

Article

Disentangling Multiannual Air Quality Profiles Aided by Self-Organizing Map and Positive Matrix Factorization

Stefano Fornasaro ¹, Aleksander Astel ^{2,*}, Pierluigi Barbieri ¹ and Sabina Licen ^{1,*}

¹ Department of Chemical and Pharmaceutical Sciences, University of Trieste, Via Giorgieri 1, 34127 Trieste, Italy; sfornasaro@units.it (S.F.); barbierp@units.it (P.B.)

² Department of Environmental Chemistry and Toxicology, Pomeranian University in Słupsk, 22a Arciszewskiego Str., 76-200 Słupsk, Poland

* Correspondence: aleksander.astel@upsl.edu.pl (A.A.); slicen@units.it (S.L.); Tel.: +39-040-558-3942 (S.L.)

Abstract: The evaluation of air pollution is a critical concern due to its potential severe impacts on human health. Currently, vast quantities of data are collected at high frequencies, and researchers must navigate multiannual, multisite datasets trying to identify possible pollutant sources while addressing the presence of noise and sparse missing data. To address this challenge, multivariate data analysis is widely used with an increasing interest in neural networks and deep learning networks along with well-established chemometrics methods and receptor models. Here, we report a combined approach involving the Self-Organizing Map (SOM) algorithm, Hierarchical Clustering Analysis (HCA), and Positive Matrix Factorization (PMF) to disentangle multiannual, multisite data in a single elaboration without previously separating the sites and years. The approach proved to be valid, allowing us to detect the site peculiarities in terms of pollutant sources, the variation in pollutant profiles during years and the outliers, affording a reliable interpretation.

Keywords: pollution; ambient air; particulate matter; NO_x; multivariate analysis; self-organizing map; hierarchical clustering; positive matrix factorization; COVID-19



Academic Editors: Qian Zhang, Hongmei Xu, Tian Chang and Peng Wei

Received: 6 January 2025

Revised: 31 January 2025

Accepted: 13 February 2025

Published: 14 February 2025

Citation: Fornasaro, S.; Astel, A.; Barbieri, P.; Licen, S. Disentangling Multiannual Air Quality Profiles Aided by Self-Organizing Map and Positive Matrix Factorization. *Toxics* **2025**, *13*, 137. <https://doi.org/10.3390/toxics13020137>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Air pollution assessment is a fundamental issue because air quality can have serious consequences for human health [1–3]. Nowadays, a huge amount of data are recorded at high frequency and, when trying to extract useful information, the researcher needs to cope with multiannual, multisite data as well as the possible presence of noise and sparse missing data [4,5].

Multivariate data analysis is widely used to face this challenge: beside well-established chemometrics such as Principal Component Analysis (PCA), Hierarchical Clustering Analysis (HCA), and k-means clustering (KM), there is a growing interest in neural networks and deep learning networks used for both analysis and prediction [5,6]. In this context, the Self-Organizing map algorithm has been successfully used for unsupervised analysis of large datasets [7–10]. The SOM algorithm is able to deal with large datasets and non-linear relationships among the variables and it is not significantly affected by outliers [11–13].

Another aim of air quality studies is to identify possible pollutant sources. In this context, the so-called receptor models are used [14,15]. Among them, Positive Matrix Factorization is widely used [16]. Nevertheless, there are few studies that use both SOM and PMF for environmental quality characterization, and they are mainly focused on water [17,18] and soil/sediment matrices [19,20].

We found only two papers using this multivariate analysis combination for assessing air quality [21,22]. In both of them, the algorithms were separately applied to the dataset and then the results were compared and combined for assessing the conclusions.

In this study, we wanted to take advantage of SOM's capability to extract recurrent variable profiles from the experimental dataset, which usually are one or two orders of magnitude less than those of the input dataset, to obtain a reduced "cleaner" dataset to be used for PMF input. In fact, the reduced dataset obtained by SOM contains less noise and "smoothed" outliers; thus, it is more suitable for further investigation than the original dataset. To the best of our knowledge this is the first time that the abovementioned method has been used to assess air quality and identify possible pollution sources and outliers, such as desert dust events.

To prove the capabilities of this method, we chose a multisite, multiannual dataset containing data from the year 2020, when the COVID-19 pandemic forced governments worldwide to subject their populations to lockdown periods [23]. Several studies have been conducted worldwide to assess the impact of lockdown on air quality. Most of the studies showed that substantial reductions in NO₂ and NO_x were associated with reduced mobility, and thus with reduced amounts of traffic. PM₁₀ and PM_{2.5} showed a reduction but with complex signals; more significant reductions were detected in megacities, whereas the decrease was less evident in suburban and rural sites. In some cases, the PM concentration increased [24,25].

2. Materials and Methods

2.1. Dataset

The hourly data elaborated in this study were collected in the two main cities in the Friuli Venezia-Giulia region (North-East of Italy). One of the cities is by the sea (Trieste—45°39'01" N 13°46'13" E, 200,000 inhabitants) and one is inland (Udine—46°04' N 13°14' E, 100,000 inhabitants).

The data were retrieved from the Environmental Protection Agency of Friuli Venezia-Giulia Region—Italy (ARPA-FVG) website.

We chose the data collected by monitoring stations classified as "urban traffic" and "city background" stations; thus, four monitoring stations were considered, and they were named according to the city (1 = Trieste, 2 = Udine) and the type (A = "traffic", B = "background"). The following pollutants were considered: benzene (Ben), toluene (Tol), PM₁₀, NO, and NO₂. The chosen periods were from 9th of March to 3rd of May from 2018 to 2023.

2.2. Data Analysis Method

The data were elaborated according to the method shown in Figure 1. First, the SOM algorithm was applied to the dataset, in which each sample is a vector containing a value for each considered variable. The SOM algorithm is a neural network that works in an unsupervised way and it is composed only of the input layer and the output layer, with no hidden layers. The output layer is formed by a list of vectors containing a value for each modeled variable, and each vector is called "node" or "neuron". Usually, there are ten to one thousand times fewer nodes in proportion to the number of samples, but the nodes can still be used to represent the variability of the input layer. The node values are arranged in a matrix called "codebook". Each "node" can be represented by a hexagon in a 2D map in which similar nodes are depicted close one to each other, and the evaluation of the multivariate distance between nodes allows us to identify possible clusters [26,27]. Each node represents (i.e., "is similar" in terms of multidimensional Euclidean distance) one or more experimental vectors belonging to the data matrix. Moreover, the way in which the

algorithm works reduces the noise of the data [28,29]; this feature is particularly useful when handling instrumental outputs, as was the case in the described study.

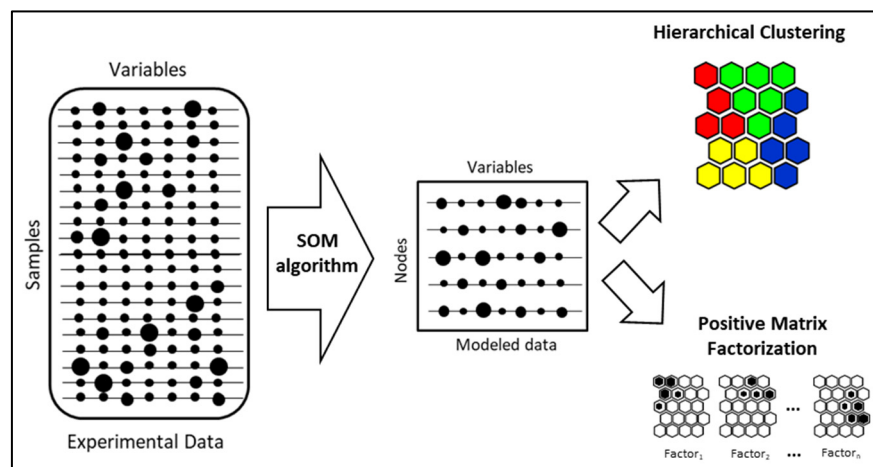


Figure 1. Scheme of data analysis method.

In this specific application, the codebook represents recurrent air quality profiles recorded at the sampling sites. The second step was to cluster the nodes applying a hierarchical clustering algorithm using the Euclidean distance and the Ward's linkage method. In many cases, a clustering method is applied to the codebook [30]. The clustering allows us to obtain “macro-groups” of nodes with similar characteristics, and the benefit of the use of SOM is that the obtained clusters can be depicted on the SOM. The centroid cluster matrix can be thus considered to represent “air quality types” (e.g., “low polluted”, “medium polluted”, highly polluted”, “background”, ...).

The centroid matrix can also give some indication about the possible pollution sources, but sometimes the sources can be mixed and there can be no clear interpretation.

The PMF algorithm is a multivariate receptor model that is widely used for identifying source contributions using data collected at the receptor sites [16,31]. We applied the PMF algorithm on the codebook to disentangle the different source contributions, and, by finding the nodes which made greater contributions to the PMF factors, we were able to provide a more complete interpretation of the “air quality types”.

The R software environment was used both for the dataset preparation and the statistical analysis. The data were elaborated using the *SOMenv* package for the R software environment [32]. The package is based on the *kohonen* package [33,34] for SOM analysis and the *openair* package [35] for managing date/time recordings. The missing data were filled in using the *mdatools* package [36]. The bidimensional HCA plots were obtained using the *pheatmap* package [37]. The EPA PMF (version 5.0 <https://www.epa.gov/air-research>, accessed on 30 November 2024) was used to perform Positive Matrix Factorization analysis.

3. Results and Discussion

3.1. Data Cleaning

The data for the four sites were gathered into a single dataset. The dataset was cleaned of all rows containing unavailable values and only the “date–time” combinations present in all four sites were retained. In this way, for each site, we obtained 7999 samples (each sample represents one hour of recording) which were almost evenly distributed in the six years of interest, leading to a dataset containing an overall number of 31,996 rows by five columns (one for each pollutant). The unavailable values were filled in using the algorithm present in the *pca.mvreplace* function contained in the *mdatools* package.

3.2. SOM Analysis

The abovementioned dataset was autoscaled by variable and used to build the SOM model, giving the model no prior knowledge about the site or year classification, to exploit the powerful unsupervised analysis potential of SOM algorithm. The algorithm initialization as well as the number of nodes and map dimensions were selected according to Vesanto's heuristic rules [26]. Several models were built using different map dimensions and the quality of the model was checked using three well-known parameters: the overall quantization error, the topographic error, and the distribution-matching error [28]. The best model was that with a 41×22 map dimension representing 902 recurrent air quality profiles. The obtained map is represented in Figure 2, in which the codebook values are shown in separate maps according to the variable using a grayscale ("heatmaps"). It can be observed that the highest pollutant values are depicted on the top and left edges of the map. In contrast, the lowest values are depicted in the bottom-right area of the map. As a rule, similar nodes are depicted close together on the map to maintain the dataset topology, but there can be some edges among nodes. The edges allow us to visualize possible node grouping and are usually represented using the distances between neighborhood nodes. The distance map represented in Figure 2 shows that there is a discontinuity (white "peak" area) spreading from the upper left part toward the map center. This is in accordance with the behavior observed in the heatmaps.

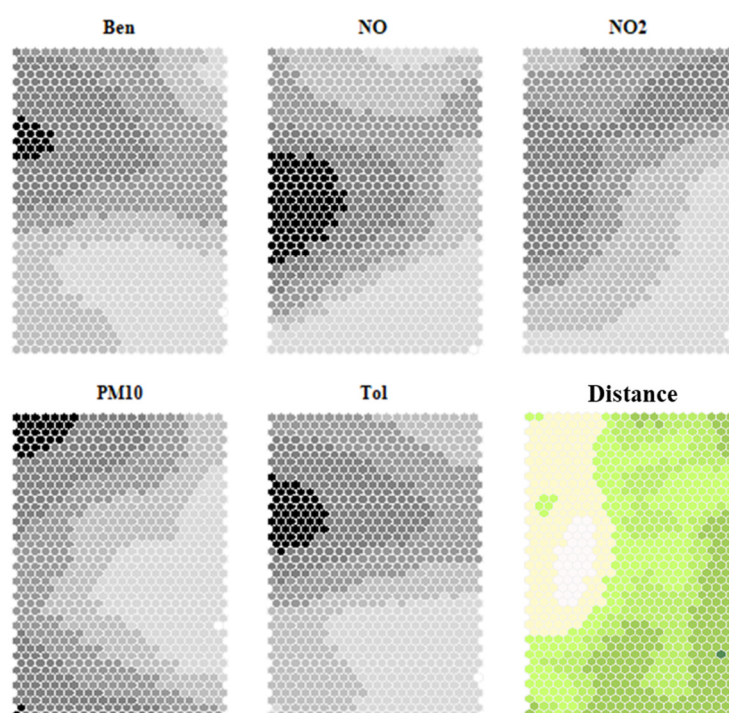


Figure 2. Distribution of the modeled variables on the SOM. The distribution of the single pollutants (Ben, NO, NO₂, Tol, PM₁₀) on each node is depicted in grayscale, from white (lower concentration values) to black (higher concentration values). In the distance map, the distance between a node and its neighbors is depicted with a scale from green to white: the higher the distance, the greater the prevalence of white shading on the scale.

3.3. Hierarchical Clustering

The nodes were then grouped using hierarchical clustering for a more in-depth and quantitative analysis, as stated in Section 2.2. The clustering algorithm was applied to the codebook. The clustering is shown in Figure 3 in a two-way mode in which both the nodes (arranged in rows) and the modeled variables (arranged in columns) were grouped.

We evaluated the quality of clustering from 2 to 10 clusters using the Davies–Bouldin index [38], which is regarded as a robust cluster validity index [39]. The best number of clusters was six. The six clusters are highlighted by colored rectangles in Figure 3; moreover, the rectangles show where the HCA “branches” were cut row-wise in order to obtain the six clusters.

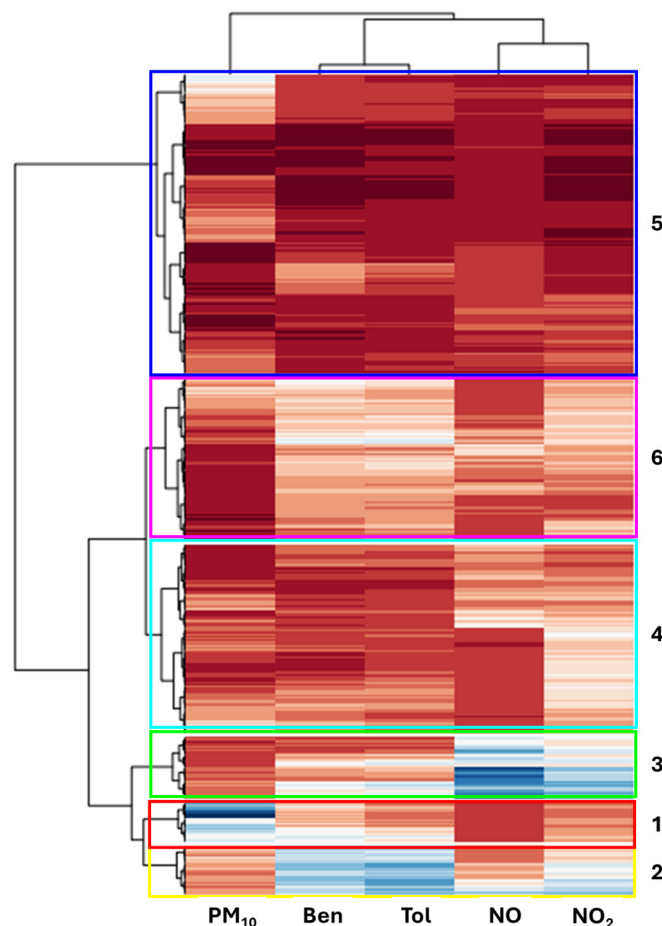


Figure 3. Clustered two-way HCA map. Each row represents a node, while each column represents the values of the modeled variables retaining the autoscaling operated before SOM analysis; thus, the color scale represents low (dark red) to high (dark blue) values. The six clusters obtained are depicted by rectangles and the assigned cluster number is indicated on the right-hand side of the figure.

In Figure 4, the grouping on the map is shown together with the “air type” variable profiles represented by radar plots. The boxplots on the right show the spread of the modeled values for each cluster. The node clusters provide insights into the characteristics present in the pollutants data. It can be observed that clusters 1, 2, and 3 (highlighted in the top-left part of the map) represent “highly polluted” air with relatively high values of PM_{10} (cluster 1), Ben and Tol (cluster 2), NO and NO_2 (cluster 3), compared to the others. By observing the boxplot, it can be noticed that for cluster 3, the variable values are fairly spread out, with the exception of PM_{10} . Clusters 1, 2, and 3 are in the same map area outlined by the distance discontinuity depicted in Figure 2. Moreover, the distribution of the variable values of the same area can be identified in the heatmaps presented in Figure 2.

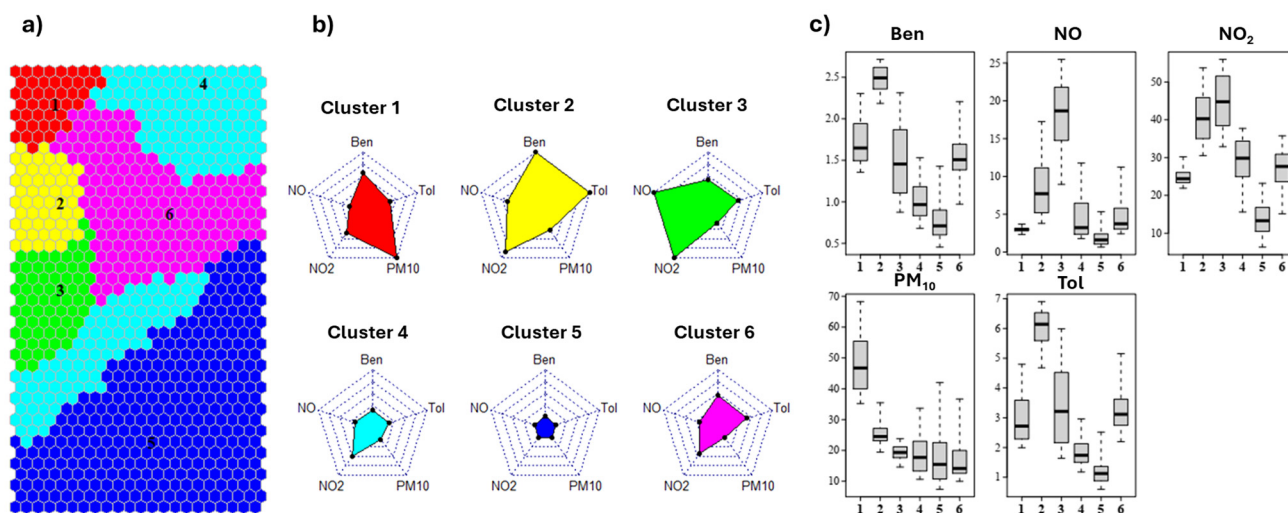


Figure 4. (a) Division of SOM nodes into 6 clusters as obtained by HCA; (b) representation of the cluster centroid values by radar plots; (c) distribution of the modeled values for each cluster, as defined by SOM. For this figure, we used the same cluster color code as the one used in Figure 3.

Cluster 5 occupies the entire bottom-right side of the map, as it contains many nodes with the smallest concentrations of pollutants among all other clusters. It can be associated with the “background” air type. A so-called “background” pollutant concentration is the lowest level of concentration that can be reached in an area, when there are no “active” sources nearby; only a level of concentration that can be attributed to long-distance transport and aged pollutant distribution in the area. Cluster 6 and 4 can be classified as “medium” and “low” polluted air, respectively.

The significance of the difference between the clusters was assessed using a Kruskal–Wallis non-parametric test. p -values of <0.001 were obtained for all the variables. Then, the Wilcoxon test with Bonferroni correction was used to assess the paired difference between clusters. Almost all the paired differences were significant; the results are reported in detail in the Supplementary Material.

The obtained clusters can also be linked to the different monitoring stations and the timeframe in which the pollutant levels were recorded. This evaluation was performed for all stations by labeling the percentage of each cluster for each day, according to the monitoring station and year. In Figure 5, the stacked daily bar plots for site A1 are represented, with the same color code provided for the SOM in Figure 4. The plots for all the sites are reported in the Supplementary Material.

Figure 5 shows the effect of the lockdown in 2020, with high prevalence of cluster 5, showing that the impact of pollutants was low for most of this period, according to results highlighted in other studies [23]. This effect can also be observed at sites B1 and A2 and, to a lesser extent, at B2. Thus, the effect is more evident in “traffic”-monitoring stations (A1 and B1), in accordance with other studies [24,25].

It can also be seen that, for the year 2020 alone, both A- and B-type sites in the respective cities show similar behavior, smoothing out the differences that can be observed in other years.

Cluster 2 seems to be a peculiar “air type” of site 1, characterized by the presence of high values of Ben and Tol, and partly of NO₂. It can be related to two possible sources. The first source, until its closure in 2020, was an integrated-cycle steel plant which, in particular, released benzene from its coke distillation ovens [40,41]. The second source, that is still active, is the presence of a harbor with a petrochemical terminal, which can release several types of hydrocarbons during unloading operations [42]. The presence of NO₂ can

be related to the combustion process for the first source and to ship stack emissions for the second one.

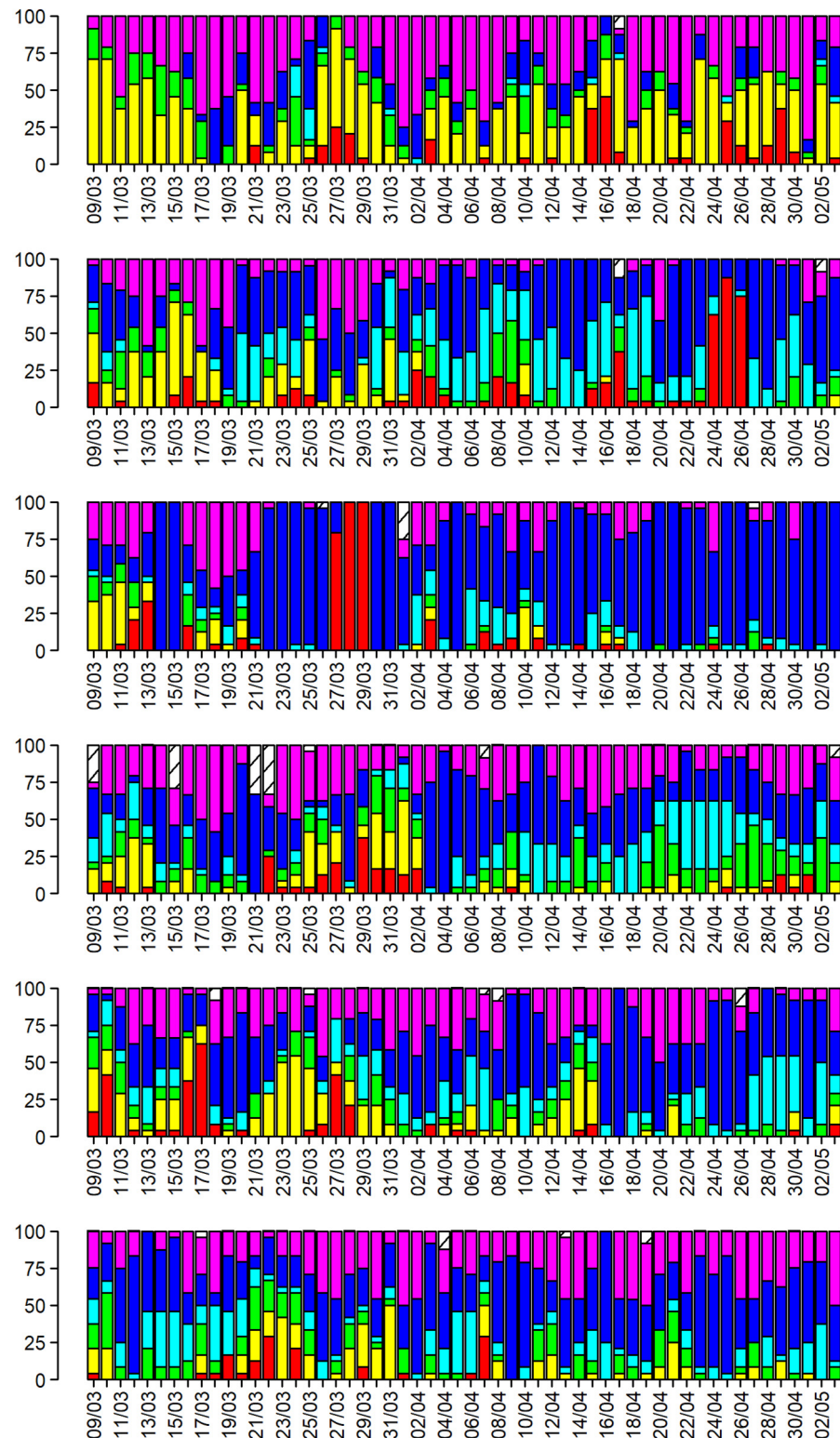


Figure 5. Barplots representing the daily percentage distribution of clusters for site A1. From the top to the bottom of the figure: years from 2018 to 2023. For this figure, we have used the same cluster color code as the one in Figure 4.

Cluster 3 shows the highest values of NO, which is a product of primary combustion. After the emission, it is oxidized to NO₂ in tens of minutes, depending on the presence

of oxidants, such as ozone, in the air [43,44]. From the daily plots, it can be observed that cluster 3 is barely visible for 2020 for all of the sites, confirming that the primary combustion source close to the monitoring stations (i.e., traffic) was largely absent, as other sources of combustion producing NO (e.g., domestic heating) that are not so close to the stations exploit their effect in NO₂ concentrations.

For 2019 and 2020, some days (at the end of March and at the end of April, respectively) show a high percentage of cluster 1, which is characterized by a high level of PM₁₀.

The SOM output allows us to detect possible outliers observing the values of the “so-called” quantization errors (QEs) [45,46].

A quantization error is the multidimensional distance of a sample from the node that best represents it (Best Matching Unit). A relatively high value of QEs means that the sample could be a possible outlier. The sample outliers were explored and reported in the Supplementary Material in stacked plots split by site and year; in 2020, all the sites showed outliers belonging to cluster 1. To a minor extent, such outliers were also present in 2019, although none were observed for site B2. The outliers correspond to two intense desert dust intrusions in the Northern Adriatic Sea area [47].

In site B2, in 2023, we noted the presence of samples with high QEs at the end of March for few hours. They could be related to a possible transient event, such as an accidental gasoline spill or road asphaltting operation.

3.4. Positive-Matrix Factorization

The codebook was used as input for the PMF algorithm and the uncertainties used for the variables were 6.2% for NO and NO₂, 10.9% for Ben and Tol, and 10% for PM₁₀, respectively. Considering the abovementioned clustering results, we tried to identify four to six emission sources/factors, obtaining better model quality parameters for five factors. The bootstrap method was employed to mitigate uncertainty and validate the precision of the PMF model. The first three factors were perfectly replicated in over 100% of the runs; factor 4 in over 87% of the runs, and factor 5 in over 42% of the runs. No runs were left unmapped, indicating that bootstrap uncertainties are interpretable and the number of factors may be suitable. For all factors, 80% of the species from the base run fell within the interquartile range (25th–75th percentile) of the bootstrap runs, thereby underscoring the robustness of the PMF modeling. The five factor profiles are shown in Figure 6 along with the nodes of the map, which mainly contributed to the factors. The factor fingerprints are reported in the Supplementary Material.

Factor 1 mainly shows an hydrocarbon source with a partial contribution of NO₂. Most of the nodes that contribute to this factor are those in cluster 2, which has already been recognized as a peculiarity of site 1. This is a mixed industrial source, as explained in detail in Section 3.3. Factor 2 shows the highest percentage of NO, indicating a primary source of combustion that, considering the monitoring station positioning described at Section 2.1, is mainly related to road transport. The nodes that mostly contribute to this factor are those belonging to cluster 3, which was described in detail in Section 3.3.

Factor 3 contains the highest percentage of PM₁₀, with nearly no percentage of the other pollutants indicating a “pure” dust source. The nodes with the greatest contribution to this factor are those belonging to cluster 1. In this cluster, the outliers represent the desert dust intrusion were found, as described in detail in Section 3.3.

Factor 4 indicates an aged combustion source, with the presence of NO₂ and PM₁₀. The nodes with the greatest contribution to this factor are those belonging to cluster 4, which represented “low polluted” air; thus, the results are consistent. Factor 5 indicates a mixed source of PM₁₀ and hydrocarbons. The nodes with the greatest contribution to this factor are fairly spread out on the SOM, with some of them gathered in a sub-area of cluster

5; moreover, this was the factor with the lowest precision. Thus, there is no straightforward interpretation of this factor.

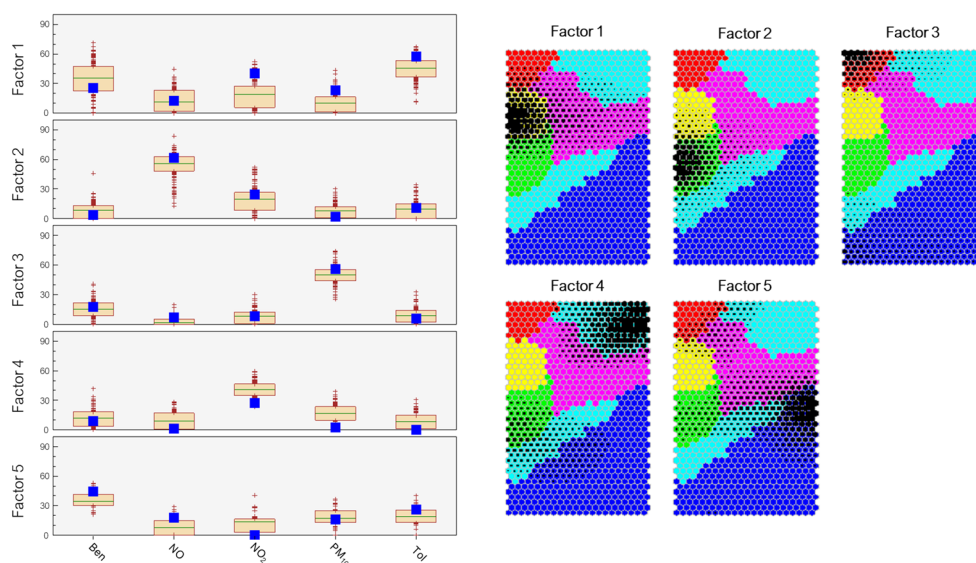


Figure 6. On the left: Variability in the % contribution of each species to the respective PMF factor (sum of factors = 100%). The base run is shown as a blue box for reference. On the right: the nodes that made greater contributions to a factor are represented in black, with a greater amount of black shading indicating a more substantial contribution.

The nodes belonging to cluster 6 did not show a clear contribution to a specific factor; thus, they probably originate from a combination of sources, in accordance with the cluster interpretation of “medium-level” pollution.

Using PMF, we recognized and confirmed the source nature of four out of six clusters identified by HCA, allowing a more in-depth interpretation of the spatial and temporal air quality variation in the two monitored cities.

4. Conclusions

In this study, we used a combined approach of multivariate analysis involving SOM, HCA, and PMF to disentangle multiannual, multisite data in a single elaboration without previously separating the sites and years. The approach proved to be valid, allowing us to detect the site peculiarities in terms of pollutant sources, the variation in pollutant profiles during years, and the outliers, affording a reliable interpretation. In detail, SOM allowed us to obtain recurrent pollutant profiles while retaining the relevant information and reducing the noise. HCA allowed us to classify different air types and evaluate their impact on the population. PMF was used to recognize and confirm the pollutant sources identified by HCA.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/toxics13020137/s1>, Table S1: Wilcoxon test results; Figure S1: Barplots representing the daily percentage distribution of clusters for site A1. From the top to the bottom of the figure: years from 2018 to 2023; Figure S2: Barplots representing the daily percentage distribution of clusters for site B1. From the top to the bottom of the figure: years from 2018 to 2023; Figure S3: Barplots representing the daily percentage distribution of clusters for site A2. From the top to the bottom of the figure: years from 2018 to 2023; Figure S4: Barplots representing the daily percentage distribution of clusters for site B2. From the top to the bottom of the figure: years from 2018 to 2023; Figure S5: Quantization error plots for site A1. From the top to the bottom of the figure: years from 2018 to 2023; Figure S6: Quantization error plots for site B1. From the top to the bottom of

the figure: years from 2018 to 2023; Figure S7: Quantization error plots for site A2. From the top to the bottom of the figure: years from 2018 to 2023; Figure S8: Quantization error plots for site B2. From the top to the bottom of the figure: years from 2018 to 2023; Figure S9: PMF results: Factor fingerprints.

Author Contributions: Conceptualization, P.B. and S.L.; methodology, S.F., A.A. and S.L.; formal analysis, S.F. and S.L.; resources, S.L. and A.A.; data curation, S.F. and S.L.; writing—original draft preparation, S.F. and S.L.; writing—review and editing, A.A., P.B., S.F. and S.L.; visualization, S.L.; supervision, P.B. and S.L.; project administration, A.A. and S.L.; funding acquisition, P.B. and S.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the European Union, NextGenerationEU project PNRR iNEST CUP J43C22000320006; and Pomeranian University in Słupsk: 7-7-16.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available on request.

Acknowledgments: We gratefully thank ARPA-FVG (Environmental Protection Agency of Friuli Venezia-Giulia region—Italy) for providing the pollutant data.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. de Vries, W.; Posch, M.; Simpson, D.; de Leeuw, F.A.A.M.; van Grinsven, H.J.M.; Schulte-Uebbing, L.F.; Sutton, M.A.; Ros, G.H. Trends and Geographic Variation in Adverse Impacts of Nitrogen Use in Europe on Human Health, Climate, and Ecosystems: A Review. *Earth Sci. Rev.* **2024**, *253*, 104789. [\[CrossRef\]](#)
2. Mahakalkar, A.U.; Gianquintieri, L.; Amici, L.; Brovelli, M.A.; Caiani, E.G. Geospatial Analysis of Short-Term Exposure to Air Pollution and Risk of Cardiovascular Diseases and Mortality—A Systematic Review. *Chemosphere* **2024**, *353*, 141495. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Markozannes, G.; Pantavou, K.; Rizos, E.C.; Sindosi, O.; Tagkas, C.; Seyfried, M.; Saldanha, I.J.; Hatzianastassiou, N.; Nikolopoulos, G.K.; Ntzani, E. Outdoor Air Quality and Human Health: An Overview of Reviews of Observational Studies. *Environ. Pollut.* **2022**, *306*, 119309. [\[CrossRef\]](#)
4. Sicard, P.; Agathokleous, E.; Anenberg, S.C.; De Marco, A.; Paoletti, E.; Calatayud, V. Trends in Urban Air Pollution over the Last Two Decades: A Global Perspective. *Sci. Total Environ.* **2023**, *858*, 160064. [\[CrossRef\]](#)
5. Tahir Bahadur, F.; Rasool Shah, S.; Rao Nidamanuri, R. Air Pollution Monitoring, and Modelling: An Overview. *Environ. Forensics* **2024**, *25*, 309–336. [\[CrossRef\]](#)
6. Havemann, S.; Kishcha, P.; Agbehadji, I.E.; Obagbuwa, I.C. Systematic Review of Machine Learning and Deep Learning Techniques for Spatiotemporal Air Quality Prediction. *Atmosphere* **2024**, *15*, 1352. [\[CrossRef\]](#)
7. Alvarez-Guerra, E.; Molina, A.; Viguri, J.R.; Alvarez-Guerra, M. A SOM-Based Methodology for Classifying Air Quality Monitoring Stations. *Environ. Prog. Sustain. Energy* **2011**, *30*, 424–438. [\[CrossRef\]](#)
8. de Oliveira, R.H.; Carneiro, C.C.; de Almeida, F.G.V.; de Oliveira, B.M.; Nunes, E.H.M.; dos Santos, A.S. Multivariate Air Pollution Classification in Urban Areas Using Mobile Sensors and Self-Organizing Maps. *Int. J. Environ. Sci. Technol.* **2019**, *16*, 5475–5488. [\[CrossRef\]](#)
9. Licen, S.; Cozzutto, S.; Barbieri, G.; Crosera, M.; Adami, G.; Barbieri, P. Characterization of Variability of Air Particulate Matter Size Profiles Recorded by Optical Particle Counters near a Complex Emissive Source by Use of Self-Organizing Map Algorithm. *Chemom. Intell. Lab. Syst.* **2019**, *190*, 48–54. [\[CrossRef\]](#)
10. Costa, E.L.R.; Braga, T.; Dias, L.A.; de Albuquerque, É.L.; Fernandes, M.A.C. Self-Organizing Maps Applied to the Analysis and Identification of Characteristics Related to Air Quality Monitoring Stations and Its Pollutants. *Neural Comput. Appl.* **2024**, *36*, 11643–11657. [\[CrossRef\]](#)
11. Song, X.H.; Hopke, P.K. Kohonen Neural Network as a Pattern Recognition Method Based on the Weight Interpretation. *Anal. Chim. Acta* **1996**, *334*, 57–66. [\[CrossRef\]](#)
12. Kohonen, T. *Self-Organizing Maps Springer Series in Information Sciences*; Springer: Berlin/Heidelberg, Germany, 2001.
13. Kohonen, T. Essentials of the Self-Organizing Map. *Neural Netw.* **2013**, *37*, 52–65. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Hopke, P.K. Review of Receptor Modeling Methods for Source Apportionment. *J. Air Waste Manag. Assoc.* **2016**, *66*, 237–259. [\[CrossRef\]](#) [\[PubMed\]](#)

15. Zhou, L.; Hopke, P.K.; Paatero, P.; Ondov, J.M.; Pancras, J.P.; Pekney, N.J.; Davidson, C.I. Advanced Factor Analysis for Multiple Time Resolution Aerosol Composition Data. *Atmos. Environ.* **2004**, *38*, 4909–4920. [\[CrossRef\]](#)
16. Paatero, P.; Tapper, U. Positive Matrix Factorization: A Non-Negative Factor Model with Optimal Utilization of Error Estimates of Data Values. *Environmetrics* **1994**, *5*, 111–126. [\[CrossRef\]](#)
17. Fan, W.; Zhou, J.; Zheng, J.; Guo, Y.; Hu, L.; Shan, R. Hydrochemical Characteristics, Control Factors and Health Risk Assessment of Groundwater in Typical Arid Region Hotan Area, Chinese Xinjiang. *Environ. Pollut.* **2024**, *363*, 125301. [\[CrossRef\]](#)
18. Zeng, J.; Liu, K.; Liu, X.; Tang, Z.; Wang, X.; Fu, R.; Lin, X.; Liu, N.; Qiu, J. Driving Factor, Source Identification, and Health Risk of PFAS Contamination in Groundwater Based on the Self-Organizing Map. *Water Res.* **2024**, *267*, 122458. [\[CrossRef\]](#)
19. Trajković, I.; Sentić, M.; Vesković, J.; Lučić, M.; Miletić, A.; Onjia, A. Source-Oriented Health Risks and Distribution of BTEXS in Urban Shallow Lake Sediment: Application of the Positive Matrix Factorization Model. *Water* **2024**, *16*, 2302. [\[CrossRef\]](#)
20. Zhang, Y.; Zhang, Q.; Chen, W.; Shi, W.; Cui, Y.; Chen, L.; Shao, J. Source Apportionment and Migration Characteristics of Heavy Metal(Loid)s in Soil and Groundwater of Contaminated Site. *Environ. Pollut.* **2023**, *338*, 122584. [\[CrossRef\]](#)
21. Hassan, M.S.; Bhuiyan, M.A.H.; Rahman, M.T. Sources, Pattern, and Possible Health Impacts of PM_{2.5} in the Central Region of Bangladesh Using PMF, SOM, and Machine Learning Techniques. *Case Stud. Chem. Environ. Eng.* **2023**, *8*, 100366. [\[CrossRef\]](#)
22. Liu, H.; Wang, Q.; Liu, S.; Zhou, B.; Qu, Y.; Tian, J.; Zhang, T.; Han, Y.; Cao, J. The Impact of Atmospheric Motions on Source-Specific Black Carbon and the Induced Direct Radiative Effects over a River-Valley Region. *Atmos. Chem. Phys.* **2022**, *22*, 11739–11757. [\[CrossRef\]](#)
23. Kumar, S. Insights on Air Pollution During COVID-19: A Review. *Aerosol Sci. Eng.* **2023**, *7*, 192–206. [\[CrossRef\]](#)
24. Sokhi, R.S.; Singh, V.; Querol, X.; Finardi, S.; Targino, A.C.; Andrade, M.d.F.; Pavlovic, R.; Garland, R.M.; Massagué, J.; Kong, S.; et al. A Global Observational Analysis to Understand Changes in Air Quality during Exceptionally Low Anthropogenic Emission Conditions. *Environ. Int.* **2021**, *157*, 106818. [\[CrossRef\]](#)
25. Bar, S.; Parida, B.R.; Mandal, S.P.; Pandey, A.C.; Kumar, N.; Mishra, B. Impacts of Partial to Complete COVID-19 Lockdown on NO₂ and PM_{2.5} Levels in Major Urban Cities of Europe and USA. *Cities* **2021**, *117*, 103308. [\[CrossRef\]](#)
26. Vesanto, J. SOM-Based Data Visualization Methods. *Intell. Data Anal.* **1999**, *3*, 111–126. [\[CrossRef\]](#)
27. Himberg, J.; Ahola, J.; Alhoniemi, E.; Vesanto, J.; Simula, O. *The Self-Organizing Map as a Tool in Knowledge Engineering*; World Scientific Publishing: Singapore, 2001; pp. 38–65.
28. Licen, S.; Astel, A.; Tsakovski, S. Self-Organizing Map Algorithm for Assessing Spatial and Temporal Patterns of Pollutants in Environmental Compartments: A Review. *Sci. Total Environ.* **2023**, *878*, 163084. [\[CrossRef\]](#)
29. Clark, S.; Sisson, S.A.; Sharma, A. Tools for Enhancing the Application of Self-Organizing Maps in Water Resources Research and Engineering. *Adv. Water Resour.* **2020**, *143*, 103676. [\[CrossRef\]](#)
30. Vesanto, J.; Alhoniemi, E. Clustering of the Self-Organizing Map. *IEEE Trans. Neural Netw.* **2000**, *11*, 586–600. [\[CrossRef\]](#)
31. Paatero, P. Least Squares Formulation of Robust Non-Negative Factor Analysis. *Chemom. Intell. Lab. Syst.* **1997**, *37*, 23–35. [\[CrossRef\]](#)
32. Licen, S.; Franzon, M.; Rodani, T.; Barbieri, P. SOMEnv: An R Package for Mining Environmental Monitoring Datasets by Self-Organizing Map and k-Means Algorithms with a Graphical User Interface. *Microchem. J.* **2021**, *165*, 106181. [\[CrossRef\]](#)
33. Melssen, W.; Wehrens, R.; Buydens, L. Supervised Kohonen Networks for Classification Problems. *Chemom. Intell. Lab. Syst.* **2006**, *83*, 99–113. [\[CrossRef\]](#)
34. Wehrens, R.; Kruisselbrink, J. Flexible Self-Organizing Maps in Kohonen 3.0. *J. Stat. Softw.* **2018**, *87*, 1–18. [\[CrossRef\]](#)
35. Carslaw, D.C.; Ropkins, K. Openair—An r Package for Air Quality Data Analysis. *Environ. Model. Softw.* **2012**, *27–28*, 52–61. [\[CrossRef\]](#)
36. Kucheryavskiy, S. Mdatools—R Package for Chemometrics. *Chemom. Intell. Lab. Syst.* **2020**, *198*, 103937. [\[CrossRef\]](#)
37. Kolde, R. Package “Pheatmap”: Pretty Heatmaps. R. package; GitHub, Inc.: San Francisco, CA, USA, 2022; pp. 1–8.
38. Davies, D.L.; Bouldin, D.W. A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1979**, *PAMI-1*, 224–227. [\[CrossRef\]](#)
39. Todeschini, R.; Ballabio, D.; Termopoli, V.; Consonni, V. Extended Multivariate Comparison of 68 Cluster Validity Indices. A Review. *Chemom. Intell. Lab. Syst.* **2024**, *251*, 105117. [\[CrossRef\]](#)
40. Licen, S.; Tolloi, A.; Briguglio, S.; Piazzalunga, A.; Adami, G.; Barbieri, P. Small Scale Spatial Gradients of Outdoor and Indoor Benzene in Proximity of an Integrated Steel Plant. *Sci. Total Environ.* **2016**, *553*, 524–531. [\[CrossRef\]](#)
41. Astel, A.M.; Giorgini, L.; Mistaro, A.; Pellegrini, I.; Cozzutto, S.; Barbieri, P. Urban BTEX Spatiotemporal Exposure Assessment by Chemometric Expertise. *Water Air Soil Pollut.* **2013**, *224*, 1503. [\[CrossRef\]](#)
42. Kiihamäki, S.P.; Korhonen, M.; Kukkonen, J.; Shiue, I.; Jaakkola, J.J.K. Effects of Ambient Air Pollution from Shipping on Mortality: A Systematic Review. *Sci. Total Environ.* **2024**, *945*, 173714. [\[CrossRef\]](#)
43. Stewart, G.B.; Dajnak, D.; Davison, J.; Carslaw, D.C.; Beddows, A.V.; Phantawesak, N.; Stettler, M.E.J.; Holloway, M.J.; Beevers, S.D. New NO_x and NO₂ Vehicle Emission Curves, and Their Implications for Emissions Inventories and Air Pollution Modelling. *Urban Clim.* **2024**, *57*, 102103. [\[CrossRef\]](#)

-
44. Ghermandi, G.; Fabbi, S.; Veratti, G.; Bigi, A.; Teggi, S. Estimate of Secondary NO₂ Levels at Two Urban Traffic Sites Using Observations and Modelling. *Sustainability* **2020**, *12*, 7897. [[CrossRef](#)]
 45. Muñoz, A.; Muruzábal, J. Self-Organizing Maps for Outlier Detection. *Neurocomputing* **1998**, *18*, 33–60. [[CrossRef](#)]
 46. Muruzábal, J.; Muñoz, A. On the Visualization of Outliers via Self-Organizing Maps. *J. Comput. Graph. Stat.* **1997**, *6*, 355–382. [[CrossRef](#)]
 47. Mifka, B.; Telišman Prtenjak, M.; Kavre Piltaver, I.; Mekterović, D.; Kuzmić, J.; Marciuš, M.; Ciglencčki, I. Intense Desert Dust Event in the Northern Adriatic (March 2020); Insights From the Numerical Model Application and Chemical Characterization Results. *Earth Space Sci.* **2023**, *10*, e2023EA002879. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.