Science of the Total Environment 1007 (2025) 180983

Contents lists available at ScienceDirect

Science of the Total Environment

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

Source apportionment of gaseous pollutants in oil and gas extraction areas: A comparison between positive matrix factorization and self-organizing maps approaches

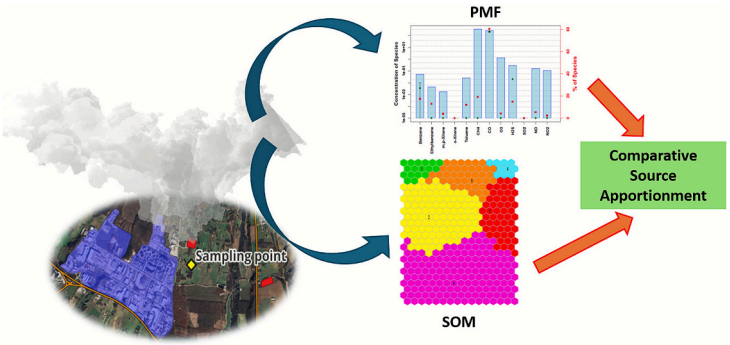
Mariassunta Biondi [a], Alessandro Zappi [a,*], Erika Brattich [b], Serena Sabia [c], Rosa Caggiano [c], Laura Tositti [a]

[a] Department of Chemistry "Giacomo Ciamician", University of Bologna, Bologna, BO, 40129, Italy
[b] Department of Physics and Astronomy "Augusto Righi", University of Bologna, Bologna, BO, 40126, Italy
[c] Institute of Methodologies for Environmental Analysis (IMAA), Italian National Research Council (CNR), Tito Scalo, PZ, 85050, Italy

HIGHLIGHTS

- Nine months of hourly data of gaseous pollutants were collected in Southern Italy
- Two source apportionment methods were applied: PMF and SOM
- Six pollution sources were highlighted by both methods
- A critical comparison of PMF and SOM results was carried out

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Atmospheric pollution
Source apportionment
Chemometrics
Data analysis
PMF and SOM comparison

ABSTRACT

Emissions from oil and gas extraction and processing may represent a significant source of atmospheric pollution, with potential hazards for human health, the environment, and the economy on a local to global scale. In this work, twelve atmospheric pollutants were quantified at hourly time resolution over eight months near the Val d'Agri Oil Center (COVA), the largest onshore oil and natural gas extraction and primary processing facility in Europe, located in a semi-rural area of southern Italy. Temporal trends and average concentration values indicate that few exceedances of regulated species were observed during the investigated period. Two multivariate statistical techniques were employed for source apportionment: the well-established Positive Matrix Factorization (PMF) and the neural-network based Self-Organizing Map (SOM) approaches were used, and their results were compared. Both PMF and SOM produced a six-source solution, identifying the main common factors responsible for air quality near COVA, including: photooxidative processes for the production of secondary gaseous species, traffic, and high-temperature COVA operations, in particular gas turbines. Interestingly, although the overlap of some of the sources emerged, likely in association with high-frequency environmental conditions, output differences emerged in the two cases. SOM succeeded in capturing highly episodic and seasonally affected sources

* Corresponding author at: Department of Chemistry "Giacomo Ciamician", University of Bologna, Via Gobetti 85, 40129, Bologna, BO, Italy.
E-mail address: alessandro.zappi4@unibo.it (A. Zappi).

https://doi.org/10.1016/j.scitotenv.2025.180983
Received 4 August 2025; Received in revised form 13 November 2025; Accepted 14 November 2025
Available online 18 November 2025

(Claus process, gas flaring), while PMF discriminated primary and secondary production of NO_x and achieved the description of fugitive emissions from extraction wells. Overall, the double approach applied provides a comprehensive description of emission sources, demonstrating that this kind of parallel source apportionment approach, when possible, would be of great benefit to environmental studies.

1. Introduction

According to the World Health Organization (WHO), as of today, nine out of ten people breathe polluted air (Ćurić et al., 2022), with the most exposed groups being children, elderly, individuals with respiratory conditions, and those in socioeconomically vulnerable situations (Martins et al., 2004). Furthermore, air pollution is a stressor for planetary ecosystems, with consequences for human health, terrestrial biodiversity, infrastructure systems, and the economy on a global scale (Lovett et al., 2009; Manisalidis et al., 2020). Air pollution sources are typically densely distributed in well-defined areas, such as urban districts, transportation hubs, and industrial districts. It is well known that pollution knows no political or geographical boundaries, and thus even rural and remote areas are not free from pollution (Majra, 2011). Moreover, rural areas can host many types of productive activities, which may consequently result in atmospheric emissions. Identifying and possibly quantifying pollution sources in the troposphere, thus, represents a focal issue, as well as a challenging task to minimize health and environmental hazards (Sicard et al., 2023).

The Agri Valley (Southern Italy) is a semirural area in southern Italy serving as an example of the development of a petrochemical facility in a region historically characterized by agricultural activities and small urban settlements (Tositti et al., 2022). Since the 1990s, this area has hosted the largest onshore oil and natural gas extraction and primary processing facility in Europe, known as the Centro Olio Val d'Agri (COVA) (Mininni, 2015). Currently, COVA operates 28 wells for the extraction of crude oil, which is conveyed via a pipeline network to the natural gas processing and initial oil processing center, with a treatment capacity of 16,500 m^3 of oil and 3,100,000 m^3 of associated natural gas per day (Calvello et al., 2017). The local oil processing facility involves the separation of the extracts into three constituent phases: natural gas, water, and crude oil (then delivered, via an oil pipeline, to the refinery facility about 120 km away, on the Ionian coast). The processing on each phase includes the removal of contaminants, the most relevant being hydrogen sulfide (H_2S), which is removed through the Claus-Scot process (Tositti et al., 2022). The COVA oil processing activity contributes to several sources of gaseous and particulate pollutants. Extraction, transportation, storage, and processing are all linked to potential fugitive emissions – particularly of aliphatic and aromatic hydrocarbons, besides sulphurated species – while auxiliary activities, like gas flaring and methane combustion in gas turbines, further impact the local airshed, as evidenced by numerous reports and studies (Buzcu-Guven et al., 2010). Given the significant emissive potential of COVA, long-term researches and monitoring on local air quality have been (and still are) conducted, highlighting its impact on the concentration of BC, and PM chemical speciation (Castagna et al., 2019; Pavese et al., 2012). In particular, the influence of multiple emission points at COVA has been investigated using exploratory multivariate statistical techniques (Calvello et al., 2017) and Positive Matrix Factorization (Tositti et al., 2022), highlighting the effects of both gas flaring and Claus process. Therefore, the use of extensive datasets and advanced computational/statistical techniques, such as those for source apportionment, allows for a detailed understanding of the atmospheric air quality conditions across the Agri Valley.

A sampling campaign including both daily PM1, chemical speciation, and hourly airborne pollutants was carried out in the Agri Valley in the period July 2017–March 2018. The aims of the present study, solely focused on hourly gas concentrations and weather parameters, are both to perform a further source apportionment for the area (Tositti et al., 2022) and to compare the performances of two source apportionment methods applied on the same dataset: Positive Matrix Factorization (PMF) and Self-Organizing Maps (SOM).

The PMF (Paatero and Tapper, 1994) is traditionally the most widely used and reliable method for source apportionment. Its main application field concerns airborne particulate matter (PM) for which it reconstructs the source profiles, also considering the mass load and producing robust output models. However, some drawbacks are present when computing PMF models, such as the possible troubles in defining the optimal number of factors, or in reliably defining the experimental uncertainties, especially when data are collected with real-time instruments (Belis et al., 2019; Huang et al., 2021). Moreover, in the case of gaseous pollutants, an experimental mass load is not available, so it is not possible to evaluate the fit of the modeled mass.

The SOM (Kohonen, 1990) method, instead, relies on unsupervised neural networks, which group data based on similarity. Recently, the potential of SOM in this domain was demonstrated in a study concerning gaseous criteria pollutants, typically found in extensive databases, in a complex emissive scenario (Zappi et al., 2024). SOM calculation is less computationally costly than PMF, and the univocal allocation of each observation of the dataset into a group allows further post-model exploitation of the model output in an easier way than PMF. However, unlike PMF, this method does not provide a quantitative estimate of source contributions. Interestingly, the emerging differences from the two applied methods did not imply a bare redistribution of the chemical parameters into sources/factors; rather, based on the meteorological variables included, SOM revealed the ability to enlighten specific circulation conditions, spatially addressing the sources emissive behavior. This factor may be very useful in cases where the number of chemical parameters is limited, but where the richness of meteorological data compensates for detailed circulation information, reflecting the actual condition for emission exposure at the receptor site.

2. Materials and methods

2.1. Study area

Agri Valley lies in the Basilicata region, Southern Italy (Fig. 1). With an area of 1723 km^2, its altitude ranges between 600 and 1200 m a.s.l.; it exhibits considerable morphological and geological diversity, associated with various types of ecosystems and rich biodiversity, imparting significant natural value to the area (Di Gilio et al., 2021).

Climatically, Agry Valley is characterized by a coexistence of a continental climate, featuring hot summers and harsh, dry winters, and a more temperate Mediterranean climate, mainly on the western coast (Giunta et al., 2022). Overall, the area experiences significant seasonal variations, with cold winters and moderately warm, dry summers. Furthermore, the height (604 m a.s.l.) and the proximity of the hills (southwest of the district) make Agri Valley rainier as compared to the whole Basilicata territory (Tositti et al., 2022). The region retains a rural character and is mostly home to small villages, one of the most populated being Viggiano, with approximately 3000 inhabitants. These villages are connected by a low-traffic road network, except for the state road (S.S.) 598, which serves as a connection between the valley and the cities of Potenza and Taranto (Di Gilio et al., 2021). The municipality of Viggiano hosts a significant industrial area, developed as a direct consequence of the COVA settlement, whose first production line has been operational since 1996 and subsequently expanded in 2001. The sampling site (Fig. 1) was placed in Viggiano (40°18.84′N, 15°54.27′E)

at 450 m E/NE from the COVA settlement. Several other potential sources of emissions are present around the station, including two COVA extraction wells, at 250 m NW and 800 m E, the S.S. 598, at 1100 m S, and the provincial road (S.P.) 103.

2.2. Sampling campaign

The concentrations of several criteria pollutants and other gaseous species associated with oil extraction activities were continuously monitored. These include carbon monoxide (CO), nitrogen oxides (NO and NO_2), ozone (O_3), BTEX compounds (benzene, toluene, ethylbenzene, and ortho-, meta-, and para-xylene), methane (CH_4), hydrogen sulfide (H_2S), and sulfur dioxide (SO_2) (Table 1). The measurements were provided by the Regional Agency for Environmental Protection of Basilicata (ARPAB), and all concentration values were standardized to a common unit of ppb/h. Monitoring was carried out from July 18, 2017, to March 31, 2018. This period was chosen to involve both warm and cold seasons, which are characterized by significant differences in meteorological conditions and atmospheric circulation patterns. Based on the moving average of temperature, using 25 °C as the threshold, the sampling period was divided into "warm" and "cold" periods (Fig. S1 in SI).

Meteorological parameters, namely temperature (°C), pressure (hPa), relative humidity, rainfall (mm), wind speed (m s^{-1}), and wind direction (°N), are also measured at the same receptor site.

Temperature ranged between −6.7 °C (28/02/2018) to 42 °C (05/08/2017), with mean and standard deviation values of (13 ± 9 °C), and a median of 11 °C. The pressure was relatively low, on account of the site altitude, ranging from 923.6 (28/12/2017) to 971 (28/01/2018) hPa (941 ± 67 hPa, median 946 hPa). The relative humidity ranged from 11 % (27/08/2017) to 99 % (28/09/2017) (65 ± 22 %, median 70 %). Wind speed was generally low, seldom exceeding 8 m s^{-1}; however, as shown in Fig. 2, most of the winds reached the sampling station from W, positioning the station downwind of the COVA settlement.

2.3. Multivariate analysis

Multivariate statistical techniques were applied to the overall dataset to extract information on the sources of the monitored gaseous species from the variability observed over nine months of hourly data. The analysis focused on identifying temporal trends, relationships, and interactions, which are influenced by atmospheric reactivity, transport mechanisms, or shared emission sources.

Table 1
Methods used for measuring the concentration of each gaseous species.

Parameter	Measurement principle
CO	Non-dispersive IR detection (NDIR)
NO, NO_2	Chemiluminescence
O_3	UV absorption
BTEX	Gas chromatography with photo-ionization detector (GC-PID)
CH_4	Photo-ionization detection (PID)
H_2S, SO_2	UV fluorescence

2.3.1. Exploratory data analysis by PCA Varimax

Principal Component Analysis (PCA), specifically its Varimax variant, was applied to the dataset using CAT software (Leardi et al., 2015) based on the R programming language. PCA is the oldest technique in multivariate statistical analysis and remains the most widely used (Abdi and Williams, 2010). It effectively manages large and heterogeneous sets of variables, revealing existing patterns of similarity/covariance by representing the most significant information from the dataset in a new set of orthogonal variables, known as Principal Components (PCs). Typically, several PCs are considered, while the others are discarded, to cumulatively account for approximately 80 % of the total variance, significantly reducing data dimensionality. The Varimax variant of PCA involves a further rotation of the computed PCs, which are converted into more easily interpretable factors. Among the results, the loading matrix allows the identification of the strongest correlations between the original variables. The matrix contains contributions with values ranging from −1 to +1. To identify the most important correlations, for each factor, the variables with the highest loading values (in absolute value) are identified and are considered as the most important in defining the corresponding factors. In some cases, the factors can be considered a first outline of the pollution sources.

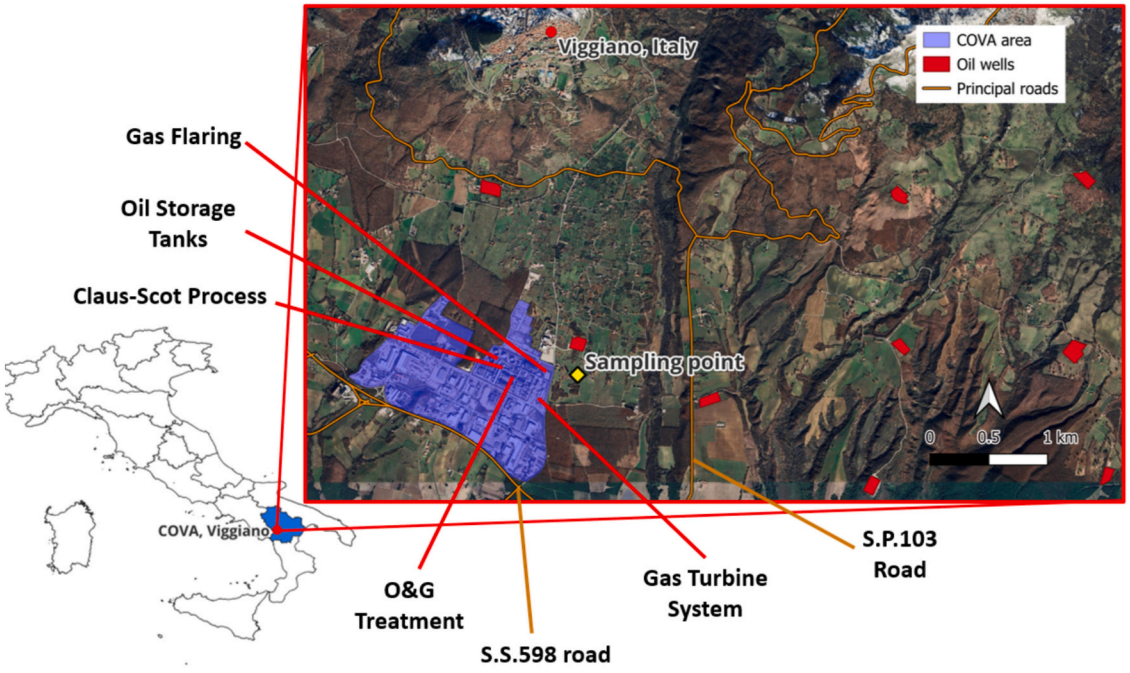


Fig. 1. Location of the sampling point and potential atmospheric emission sources in Agri Valley (Italy).

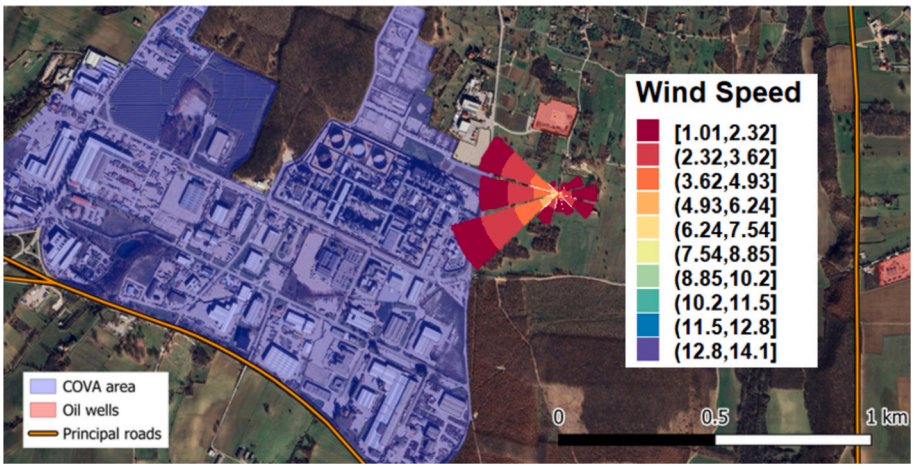


Fig. 2. Wind Rose registered at the sampling station for the whole sampling period (July, 18 - March, 31).

2.3.2. Positive matrix factorization (PMF)

Source apportionment computation was carried out using the EPA PMF 5.0 (Belis et al., 2019). PMF is the most important receptor model, designed to analyze data matrices and identify underlying latent structures, while addressing challenges such as noise and variability. Besides the data matrix, this model requires a matrix of uncertainties with the same dimensions as the data matrix. The uncertainty matrix helps to link the factor model to physical reality, enhancing the model robustness against potential noise in the data. Due to the absence of the instrumental uncertainties, the uncertainty matrix (σ_{ij}) was calculated from concentration values (x_{ij}) and instrumental detection limits (DL_j) using eq. 1 (Ito et al., 2004; Reff et al., 2007):

$$\sigma_{ij} = 0.05\, x_{ij} + DL_j \quad (1)$$

PMF, whose procedure is described elsewhere (Paatero and Tapper, 1994), is extensively used in environmental analysis to disclose the most important pollution sources and co-emissions of pollutants.

2.3.3. Self-organizing map (SOM)

Besides its traditional and diverse applications (Miljković, 2017), the SOM method is emerging as an effective technique for source apportionment in environmental science in recent years. In the present work, SOM computation was performed using the R package SOMEnv (Licen et al., 2023).

This approach employs an artificial neural network to map high-dimensional data into a lower-dimensional space. The description of the method is described elsewhere (Zappi et al., 2024). Substantially, this method produces a map, composed of a specific number of units (vectors), characterized by "regions of similarity", where similar units are located close to each other. These units can be grouped into clusters, capturing similar input variables and containing observations with comparable properties or origins. As an unsupervised learning method, SOM does not require labeled data and effectively preserves the initial relationships among the data points. In the present work, the number of units (m) in the map was determined as a function of the number of observations (n) (Nakagawa et al., 2020; Zappi et al., 2024), selecting the nearest integer to the result of eq. 2:

$$m = 5\sqrt{n} \quad (2)$$

The distribution of units in the two-dimensional space of the map was established by setting the dimensions of the axes as a proportion of the first two eigenvalues derived from a prior Principal Component Analysis (PCA) (Hentati et al., 2010).

3. Results and discussion

3.1. Preliminary data processing

Before elaborating the treated atmospheric pollutants and performing multivariate analyses, the quality of the dataset was evaluated by visual inspection of the time series of each parameter. This allowed the identification of irregularities in the baseline of CH_4, NO, NO_2, benzene, H_2S, and SO_2. Therefore, for the sake of homogeneity, a baseline correction was elaborated based on a homemade code in R (R Core Team, Vienna, Austria). To the scope, for each species, the median of the whole time series was calculated. Then, for each *i*th data point, the median of a window of 100 data points before and 100 after the *i*th point was calculated, and the difference between the two medians was added or subtracted from the *i*th point to shift it to have a baseline close to the overall median. In this way, the corrected series is vertically shifted to the same baseline value (the overall median), but the smoothing of peaks is limited, due to the large windows (201 data points) used to evaluate the difference between the medians. Fig. S2 shows an example of baseline correction on the H_2S time series. Moreover, as a preliminary data inspection, after baseline correction, some evident outliers and some missing data were replaced with the overall mean value. The outliers (22 data points) were found by visual inspection of the time series as isolated observations (one- or, at maximum, two-hours data) showing concentrations at least three-times higher than the corresponding air quality threshold limit, indicating a spike due to electric issues, rather than a real data.

3.2. Gaseous pollutants

Table 2 summarizes the basic statistics for the twelve gaseous pollutants over the entire sampling period. Figs. S3 and S4 show the time series of the pollutants, after the baseline-correction procedure. Their relative abundance is as follow (mean value ± standard deviation): CH_4 (1891.4 ± 87.7 ppb) > CO (323.9 ± 124.9 ppb) > O_3 (32.8 ± 14.3 ppb) > NO_2 (4.8 ± 4.5 ppb) > NO (3.0 ± 3.3 ppb) > H_2S (2.0 ± 1.6 ppb) > SO_2 (1.6 ± 3.6 ppb) > Benzene (0.30 ± 0.25 ppb) > Toluene (0.21 ± 0.26 ppb) > (m,p)-Xylene (0.05 ± 0.07 ppb) > Ethylbenzene (0.04 ± 0.03 ppb) > o-Xylene (0.01 ± 0.02 ppb).

The maxima reported in Table 2 have been compared to the hourly limit values established by EU Directive 2008/50/CE (European Parliament, 2008) for air quality regulation. During the studied period, no exceedance for CO and NO_2 was observed (thresholds: 8730 and 106 ppb respectively) and one exceedance (the maximum value reported in Table 2) was found for SO_2 (threshold: 134 ppb). Benzene, instead, showed 38 hourly data exceeding the threshold value (1.57 ppb), but, in

Table 2
Descriptive statistics of the gaseous pollutants concentrations (ppb) data: arithmetic mean, standard deviation (SD), minimum (Min) and maximum (Max), median, first and third quartile (Q1 and Q3), Inter-Quartile (Q3–Q1), and Max-Min ranges.

ppb	CH_4	CO	O_3	NO_2	NO	H_2S	SO_2	Benz	Tol	m.p-Xil	Etbenz	o-Xil
Mean	1891.4	323.9	32.8	4.8	3.0	2.03	1.6	0.30	0.21	0.05	0.04	0.01
SD	87.7	124.9	14.3	4.5	3.3	1.58	3.6	0.25	0.26	0.07	0.03	0.02
Min	1.1	0.00	0.8	0.0	0.0	0.18	0.0	0.00	0.00	0.00	0.00	0.00
Q1	1853.8	233.2	24.0	1.9	1.3	1.60	0.9	0.21	0.10	0.02	0.02	0.00
Median	1876.2	324.5	32.8	3.1	1.7	1.86	1.2	0.25	0.15	0.03	0.03	0.01
Q3	1908.9	404.9	40.7	6.0	3.1	2.13	1.6	0.34	0.24	0.06	0.04	0.01
Max	3256.1	826.6	375.1	31.1	31.6	47.95	149.9	5.91	5.42	1.94	0.57	0.64
Q3 – Q1	55.1	171.6	16.8	4.1	1.7	0.53	0.7	0.13	0.14	0.04	0.03	0.01
Max-Min	3255.1	826.6	374.3	31.1	31.6	47.77	149.9	5.91	5.42	1.94	0.57	0.64

this case, EU Directive 2008/50/CE indicates an annual threshold with a maximum of 10 % exceedances, which is respected in our dataset (0.6 % of exceedances). Tables S1 and S2 provided in the SI summarize the descriptive statistics calculated on a seasonal basis and, respectively, for the above-described "cold" and "warm" sampling periods (Fig. S1). Significant differences are primarily observed for O_3, with a higher relative mean concentration during the warm period. This is consistent with the well-known seasonal pattern of tropospheric O_3 (also evident from the O_3 time plot, Fig. S3), a secondary pollutant produced by the photo-oxidation mechanisms of NO_x and VOC (Chen et al., 2023). On the other hand, CO and NO show greater relative abundance in the cold period, likely driven by a combination of meteorological conditions and chemical processes, since cold temperatures often result in a shallower boundary layer and temperature inversions. Such conditions may trap pollutants near the surface, inhibiting dispersion and causing CO and NO accumulation (Kenagy et al., 2018). Moreover, CO is primarily removed from the atmosphere through reactions with hydroxyl (OH) radicals, produced by sunlight-driven processes. Shorter daylight hours and lower solar intensity in winter reduce OH concentrations, slowing the removal of CO and increasing its atmospheric lifetime. Similarly, the conversion of NO to other nitrogen oxides and secondary pollutants (like ozone and nitrates) is less efficient in winter due to lower temperatures and diminished sunlight, leading to higher NO concentrations (Kenagy et al., 2018).

3.3. Multivariate analysis

3.3.1. Exploratory data analysis by PCA Varimax

The loading matrix from the PCA-Varimax model calculated on the entire data set is reported in Table 3. The first four factors were selected, accounting for 74 % of the total explained variance (EV).

The first factor (31.8 % EV) relates to BTEX compounds, indicating a likely shared atmospheric source. Their emissions are generally primarily linked to fugitive emissions from motor vehicle fuel tanks, as well as from extraction wells, transport, storage, and processing associated with fossil fuel facilities (de Castro et al., 2015). The second factor (16.3 % EV) is characterized by contributions from NO_x (NO, NO_2), primarily associated with high-temperature combustion processes (Seinfeld, 2015). The third factor (14.3 % EV) is dominated by contributions from sulfureted gases (H_2S, SO_2), which can be connected to several conditions, including oil wells (H_2S), co-emission from the Claus process at COVA, all followed by the spontaneous atmospheric conversion of H_2S into SO_2 (Tositti et al., 2022). This agrees with our previous results from the source apportionment of chemically speciated PM_1 from the same area, where COVA plant emissions were revealed by the occurrence of elemental sulfur species from the catalytic processing of H_2S (Tositti et al., 2022). Finally, the fourth factor (11.6 % EV) appears to be dominated by the photo-oxidative relationship among pollutants, as revealed by the strong negative correlation of CH_4 and CO compared to O_3. The positive correlation between CH_4 and CO is likely due to shared local emission sources (e.g., gas flaring and the local thermal oxidizers), though CO might also have a secondary origin from the oxidation of the volatile hydrocarbons (Grant et al., 2010). The inverse relationship between these species and O_3 is related to the origin of ozone, a secondary gaseous pollutant produced through complex oxidation processes that connect hydrocarbons and NO_x, mediated by tropospheric photochemistry(Fiore et al., 2024).

Table 3
Loading matrix from the PCA-Varimax model. The most important variables for each factor are highlighted.

	Factor 1 (31.8 EV%)	Factor 2 (16.3 EV%)	Factor 3 (14.3 EV%)	Factor 4 (11.6 EV%)
Benz	**0.44**	−0.07	−0.04	−0.07
Etbenz	**0.44**	−0.01	0.05	−0.06
m,p-Xil	**0.46**	0.04	0.01	0.08
o-Xil	**0.44**	0.04	0.01	0.13
Tol	**0.46**	−0.01	−0.03	−0.06
CH4	0.02	0.11	−0.21	**−0.40**
CO	−0.01	−0.13	−0.04	**−0.69**
O_3	−0.02	−0.17	−0.19	**0.56**
H_2S	−0.01	0.00	**−0.69**	−0.02
SO_2	0.01	0.01	**−0.66**	0.04
NO	−0.01	**0.70**	0.02	0.04
NO_2	0.00	**0.67**	−0.03	0.02

3.3.2. Positive matrix factorization

Positive Matrix Factorization (PMF) was subsequently applied to hourly air pollution data over the entire sampling period to identify specific local sources and quantify their contributions, thereby complementing the exploratory framework provided by PCA. Data was then analyzed using EPA PMF software (Belis et al., 2019). First, data gaps were filled by calculating the average of the preceding and subsequent observations in the data set. In each case, the average value calculated was associated to an uncertainty of 400 %. Variables such as ethylbenzene, ortho-xylene, meta-para-xylene, H_2S, and SO_2 were classified as "weak" due to an S/N below 1, tripling their associated uncertainty. Additionally, a 5 % extra modeling uncertainty was added to account for other potential errors not included in the uncertainty matrix (Belis et al., 2019).

The model was tested with a range of 3 to 7 factors, and the 6-factor model was then deemed the best solution. This decision was based on the strong agreement of Q(true) and Q(robust) with Q(expected), as well as the intra-run variability and the normal distribution of the residuals, without excessive large-scaled values, the comparison between modeled and observed concentrations of pollutants, and the extraction of realistic and interpretable source profiles. The decision about the optimal solution were based on JRC Scientific and Technical Report (Comero et al., 2009) and on the European Guide on Source Apportionment (Belis et al., 2019). Additionally, this model struck the optimal balance between reducing the initial dimensionality of the data and delivering a clear, interpretable result, also aligned with the physical reality and existing knowledge about the studied location. The calculation was then repeated with 6 factors and 200 runs, to identify the absolute minimum in the solution space (Belis et al., 2019). Rotational tools were then

applied to the selected run (200), testing Fpeak values from −2 to +2 and ultimately selecting +0.5 based on %dQ and the best alignment between the model results and the physical reality of the problem (Belis et al., 2019). Factor-profiles and Factor-contributions charts for the selected solution are reported in Fig. 3.

At first glance, the six factors that emerged appear very different and highly detailed compared to the PCA Varimax analysis. They have therefore been used to identify local pollution sources, based on their chemical fingerprints, on the correlations between pollutants, and on their relationship with wind speed and direction.

More specifically, conditional bivariate probability functions (CBPFs) were calculated to find the direction of the source contributions (Kim et al., 2003). CBPF analysis is presented here in the form of monthly polar plot (Fig. 4), calculated by linking the fractional factor profiles from PMF to hourly wind speed and direction data (Crilley et al., 2015). Observations with wind speeds <1 m s^{-1} were excluded from the analysis due to low accuracy in determining both speed and direction at such a level.

Factor 1 (Fig. 3a) is primarily characterized by carbon monoxide (CO) and, to a lesser extent, benzene. The CBPF chart (Fig. 4a) highlights significant contributions in the winter months, predominantly from the Eastern sector, where a busy road (S.P. 103) is located, identifying traffic as the main source. A further contribution from the W-NW sector identifies gas flaring and thermal destructors as additional sources of the air pollutants involved. The relatively higher contributions during the winter months suggest the influence of a shallow boundary layer and of the frequent thermal inversions, both typical of the colder period, leading to the accumulation of airborne pollution near the ground.

Factor 2 (Fig. 3b) is associated with O_3, a strong oxidant and a relevant greenhouse gas of secondary origin, and SO_2. This factor features a substantially uniform temporal pattern, despite a relatively higher contribution in the summer months (Fig. 4b). The occurrence of

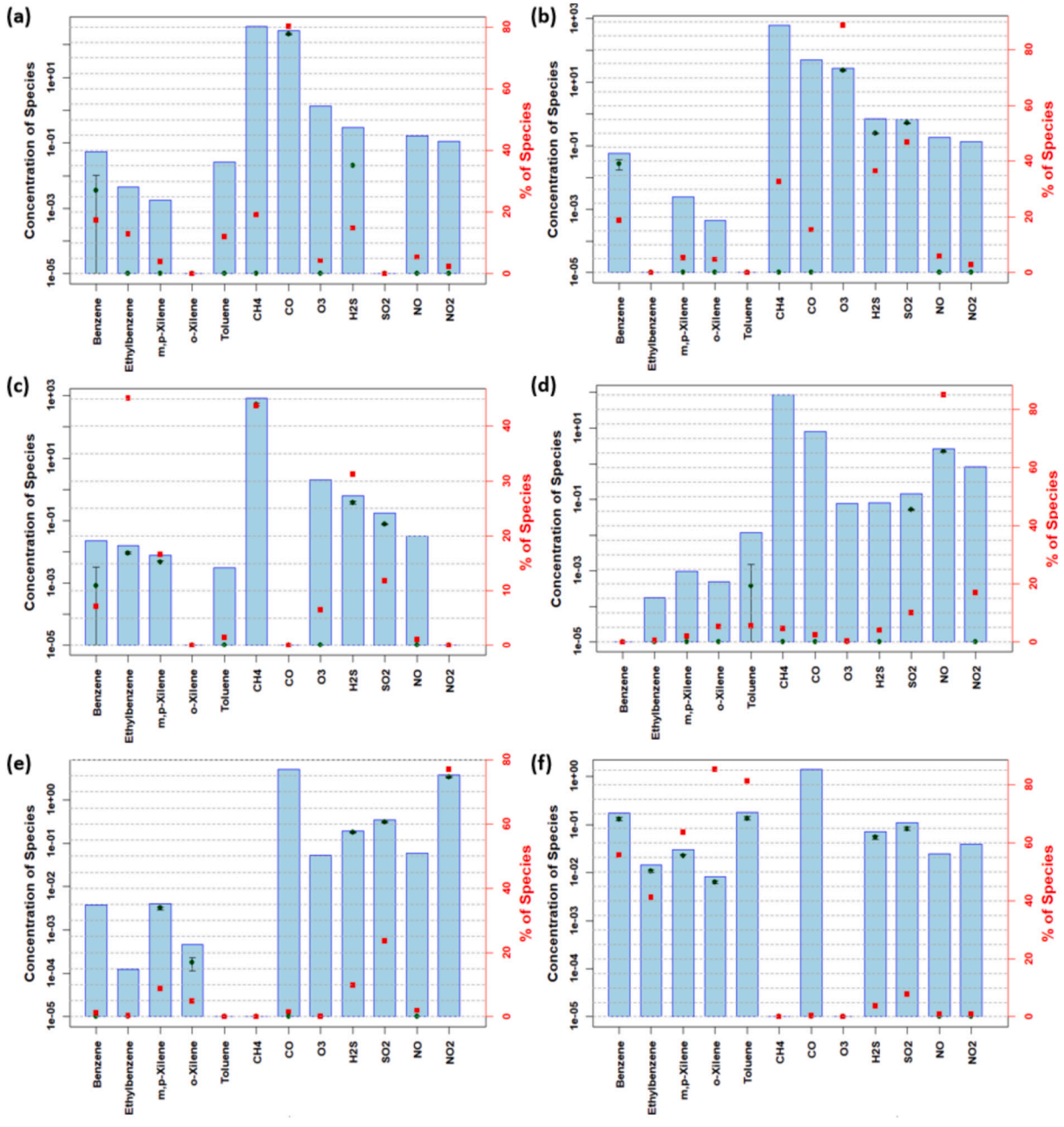


Fig. 3. Factor Profiles from PMF (run 200, Fpeak 0.5) analysis of hourly pollution data.

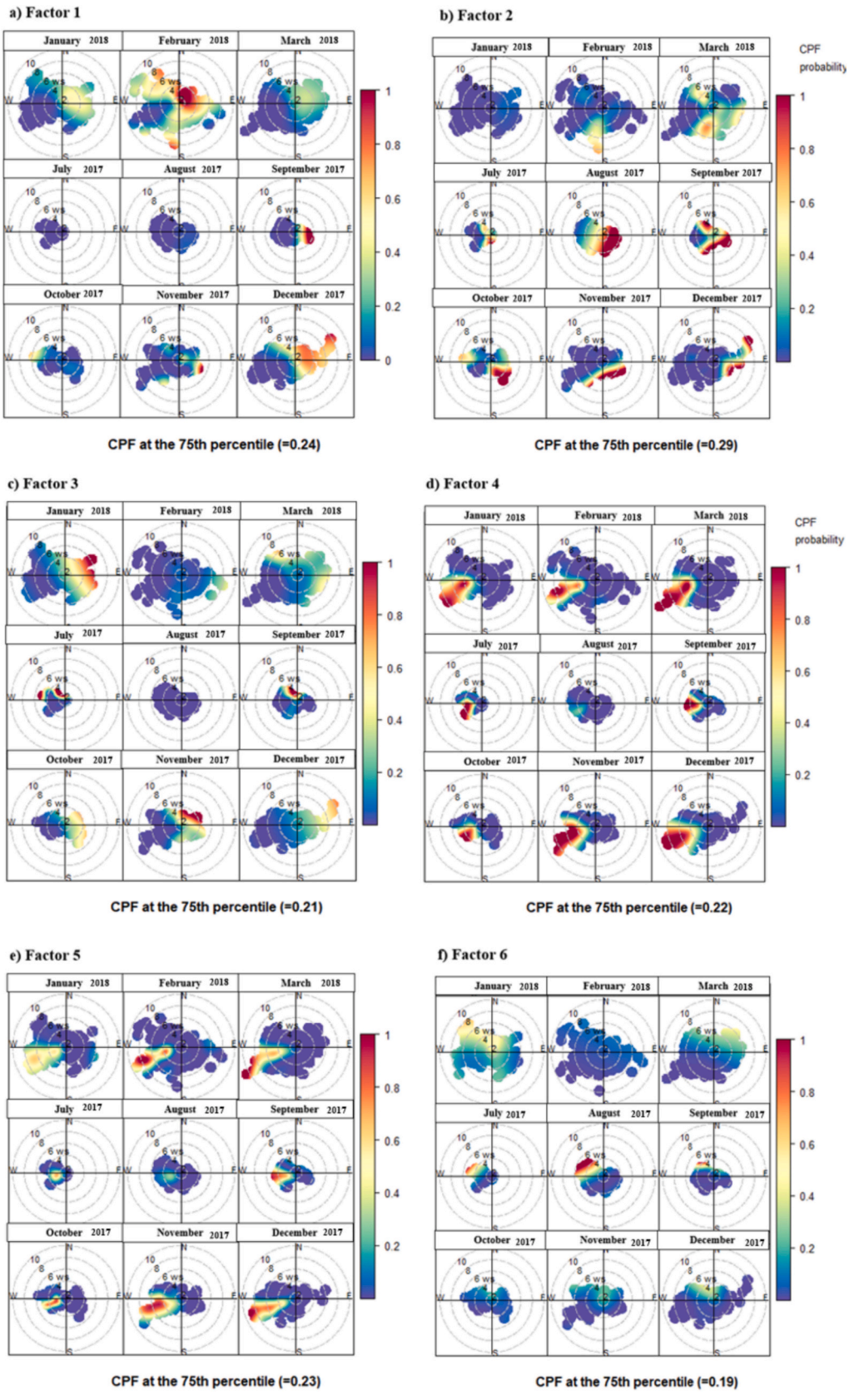


Fig. 4. Hourly CBPF polar plots for the hourly fractional contributions to the six PMF factors.

SO_2 in this factor suggests that this gas is largely of secondary origin in this area, because of the atmospheric oxidation of fugitive H_2S from several potential sources across COVA. The environmental fate of H_2S is rather well-known, with an average tropospheric lifetime of 2 days, decreasing to a few hours in the OH radical-rich conditions, as expected in the Mediterranean region during summer daytime, promoting a fast conversion into SO_2 (Davies et al., 1992; Seinfeld, 2015; Tositti et al., 2022). The likelihood of a secondary origin is consistent with the smooth temporal pattern of the factor analyzed, due to the post-emission conversion of the chemical precursors of ozone (VOCs and NO_x) and H_2S, requiring mixing of the emitted plumes within a weak circulation framework, as observed in this district.

The compositional profile of Factor 3 (Fig. 3c) is primarily associated with reduced species, i.e., CH_4, H_2S, and BTEX, especially ethylbenzene. Methane, with an atmospheric lifetime of about 9–12 years, is more persistent than H_2S and all the other VOCs. It is the main component of natural gases, present in oil-related fugitive emissions, along with H_2S (El Hachem and Kang, 2022) and BTEX compounds such as benzene (atmospheric lifetime: 9.4 d), toluene (1.9 days), xylenes (0.6–0.8 days), and ethylbenzene (1.6 days) (Costa-Gómez et al., 2023; Esswein et al., 2014). The polar plots of the CBPFs (Fig. 4c) indicate a source originating from N-NE, coinciding with one of the operative extraction wells of the COVA oil field (Fig. 1), approximately 200 m far from the receptor site and, possibly, though to a lesser extent, toward a second well (about 800 m east of the sampling point). This suggests that the factor may be linked to fugitive emissions from the extraction wells, whose proximity is not sufficient for a prompt tropospheric oxidation of the species.

Factor 4 (Fig. 3d) relates to nitrogen monoxide (NO) and exhibits significant contributions during the winter months (Fig. 4d), again in connection with the seasonal lower temperature and its influence on the behavior of the boundary layer, as discussed for factor 1. Nitrogen oxides (NO_x, defined as the sum of NO and NO_2) represent a significant fraction of the atmospheric emissions from high-temperature utilities, with NO accounting for 95 % of the NO_x at the source emission (Zhang et al., 2022). The CBPF functions (Fig. 4d) elucidate the relationship between the factor and a W-SW circulation, particularly intense during the winter months. Under these conditions, air masses can thus intercept and upload the emissions from the COVA gas turbine facility. Due to the short distance between the emission source and the receptor site, secondary NO_2 formation through the complexity of tropospheric chemistry is still limited, resulting in its negligible contribution to the factor (Fiore et al., 2024).

Factor 5 (Fig. 3e) is characterized by a high contribution of NO_2 and SO_2. The monthly CBPF (Fig. 4e) highlights significant contributions to the factor, especially from observations belonging to the cold period months, associated with circulation from the SW. This factor could therefore describe secondary emissions from the entire hydrocarbon processing area of COVA (SW of the sampling point). NO_2 is indeed notoriously the main product of the atmospheric photo-oxidation of NO, already indicated in Factor 4 as the most abundant primary nitrogen oxide produced by high-temperature combustion processes. Similarly, SO_2 could be produced as a secondary pollutant from the atmospheric photo-oxidation of H_2S, potentially emitted by various facilities in the COVA network. The winter seasonality of Factor 5 could describe the photo-oxidative chemistry of the primary emissions from COVA, which retains reactive intermediates (NO_2 and SO_2) and concentrates them near the ground, due to thermal stability and a shallow boundary layer, allowing these species to be intercepted at the receptor site.

The profile of Factor 6 (Fig. 3f) is characterized by BTEX (benzene, toluene, ortho, meta, para-xylene, and ethylbenzene). The monthly CBPF polar plots (Fig. 4f) appear scarcely correlated with wind except for weak NW correlations in July, August, and, to a much lesser extent, January. Considering that toluene and benzene arise from various common sources, characterized by the emission of these compounds in distinct ratios, the toluene/benzene concentration ratio (T/B) emerges as a valuable diagnostic tool for identifying emission sources. A previous work (Halliday et al., 2016) have notably emphasized that this ratio can effectively distinguish between fresh traffic emissions (T/B $\geq$ 2) and oil and gas activities (O&NG) in specific regions, which instead typically display lower T/B values.

For this reason, the T/B ratio was computed for the months of July and August, during which the observed values exhibited the highest correlations with the factor. Comprehensive pollution rose for the two months (Fig. 5) was constructed using the ratio values. Fig. 5 shows that, for the studied months, T/B $<$ 2, indicating emissions from O&NG extraction and treatment (Halliday et al., 2016). The most significant contributions come from W-SW, indicating the main sources as the facilities of the COVA network. A smaller contribution also comes from E, where one of the currently active extraction wells is located.

3.3.3. *Self-organizing maps*

Self-Organizing Maps (SOM) were also calculated using air pollution data from the whole sampling campaign, according to the methodology outlined in the experimental section. The SOM was computed with the R package SOMEnv (Licen et al., 2023), using a 24 $\times$ 16 map with 100 computation epochs. After the analysis, each unit contained at least one observation from the dataset.

Subsequently, the results were subjected to a clustering process aimed at creating groups of observations characterized by close multi-pollutant profiles, potentially conducive to their emission sources or of shared environmental processes (Pearce et al., 2014). The clustering computation was tested for two to eight clusters, ultimately selecting a six-cluster solution as the optimal one, based on the Davies-Bouldin index (Licen et al., 2023). The heatmaps reported in Fig. 6a represent the distribution of observations within the individual units of the map for each modeled variable. The content of each unit is expressed in terms of basic statistical parameters. Fig. 6b represents the final SOM map, after clustering computation, where each color identifies a specific cluster.

To evaluate the chemical profile of each cluster, boxplots for the modeled variables for each cluster were calculated (Fig. 7). To do so, data were first auto-scaled (each data point of the original dataset was subtracted by the column mean and divided by the column standard deviation). In this way, zero values in Fig. 7 correspond to the mean of

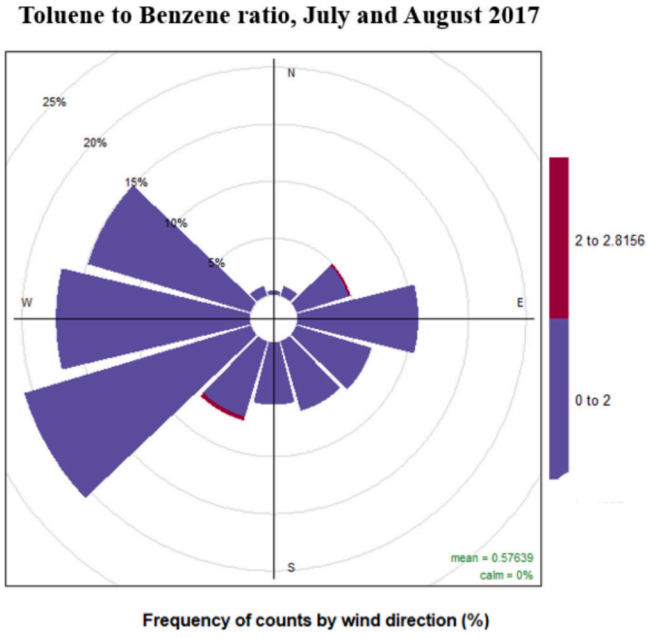


Fig. 5. Pollution rose of the toluene to benzene ratio for the month of July 2017 and August 2017. The colors identify the different ranges (blue $<$ 2, red $>$ 2).

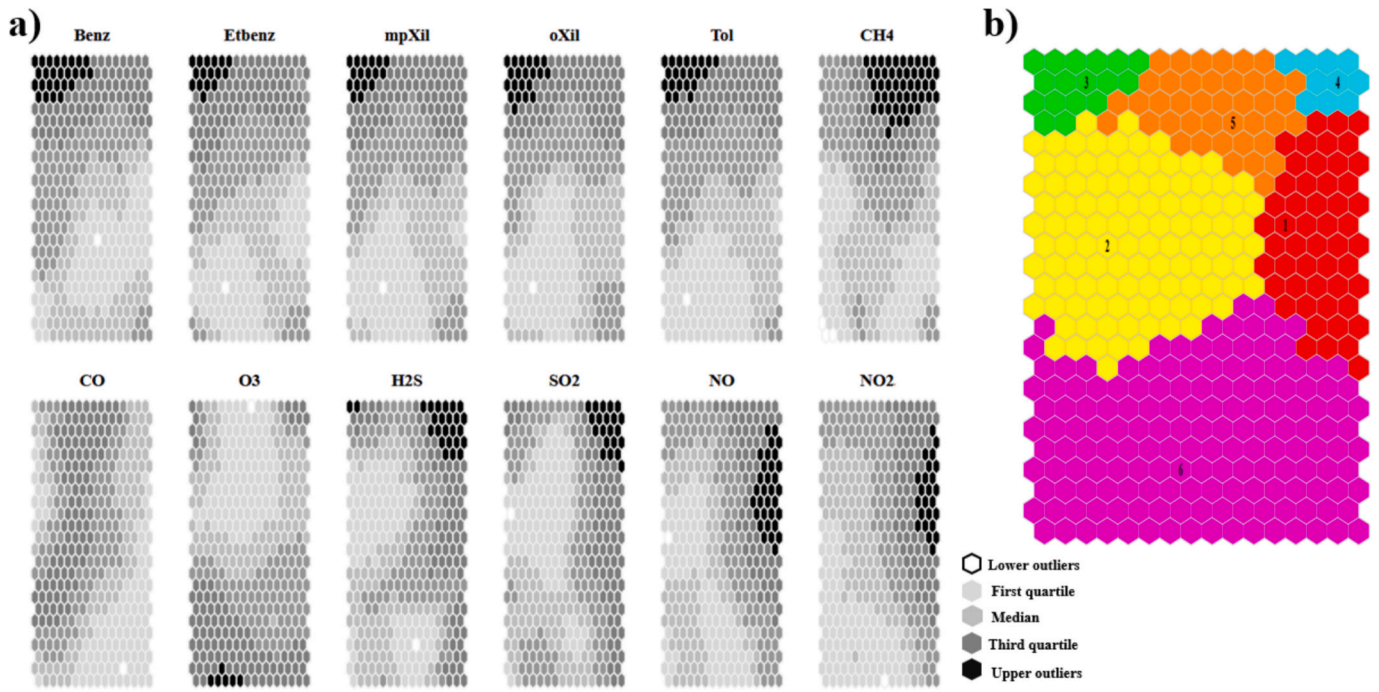


Fig. 6. a) Heatmaps of the quartile distribution of chemical variables; b) SOM map for units' distribution across six clusters, with cluster numbers placed into the cluster-centroid unit.

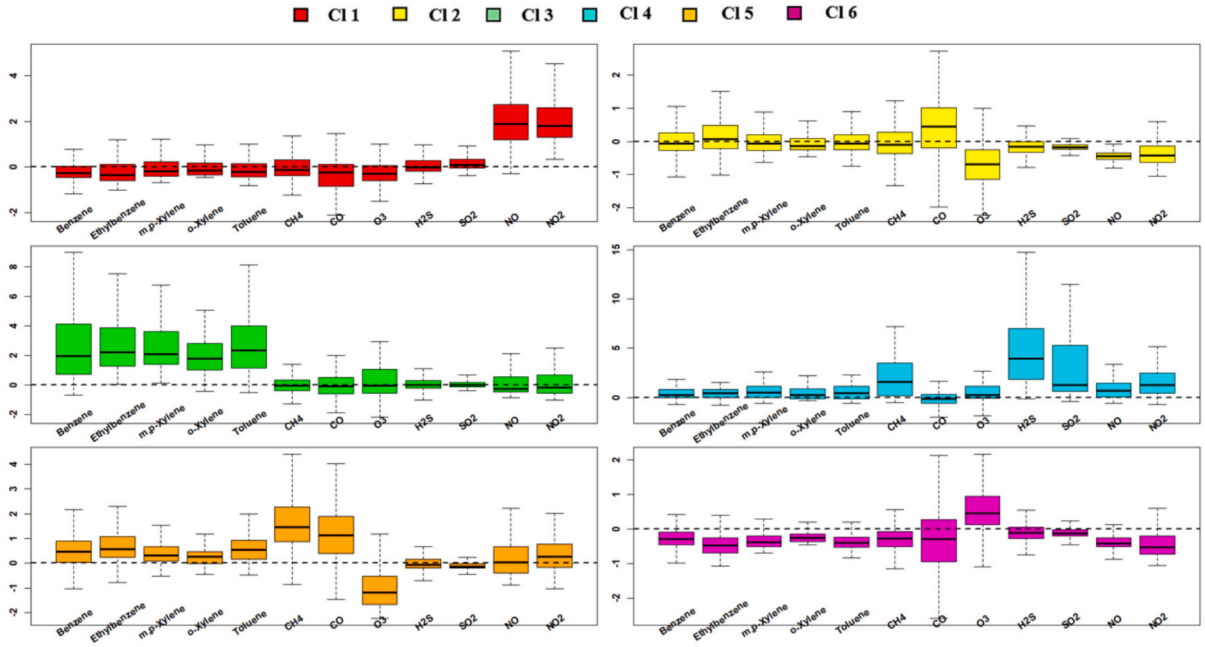


Fig. 7. Boxplots of auto-scaled chemical variables based on the cluster grouping. Colors are based on SOM map as depicted in Fig. 6.

the variable in the whole dataset. Consequently, median values higher than 0 indicate a relative abundance of that species within the respective cluster, while median values lower than 0 indicate a deficiency of that species. Boxplots were also calculated for temperature (°C), pressure (hPa), and humidity (HR%) (Fig. 8a), as well as wind roses (Fig. 8b) for each cluster to evaluate the seasonal and meteorological influences on each cluster. Finally, monthly averaged time variations were computed for the diagnostic concentration ratios between toluene and benzene, H_2S and SO_2, and NO and NO_2.

Cluster 1 includes 670 observations, most of which (460) belonging to the cold season, which, in turn, includes most of the total observations of the dataset. Consequently, for the following clusters, only cases related to the warm period will be highlighted. Observations are characterized by SW circulation (Fig. 8b), identifying nitrogen oxides (NO, NO_2) as significant byproducts of methane combustion in the gas turbines at COVA.

Cluster 2 (1757 observations) seems to reflect the accumulation of carbon monoxide (CO), emitted by all the local combustion sources (Fig. 7), due to limited photochemical activity during the winter months, especially at night (Fishman and Seiler, 1983). In particular, the primary

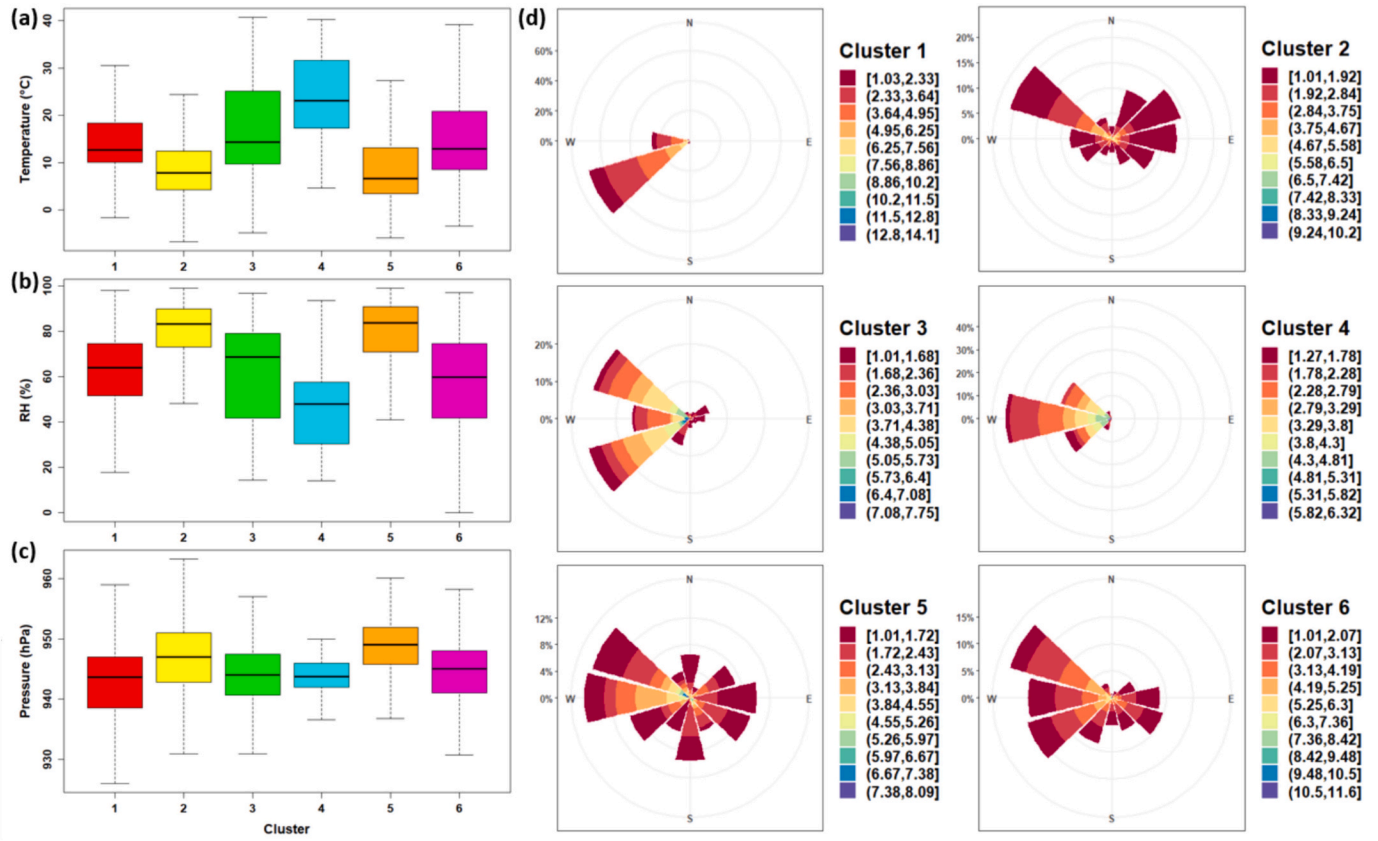


Fig. 8. a) Boxplots of the meteorological variables (temperature, pressure, relative humidity; b) wind roses for each cluster.

source seems to be traffic along S.P.103 (east) as far as the E in component is concerned, followed by thermal oxidizers activities at COVA, situated to the W-NW of the sampling point.

Cluster 3, which includes a limited number of observations (204), is characterized by relatively intense winds (up to approximately 7 m s^{-1}) originating from NW and SW (Fig. 8b). This cluster essentially describes the co-emission and strong correlation of BTEX compounds (Fig. 7). The monthly behavior of the toluene-benzene ratio (Fig. 9), as a function of wind direction (Fig. 8b), highlights that, for this cluster and throughout the examined period, the main source of BTEX is high-temperature O&NG processes (T/B $<$ 2) (Halliday et al., 2016) within the COVA perimeter.

Cluster 4 is the least populated (71 observations). It is characterized by the highest temperatures observed during the campaign (Fig. 8a), mostly occurring in the warm season, i.e. July, August, September, and October (53 observations) and daytime hours (54 observations) with a moderate W circulation (Fig. 8b). The concentration profiles show elevated levels of sulfurate gases (H_2S and SO_2), as well as methane (CH_4) and nitrogen oxides (NO and NO_2) (Fig. 7). The very limited number of associated observations complicates the interpretation of the cluster. The sparse population, along with the compositional profile, suggests a potential origin from flaring activity at the COVA torches, characterized by episodic downward transport events from their high emission point (65 m). Such transport phenomena are difficult to observe at ground level, except under conditions of thermal convective mobility of air masses, such as those promoted by high summer temperatures. Alternatively, the preferential circulation from W, coupled with a compositional profile rich in sulfur gases and characterized by the highest H_2S/SO_2 ratio (Fig. 9), might suggest emissions from the sulfur recovery facility.

Cluster 5 (387 observations) primarily pertains to daytime hours (226 observations). It is characterized by the highest values of RH% (Fig. 8a) and winds that are quite uniform in direction, although the most intense (up to 8 m s^{-1}) come from W and NW (Fig. 8b). The compositional profile indicates elevated concentrations of CH_4, CO, BTEX, and NO_2 (Fig. 7), and the average monthly toluene-benzene ratio (Fig. 9) indicates high-temperature oil and gas treatment processes (T/B $<$ 2; Halliday et al., 2016). All these evidences suggest that this cluster could describe the wintertime accumulation at ground level of pollutants from local high temperature processes (conducted most of all within the COVA perimeter), promoted by high values of RH% (Vaishali et al., 2023).

Cluster 6 is the most populated, comprising 3067 observations, with the majority occurring during daytime hours (1700 observations). Wind speeds are moderate, and the circulation appears to originate from the western area (Fig. 8b). This cluster is relatively the "cleanest", except for a notable abundance of ozone (O_3). The NO/NO_2 ratio is the highest among the clusters (Fig. 9), suggesting a deficiency in NO_2, the principal O_3 precursor (Khoder, 2009; Monks et al., 2015). Therefore, this cluster describes the effects of daytime photochemical oxidation cycles on atmospheric composition.

3.3.4. *Comparison between PMF and SOM source identifications*

The main scope of the present work is to carry out a critical comparison between the results of two distinct multivariate analyses, i.e., PMF and SOM computations, on a high-resolution air quality dataset from an area with low urban emissions, but close to a complex industrial facility constituted by a multiplicity of active emission sources.

Varimax analysis, although giving some suggestions about the correlations between variables, does not provide a sufficient level of detail to compare its results with those of the other methods.

The comparison between PMF and SOM is not straightforward, due to their different approaches, leading, however, to a significantly higher level of detailed information. SOM, indeed, is observations-oriented: its

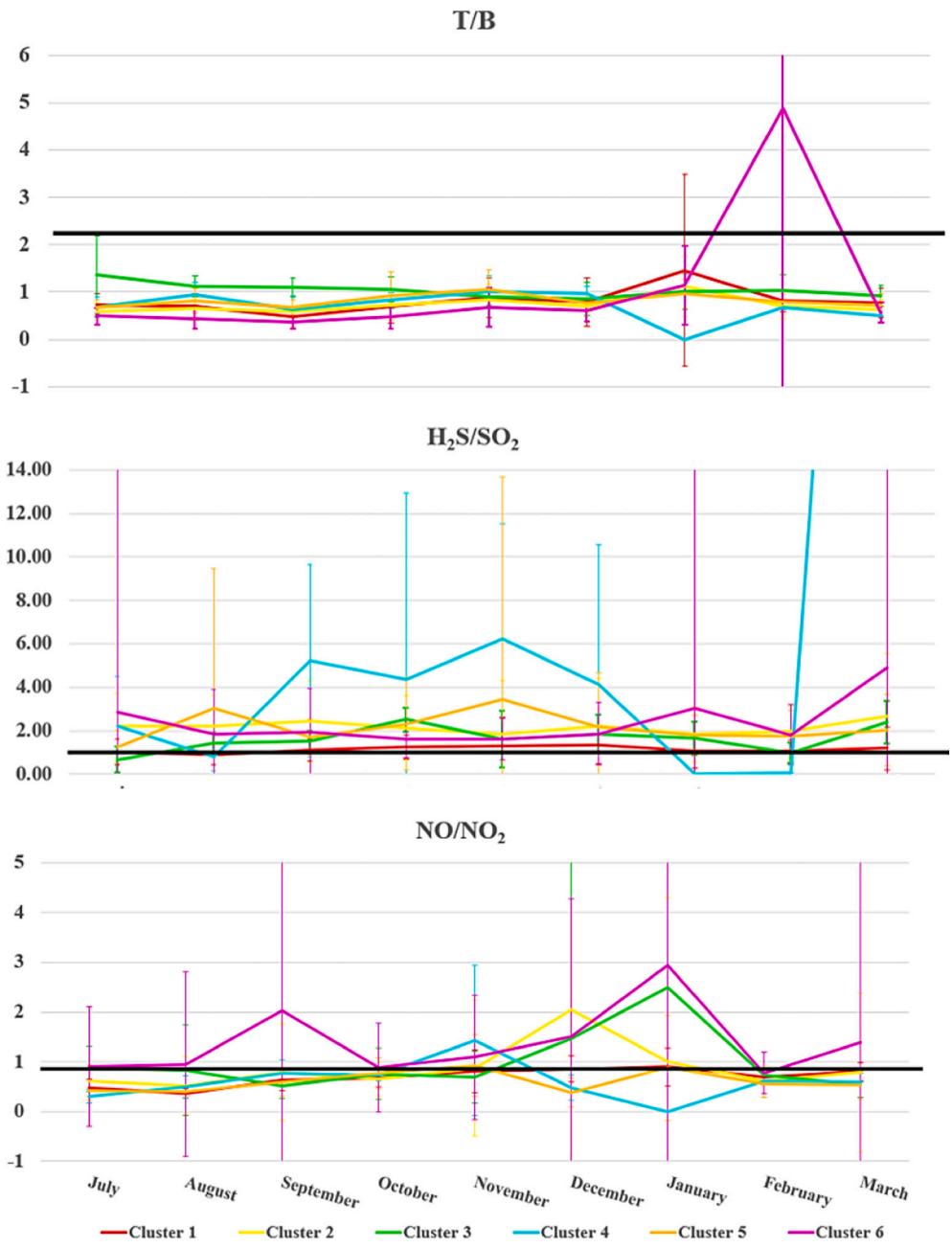


Fig. 9. Monthly time variations for the toluene to benzene, H_2S to SO_2 and NO to NO_2 concentration diagnostic ratios based on the cluster division.

algorithm learns from data, and, after each computation epoch, the model is adapted to best describe the behavior of the observations of the dataset. PMF, instead, is source-oriented: it aims to find the co-emissive profiles of the modeled pollutants, sacrificing the description of each single observation toward a more general behavior of the pollution sources. Such differences in the methods bring results that share some similarities but also some peculiarities for one method compared to the other.

Similarities between the results can be found, for example, for CO, which is described by PMF-F1 and SOM-Cl2 (although PMF shows a partial co-emission with benzene and H_2S), as deriving in both cases from an eastern source, with a minor contribution from NW, mainly during the cold period. In both cases, CO has been associated with traffic and high-temperature operations of the COVA plant. In addition, BTEX species showed a common fingerprint in both models (PMF-F6 and SOM—Cl3) with a westerly origin, mainly associated with O&NG processes of COVA.

O_3 and CH_4 are probably the most controversial species in the reported analysis. Regarding O_3, both models describe most of all its secondary nature, although in a different way. In PMF, it is associated, in F2, with SO_2, another secondary species, in a photochemical source that is stronger during the warm period. Thus, PMF effectively describes the atmospheric patterns of photo-oxidation of VOCs, with formation of peroxy-radicals, and consequent conversion of NO to NO_2 (Fiore et al., 2024), resulting in the formation of O_3 and the parallel photo-oxidation of H_2S to SO_2 (Cox and Sandalls, 1974). SOM, on the other hand, shows an increase in O_3 concentration in Cl6, a peculiar result of the SOM analysis. Indeed, Cl6 is a "clean" or "no-source" cluster, already described in previous studies (Zappi et al., 2024), and is not present among the PMF results due to the absence of a specific source. Weather conditions do not allow this cluster to be attributed to a specific season, but the relative abundance of O_3 enables finding similarities with PMF-

F2 and associating SOM-Cl6 with the aforementioned photo-oxidative patterns. The depletion of all pollutants, except for O_3, in the cluster could also describe the anti-correlation of this species with its precursors such as NO_x and, to a lesser extent, VOCs.

CH_4, instead, is modeled differently in the two cases. Based on PMF—F3, it can be reasonably ascribed to fugitive emissions, probably from a northern extraction well, also characterized by the presence of H_2S and, to some extent, BTEX, particularly ethylbenzene. The closest SOM cluster, based on chemical fingerprint, is Cl4, which, however, describes a different and very specific situation. It is the less populated cluster, with mainly warm-season characteristics and westerly origin, probably associated with seasonal emissions from the COVA sulfur recovery unit, or maybe with highly episodic emissions from gas flaring activities, detectable at ground level during summer because of convective motion of air masses with downward transport of pollutants emitted at high altitudes, promoted by high temperatures. Also, SOM-Cl5 shows a higher concentration of CH_4, but with a different chemical fingerprint (besides BTEX, also CO and NO_2 are present) and a more generic origin, probably associated with all the local high-temperature processes. In this case, perhaps due to the low-frequency nature of these occurrences and the lower number of warm-season cases, SOM fails in identifying fugitive emissions, while PMF suggests their presence and also their origin from W-NW from the sampling point. Conversely, SOM can effectively describe the role of CH_4 in the chemical fingerprint of seasonal emissions from highly specific facilities, which is only evident in a limited group of observations.

Finally, nitrogen oxides deserve further discussion. It is well-known that combustion can be considered the common source of NO and NO_2, although the primary product of combustion is NO that is rapidly converted into NO_2 (Fiore et al., 2024; Seinfeld, 2015). The relatively low lifetime of NO generally produces a high correlation between these two species, as is the case with our SOM model, which describe them together in Cl1 as originating from methane combustion by COVA. PMF, however, in this case seems to be able to enhance the description of these two species, by dividing them into two distinct factors: NO in F4, with only a slight contribution of NO_2, and NO_2 in F5, with a contribution of SO_2. Contributions of F4 are weaker in the warm period, showing some peaks during the cold one, while those of F5 are almost constant in both warm and cold periods (Fig. 4e). In comparison, SOM-Cl1 is not affected by seasonality. These observations support the hypothesis that PMF is able to distinguish between primary and secondary emissions from COVA high-temperature facilities.

In conclusion, both models have shown strong and weak points, reflecting the different approaches to data analysis on which they are based. The PMF, indeed, showed its reliability in apportioning pollution sources. Although the considered species are criteria pollutants, with no mass closure, PMF was able to discriminate primary and secondary sources and point sources due to fugitive emissions from COVA. However, the attribution to the specific sources of each factor with PMF is mostly based on the chemical fingerprint, because the spatial location of the source can be estimated only by the CBPF probability of wind pattern, and the temporal trend can be estimated only by the factor contributions. On the other side, the attribution of each observation to a specific cluster carried out by SOM allows for better identification of the spatial allocation and temporal behavior of each identified source, using the meteorological parameters to optimize the clusters description. However, the identified sources, due to the absence of uncertainties in the computation, may be less specific than those obtained by PMF.

We believe that the synergistic use of PMF and SOM methods, when possible, would be of great benefit to environmental studies. Indeed, when the two methods agree about a certain source, their behavior can be fully disclosed and described, using the strengths of both. When, instead, the two methods seem to be in contrast, they are likely describing the same phenomena from different points of view.

4. Conclusions

Nine months of air quality data were collected in the Agri Valley, a semi-rural area in southern Italy, slated to host Europe's largest onshore fossil fuel extraction and initial processing facility, known as COVA. The dataset reported here covers hourly monitoring of concentrations of twelve different gaseous pollutants (benzene, ethylbenzene, meta-, para-, ortho-xylene, toluene, CH_4, CO, O_3, H_2S, SO_2, NO, NO_2), including highly specific markers of oil and gas processing processes, as well as meteorological parameters. During the study period, hourly concentrations never exceeded regulatory limits.

To profile the local emission sources of the monitored pollutants, several multivariate statistical techniques were applied, allowing for a comparison of the results obtained. Varimax PCA, a widely used multivariate tool, provided an initial indication of the correlations between the studied variables. However, it was unable to identify potential local sources of the pollutants studied. To this end, the PMF and SOM methods were employed. These methods, thanks to their greater computational power, not only proved to be more efficient in improving the discrimination potential and the level of detail of the latent information in the dataset, but also demonstrated a good degree of concordance, describing some of the main sources and local atmospheric mechanisms similarly. However, the different computational approaches of the two methods led to the observation of some discrepancies. Thanks to its unique assignment of observations to clusters, SOM effectively clarifies the complex relationships between pollutant concentrations and spatial and meteorological variables, as well as episodic and/or complex temporal trends. Conversely, PMF proved to be a useful method for reliable source apportionment, particularly thanks to the integration of data uncertainties, which allows for the distinction between primary and secondary emissions and the identification of low-frequency emission sources. Overall, both methods have strengths and weaknesses, demonstrating the ability to provide a complete descriptive picture of the local atmosphere and associated phenomena only when used synergistically.

CRediT authorship contribution statement

Mariassunta Biondi: Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **Alessandro Zappi:** Writing – original draft, Validation, Supervision, Software, Formal analysis, Data curation. **Erika Brattich:** Writing – review & editing, Visualization, Validation, Supervision, Conceptualization. **Serena Sabia:** Writing – review & editing, Data curation, Conceptualization. **Rosa Caggiano:** Writing – review & editing, Resources, Project administration, Data curation, Conceptualization. **Laura Tositti:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare no competing interests.

Acknowledgements

We are thankful to the Agenzia Regionale per la Protezione dell'Ambiente della Basilicata (ARPAB, https://www.arpab.it/) for providing the data.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2025.180983.

Data availability

Data will be made available on request.

References

Abdi, H., Williams, L.J., 2010. Principal component analysis. WIREs Comput. Stat. 2, 433–459. https://doi.org/10.1002/wics.101.

Belis, C.A., Favez, O., Mircea, M., Diapouli, E., Manousakas, M.-I., Vratolis, S., Gilardoni, S., Paglione, M., Decesari, S., Mocnik, G., Mooibroek, D., Salvador, P., Takahama, S., Vecchi, R., Paatero, P., 2019. European Guide on Air Pollution Source Apportionment with Receptor Models: Revised Version 2019. Publications Office of the European Union.

Buzcu-Guven, B., Harriss, R., Hertzmark, D., 2010. Gas Flaring and Venting: Extent, Impacts, and Remedies Energy Market Consequences of an EmergIng U.S. Carbon Management PolIcy.

Calvello, M., Caggiano, R., Esposito, F., Lettino, A., Sabia, S., Summa, V., Pavese, G., 2017. IMAA (integrated measurements of aerosol in Agri valley) campaign: multi-instrumental observations at the largest European oil/gas pre-treatment plant area. Atmos. Environ. 169, 297–306. https://doi.org/10.1016/j.atmosenv.2017.09.026.

Castagna, J., Calvello, M., Esposito, F., Pavese, G., 2019. Analysis of equivalent black carbon multi-year data at an oil pre-treatment plant: integration with satellite data to identify black carbon transboundary sources. Remote Sens. Environ. 235, 111429. https://doi.org/10.1016/j.rse.2019.111429.

Chen, Z., Xie, Y., Liu, J., Shen, L., Cheng, X., Han, H., Yang, M., Shen, Y., Zhao, T., Hu, J., 2023. Distinct seasonality in vertical variations of tropospheric ozone over coastal regions of southern China. Sci. Total Environ. 874, 162423. https://doi.org/10.1016/j.scitotenv.2023.162423.

Comero, S., Capitani, L., Gawlik, B.M., 2009. Positive matrix factorisation (PMF): an introduction to the chemometric evaluation of environmental monitoring data using PMF. Publications Office. doi/. https://doi.org/10.2788/2497.

Costa-Gómez, I., Caracena, A.B., Durán-Amor, M., Banon, D., 2023. BTEX proportions as an indicator of benzene hotspots and dispersion tends in cities where sea and land breezes dominate. Air Qual. Atmos. Health 16, 733–744. https://doi.org/10.1007/s11869-023-01306-3.

Cox, R.A., Sandalls, F.J., 1974. The photo-oxidation of hydrogen sulphide and dimethyl sulphide in air. Atmospheric Environment (1967) 8, 1269–1281. https://doi.org/10.1016/0004-6981(74)90006-7.

Crilley, L.R., Bloss, W.J., Yin, J., Beddows, D.C.S., Harrison, R.M., Allan, J.D., Young, D. E., Flynn, M., Williams, P., Zotter, P., Prevot, A.S.H., Heal, M.R., Barlow, J.F., Halios, C.H., Lee, J.D., Szidat, S., Mohr, C., 2015. Sources and contributions of wood smoke during winter in London: assessing local and regional influences. Atmos. Chem. Phys. 15, 3149–3171. https://doi.org/10.5194/acp-15-3149-2015.

Ćurić, M., Zafirovski, O., Spiridonov, V., 2022. Air quality and health. In: Ćurić, M., Zafirovski, O., Spiridonov, V. (Eds.), Essentials of Medical Meteorology. Springer International Publishing, Cham, pp. 143–182. https://doi.org/10.1007/978-3-030-80975-1_8.

Davies, T.D., Kelly, P.M., Low, P.S., Pierce, C.E., 1992. Surface ozone concentrations in Europe: links with the regional-scale atmospheric circulation. J. Geophys. Res. Atmos. 97, 9819–9832. https://doi.org/10.1029/92JD00419.

de Castro, B.P., de Souza Machado, G., Bauerfeldt, G.F., Nunes Fortes, J.D., Martins, E. M., 2015. Assessment of the BTEX concentrations and reactivity in a confined parking area in Rio de Janeiro, Brazil. Atmos. Environ. 104, 22–26. https://doi.org/10.1016/j.atmosenv.2015.01.013.

Di Gilio, A., Palmisani, J., Petraccone, S., de Gennaro, G., 2021. A sensing network involving citizens for high spatio-temporal resolution monitoring of fugitive emissions from a petroleum pre-treatment plant. Sci. Total Environ. 791. https://doi.org/10.1016/j.scitotenv.2021.148135.

El Hachem, K., Kang, M., 2022. Methane and hydrogen sulfide emissions from abandoned, active, and marginally producing oil and gas wells in Ontario. Canada. Science of The Total Environment 823, 153491. https://doi.org/10.1016/j.scitotenv.2022.153491.

Esswein, E.J., John, S., Bradley, K., Michael, B., Marissa, A.-S., Kiefer, M., 2014. Evaluation of some potential chemical exposure risks during Flowback operations in unconventional oil and gas extraction: preliminary results. J. Occup. Environ. Hyg. 11, D174–D184. https://doi.org/10.1080/15459624.2014.933960.

European Parliament, 2008. Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe. OJ L 152, 1–44, 11.6.2008.

Fiore, A.M., Mickley, L.J., Zhu, Q., Baublitz, C.B., 2024. Climate and tropospheric oxidizing capacity. Annu. Rev. Earth Planet. Sci. 52, 321–349. https://doi.org/10.1146/annurev-earth-032320-090307.

Fishman, J., Seiler, W., 1983. Correlative nature of ozone and carbon monoxide in the troposphere: implications for the tropospheric ozone budget. J. Geophys. Res. 88, 3662–3670. https://doi.org/10.1029/JC088iC06p03662.

Giunta, G., Ceppi, A., Salerno, R., 2022. Local-scale weather forecasts over a complex terrain in an early warning framework: performance analysis for the Val D'agri (southern Italy) case study. Adv. Meteorol. 2022, 2179246. https://doi.org/10.1155/2022/2179246.

Grant, A., Archibald, A.T., Cooke, M.C., Shallcross, D.E., 2010. Modelling the oxidation of seventeen volatile organic compounds to track yields of CO and CO2. Atmos. Environ. 44, 3797–3804. https://doi.org/10.1016/j.atmosenv.2010.06.049.

Halliday, H.S., Thompson, A.M., Wisthaler, A., Blake, D.R., Hornbrook, R.S., Mikoviny, T., Müller, M., Eichler, P., Apel, E.C., Hills, A.J., 2016. Atmospheric benzene observations from oil and gas production in the Denver-Julesburg Basin in July and August 2014. J. Geophys. Res. Atmos. 121 (11), 11–55, 74. https://doi.org/10.1002/2016JD025327.

Hentati, A., Kawamura, A., Amaguchi, H., Iseri, Y., 2010. Evaluation of sedimentation vulnerability at small hillside reservoirs in the semi-arid region of Tunisia using the self-organizing map. Geomorphology 122, 56–64. https://doi.org/10.1016/j.geomorph.2010.05.013.

Huang, J., Wu, Y., Sun, J., Li, X., Geng, X., Zhao, M., Sun, T., Fan, Z., 2021. Health risk assessment of heavy metal(loid)s in park soils of the largest megacity in China by using Monte Carlo simulation coupled with positive matrix factorization model. J. Hazard. Mater. 415, 125629. https://doi.org/10.1016/j.jhazmat.2021.125629.

Ito, K., Xue, N., Thurston, G., 2004. Spatial variation of PM2.5 chemical species and source-apportioned mass concentrations in new York City. Atmos. Environ. 38, 5269–5282. https://doi.org/10.1016/j.atmosenv.2004.02.063.

Kenagy, H.S., Sparks, T.L., Ebben, C.J., Wooldrige, P.J., Lopez-Hilfiker, F.D., Lee, B.H., Thornton, J.A., McDuffie, E.E., Fibiger, D.L., Brown, S.S., Montzka, D.D., Weinheimer, A.J., Schroder, J.C., Campuzano-Jost, P., Day, D.A., Jimenez, J.L., Dibb, J.E., Campos, T., Shah, V., Jaeglé, L., Cohen, R.C., 2018. NOx lifetime and NOy partitioning during WINTER. J. Geophys. Res. Atmos. 123, 9813–9827. https://doi.org/10.1029/2018JD028736.

Khoder, M.I., 2009. Diurnal, seasonal and weekdays–weekends variations of ground level ozone concentrations in an urban area in greater Cairo. Environ. Monit. Assess. 149, 349–362. https://doi.org/10.1007/s10661-008-0208-7.

Kim, E., Hopke, P.K., Edgerton, E.S., 2003. Source identification of atlanta aerosol by positive matrix factorization. J. Air Waste Manage. Assoc. 53, 731–739. https://doi.org/10.1080/10473289.2003.10466209.

Kohonen, T., 1990. The self-organizing map. Proc. IEEE 78, 1464–1480. https://doi.org/10.1109/5.58325.

Leardi, R., Melzi, C., Polotti, G., 2015. CAT (Chemometric Agile Tool). freely downloadable from. http://gruppochemiometria.it/index.php/software (last accessed on September 30, 2025).

Licen, S., Astel, A., Tsakovski, S., 2023. Self-organizing map algorithm for assessing spatial and temporal patterns of pollutants in environmental compartments: a review. Sci. Total Environ. 878, 163084. https://doi.org/10.1016/j.scitotenv.2023.163084.

Lovett, G.M., Tear, T.H., Evers, D.C., Findlay, S.E.G., Cosby, B.J., Dunscomb, J.K., Driscoll, C.T., Weathers, K.C., 2009. Effects of air pollution on ecosystems and biological diversity in the eastern United States. Ann. N. Y. Acad. Sci. 1162, 99–135. https://doi.org/10.1111/j.1749-6632.2009.04153.x.

Majra, J.P., 2011. Air quality in rural areas. In: Mazzeo, N. (Ed.), Chemistry, Emission Control, Radioactive Pollution and Indoor Air Quality. IntechOpen, Rijeka. https://doi.org/10.5772/16890 p. Ch. 23.

Manisalidis, I., Stavropoulou, E., Stavropoulos, A., Bezirtzoglou, E., 2020. Environmental and health impacts of air pollution: a review. Front. Public Health. https://doi.org/10.3389/fpubh.2020.00014.

Martins, M.C.H., Fatigati, F.L., Véspoli, T.C., Martins, L.C., Pereira, L.A.A., Martins, M.A., Saldiva, P.H.N., Braga, A.L.F., 2004. Influence of socioeconomic conditions on air pollution adverse health effects in elderly people: an analysis of six regions in São Paulo, Brazil. J. Epidemiol. Community Health 58 (1978), 41. https://doi.org/10.1136/jech.58.1.41.

Miljković, D., 2017. Brief review of self-organizing maps, in: 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1061–1066. https://doi.org/10.23919/MIPRO.2017.7973581.

Mininni, M., 2015. How to manage conflicts between resources' exploitation and identity values. In: Gambino, R., Peano, A. (Eds.), Nature Policies and Landscape Policies: Towards an Alliance. Springer International Publishing, Cham, pp. 469–477. https://doi.org/10.1007/978-3-319-05410-0_54.

Monks, P.S., Archibald, A.T., Colette, A., Cooper, O., Coyle, M., Derwent, R., Fowler, D., Granier, C., Law, K.S., Mills, G.E., Stevenson, D.S., Tarasova, O., Thouret, V., von Schneidemesser, E., Sommariva, R., Wild, O., Williams, M.L., 2015. Tropospheric ozone and its precursors from the urban to the global scale from air quality to short-lived climate forcer. Atmos. Chem. Phys. 15, 8889–8973. https://doi.org/10.5194/acp-15-8889-2015.

Nakagawa, K., Yu, Z.-Q., Berndtsson, R., Hosono, T., 2020. Temporal characteristics of groundwater chemistry affected by the 2016 Kumamoto earthquake using self-organizing maps. J. Hydrol. (Amst.). 582, 124519. https://doi.org/10.1016/j.jhydrol.2019.124519.

Paatero, P., Tapper, U., 1994. 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.

Pavese, G., Calvello, M., Esposito, F., 2012. Black carbon and organic components in the atmosphere of southern Italy: comparing emissions from different sources and production processes of carbonaceous particles. Aerosol Air Qual. Res. 12, 1146–1156. https://doi.org/10.4209/aaqr.2011.12.0236.

Pearce, J.L., Waller, L.A., Chang, H.H., Klein, M., Mulholland, J.A., Sarnat, J.A., Sarnat, S.E., Strickland, M.J., Tolbert, P.E., 2014. Using self-organizing maps to develop ambient air quality classifications: a time series example. Environ. Health 13, 56. https://doi.org/10.1186/1476-069X-13-56.

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

Seinfeld, J.H., 2015. Atmospheric chemistry. In: reference module in chemistry, Molecular Sciences and Chemical Engineering. Elsevier. https://doi.org/10.1016/B978-0-12-409547-2.11626-9.

Sicard, P., Agathokleous, E., Anenberg, S.C., De Marco, A., Paoletti, E., Calatayud, V., 2023. Trends in urban air pollution over the last two decades: a global perspective. Sci. Total Environ. 858, 160064. https://doi.org/10.1016/j.scitotenv.2022.160064.

Tositti, L., Morozzi, P., Brattich, E., Zappi, A., Calvello, M., Esposito, F., Lettino, A., Pavese, G., Sabia, S., Speranza, A., Summa, V., Caggiano, R., 2022. Apportioning PM1 in a contrasting receptor site in the Mediterranean region: aerosol sources with an updated sulfur speciation. Sci. Total Environ. 851. https://doi.org/10.1016/j.scitotenv.2022.158127.

Vaishali, Verma, G., Das, R.M., 2023. Influence of temperature and relative humidity on PM2.5 concentration over Delhi. Mapan - Journal of Metrology Society of India 38, 759–769. https://doi.org/10.1007/s12647-023-00656-8.

Zappi, A., Brattich, E., Biondi, M., Tositti, L., 2024. How to use efficiently airborne criteria pollutants and radon-222 in source apportionment: a self-organizing maps approach. Chemosphere 367, 143619. https://doi.org/10.1016/j.chemosphere.2024.143619.

Zhang, X.-C., Sha, Q.-E., Lu, M.-H., Wang, Y.-Z., Rao, S.-J., Ming, G.-Y., Li, Q.-Q., Wu, S.-Z., Zheng, J.-Y., 2022. Volatile organic compound emission characteristics and influences assessment of a petrochemical Industrial Park in the Pearl River Delta region. Huan Jing Ke Xue 43, 1766–1776. https://doi.org/10.13227/j.hjkx.202107184.