THE EFFECT OF AIR POLLUTANTS ON CHRONIC RESPIRATORY DISEASE PREVALENCE AND MORTALITY

Veljko Petrovic

STUDENT NUMBER 12304257

Introduction and motivation

This assignment aims to explore the critical issue of how air pollution impacts the prevalence of chronic respiratory diseases (CRD) in metropolitan France. As the country transitions to greener energy sources, understanding the health effects of various air pollutants on French citizens is a pressing concern.

In this study, we use several annually measured air pollutants as proxies for air pollution: Nitrogen Dioxide (NO_2), Ozone (O_3), Particulate Matter (PM10), and Fine Particulate Matter (PM2.5). The objective is to determine whether these pollutants are associated with mortality rates attributed to respiratory causes.

Explore the data in R

Motivated by the goal of examining the relationship between air pollution and chronic respiratory diseases within the provided dataset ("Dataset"), I explored the availability of data on Average PM2.5 and Average PM10 levels. Using the is.na function, I identified missing values in both categories. Additionally, data on the prevalence of chronic respiratory diseases were incomplete, as were critical details on age and gender, which are essential for understanding how pollutants might differentially affect various demographic groups. To address these gaps, I chose to expand the analysis to include the mentioned elements and to incorporate a more detailed examination of different air pollutants beyond PM2.5 and PM10.

Enriching the dataset

To enhance the dataset, I incorporated additional data sources. Beyond addressing the missing values for Average PM2.5 and Average PM10, I included data on Nitrogen Dioxide (NO_2), Ozone (O_3), Particulate Matter (PM10), and Fine Particulate Matter (PM2.5). These data were sourced from the Ineris website[1], providing annual averages per department, weighted by population, and accounting for urban, peri-urban, and non-urban areas. Areas with significant industrial or traffic activity were excluded.

Medical data were supplemented with the number and prevalence of chronic respiratory diseases (excluding cystic fibrosis), coded as *Ntop and prev*, respectively, and adjusted for age and gender. Additionally, data on the number and density of pulmonologists (coded as *effectif* and *Density*) were included[2]. Finally, mortality rates caused by respiratory diseases (excluding COVID-19), coded as *Total_Deaths*, were added[3].

Since most of the relevant data were sourced externally, I created a “new” dataset by merging the mentioned data sources rather than using the original dataset as the base. This decision was made to avoid retaining an excessive number of irrelevant variables present in the original dataset. However, the same merging techniques (e.g., inner_join, left_join) could have been applied if the original dataset had been used as the foundation. The variables *Department* and *Year* served as common keys for merging. During the merging process, variable names were modified to ensure uniqueness (e.g., *Department*) and facilitate integration. The final variable names referenced in this document correspond to the **pdata** dataset used for the econometric analysis. All changes are documented in the R code to ensure transparency and allow the reader to follow the workflow easily. Additionally, the **Age** variable was standardized to have consistent ranges across all datasets.

Define a relevant and appropriate econometric question

The main econometric question addressed in this analysis is how air quality impacts the prevalence of chronic respiratory diseases (CRD) and mortality rates. It is essential to acknowledge that this approach faces many confounding factors, making it challenging to argue for strong causal relationships. However, like pieces of a puzzle, each analysis, with its specific controls, contributes to the broader body of literature on this critical issue. Finding significant results in the hypothesized direction can provide evidence for further, more controlled research. Therefore, this type of high-level analysis holds importance, even if the findings are more correlational than causal.

The dependent variables of interest are the prevalence of CRD **(**coded: *prev***)** and the mortality rate **(**coded: *Total_Deaths***)**. These are calculated by dividing the number of registered patients with CRD and the number of deaths from respiratory diseases by the population covered by insurance (coded: *Npop*). This approach adjusts for the size differences among French departments, as population size is not explicitly included as a control variable.

The independent variables are population-weighted air pollutants, specifically NO_2 (*NO2_PopWeighted*), O_3 (*O3_PopWeighted*), SOMO35 (*SOMO35_PopWeighted*), PM10 (*PM10_PopWeighted*), and PM2.5 (*PM2_5_PopWeighted***)**. A detailed description of these pollutants is provided in the appendix. The model also includes **control variables**: the density of pulmonologists per 100,000 people (coded: *Density*), gender (a dummy variable coded: Gender_Dummy, where 1 = "Hommes" and 0 = "Femmes"), and **age**.

Age is included as a categorical variable, as being older is strongly associated with increased death rates. The age categories range from **0-24 years** (the first group) to **95+ years**, with each subsequent range spanning 10 years (e.g., 25-34). The decision to group ages in this manner is both practical and theoretical. Practically, detailed age granularity for mortality data was not available. Theoretically, older individuals face significantly higher risks of developing CRD and mortality, warranting a more nuanced analysis of these age groups. While this approach does not provide detailed insights into CRD prevalence among children, this limitation is not critical for the current research focus. Future studies could examine CRD prevalence specifically in pediatric populations.

Analyse (with descriptive statistics) the subset of data of interest/selected.

Using the *summary* function, we gain a general understanding of the variables of interest. However, to implement an appropriate econometric approach, it is essential to ensure significant variability in the data. This is assessed using the coefficient of variation (CV), a statistical measure that compares the ratio of the standard deviation to the mean, often expressed as a percentage, to evaluate relative variability. Given the panel nature of the dataset, variability is examined both within departments (over time) and between departments (in a single year). Most variables of interest show a CV above the commonly used 0.1 threshold. However, the density of pulmonologists (Density) remains relatively constant, suggesting limited variability.

Another critical concern is the high likelihood of multicollinearity among the air pollutant variables, which can inflate standard errors and reduce the precision of the results. To address this, a correlation matrix (included in the appendix) was first created to identify potential

collinear relationships. Subsequently, the Variance Inflation Factor (VIF) test was conducted to quantify how much multicollinearity inflates the variance of regression coefficients. Based on this analysis, the variables retained in the model are *NO2_PopWeighted*, *SOMO35_PopWeighted*, and *PM2_5_PopWeighted*, with VIF values of 2.15, 1.01, and 2.15, respectively. The iterative selection process is documented in the accompanying code.

In addition to multicollinearity, the presence of non-normal residuals (explored further in the next sections) necessitated the use of robust standard errors to ensure reliable inference and address potential violations of regression assumptions.

Implement and explain your strategy in terms of econometric approach.

Many factors can influence the development of chronic respiratory diseases (CRD) and their relationship with air pollution, including personal habits (e.g., smoking, physical activity), climate (e.g., average temperature), and economic well-being (e.g., average income per department). Many of these factors are geographically specific, such as climate, or tied to the cultural identity of particular departments. France is one of the most diverse countries in Europe, with regional variations in habits, cultural activities, and dietary patterns, which may affect health outcomes differently. Additionally, some departments are more rural while others are urbanized. These factors could represent time-invariant and unobservable characteristics that impact the variables of interest.

Given this context, a fixed effects (FE) model is econometrically appropriate. The FE model accounts for time-invariant entity-specific effects, which are assumed to influence the dependent variables but not vary over time. This model essentially removes these effects by de-meaning the data, focusing the analysis on variations within entities (departments) over time. On the other hand, random effects (RE) models assume that these entity-specific effects are random and uncorrelated with the predictor variables, treating them as part of the model's error term. Based on the nature of the data, the FE model appears to be more suitable.

To formally test this intuition, I used the Hausman test to determine whether a FE or RE model is more appropriate. The Hausman test evaluates whether the unique errors (random effects) are correlated with the predictors. If they are, the FE model is preferred, as the RE model would yield biased results. If there is no correlation, the RE model is more efficient. The results of the Hausman test rejected the null hypothesis (H0) for both dependent variables, indicating that the FE model is statistically and intuitively more appropriate (prev: chisq = 28.819, df = 12, p-value = 0.00419; Total_Deaths: chisq = 50.102, df = 12, p-value < 0.00001).

To implement this analysis, I created four models: two FE models (one for each dependent variable, **prev** and **Total_Deaths**) and two RE models for comparison. After confirming the suitability of the FE model, further steps addressed multicollinearity and heteroscedasticity by computing robust standard errors. Robust standard errors also accounted for the violation of the normality assumption for residuals. Normality was assessed using the Kolmogorov-Smirnov test (as the sample size of 5000 data points was too large for the Shapiro-Wilk test). For both dependent variables, the normality assumption was violated (prev: D = 0.032593, p-value < 0.00001; Total_Deaths: D = 0.1337, p-value < 0.00001).

Interpret your results and answer to the initial research question.

Prevelence:

The analysis indicates a small but highly significant positive effect of population-weighted NO_2 concentration (*NO2_PopWeighted*) on the prevalence of chronic respiratory diseases (Estimate: 0.00037, p-value = 0.00000). Specifically, a 1-unit increase in population-weighted NO_2 concentration is associated with a 0.037% increase in CRD prevalence, holding other factors constant. Additionally, males are found to have a 1.69% higher likelihood of developing CRD compared to females (Estimate: 0.01697, p-value = 0.00000).

Regarding age, the reference group is the 0–24 age bracket. As expected, CRD prevalence increases with age starting from 45 years onwards, while the 25–44 age groups show a lower likelihood of CRD. This finding may reflect a stylized fact specific to the dataset or suggest that children and young adults are vulnerable to CRD, while physically robust adults are less likely to develop the condition. These patterns warrant further investigation, including potential interaction effects between age and other variables.

Mortality

The mortality analysis provides stronger conclusions. It identifies population-weighted NO_2 concentration (*NO2_PopWeighted*) as a strong and highly significant predictor of mortality, with a 1-unit increase in NO_2 associated with a 0.717 increase in the mortality rate (p-value = 0.00000). Similarly, SOMO35 shows a positive and significant effect (Estimate: 0.00091, p-value = 0.00001), though its magnitude is relatively small. Unexpectedly, PM2.5 exhibits a significant negative relationship with mortality (Estimate: −0.691, p-value = 0.00001), which may be attributed to multicollinearity or other confounding factors requiring further exploration.

Gender is not found to be a significant predictor of mortality (p-value = 0.29870), whereas age emerges as a major determinant. Older age groups exhibit significantly higher mortality rates, as expected, with the magnitude of the effect increasing substantially in the oldest cohorts.

For a more detailed breakdown, refer to the table in the appendix.

Critically discuss your work

As discussed at the outset, this approach has significant limitations in establishing a causal relationship between air pollutants and chronic respiratory diseases (CRD). Disentangling the effects of different air pollutants is inherently challenging, and individual interpretations of their impacts should be approached with caution. High multicollinearity among the pollutants reduces the precision of the estimates. While robust standard errors mitigate some of these issues, they do not fully resolve them, potentially contributing to surprising results, such as the negative effect of PM2.5 on mortality. Nevertheless, when taken together, the air pollutants serve as a reasonable proxy for air quality, offering valuable insights into the research question.

The measures of air pollution themselves are imperfect, as they rely on annual averages that may not capture temporal fluctuations or spatial concentrations of pollutants. Population density and pollution intensity are key factors, and while population-weighted measures were computed to address this issue, a more nuanced approach would provide deeper insights.

Additionally, many individual-level factors, such as differences in smoking rates, were not accounted for due to the lack of granular data. Including these variables could have enhanced the analysis but was not feasible given the available data.

Lastly, time lags in the relationship between air pollution and CRD development were not addressed. The study period (2015–2020) may be too short to fully capture the long-term effects of air pollution, as CRD typically develops over an extended period. Future research should consider longer time horizons to better understand these dynamics.

Appendix:

Air pollutant explanation:

Annual average concentration of NO_2 ($\mu g/m^3$): The average yearly concentration of nitrogen dioxide (NO_2) in the air, measured in micrograms per cubic meter. This is a key air pollutant that affects respiratory health.

Population-weighted annual average concentration of NO_2 ($\mu g/m^3$): This is the annual average NO_2 concentration adjusted for population distribution, giving more weight to areas where more people live.

Annual average concentration of O_3 ($\mu g/m^3$): The average yearly concentration of ozone (O_3) in the air, also measured in micrograms per cubic meter. Ozone at ground level is a harmful air pollutant.

Population-weighted annual average concentration of O_3 ($\mu g/m^3$): Similar to NO_2, this measures the average ozone concentration weighted by the population distribution.

Annual average of SOMO35 ($\mu g/m^3 \cdot day$): SOMO35 (Sum of Ozone Means Over 35 ppb) is an indicator of ozone exposure. It represents the total of daily maximum 8-hour average ozone concentrations exceeding 35 parts per billion (ppb).

Population-weighted annual average of SOMO35 ($\mu g/m^3 \cdot day$): The SOMO35 metric weighted by population, highlighting areas where ozone exposure affects more people.

Annual average of AOT40 ($\mu g/m^3 \cdot hour$): AOT40 (Accumulated Ozone exposure over a Threshold of 40 ppb) is used to estimate ozone exposure affecting vegetation and crops. It sums the excess concentrations over 40 ppb during daylight hours.

Annual average concentration of PM10 ($\mu g/m^3$): The average yearly concentration of particulate matter (PM10), particles with a diameter of 10 micrometers or smaller, which can penetrate the respiratory system.

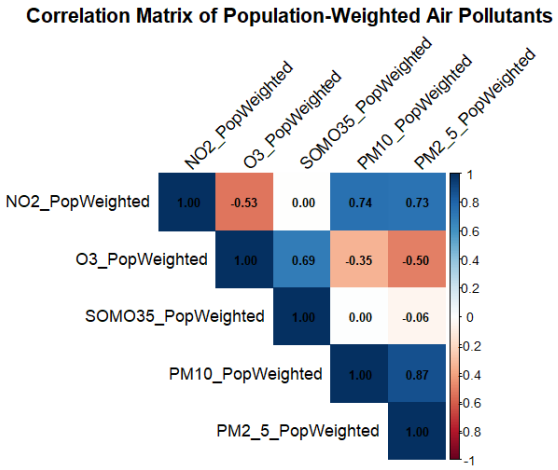
Population-weighted annual average concentration of PM10 ($\mu g/m^3$): The PM10 metric adjusted for population distribution.

Annual average concentration of PM2.5 ($\mu g/m^3$): The average yearly concentration of fine particulate matter (PM2.5), which consists of particles smaller than 2.5 micrometers. These are particularly harmful as they can enter the bloodstream.

Population-weighted annual average concentration of PM2.5 (μg/m³): The PM2.5 concentration adjusted to reflect population distribution.

Fixed Effects Models with Robust Standard Errors

	Dependent variable:	
	Prevalence	Mortality
	(1)	(2)
NO2 (Pop. Weighted)	0.0004***	0.717***
	(0.0001)	(0.084)
SOMO35 (Pop. Weighted)	-0.00000*	0.001***
	(0.00000)	(0.0001)
PM2.5 (Pop. Weighted)	-0.00004	-0.691***
	(0.0001)	(0.129)
Gender (Dummy)	0.017***	0.291
	(0.001)	(0.280)
Age: 25-34	-0.013***	-0.175***
	(0.001)	(0.035)
Age: 35-44	-0.005***	0.387***
	(0.001)	(0.055)
Age: 45-54	0.008***	2.606***
	(0.001)	(0.226)
Age: 55-64	0.031***	10.197***
	(0.001)	(0.866)
Age: 65-74	0.053***	24.185***
	(0.001)	(1.818)
Age: 75-84	0.081***	52.120***
	(0.001)	(3.638)
Age: 85-94	0.098***	99.681***
	(0.001)	(6.437)
Age: 95+	0.090***	26.499***
	(0.002)	(1.718)
Observations	10,368	10,368
R2	0.890	0.651
Adjusted R2	0.889	0.648
Note:	*p<0.1; **p<0.05; ***p<0.01	
	Robust standard errors in parentheses.	



References:

Dataset links

1)https://www.ineris.fr/fr/recherche-appui/risques-chroniques/mesure-prevision-qualite-air/qualite-air-france-metropolitaine

2) https://data.ameli.fr/explore/dataset/effectifs/information/

3) https://opendata-cepidc.inserm.fr/

NB: Some of the dataset have been modified in excel mostly to delete excess columns and translate to English.