frontiers | Frontiers in Public Health

TYPE Original Research
PUBLISHED 07 May 2024
DOI 10.3389/fpubh.2024.1344865

Check for updates

OPEN ACCESS

EDITED BY
Uttara Saran,
University of Texas MD Anderson Cancer Center, United States

REVIEWED BY
Arlette Setiawan,
Padjadjaran University, Indonesia
Karthikeyan Bagavathy Shanmugam,
Emory University, United States

*CORRESPONDENCE
Roberto Cilli
✉ roberto.cilli@uniba.it
Sabina Tangaro
✉ sabina.tangaro@uniba.it

†These authors have contributed equally to this work

RECEIVED 26 November 2023
ACCEPTED 08 April 2024
PUBLISHED 07 May 2024

CITATION
Romano D, Novielli P, Cilli R, Amoroso N, Monaco A, Bellotti R and Tangaro S (2024) Air pollution and mortality for cancer of the respiratory system in Italy: an explainable artificial intelligence approach.
Front. Public Health 12:1344865.
doi: 10.3389/fpubh.2024.1344865

COPYRIGHT
© 2024 Romano, Novielli, Cilli, Amoroso, Monaco, Bellotti and Tangaro. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Air pollution and mortality for cancer of the respiratory system in Italy: an explainable artificial intelligence approach

Donato Romano[1,2], Pierfrancesco Novielli[1,2], Roberto Cilli[2,3]*, Nicola Amoroso[2,4], Alfonso Monaco[2,3], Roberto Bellotti[2,3†] and Sabina Tangaro[1,2*†]

[1]Dipartimento di Scienze del Suolo, della Pianta e degli Alimenti Universita' degli Studi di Bari Aldo Moro, Bari, Italy, [2]Istituto Nazionale di Fisica Nucleare Sezione di Bari, Bari, Italy, [3]Dipartimento Interateneo di Fisica, "M. Merlin" Universita' degli Studi di Bari Aldo Moro, Bari, Italy, [4]Dipartimento di Farmacia Scienze, del Farmaco Universita' degli Studi di Bari Aldo Moro, Bari, Italy

Respiratory system cancer, encompassing lung, trachea and bronchus cancer, constitute a substantial and evolving public health challenge. Since pollution plays a prominent cause in the development of this disease, identifying which substances are most harmful is fundamental for implementing policies aimed at reducing exposure to these substances. We propose an approach based on explainable artificial intelligence (XAI) based on remote sensing data to identify the factors that most influence the prediction of the standard mortality ratio (SMR) for respiratory system cancer in the Italian provinces using environment and socio-economic data. First of all, we identified 10 clusters of provinces through the study of the SMR variogram. Then, a Random Forest regressor is used for learning a compact representation of data. Finally, we used XAI to identify which features were most important in predicting SMR values. Our machine learning analysis shows that NO, income and O3 are the first three relevant features for the mortality of this type of cancer, and provides a guideline on intervention priorities in reducing risk factors.

KEYWORDS

explainable artificial intelligence, air pollution, lung cancer, respiratory disease, socio-economic indices, public health, remote sensing 2010 MSC: 00-01, 99-00

1 Introduction

This study aims to investigate the relationship between air pollution and respiratory system cancer mortality in Italian provinces. Air pollution is a pressing issue of the modern world, impacting human health and the environment in numerous ways (1). One of its significant consequences is its link to respiratory system cancer, a malignancy that claims millions of lives globally each year (2). Air pollution, which mainly consists of fine particulate matter and toxic gases, can enter the lung tissue and trigger a series of chronic inflammatory and oxidative damage to the cells, leading to malignant transformation.

Respiratory system cancer is a type of cancer that affects the lungs, bronchi and trachea. The most common types of respiratory system cancer are lung cancer and bronchial cancer. (3). While smoking is a primary risk factor, there are other important contributing factors, such as secondhand smoke, occupational exposure, and air pollution. This sets the stage for the focus of the study (3). It is estimated that 14% of lung cancer deaths are attributable

to environmental air pollution (4). Symptoms of respiratory system cancer can include persistent cough, chest pain, shortness of breath, hoarseness, fatigue, and weight loss.

A number of works have dealt with long and short-term effect of air pollution and cancer (5, 6), highlighting a multitude of contributing factors such as the type of pollutant, exposure time and frequency, and individual susceptibility (7): an increase of 10 $\mu g/m3$ in PM10 concentration raises the average likelihood of developing lung cancer by 20%, whereas a 5 $\mu g/m3$ rise in PM2.5 elevates the risk by 30% (8). Recognizing and addressing the link between air pollution and respiratory system cancer is critical for protecting public health on a global scale. It informs evidence-based policies, encourages international collaboration, and empowers individuals and communities to take actions that can mitigate the impact of air pollution on respiratory health.

The goal of this work was to implement a Machine Learning (ML) algorithm predicting the standard mortality ratio (SMR) for lung, bronchi and trachea cancer of the Italian provinces by using air pollution data downloaded from Copernicus Atmosphere Monitoring Service (CAMS) and socio-economic data downloaded from ISTAT. ML is the discipline dealing with the replication of the learning mechanisms of the human brain through statistical algorithms (9). Random forest is a machine learning algorithm used for both classification and regression tasks. It is an ensemble learning method that creates several decision trees and combines their predictions to make a final decision or prediction. Recent advancements in ML techniques resulted in the introduction of eXplainable Artificial Intelligence (XAI) which allows for the identification of the crucial attributes for each instance (10–12). Explainable AI provides clarity and understanding into the decision-making processes of AI models. It helps improve transparency, trust, accountability, and compliance with regulations. XAI methods has been applied to ML algorithms to provide a clear picture of the relevant features affecting the performance of the models, their relations with the outcomes, with each other's and both their local and global effects. The main intent of our study was to present a framework to determine which pollution indices, based on remote sensing data, are most associated to the mortality from cancer of the respiratory system. We studied the mortality in the Italian provinces given its heterogeneity in terms of ecological and environmental features. In order to do this, we evaluated to what extent mortality from cancer of the respiratory system can be predicted based on environmental pollution and socio-economic indices.

2 Materials

The study was conducted using mortality data from ISTAT, that is the Italian Institute of Statistics, responsible for collecting, analyzing, and disseminating official statistics on the country's population, economy, and society.[1] ISTAT's main functions include conducting population censuses, compiling and publishing official statistics on topics such as, employment, and economic indicators, and providing support and expertise to other public institutions and organizations in the field of statistics. In particular, the respiratory system cancer mortality at the provinces level in 2019 has been considered.

In this work, we mainly exploited pollutants' concentration from the Copernicus Air Monitoring Service (CAMS). An outline of the data preparation workflow is presented in Figure 1. In order to get an analysis-ready data table we firstly collected daily air quality maps from the Copernicus Data Store for the year 2019.[2]

2.1 Input data preparation

The pollution data of year 2019 has been downloaded from Copernicus Atmosphere Monitoring Service (CAMS). It is a European Union program that provides comprehensive information on air quality and the Earth's atmosphere.

It aims to improve air quality forecasts and support decision-making related to air quality management and environmental policy-making in Europe and around the world. CAMS uses a wide range of independent monitoring and modeling systems to collect and analyze data on atmospheric composition, air quality, and weather patterns. CAMS provides annual air quality reanalyzes for Europe based on both unvalidated and validated observations.[3] Since the downloaded data covered a larger area than that of interest, only pollution data of the Italian peninsula were extracted using the Python library GeoPandas.[4] Then, for each selected pollutant, both annual average and standard deviation were computed. The polluting substances considered were: carbon monoxide (CO), nitrogen monoxide (NO), nitrogen dioxide (NO_2), ozone (O_3), particulate matter 10 (pm10), particulate matter 2.5 (pm2.5) and sulfur dioxide (SO_2). Pollutant values are the result of an ensemble median of 11 state-of-the-art numerical air quality models developed in Europe: CHIMERE from INERIS (France) (13), EMEP from MET Norway (Norway) (14), EURAD-IM from Jülich IEK (Germany) (15), LOTOS-EUROS from KNMI and TNO (Netherlands) (16), MATCH from SMHI (Sweden) (17), MOCAGE from METEO-FRANCE (France) (18), SILAM from FMI (Finland) (19), DEHM from AARHUS UNIVERSITY (Denmark) (20), GEM-AQ from IEP-NRI (Poland) (21), MONARCH from BSC (Spain) (22) and MINNI from ENEA (Italy) (23). The yearly ensemble reanalyzes are available with a time resolution of 1 h from step 1st January to the 31st of December, while the horizontal resolution of ensemble reanalyzes is on a $0.1^\circ \times 0.1^\circ$ regular latitude-longitude grid. The analysis was conducted using 2019 data since pollution data available on the CAMS platform starts from April 2018. Additionally, years following 2019 were not considered due to the COVID epidemic that

1 Available online at: https://www.istat.it/it/archivio/222527.

2 Available online at: https://atmosphere.copernicus.eu/.

3 Available online at: https://ads.atmosphere.copernicus.eu/#!/search?text=andtype$=$datasetandkeywords$=$((%20%22Product%20type:%20Reanalysis%22%20)%20AND%20(%20%22Variable%20domain:%20Atmosphere%20(composition)%22%20)%20AND%20(%20%22Spatial%20coverage:%20Europe%22%20)%20AND%20(%20%22Temporal%20coverage:%20Past%~22%20)).

4 Available online at: https://geopandas.org/en/stable/.

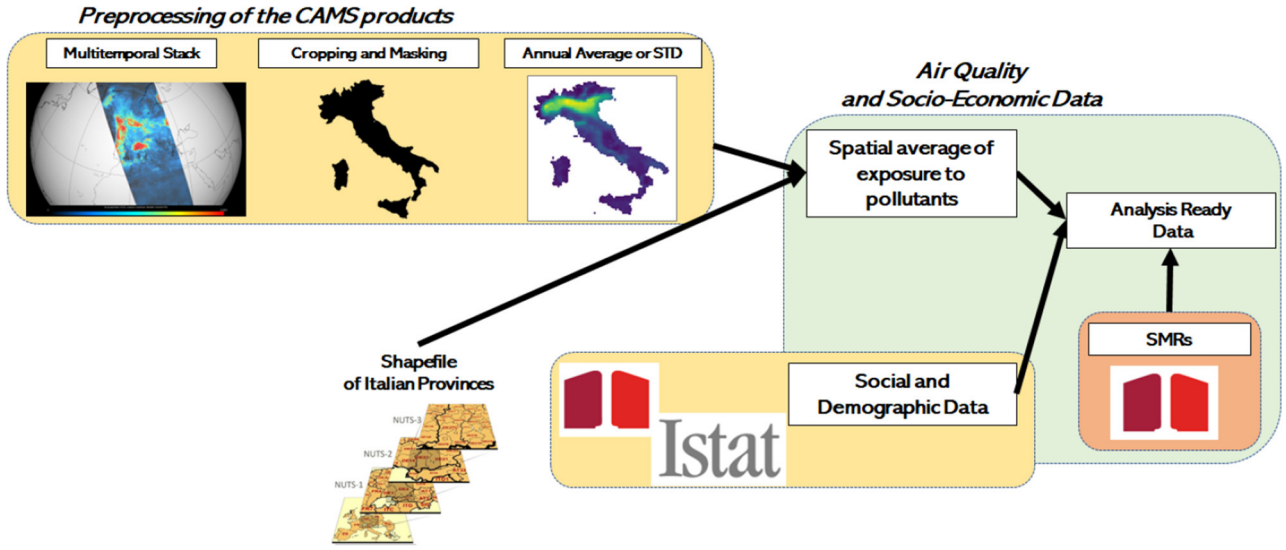

FIGURE 1
Data preparation workflow.

TABLE 1 Mean and standard deviation values of air pollution data by province and pollutant.

Pollutants	Northern Italy	Center Italy	Southern Italy
Pm2.5$(\frac{\mu g}{m^3})$	15.55 ± 5.42	10.94 ± 1.11	11.37 ± 2.15
pm10 $(\frac{\mu g}{m^3})$	19.09 ± 6.03	15.56 ± 1.37	16.33 ± 2.29
CO$(\frac{\mu g}{m^3})$	194.04 ± 45.17	156.58 ± 9.24	150.85 ± 20.63
NO$(\frac{\mu g}{m^3})$	1.23 ± 1.21	0.34 ± 0.16	0.33 ± 0.3
$NO_2(\frac{\mu g}{m^3})$	13.15 ± 6.49	7.32 ± 2.55	6.17 ± 3.5
$SO^2(\frac{\mu g}{m^3})$	1.56 ± 0.66	1.01 ± 0.27	1.33 ± 0.51
$O^3(\frac{\mu g}{m^3})$	58.88 ± 7.79	67.82 ± 5.88	69.63 ± 7.29

began in the early months of 2020, a disease that increased deaths from respiratory illnesses. The average values and standard deviations of pollutants for the provinces of Northern, Central, and Southern Italy are reported in Table 1. Subsequently, in order to get a structured data table to be added to socio-demographic descriptors, a spatial average of these intermediary maps have been computed within the boundaries of each Italian province. Other pollution-related and socio-economic variables were considered to enrich the data set, in particular: cultivated areas, urban areas, benzene, temperature, N fertilizer, P4010 fertilizer, microelement fertilizer, organic fertilizer, bed number, which represents the number of hospital beds available, lifetime, income, life quality, instruction, vehicles total, urban traffic, photovoltaic panel, green urban, electric consumption, noise and wastes (24).

The complete list of the selected descriptors is reported in Table 2.

2.2 Output data preparation: the indirect standardization

Indirect standardization is a statistical method used to compare the rates of events or conditions in two or more populations while controlling for differences in population characteristics such as age, sex or socioeconomic status. This method calculates expected rates for each population by applying the distribution of a particular population characteristic (for example age) to a reference population with known rates (25). The observed rates of a particular outcome in each population are then compared to the expected rates, which have been adjusted for any differences in population characteristics. In this work we computed the standard mortality ratio (SMR), the ratio between the deaths observed in a territory and those expected in the same. The expected deaths were calculated by applying the corresponding specific mortality ratios of the population assumed as standard (the national one in this case) to the average annual population by age classes of each territorial unit. The average standardized mortality ratio (SMR for cancer of the respiratory system) distribution within Italian provinces is shown in Figure 2A. Moreover, the province of Sondrio was removed from our analysis since it appears to be an outlier in the SMR distribution, as shown in the Figure 2B.

3 Methods

The main goal of this work is to find through an XAI algorithm, which air pollutants and/or socio-economic index contribute most to mortality from lung, trachea and bronchial cancer in Italian provinces. A cross-validation framework has been implemented to train a Random Forest regressor (RF) (26) of the standard mortality ratio for respiratory cancer in Italian provinces; then, by using SHapley Additive exPlanations (27), a method based on cooperative game theory and used, we performed an explain ability analysis

TABLE 2 Overview of the variables selected for this study.

Data	Type	Source	Value	Coding
Pollutant	Time series $\left(\frac{\mu g}{m^3}\right)$	CAMS	Mean $\left(\frac{\mu g}{m^3}\right)$, Standard deviation $\left(\frac{\mu g}{m^3}\right)$	mean pm2.5, std pm2.5, mean pm10, std pm10, mean CO, std CO, mean NO, std NO, mean NO_2, std NO_2, mean O_3, std O_3, mean SO_2, std SO2
Anthropogenic		ISTAT	Mean	N fertilizer, P410 fertilizer, Microelement fertilizer, Organic fertilizer, bed number, life time, income, life quality, instruction, Vehicles total, urban traffic, Electric Consumption, noise, wastes
Environment		ISTAT	Mean	cultivated areas, urban areas, benzene, temperature, photovoltaic panel, green urban

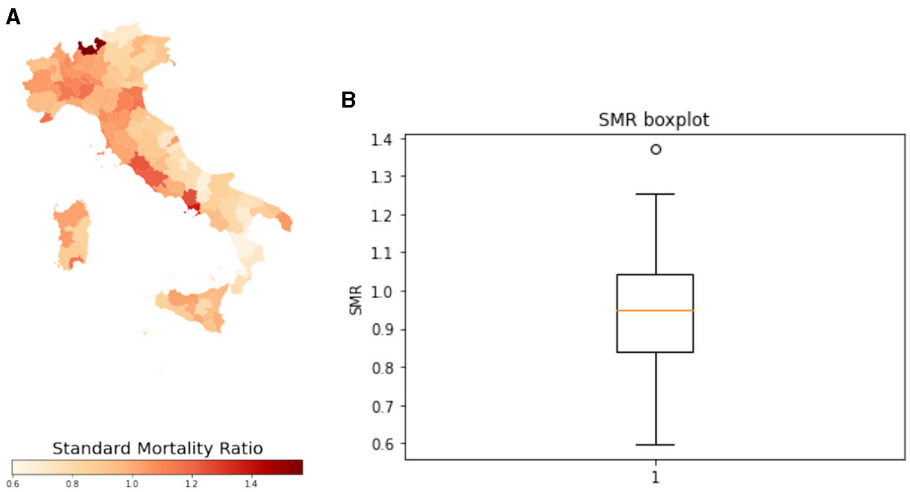

FIGURE 2
(A) Standardized mortality ratio distribution in 2019 within Italian provinces. **(B)** Boxplot of standardized mortality ratio for cancer of the respiratory system of the Italian provinces in 2019.

to increase transparency and interpretability of machine learning model. The Figure 3 shows a schematic overview of the methods adopted in the present work.

3.1 Random forest regressor of standard mortality ratio

Random Forest (RF) operates as an ensemble learning classifier rooted in the concept of classification trees. RF essentially creates a collection of classification trees, wherein each tree is trained on a bootstrapped sample from the available data. To prevent biased estimations, one-third of the available examples are excluded and utilized for the out-of-bag error estimation.

In the process of tree growth, the ideal split at each node relies on a random selection of M/3 descriptors where M represents the total available descriptors. It has been shown that the classification error is influenced by two primary factors: the interdependence among trees in the forest and the individual predictive power of each tree. Managing these factors involves adjusting the number of trees in the forest and the quantity of features sampled per split. The accuracy of RF models substantially depends on two parameters, the number of sampled features f and the number of the forest trees T.

In this study, RF was implemented using the random Forest routine from the scikit-learn package (v 1.2.1) (28) with its default configuration.

3.2 Hierarchical spatial clustering and model performance assessment

In order to avoid information leakage between spatially dependent observations, we preliminary found a partition of clusters of adjacent provinces to keep apart during the training and validation steps.

The spatial clusters of provinces adopted in the cross-validation scheme were found through a combined use of a hierarchical clustering algorithm applied to the matrix of the euclidean distances between provinces, and the semivariogram plot of the SMRs. A semivariogram plot is a tool used in geostatistics to assess the spatial dependence of a variable. It is usually plotted on a graph where the x-axis corresponds to the distance between a pair of selected locations (also known as spatial lag), while the y-axis corresponds to the average

squared difference of the variable values computed for those pairs of location within a given distance interval. In this work, we estimated the spatial range of the semivariogram, i.e., the distance after which the observations are supposed to be no longer correlated, using a discretionary approach by a visual identification of the plateau (i.e., when no further increase in variance is observed).

Establishing the range of the empirical semivariogram of the SMRs allowed us to delineate ten clusters after thresholding the dendrogram obtained by the hierarchical clustering of the geographical distances of the Italian provinces (Figure 4).

We utilized a leave-one-cluster-out cross-validation approach to mitigate spatial data bias when assessing the regression performance of an ensemble Random Forest regressor. This validation scheme was implemented to prevent overestimation of performance. In fact, spatial autocorrelation in two adjacent provinces, one selected in the training set and the other in validation, may lead to overoptimistic results.

Finally, we evaluated the performance of our machine learning model, by computing the coefficient of determination r^2 (Equation 1) and the mean absolute error *MAE* (Equation 2), whose definition is provided in the following:

Coefficient of determination:

$$r^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (1)$$

Mean absolute error:

$$MAE = \frac{1}{n}\sum_{i-1}^{n}\left|(y_i - \hat{y}_i)\right| \quad (2)$$

where $\hat{y}_i$ are the predicted values, $\bar{y}$ is their average, while y_i are the observed values of the SMR in the validation set.

3.3 Features explain ability

Epidemiological studies are crucial in understanding and controlling the spread of diseases, but often require large sets of data that are difficult for humans to analyze efficiently. Artificial intelligence (AI) has emerged as a useful tool in epidemiological studies, with promising applications in predicting disease outbreaks, identifying risk factors, and developing targeted interventions. However, as AI becomes more prevalent in the field, there is a growing concern about its lack of transparency and explain ability, which can limit its utility and undermine the trust in its results. Explainable artificial intelligence (XAI) can address these concerns by providing interpretable models, transparent decision-making processes, and clear explanations of the AI's predictions and recommendations. In this work, we used SHapley Additive exPlanations (SHAP) method, a XAI algorithm borrowed from game theory (29, 30). The SHapley Additive exPlanations (SHAP) method is a model-agnostic approach to interpret the output of any machine learning model. It provides a unified framework for interpreting the predictions of any model by assigning a feature importance value to each input feature (Equation 3).

SHAP method values how a feature affects the performance of the model on the validation set by including and removing it from the model:

$$\Phi j(x) = \sum_{F \subseteq S - \{j\}} \frac{|F|!\,(|S| - |F| - 1)!}{|S|!}\left[fx(F \cup j) - fx(F)\right] \quad (3)$$

where x is an instance, the sum is over all the subsets S of features which include the feature j, $\frac{|F|!(|S|-|F|-1)!}{|S|!}$ is a weight parameter that multiplies all of the permutations of S! by the potential permutations of the remaining class that doesn't belong to S, while $fx(F \cup j)$ and $fx(F)$ denote respectively the regression score obtained by including and non-including feature j.

4 Results

The goal of this work was to evaluate, through explainable machine learning models, whether and to what extent air pollutants and socio-economic descriptors are associated with mortality due to respiratory cancer.

Following the procedures described in the methods section, we delineated 10 clusters when using the spatial range value from the semivariogram plot to split the dendrogram of the geographical distances between the Italian provinces. Then, a leave-one cluster-out validation scheme was adopted in order to get a robust assessment of the model performance. We quantitatively assessed the model performance in terms of the average metrics obtained on the validation set; our model achieved an $r2$ value of 0.28 and a mean absolute error (*M AE*) value of 0.10. Moreover, the importance of all available variables included in the analysis was computed by exploiting the permutation feature importance algorithm and plotted in Figure 5A. Finally, a scatter plot displaying the mutual agreement between the actual and predicted Standard Mortality Ratios is shown in Figure 5B. The provinces characterized by the highest SMR in the 2019 are located in the right-most part of the scatterplot; these are the provinces of Napoli, Caserta, Viterbo, Roma, Piacenza, Imperia, Ravenna, Cagliari, Alessandria and Ferrara.

Figure 6 shows the most important features for regression according to the SHAP algorithm. This summary plot offers an overview of the varying degrees of influence of every feature on the model's predictions, thereby enabling a better grasp of the comprehensive significance and effect of distinct features in the analysis.

According to the SHAP summary plot, the top 5 most important variables include three related to pollution (std NO, mean O_3, mean NO_2), one associated with climate (temperature), and another tied to social factors (income).

5 Discussion

The performances obtained in terms of r2 and MAE are respectively 0.28 and 0.10. This result shows that environmental pollution is associated to the considered type of cancer. From the analysis conducted, it emerges that there is an association between certain pollutants and the incidence of respiratory system cancer.

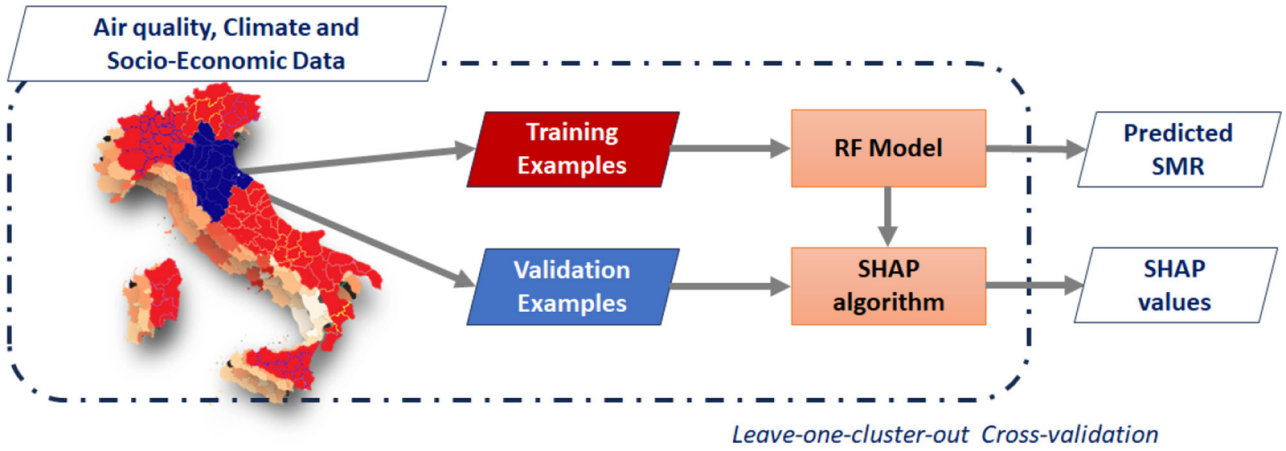

FIGURE 3
Methodological overview: climate, annual average and standard deviation pollution concentrations and socio-economic variables are used to model the standard mortality ratios (SMRs) in the Italian provinces due to cancer of the respiratory system. Then, the SHAP algorithm is used to compute the contribution of each input variable to the predicted SMRs.

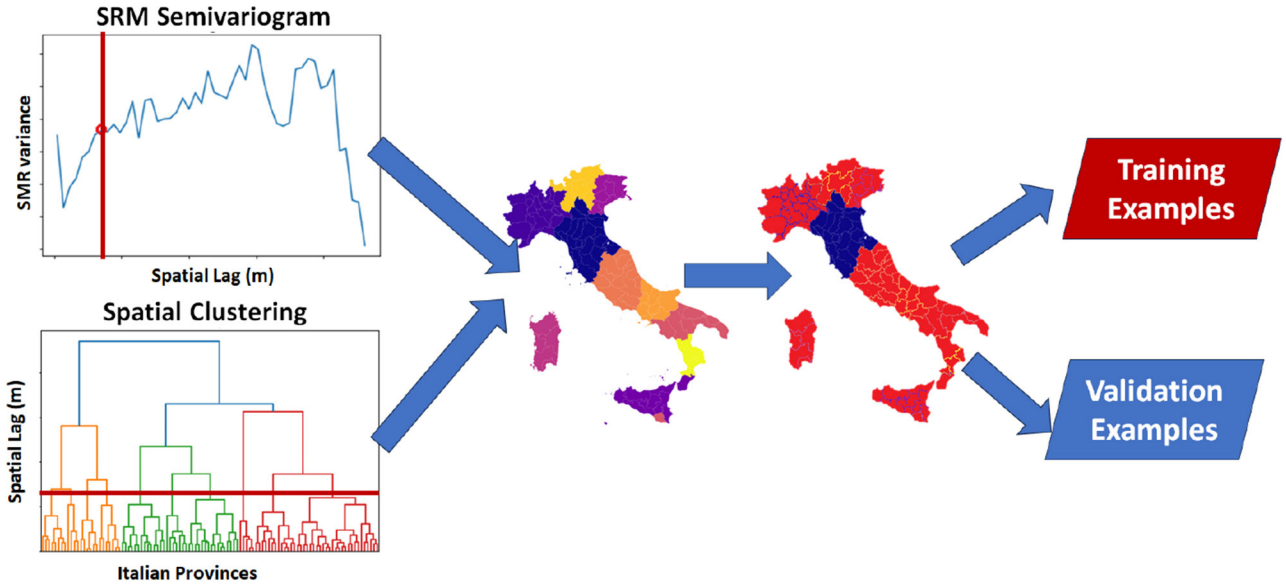

FIGURE 4
The hierarchical clustering algorithm used in combination with the semivariogram plot of the SMR rates delineated 10 spatially-contiguous community that alleviated the spatial data leakage that might occur in epidemiological studies.

This information should be included alongside other risk factors in studies investigating risk factors for individual subjects, including personalized ones (31, 32).

The spatial auto-correlation analysis, performed by using the variogram of the standard mortality ratio, shown ten clusters of provinces. The cross validation performed by using these clusters, obtained an r2 lower than one with a random cross-validation, by showing a spatial bias that overestimated the random forest regressor performance.

XAI estimated which are the most important global and local features in predicting SMR. In particular, for the present case, the main role played by the standard deviation of NO and mean O_3 were revealed. This result is consistent with the literature, as it is known that a greater exposure to NO_2 is correlated with a greater risk of developing lung cancer (33). According to our study, a low average O_3 concentration is linked to a greater SMR for tumors of the respiratory system. Such a spurious association have been observed before in previous works, including Dutton et al. (34) and Travaglio et al. (35), and we believe that a plausible interpretation is that increased ozone exposure is acting a as proxy of residing in rural areas (36, 37). Moreover, it is also known that the concentration of O_3 is anti-correlated to the concentration

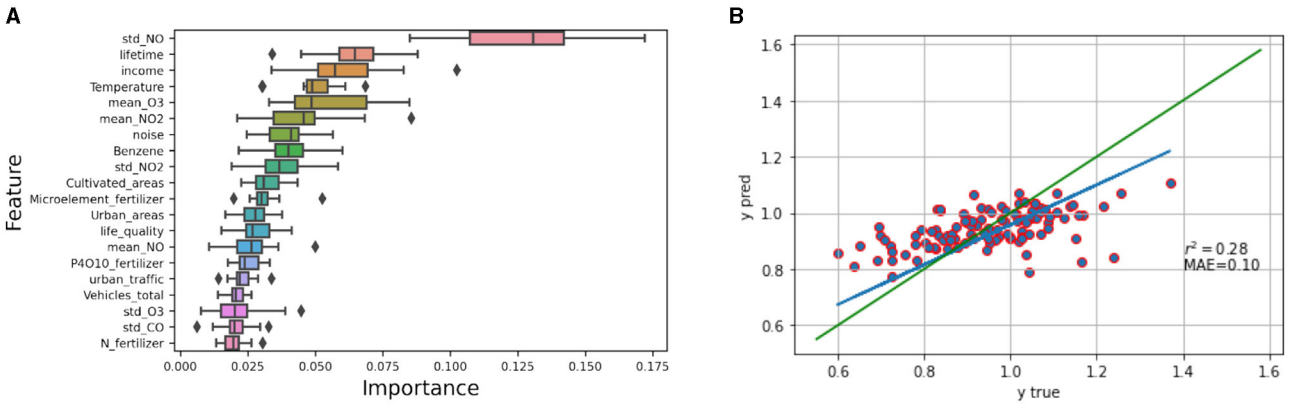

FIGURE 5
(A) Random forest features importance. **(B)** Scatter plot of predicted Standard Mortality Ratio (y pred) vs. true Standard Mortality Ratio (y true). The blue line represents the best fit for the plotted points, whereas the green one represents where the points would fall if all predicted values perfectly matched the observed ones.

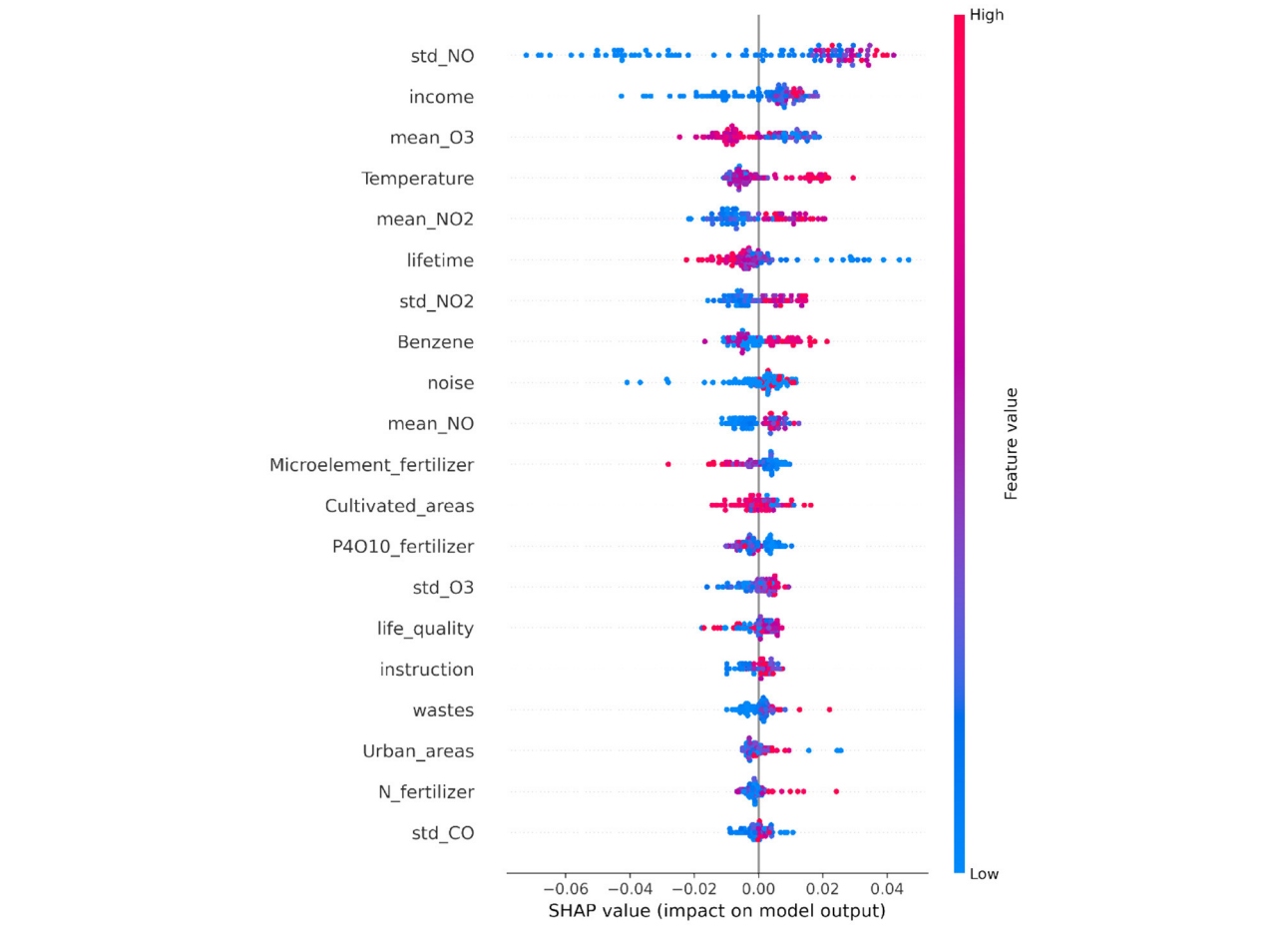

FIGURE 6
SHAP summary plot illustrating the SHAP values for each feature. Each point on the plot represents a Shapley value of a provinces, with the y-axis indicating the corresponding feature and the x-axis representing the Shapley value itself. The color gradient reflects the feature value, ranging from low to high. The features are ordered based on their mean importance, with more important features positioned toward the top.

of NO_2 (36–42). This spatial pattern between rural areas and O_3 concentration is rather general, as O_3 in troposphere is a secondary pollutant that is produced after NO_2 reacting with UV light (40, 43). We also believe that the observed negative correlation between NO_2 and O_3 can be largely attributable to this causal link.

Concerning the socio-demographic descriptors, income and lifetime were the most important. Higher income appears to be positively associated with SMR, implying greater cancer exposure among wealthier individuals. Such association is in stark contrast with the part of the current literature supporting the evidence of how social inequalities may increase exposure to poor air quality and vulnerability to respiratory diseases (34, 44, 45). Richardson et al. (44) claimed evidence of income-related inequalities in exposure to pollutants. Studies conducted to individual-level demographic data that accounts social exclusion and ethnicity (34, 45), evidenced that individual belonging to ethnic minorities are disproportionately exposed to poor air quality. However, this pattern has not been observed in Italy. Germani et al. (46), provided an empirical analysis conducted on the Italian provinces (NUTS3) claiming that air pollution increases with the average income per administrative unit. Since our study relies on provinces, coarse air quality products and no individual-level data is included to account for social and gender, we believe that the average income per province may act as a proxy for both human activities and may not point to social exclusion due to the coarse resolution of the proposed study.

Lifetime duration shows a consistent trend with expected associations: shorter lifespans correspond to higher mortality rates from respiratory diseases. We might speculate that respiratory diseases significantly impact life expectancy in Italy, particularly in certain provinces, potentially influencing the overall health of the population.

The presented study has limitations that we aim to overcome in future research. In particular, the database considered here examined the remote sensing observations of 2019 of exposure to air pollutants. An extension of the time range of our study would guarantee greater robustness to the analyses. Moreover, another limitation of this study is the failure to account for the population density differences within the provinces, as it varies between small urban centers and large cities, implying uniform exposure to air pollutants for the entire population of a province.

6 Conclusion

The presented analysis reveals a correlation between specific pollutants and socio-economic indices and the occurrence of respiratory system cancer. These finding could still offer valuable insights for further epidemiological studies as our results may suggest which variables to gather to perform analyses on individual level dataset that could lead to stronger and more conclusive results. Moreover, this study opens up future prospects for similar research on other types of cancer related to environmental pollution, as well as other types of diseases such as neurodegenerative ones.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

DR: Writing—review & editing, Writing—original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. PN: Writing—review & editing, Validation, Writing—original draft, Methodology, Investigation, Conceptualization. RC: Writing—review & editing, Writing—original draft, Validation, Methodology. NA: Writing—review & editing. AM: Writing—review & editing, Validation, Methodology. RB: Writing—review & editing, Supervision, Methodology, Funding acquisition. ST: Writing—review & editing, Writing—original draft, Visualization, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This paper has been supported by the TEBAKA (TErritorial BAsic Knowledge Acquisition project Avviso MIUR n.1735 del 13/07/2017, the National Institute for Nuclear Physics (INFN), next AIM (Artificial Intelligence in Medicine: next steps) research project (INFN-CSN5), https://www.pi.infn.it/aim (accessed on 30 October 2023); The National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.4-Call for tender No. 3138 of 16 December 2021 of Italian Ministry of University and Research funded by the European Union-NextGenerationEU, award number: Project code: CN00000013, Concession Decree no. 1031 of 17 February 2022 adopted by the Italian Ministry of University and Research, CUP H93C22000450007, Project title: National Centre for HPC, Big Data and Quantum Computing.

Acknowledgments

Authors would like to thank the resources made available by ReCaS, a project funded by the MIUR (Italian Ministry for Education, University and Research) in the PON Ricerca e Competitivita' 2007–2013-Azione I-Interventi di rafforzamento strutturale *PONa*3 00052, Avviso 254/Ric, University of Bari. The authors would like to thank Dr. Nicola Carelli of ARPA-Puglia for discussion of presented analysis results.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Fajersztajn L, Veras M, Barrozo LV, Saldiva P. Air pollution: a potentially modifiable risk factor for lung cancer. *Nat Rev Cancer.* (2013) 13:674–8. doi: 10.1038/nrc3572

2. Huang Y, Zhu M, Ji M, Fan J, Xie J, Wei X, et al. Air pollution, genetic factors, and the risk of lung cancer: a prospective study in the UK Biobank. *Am J Respir Crit Care Med.* (2021) 204:817–25. doi: 10.1164/rccm.202011-4063OC

3. Manisalidis I, Stavropoulou E, Stavropoulos A, Bezirtzoglou E. Environmental and health impacts of air pollution: a review. *Front Pub Health.* (2020) 8:505570. doi: 10.3389/fpubh.2020.00014

4. GBD 2016 Risk Factors Collaborators. Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet.* (2017) 390:1345. doi: 10.1016/S0140-6736(17) 32366-8

5. Pershagen G. Air pollution and cancer. *IARC Sci Pub.* (1990) 104:240–51.

6. Krewski D, Jerrett M, Burnett RT, Ma R, Hughes E, Shi Y et al. *Extended Follow-up and Spatial Analysis of the American Cancer Society Study Linking Particulate Air Pollution and Mortality.* Boston, MA: Health Effects Institute (2009).

7. Cohen AJ, Pope CA. Lung cancer and air pollution. *Environ Health Perspect.* (1995) 103:219–24. doi: 10.1289/ehp.95103s8219

8. Santos UDP, Arbex MA, Braga ALF, Mizutani RF, Cançado JED, Terra-Filho M, et al. Environmental air pollution: respiratory effects. *J Brasileiro de Pneumologia.* (2021) 47:e20200267. doi: 10.36416/1806-3756/e20200267

9. Hastie T, Tibshirani R, Friedman JH, Friedman JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York, NY: Springer (2009).

10. Lombardi A, Diacono D, Amoroso N, Monaco A, Tavares JMR, Bellotti R, et al. Explainable deep learning for personalized age prediction with brain morphology. *Front Neurosci.* (2021) 15:674055. doi: 10.3389/fnins.2021.674055

11. Lombardi A, Diacono D, Amoroso N, Biecek P, Monaco A, Bellantuono L, et al. A robust framework to investigate the reliability and stability of explainable artificial intelligence markers of Mild Cognitive Impairment and Alzheimer's Disease. *Brain Inf.* (2022) 9:17. doi: 10.1186/s40708-022-00165-5

12. Amoroso N, Quarto S, La Rocca M, Tangaro S, Monaco A, Bellotti R. An explainability artificial intelligence approach to brain connectivity in Alzheimer's disease. *Front Aging Neurosci.* (2023) 15. doi: 10.3389/fnagi.2023.1238065

13. Thunis P, Degraeuwe B, Pisoni E, Meleux F, Clappier A. Analyzing the efficiency of short-term air quality plans in European cities, using the CHIMERE air quality model. *Air Q Atmosp Health.* (2017) 10:235–48. doi: 10.1007/s11869-016-0427-y

14. Hass H, Ebel A, Feldmann H, Jakobs HJ, Memmesheimer M. Evaluation studies with a regional chemical transport model (EURAD) using air quality data from the EMEP monitoring network. *Atmosp Environ Part A Gen Topics.* (1993) 27:867–87. doi: 10.1016/0960-1686(93)90007-L

15. Duarte EDSF, Franke P, Lange AC, Friese E, da Silva Lopes FJ, da Silva JJ, et al. Evaluation of atmospheric aerosols in the metropolitan area of São Paulo simulated by the regional EURAD-IM model on high-resolution. *Atmos Pollut Res.* (2021) 12:451–69. doi: 10.1016/j.apr.2020.12.006

16. Hinestroza-Ramirez JE, Lopez-Restrepo S, Yarce Botero A, Segers A, Rendon-Perez AM, Isaza-Cadavid S, et al. Improving air pollution modelling in complex terrain with a coupled WRF–LOTOS–EUROS approach: a case study in Aburrá Valley, Colombia. *Atmosphere.* (2023) 14:738. doi: 10.3390/atmos140 40738

17. Persson C, Langner J, Robertson L. Air pollution assessment studies for Sweden based on the MATCH model and air pollution measurements. *Air Pollut Modeling Appl.* (1996) 9:127–34. doi: 10.1007/978-1-4615-5841-5_15

18. Joly M, Josse B, Plu M, Arteta J, Guth J, Meleux F. *High-Resolution air Quality Forecasts With MOCAGE Chemistry Transport Model. Air Pollution Modeling and its Application XXIV 2016.* Cham: Springer International Publishing (2016). p. 563–5.

19. Ots R, Loot A, Kaasik M. Scale-dependent and seasonal performance of SILAM model in Estonia. In: *Air Pollution Modeling and its Application XXII 2014.* Geneva: Springer Netherlands (2014). p. 593–7.

20. Van Loon M, Vautard R, Schaap M, Bergström R, Bessagnet B, Brandt J, et al. Evaluation of long-term ozone simulations from seven regional air quality models and their ensemble. *Atmospheric Environ.* (2007) 41:2083–97. doi: 10.1016/j.atmosenv.2006.10.073

21. Neary L, Kaminski JW, Lupu A, McConnell JC. Developments and Results From a global multiscale air quality model (GEM-AQ). In: *Air Pollution Modeling and Its Application XVII.* Cham: Springer (2007). p. 403–10.

22. Klose M, Jorba O, Gonçalves Ageitos M, Escribano J, Dawson ML, Obiso V, Di Tomaso E. Mineral dust cycle in the Multiscale Online Nonhydrostatic AtmospheRe CHemistry model (MONARCH) version 2.0. *Geosci Model Dev Discussions.* (2021) 2021:1–59. doi: 10.5194/gmd-14-6403-2021

23. Mircea M, Zanini G, Briganti G, Cappelletti A, Pederzoli A, Vitali L, et al. Modeling Air Quality Over Italy With MINNI Atmospheric Modeling System: From Regional to Local Scale. In: *Air Pollution Modeling and its Application XXI 2012.* Cham: Springer (2012). p. 491–8.

24. Gatti RC, Di Paola A, Monaco A, Velichevskaya A, Amoroso N, Bellotti R. The spatial association between environmental pollution and long-term cancer mortality in Italy. *Sci Total Environ.* (2023) 855:158439. doi: 10.1016/j.scitotenv.2022.158439

25. Wilcox AJ, Russell IT. Birthweight and perinatal mortality: III. Towards a new method of analysis. *Int J Epidemiol.* (1986) 15:188–96. doi: 10.1093/ije/15.2.188

26. Breiman L. Random forests. *Mach Learn.* (2001) 45:5–32. doi: 10.1023/A:1010933404324

27. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion.* (2020) 58:82–115. doi: 10.1016/j.inffus.2019.12.012

28. Bisong E, Bisong E. *Introduction to Scikit-learn. Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*, 215–229. doi: 10.1007/978-1-4842-4470-8_18

29. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in neural information processing systems. (2017) p. 30.

30. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Int.* (2020) 2:56–67. doi: 10.1038/s42256-019-0138-9

31. Malhotra J, Malvezzi M, Negri E, La Vecchia C, Boffetta P. Risk factors for lung cancer worldwide. *Eur Resp J.* (2016) 48:889–902. doi: 10.1183/13993003.00359-2016

32. Parikh AR. Lung Cancer Genomics. *Acta Medica Academica.* (2019) 48:244. doi: 10.5644/ama2006-124.244

33. Nyberg F, Gustavsson P, Järup L, Bellander T, Berglind N, Jakobsson R, et al. Urban air pollution and lung cancer in Stockholm. *Epidemiology.* (2000) 11:487–95. doi: 10.1097/00001648-200009000-00002

34. Dutton A. *Coronavirus (COVID-19) Related Mortality Rates and the Effects of Air Pollution in England.* London: Office of National Statistics (2020).

35. Travaglio M, Yu Y, Popovic R, Selley L, Leal NS, Martins LM. Links between air pollution and COVID-19 in England. *Environ Pollut.* (2021) 268:115859. doi: 10.1016/j.envpol.2020.115859

36. Melkonyan A, Kuttler W. Long-term analysis of NO, NO_2 and O_3 concentrations in North Rhine-Westphalia, Germany. *Atmos Environ.* (2012) 60:316–26. doi: 10.1016/j.atmosenv.2012.06.048

37. Hagenbjörk A, Malmqvist E, Mattisson K, Sommar NJ, Modig L. The spatial variation of O_3, NO, NO_2 and NO x and the relation between them in two Swedish cities. *Environ Monit Assess.* (2017) 189:1–12. doi: 10.1007/s10661-017-5872-z

38. Domínguez-López D, Adame JA, Hernández-Ceballos MA, Vaca F, De la Morena BA, Bolívar JP. Spatial and temporal variation of surface ozone, NO and NO 2 at urban, suburban, rural and industrial sites in the southwest of the Iberian Peninsula. *Environ Monitor Assessm.* (2014) 186:5337–51. doi: 10.1007/s10661-014-3783-9

39. Clapp LJ, Jenkin ME. Analysis of the relationship between ambient levels of O_3, NO_2 and NO as a function of NOx in the UK. *Atmos Environ.* (2001) 35:6391–405. doi: 10.1016/S1352-2310(01)00378-8

40. Gorrochategui E, Hernandez I, Tauler R. A model for simultaneous evaluation of NO_2, O_3, and PM 10 pollution in urban and rural areas: handling incomplete

data sets with multivariate curve resolution analysis. *Atmospheric Chem Phys.* (2022) 22:9111–27. doi: 10.5194/acp-22-9111-2022

41. Fernández-Guisuraga JM, Castro A, Alves C, Calvo A, Alonso-Blanco E, Blanco-Alegre C, et al. Nitrogen oxides and ozone in Portugal: trends and ozone estimation in an urban and a rural site. *Environ Sci Pollut Res.* (2016) 23:17171–82. doi: 10.1007/s11356-016-6888-6

42. Safieddine S, Clerbaux C, George M, Hadji-Lazaro J, Hurtmans D, Coheur PF, et al. Tropospheric ozone and nitrogen dioxide measurements in urban and rural regions as seen by IASI and GOME-2. *J Geophys Res Atmos.* (2013) 118:10–555. doi: 10.1002/jgrd.50669

43. Reeves CE, Penkett SA, Bauguitte S, Law KS, Evans MJ, Bandy BJ, Kley D. Potential for photochemical ozone formation in the troposphere over the North Atlantic as derived from aircraft observations during ACSOE. *J Geophys Res Atmos.* (2002) 107:ACH-14. doi: 10.1029/2002JD002415

44. Richardson EA, Pearce J, Tunstall H, Mitchell R, Shortt NK. Particulate air pollution and health inequalities: a Europe-wide ecological analysis. *Int J Health Geogr.* (2013) 12:1–10. doi: 10.1186/1476-072X-12-34

45. van den Brekel L, Lenters V, Mackenbach JD, Hoek G, Wagtendonk A, Lakerveld J, et al. Ethnic and socioeconomic inequalities in air pollution exposure: a cross-sectional analysis of nationwide individual-level data from the Netherlands. *The Lancet Planetary Health.* (2024) 8:e18–29. doi: 10.1016/S2542-5196(23)00258-9

46. Germani AR, Morone P, Testa G. Environmental justice and air pollution: a case study on Italian provinces. *Ecol Econ.* (2014) 106:69–82. doi: 10.1016/j.ecolecon.2014.07.010