptions and the minimization term in WLS closely align with those in OLS. The only difference is that WLS accounts for heteroscedasticity by introducing a weighting term for individual OP observations whose variance is assumed to be related to the variance of the residuals. The estimation equation in WLS is the same as that of OLS, but the objective to minimize is expressed as

$$\text{Minimize } \sum_{i=1}^m (y_i - \hat{y}_i)^2 \cdot w_i, \\ w_i = \frac{1}{\text{SD}_i^2}, \quad (6)$$

where w_i is the weight assigned to each observation and SD_i is the OP analysis variance of each observation.

2.5.3 Positive least squares (PLS)

The assumptions for PLS primarily include linearity, independence, and normality. PLS can be applied with weighting if there is heteroscedasticity in the data. PLS extends OLS with the constraint that the regression coefficients must be non-negative. The estimation equation and the error term, PLS, are similar to OLS (without weighting) and WLS (applying weighting). To ensure the positivity of coefficients, a specific condition must be met.

$$\beta_i \geq 0, \forall i \text{ in PM sources} \quad (7)$$

2.5.4 Ridge

Shrinkage methods such as Ridge regression try to produce a more interpretable model or reduce error in the presence of collinearity by selecting a subset of the independent variables. Ridge regression is introduced by Hoerl and Kennard (1970), and it incorporates a penalty term that shrinks the coefficients towards 0. Ridge regression minimizes the residual sum of squares plus a penalty term proportional to the sum of squares of the coefficients (L2 regularization) as shown in Eqs. (8) and (9). Consequently, Ridge regression reduces the influence of a PM_{10} source that exhibits minimal impact on OP prediction without excluding it from the model.

$$\text{Minimize } \sum_{i=1}^m (y_i - \hat{y}_i)^2 + \lambda \cdot \sum_{j=1}^p \beta_j^2 \quad (8)$$

$$\text{Minimize } \frac{1}{2m} \sum_{i=1}^m w_i (y_i - \hat{y}_i)^2 + \lambda \cdot \sum_{j=1}^p \beta_j^2 \quad (9)$$

Here, λ is the parameter representing the amount of shrinkage; the larger λ , the greater the shrinkage. The hyperparam-

eter tuning was implemented with different values of λ (5, 1, 0.5, 0.1, 0.01, 0.005, 0.001, 0.0005, and 0.0001). The best λ for every site varied from 0.005 to 0.01, and in this study, 0.01 was selected. Ridge can be applied with weighting to account for heteroscedasticity.

2.5.5 Least absolute shrinkage and selection operator (Lasso)

Lasso (Tibshirani, 1996) is a shrinkage method that uses a penalty term proportional to the sum of the absolute regression coefficients (L1 regularization). This penalty term shrinks the coefficients of a source with a low impact on OP prediction to 0, effectively removing it from the model. This results in a sparse model that may be easier to interpret and may reduce error on out-of-sample data. However, Lasso is more sensitive to outliers than Ridge regression and is less stable when data are collinear. Lasso can be applied with weighting to account for heteroscedasticity.

$$\text{Minimize } \sum_{i=1}^m (y_i - \hat{y}_i)^2 + \lambda \cdot \sum_{j=1}^p |\beta_j| \quad (10)$$

$$\text{Minimize } \frac{1}{2m} \sum_{i=1}^m w_i (y_i - \hat{y}_i)^2 + \lambda \cdot \sum_{j=1}^p |\beta_j| \quad (11)$$

Similar to Ridge, λ is the parameter representing the amount of shrinkage; λ is selected as 0.01 in this study by running the hyperparameter tuning using the same values as for Ridge.

2.5.6 Generalized linear model (GLM)

Generalized linear models, as introduced by McCullagh (1989), provide a framework for regression analysis that can contain non-normal error distributions and capture non-linear relationships between OP activities and PM_{10} sources. GLMs allow for error variance that is a function of the predicted value, hence accounting for heteroscedasticity. Key assumptions underlying GLM include (1) independence, (2) the non-normal distribution of OP, and (3) that the relationship between the PM_{10} sources and the transformed OP (logarithm in this study) is linear. The mathematical expression for GLM can be represented as follows:

$$\log(\hat{y}) = \beta_0 + \sum_{i=1}^p \beta_i \cdot x_i, \quad (12)$$

where β_0 denotes the intercept, β_i represents the regression coefficient of source i , and x_i is the concentration of source i .

2.5.7 Random forest (RF)

RF, an ensemble learning method introduced by Breiman (2001), combines multiple decision trees to make predictions. In the reference implementation, each tree is grown

on a bootstrap sample of the data, and a random subset of the available features is evaluated at each node to choose the best split. The predictions of all trees are averaged to give the RF final prediction. RF is customizable via hyperparameters such as the number of trees, the size of the bootstrap sample, and the number of features to evaluate at each node. For hyperparameter tuning, 5-fold cross-validation was used on the training data. The training dataset was separated into five parts: four parts were used for training and the remaining part was used for validation. This process was repeated five times, and the hyperparameter value producing the lowest mean RMSE across the five parts was selected. The hyperparameter tuning is shown in Sect. S1.1 in the Supplement.

RF does not assume a specific equation to express the relationship between OP activities and PM₁₀ sources, with the result that intrinsic OP could not be computed in this regression model. Nevertheless, RF can estimate the relative importance of each PM₁₀ source in OP prediction. This study estimated the permutation importance of each PM₁₀ source as the mean increase in the mean squared error of predicted OP when the values of the PM₁₀ source were permuted.

2.5.8 Multilayer perception (MLP)

MLP is an artificial neural network that consists of multiple layers of interconnected nodes or neurons organized in a feedforward structure (Akhtar et al., 2018; Chianese et al., 2018; Bourlard and Wellekens, 1989). These layers include an input layer (PM₁₀ sources), one layer or several hidden layers, and an output layer (OP_{AA} or OP_{DTT} activities). In MLP, the neurons in the hidden layers are linked to the previous neurons by the connection weight, where every neuron is independent and has a different weight. The output of each neuron depends on its inputs and an activation function, which, if non-linear, allows the model to capture non-linear relationships. The implementation of MLP includes three steps: (1) forward pass to training model – the input is passed to the model, multiplied with an initial weight, bias is added at every layer, and the output of the model is then calculated; (2) error calculation – after applying step 1, the output of the model and the observed data are used to calculate the error; (3) backward pass – the error is propagated back through the network, and the weights are then adjusted to minimize overall error. These three steps are repeated until the error is minimized.

The choice of hyperparameters to ensure the MLP model robustness is processed by hyperparameter tuning using 5-fold cross-validation, as shown in Sect. S1.2. Thanks to hyperparameter tuning, the two hidden layers and a logistic sigmoid activation function were selected in this study to capture the non-linear relationships between OP activities and PM₁₀ sources.

All regression models were performed using the Python package statsmodels 0.14.0 (Seabold and Perktold, 2010) and scikit-learn 1.3.1 (Pedregosa et al., 2011).

2.5.9 Performance of the models

The performance metrics *R*-square (R^2), mean absolute error (MAE), and root mean square error (RMSE) were used to assess the goodness of fit of the models, as described by Kuhn and Johnson (2013). R^2 quantifies the ability of the model to explain the variance in the data. $R^2 = 1$ indicates a perfect fit. RMSE represents the aggregation of the individual differences between predicted OP and measured OP, while MAE assesses the average magnitude of errors between them. Lower RMSE and MAE values indicate a better fit, with a perfectly fitting model yielding an RMSE or MAE of 0. Equations (13), (14), and (15), respectively, define R^2 , MAE, RMSE. These indicators are computed for the training and testing data of each sampling iteration and averaged across the 500 sampling iterations.

$$R^2 = 1 - \frac{\text{Sum of Squared Residuals}}{\text{Total Sum of Squares}} = 1 - \frac{\sum_{i=0}^m (y_i - \hat{y}_i)^2}{\sum_{i=0}^m (y_i - \bar{y})^2} \quad (13)$$

$$\text{MAE} = \frac{\sum_{i=0}^m |y_i - \hat{y}_i|}{m} \quad (14)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=0}^m (y_i - \hat{y}_i)^2}{m}} \quad (15)$$

3 Result and discussion

Assessments of collinearity and homoscedasticity are addressed in Sect. 3.1. Model performance, including key performance metrics and identification of the optimal model, is detailed in Sect. 3.2. Section 3.3 compares the intrinsic OP estimated by the different models. Section 3.4 compares the intrinsic OP between the combined best-fit and reference models. Lastly, Sect. 3.5 proposes recommendations for selecting an appropriate model.

3.1 Dataset characteristics

The contributions of identified sources ($\mu\text{g m}^{-3}$) and the OP_v activities ($\text{nmol min}^{-1} \text{m}^{-3}$) in each site are presented in Fig. 3, illustrating variations in annual average OP activities and PM₁₀ source contributions by site. Most sites, including traffic and industrial sites, show higher OP_{DTT} activities than OP_{AA}. Conversely, for the alpine valley sites, CHAM presents higher OP_{AA} than OP_{DTT}, while GRE-fr experiences similar levels of OP_{AA} and OP_{DTT}. Additionally, the average OP activities in every site are not proportional to the average PM concentration. For instance, CHAM and NIC

had lower PM₁₀ concentrations but higher OP activities than other sites, while TAL showed high PM₁₀ concentrations but relatively lower OP activities.

The variations observed in the levels of PM₁₀ and OP across six sites can be attributed to distinctions in the identified sources and their respective contributions. These disparities are contingent upon the unique typologies of each site, which are discussed in Weber et al. (2021). Further, we can observe a significant seasonality in the OP activities (Table S1 in the Supplement). The strong seasonality of OP in alpine valley sites has been addressed in previous studies (Borlaza et al., 2021; Dominutti et al., 2023; Weber et al., 2018, 2021), with thermal inversions during winter increasing pollutant concentrations and OP activities compared with summer. Conversely, OP activities in cold and warm periods in other sites are not significantly different.

The PM₁₀ sources and their distribution vary among sites (Fig. 3) because of the difference in typology and local activities. For instance, in the industrial site (PdB), two specific sources are identified: shipping emissions (HFO), with an annual mean contribution of 1.39 $\mu\text{g m}^{-3}$, and industrial sources at 0.86 $\mu\text{g m}^{-3}$. The urban background site TAL also appears to be influenced by industrial sources (2.34 $\mu\text{g m}^{-3}$), which might, however, be partly due to biases induced by the application of the harmonized receptor model protocol (Weber et al., 2019). Note that the application of a site-specific PMF procedure for this site leads to a much lower contribution of this source category but relatively similar contributions of other sources (Favez, 2017). GRE-fr, an urban background site in an alpine valley, presents significant long-range transport sources, with secondary sulfate contributing 3.90 $\mu\text{g m}^{-3}$ followed by biomass burning at 2.21 $\mu\text{g m}^{-3}$. As expected, biomass burning is an abundant source in CHAM, accounting for 7.28 $\mu\text{g m}^{-3}$ of the PM contribution, while the traffic sites RBX and NIC displayed high contributions of traffic sources (at 2.43 and 1.45 $\mu\text{g m}^{-3}$, respectively).

The presence of multicollinearity and homoscedasticity was tested to assess the data characteristics of every site. The only site with evidence of collinearity was NIC, where the VIF of the traffic source was equal to 5.0. For all other sites, VIF values are below 5, indicating limited collinearity among sources. This is expected, as the PMF analysis is constrained to avoid collinearity between sources. VIF values for each site can be found in Table S2.

The presence of heteroscedasticity is commonly found when the dependent variable (or OP in this study) exhibits a large difference between the minimum and maximum values or when the error variance varies proportionally with an independent variable (PM₁₀ sources). The heteroscedasticity was assessed by applying the Goldfeld–Quandt test. Table 2 presents the *p* values of the Goldfeld–Quandt test, indicating homoscedasticity of OP prediction when *p* > 0.05. This test reveals that heteroscedasticity was detected in CHAM, in GRE-fr, in NIC for OP_{AA}, and in CHAM and TAL for OP_{DTT} (Table 2). We observed a large difference between the cold

Table 2. The *p* value of the Goldfeld–Quandt heteroscedasticity test.

	PdB	TAL	GRE-fr	CHAM	RBX	NIC
AA	0.15	0.78	$\ll 0.001$	$\ll 0.001$	0.44	0.002
DTT	0.59	$\ll 0.001$	0.189	$\ll 0.001$	0.56	0.91

and warm periods for both OP_{AA} and OP_{DTT} in CHAM, similar to what was seen for OP_{AA} in GRE-fr (Table S1), which can be the reason for the presence of heteroscedasticity. For NIC and TAL, there is an insignificant difference between the cold and warm periods, which indicates the presence of heteroscedasticity may be because of the relationship between the PM₁₀ sources and error variance. When heteroscedasticity is detected, unweighted regression for OP prediction according to the source may not accurately reflect the uncertainty in the intrinsic OP of each source. The scatterplots representing the relationship between the regression analysis residuals and the fitted values (for observed OP) are available in Figs. S1 and S2 in the Supplement.

3.2 Performance of regression models

The 11 regression models, with or without the weighting for some of them, were tested by comparing their performance metrics between the measured and reconstructed OPs. For each run (*n* = 500 iterations), the *R*², RMSE, and MAE were computed for the testing and training dataset, resulting in 500 values for each performance metric. Figure 4 presents the mean *R*² values of the training datasets as well as the mean and the standard deviation of the testing datasets of the OP_{AA} models across the 500 sampling iterations, and Fig. 5 presents the mean RMSE and MAE. The same result pattern was found for OP_{DTT}, as presented in Tables S3, S4, S5. The WLS, wPLS, wRidge, and wLasso models incorporated weighting, while the OLS, PLS, Ridge, Lasso, GLM, RF, and MLP models were unweighted.

OP predictions across all sites are statistically validated, with testing *R*² values in RBX, NIC, PdB, TAL, CHAM, and GRE-fr observed to be 0.66, 0.76, 0.76, 0.78, 0.87, and 0.90, respectively. The lowest mean test set RMSE values are 0.70, 0.28, 0.21, 0.37, 0.70, and 0.31 nmol min^{−1} m^{−3}, respectively, for the same sites. The lowest mean test set MAE values are 0.49, 0.23, 0.14, 0.25, 0.45, and 0.21 nmol min^{−1} m^{−3}, respectively. Notably, the GLM model exhibits the lowest *R*² values and the highest RMSE for all sites (Tables S3–S5). These results strongly suggest that the relationship between OP_{AA} and PM₁₀ sources is not log-linear.

Differences in MAE, RMSE, and *R*² between the training and testing database for RF and MLP are significant across the sites. Notably, RF displays a large difference in *R*², with a gap of up to 0.6 in RBX (*R*² training: 0.92; *R*² testing: 0.27). Similar gaps were found in RMSE and MAE. RF con-

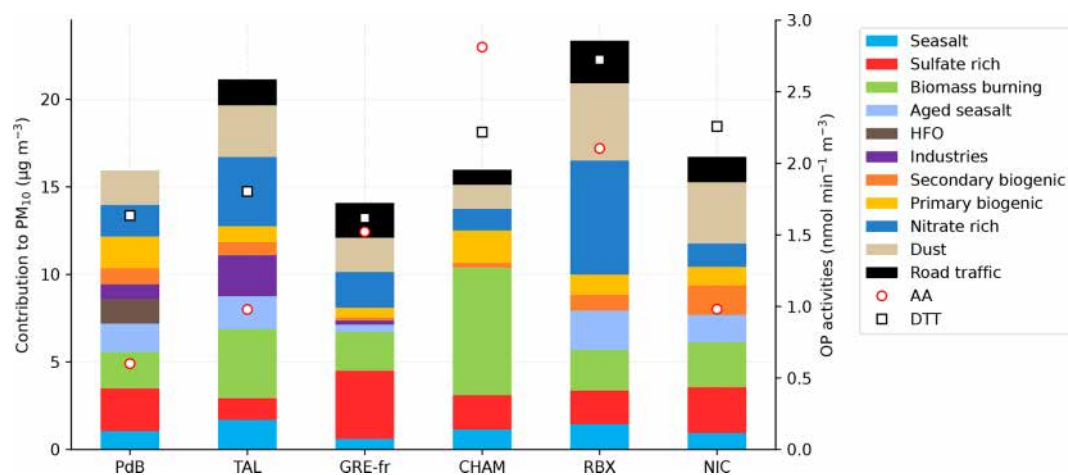


Figure 3. The contribution of sources to PM_{10} and the OP activities in six sites. The left y axis and bar show the contribution of PM sources in $\mu\text{g m}^{-3}$. The right y axis, circles, and squares show the mean OP_v activities in $\text{nmol min}^{-1} \text{m}^{-3}$, with the red circle indicating OP_{AA} and the black square indicating OP_{DTT} .

sistently performed best on the training set, characterized by the highest R^2 and the lowest MAE and RMSE values, but had lower test set R^2 values than the other models (except GLM). Conversely, MLP exhibited training R^2 values comparable to other models but lower test R^2 values. These findings suggest overfitting: the flexible algorithms identify relationships in the training data that do not generalize to the testing data. This observation may be attributed to the limitations of data coverage, possibly failing to fully represent the underlying relationships, leading to poor performance in the testing datasets (Matsuki et al., 2016; Benkendorf and Hawkins, 2020; Stockwell and Peterson, 2002; Wisz et al., 2008; Hernandez et al., 2006; Hawkins, 2004; Raudys and Jain, 1991). Pearce and Ferrier (2000) recommended that the minimum number of samples for robust performance should be over 250 for GLM model, while Raudys and Jain (1991) showed that the minimum number of samples is based on the complexity of the model and the number of predictors. Additionally, Harrell (2016) suggested that the number of predictors (PM sources) should be below the number of samples divided by 15, a threshold not reached in this analysis. For example, in NIC, the minimum number of samples should be 135 for the training set ($9 \text{ PM sources} \times 15$), while in total, we have only 107 samples. Therefore, we can also recommend that, for optimal performance of RF and MLP, the number of samples and PM sources should satisfy these thresholds.

The WLS, OLS, wPLs, wRidge, and wLasso models show more robust performances with fewer differences between the training and testing data. At most sites, there is very little difference between the R^2 , RMSE, and MAE of OLS and Ridge, with or without weighting, and often between PLS and Lasso as well. This consistency is observed even in the collinearity case of NIC, where $\text{VIF} = 5$. The difference between these models is a maximum of 0.06 in R^2 , 0.01 in

MAE, and 0.1 in RMSE, indicating that these models work well for OP prediction. Nevertheless, it is worth noting that every model exhibits different assumptions that have to be respected. The assumption violations may lead to unreliable regression coefficients (intrinsic OP) even though the prediction is good (Williams et al., 2013; Cohen et al., 2002).

The best model for each site was selected based on both data characteristics (collinearity and heteroscedasticity) and testing data performance. For sites with collinearity, Ridge and Lasso were considered the most appropriate. For sites with heteroscedasticity, models with weights were considered the most appropriate. For sites with neither collinearity nor heteroscedasticity, OLS and PLS were considered the most appropriate. Tables 3 and 4 present the best OP_{AA} and OP_{DTT} prediction models for each site. It follows that the best model is not necessarily the same one for both series of OP for a given site. As a rule, the model that exhibits the best performance metrics (the best model by error in Table 3 for OP_{AA} and Table 4 for OP_{DTT}) is suited to be the best model chosen by data characteristics; therefore, choosing a model according to data characteristics help to obtain more reliable OP predictions.

3.3 Effect of the choice of a model on intrinsic OP

It is particularly important to try to define the best way of calculating the more accurate intrinsic OP of PM sources and the contribution of sources to OP, since these values are fundamental inputs in all the works of large-scale modeling of OP with chemical transport models (CTM) (Daellenbach et al., 2020; Vida et al., 2024). Figures 6 and 7 show the variations in intrinsic OP for all the models, focusing on the results of NIC as an example. The evaluation of the five other sites is presented in Figs. S3–S7 for OP_{AA} and Figs. S8–S12 for OP_{DTT} . The differences in equations, error term minimiza-

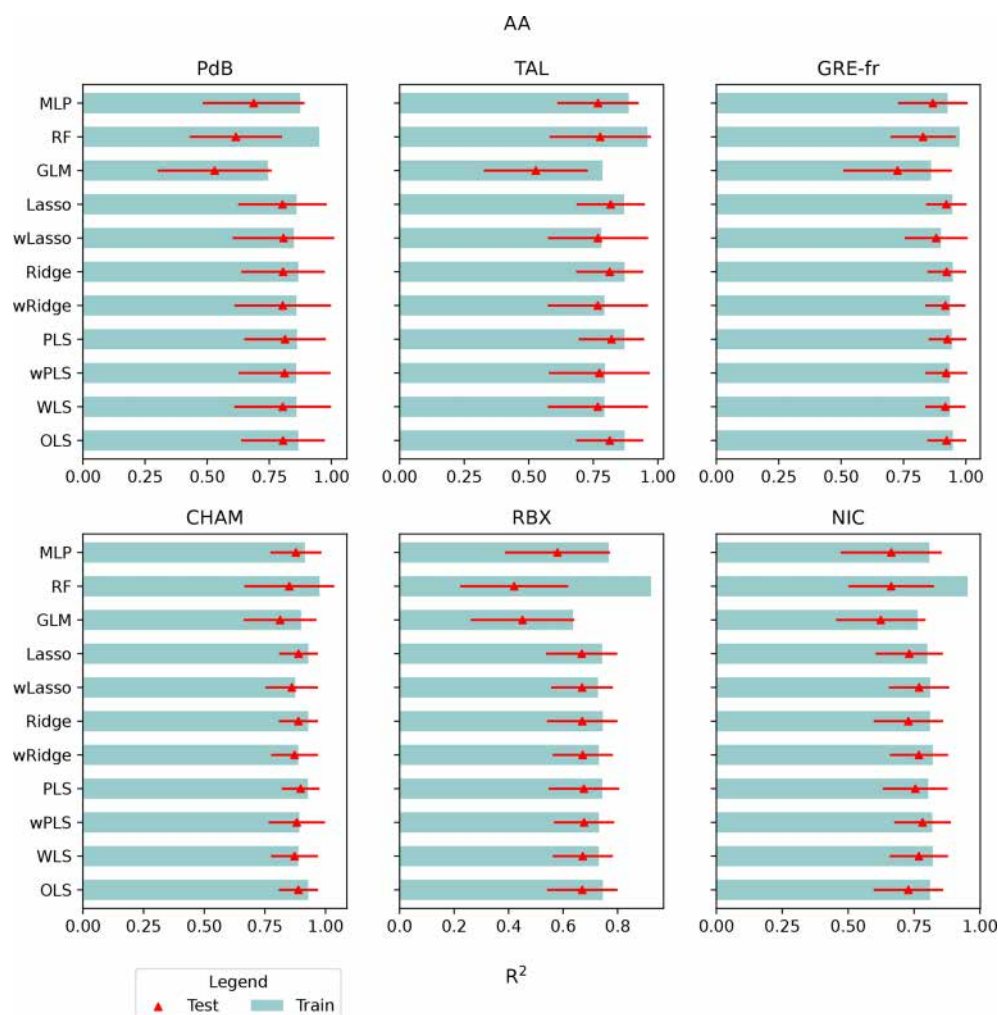


Figure 4. The R^2 values of 11 OP_{AA} models in six sites. The mean R^2 of the training data is shown by the blue bars, the mean R^2 of the testing data is shown by the red triangle, and the standard deviation of the R^2 of the testing data is shown by the red bars. The y axis represents the models, and the x axis denotes the R^2 values.

Table 3. Criteria for selection of the best model for OP_{AA} .

	PdB	TAL	GRE-fr	CHAM	RBX	NIC
Collinearity	No	No	No	No	No	Yes
Heteroscedasticity	No	No	Yes	Yes	No	Yes
Best model by characteristic	OLS/PLS	OLS/PLS	WLS/wPLS	WLS/wPLS	OLS/PLS	wRidge/wLasso
Best model by error	PLS	PLS	wPLS	wPLS	OLS	wRidge

tions, and assumptions can explain the differences in intrinsic OP per μg of source among the eight regression models. While the R^2 , RMSE, and MAE values are similar among models (except for GLM, RF, and MLP), the intrinsic OP values significantly differ between the models with and without weighting and between the linear and non-linear regression

models. The average intrinsic OP of 500 iterations is discussed in this section, since these values are generally used to calculate the contribution of the PM_{10} source to OP in prior studies (Borlaza et al., 2021; Dominutti et al., 2023; Weber et al., 2018). The mean and standard deviation of intrinsic

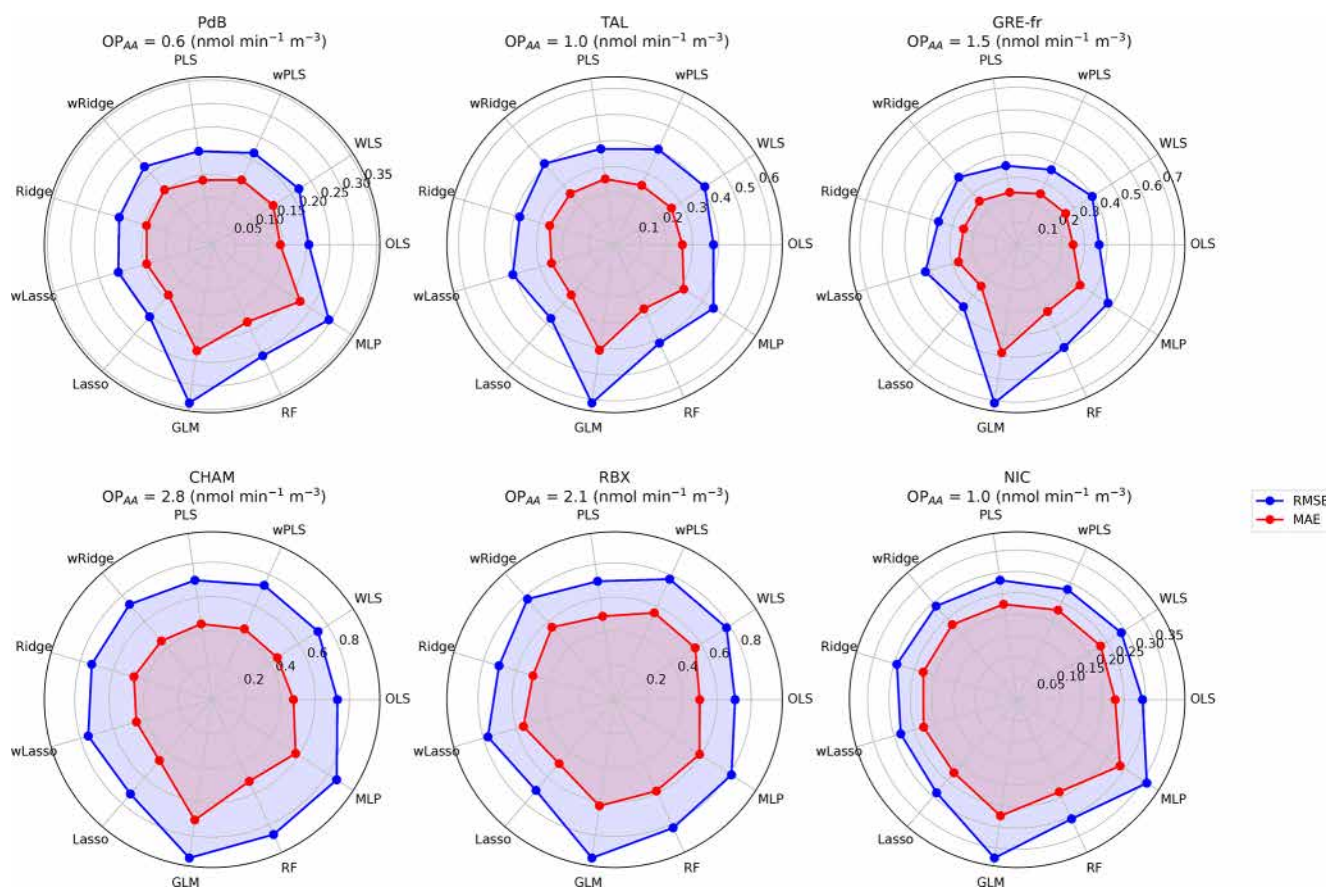


Figure 5. The MAE and RMSE of 11 OP_{AA} models in every site for the testing data. Blue and red lines represent the RMSE and the MAE, respectively. The values in the figure are the mean of the RMSE and MAE of 500 iterations.

Table 4. Criteria for selection of the best model for OP_{DTT} .

	PdB	TAL	GRE-fr	CHAM	RBX	NIC
Collinearity	No	No	No	No	No	Yes
Heteroscedasticity	No	Yes	No	Yes	No	No
Best model by characteristic	OLS/PLS	WLS/wPLS	OLS/PLS	WLS/wPLS	OLS/PLS	Ridge/Lasso
Best model by error	OLS	wPLS	PLS	wPLS	PLS	Ridge

OP_{AA} and OP_{DTT} for the six sites are shown in Tables S6 and S7, respectively.

The intrinsic OP_{AA} of PM_{10} sources at NIC is the same between WLS and wRidge and between the OLS and Ridge, revealing that the moderate collinearity of the road traffic source did not affect the estimated intrinsic OP_{AA} . PLS sets the intrinsic OP_{AA} of some sources to 0, therefore producing slightly different results. Lasso regression sets the intrinsic OP_{AA} of some sources to 0 and shrinks the estimates for all other sources toward 0. GLM produces intrinsic OP_{AA} values that represent a multiplicative change on the log scale,

and thus they are not directly comparable to the other models. However, the direction and importance of the sources are similar to the other models. Whatever the model, road traffic appears as the source with the highest intrinsic OP_{AA} , followed by biomass burning, aged salt, salt, and sulfate-rich sources, in NIC. Traffic and biomass burning sources have been similarly recognized as significant contributors to OP_{AA} in prior studies (Borlaza et al., 2021; Dominutti et al., 2023; Stevanović et al., 2023). The intrinsic OP of the dominant sources is stable, indicating that all these models could give the same information about the intrinsic OP of the

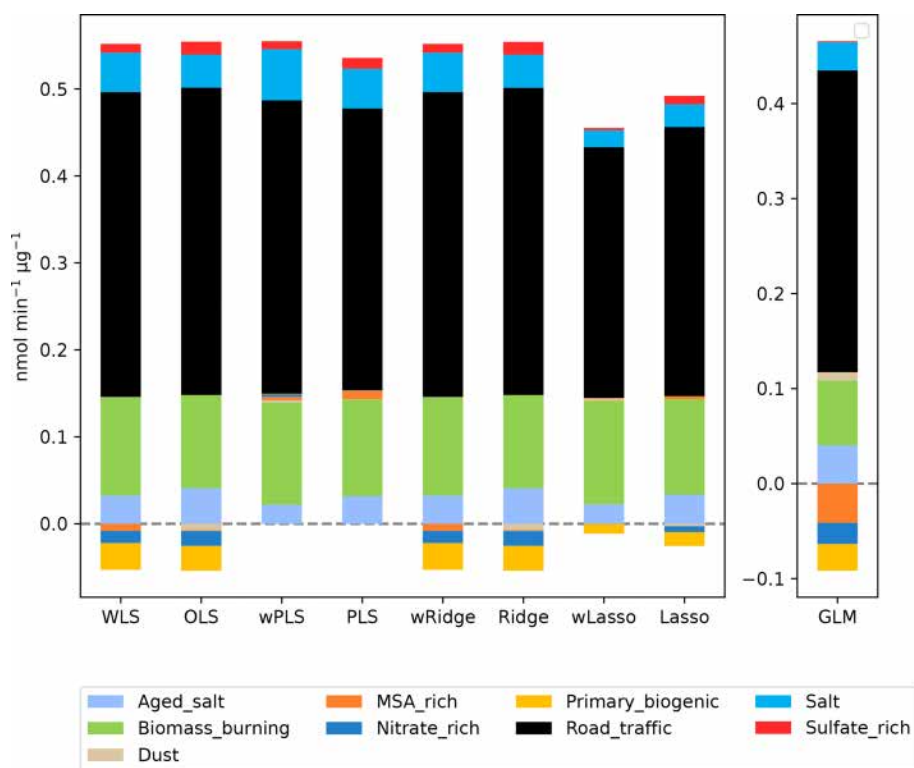


Figure 6. Intrinsic OP_{AA} values of the different PM_{10} sources at Nice obtained with the different models.

main sources. Conversely, the differences are larger between models for the sources with small to very small intrinsic OP (MSA-rich, primary biogenic, nitrate-rich, and dust sources), whose intrinsic OP varies from positive to negative among models.

The OP_{DTT} intrinsic values in NIC (Fig. 7) display minimal variation among the WLS and wPLS. This consistency is linked to the absence of negative intrinsic values. On the other hand, even though there is the presence of moderate collinearity, wRidge still has the same result as WLS and wPLS. In line with the OP_{AA} results, the wLasso and GLM models exhibit distinct responses compared with the other models. The intrinsic OP_{DTT} of all sources varies depending on the presence or absence of weighting. While the WLS models tend to amplify the influence of some sources (aged sea salt, primary biogenic, sea salt, and sulfate-rich sources), the OLS reduces the intrinsic OP_{DTT} of these sources. Conversely, MSA-rich, nitrate, and road traffic sources undergo less influence in WLS but more influence in OLS. Different from OP_{AA} , OP_{DTT} prediction shows more variation among models, highlighting the effect of choosing a model on evaluating the intrinsic OP_{DTT} of PM_{10} sources.

The comparison of intrinsic OP among regression models in NIC demonstrated that OP_{DTT} and OP_{AA} intrinsic values exhibit variation across different models with and without weighting, illustrating that the choice of the model significantly influences the values obtained for intrinsic OP of

PM_{10} sources (a similar pattern is observed for all other sites and shown in Figs. S3–S7 for OP_{AA} and Figs. S8–S12 for OP_{DTT}). Because of the difference in intrinsic OP across models, a comparison between the best-performing and most commonly used models (OLS) is presented in the following section to elucidate the advantage of choosing a model based on data characteristics (Sect. 3.4).

3.4 Comparisons between the best site-specific model and OLS

In this section, the intrinsic OP of the best model is selected for each site as discussed in Sect. 3.2, and the intrinsic values of each source are compared with the ones returned by the OLS model. The OLS model is used as a representative of usual practices that do not consider the database characteristics (Williams et al., 2013). The average intrinsic OP value of each PM_{10} source is calculated from all 500 bootstrapping iterations for all sites where that particular source is identified. Intrinsic OP values obtained in this way from the best model (the best model presented in Table 3 for OP_{AA} and Table 4 for OP_{DTT}) encompassing all six sites are called “the intrinsic OP of the best model”, and the intrinsic OP values derived from the OLS from all six sites are called the “intrinsic OP of the reference model”.

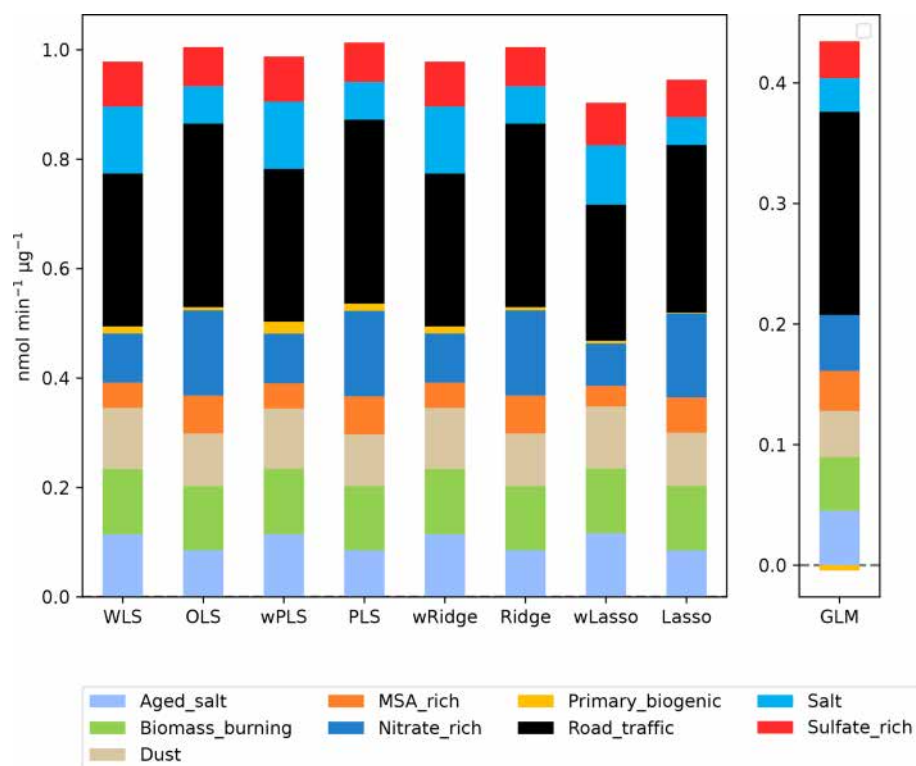


Figure 7. Variations in the intrinsic OP_{DTT} of the different PM_{10} sources at Nice obtained with the different models.

A meaningful comparison of the two series of intrinsic values requires two conditions. First, intrinsic OP should be consistent across all sites. While recognizing that intrinsic OP values depend on diverse factors, we assumed the sites share fairly uniform PM_{10} chemical source profiles in France. This is demonstrated by evaluating the Pearson distance and standardized identity distance similarity indicators of the source chemical profiles (Belis et al., 2015; Weber et al., 2019), and Fig. S13 indicates consistent profiles of sources for the six sites. Consequently, we could expect to observe minimal divergence in intrinsic OP values among these sites. Second, we postulate that negative intrinsic OP values are possible since previous studies have reported that total PM_{10} intrinsic OP can be modulated due to the synergistic/antagonistic effects involving, for example, soluble copper, quinones, and bacteria (Borlaza et al., 2021; Pietrogrande et al., 2022; Samake et al., 2017; S. Wang et al., 2018; Xiong et al., 2017). Samake et al. (2017) demonstrated that the presence of bacterial cells in aerosol decreases the redox activity of Cu and 1,4-naphthoquinone, with a maximum decrease of 60 % compared with the oxidative reactivity considered individually. Pietrogrande et al. (2022) indicated that the mixture of Cu, Fe, 9,10-phenanthrene quinone, and 1,2-naphthoquinone reduces the rate consumption of AA and DTT by up to 50 % depending on the quantity of each chemical. Wang et al. (2018) reported that the mixing of Cu, naphthalene secondary organic aerosol (SOA), and phenanthrene

SOA only achieved half of the DTT rate consumption compared with the separately considered consumption. Xiong et al. (2017) showed the presence of antagonists in the interaction of Fe and quinones; nevertheless, it was much lower than those in the other studies (under 10 %). These references reported that the antagonistic effects of a mixture can significantly reduce the consumption rate of OP_{DTT} and OP_{AA} , and this impact varies widely from 10 % to 60 % depending on the type of chemical species and the quantity of each species in the mixture. Consequently, we consider here that the intrinsic OP value of an individual site for a given source could be negative only within a range of at most 60 % of the mean combined intrinsic OP value of this source across all sites. Negative intrinsic OP exceeding this criterion may result from the mathematical construction of the model. The comparison between the intrinsic OP_{AA} of the best model and the reference model is presented in Sect. 3.4.1 and that of OP_{DTT} is shown in Sect. 3.4.2.

3.4.1 OP_{AA} activities

The results of the comparison of OP_{AA} intrinsic values (Fig. 8 and Table S8) show that the anthropogenic sources have the highest intrinsic OP values in both the best model and the reference model. Among these sources, road traffic appears as the most prominent potent fraction, followed by biomass burning, HFO, and industrial sources. These results

are aligned with prior research (Calas et al., 2019; Daellenbach et al., 2020; Dominutti et al., 2023; Fadel et al., 2023; Fang et al., 2016; in't Veld et al., 2023; Weber et al., 2018; Zhang et al., 2020) which has highlighted the sensitivity of OP_{AA} to concentrations of metals, black carbon, and organic carbon. The differences between the best model and the reference model were insignificant for these sources, demonstrating that the best model and the reference model consistently captured similar patterns for the most critical sources of OP activities.

However, the interquartile ranges (IQR) of the intrinsic OP values are consistently narrower for the best model across all sources, accounting for less divergence in intrinsic OP values across sites. Moreover, the median intrinsic OP values obtained from the best model closely approximated the mean values, indicating the absence of extreme intrinsic OP values. For instance, in the case of road traffic, the mean and median values were 0.24 and $0.23 \text{ nmol min}^{-1} \mu\text{g}^{-1}$, respectively. Conversely, the reference model exhibited a large difference between the mean and median values, implying lower consistency across sites and sampling iterations. The same result was observed in biomass burning source, in which the median and mean intrinsic OP in the best model had fewer discrepancies. Further, the biomass burning intrinsic OP in GRE-fr of the best model is more consistent with those in other sites (best: $0.30 \text{ nmol min}^{-1} \mu\text{g}^{-1}$; reference: $0.35 \text{ nmol min}^{-1} \mu\text{g}^{-1}$).

When considering sources with low intrinsic OP, the variability can be larger between the two methods. As an example, for the sulfate-rich sources, the median intrinsic OP values were positive ($0.002 \text{ nmol min}^{-1} \mu\text{g}^{-1}$), while the mean intrinsic OP values were negative ($-0.008 \text{ nmol min}^{-1} \mu\text{g}^{-1}$). The mean intrinsic OP in the best model exhibited fewer negative values in individual sites than in the reference model (for aged salt, salt, primary biogenic, MSA-rich, sulfate-rich, and nitrate-rich sources). In addition, the best model showed the least disparate intrinsic OP among individual sites, for instance, the aged salt sources in GRE-fr and the primary biogenic and salt sources in CHAM. Furthermore, the best model displayed an intrinsic OP meaningful in terms of geochemical validation, which was shown in the salt, primary biogenic, and sulfate-rich sources. For instance, in the reference model, the average intrinsic OP of the primary biogenic source in NIC ($-0.03 \text{ nmol min}^{-1} \mu\text{g}^{-1}$), the intrinsic OP of salt in GRE-fr ($-0.07 \text{ nmol min}^{-1} \mu\text{g}^{-1}$), and the sulfate-rich source in CHAM ($-0.05 \text{ nmol min}^{-1} \mu\text{g}^{-1}$) represented a 100 % reduction compared with the mean intrinsic OP of all sites. Moreover, negative intrinsic OP was observed in NIC (primary biogenic), and some extreme values were observed in GRE-fr (aged salt, salt) and CHAM (salt, primary biogenic, MSA-rich; where heteroscedasticity was presented) in the OLS model, which underscores that the model assumptions on data characteristics proving false could impact the accu-

racy of OP prediction. Consequently, these results highlight the advantage of considering the data in the model selection.

The detailed comparison of intrinsic OP_{AA} between the best model and the reference model is categorized into four groups and discussed in detail in Sect. S9. These groups include (1) anthropogenic sources without nitrate and sulfate (road traffic, biomass burning, HFO, and industrial sources); (2) natural inorganic sources (aged sea salt, sea salt, dust); (3) biogenic sources (primary biogenic and MSA-rich sources); and (4) nitrate and sulfate-rich sources.

3.4.2 OP_{DTT} activities

Similar to OP_{AA} , for OP_{DTT} the IQR of the best model is narrower for most of the sources than the IQR of the reference model (OLS). Except for road traffic, industrial, and MSA-rich sources, the IQR is slightly higher in the best model (Fig. 9 and Table S9). In the two models, the mean intrinsic OP is essentially unchanged, where traffic is the most critical source (0.27 ± 0.10) followed by HFO (0.18 ± 0.01), biomass burning (0.12 ± 0.03), dust (0.12 ± 0.07), primary biogenic (best: 0.10 ± 0.06 ; reference: 0.12 ± 0.08) and MSA-rich sources (best: 0.11 ± 0.09 ; reference: 0.09 ± 0.09). The minimum difference between the two models in the dominant sources again confirms the conclusion in the OP_{AA} comparison, demonstrating the similar pattern of the best model and the reference model in the most crucial sources of OP. For both the best and the reference models, OP_{DTT} activities showed sensitivity to more sources compared with OP_{AA} , as discussed in previous studies (Borlaza et al., 2021; Calas et al., 2019; Dominutti et al., 2023; Fadel et al., 2023).

While the best model and reference model give the same mean intrinsic OP_{DTT} for all sites, the mean OP_{DTT} at each individual site can vary substantially between the two models. The best model exhibited the positive intrinsic OP for all sources, while the reference model displayed negative intrinsic OP in RBX (MSA-rich and sulfate-rich sources). Especially in the case of sulfate-rich sources in RBX, the negative intrinsic OP in the reference model passed the threshold of the negative value, which presented a 110 % reduction compared to the mean intrinsic OP of all sites. This is also found in the OP_{AA} comparison, which confirmed that the best model generates a geochemical meaningful intrinsic OP. In addition, the best model exhibited consistent intrinsic OP across sites, especially for the dust, salt, primary biogenic, and sulfate-rich sources in TAL (heteroscedasticity is presented in this site), where intrinsic OP in TAL in the best model is more similar to the other sites. For instance, the reference model showed that the intrinsic OP in TAL is $0.20 \text{ nmol min}^{-1} \mu\text{g}^{-1}$, far from the mean of all sites ($0.07 \text{ nmol min}^{-1} \mu\text{g}^{-1}$). We observed the same for the intrinsic OP of the nitrate-rich source in CHAM (where the heteroscedasticity is detected), which displayed a less dissimilar OP in CHAM compared with the other site in the best

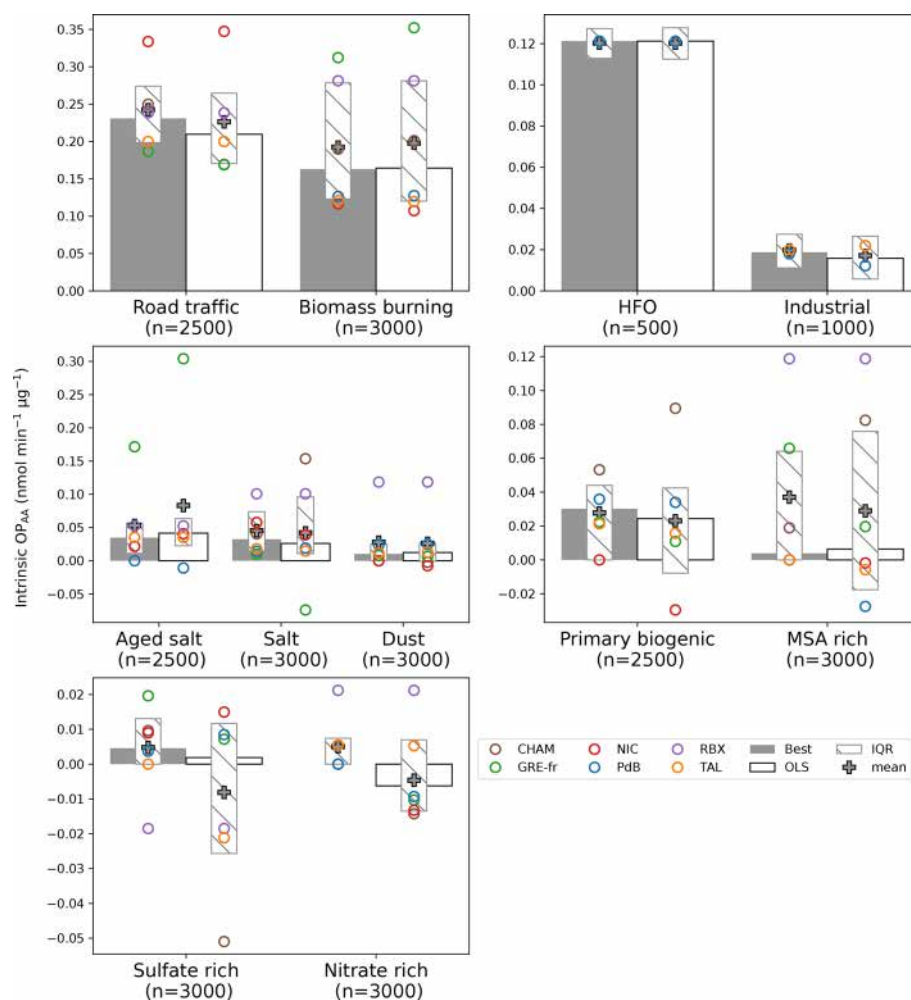


Figure 8. Intrinsic OP_{AA} estimated by the best model and the reference method in the six sites. The y axis represents the intrinsic OP values in $\text{nmol min}^{-1} \mu\text{g}^{-1}$, and the x axis represents the sources. The gray bars are the median intrinsic OP values of the best models in the six sites ($n = 500$ bootstrapping \times number of sites where the given source is detected) for each source. The white bars are the same median intrinsic OP values for the reference (OLS) model. The gray plus symbol represents the mean of intrinsic OP values. The hatched bars are the interquartile ranges of the intrinsic OP values. The dots represent the mean intrinsic OP of all sites, including gray – Chamonix, green – Grenoble, red – Nice, blue – Port-de-Bouc, purple – Roubaix, and orange – Talence.

model. This again validates the conclusion in the OP_{AA} comparison, demonstrating that respecting the model assumption is essential for obtaining a robust OP SA result.

The comparison of intrinsic OP between the best model and the reference model highlights the importance of considering the database characteristics when selecting a model for OP SA. For all the datasets studied here, using the best model for each site delivered more robust results, with reduced uncertainty and reduced differences in intrinsic OP across sites, and it provided a more geochemically meaningful intrinsic OP. The recommendation for selecting a model based on the characteristics of the database is presented in Sect. 3.5.

3.5 Guidelines for the selection of a regression model for OP SA

Our results have highlighted the benefits of choosing a model that matches the characteristics of the data to improve the robustness of the OP SA method. For this reason, this section develops a workflow to help make model selection decisions. Before selecting a regression for OP SA, the first question is whether the PM_{10} sources are collinear and the second is whether the residual variance of the regression between OP and PM_{10} mass is constant. These two questions represent the characteristics of PM_{10} sources and OP activities, which vary according to the study site.

For data exhibiting collinearity between sources and generating a residual variance that varies according to the value of the PM_{10} sources, weighted regularization regression can

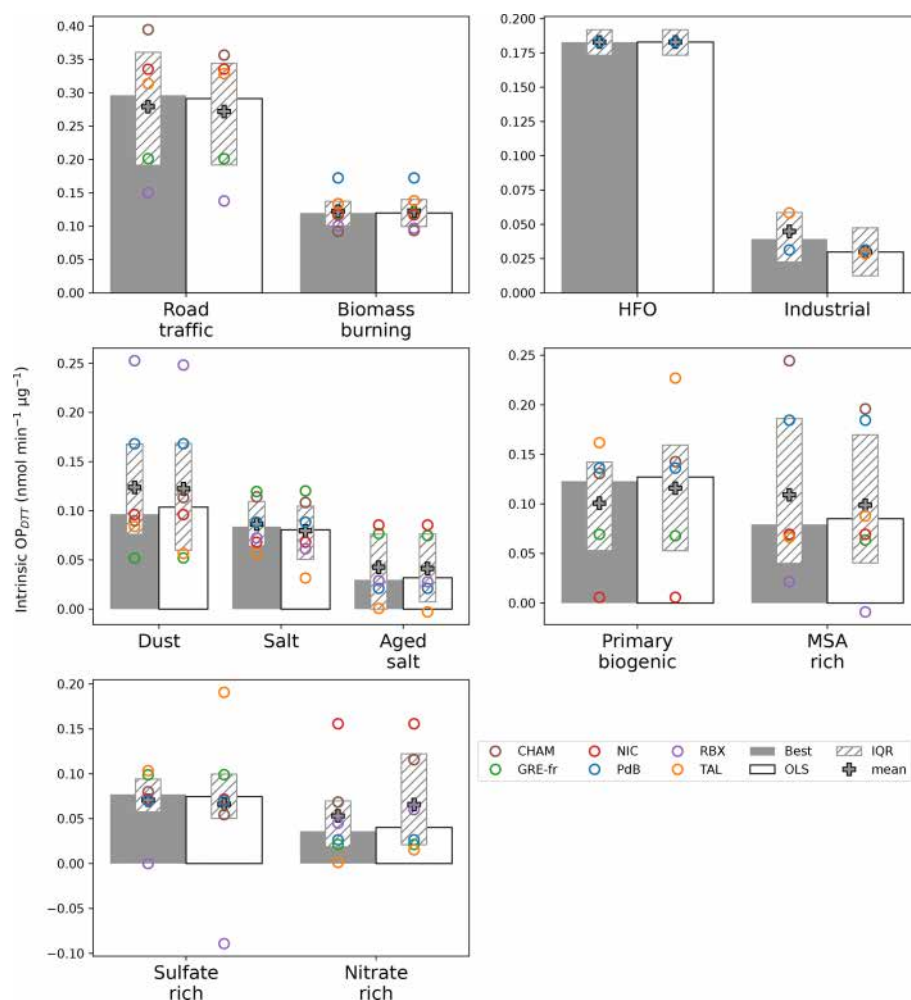


Figure 9. Intrinsic OP_{DTT} estimated by the best model and the reference model in the six sites. The y axis represents the intrinsic OP values in $\text{nmol min}^{-1} \mu\text{g}^{-1}$, and the x axis represents the sources. The gray bars are the median intrinsic OP values of the best models in the six sites ($n = 500$ bootstrapping \times number of sites where the given source is detected) for each source. The white bars are the same median intrinsic OP values for the reference (OLS) model. The gray plus symbol represents the mean of intrinsic OP values. The hatched bars are the interquartile ranges of the intrinsic OP values. The dots represent the mean intrinsic OP of all sites, including gray – Chamonix, green – Grenoble, red – Nice, blue – Port-de-Bouc, purple – Roubaix, and orange – Talence.

help to reduce collinearity and to match the model assumption about the residual. On the other hand, unweighted Ridge and Lasso are introduced for data showing collinearity and homoscedasticity. Additionally, data with no collinearity are suitable for OLS and unweighted PLS in the case of homoscedasticity, while WLS and weighted PLS are used for data with heteroscedasticity.

If the number of predictors (PM_{10} sources) is below the number of samples divided by 15, RF and MLP can also be employed to capture possible non-linear relationships between the OP and PM_{10} sources. However, cross-validation must be used to ensure that there is no overfitting. In addition, these models do not estimate intrinsic OP ($\text{nmol min}^{-1} \mu\text{g}^{-1}$) but only the importance of each PM_{10} source to the OP prediction. This is a major drawback, since

the intrinsic OP of sources is a prerequisite for the modeling effort of OP with CTM. However, RF and MLP could be useful for OP prediction in the case of larger datasets generated by online instruments.

For each data characteristic there is more than one model that is suitable. Out-of-sample performance metrics should be employed to identify the most accurate of these models.

Finally, these techniques of OP apportionment could not be performed well with uncertain PMF-derived sources. The PMF results sometimes do not adequately represent PM mass concentration for several reasons, such as the lack of a trace species to identify a source, an insufficient sample size, the source contribution being too small to be identified (under 1 %), or collinearity issues. Important information could be missed because of these problems in PMF implementation,

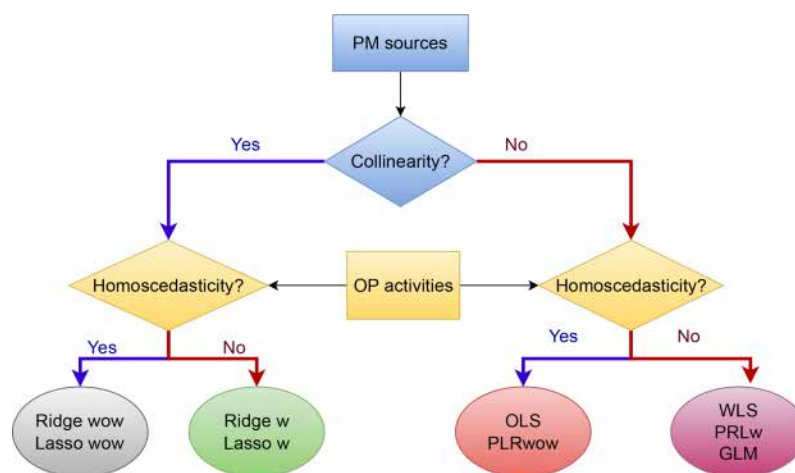


Figure 10. Workflow in model selection considering the data characteristics.

which is explained by the model's low accuracy. Our study did not encounter this problem, since the PMF is harmonized and performed according to European recommendations which could perform the regression technique well and make it possible to obtain a very satisfactory successive OP modeled in comparison with observations after regression techniques (R^2 from 0.7 to 0.9). However, this problem could potentially happen, and for these cases, we could recommend either subtracting the total source contribution from the PM mass concentration to get the part that PMF cannot simulate. The information in this part may contain vital sources. Alternatively, it is possible to re-execute the PMF to validate the result and ensure the robustness of the chemical profile and the contribution of sources.

Limitations and perspectives of the study:

- This study compares eight regression models but is not exhaustive; further research could add more regression techniques to evaluate result variations across models. The potential techniques that could be applied for OP SA are gradient boosting techniques for resolving regression models or supervised machine learning techniques which enable the investigation of linear and non-linear regression relationships. However, the consistently strong performance of ordinary linear regression across six locations in France suggests that there may be little to gain from applying more complex models in areas with similar PM_{10} sources.
- PMF coupled with a regression model remains a popular approach for OP SA. Notably, the uncertainties in PMF are typically addressed in chemical profiles but not in contributions. Incorporating uncertainty from variations in contributions into models could enhance their robustness compared with relying only on absolute PMF results.

- Observations ranged between 100 and 200 samples at each site, which may be insufficient to obtain a fair performance of GLM, decision tree, and neural network models, even though this number of samples is sufficient to address SA through the PMF model for offline analyses. Therefore, this study outlines well the limitations of GLM, RF, and MLP for offline datasets. Future investigations should be performed in an extended dataset, such as long-term or real-time measurement data, to investigate the performance of machine learning algorithms.
- This study only focused on the two most popular OP assays of PM_{10} (OP_{DTT} and OP_{AA}). However, there are various OP assays available, such as OP_{DCFH} , OP_{OH} , OP_{FOX} , OP_{GSH} and OP_{ESR} , and different sizes of PM (PM_1 , $PM_{2.5}$, PM_5). Further research should include more OP assays, which can be helpful in evaluating the performance of various regression models for different OP and different PM sizes.
- This study used the analytical uncertainty as the weighting for the weighted model. However, the weighting can be selected based on different ways, as reported by Montgomery et al. (2012): (1) prior information from the theoretical model, (2) using the residual extracted from the OLS model, (3) selecting the weighting based on the uncertainty of the instrument if the dependent variable is measured by a different method, and (4) selecting the weighting based on the error of these observations if the dependent variable is the average of different observations.

4 Conclusion

The results of the OP SA marked an important milestone as they were revealed for the first time through the use of eight

regression models, including OLS, WLS, PLS, GLM, Ridge, Lasso, RF, and MLP. This in-depth analysis was carried out on a complete set of data collected from six sites with different characteristics. The approach of selecting a suitable model for each site based on specific data characteristics resulted in a more consistent intrinsic OP across sites, in stark contrast to the variation observed when using the basic OLS model. The revelations of the study have provided concrete recommendations for the judicious selection of an appropriate regression model based on the unique characteristics of the dataset. These guidelines should help to improve the accuracy of OP assessments and contribute to the refinement of air quality assessment methods. In addition, the implications of this research extend to the implementation of OP monitoring as a new measure of air quality, particularly on European supersites. As this initiative aligns with the ongoing revision process of European Directive 2008/50/CE, the findings of the study assume a pivotal role in shaping the methodologies underpinning air quality assessments on a broader regulatory level.

Code availability. The code is available at <https://doi.org/10.5281/zenodo.11071884> (Ngoc Thuy, 2024).

Data availability. The datasets could be made available upon request by contacting the corresponding author.

Supplement. The supplement related to this article is available online at: <https://doi.org/10.5194/acp-24-7261-2024-supplement>.

Author contributions. VDNT performed the data analysis for the OP source apportionment setup. GU and JLJ carried out the mentoring, supervision, and validation of the methodology and results. IH, PAD, and VDNT worked on the result visualization. OF, JLJ, and GU acquired funding for the original PM sampling and analysis. VDNT wrote the original draft. All authors reviewed and edited the manuscript.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

Acknowledgements. The authors would like to express their sincere gratitude to many people of the Air-O-Sol analytical platform at IGE (including Sophie Darfeuil, Rhabira Elazzouzi, and Takoua Madhbi), to Robin Aujay (Ineris) for sample management at TAL and RBX, to Laurent Alleman (IMT Nord-Europe) and Nicolas Bonnaire (LSCE) for part of the chemical analyses for some sites, and to all the personnel in the AASQA in charge of the sites for their contribution in conducting the dedicated sample collection. The authors would like to thank Samuel Weber for running the PMF model in his previous professional post.

Financial support. The PhD grant of Vy Dinh Ngoc Thuy was funded by grant nos. PR-PRE-2021, UGA-UGA 2022-16 FUGA-Fondation Air Liquide, and ANR ABS (grant no. ANR-21-CE01-0021-01). Analytical work on OP was funded through ANR GET OP STAND (grant no. ANR-19-CE34-0002), MOBILAIR, and ACME IDEX projects at UGA (grant no. ANR-15-IDEX-02). The sampling and chemical analyses performed at the TAL, GRE, RBX, PdB, and NIC sites have been partly funded by the French Ministry of Environment in the frame of the CARA program. The present work was also supported by the European Union Horizon 2020 research and innovation program under grant no. 101036245 (RI-URBANS) for the postdoctoral salary of Pamela Dominutti.

Review statement. This paper was edited by Arthur Chan and reviewed by four anonymous referees.

References

- Akhtar, McWhinney, R. D., Rastogi, N., Abbatt, J. P. D., Evans, G. J., and Scott, J. A.: Cytotoxic and proinflammatory effects of ambient and source-related particulate matter (PM) in relation to the production of reactive oxygen species (ROS) and cytokine adsorption by particles, *Inhal. Toxicol.*, 22, 37–47, <https://doi.org/10.3109/08958378.2010.518377>, 2010.
- Akhtar, A., Islamia, J. M., Masood, S., Islamia, J. M., Masood, A., and Islamia, J. M.: Prediction and Analysis of Pollution Levels in Delhi Using Multilayer Perceptron, *Adv. Intell. Syst.*, 542, 563–572, <https://doi.org/10.1007/978-981-10-3223-3>, 2018.
- Ayres, J. G., Borm, P., Cassee, F. R., Castranova, V., Donaldson, K., Ghio, A., Harrison, R. M., Hider, R., Kelly, F., Kooter, I. M., Marano, F., Maynard, R. L., Mudway, I., Nel, A., Sioutas, C., Smith, S., Baeza-Squiban, A., Cho, A., Duggan, S., and Froines, J.: Evaluating the toxicity of airborne particulate matter and nanoparticles by measuring oxidative stress potential – A workshop report and consensus statement, *Inhal. Toxicol.*, 20, 75–99, <https://doi.org/10.1080/08958370701665517>, 2008.
- Bates, J. T., Weber, R. J., Abrams, J., Verma, V., Fang, T., Klein, M., Strickland, M. J., Sarnat, S. E., Chang, H. H., Mulholland, J. A., Tolbert, P. E., and Russell, A. G.: Reactive Oxygen Species Generation Linked to Sources of Atmospheric Particulate Matter and Cardiorespiratory Effects, *Environ. Sci. Technol.*, 49, 13605–13612, <https://doi.org/10.1021/acs.est.5b02967>, 2015.
- Bates, J. T., Weber, R. J., Verma, V., Fang, T., Ivey, C., Liu, C., Sarnat, S. E., Chang, H. H., Mulholland, J. A., and Russell, A.: Source impact modeling of spa-

- tiotemporal trends in PM_{2.5} oxidative potential across the eastern United States, *Atmos. Environ.*, 193, 158–167, <https://doi.org/10.1016/j.atmosenv.2018.08.055>, 2018.
- Bates, J. T., Fang, T., Verma, V., Zeng, L., Weber, R. J., Tolbert, P. E., Abrams, J. Y., Sarnat, S. E., Klein, M., Mulholland, J. A., and Russell, A. G.: Review of Acellular Assays of Ambient Particulate Matter Oxidative Potential: Methods and Relationships with Composition, Sources, and Health Effects, *Environ. Sci. Technol.*, 53, 4003–4019, <https://doi.org/10.1021/acs.est.8b03430>, 2019.
- Beelen, R., Stafoggia, M., Raaschou-Nielsen, O., Andersen, Z. J., Xun, W. W., Katsouyanni, K., Dimakopoulou, K., Brunekreef, B., Weinmayr, G., Hoffmann, B., Wolf, K., Samoli, E., Houthuijs, D., Nieuwenhuijsen, M., Oudin, A., Forsberg, B., Olsson, D., Salomaa, V., Lanki, T., Yli-Tuomi, T., Oftedal, B., Aamodt, G., Nafstad, P., De Faire, U., Pedersen, N. L., Östenson, C. G., Fratiglioni, L., Penell, J., Korek, M., Pyko, A., Eriksen, K. T., Tjønneland, A., Becker, T., Eeftens, M., Bots, M., Meliefste, K., Wang, M., Bueno-De-Mesquita, B., Sugiri, D., Krämer, U., Heinrich, J., De Hoogh, K., Key, T., Peters, A., Cyrys, J., Concin, H., Nagel, G., Ineichen, A., Schaffner, E., Probst-Hensch, N., Dratva, J., Ducret-Stich, R., Vilier, A., Clavel-Chapelon, F., Stempfelet, M., Grioni, S., Krogh, V., Tsai, M. Y., Marcon, A., Ricceri, F., Sacerdote, C., Galassi, C., Migliore, E., Ranzi, A., Cesaroni, G., Badaloni, C., Forastiere, F., Tamayo, I., Amiano, P., Dorronsoro, M., Katsoulis, M., Trichopoulou, A., Vineis, P., and Hoek, G.: Long-term exposure to air pollution and cardiovascular mortality: An analysis of 22 European cohorts, *Epidemiology*, 25, 368–378, <https://doi.org/10.1097/EDE.000000000000076>, 2014.
- Belis, C. A., Karagulian, F., Larsen, B. R., and Hopke, P. K.: Critical review and meta-analysis of ambient particulate matter source apportionment using receptor models in Europe, *Atmos. Environ.*, 69, 94–108, <https://doi.org/10.1016/j.atmosenv.2012.11.009>, 2013.
- Belis, C. A., Karagulian, F., Amato, F., Almeida, M., Artaxo, P., Beddows, D. C. S., Bernardoni, V., Bove, M. C., Carbone, S., Cesari, D., Contini, D., Cuccia, E., Diapouli, E., Eleftheriadis, K., Favez, O., El Haddad, I., Harrison, R. M., Hellebust, S., Hovorka, J., Jang, E., Jorquera, H., Kammermeier, T., Karl, M., Lucarelli, F., Mooibroek, D., Nava, S., Nøjgaard, J. K., Paatero, P., Pandolfi, M., Perrone, M. G., Petit, J. E., Pietro-dangelo, A., Pokorná, P., Prati, P., Prevot, A. S. H., Quass, U., Querol, X., Saraga, D., Sciare, J., Sfetsos, A., Valli, G., Vecchi, R., Vestenius, M., Yubero, E., and Hopke, P. K.: A new methodology to assess the performance and uncertainty of source apportionment models II: The results of two European intercomparison exercises, *Atmos. Environ.*, 123, 240–250, <https://doi.org/10.1016/j.atmosenv.2015.10.068>, 2015.
- Bell, M. L., Samet, J. M., and Dominici, F.: Time-series studies of particulate matter, *Annu. Rev. Publ. Health*, 25, 247–280, <https://doi.org/10.1146/annurev.publhealth.25.102802.124329>, 2004.
- Benkendorf, D. J. and Hawkins, C. P.: Effects of sample size and network depth on a deep learning approach to species distribution modeling, *Ecol. Inform.*, 60, *Ecol. Inform.*, 60, 101137, <https://doi.org/10.1016/j.ecoinf.2020.101137>, 2020.
- Borlaza, L. J. S., Weber, S., Uzu, G., Jacob, V., Cañete, T., Micallef, S., Trébuchon, C., Slama, R., Favez, O., and Jaffrezo, J.-L.: 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, <https://doi.org/10.5194/acp-21-5415-2021>, 2021a.
- Borlaza, L. J. S., Weber, S., Jaffrezo, J.-L., Houdier, S., Slama, R., Rieux, C., Albinet, A., Micallef, S., Trébluchon, C., and Uzu, G.: Disparities in particulate matter (PM₁₀) origins and oxidative potential at a city scale (Grenoble, France) – Part 2: Sources of PM₁₀ oxidative potential using multiple linear regression analysis and the predictive applicability of multilayer perceptron neural network analysis, *Atmos. Chem. Phys.*, 21, 9719–9739, <https://doi.org/10.5194/acp-21-9719-2021>, 2021b.
- Bourlard, H. and Wellekens, C. J.: Speech pattern discrimination and multilayer perceptrons, *Comput. Speech Lang.*, 3, 1–19, [https://doi.org/10.1016/0885-2308\(89\)90011-9](https://doi.org/10.1016/0885-2308(89)90011-9), 1989.
- Breiman, L.: RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis, *Mach. Learn.*, 45, 5–32, https://doi.org/10.1007/978-3-030-62008-0_35, 2001.
- Brook, R. D., Rajagopalan, S., Pope, C. A., Brook, J. R., Bhatnagar, A., Diez-Roux, A. V., Holguin, F., Hong, Y., Luepker, R. V., Mittleman, M. A., Peters, A., Siscovick, D., Smith, S. C., Whitsel, L., and Kaufman, J. D.: Particulate matter air pollution and cardiovascular disease: An update to the scientific statement from the american heart association, *Circulation*, 121, 2331–2378, <https://doi.org/10.1161/CIR.0b013e3181d8bec1>, 2010.
- Brown, S. G., Eberly, S., Paatero, P., and Norris, G. A.: 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, <https://doi.org/10.1016/j.scitotenv.2015.01.022>, 2015.
- Calas, A., Uzu, G., Martins, J. M. F., Voisin, Di., Spadini, L., Lacroix, T., and Jaffrezo, J. L.: The importance of simulated lung fluid (SLF) extractions for a more relevant evaluation of the oxidative potential of particulate matter, *Sci. Rep.-UK*, 7, 1–12, <https://doi.org/10.1038/s41598-017-11979-3>, 2017.
- Calas, A., Uzu, G., Kelly, F. J., Houdier, S., Martins, J. M. F., Thomas, F., Molton, F., Charron, A., Dunster, C., Ollie, A., Jacob, V., Besombes, J.-L., Chevrier, F., and Jaffrezo, J.-L.: Comparison between five acellular oxidative potential measurement assays performed with detailed chemistry on PM₁₀ samples from the city of Chamonix (France), *Atmos. Chem. Phys.*, 18, 7863–7875, <https://doi.org/10.5194/acp-18-7863-2018>, 2018.
- Calas, A., Uzu, G., Besombes, J. L., Martins, J. M. F., Redaelli, M., Weber, S., Charron, A., Albinet, A., Chevrier, F., Brulfert, G., Mesbah, B., Favez, O., and Jaffrezo, J. L.: Seasonal variations and chemical predictors of oxidative potential (OP) of particulate matter (PM), for seven urban French sites, *Atmosphere-Basel*, 10, 698, <https://doi.org/10.3390/atmos10110698>, 2019.
- Chianese, E., Camastra, F., and Ciamarella, A.: Spatio-temporal learning in predicting ambient particulate matter concentration by multi-layer perceptron Spatio-temporal Learning in Predicting Ambient Particulate Matter Concentration by Multi-Layer, *Ecol. Inform.*, 49, 54–61, <https://doi.org/10.1016/j.ecoinf.2018.12.001>, 2018.
- Cho, A., Sioutas, C., Miguel, A. H., Kumagai, Y., Schmitz, D. A., Singh, M., Eiguren-Fernandez, A., and Froines, J. R.: Redox activity of airborne particulate matter at different sites in the Los Angeles Basin, *Environ. Res.*, 99, 40–47, <https://doi.org/10.1016/j.envres.2005.01.003>, 2005.

- Cohen, J., Cohen, P., West, S. G., and Aiken, L. S.: Applied multiple regression/correlation analysis for the behavioral sciences, Routledge, 536 pp., <https://doi.org/10.4324/9780203774441>, 2002.
- Craney, T. A. and Surles, J. G.: Model-dependent variance inflation factor cutoff values, *Qual. Eng.*, 14, 391–403, <https://doi.org/10.1081/QEN-120001878>, 2002.
- Crobeddu, B., Aragao-Santiago, L., Bui, L. C., Boland, S., and Baeza Squiban, A.: Oxidative potential of particulate matter 2.5 as predictive indicator of cellular stress, *Environ. Pollut.*, 230, 125–133, <https://doi.org/10.1016/j.envpol.2017.06.051>, 2017.
- Crouse, D. L., Peters, P. A., van Donkelaar, A., Goldberg, M. S., Villeneuve, P. J., Brion, O., Khan, S., Atari, D. O., Jerrett, M., Pope, C. A., Brauer, M., Brook, J. R., Martin, R. V., Stieb, D., and Burnett, R. T.: Risk of nonaccidental and cardiovascular mortality in relation to long-term exposure to low concentrations of fine particulate matter: A Canadian national-level cohort study, *Environ. Health Persp.*, 120, 708–714, <https://doi.org/10.1289/ehp.1104049>, 2012.
- Crouse, D. L., Peters, P. A., Hystad, P., Brook, J. R., van Donkelaar, A., Martin, R. V., Villeneuve, P. J., Jerrett, M., Goldberg, M. S., Arden Pope, C., Brauer, M., Brook, R. D., Robichaud, A., Menard, R., and Burnett, R. T.: Ambient PM_{2.5}, O₃, and NO₂ exposures and associations with mortality over 16 years of follow-up in the Canadian census health and environment cohort (CanCHEC), *Environ. Health Persp.*, 123, 1180–1186, <https://doi.org/10.1289/ehp.1409276>, 2015.
- Daellenbach, K. R., Uzu, G., Jiang, J., Cassagnes, L.-E., Leni, Z., Vlachou, A., Stefenelli, G., Canonaco, F., Weber, S., Segers, A., Kuenen, J. J. P., Schaap, M., Favez, O., Albinet, A., Aksoyoglu, S., Dommen, J., Baltensperger, U., Geiser, M., El Haddad, I., Jaffrezo, J.-L., and Prévôt, A. S. H.: Sources of particulate-matter air pollution and its oxidative potential in Europe of particulate-matter air pollution and its oxidative potential in Europe, *Nature*, 587, 414–419, <https://doi.org/10.1038/s41586-020-2902-8>, 2020.
- Deng, M., Chen, D., Zhang, G., and Cheng, H.: Policy-driven variations in oxidation potential and source apportionment of PM_{2.5} in Wuhan, central China, *Sci. Total Environ.*, 853, 158255, <https://doi.org/10.1016/j.scitotenv.2022.158255>, 2022.
- Dominici, F.: Time-series analysis of air pollution and mortality: a statistical review, *Res. Rep. Health. Eff. Inst.*, 123, 3–27, 2004.
- Dominutti, P. A., Borlaza, L., Sauvain, J. J., Ngoc Thuy, V. D., Houdier, S., Suarez, G., Jaffrezo, J. L., Tobin, S., Trébouchon, C., Socquet, S., Moussu, E., Mary, G., and Uzu, G.: Source apportionment of oxidative potential depends on the choice of the assay: insights into 5 protocols comparison and implications for mitigation measures, *Environ. Sci. Atmos.*, 3, 1497–1512, <https://doi.org/10.1039/d3ea00007a>, 2023.
- Elangasinghe, M. A., Singhal, N., Dirks, K. N., and Salmond, J. A.: Development of an ANN-based air pollution forecasting system with explicit knowledge through sensitivity analysis, *Atmos. Pollut. Res.*, 5, 696–708, <https://doi.org/10.5094/APR.2014.079>, 2014.
- Fadel, M., Courcot, D., Delmaire, G., Roussel, G., Affif, C., and Ledoux, F.: Source apportionment of PM_{2.5} oxidative potential in an East Mediterranean site, *Sci. Total Environ.*, 900, 165843, <https://doi.org/10.1016/j.scitotenv.2023.165843>, 2023.
- Fang, T., Verma, V., Bates, J. T., Abrams, J., Klein, M., Strickland, M. J., Sarnat, S. E., Chang, H. H., Mulholland, J. A., Tolbert, P. E., Russell, A. G., and Weber, R. J.: Oxidative potential of ambient water-soluble PM_{2.5} in the southeastern United States: contrasts in sources and health associations between ascorbic acid (AA) and dithiothreitol (DTT) assays, *Atmos. Chem. Phys.*, 16, 3865–3879, <https://doi.org/10.5194/acp-16-3865-2016>, 2016.
- Favez, O.: Traitement harmonisé de jeux de données multi-sites pour l'étude des sources de PM par Positive Matrix Factorization, Technical Report, <https://docplayer.fr/124547484-Traitement-harmonise-de-jeux-de-donnees-multi-sites-pour-l-etude-des-sources-de-pm-par-positive-matrix-factorization.html> (last access: 18 June 2024), 2017.
- Godri, K. J., Harrison, R. M., Evans, T., Baker, T., Dunster, C., Mudway, I. S., and Kelly, F. J.: Increased oxidative burden associated with traffic component of ambient particulate matter at roadside and Urban background schools sites in London, *PLoS One*, 6, e21961, <https://doi.org/10.1371/journal.pone.0021961>, 2011.
- Goldfeld, S. M. and Quandt, R. E.: Some tests for homoscedasticity, *J. Am. Stat. Assoc.*, 60, 539–547, 1965.
- Harrell: Regression Modeling Strategies, *Technometrics*, 45, 170–170, <https://doi.org/10.1198/tech.2003.s158>, 2016.
- Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H.: Springer Series in Statistics The Elements of Statistical Learning, *Math. Intell.*, 27, 83–85, 2009.
- Hawkins, D. M.: The Problem of Overfitting, *J. Chem. Inf. Comp. Sci.*, 44, 1–12, <https://doi.org/10.1021/ci0342472>, 2004.
- Hernandez, P. A., Graham, C. H., Master, L. L., and Albert, D. L.: The effect of sample size and species characteristics on performance of different species distribution modeling methods, *Ecography*, 29, 773–785, <https://doi.org/10.1111/j.0906-7590.2006.04700.x>, 2006.
- Hoerl, A. E. and Kennard, R. W.: Ridge Regression: Applications to Nonorthogonal Problems, *Technometrics*, 12, 69–82, <https://doi.org/10.2307/1267352>, 1970.
- in't Veld, M., Pandolfi, M., Amato, F., Pérez, N., Reche, C., Dominutti, P., Jaffrezo, J., Alastuey, A., Querol, X., and Uzu, G.: Discovering oxidative potential (OP) drivers of atmospheric PM₁₀, PM_{2.5}, and PM₁ simultaneously in North-Eastern Spain, *Sci. Total Environ.*, 857, 159386, <https://doi.org/10.1016/j.scitotenv.2022.159386>, 2023.
- Janssen, N. A. H., Yang, A., Strak, M., Steenhof, M., Hellack, B., Gerlofs-Nijland, M. E., Kuhlbusch, T., Kelly, F., Harrison, R., Brunekreef, B., Hoek, G., and Cassee, F.: Oxidative potential of particulate matter collected at sites with different source characteristics, *Sci. Total Environ.*, 472, 572–581, <https://doi.org/10.1016/j.scitotenv.2013.11.099>, 2014.
- Kelly, F. J. and Mudway, I. S.: Protein oxidation at the air-lung interface, *Amino Acids*, 25, 375–396, <https://doi.org/10.1007/s00726-003-0024-x>, 2003.
- Kuhn, M. and Johnson, K.: Applied predictive modeling, New York, Springer, 600 pp., <https://doi.org/10.1007/978-1-4614-6849-3>, 2013.
- Leni, Z., Cassagnes, L. E., Daellenbach, K. R., Haddad, I. El, Vlachou, A., Uzu, G., Prévôt, A. S. H., Jaffrezo, J. L., Baumann, N., Salathe, M., Baltensperger, U., Dommen, J., and Geiser, M.: Oxidative stress-induced inflammation in susceptible airways by anthropogenic aerosol, *PLoS One*, 15, e0233425, <https://doi.org/10.1371/journal.pone.0233425>, 2020.

- Li, N., Xia, T., and Nel, A. E.: The role of oxidative stress in ambient particulate matter-induced lung diseases and its implications in the toxicity of engineered nanoparticles, *Free Radic. Biol. Med.*, 44, 1689–1699, <https://doi.org/10.1016/j.freeradbiomed.2008.01.028>, 2008.
- Li, J., Zhao, S., Xiao, S., Li, X., Wu, S., Zhang, J., and Schwab, J. J.: Source apportionment of water-soluble oxidative potential of PM_{2.5} in a port city of Xiamen, Southeast China, *Atmos. Environ.*, 314, 120122, <https://doi.org/10.1016/j.atmosenv.2023.120122>, 2023.
- Liu, F. and Ng, N. L.: Toxicity of Atmospheric Aerosols: Methodologies & Assays, *Am. Chem. Soc.*, Chap. 3, ISBN 9780841299818, <https://doi.org/10.1021/acsinfocus.7e7012>, 2023.
- Liu, W. J., Xu, Y. S., Liu, W. X., Liu, Q. Y., Yu, S. Y., Liu, Y., Wang, X., and Tao, S.: Oxidative potential of ambient PM_{2.5} in the coastal cities of the Bohai Sea, northern China: Seasonal variation and source apportionment, *Environ. Pollut.*, 236, 514–528, <https://doi.org/10.1016/j.envpol.2018.01.116>, 2018.
- Lodovici, M. and Bigagli, E.: Oxidative stress and air pollution exposure, *J. Toxicol.*, 2011, 487074, <https://doi.org/10.1155/2011/487074>, 2011.
- Matsuki, K., Kuperman, V., and Van Dyke, J. A.: The Random Forests statistical technique: An examination of its value for the study of reading, *Sci. Stud. Read.*, 20, 20–33, <https://doi.org/10.1080/10888438.2015.1107073>, 2016.
- McCullagh: Generalized linear models, *Eur. J. Oper. Res.*, 16, 285–292, <https://doi.org/10.1201/9780203738535>, 1989.
- Montgomery, C. D., Peck, A. E., and Vining, G. G.: *Introducing To Linear Regression Analysis*, John Wiley & Sons, 5th edn., ISBN 9781119180173, 2012.
- Mudway, I. S., Kelly, F. J., and Holgate, S. T.: Oxidative stress in air pollution research, *Free Radical Bio. Med.*, 151, 2–6, <https://doi.org/10.1016/j.freeradbiomed.2020.04.031>, 2020.
- Nelin, T. D., Joseph, A. M., Gorr, M. W., and Wold, L. E.: Direct and indirect effects of particulate matter on the cardiovascular system, *Toxicol. Lett.*, 208, 293–299, <https://doi.org/10.1016/j.toxlet.2011.11.008>, 2012.
- Ngoc Thuy, V. D.: Regression techniques applied to particulate matter oxidative potential source apportionment, Zenodo [code], <https://doi.org/10.5281/zenodo.11071884>, 2024.
- O'Brien, R. M.: A caution regarding rules of thumb for variance inflation factors, *Qual. Quant.*, 41, 673–690, <https://doi.org/10.1007/s11135-006-9018-6>, 2007.
- Paatero, P. and Hopke, P. K.: Rotational tools for factor analytic models, *J. Chemometr.*, 23, 91–100, <https://doi.org/10.1002/cem.1197>, 2009.
- Paatero, P. and Tappert, U.: Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values, *Environmetrics*, 5, 111–126, <https://doi.org/10.1002/env.3170050203>, 1994.
- Pearce, J. and Ferrier, S.: An evaluation of alternative algorithms for fitting species distribution models using logistic regression, *Ecol. Modell.*, 128, 127–147, [https://doi.org/10.1016/S0304-3800\(99\)00227-6](https://doi.org/10.1016/S0304-3800(99)00227-6), 2000.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, *J. Mach. Learn. Res.*, 12, 2825–2830, 2011.
- Pelucchi, C., Negri, E., Gallus, S., Boffetta, P., Tramacere, I., and La Vecchia, C.: Long-term particulate matter exposure and mortality: A review of European epidemiological studies, *BMC Public Health*, 9, 1–8, <https://doi.org/10.1186/1471-2458-9-453>, 2009.
- Peng, R. D., Bell, M. L., Geyh, A. S., McDermott, A., Zeger, S. L., Samet, J. M., and Dominici, F.: Emergency admissions for cardiovascular and respiratory diseases and the chemical composition of fine particle air pollution, *Environ. Health Persp.*, 117, 957–963, <https://doi.org/10.1289/ehp.0800185>, 2009.
- Pietrogrande, M. C., Romanato, L., and Russo, M.: Synergistic and Antagonistic Effects of Aerosol Components on Its Oxidative Potential as Predictor of Particle Toxicity, *Toxics*, 10, 196, <https://doi.org/10.3390/toxics10040196>, 2022.
- Pope, C. A. and Dockery, D. W.: Health effects of fine particulate air pollution: Lines that connect, *J. Air Waste Manage.*, 56, 709–742, <https://doi.org/10.1080/10473289.2006.10464485>, 2006.
- Rao, X., Zhong, J., Brook, R. D., and Rajagopalan, S.: Effect of Particulate Matter Air Pollution on Cardiovascular Oxidative Stress Pathways, *Antioxidants Redox Signal*, 28, 797–818, <https://doi.org/10.1089/ars.2017.7394>, 2018.
- Raudys, S. J. and Jain, A. K.: Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners, *IEEE T. Pattern Anal.*, 13, 252–264, <https://doi.org/10.1109/34.75512>, 1991.
- Rosenblad, A.: The Concise Encyclopedia of Statistics, *J. Appl. Stat.*, 38, 867–868, <https://doi.org/10.1080/02664760903075614>, 2011.
- Samake, A., Uzu, G., Martins, J. M. F., Calas, A., Vince, E., Parat, S., and Jaffrezou, J. L.: The unexpected role of bioaerosols in the Oxidative Potential of PM, *Sci. Rep.-UK*, 7, 10978, <https://doi.org/10.1038/s41598-017-11178-0>, 2017.
- Seabold, S. and Perktold, J.: Statsmodels: Econometric and statistical modeling with python, in: 9th Python in Science Conference, Vol. 7, p. 1, 28 June–3 July, <http://conference.scipy.org.s3-website-us-east-1.amazonaws.com/proceedings/scipy2010/pdfs/seabold.pdf> (last access: 15 May 2024), 2010.
- Shangguan, Y., Zhuang, X., Querol, X., Li, B., Moreno, N., Trechera, P., Sola, P. C., Uzu, G., and Li, J.: Characterization of deposited dust and its respirable fractions in underground coal mines: Implications for oxidative potential-driving species and source apportionment, *Int. J. Coal Geol.*, 258, 104017, <https://doi.org/10.1016/j.coal.2022.104017>, 2022.
- Stevanović, S., Jovanović, M. V., Jovašević-Stojanović, M. V., and Ristovski, Z.: Source apportionment of oxidative potential What We Know So Far, *Therm. Sci.*, 27, 2347–2357, <https://doi.org/10.2298/TSCI221107111S>, 2023.
- Stockwell, D. R. B. and Peterson, A. T.: Effects of sample size on accuracy of species distribution models, *Ecol. Modell.*, 148, 1–13, [https://doi.org/10.1016/S0304-3800\(01\)00388-X](https://doi.org/10.1016/S0304-3800(01)00388-X), 2002.
- Szigeti, T., Óvári, M., Dunster, C., Kelly, F. J., Lucarelli, F., and Zárny, G.: Changes in chemical composition and oxidative potential of urban PM_{2.5} between 2010 and 2013 in Hungary, *Sci. Total Environ.*, 518–519, 534–544, <https://doi.org/10.1016/j.scitotenv.2015.03.025>, 2015.
- Szigeti, T., Dunster, C., Cattaneo, A., Cavallo, D., Spinazzè, A., Saraga, D. E., Sakellaris, I. A., de Kluizenaar, Y., Cornelissen,

- E. J. M., Hänninen, O., Peltonen, M., Calzolari, G., Lucarelli, F., Mandin, C., Bartzis, J. G., Záray, G., and Kelly, F. J.: Oxidative potential and chemical composition of PM_{2.5} in office buildings across Europe – The OFFICAIR study, *Environ. Int.*, 92–93, 324–333, <https://doi.org/10.1016/j.envint.2016.04.015>, 2016.
- Tibshirani, R.: Regression Shrinkage and Selection Via the Lasso, *J. R. Stat. Soc. B*, 58, 267–288, <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>, 1996.
- Verma, V., Fang, T., Guo, H., King, L., Bates, J. T., Peltier, R. E., Edgerton, E., Russell, A. G., and Weber, R. J.: Reactive oxygen species associated with water-soluble PM_{2.5} in the southeastern United States: spatiotemporal trends and source apportionment, *Atmos. Chem. Phys.*, 14, 12915–12930, <https://doi.org/10.5194/acp-14-12915-2014>, 2014.
- Viana, M., Kuhlbusch, T. A. J., Querol, X., Alastuey, A., Harrison, R. M., Hopke, P. K., Winiwarter, W., Vallius, M., Szidat, S., Prévôt, A. S. H., Hueglin, C., Bloemen, H., Wählin, P., Vecchi, R., Miranda, A. I., Kasper-Giebl, A., Maenhaut, W., and Hitenberger, R.: Source apportionment of particulate matter in Europe: A review of methods and results, *J. Aerosol Sci.*, 39, 827–849, <https://doi.org/10.1016/j.jaerosci.2008.05.007>, 2008.
- Vida, M., Foret, G., Siour, G., Coman, A., Weber, S., Favez, O., Jaffrezo, J., Pontet, S., Mesbah, B., Gille, G., Zhang, S., Chevrier, F., Pallares, C., Uzu, G., and Beekmann, M.: Oxidative potential modelling of PM₁₀: a 2-year study over France, *Atmos. Chem. Phys.*, in preparation, 2024.
- Wang, D., Yang, X., Lu, H., Li, D., Xu, H., Luo, Y., Sun, J., Hang Ho, S. S., and Shen, Z.: Oxidative potential of atmospheric brown carbon in six Chinese megacities: Seasonal variation and source apportionment, *Atmos. Environ.*, 309, 119909, <https://doi.org/10.1016/j.atmosenv.2023.119909>, 2023.
- Wang, J., Jiang, H., Jiang, H., Mo, Y., Geng, X., Li, J., Mao, S., Bualert, S., Ma, S., Li, J., and Zhang, G.: Source apportionment of water-soluble oxidative potential in ambient total suspended particulate from Bangkok: Biomass burning versus fossil fuel combustion, *Atmos. Environ.*, 235, 117624, <https://doi.org/10.1016/j.atmosenv.2020.117624>, 2020.
- Wang, S., Ye, J., Soong, R., Wu, B., Yu, L., Simpson, A. J., and Chan, A. W. H.: Relationship between chemical composition and oxidative potential of secondary organic aerosol from polycyclic aromatic hydrocarbons, *Atmos. Chem. Phys.*, 18, 3987–4003, <https://doi.org/10.5194/acp-18-3987-2018>, 2018.
- Wang, Y., Wang, M., Li, S., Sun, H., Mu, Z., Zhang, L., Li, Y., and Chen, Q.: Study on the oxidation potential of the water-soluble components of ambient PM_{2.5} over Xi'an, China: Pollution levels, source apportionment and transport pathways, *Environ. Int.*, 136, 105515, <https://doi.org/10.1016/j.envint.2020.105515>, 2020.
- Weber, S., Uzu, G., Calas, A., Chevrier, F., Besombes, J.-L., Charon, A., Salameh, D., Ježek, I., Močnik, G., and Jaffrezo, J.-L.: An apportionment method for the oxidative potential of atmospheric particulate matter sources: application to a one-year study in Chamonix, France, *Atmos. Chem. Phys.*, 18, 9617–9629, <https://doi.org/10.5194/acp-18-9617-2018>, 2018.
- 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., and Favez, O.: Comparison of PM₁₀ sources profiles at 15 french sites using a harmonized constrained positive matrix factorization approach, *Atmosphere-Basel*, 10, 310, <https://doi.org/10.3390/atmos10060310>, 2019.
- Weber, S., Uzu, G., Favez, O., Borlaza, L. J. S., Calas, A., Salameh, D., Chevrier, F., Allard, J., Besombes, J.-L., Albinet, A., Pontet, S., Mesbah, B., Gille, G., Zhang, S., Pallares, C., Leoz-Garziandia, E., and Jaffrezo, J.-L.: Source apportionment of atmospheric PM₁₀ oxidative potential: synthesis of 15 year-round urban datasets in France, *Atmos. Chem. Phys.*, 21, 11353–11378, <https://doi.org/10.5194/acp-21-11353-2021>, 2021.
- WHO: WHO global air quality guidelines, <https://iris.who.int/handle/10665/345329> (last access: 18 June 2024), 2021.
- Williams, M., Gomez Grajales, C. A., and Kurkiewicz, D.: Assumptions of Multiple Regression: Correcting Two Misconceptions – Practical Assessment, *Res. Evaluat.*, 18, 1–16, 2013.
- Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., Guisan, A., Elith, J., Dudík, M., Ferrier, S., Huettmann, F., Leathwick, J. R., Lehmann, A., Lohmann, L., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M. C., Phillips, S. J., Richardson, K. S., Scachetti-Pereira, R., Schapire, R. E., Soberón, J., Williams, S. E., and Zimmermann, N. E.: Effects of sample size on the performance of species distribution models, *Divers. Distrib.*, 14, 763–773, <https://doi.org/10.1111/j.1472-4642.2008.00482.x>, 2008.
- Xiong, Q., Yu, H., Wang, R., Wei, J., and Verma, V.: Rethinking Dithiothreitol-Based Particulate Matter Oxidative Potential: Measuring Dithiothreitol Consumption versus Reactive Oxygen Species Generation, *Environ. Sci. Technol.*, 51, 6507–6514, <https://doi.org/10.1021/acs.est.7b01272>, 2017.
- Yang, A., Jedynska, A., Hellack, B., Kooter, I., Hoek, G., Brunekreef, B., Kuhlbusch, T. A. J., Cassee, F. R., and Janssen, N. A. H.: Measurement of the oxidative potential of PM_{2.5} and its constituents: The effect of extraction solvent and filter type, *Atmos. Environ.*, 83, 35–42, <https://doi.org/10.1016/j.atmosenv.2013.10.049>, 2014.
- Yu, P., Guo, S., Xu, R., Ye, T., Li, S., Sim, M. R., Abramson, M. J., and Guo, Y.: Cohort studies of long-term exposure to outdoor particulate matter and risks of cancer: A systematic review and meta-analysis, *Innovation*, 2, 100143, <https://doi.org/10.1016/j.xinn.2021.100143>, 2021.
- Yu, S. Y., Liu, W. J., Xu, Y. S., Yi, K., Zhou, M., Tao, S., and Liu, W. X.: Characteristics and oxidative potential of atmospheric PM_{2.5} in Beijing: Source apportionment and seasonal variation, *Sci. Total Environ.*, 650, 277–287, <https://doi.org/10.1016/j.scitotenv.2018.09.021>, 2019.
- Zhang, Y., Albinet, A., Petit, J. E., Jacob, V., Chevrier, F., Gille, G., Pontet, S., Chrétien, E., Dominik-Sègue, M., Levigoureux, G., Močnik, G., Gros, V., Jaffrezo, J. L., and Favez, O.: Substantial brown carbon emissions from wintertime residential wood burning over France, *Sci. Total Environ.*, 743, 140752, <https://doi.org/10.1016/j.scitotenv.2020.140752>, 2020.