LIFESTYLE, ENVIRONMENTAL CONDITIONS AND MORTALITY IN EUROPEAN COUNTRIES AND IN ITALIAN REGIONS

Simona Cafieri, Gianmarco Borrata

Abstract. This paper uses state-of-the-art machine learning techniques to study the relationship between environmental pollution, life expectancy and lifestyle variables in European countries, with a focus on the Italian regions.
K-means clustering allowed to analyse the impact of the pandemic on socio-economic variables between 2019 and 2021, showing how countries position themselves with respect to these changes.
Different regional typologies were outlined, reflecting the diversity of environmental and health challenges.
Furthermore, a random forest analysis was used to predict life expectancy in European countries and Italian regions based on the presence of the most polluting and health-damaging substances.
This methodological approach offers new ways of identifying priorities for intervention, combining environmental mitigation with targeted prevention and treatment strategies.

1. Introduction

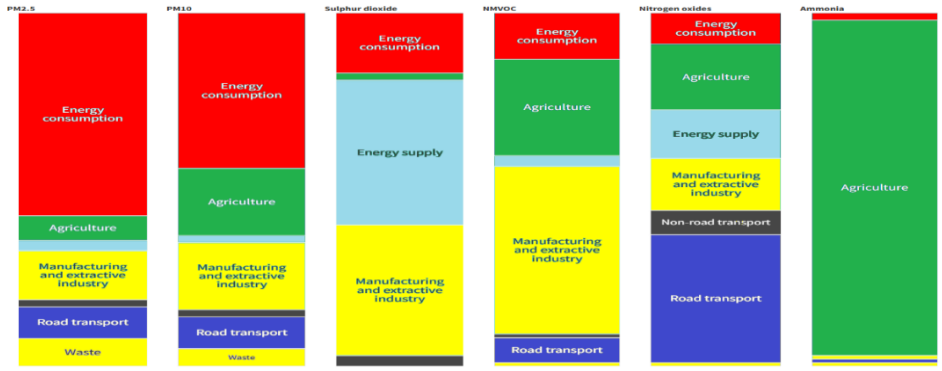
According to the World Health Organization, Europe is the most affected region in the world by non-communicable diseases such as cancer, cardiovascular diseases and chronic respiratory diseases, with a relatively small group of health conditions accounting for a large proportion of the disease burden.

Since 2013, European Commission reports have identified the main causes and effects of health inequalities, including living conditions, health-related behaviours, education, occupation and income. The report recognises the role that exposure to air pollution can play in health inequalities.

Air pollution is the most important environmental health risk factor, independent of an individual's lifestyle. It remains a major cause of ill health, contributing not only to the development of cancers but also to respiratory and cardiovascular diseases, and is thought to be responsible for more than 6.5 million deaths per year worldwide.

According to the European Environment Agency (EEA), 253,000 deaths in the EU-27 in 2021 were attributable to exposure to concentrations of fine particulate matter (PM2.5), and coarse particulate matter (PM10). A further 52,000 deaths were associated with exposure to other key air pollutants, such as high NO2 (nitrogen dioxide) and O3 (ozone) concentrations. According to the EEA, energy consumption in residential, commercial and institutional buildings was the main source of PM10 and PM2.5. Agriculture was the main source of ammonia and methane, accounting for 94% and 55% of emissions respectively.

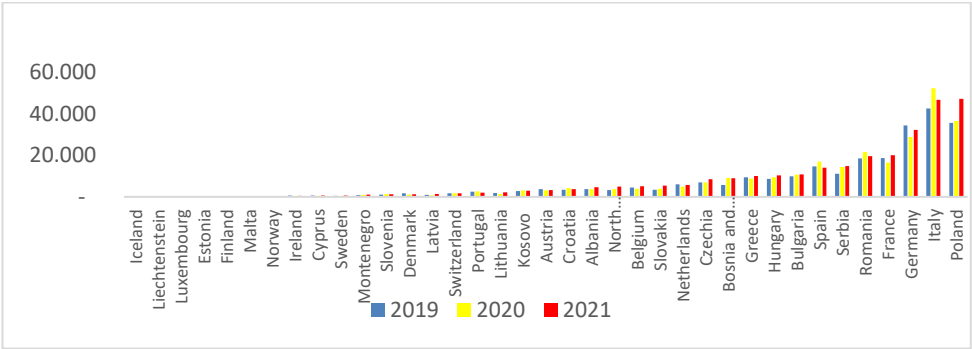
Figure 1 – *Main sources of air pollutants in Europe, 2021.*



Source: European Environment Agency.

In 2021 road transport was the main source of nitrogen oxides, accounting for 39% of emissions. In the same year, the energy sector was the main source of sulphur dioxide, responsible for 46% of emissions. Manufacturing and extractive industries were the main emitters of heavy metals to air.

Figure 2 – *N. of deaths attributable to the environment in EU countries (2019 -2021).*



Source: European Environment Agency

Figure 2 presents the number of deaths attributable to the environment: Eastern European countries carry a significantly higher environmental burden of disease and mortality than western European countries, exacerbating economic inequalities across the European region. The relationship between socio-economic status and health inequality is unequivocal[1].

[1] Pickett and Wilkinson, (2015)

Societies with wide disparities in socio-economic status also have wide disparities in health and mortality outcomes. Social disparities are evident in both lifestyle choices and dietary patterns. As incomes decline and household budgets tighten, food choices often shift towards cheaper, more energy-dense options. High-quality protein, whole grains, fruits, and vegetables are typically the first to be sacrificed. Over time, this dietary pattern can lead to an increased risk of chronic diseases and premature mortality.

2. Methodology: preliminary consistency analysis

This study aims to identify potential significant correlations between environmental pollution, life expectancy and lifestyle variables using cutting-edge machine learning techniques. For a preliminary consistency analysis, a K-means cluster analysis has been applied to look for likely relationships within the data without any prior assumption. The K-Means clustering technique is an unsupervised learning algorithm widely used in data science to identify hidden structures in data and break them down into homogeneous groups.K-Means algorithm, minimizing within-cluster inertia, partitions quantitative observations into K clusters, with K random initial centroids, thus associating each observation to the nearest centroid according to the Euclidean distance:

$$\mathrm{dE} = \sum\sqrt{(\mathrm{xi} - \mathrm{yi})2\mathrm{Ni} = 1} \quad (1)$$

Where xi and yi are coordinates of two observations, X = (x1,2, ..., xN) and Y = (y1,y2, ..., yN), respectively. Then, the algorithm computes the centroids of the new groupings iteratively until convergence.

The K-means algorithm was implemented for regional analysis using independent variables that concern some socioeconomic and lifestyle variables at the regional scale:

The feature **"Meat consumption"** derives from five features concerning the consumption of animal-source food meats, defined as the percentage of individuals who declare that they - consume cured meats, chicken, turkey, beef, pork, more than once a week[2].

The five variables have been summarized on animal-source food into a single feature ("Meat consumption"), namely the first component resulting from a **Principal Component Analysis**.

PCA are made as a linear combination, orthogonal to each other, of the dataset features; the first principal component PC1 represents a large fraction of variation (variance explained) in the sample, and successive PCs account for decreasing portions of the remaining variation.

[2] For Italy data are provided by the 2021 Aspects of Daily Life (AVQ) Istat survey, Data for other countries, are provided by surveys carried out from other National Institutes of Statistics and collected by Eurostat. in Regions database

Therefore, to reduce the variability to get a good representation of the data and to build the feature **"Meat consumption"**, only the first few PCs have been used only PC1, which contained almost 80 % of the data variance.

Table 1 – *Independent variables relating to socio-economic aspects and lifestyle.*

Variable	Description
Environmental pollution	the value of the annual average concentration of PM10 and PM2.5
Available hospital beds	the average number of beds in public hospitals, per regional population
Income	the mean annual income per family provided by the EuSilc survey
Life expectancy	an expect to live at birth
Overweight	the average percentage of people in excess weight (overweight and obese individual[3])
Smoke	the average percentage of smokers
Fruit and vegetable consumption	the percentage of subjects who declare that they eat fruit or vegetables at least once a day

3. Results of European regions between 2019 and 2021

As a result of the K-means cluster analysis for European regions based on environmental and socio-economic variables in 2019 and 2021, three clusters were identified as the optimal number according to the Elbow method[4].

Cluster 1 (Red in Fig.3): Cluster 1 includes Eastern European countries such as Romania, Bulgaria, and Hungary. These states remain characterized by socio-economic challenges, including lower income levels and limited access to advanced healthcare services. Unfavorable health behaviors such as higher smokers' prevalence and overweight rates are notable in this group. Additionally, this cluster shows moderate life expectancy and significant environmental challenges (Fig: 4a, 4b). Between 2019 and 2021,Cipro and Greece join this cluster.

Cluster 2 (Blue in Fig: 3) predominantly includes Western European countries such as France, and Switzerland. These countries demonstrate strong socio-economic indicators, such as high income levels and advanced healthcare systems., lower smokers' rates and higher fruit and vegetable consumption, environmental pollution and meat consumption are moderate concerns in some countries within this group, while life expectancy is generally high (Fig: 4c, 4d). Between 2019 and 2021 Norway, Germany, Belgium and Austria will leave this cluster, while Italy, Spain and Sweden will enter.

Cluster 3 (Green in Fig.3): This cluster features Nordic and Southern European countries. These states present a balanced profile, with moderate to high income levels and positive lifestyle factors (Fig: 4e, 4f). Life expectancy remains relatively high overall, reflecting positive health outcomes in these states.

[3] Overweight" means people with a body mass index (BMI) between 25 and 29,9 while people with a BMI of 30 are included in the "obese" group.

[4] Elbow

Figure 3 – *Cluster in EU regions in 2019 and 2021*

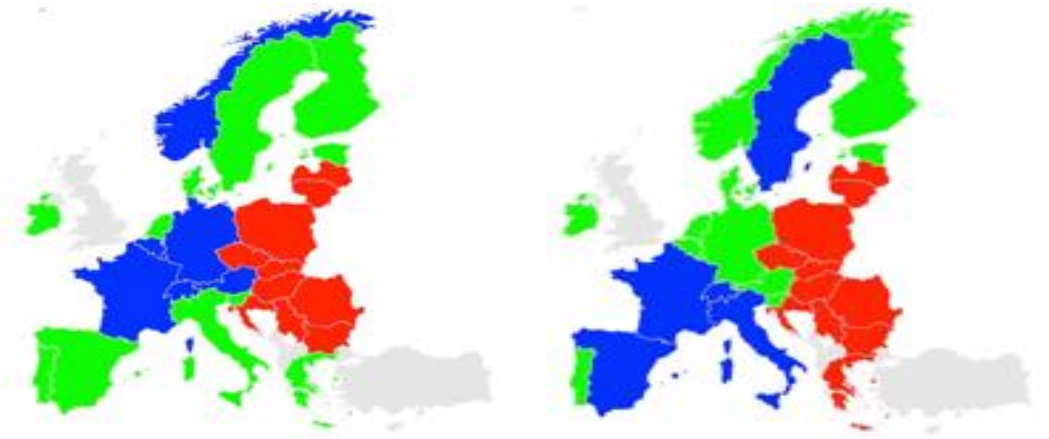
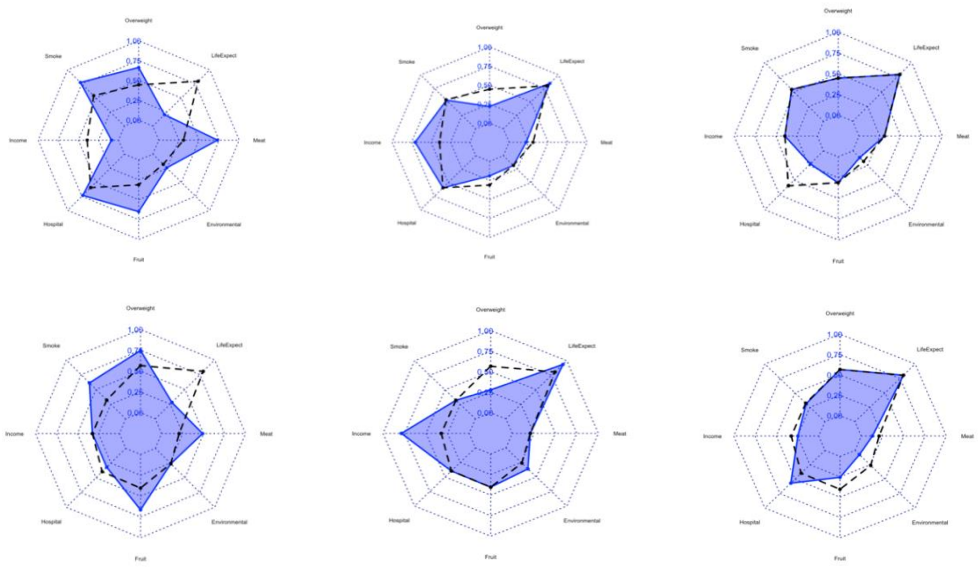


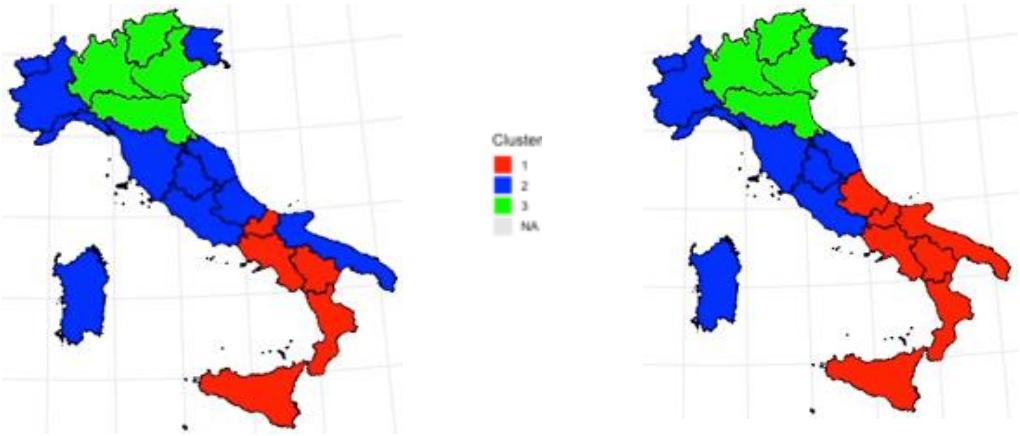
Figure 4 – *Cluster characteristics in the EU regions in 2019(a,c,e) and 2021(b,d,f)*



4. Focus on Italy

Focusing on the Italian regions to perform a cluster analysis based on environmental and socio-economic variables in 2019 and 2021, again using Elbow methods, three clusters were identified as the optimal number.

Figure 5 – *Cluster in Italian regions in 2019 and 2021*.



Cluster 1:(Red in Fig: 5): Composed of southern regions, this cluster highlights socio-economic challenges. It is characterized by lower income levels, higher smoking prevalence, and significant rates of overweight individuals and meat consumption. The life expectancy is very low in this group. (Fig: 6a, 6b)

Cluster 2: (Blue in Fig: 5): Includes the north-west and central regions, which have moderate socio-economic indicators. This cluster is characterized by healthier lifestyles, including lower prevalence of overweight individuals and high consumption of fruit and vegetables, reflecting a relatively balanced health profile, with high life expectancy. (Fig: 6c, 6d)

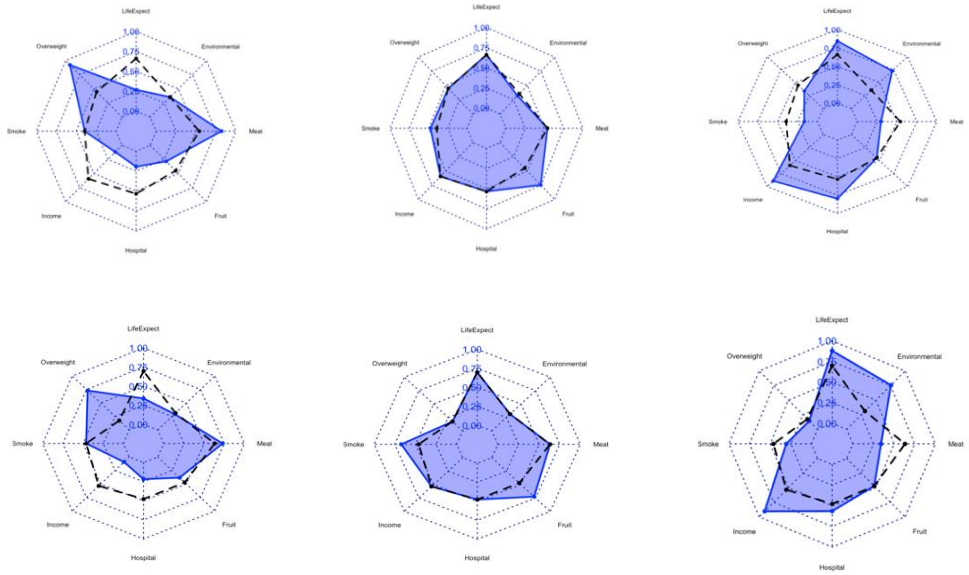
Cluster 3 (Green in Fig: 5): This cluster groups northern regions. These regions demonstrate high income levels and favorable health behaviors, such as higher fruit and vegetable consumption and lower smoking prevalence with life expectancy remaining notably high compared to other regions. (Fig: 6e, 6f).

5. A deeper level of analysis: prediction of future mortality incidence

To delve deeper into the specific air pollutants linked to mortality in EU and Italian regions, a secondary analysis was conducted using a Random Forest[5] model.

[5] Babu S, Thomas B (2023)

Figure 6 – *Cluster characteristics in the Italian regions in 2019(a,c,e)and 2021(b,d,f).*



The Random Forest technique is a regression tree technique which combines multiple decision trees and randomization of predictors to achieve a high degree of predictive accuracy. Each tree in the forest builds from a different subset of the data and makes its independent prediction. The final prediction for input is based on the average of all the individual trees' predictions. The RF regression is one of the most popular machine learning algorithms and has been success fully applied to both classification and regression in many different fields.

A Random Forest analysis was developed to predict life expectancy in the six identified clusters, using major environmental pollutants (associated to life expectancy reduction) such as PM2.5, PM10, O3 and NO2 as predictors according to EEA and Ispra data.

The prediction was made from socio-economic data and environmental variables in 2021. The importance of each variable and its predictive accuracy were assessed. The evaluation used leave-one-out cross-validation, highlighting the predictive ability of the model within each cluster.

Table 2 – *Mean Values of Air Pollutants for Different Clusters of Italian Regions.*

	NO2	O3	PM10	PM2.5
Cluster 1	15.55	54.39	22.10	11.53
Cluster 2	16.27	57.98	19.95	11.87
Cluster 3	22.54	53.12	24.26	16.59

Table 3 – *Mean Values of Air Pollutants for Different Clusters of European States.*

	NO2	O3	PM10	PM2.5
Cluster 1	7.80	61.77	16.76	12.84
Cluster 2	5.88	60.74	14.52	8.25
Cluster 3	6.59	65.75	13.94	7.73

After identifying the relevant characteristics (pollutants), cleaning and pre-processing the data, dealing with missing values and normalizing pollutant levels, the following steps were taken:

- Model Training: Data are split into training and test sets. Random Forest model is trained with pollutants as features and life expectancy rates as the target.
- Model Evaluation: The model's performance is assessed on the test set using metrics like Mean Squared Error (MSE)[6] Mean absolute error (MAE)[7] and Root Mean Squared Error **(**RMSE**)[8] .**
- Future Data Prediction: future pollutant data are collected and input into the trained model.
- Life expectancy prediction: The model predicts future life expectancy based on the input pollutant data, providing insights into potential future health impacts in the region based on the levels of the main pollutants detected in 2021 according to EEA data.

The following table illustrates how the model's predictions assesses the model's accuracy using the MSE, MAE, RMSE

Table 4 – *Performance Measures to predict Life Expectancy in European countries.*

	Cluster 1	Cluster 2	Cluster 2
MSE	0.038	0.003	0.018
MAE	0.145	0.056	0.097
RMSE	0.196	0.059	0.136

In Cluster 1, NO2 was identified as the most significant predictor, with PM10 and PM2.5 also contributing notably. O3 had a limited influence on life expectancy predictions (Fig: 7a). This cluster demonstrated a MSE of 0.038, MAE of 0.145, and RMSE of 0.196, indicating reliable predictions (Tab: 4).

6 $MSE = \frac{1}{n}\sum e_t^2$

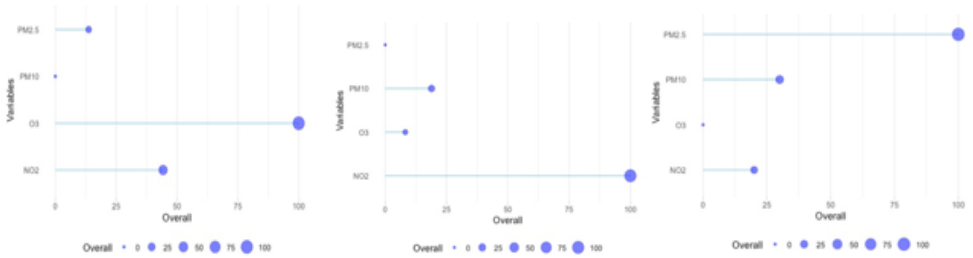
7 $MAE = \frac{1}{n}\sum |e_t|$

8 $RMSE = \sqrt{\frac{1}{n}\sum e_t^2}$

For Cluster 2, PM2.5 emerged as the most critical variable, highlighting the substantial impact of fine particulate matter on life expectancy. The variables PM10, and NO2 played meaningful roles, while O3 had minimal relevance (Fig: 7b). This cluster achieved the best predictive performance, with the lowest MSE (0.003), MAE (0.056), and RMSE (0.059), reflecting highly accurate predictions (Tab: 4).

In Cluster 3, NO2 and O3 were the dominant variables, underlining their importance in regions characterized by diverse environmental conditions. PM2.5 and PM10 were less influential compared to other clusters (Fig: 7c). The predictive metrics for this cluster were strong, with an MSE of 0.018, MAE of 0.097, and RMSE of 0.136, showcasing good model performance (Tab: 4).

Figure 7 – *Predictors of life expectancy in EU regions grouped in 3 clusters(a, b, c).*



A similar Random Forest model can be build to predict life expectancy by analyzing the three clusters previously identified, with environmental variables levels (main pollutants detected in 2021) as predictors according to Ispra data.

Table 5 – *Performance Measures to predict life expectancy in Italian regions.*

	Cluster 1	Cluster 2	Cluster 2
MSE	0,032	0,011	0,008
MAE	0,124	0,104	0,084
RMSE	0,178	0,104	0,089

The predictions were evaluated using the leave-one-out cross-validation method, and the results provide insights into the predictive performance and importance of different variables across clusters.

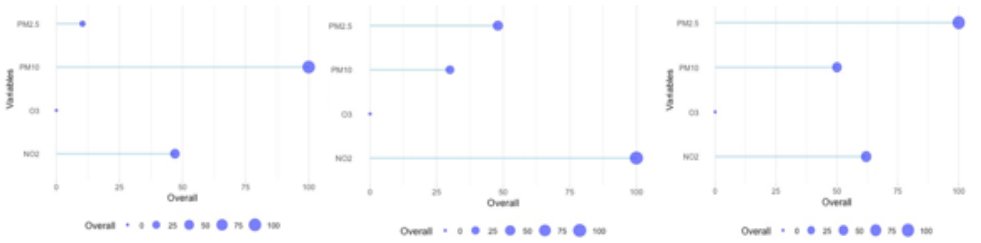
In Cluster 1, the PM10 was the most critical predictor, with the highest importance score, indicating its strong association with life expectancy in this cluster. NO2 and PM2.5 also played significant roles, while O3 was less influential (Fig: 8d) Cluster 1 exhibited the highest error

metrics among the three groups, with an MSE of 0.032, an MAE of 0.124, and an RMSE of 0.178, but the overall predictive performance is still acceptable.

For the cluster 2, NO2 emerged as the most important predictor, emphasizing the role of traffic-related pollution in influencing mortality rates. PM10 and PM2.5 were also relevant, while O3 showed minimal importance (Fig: 8e). Cluster 2 performed well, with an MSE of 0.011 and an RMSE of 0.104.

Finally, in cluster 3 the PM2.5 was the dominant variable, reflecting the relevance of fine particulate matter in regions with higher industrial activities. PM10 and NO2 followed in importance, with O3 having a negligible impact (Fig: 8c). Cluster 3 achieved the best results, showing the lowest error metrics, including a MSE of 0.008, a MAE of 0.084, and a RMSE of 0.089.

Figure 8 – *Predictors of life expectancy in EU regions grouped in 3 clusters(d, e, f).*



6. Results and discussion

The results highlight the crucial role of environmental and socio-economic conditions in determining health inequalities. Regional classification using K-means clustering allowed the identification of three macro-groups with distinct socio-economic and health characteristics. Eastern European regions, although less polluted than some in Western Europe, show greater vulnerability due to lower income levels and limited access to health services. Conversely, Western European countries, while benefiting from advanced health care systems and generally healthier lifestyles, face growing concerns about air pollution, as evidenced by the effects of fine particulate matter (PM2.5 and PM10).

An important finding is the importance of particulate matter and nitrogen dioxide (NO2) as primary predictors of mortality and life expectancy, independent of socio-economic factors. This underlines the fact that in high-pollution contexts, even a healthy lifestyle cannot offset the negative effects of prolonged exposure to pollutants. In Italy, regional disparities are particularly pronounced: northern regions show better health indicators due to higher incomes and healthier lifestyles, while southern regions face a combination of adverse factors, including lower incomes and higher rates of obesity and smoking. The Random Forest analysis confirmed the key role of specific pollutants, with PM2.5 emerging as the most significant predictor of reduced life expectancy in many clusters. These findings underscore the importance of targeted

interventions to reduce air pollution and integrated policies addressing both socio-economic inequalities and environmental risks.

7. Conclusions

This paper demonstrated the effectiveness of advanced techniques, such as K-means clustering and the Random Forest model, in identifying relationships between environmental pollution, socio-economic factors, and life expectancy. The emergence of PM2.5 and NO2 as key predictors highlights the need for specific policies aimed at improving air quality, particularly in the most vulnerable areas.

These findings underscore the importance of targeted interventions to reduce air pollution and integrated policies addressing both socio-economic inequalities and environmental risks.

Moreover, the observed disparities call for focused strategies to promote health equity and to reduce pollution. A polluted environment can impede our efforts to achieve optimal health, just as a polluted train carriage can impede our comfort on the journey of our lives.

References

BABU S, THOMAS B**.** 2023.A survey on air pollutant PM2.5 prediction using random forest model**,** *Environmental Health Engineering and Management Journal*, Vol. 10, No. 2, pp. 157–163

BALAKRISHNAN K., FULLER R., AND OTHER. 2022. Pollution and health: a progress update, *The Lancet Health Review*, Vol. 6, No. 6, e535-e547.

BREIMAN L**.** 1996**.** Bagging predictors. *Machine Learning.* Vol. 24, pp. 123–140. Springer, Berlin/Heidelberg, Germany.

BREIMAN L. 2001. Random forests, *Machine Learning*, pp. 5–32.

DI LORENZO G., FEDERICO P., DE PLACIDO S., BUONERBA C. 2015 Increased risk of bladder cancer in critical areas at high pressure of pollution of the Campania region in Italy: a systematic review, *Critical reviews in oncology/hematology*, Vol. 96, pp. 534–541

EUROPEAN ENVIRONMENT AGENCY 2019 *Healthy Environment, healthy lives: how the Environment Influences Health in Europe*, Report N. 21/2019

EUROPEAN ENVIRONMENT AGENCY 2023 *Air Quality e-Reporting*

GURURAJ T., VISHRUTHA Y. M., UMA M., RAJESHWARI D., RAMYA B. K. 2020. Prediction of Lung Cancer Risk using Random Forest Algorithm Based on Kaggle Data Set. *International Journal of Recent Technology and Engineering (IJRTE)*, Vol. 8, No. 6, March 2020. ISSN: 2277-3878 (Online),

IARC 2022. https://monographs.iarc.who.int/agents-classified-by-the-iarc/

ISTAT 2022. *Aspetti della vita quotidiana.*

ISTAT 2021. *Dati ambientali nelle città.*

ISTAT 2021. *Indagine sulle cause di morte.*

ISTAT 2021. *Indagine sui prodotti fitosanitari destinati all'uso agricolo.*

JAIN A.K., DUBES RC. 1988. *Algorithms for Clustering Data.* Prentice-Hall, New Jersey.
JACKSON E.J.1991, *A User's guide to principal components*, John Wiley &Sons, p.569.
KASSOMENOS, P., PETRAKIS M., SARIGIANNIS D., GOTTI A., KARAKITSIOS S. 2011. Identifying the contribution of physical and chemical stressors to the daily number of hospital admissions implementing an artificial neural network model, *Air Quality, Atmosphere & Health,* Vol, 4, pp. 263–272.
LIAW A., WIENER M. 2002. Classification and regression by random Forest, *R News,* Vol. 2 , No. 3, pp.18–22.
LIN Y.-C., TSAI C.-H., HSU H.-T., LIN C.-H. 2021. Using Machine Learning to Analyze and Predict the Relations Between Cardiovascular Disease Incidence, Extreme Temperature and Air Pollution. *IEEE 3rd Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS)*, pp. 234–237.
MARIEN L., VALIZADEH M., ZU CASTELL W., NAM C., RECHID D., SCHNEIDER A., MEISINGER C., LINSEISEN J., WOLF K., BOUWER L. 2022. Machine learning models to predict myocardial infarctions from past climatic and environmental conditions, *Natural Hazards and Earth System Sciences*, pp. 1–36.
PAVONE P., PAGLIACCI F., RUSSO M., RIGHI S., GIORGI A. 2021. Multidimensional clustering of EU regions. A contribution to orient public policies in reducing regional disparities, *Social Indicators Research*, April 2020.
PICKETT K. E, WILKINSON R. G. 2015. Income inequality and health: A causal review, *Social Science & Medicine*, Vol. 128, pp. 316-326.
QUINLAN J.R. 1986. Introduction of decision trees, *Machine Learning*, Vol. 1 , pp. 81–106.
SYAKUR M.A., KHOTIMAH B.K, ROCHMAN E.M.S, SATOTO B.D. 2018, Integration K-means clustering method and Elbow method for identification of the best customer profile cluster. In *IOP conference series: materials science and engineering*, Vol. 335, pp. 12-17.

Simona CAFIERI, Istat, cafieri@istat.it
Gianmarco BORRATA, Università di Napoli "Federico II", gianmarco.borrata@unina.it