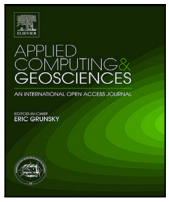
Applied Computing and Geosciences 23 (2024) 100184

Contents lists available at ScienceDirect

Applied Computing and Geosciences

journal homepage: www.sciencedirect.com/journal/applied-computing-and-geosciences



An in-depth multivariate analysis of $PM_{2.5}$ concentration and associated premature deaths in Europe and its strategic relationship with sustainability

Laura Sanz-Martín *, Javier Parra-Domínguez, Juan Manuel Corchado

Universidad de Salamanca, Edificio I+D+i, C/ Espejo, s/n, Salamanca, 37007, Salamanca, Spain

ARTICLE INFO

MSC:
62P12
62G10
62M10

Keywords:
$PM_{2.5}$
Air pollution
Machine learning
Public health
Premature deaths
Sustainability

ABSTRACT

The strategic importance of sustainability is evident when it comes, for example, to health. Public policies aimed at mitigating the effects of harmful substances, such as fine particulate matter (PM 2.5), are justified by the direct link between fine particulate matter and the health of citizens, in this case, premature deaths. An advanced statistical and exhaustive analysis of different areas and countries shows a strong link between exposure to $PM_{2.5}$, premature deaths in other countries, and significant differences in $PM_{2.5}$ levels between urban and rural areas.

Although $PM_{2.5}$ concentration has decreased in most countries studied, this effort must be continued and aligned with the Sustainable Development Goals of the 2030 Agenda, underlining the need to implement effective air pollution control policies to reduce the health risks associated with $PM_{2.5}$ exposure. To this end, identifying temporal trends and geographical patterns can guide the development of specific interventions tailored to the needs of each region.

1. Introduction

Concern for sustainability is maintained as a prevalent topic in current society (Mensah, 2019). In today's environment, this concern is mainly in the private sector (Rashed and Shah, 2021), where new concepts are being integrated at the European level. This is significantly evidenced by the introduction of the new Directive (EU) 2022/2464[1] on corporate sustainability reporting, known as CSRD. This directive represents a firm step by the European Commission within the framework of the European Green Deal and the so-called Agenda for Sustainable Financing, with the goal of complementing and perfecting the existing legislation by amending Regulation (EU) No. 537/2014,[2] Directive 2004/109/EC,[3] Directive 2006/43/EC,[4] and Directive 2013/34/EU.[5]

The necessary development concerning sustainability and its reporting in the public sector has been highlighted by the recent implementation of mandatory reporting and presentation of non-financial information in the private sector (Fusco and Ricci, 2019). Reporting that allows for homogenization and, thus, subsequent verification is pursued by different institutions (Vitolla et al., 2019).

In the context of the Sustainable Development Goals (SDGs), homogenization is the tendency to standardize practices, policies, and strategies globally to address sustainable development challenges effectively (Parra-Domínguez et al., 2022). The SDGs, adopted by all Member States of the United Nations in 2015, are recognized as a universal call to eradicate poverty, protect the planet, and ensure that all people enjoy peace and prosperity by 2030 (Tremblay et al., 2020). A crucial role is played by homogenization in this effort by promoting a

* Corresponding author.
E-mail address: laurasanzmartin@usal.es (L. Sanz-Martín).

[1] (Directive (EU) 2022/2464 of the European Parliament and of the Council of 14 December 2022 amending Regulation (EU) No 537/2014, Directive 2004/109/EC, Directive 2006/43/EC and Directive 2013/34/EU, as regards corporate sustainability reporting).

[2] Regulation (EU) No 537/2014 of the European Parliament and of the Council of 16 April 2014 on specific requirements regarding statutory audit of public-interest entities and repealing Commission Decision 2005/909/EC Text with EEA relevance.

[3] DIRECTIVE 2004/109/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 15 December 2004 on the harmonization of transparency requirements in relation to information about issuers whose securities are admitted to trading on a regulated market and amending Directive 2001/34/EC.

[4] Directive 2006/43/EC of the European Parliament and of the Council of 17 May 2006 on statutory audits of annual accounts and consolidated accounts, amending Council Directives 78/660/EEC and 83/349/EEC and repealing Council Directive 84/253/EEC (Text with EEA relevance).

[5] Directive 2013/34/EU of the European Parliament and of the Council of 26 June 2013 on the annual financial statements, consolidated financial statements and related reports of certain types of undertakings, amending Directive 2006/43/EC of the European Parliament and of the Council and repealing Council Directives 78/660/EEC and 83/349/EEC (Text with EEA relevance).

https://doi.org/10.1016/j.acags.2024.100184
Received 20 March 2024; Received in revised form 30 July 2024; Accepted 30 July 2024
Available online 6 August 2024

unified and coherent approach to achieving these goals. However, this process must balance the need for common standards with respect for various regions' cultural, economic, and political diversity.

Sustainability is a crucial principle in global development, guiding towards a more equitable and resilient future (Mensah, 2019). It balances current needs without compromising future generations' ability to meet their own. It involves strategic management of resources, minimizing ecological footprints, and promoting well-being. Sustainability addresses environmental, economic, and social dimensions, recognizing the interdependence between growth, social inclusion, and environmental protection (Ahmad et al., 2023). The link between sustainability and the Sustainable Development Goals is intrinsic and multifaceted.

In our case study, we will focus on SDG 11. It seeks to make cities and human settlements inclusive, safe, resilient and sustainable. It encompasses a broad range of targets that focus on enhancing urban living conditions while minimizing the environmental impact of cities (Takase, 2018). By 2030, SDG 11 aims to ensure that all people have access to adequate, safe and affordable housing and essential services, and to improve slums. A crucial component of this goal is to decrease the per capita environmental impact of cities, focusing especially on air quality and municipal and other waste management (Nabiyeva et al., 2023). The goal also includes significantly increasing the number of cities and settlements that adopt and implement integrated policies. These policies focus on inclusiveness, resource efficiency, climate change mitigation and adaptation, disaster resilience, and the development and implementation of integrated disaster risk management at all levels.

SDG 11 aims to ensure that all people, especially women, children, the elderly and persons with disabilities, have access to safe, inclusive and green public spaces. It promotes sustainable cities with essential services, energy, housing, and transportation. Recognizing cities as economic growth engines, it addresses challenges like urban sprawl, pollution, poverty, and lack of well-planned urbanization.

According to the World Health Organization (WHO) in 2019, the 10 leading causes of death accounted for 55 per cent of the 55.4 million deaths worldwide. Lower respiratory tract infections were the second leading cause of death in low-income countries, fifth in upper-middle-income countries and sixth in high-income countries. This ranking underlines the importance of addressing factors such as fine particles, especially $PM_{2.5}$, which have a significant ability to penetrate the human respiratory system and bloodstream. These particles are associated with a range of adverse effects, including acute and chronic respiratory diseases such as cardiovascular disease, lung cancer and cerebrovascular disease. Reducing $PM_{2.5}$ pollution is not only crucial to protect public health, but also to promote sustainable and vital urban environments. WHO considers air pollution to be the second most important risk factor for non-communicable diseases. The WHO Global Air Quality Guidelines set global standards and limits for major air pollutants, with interim targets such as interim target 1 to reduce levels of $PM_{2.5}$ to 35 $\mu g/m^3$, marking a significant step towards protecting public health.

This study aims to understand the impact of air pollution, specifically $PM_{2.5}$ concentration, on public health and sustainable development. Our study focuses on SDG 11 of the United Nations Sustainable. It focuses on European nations' premature mortality rates, aiding in developing public health policies and raising public awareness of health risks associated with air pollution. This work is organized as follows: first, we introduce the data used and its descriptive statistics. Then, we explain the methods employed. Next, we present the findings, and finally, the conclusions.

2. Methodology

2.1. Data

A multivariate analysis of $PM_{2.5}$ and associated premature deaths of 23 countries in Europe was performed for this work. This is part of SDG 11, namely 11.6.2, which aims to reduce the number of premature deaths due to $PM_{2.5}$ by 55% by 2030, compared to 2005. In order to perform this multivariate analysis, we used data from the World Health Organization for the concentration of particles and from Eurostat for the ratio and number of premature deaths due to exposure to fine particulate matter (PM2.5).

The data from the WHO contained the $PM_{2.5}$ with a confident interval (CI) of 95% in urban areas, rural areas, cities, towns and in each country; and the data from Eurostat[6] contained the number of deaths and the rate for each country. The countries under study were Austria, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Czechia, Denmark, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Netherlands, Norway, Poland, Portugal, Slovakia, Slovenia, Spain, Sweden and Switzerland.[7]

Table 1 shows the descriptive analysis performed each year to determine whether the concentration of $PM_{2.5}$ has decreased. We could confirm that the average concentration of $PM_{2.5}$ has decreased over the years, namely by 26.87%. Although we observed that the average of $PM_{2.5}$ has had ups and downs, it has decreased progressively until 2016, then increasing not only in 2017 but also in 2018, and decreasing almost 2 points in 2019.

Table 1
Descriptive statistics for $PM_{2.5}$.

Year	Mean	Std	Min	50%	Max
2010	17.04	6.99	6.45	15.87	42.92
2011	17.58	7.27	6.40	16.12	45.06
2012	15.86	6.84	5.78	14.53	43.85
2013	15.23	6.27	5.64	14.22	38.98
2014	14.70	6.62	5.27	13.01	39.60
2015	14.70	6.16	5.33	13.31	37.95
2016	13.76	6.01	4.93	12.28	36.78
2017	14.04	6.40	4.94	12.32	40.28
2018	14.23	6.59	5.13	12.45	39.63
2019	12.46	5.06	5.06	11.34	32.53

In Table 2, we can see that there is a difference between the areas; the concentration in urban areas is more or less the average between cities and towns. The difference between urban and rural areas is 2.1 points, with the average concentration of $PM_{2.5}$ lower in rural areas than in urban areas in Europe. When comparing large cities with small ones, the average difference is also remarkable.

Table 2
Descriptive statistics for each area.

Area	Mean	Std	Min	50%	Max
Cities	16.27	7.24	6.4	15.31	45.06
Towns	14.92	6.57	5.95	13.54	39.61
Urban areas	15.53	6.74	6.17	14.57	41.37
Rural areas	13.43	5.84	4.93	12.23	33.48

Table 3 presents descriptive statistics for premature death rates in Europe over the years under study. From 2010 to 2019, the rate of premature deaths has decreased by 37.01%, 10% more than the concentration of $PM_{2.5}$. A closer look shows that over the years the average death rate in Europe has not decreased progressively. In fact, the death rate peaked in 2011, averaging 92.09. From that year onwards, the average gradually decreased until 2014. Subsequently, it has experienced ups and downs, reaching the minimum in 2019, with an average ratio in Europe of 53.65. The difference between minimum and maximum values is notable, indicating variability between countries, which will be explored in more detail in the next graphs.

Fig. 1 shows the concentration of $PM_{2.5}$ for each country. The high concentration value of Bosnia and Herzegovina is remarkable, its minimum value being higher than the maximum of most countries. Bulgaria, Poland, Slovakia and Slovenia do not have such high values, but they

[6] https://ec.europa.eu/eurostat.
[7] Data available on January 15, 2024.

Table 3
Descriptive statistics for the rates of premature deaths.

Year	Mean	Std	Min	50%	Max
2010	85.17	50.92	19.0	84.0	215.0
2011	92.09	60.83	17.0	80.0	272.0
2012	75.13	49.74	14.0	64.0	219.0
2013	71.00	44.14	7.0	72.0	187.0
2014	64.57	43.36	14.0	56.0	197.0
2015	71.09	50.96	4.0	57.0	204.0
2016	67.65	55.52	4.0	51.0	213.0
2017	70.78	58.50	2.0	50.0	192.0
2018	69.17	53.44	8.0	57.0	209.0
2019	53.65	42.36	3.0	41.0	166.0

are worrying too. Also noteworthy are the low concentration values in northern European countries such as Finland, Norway and Sweden. Countries from the center and south of Europe like France, Germany, Netherlands or Spain have values between 10 and 15 micrograms per cubic meter, target 3.

Figs. 2 and 3 show both the rate and the number of premature deaths in the countries studied. It is evident that the number of deaths is notably higher in France, Germany, Italy and Poland compared to the rest of the countries. On the other hand, Denmark, Finland, Ireland, Norway, Slovenia and Sweden have a very low number of premature deaths, with the lowest number of premature deaths in these countries. However, the rates show a significant variation. We note that the highest values are in Bosnia and Herzegovina and Bulgaria, followed closely by Croatia, Czech Republic, Greece, Hungary, Poland and Slovakia. It would be expected that these countries would have the highest rates of deaths, given that they have the highest concentrations of fine particulate matter. As noted above, France was the country with the highest number of premature deaths. If we analyze its rate, although it is not the lowest, it is below the average. This is because the rate takes into account the number of inhabitants in each country. It is normal that the number of deaths is higher in countries with a larger population. For this reason, we will perform all analyses using the death rate, so as not to bias the results or influence the number of inhabitants of each country.

Upon revision of descriptive analyses and box-plots, it is evident that there are differences in particle concentration, ratio, and number of deaths across different areas and countries.

2.2. Method

In order to see if our sample follows a normal distribution and decide if we can use parametrical or non-parametrical tests, we perform Shapiro–Wilk test. Given an ordered random sample, $x_1 \leq x_2 \leq \cdots \leq x_n$, the original Shapiro–Wilk test statistic (Shapiro and Wilk, 1965) is defined as,

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_i\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{1}$$

where x_i is the ith order statistic, $\bar{x}$ is the sample mean, $a = (a_1, \ldots, a_n)$ is the vector of coefficients given by $\left(\prod_{j=1}^{k}(1 - \frac{u_j^2}{2})\right) / \sqrt{\sum_{j=1}^{k} u_j^2}$, $u = (u_1, \ldots, u_n)$ are the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution and V is the covariance matrix of those order statistics and the value of W lies between zero and one. Small values of W lead to the rejection of normality whereas a value of one indicates normality of the data (Mohd Razali and Yap, 2011).

After seeing if the sample follows a normal distribution, we are going to analise if we can perform the ANOVA test, which assumes two premises: the sample has to follow a normal distribution and it has to be homocedastic, this means uniform variances across groups or samples. To analise this, we perform Levene's test (Levene, 1960). It serves to examine whether a set of k samples exhibits equal variances, a property referred to as homogeneity of variance. The hypotheses for Levene's test are expressed as follows: H_0: homogeneity of variances across all groups, H_a: heterogeneity of variances for at least one pair of groups.

The test statistic, denoted as W, is calculated using the formula:

$$W_0 = \frac{\sum_{i=1}^{k} N_i(\bar{z}_{i.} - \bar{z}_{..})^2/(g-1)}{\sum_{i}^{k} N_i(\bar{z}_{i.} - \bar{z}_{..})^2 / \sum_i (n_i - 1)} \tag{2}$$

where $\bar{z}i = \sum zij/n_i$ represents the mean within the ith group and $\bar{z}.. = \sum\sum zij/\sum n_i$ represents the overall mean.

If the calculated test statistic W surpasses the critical value from the F-table with degrees of freedom $g-1$ and $\sum_i(n_i - 1)$, we reject the null hypothesis H_0 at the significance level α. This decision implies evidence of heterogeneity of variances among the considered groups (Brown and Forsythe, 1974).

After performing these two tests, in neither of the cases under study we cannot perform the ANOVA test, so we use the nonparametric test Kruskal–Wallis. Kruskal and Wallis (1952) addressed the practical statistical challenge of determining whether samples should be considered as originating from the same population. When faced with the variations often observed among samples, the key question is whether these differences indicate variations between populations or are simply chance fluctuations expected among random samples from the same population. In such situations, it is common to assume that the populations are roughly similar in shape, suggesting that any differences are due to a shift or translation.

The test statistic, denoted as H, is computed as follows:

$$H = \frac{12}{C}\left(\sum_{i=1}^{C} R_i^2 - \frac{3(N+1)}{N}\right) \tag{3}$$

where C represents the number of samples, n_i is the number of observations in the ith sample, N is the total number of observations and R_i is the sum of ranks in the ith sample. Large values of H lead to the rejection of the null hypothesis.

Once we have seen if there are differences between groups, we want to see in which groups are differences. In order to analise this, we perform Dunn test. The Dunn z-test statistics (Dunn, 1961) is an approximation of the exact rank sum test statistics. It uses the mean classifications of the results within each group obtained from the previous Kruskal–Wallis test $W_i = \frac{W_i}{n_i}$, where W_i represents the sum of ranges and n_i is the sample size for the i-nth group. The inference is based on differences in mean classifications between groups. For the comparison between group A and group B, the calculation is expressed as:

$$z_i = \frac{y_i}{\sigma_i} \tag{4}$$

where i refers to one of the 1 to m multiple comparisons, $y_i = W_A - W_B$ and $sigma_i$ denotes the standard deviation of y_i.

After performing statistical tests, we can apply machine learning techniques to analyze the data in more depth. A commonly used technique is the k-means algorithm in data mining for clustering large data sets. Proposed by MacQueen in 1967, this algorithm sorts data objects into different k clusters iteratively. The choice of this algorithm is justified by its efficiency and effectiveness in clustering multidimensional populations into k sample-based sets. The k-means process tends to generate partitions that are efficient in terms of the variance within each class. This is evidenced by the W2(S) metric, which tends to be low for partitions generated by the method, indicating good separation between clusters. In addition, it is an easy-to-program and computationally inexpensive algorithm, allowing very large samples to be processed on standard computers. Its potential applications are diverse, including similarity-based clustering methods, non-linear prediction, approximation of multivariate distributions and non-parametric tests for independence between variables (MacQueen, 1967).

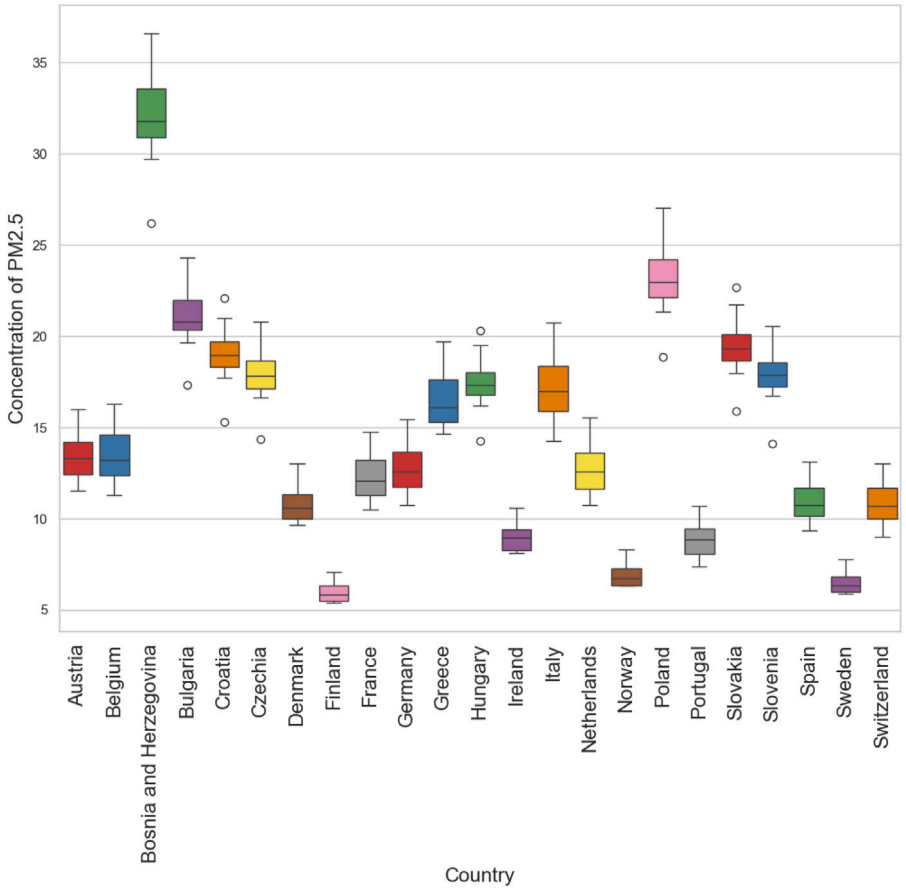


Fig. 1. Concentration of $PM_{2.5}$ for each country.

Initially, k cluster centers are chosen at random and then each data object is assigned to the nearest cluster center according to Euclidean distance. After each assignment, the cluster centers are recalculated and the procedure is repeated until the cluster centers converge or the criterion function reaches a local minimum. The Euclidean distance is used to calculate the distance between cluster centers and data objects (Na et al., 2010).

The criterion function, E, is defined as the sum of the squared errors of all objects in the database. This function helps to evaluate the effectiveness of the clustering performed by the k-means algorithm. In addition to k-means, there are numerous other machine learning algorithms used for clustering, classification, regression tasks and more, which can be tailored according to specific data analysis needs. Machine learning plays a crucial role in extracting meaningful information from large data sets, enabling predictions, identifying hidden patterns and automating complex processes in a variety of fields, from medicine to finance and beyond.

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \tag{5}$$

where μ_i is the center of the group C_i, and $\|x - \mu_i\|$ represents the Euclidean distance between a point x and the center μ_i of the group C_i.

A time series is a set of observations ordered in time that show the evolution of a phenomenon or variable over time. The purpose of analyzing a time series with data at regular intervals is to understand its pattern of behavior in order to predict the future, assuming that conditions will remain constant.

In general, random phenomena are associated with the series of interest, so the analysis of its past behavior only allows to approximate the structure or probabilistic model for future prediction (Viñals, 2009). This also applies to time series regression modeling, as explained by Kedem and Fokianos (2005). In this context, the time series analyst deals with stochastic relationships, i.e., those that include error terms in the model specification. The simplest form of such a model between two variables, say Y_t and X_t, is called the simple time series regression model. This approach allows not only to understand the past behavior of the series, but also to use machine learning techniques to predict its future behavior and make informed decisions in various fields such as finance, marketing, and resource management, among others.

$$Y_t = a + bX_t + e_t \tag{6}$$

The equation represents a simple time series regression model, where:

- Y_t is the endogenous variable at time t
- X_t is the exogenous variable at time t, which is an independent variable that can influence Y_t
- a and b are unknown parameters that determine the relationship between X_t and Y_t
- e_t is the random perturbation term at time t, which represents the effect of all factors not included in the model

It is important to clarify that in this context, the variables X_t and Y_t represent values at a specific time step t. This equation does not model the dependency between different time steps (e.g., X_{t-1}, X_{t-2}, …, X_{t-n}). Instead, it captures the relationship between the two variables at the same time point. To model the dependency over multiple time steps, more complex time series models such as ARIMA or LSTM are required.

Linear regression is a technique commonly used to predict a quantitative response variable from a quantitative predictor variable. In

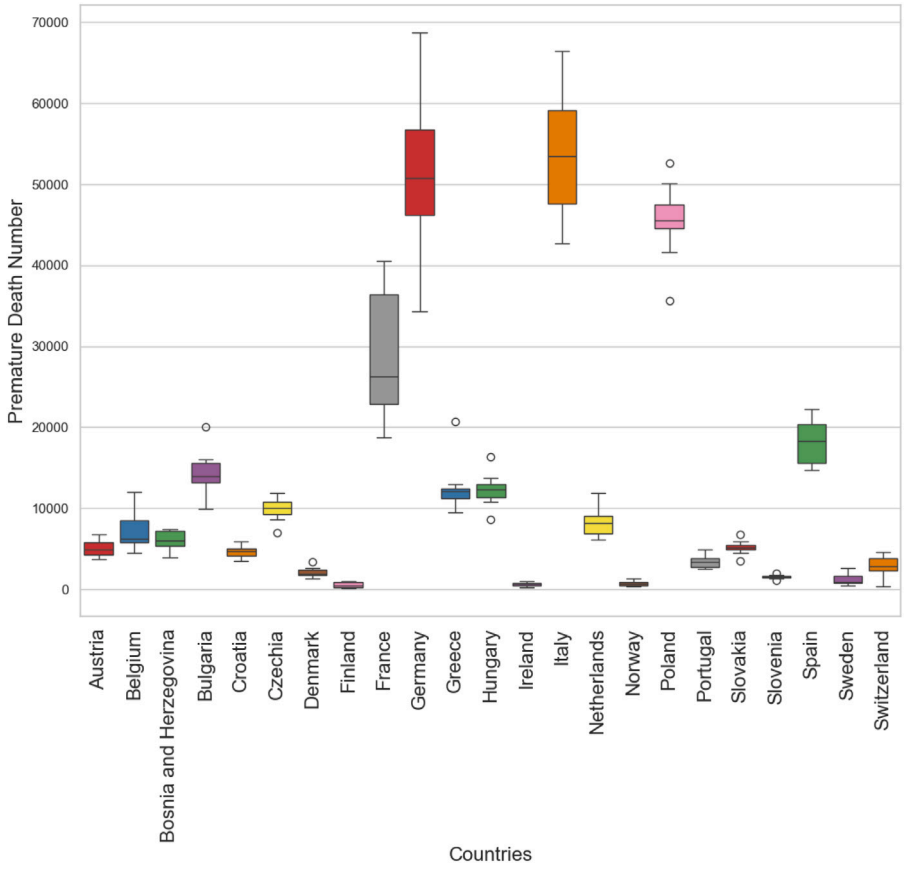


Fig. 2. Number of premature deaths for each country.

this technique, the relationship between the variables is modeled as a straight line, where one seeks to fit the line that best fits the observed data. The linear regression model has the general form shown in Eq. (7), where Y is the response variable, X is the predictor variable, β_0 is the intercept and β_1 is the slope of the line. The term $varepsilon$ represents the error in the estimate, which is the difference between the predicted value and the observed value.

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{7}$$

The data were split into a training set and a test set using a temporal split, where data from 2010 to 2017 were used for training and data from 2018 to 2019, both inclusive, were used for model evaluation. In this way, we ensured that 20% of the data were for the test.

Another regression model was created to capture the evolution of the predicted annual fatality rate in Europe. To capture potential non-linear relationships, polynomial features of degree 2 are introduced to the Concentration predictor variable using scikit-learn's PolynomialFeatures. This transformation expands the model's capability to account for quadratic effects in the predictor variable. In model development, Leave-One-Out Cross-Validation (LOO-CV) is implemented to robustly evaluate the model's performance. This technique involves iteratively training the model on all years except the current one and then predicting outcomes for the omitted year. Ridge Regression, integrated within scikit-learn's Ridge module, is utilized for its regularization capabilities, which help stabilize the model and prevent overfitting by penalizing large coefficient values.

Evaluation of model predictions is performed using standard regression metrics, including R^2 (coefficient of determination), Mean Squared Error (MSE), Mean Absolute Error (MAE), and Median Absolute Error (MedAE). These metrics collectively assess the accuracy and reliability of the model in predicting the Rate variable based on the Concentration predictor across the dataset.

In this paper we used the above mathematical procedure using the functionalities of the Python library scikit-learn, which approaches the problem as a particular case of linear regression (Golondrino et al., 2020).

All statistical analyses, contrasts, graphs and models used in this study were performed using the Python programming language and the Spyder development environment. In particular, the following libraries were used: Pandas for data manipulation, NumPy for numerical operations, Seaborn and Matplotlib for data visualization, Scikit-learn for linear regression model building and data analysis, and SciPy for statistical testing.

3. Results and discussion

Differences between various areas were investigated. Non-parametric tests were employed due to the absence of a normal distribution, despite observing homoscedasticity. The Kruskal–Wallis test yielded highly significant results, indicating significant differences between medians in at least a couple of areas. To pinpoint these differences, the Dunn test was conducted (see Fig. 4). Significant disparities were noted between cities and urban areas, as well as between urban and rural areas. At a 95% confidence level, differences between large and small cities, as well as between small cities and urban areas, could not be conclusively affirmed. However, a 10% risk would suggest differences in these cases.

The data presented in Table 4 provides a detailed overview of the distribution of rural and urban populations across various European countries.[8] This data is essential for placing the findings of our study in perspective because it shows how much of the population lives in

[8] https://commission.europa.eu/.

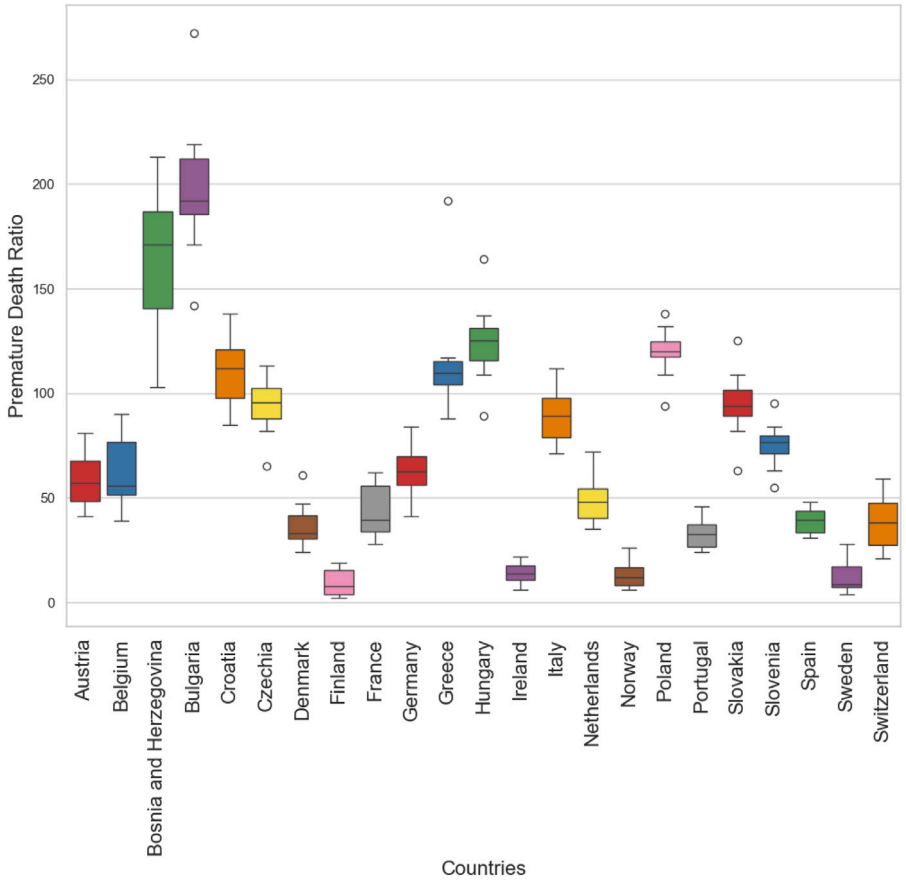


Fig. 3. Premature death rate for each country.

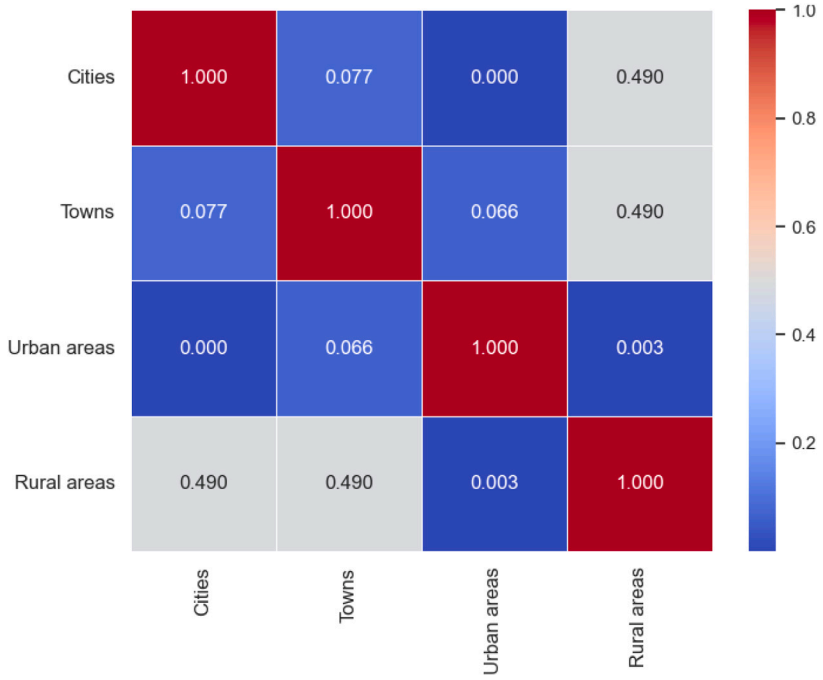


Fig. 4. Dunn test p-values matrix for medium differences between areas.

rural and urban areas and how that affects health and air quality. Determining the percentage of people living in rural as opposed to urban regions aids in clarifying the causes of air pollution and the varying effects it has. As an example, nations with greater rates of urbanization, like Belgium (91.50%) and the Netherlands (92.00%), may have different dynamics of air pollution than those with higher rates of ruralization, like Slovenia (58.12%) and Ireland (56.88%). Because of increased traffic, industry, and population density, urban

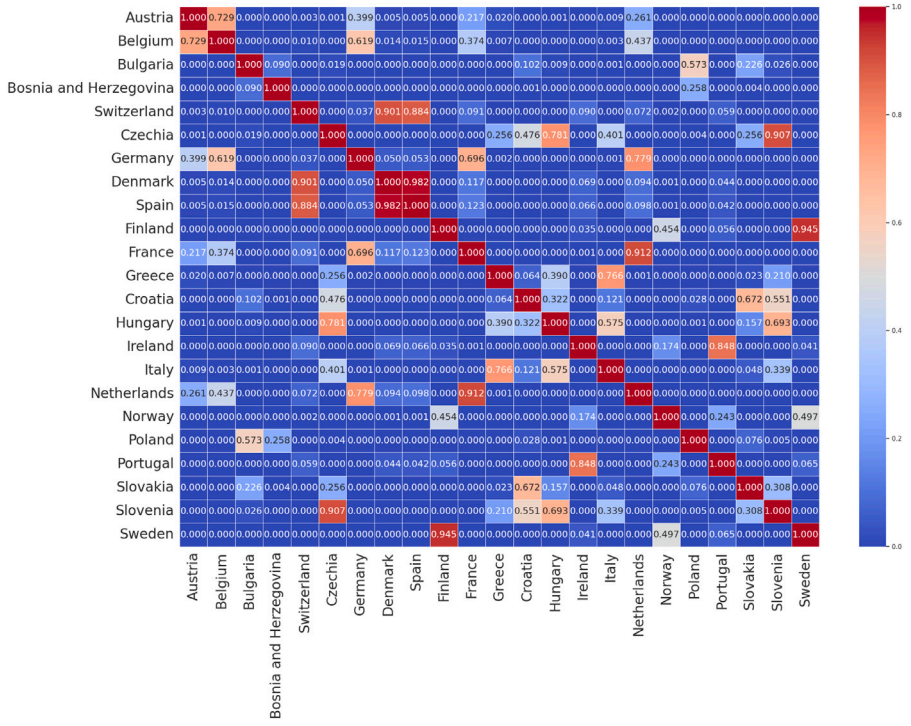


Fig. 5. Dunn test p-values matrix for medium differences between countries.

areas usually have greater concentrations of pollutants, which can aggravate health problems connected to air quality. The interpretation of the health data examined in our study depends critically on these demographic distributions as well. The observed variations in $PM_{2.5}$ levels and related health consequences, such as premature death rates, may be explained in part by the variance in the percentages of the urban and rural populations among the nations under study. For instance, high population density urban regions may have greater $PM_{2.5}$ concentrations, which can exacerbate respiratory and cardiovascular problems, whereas lower $PM_{2.5}$ concentrations in rural locations may present with other health issues. We can better understand the complex effects of air pollution and create focused strategies for both rural and urban settings to successfully reduce these effects by comparing this demographic data with our health outcome results.

Table 4
Rural and urban population distribution in european countries.

Country	% Rural population	% Urban population
Austria	40.16%	59.84%
Belgium	8.50%	91.50%
Bosnia and Herzegovina	48.52%	51.48%
Bulgaria	12.96%	87.04%
Croatia	42.49%	57.51%
Czechia	21.19%	78.81%
Denmark	28.26%	71.74%
Finland	39.26%	60.74%
France	27.95%	72.05%
Germany	15.62%	84.38%
Greece	31.25%	68.75%
Hungary	18.56%	81.44%
Ireland	56.88%	43.12%
Italy	9.87%	90.13%
Netherlands	8.00%	92.00%
Norway	16.99%	83.01%
Poland	35.68%	64.32%
Portugal	30.82%	69.18%
Slovakia	37.31%	62.69%
Slovenia	58.12%	41.88%
Spain	3.33%	96.67%
Switzerland	26.00%	74.00%
Sweden	8.96%	91.04%

$PM_{2.5}$ level disparities across countries were examined using the Kruskal–Wallis test, as the assumptions for the ANOVA test were not met. The obtained statistic was highly significant, suggesting differences between at least two countries. Post-hoc contrasts were conducted, and the Dunn test was analyzed, resulting in a graphical representation of the p-values associated with the resulting matrix. It was evident that significant differences were present among most countries. However, only a few distinctions were observed between certain pairs, such as Belgium and Austria, Belgium and Germany, Denmark and Switzerland, the Czech Republic and Slovenia, Sweden and Finland, the Netherlands and France, the Netherlands and Germany, Hungary and Slovenia, among others (see Figs. 5 and 6).

The classification of similar countries in these clusters leads us to consider that a factor influencing the concentration of fine particles could be the geographical location of the countries, elements such as climate, or even similar policies may be influencing.

In Fig. 7 we can appreciate how the relationship between the concentration of fine particles and the premature mortality rate for each country follows a regression line until the concentration reaches the value 23 and the premature mortality rate the value 125. After this point, the relationship is more dispersed. It is remarkable how Bosnia and Herzegovina has the highest concentration of fine particles and its premature mortality rate is lower than it should be compared to other countries. Worrying, that in the case of Bulgaria the opposite happens, its death rate is higher with a lower concentration of particles.

A clustering analysis was performed to categorize countries based on their levels of $PM_{2.5}$ concentration and rates of premature deaths. Our goal was to uncover patterns and similarities among countries concerning air quality and health outcomes.

Each cluster represents a grouping of countries with similar characteristics concerning air pollution and its rate of premature deaths. The clusters consists on countries that exhibit comparable levels of $PM_{2.5}$ and premature rates of deaths.

- Cluster 0: Bosnia and Herzegovina and Bulgaria
- Cluster 1: Finland, Ireland, Norway, Sweden and Portugal
- Cluster 2: Croatia, Greece, Italy, Slovakia, Poland

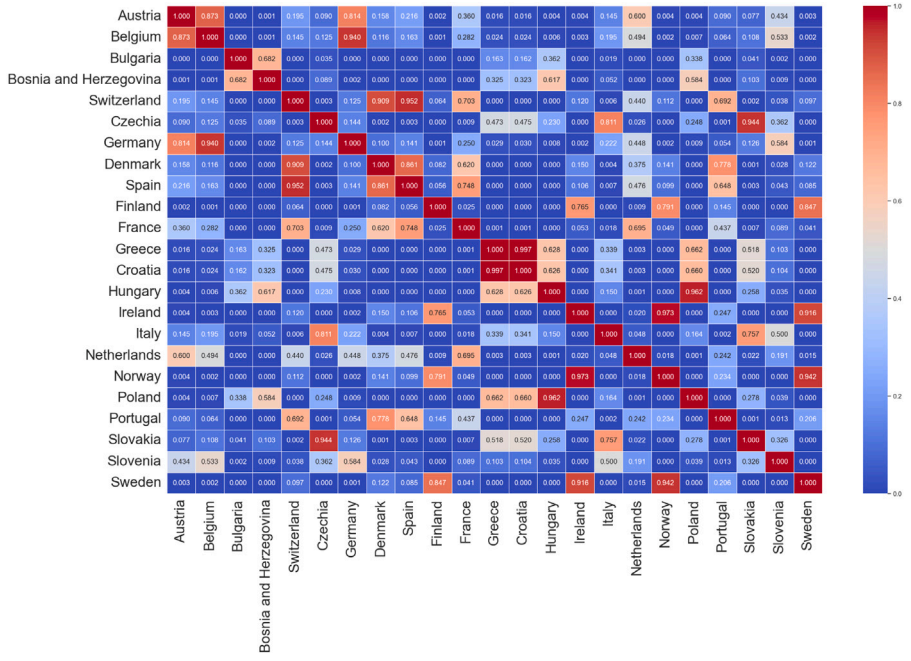


Fig. 6. Dunn test p-values matrix for medium differences between countries.

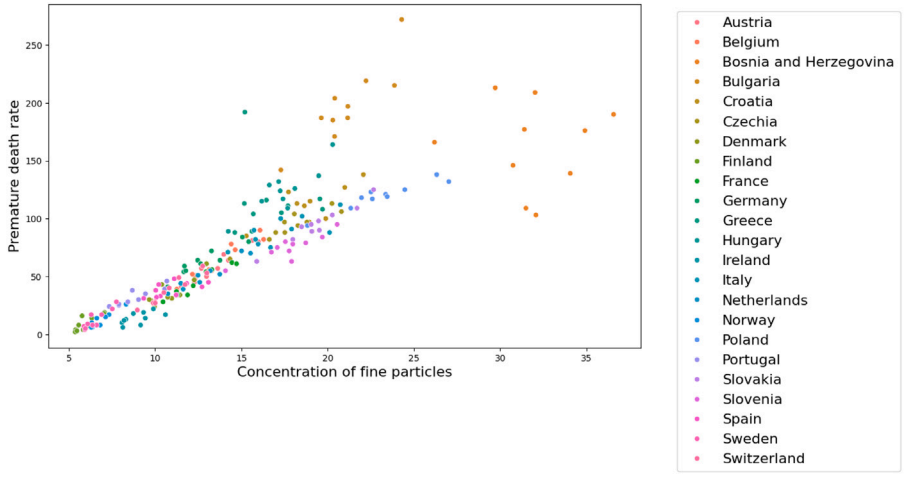


Fig. 7. Relation between the concentration of fine particles and premature death rate.

- Cluster 3: Netherlands, France, Belgium, Germany, Spain, Denmark and Switzerland

It is noteworthy that Austria, Czechia and Hungary are not clustered according to the analysis conducted. Further investigation is warranted to understand the underlying factors contributing to this outcome.

In Fig. 8, the distribution of each cluster in relation to the concentration of fine particulate matter and the rate of premature deaths can be seen. And in Tables 5 and 6, the descriptive statistics for concentration and rate in each cluster. It is notable the differences between cluster 0, which has the highest values, and cluster 3, which has the lowest.

Table 5
Descriptive statistics for concentration in each cluster.

Cluster	Mean	Median	Std
0	26.42	26.26	5.81
1	12.26	12.16	1.42
2	18.67	18.52	2.37
3	7.77	7.81	1.55

Table 6
Descriptive statistics for rate in each cluster.

Cluster	Mean	Median	Std
0	178.50	186.00	38.44
1	49.28	48.00	12.00
2	99.19	97.00	17.82
3	17.48	17.00	9.72

Fig. 9 shows a map of Europe showing how the clusters are geographically distributed. It is notable that countries with similar locations tend to cluster together. Those in the north and west have lower levels of $PM_{2.5}$ and lower rates of premature deaths. At the center, the values are moderate, while in Eastern European countries, marked in orange, worrying concentrations are observed. Finally, in red, there are the countries with the highest concentrations of particles and highest rates of premature deaths.

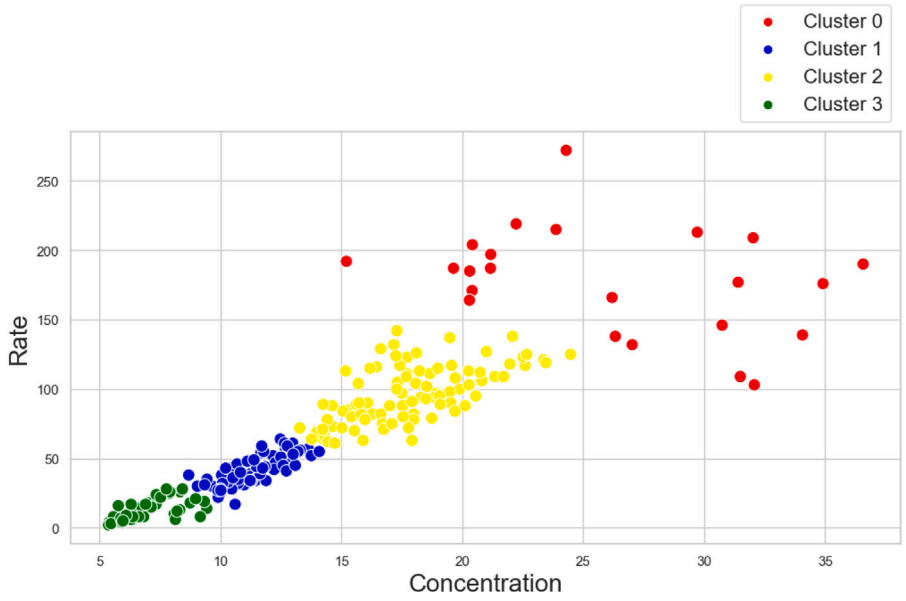


Fig. 8. Relation between $PM_{2.5}$ and premature death rate for each cluster.

Fig. 9. Map of the clusters.

We have analyzed the time trends for the concentration of fine particles for each country, as well as the ratio of premature deaths for these same particles. These analyses are shown in Figs. 10 and 11, respectively. For this we have made two graphs in which we can see how the concentration of $PM_{2.5}$ has decreased over the years and it is remarkable how with more or less particle value, all countries follow the same pattern. Opposite to the time trends for the premature death rate which not follow any pattern and not all countries have decreased the rate over the years under study. Notably, in Greece in the year 2017, the death rate increased its value 50 points reaching its maximum value.

Table 7 presents the time trends of fine particulate matter ($PM_{2.5}$) concentration and mortality rate per country. Some countries, such as Bosnia and Herzegovina, Bulgaria and Greece, show a significant decrease in the concentration of $PM_{2.5}$ over time, while other countries experience less marked changes or even increases in concentration. An apparent relationship between $PM_{2.5}$ concentration and mortality is observed in some countries. For example, Bosnia and Herzegovina and Greece show a reduction in the concentration of $PM_{2.5}$, but an increase in the mortality rate, suggesting the influence of factors other than air quality on mortality rates. This observation highlights the complexity of mortality trends, where multiple factors such as healthcare quality, socioeconomic conditions, and lifestyle changes also play significant roles.

Furthermore, air quality datasets, such as those from Eurostat and WHO, may vary due to differences in data collection methods, spatial coverage, and temporal resolution across countries. For example, the WHO estimates annual urban mean concentrations of $PM_{2.5}$ using a combination of ground measurements, satellite remote sensing data, and modeling techniques (WHO, 2021). In contrast, Eurostat relies

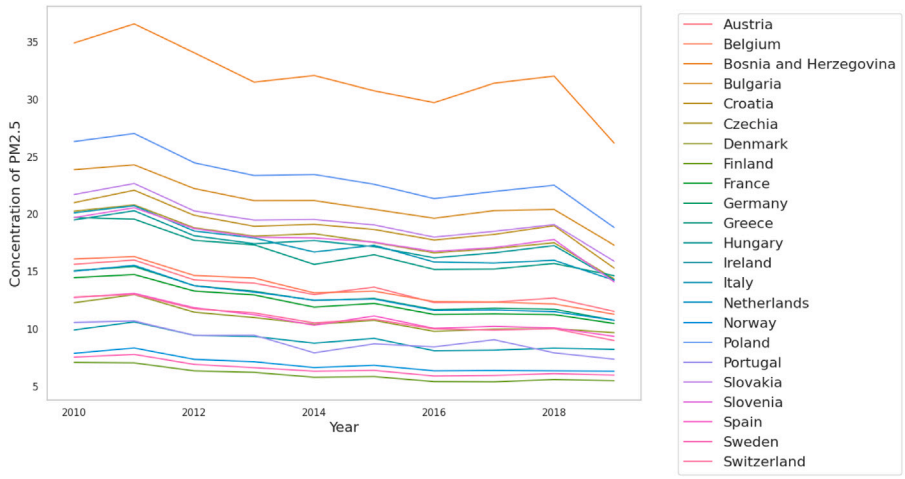


Fig. 10. Concentration of $PM_{2.5}$ over the years.

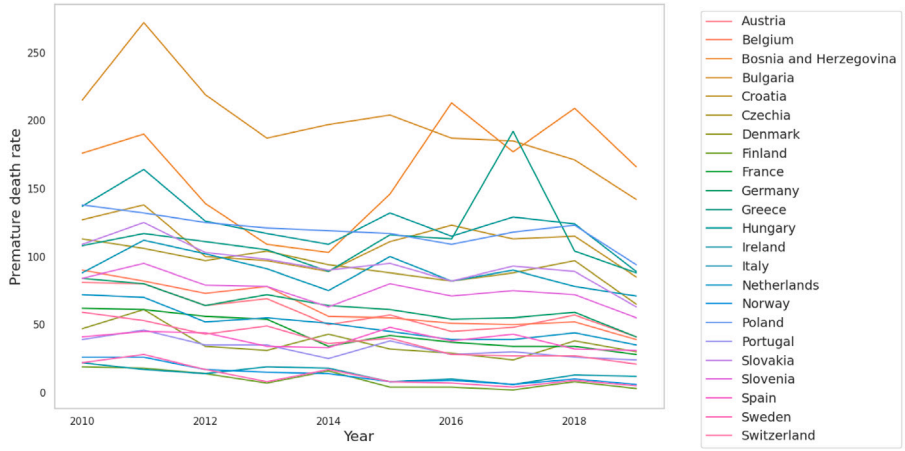


Fig. 11. Rate of premature deaths over the years.

on data from the European Environment Agency (EEA), which compiles information from fixed-site monitors and other sources (Eurostat, 2023).

On the other hand, some countries show a steady decrease in mortality rate over time, such as Austria, Belgium, Bulgaria, Germany, the Netherlands, Poland, Slovakia, Portugal and Switzerland. This could indicate improvements in medical care, advances in public health or lifestyle changes that are contributing to a reduction in mortality. The table also highlights the worrying increase in the mortality rate in Bosnia and Herzegovina each year, despite the decrease in the concentration of $PM_{2.5}$. This may require further research to understand the underlying causes of this phenomenon and to take appropriate corrective action.

The inclusion of CI in the table provides a measure of precision to the estimates and allows for a deeper understanding of the variability of the data. The confidence interval for the reduction in concentration of $PM_{2.5}$ in Austria is $[-0.5940, -0.3061]$, indicating with a high degree of certainty that the true reduction is within this range. This information is essential to interpret the results more accurately and to notify decisions based on the data presented. CI provide a robust statistical framework that supports the reliability of the conclusions and contextualizes the observed changes.

A linear regression model was developed to predict the premature death rate based on $PM_{2.5}$ particulate matter concentration. $PM_{2.5}$ concentration is considered as the independent variable, while premature death rate is treated as the dependent variable. The dataset is organized by country and year, although these variables are not directly part of the regression model, but are used to filter and structure the data. The aim is to understand how the premature death rate varies as a function of $PM_{2.5}$ particulate matter concentration, providing valuable information to address public health and environmental issues.

The model showed an R-squared value of 0.8857, indicating that approximately 88.57% of the variability in premature death rates can be explained by $PM_{2.5}$ concentration. In addition, the model has a Mean Squared Error of 261.09. This model offers a valuable tool for predicting premature death rates in relation to $PM_{2.5}$ concentration, as evidenced by the significant relationship depicted in Fig. 12. Understanding the coefficients allows insight into how $PM_{2.5}$ concentration affects premature death rates. However, it is important to acknowledge potential limitations, such as assuming linearity in the $PM_{2.5}$-death rate relationship, and rigorous validation is crucial for generalizability. Overall, this lineal regression model serves as a valuable asset for informing public health and environmental policy decisions.

Another regression model was created with the objective of predicting the rate of premature deaths for each year under study. To assess the predictive ability of the polynomial regression model in relation to the premature death rate and the concentration of $PM_{2.5}$, the Leave-One-Out cross-validation (LOO-CV) technique was employed.

Table 7
Trends in $PM_{2.5}$ concentration and death rate by country with confidence intervals.

Country	Fine particulates matter	Death rate
Austria	−0.4501 ([−0.5940, −0.3061])	−4.0364 ([−5.9095, −2.1632])
Belgium	−0.5443 ([−0.6607, −0.4279])	−5.2485 ([−6.7064, −3.7906])
Bosnia and Herzegovina	−0.7897 ([−1.2229, −0.3565])	3.5636 ([−6.2904, 13.4177])
Bulgaria	−0.6145 ([−0.8386, −0.3905])	−9.2545 ([−14.5291, −3.9799])
Croatia	−0.5172 ([−0.7747, −0.2596])	−2.2667 ([−6.4651, 1.9317])
Czechia	−0.5482 ([−0.7599, −0.3364])	−3.7091 ([−5.7977, −1.6205])
Denmark	−0.3321 ([−0.4609, −0.2032])	−2.3091 ([−4.5532, −0.0650])
Finland	−0.1927 ([−0.2627, −0.1226])	−1.7879 ([−2.8065, −0.7692])
France	−0.4554 ([−0.5682, −0.3426])	−3.9273 ([−5.1748, −2.6797])
Germany	−0.4822 ([−0.6051, −0.3594])	−3.8545 ([−5.2295, −2.4795])
Greece	−0.5506 ([−0.7449, −0.3563])	1.1212 ([−6.6629, 8.9053])
Hungary	−0.4852 ([−0.7046, −0.2658])	−4.1212 ([−8.1758, −0.0666])
Ireland	−0.2495 ([−0.3534, −0.1455])	−1.1818 ([−2.1638, −0.1999])
Italy	−0.6422 ([−0.8167, −0.4676])	−2.7455 ([−5.4279, −0.0630])
Netherlands	−0.4970 ([−0.6216, −0.3724])	−3.8424 ([−5.1845, −2.5003])
Norway	−0.2122 ([−0.2940, −0.1305])	−2.2485 ([−3.0761, −1.4209])
Poland	−0.7179 ([−0.9788, −0.4569])	−3.2242 ([−5.1384, −1.3101])
Portugal	−0.3187 ([−0.4690, −0.1684])	−1.8667 ([−3.0606, −0.6727])
Slovakia	−0.5527 ([−0.7882, −0.3172])	−4.6000 ([−6.9414, −2.2586])
Slovenia	−0.4987 ([−0.7374, −0.2600])	−2.7030 ([−4.7038, −0.7023])
Spain	−0.3778 ([−0.5023, −0.2532])	−0.9636 ([−2.4118, 0.4845])
Sweden	−0.1982 ([−0.2772, −0.1191])	−2.2000 ([−3.4149, −0.9851])
Switzerland	−0.4130 ([−0.5136, −0.3125])	−4.0182 ([−4.9935, −3.0428])

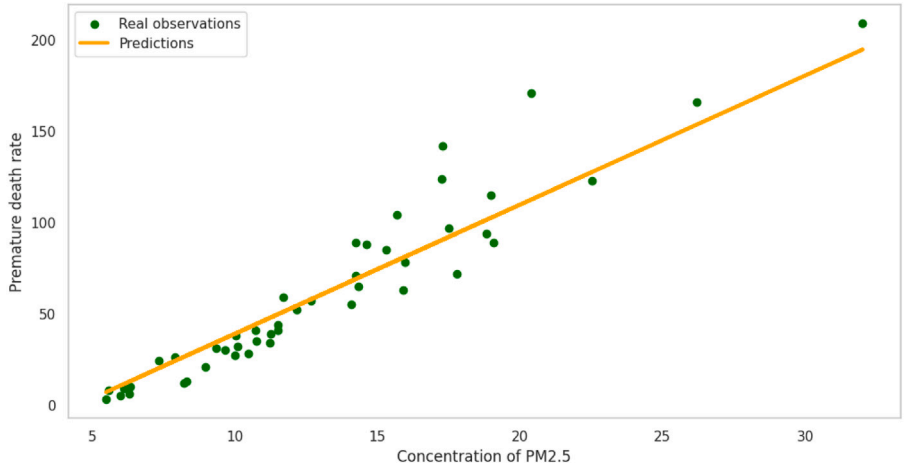


Fig. 12. Prediction of premature death from the concentration of fine particle matter in test period.

This approach allows for a rigorous evaluation of the model, ensuring that each observation is used for both training and testing, thus guaranteeing an accurate measure of its predictive performance.

Leave-One-Out validation involved excluding a specific year's data from the training set and using this excluded data for testing. This process was repeated for each year of the period 2010 to 2019. A degree 2 polynomial transformation was used for $PM_{2.5}$ concentrations, capturing possible non-linear relationships between this variable and the premature death rate. Each iteration of the model was trained with data from all years except the test year, making predictions for the excluded year.

The polynomial regression model showed an R-squared (R^2) value of 0.78, indicating that 78% of the variability in premature death rates can be explained by the concentration of $PM_{2.5}$. The model also had a MSE of 576.45, a MAE of 15.80, and a MedAE of 10.75.

Fig. 13 shows how our prediction model captures trends in the rate of premature deaths in Europe. A clear decrease is evident since 2010, although, as mentioned above, this trend has fluctuated over time.

4. Conclusions

The quality of life and general well-being of people throughout the world are seriously threatened by air pollution, a global public health issue. Our research has demonstrated the strong correlation between air pollution and the rates of premature death, highlighting the pressing need for a comprehensive and efficient response to this issue. The information provided here shows that exposure to airborne fine particulate matter ($PM_{2.5}$) is strongly linked to a dramatic rise in the number of premature deaths in several different nations and areas. This concerning association highlights how vital it is to put policies and procedures in place to mitigate air pollution in order to safeguard public health and lessen the devastation caused by these diseases linked to air pollution.

After analyzing statistically, the areas and countries in our database, it was observed that there were significant differences in $PM_{2.5}$ levels between the different urban, rural and urban areas, as well as between the different countries under study. Post-hoc tests revealed that the

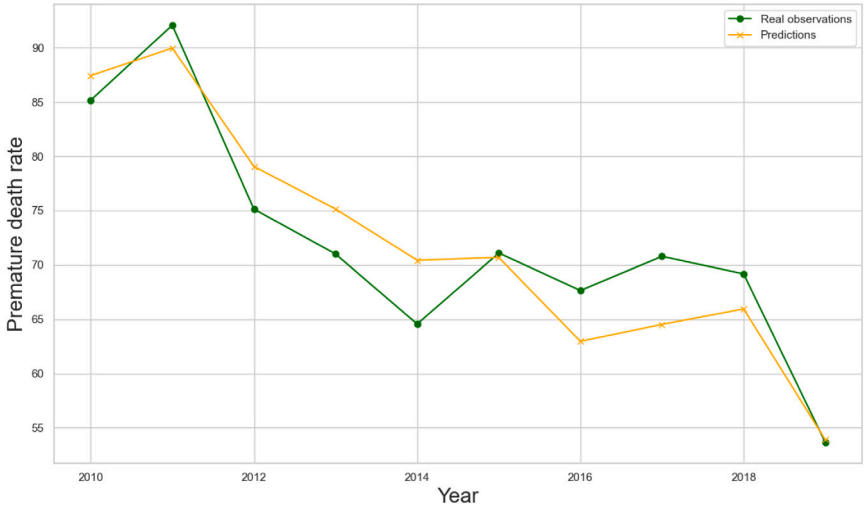


Fig. 13. Prediction of premature death for each year.

largest differences were between urban and rural areas, as well as between cities and urban areas, but we cannot affirm the existence of differences between large and small cities, as well as between small cities and urban areas. Similarly, we observed that countries with significant influence on air pollution levels exhibited similar patterns, likely due to their proximity and neighborhood.

We also analyzed the temporal trends, and the concentration of $PM_{2.5}$ has decreased in most of the countries studied, suggesting possible improvements in air quality. This is something we could intuit, as it aligns with the goals set out by the Sustainable Development Goals (SDGs) of the 2030 Agenda. However, this decrease does not occur in premature mortality rates, which do not follow a uniform pattern and some countries experience worrying increases in some periods. The lack of corresponding reductions in premature mortality rates in some countries indicates that achieving the broad objectives of SDG 11 requires more than just improvements in air quality. The complex relationship between air quality and health outcomes suggests that other factors, such as healthcare quality, socioeconomic conditions, and existing health disparities, play significant roles. Therefore, while reducing $PM_{2.5}$ levels is a critical step, it must be complemented by comprehensive policies that address these broader determinants of health.

As demonstrated in the lineal regression model, a significant association was found between $PM_{2.5}$ concentration and premature mortality rates. The high precision of the model created highlights the importance of air quality in public health, as it predicts mortality rates from concentration.

These findings underscore the need to implement effective air pollution control policies to reduce health risks associated with exposure to $PM_{2.5}$. Given the strong correlation between $PM_{2.5}$ concentration and premature mortality rates, it is imperative that governments and public health organizations prioritize the development and enforcement of stringent air quality regulations. These policies should aim not only to limit emissions from major sources such as industrial activities, transportation, and residential heating but also to promote cleaner technologies and renewable energy sources. In addition, the identification of temporal trends and geographical patterns can inform the development of specific interventions adapted to the needs of each region.

It is important to consider the limitations of the study, such as data availability and underlying assumptions in the statistical models used. Further research is needed to better understand the underlying factors contributing to observed differences in air pollution and premature mortality rates. Continuous validation of the regression model and the collection of up-to-date data are essential to improve prediction accuracy and maintain the relevance of conclusions over time.

CRediT authorship contribution statement

Laura Sanz-Martín: Writing – review & editing, Writing – original draft, Visualization, Validation, Software,Methodology, Investigation. **Javier Parra-Domínguez:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation. **Juan Manuel Corchado:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This research is part of the International Chair Project on Trustworthy Artificial Intelligence and Demographic Challenge within the National Strategy for Artificial Intelligence (ENIA), in the framework of the European Recovery, Transformation and Resilience Plan. Reference: TSI-100933-2023-0001. This project is funded by the Secretary of State for Digitalization and Artificial Intelligence, Spain and by the European Union (Next Generation).

References

Ahmad, H., Yaqub, M., Lee, S.H., 2023. Environmental-, social-, and governance-related factors for business investment and sustainability: A scientometric review of global trends. Environ. Dev. Sustain. 1–23.

Brown, M.B., Forsythe, A.B., 1974. Robust tests for equality of variances. J. Amer. Statist. Assoc. 69, 364–367. http://dx.doi.org/10.1080/01621459.1974.10482955.

Dunn, O.J., 1961. Multiple comparisons among means. J. Amer. Statist. Assoc. 56, 52–64. http://dx.doi.org/10.2307/2282330.

Eurostat, 2023. Air pollution statistics - emission inventories.

Fusco, F., Ricci, P., 2019. What is the stock of the situation? A bibliometric analysis on social and environmental accounting research in public sector. Int. J. Public Sect. Manag. 32 (1), 21–41.

Golondrino, G.C., Muñoz, W.Y.C., Martínez, L.M.S., 2020. Aplicación de la regresión polinomial para la caracterización de la curva del COVID-19, mediante técnicas de machine learning. Invest. Innov. Ingen. 8 (2), 87–105.

Kedem, B., Fokianos, K., 2005. Regression Models for Time Series Analysis. John Wiley & Sons.

Kruskal, W.H., Wallis, W.A., 1952. Use of ranks in one-criterion variance analysis. J. Amer. Statist. Assoc. 47 (260), 583–621, https://www.jstor.org/stable/2280779.

Levene, H., 1960. In: Olkin, I., et al. (Eds.), Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling. Stanford University Press, pp. 278–292.

MacQueen, J.B., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of 5-Th Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1, pp. 281–297, (14).

Mensah, J., 2019. Sustainable development: Meaning, history, principles, pillars, and implications for human action: Literature review. Cogent Social Sci. 5 (1), 1653531.

Mohd Razali, N., Yap, B.W., 2011. Power comparisons of shapiro–wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests. J. Statist. Model. Anal. 2 (1), 21–33.

Na, S., Xumin, L., Yong, G., 2010. Research on k-means clustering algorithm: An improved k-means clustering algorithm. In: 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, Jian, China. pp. 63–67. http://dx.doi.org/10.1109/IITSI.2010.74.

Nabiyeva, G.N., Wheeler, S.M., London, J.K., Brazil, N., 2023. Implementation of sustainable development goal 11 (sustainable cities and communities): initial good practices data. Sustainability 15 (20), 14810.

Parra-Domínguez, J., Gil-Egido, A., Rodríguez-González, S., 2022. SDGs as one of the drivers of smart city development: The indicator selection process. Smart Cities 5 (3), 1025–1038.

Rashed, A.H., Shah, A., 2021. The role of private sector in the implementation of sustainable development goals. Environ. Dev. Sustain. 23, 2931–2948.

Shapiro, S.S., Wilk, M.B., 1965. An analysis of variance test for normality (complete samples). Biometrika 52 (3/4), 591–611. http://dx.doi.org/10.2307/2333709.

Takase, C., 2018. Implementing SDG 11–Key Elements, Challenges and Opportunities. The United Nations Centre for Regional Development.

Tremblay, D., Fortier, F., Boucher, J.F., Riffon, O., Villeneuve, C., 2020. Sustainable development goal interactions: An analysis based on the five pillars of the 2030 agenda. Sustain. Dev. 28 (6), 1584–1596.

Viñals, M.P., 2009. vol. 64, Univ. Politèc. de Catalunya.

Vitolla, F., Raimo, N., Rubino, M., 2019. Appreciations, criticisms, determinants, and effects of integrated reporting: A systematic literature review. Corp. Soc. Responsib. Environ. Manage. 26 (2), 518–528.

World Health Organization, 2021. WHO Global Air Quality Guidelines: Particulate Matter (PM2.5 and PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide. World Health Organization, Geneva.