Hidden Damages: Facing Heterogeneity in the Effects of Air Pollution on Regional GDP*

John Gomez-Mahecha[†1,2], Lara Aleluia Reis[1,2,3] and Cristina Cattaneo[1,2]

[1] *RFF-CMCC European Institute on Economics and the Environment, Italy*

[2] *CMCC Foundation – Euro-Mediterranean Center on Climate Change, Italy*

[3] *CENSE – Center for Environmental and Sustainability Research & CHANGE - Global Change and Sustainability Institute, NOVA School of Science and Technology, Portugal*

June 4, 2026

Click here for the latest version

Abstract

Standard estimates of the economic costs of air pollution assume (directly or indirectly) that causal effects and instrument strength are homogeneous across regions. We relax this assumption and show that it matters. Using a panel of European NUTS3 regions from 2000 to 2022, we estimate the effect of $PM_{2.5}$ on regional GDP per capita through a strategy that jointly accounts for heterogeneity in both the instruments' first stage and the causal response. For the average effect, we implement the Flexible IV (FIV) estimator from the Double-Debiased Machine Learning literature, allowing instrument strength to vary across biogeographical regions. Our preferred estimate is a 0.3 percent decline in GDP per capita per $\mu g/m^3$ increase in annual $PM_{2.5}$ — approximately one-third of the reported by conventional two-stage least squares specifications in previous work. Coupling the FIV with an IV Causal Forest yields conditional causal responses that span -1.22 to 0.1 percent, with a long left tail. Statistically significant heterogeneity is driven by regional dependence on agriculture and manufacturing, and by historical exposure to pollution, reflecting the interplay between vulnerability and adaptive capacity. We then allocate the localized damage estimates across income quintiles using microdata from the EU Labor Force Survey and the Survey on Income and Living Conditions. The resulting distributional pattern is non-monotonic: middle-income quintiles bear the largest proportional losses while the top quintile is the least affected.

*This project has received funding from the European Union's Horizon Europe research and innovation program under the AdJUST project (grant agreement no. 101069880 – CUPC53C21000280006). 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. All errors are of our own.

[†]Corresponding coauthor: john.gomezmahecha@cmcc.it

1 Introduction

Air pollution is a major threat to global health. For example, only in 2023, the Health Effects Institute (HEI) reports, globally, 7.9 million deaths and the loss of 232 healthy-years-of-life (Health Effects Institute, 2025). In addition, higher levels of pollution have been associated with a decrease in labor participation (Aragón et al., 2017; Hanna and Oliva, 2015; Montt, 2018), absenteeism (Zivin and Neidell, 2012; Aragón et al., 2017; Fan and Grainger, 2023), loss of labor and capital productivity (Li et al., 2017; Avnery et al., 2011; Chang et al., 2016, 2019; Dong et al., 2021; Lichter et al., 2017; Sarmiento, 2022), within other intellectual and physical abilities (Aguilar-Gomez et al., 2022; Zivin and Neidell, 2018). Together, these impacts generate substantial losses in economic activity, in addition to the even greater toll on human lives.

Estimating the impact of pollution on regional economic output (Dechezleprêtre et al., 2020a,b; Gan et al., 2021; Behrer et al., 2023; Avila-Uribe, 2023) is key to assessing the consequences of policies affecting (directly or indirectly) air pollution. However, identifying a causal relationship between pollution and economic activity is challenging because pollution levels are themselves influenced by economic conditions. To address this endogeneity, the literature has usually employed a relatively standard strategy based on an instrumental variable (IV) approach. By isolating exogenous variation in air pollution, this strategy identifies average causal effects on economic output.

However, this literature has often understudied potential heterogeneity in these estimations. A growing body of evidence suggests that both exposure to and damages from air pollution are uneven across space and socio-economic groups (Health Effects Institute, 2025; European Environmental Agency, 2018; Rentschler and Leonova, 2023). Regional air pollution levels reflect differences in industrial structure, geography, and meteorology. Meanwhile, vulnerability, the scope for adaptation and awareness, determine how a given exposure translates into economic losses. Together, these factors make air pollution a plausible driver of regional differences and distributional inequality. Effective and equitable environmental policy requires going beyond average effects, which mask economically meaningful heterogeneity across regions and socioeconomic characteristics. To understand where damages are particularly severe and why, it is paramount to identify the most vulnerable groups, deepen our understanding of the mechanisms driving these disparities, and clarify the underlying factors shaping them.

Estimating heterogeneous causal effects of pollution on regional GDP (or other economic aggre-

gates) in an IV setting faces three compounding methodological challenges that, to our knowledge, have not been jointly addressed in the existing literature. First, modeling heterogeneity in the effect of pollution on GDP, i.e, heterogeneity in IVs' second stage, requires the fragmentation of the sample into pre-specified subgroups defined by a set of heterogeneity covariates. The existing heterogeneity analyses have been typically limited to a subset of these variables and to one variable at a time, because of concerns of multiple testing and statistical power. However, when heterogeneity is driven by interactions among multiple covariates, these approaches will most likely fail to detect the true heterogeneous structure (Athey et al., 2019; Wager and Athey, 2018; Athey et al., 2025). Second, heterogeneity is expected not only in the second stage, but also in the IVs' first stage. Just as the effect of pollution changes based on the regions' particular combination of vulnerability and exposure, the relevance of the instruments to account for changes in pollution is also likely to vary spatially. Instruments based on atmospheric conditions—such as thermal inversions or vertical and horizontal ventilation—may generate differential changes in pollution across regions, as the same atmospheric shock interacts with region-specific characteristics that determine pollutant formation, concentration, and dispersion. Abadie et al. (2024) show that heterogeneity in the First Stage is prone to inefficiencies, but in extreme forms could prevent the causal interpretation of the estimates of the average effects. Finally, these two sources of heterogeneity interact as the ability to estimate heterogeneous treatment effects in an IV setting depends critically on maintaining first-stage strength over the support of the heterogeneity covariates (Athey et al., 2025). Therefore, accounting for heterogeneity in the First Stage to limit the risk of weak first stages will be vital to address the challenge imposed by the description of the second-stage heterogeneity structure.

In this paper, we estimate the effect of $PM_{2.5}$ air pollution on GDP per capita in NUTS3 regions in Europe, from 2000 to 2022, allowing for the expression of heterogeneity in the first and second stages. $PM_{2.5}$ is one of the most commonly used indicators of ambient air pollution in the literature, given that its health consequences are especially well established and severe.[1] We propose a novel strategy that jointly tackles all three methodological challenges discussed above. This strategy allows us to improve the current estimates of the average effect, provide novel estimates of individualized (region-by-year) effects and use them to formulate informative indicators of disproportionate costs of pollution across relevant socio-economic groups.

In particular, this paper achieves three goals. First, it provides a set of robust estimators of the average effect of pollution on the GDP per capita, in this case, the Average Causal Response

[1]For example, Health Effects Institute (2025) attributes more than half of the premature deaths in 2023 to $PM_{2.5}$

(ACR), that account for heterogeneity in the First Stage. To do so, we implement the Flexible IV (FIV) estimator, from the Double-Debiased Machine Learning (DDML) literature (Chernozhukov et al., 2018; Ahrens et al., 2025), allowing the first stage to vary across space, and in particular bio-regions, while preventing the fall into the many-IV bias (Abadie et al., 2024). Our preferred estimate suggests that an increase of 1 $\mu g/m^3$ in the annual average $PM_{2.5}$ decreases regional GDP by an average of 0.3% across Europe. Our estimate is smaller than the less-flexible homogeneous model underlying, the Two-Stage Least Squares (2SLS) estimator, would suggest, with previous papers reporting values between 0.81 and 1.06 percent. We find our results are robust to different specifications and data sources.

Second, we provide estimators for individualized estimates of the annual effects of pollution on regional GDP, namely the Conditional Average Causal Responses (CACRs) to pollution. In particular, we propose to couple the FIV with the localized moment equations estimator of the IV Causal Forest (IVforest) proposed by Athey et al. (2019), to allow flexibility in the first stage while accounting for possible heterogeneity in the second. To the best of our knowledge, this is the first paper proposing such coupling. In particular, we found that the effects of pollution on the GDP per capita are non-symmetrically distributed, with a long left tail and values ranging from -1.22% to 0.1%. We not only found statistical evidence of heterogeneity but also showed that this heterogeneity responds to variables such as the degree of economic and employment dependence on the agricultural and manufacturing sectors, and the historical relative level of pollution. We conclude that the heterogeneity results from a combination of vulnerability to pollution and adaptability.

Finally, we exploit the localized estimates to construct distributional damage indicators by allocating regional GDP losses across income quintiles. Overall, rather than focusing solely on average pollution damages, we explore where the economic costs of air pollution arise and who ultimately bears them. Using EU microdata from the Labor Force Survey (LFS) and the Survey on Income and Living Conditions (SILC), we allocate regional GDP to income groups and evaluate how each group's share responds to changes in pollution. We found that, in fact, the top income quintile is less affected by air pollution than the other income groups. Nonetheless, the middle quintile appears to be the most affected. This provides an example of how assessing heterogeneity in the Second Stage unveils policy-relevant indicators of regional vulnerability and adaptation.

This paper is organized as follows. Section 2 formalizes our empirical strategy for both, estimating robust average effects under first stage heterogeneity and estimating the conditional causal effects. Section 3 presents the results for our three objectives, along with a series of proposed

robustness exercises. Finally, Section 4 concludes with a discussion of the new insights behind the paper and the most important policy conclusions behind our results.

2 Methodology

Given the potential endogeneity of pollution—arising because local pollution levels are jointly determined with economic activity and may also be correlated with omitted regional shocks—the effect of air pollution on regional Gross Domestic Product (GDP) has typically been estimated using linear panel-data instrumental variable (IV) models. In this setting, the IV coefficient can be interpreted as an Average Causal Response (ACR). While this framework delivers a well-defined average effect under proper assumptions on the instruments, it provides limited guidance on how the causal response varies systematically with observable regional characteristics.

Pollution may affect economic aggregates through multiple margins simultaneously: labor input (Hanna and Oliva, 2015; Aragón et al., 2017; Montt, 2018; Holub et al., 2016; Fan and Grainger, 2023; Borgschulte et al., 2022), labor efficiency (Zivin and Neidell, 2012; Chang et al., 2016; Hill et al., 2023; Chang et al., 2019; Batheja et al., 2023; Dong et al., 2021; Archsmith et al., 2018; Lichter et al., 2017; Wang et al., 2022; Adhvaryu et al., 2019; Sarmiento, 2022; Sarmiento and Nowakowski, 2023), sector-specific production technologies(Manono et al., 2025; Rai, 2016; Wu et al., 2025; Li et al., 2017; Kim, 2024; Avnery et al., 2011; Li et al., 2023; Wu et al., 2025; Manono et al., 2025), firm sales (Leroutier and Ollivier, 2024), household income(Borgschulte et al., 2022; Lin et al., 2024), and possibly innovation (Tan and Yan, 2021; Chen et al., 2023; Bracht and Verhoeven, 2022). Whether these micro and meso-level disruptions translate into observable GDP losses depends on sectoral composition, factor mobility, substitution margins, and adaptation. This makes the aggregate effect theoretically ambiguous in magnitude ex ante and likely heterogeneous across regions.

The literature points to three key dimensions of heterogeneity: (1) population vulnerability, (2) economic structure, and (3) baseline exposure levels. These dimensions also offer a useful lens for understanding the underlying mechanisms. The first dimension, population vulnerability, captures the idea that sensitivity to air pollution depends on the characteristics of the affected population. Studies on income and earnings have found evidence of disparities across racial and age compositions (Borgschulte et al., 2022), while Leroutier and Ollivier (2024) find that pollution damages are higher in smaller firms, suggesting that firm-level adaptive capacity moderates the

effect.

The second dimension considers that some economic sectors are more susceptible than others. Authors such as Borgschulte et al. (2022), Lin et al. (2024), Dechezleprêtre et al. (2020b), Dechezleprêtre et al. (2020a), Behrer et al. (2023), and Leroutier and Ollivier (2024), have suggested different effects among economic sectors. The manufacturing and agricultural sectors are often considered to have distinct effects; however, there is still no consensus on which is more affected. For example, results from Leroutier and Ollivier (2024) suggest that it is more important whether sectors sell to final consumers, as in these sectors, productivity shocks can be further amplified by additional pollution-related effects on the demand.

The third dimension is based on the idea that pre-existing levels of pollution could aggravate the negative consequences. The evidence in this case is more conflicting, as the current literature does not agree on the direction or on the marginality of the effect (e.g., Borgschulte et al. (2022) and Behrer et al. (2023) on one side, and Dechezleprêtre et al. (2020a) and Avila-Uribe (2023) on the other).

Notice that, vulnerability, economic structure and exposure dimensions are likely interacting. A region with high exposure but strong adaptive capacity, through sector composition, income levels, previous consciousness of higher exposure or institutional quality, may show smaller damages than a region with moderate exposure but high structural vulnerability. Similarly, rural areas could be more affected by pollution than urban areas because they depend more on a vulnerable economic system. On the other hand, urban areas are more likely to be more exposed. Is only the final integration of multiple factors/covariates that could eventually determine the effect of rurality on the heterogeneity. Such ambiguity can be found in current literature where Borgschulte et al. (2022) and Avila-Uribe (2023) find evidence for more serious damages in rural areas, while Dechezleprêtre et al. (2020a) find evidence on the other direction. Only a multivariate heterogeneity estimator can account for such interactions.

Approaches in the literature that aim to detect heterogeneity while accounting for endogeneity generally rely on one of three strategies: (i) estimating a common specification across subsamples and comparing coefficients (Lin et al., 2024; Gan et al., 2021; Fu et al., 2021; Leroutier and Ollivier, 2024); (ii) augmenting the model with interactions between the pollution variable and group or subsample indicators (Borgschulte et al., 2022; Dechezleprêtre et al., 2020b,a); iii) relying on the availability of outcome variables for each group — for example, sector-specific GDP or sector-specific employment — to infer differential effects without explicitly splitting the sample (Borgschulte

et al., 2022; Lin et al., 2024; Dechezleprêtre et al., 2020b,a; Avila-Uribe, 2023). In some cases, combinations of these strategies are used, or they could be indistinguishable. Their main indicator of heterogeneity is then the differences in the predicted averages across subgroups.

However, all three approaches are unlikely to be satisfactory when heterogeneity is driven by multiple, interacting regional characteristics. These strategies are informative about selected margins but are poorly suited to recovering a general heterogeneity surface. For example, the use of outcome variables for each group is only possible for specific heterogeneity covariates (such as economic sector), and only under a particular study design, making it unsuitable as a general approach. Meanwhile, analysis per subsamples requires the fragmentation of the data, so the analysis can be performed only for a subset of the heterogeneity variables, one variable at a time, as the sample fragmentation raises concerns about multiple testing and statistical power. In addition, when heterogeneity is driven by interactions among multiple of these covariates, these approaches will generally fail to detect the true heterogeneous structure (Athey et al., 2019; Wager and Athey, 2018; Athey et al., 2025), and confound the participation of each of them. Finally, because some of these strategies are not represented as nested models, formal testing of heterogeneity is usually missing. Thus, it is difficult to determine whether observed differences across subgroups reflect genuine treatment effect variation or sampling variability.

At the same time, none of the previous literature that accounted for heterogeneity in an IV setting, has addressed potential heterogeneity in the first stage. More importantly, differences in the strength of the instruments that several authors have already identified, even in their main set of regressions or in the regressions on subsamples (Borgschulte et al., 2022; Dechezleprêtre et al., 2020b,a; Avila-Uribe, 2023), can be the result of heterogeneity in the first stage. In particular, the meteorological instruments used to generate exogenous variation in PM2.5, thermal inversions, horizontal ventilation coefficients, and boundary layer height, are unlikely to affect pollution uniformly across regions, as their strength depends on region-specific characteristics such as topography, prevailing atmospheric conditions, which determine how strongly a given atmospheric shock translates into changes in PM2.5 concentration. As we will discuss in the next section, first stage heterogeneity has the potential to introduce issues in the estimation on the average effect and on disentangle its heterogeneity.

These limitations motivate an estimation strategy that can accommodate flexible first-stage heterogeneity while recovering conditional causal responses over a high-dimensional set of regional characteristics.

To formalize the discussion, consider the following simplified heterogeneous-effects representation:

$$\log GDPpc_{i,t} = \tau_{i,t}\, PM2.5_{i,t} + g(W_{i,t}) + \eta_{c(i),t} + \alpha_i + e_{i,t}, \tag{1}$$

$$PM2.5_{i,t} = Z'_{i,t}\ \rho_{i,t} + h(W_{i,t}) + \eta^{P}_{c(i),t} + \alpha^{P}_{i} + u_{i,t}, \tag{2}$$

where i, t, and $c(i)$ index NUTS3 regions, years, and countries, respectively. Meanwhile, $W_{i,t}$ represents possible observable covariates (such as weather variables), and η and α represent temporal and regional unobservable heterogeneity (potentially different in the first and second stage). $Z_{i,t}$ identifies a set of atmospheric instrumental variables commonly used in the literature, such as thermal inversions, horizontal ventilation coefficients, and boundary layer height.

Explicitly accounting for heterogeneity in this setting raises three related identification challenges. First, the causal effect of pollution on GDP, represented by $\tau_{i,t}$, may vary systematically across regions and over time (second-stage heterogeneity). In that case, standard linear IV estimators, under proper assumptions on the instrumental variable, recover an average causal response, but do not by themselves characterize how the marginal effect could vary with a set of observed heterogeneity covariates, which we will denote as X^{het}. Recovering this structure requires sufficient support over the heterogeneity space and an estimation strategy that can flexibly accommodate multi-dimensional interactions; however, some restriction on the form of heterogeneity may be needed.

Second, the relevance of the instruments, $\rho_{i,t}$ in Equation 2, may also vary across regions and over time (first-stage heterogeneity). In our context, this is especially plausible because atmospheric instruments may induce differential changes in pollution depending on regional characteristics that affect pollutant formation, concentration, and dispersion. As shown by Abadie et al. (2024), ignoring first-stage heterogeneity can reduce efficiency, and in sufficiently extreme cases may undermine the causal interpretation of the resulting IV estimand if it results on violating monotonicity assumptions (Angrist and Imbens, 1995; Angrist et al., 2000).

Third, these two forms of heterogeneity interact. Even if the causal effect varies with observable regional characteristics, such heterogeneity cannot be credibly recovered when instrument relevance weakens across the covariate space. The ability to estimate heterogeneous treatment effects in an IV setting depends critically on maintaining first-stage strength over the support of the heterogeneity

covariates (Athey et al., 2025). The next section develops our strategy in details.

2.1 Estimating average and conditional causal responses under first- and second-stage heterogeneity

Addressing first-stage heterogeneity is a key preliminary step to uncover complex forms of heterogeneity in the second stage. Inspired by Abadie et al. (2024), we argue that the atmospheric phenomena used as instruments may affect $PM_{2.5}$ differently across space, because their effect depends on relatively stable regional characteristics, such as topography, surface energy balance, land-sea interactions, and related features. We assume that these characteristics are partly summarized by bioregions, so that allowing instrument relevance to vary by them captures an important component of the first-stage heterogeneity. In this case, the first-stage equation in Equation 2 can be written as

$$PM2.5_{i,t} = Z'_{i,t}\rho_{g(i)} + h(W_{i,t}) + \eta^{P}_{c(i),t} + \alpha^{P}_{i} + u_{i,t} \tag{3}$$

where $g(i)$ denotes the bioregion of NUTS3 region i, and $\rho_{g(i)}$ allows the relevance of the instruments to vary across them. We assume that each NUTS3 region belongs to one and only one bioregion.

Abadie et al. (2024) show that, when first-stage relevance varies across groups, fully interacting instruments with group indicators under the traditional Two Stage Least Squares (2SLS) estimator is susceptible to the many-IV bias in finite settings, especially when some group-specific instruments are weak or irrelevant. In our setting, this concern is particularly salient because we would like to consider several atmospheric instruments and their interactions with bioregion indicators. Some form of regularization is therefore useful to discipline the effective dimensionality of the first stage and to limit the influence of weak or uninformative instrument-bioregion interactions.

We therefore propose to employ a Flexible Partially Linear Instrumental Variable (FIV) estimator based on the Double/Debiased Machine Learning (DDML) framework of Chernozhukov et al. (2018); Ahrens et al. (2025) and the implementation described by Ahrens et al. (2024). The estimator uses flexible machine-learning methods to estimate nuisance components, including the relationship between $PM_{2.5}$, the instruments, controls, and instrument-bioregion interactions, while employing orthogonalization and cross-fitting to reduce the ACR estimator's sensitivity to first-stage estimation error. This allows us to include a rich set of instrument-bioregion interactions without estimating a fully saturated many-instrument specification by conventional 2SLS. Under

this strategy, Equation 3 will take a more flexible form

$$PM2.5_{i,t} = f(Z_{i,t}, I_{g(i)} \cdot Z_{i,t}) + h(W_{i,t}) + \eta^{P}_{c(i),t} + \alpha^{P}_{i} + u_{i,t}, \tag{4}$$

where $I_{g(i)} \cdot Z_{i,t}$ corresponds to the interactions of bio-region indicators variables and the instruments.

Our implementation of the FIV setting follows the strategy suggested by Ahrens et al. (2024), adapted to our scenario. The nuisance component predictions for the FIV are made using an ensemble estimator that weights predictions from elastic net regressions, random forests, and gradient boosting, combined with five-fold out-of-sample prediction. To account for regional and annual-country fixed effects, we residualized all variables (outcome, endogenous variables, instruments, and covariates) before implementation (see A.2 for details). Finally, the regressions reported after accounting for the nuisance factors are all estimated with clustered standard errors at the NUTS3 level, to account for the panel structure of our data. Further details on the implementation and simulation of the proposed strategy in relevant synthetic scenarios are presented in the appendix (section A.2).

The FIV estimator proposed by Ahrens et al. (2024) has the additional advantage of allowing flexibility in the first stage while accommodating potential heterogeneity in the second stage. We propose to measure potential heterogeneity in the effect of pollution on GDP by implementing the IV Causal Forest (IVforest) proposed by Athey et al. (2019) using the transformed GDP per capita, $PM_{2.5}$, and the set of instruments behind the implementation from Ahrens and colleagues. The Causal Forest is a strategy that combines the estimation of local moment conditions with modifications to the traditional Random Forest algorithm. In particular, the forest is used to build a similarity score based on a set of covariates, which we will call heterogeneity covariates (X^{het}), defining the heterogeneity of the causal parameter of interest. That is, in order to provide evidence of heterogeneity in the effect of pollution on regional GDP, we will compromise on a version of the structural model behind Equation 1, of the form:

$$\log GDPpc_{i,t} = \tau(X^{het}_{i,t}) \cdot PM2.5_{i,t} + g\left(W_{i,t}\right) + \eta_{c(i),t} + \alpha_i + e_{it} \tag{5}$$

where the heterogeneity of the τ can be defined by a set of heterogeneity covariates X^{het}. Athey et al. (2019) show that this strategy will allow the estimation of Conditional (Local) Average Treatment Effects - CATE. In the case of a continuous treatment, this will correspond to Conditional

Average Causal Response (CACR) of the form $E[\tau|X^{het} = x^{het}]$.

These heterogeneity covariates (X^{het}) seek to represent the sources of variation of the effect of pollution on the GDP. As discussed in the previous section, how regions will be affected by pollution is expected to result from the interaction of multiple factors related to population vulnerability, the underlying economic structure, and exposure levels. That is, regions vary in the economic composition of production, with differential importance of vulnerable sectors or production factors, or diversity of economic activities. They also have distinct pollution exposure due to historical differences in pollution levels, different urban-to-rural ratios, different ratios of vulnerable populations in the economically active population, or distinct topographic characteristics.

To the best of our knowledge, this is the first paper to propose such an integration, which combines flexibility in the first stage and heterogeneity in the second stage. The contribution of the coupling is not the transformation itself, which IVforests already require, but the use of the Flexible IV estimator to perform that transformation. An additional nice attribute of the proposed estimator is that X^{het} does not require overlap with the regular IV covariates ($W_{i,t}$). The integration of the FIV with the IVforest would allow us to test the form and composition of this set of variables, check for the existence of heterogeneity, and gain insights into the specific impacts of each variable. Furthermore, as X^{het} potentially varies across regions and time, we could also interpret the estimates as individualized Causal Responses, yielding an estimate of the effect for each NUTS3 region and time in our sample.

It is important to note that this greater flexibility comes with a price. For example, the assumptions required for the validity of the instrument are stronger under the FIV estimator than the standard Two-Stage Least Squares (2SLS). In particular, the possibility of including non-linear relations between instruments and pollution requires full conditional independence (i.e., $E[e|Z, W] = E[e|W] = 0$), stringent than the standard IV orthogonality condition (i.e., $E[Z'e|W] = 0$). In our case, this will not be a significant concern, as it is highly unlikely that the atmospheric phenomenon would affect GDP through other channel than by pollution, especially after conditioning on the main weather indicators in $W_{i,t}$, but it could be a significant concern in other implementations. Another possible concern is that our ability to understand the heterogeneity will depend on having a rich set of observable factors accounted for in X^{het}, but this will depend on data availability at the scale and temporal level of interest.

Data and Implementation Details We implemented the proposed design for a sample of NUTS3 European regions, with annual data between 2000 and 2022. Regional economic outcomes, mainly

GDP, are obtained from Eurostat's regional accounts.[2] Measures of air pollution are derived from the reanalysis of satellite-, simulation-, and monitor-based sources of monthly average $PM_{2.5}$ concentrations developed by Donkelaar et al. (2021) and retrieved from the Washington University in Saint Louis,[3] which we aggregate to the NUTS3 level using population weights[4]. To construct the instrumental variables ($Z_{i,t}$) and weather covariates ($W_{i,t}$), we use high-resolution meteorological data from Copernicus' ERA5 reanalysis from the CDS.[5]. Detailed description of the construction of the dataset can be found in the data subsection, in the methodological appendix (Appendix A.2). Figure 1 summaries the different sources and their corresponding transformation up to the NUTS3-year resolution. Further details on the construction of the data and basic description, can be found in the Appendix A.2, data section.

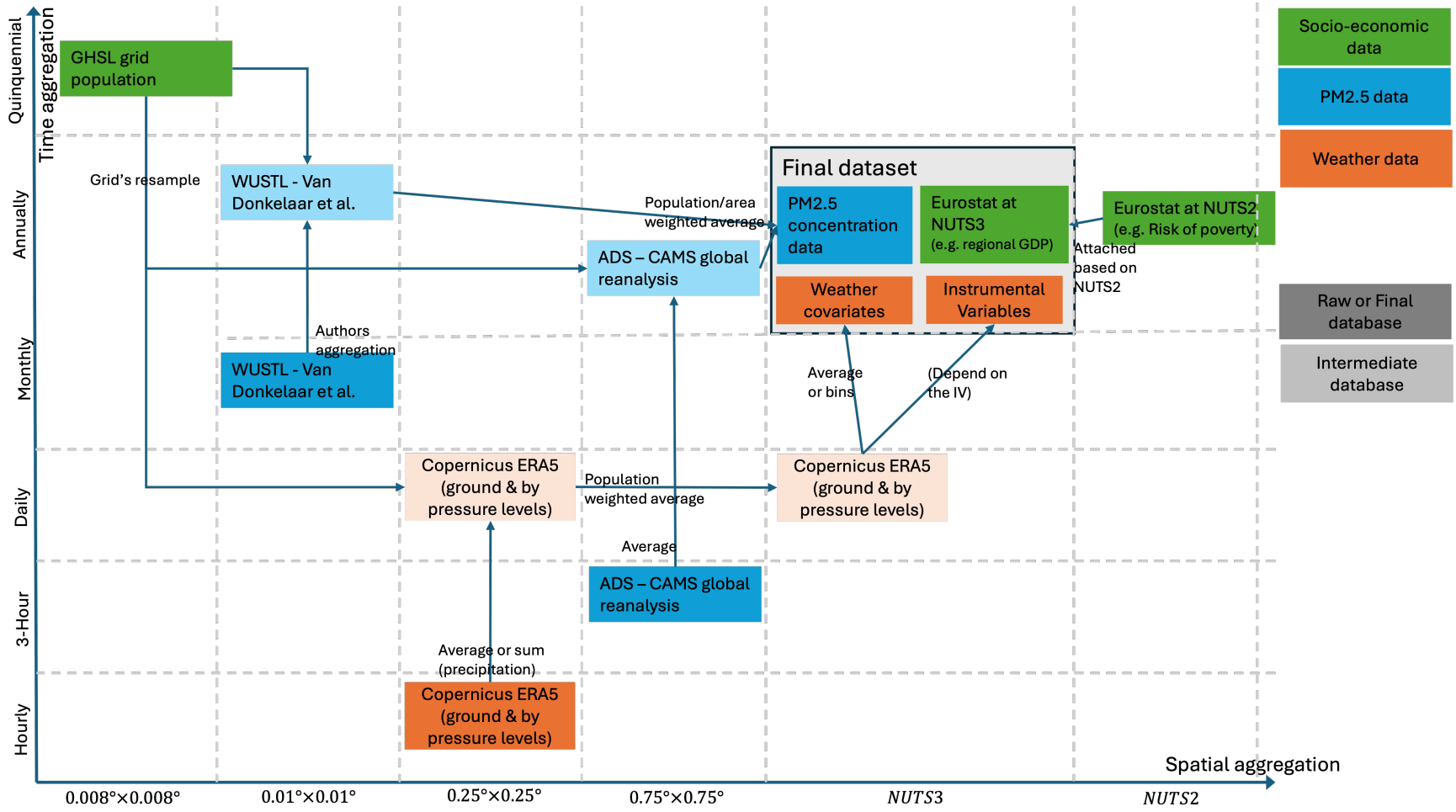
Our set of IV variables includes three potential instruments. We rely on Dechezleprêtre et al. (2020b) and define as the first potential instrument the proportion of days with thermal inversion episodes (TIs). This variable is expected to have a positive effect on air pollution, as inversions prompt the accumulation of air pollutants in the lower atmosphere, including PM2.5. We built a second potential instrument using the planetary boundary layer height (BLH). Previous research has shown that BLH and the concentration of the main pollutant (such as PM2.5 and NO_x) exhibit a negative correlation under regular conditions (Yuval et al., 2020). We use the skewness of the annual distribution, which is positively related to pollution, because higher skewness indicates more days with lower heights and therefore reduced vertical diffusion. Finally, following the strategy proposed by Bracht and Verhoeven (2022), we define a daily ventilation coefficient as a measure of vertical and horizontal ventilation. In our case, we aggregate the daily ventilation coefficient in annual means, which are negatively correlated with pollution, as higher ventilation favors the dispersion of the pollutants, reducing their concentration. Finally, we built a third instrument The literature has extensively discussed the validity of these variables as instruments. Identification relies on the fact that these phenomena are outside human manipulation within the time frame and spatial scale typically used in the statistical models.

[2]Retrieved on July 2024 from here.

[3]Retrieved on July 2024 from here. Version V5.GL.04.

[4]Grid population data is taken from the Global Human Settlement Layer (GHSL), produced by the Joint Research Center (JRC, European Commission). Since GHSL data are produced every five years, we use the nearest past quinquennial benchmark for each analysis year

[5]Retrieved October 2024 from here



Notes: The figure identifies sources for the (1) socio-economic data (green boxes), (2) $PM_{2.5}$ concentration data (blue boxes), and (3) weather data (orange boxes). Vertical axes represent temporal resolutions, and the horizontal axis represents spatial resolutions. For each source, the diagram traces the path from the raw datasets to the final dataset at the desired NUTS3-year level (darker boxes). The arrows represent transformations applied to the data when changing its resolution, and the light boxes represent intermediate transformations. Detailed description of the boxes can be found in Appendix A.2, data section.

Figure 1: Data Sources and Spatial and Time Aggregations

Our set of heterogeneous covariates (X^{het}) intends to capture the main dimensions of heterogeneity suggested by the literature, namely differences in economic structure, population vulnerability, and environmental exposure. For that, we first include the shares of employment and gross value added (GVA) of the primary and secondary sectors in their corresponding totals.[6] These variables proxy for the sectoral composition of regional economies and account for differences in vulnerability to pollution. Agricultural production is directly exposed to environmental conditions, and pollution can affect crop yields and outdoor labor productivity, while manufacturing activity may be affected through labor productivity, operational disruptions, and exposure of industrial processes to environmental conditions. Second, we include population growth as an indicator of regional demographic and economic dynamics.[7] Population growth may reflect differences in urbanization, economic expansion, and the capacity of regions to adapt to environmental shocks, all of which can influence how pollution translates into economic outcomes. Finally, we capture differences in baseline environmental conditions by including the relative pollution level, measured as pollution quantiles. This approach allows the estimated effects to vary across regions with different pollution environments, thereby accounting for potential nonlinearities in pollution damages and differences in adaptation across areas with persistently higher exposure. All variables are expressed as five-year rolling averages excluding the current year t, to prevent contemporaneous endogeneity. In addition, we include year and NUTS2 region indicators to X^{het}, to allow for time-constant or NUTS2-constant heterogeneity components.

2.2 Constructing distributional indicators of pollution damages from estimated CACRs

Modeling heterogeneity is not only key for identification but also for a better understanding of who bears the costs of pollution, and hence its distributional consequences. The estimation of heterogeneous effects, beyond average effects, could allow to uncover substantial variation in the economic damages of air pollution across regions, as we have seen in the sections above. This reflects differences in vulnerability, exposure and adaptive capacity. A key contribution of our approach is that we will exploit these heterogeneous effects to inform policy relevant indicators. In this section, we describe our proposal for using our estimates to construct measures of how the economic costs of air pollution are distributed across income groups, providing new evidence on unequal pollution

[6] GVA retrieved on June 2025 from here. Employment data retrieved on June 2025 from here

[7] Data on population retrieved from here.

burdens.

There is already substantial research on how air pollution affects certain groups of the population more than others. However, most of this literature has focused on exposure rather than vulnerability. Although we expect them to be correlated, the two phenomena may differ. For example, as discussed above, a household that is highly exposed because it lives in a heavily polluted area may nevertheless experience limited effects if it has the resources and knowledge to adapt, for instance through the use of air purifiers, preventive behavior or preventive medicine. In this case, higher exposure does not translate into indoor exposure. Moreover, its income may depend on sectors that are less affected, making the household only weakly vulnerable.

We propose measuring vulnerability by combining the spatial variability in the locations of low-income households and our novel estimates of the pollution effect on regional GDP (i.e., the τ). Let us define the portion of the region j's GDP attributable to the income group q as $GDP_j(q) = \phi_j(q) \times rGDP_j$, where $rGDP_j$ is the regional GDP for region j, and $\phi_j(q)$ is the proportion of the total income of region j that comes from households in income group q. Therefore, the total product coming from income group q across the n_j European regions could be approximated as

$$\begin{aligned} GDP(q) &= \sum_{j=1}^{n_j} GDP_j(q) \\ &= \sum_{j=1}^{n_j} \phi_j(q) \times rGDP_j. \end{aligned} \tag{6}$$

We proposed that the vulnerability of a given income group to pollution shocks can be measured as the derivative of its product with respect to pollution. That is, we define the vulnerability of income group q as:

$$V(q) \equiv \frac{\partial GDP(q)}{\partial\ pollution} \tag{7}$$

$$= \sum_{j=1}^{n_j} \frac{\partial GDP_j(q)}{\partial\ pollution}. \tag{8}$$

We could approximate the effect of air pollution on the portion of the region j's GDP attributable to the income group q (i.e., $\frac{\partial GDP_j(q)}{\partial\ pollution}$) by

$$\frac{\partial GDP_j(q)}{\partial\ pollution} \approx \phi_j(q) \times \tilde{\tau}_j, \tag{9}$$

where $\tilde{\tau}_j$ accounts for the individualized effect of pollution on regional GDP in region j and could be approximated by our estimates of the CACR of the regional GDP (i.e., the $\tau(X^{het}_{i,t})$ of Equation 5) discussed in the previous section.[8]

Note that j is a geographical reference level, which may or may not match the NUTS3 level (our is) used for the IV estimations proposed in the previous section. As described above, the indicator relies on the spatial variability in $GDP_j(q)$ and $\tilde{\tau}_j$, so we would like a finer spatial resolution. On the other hand, there are data limitations on the extent to which we can provide a reliable measure of $\phi_j(q)$, as we need disaggregated income information at the j-level. We proposed building the indicator at the NUTS2 level (i.e., n_j accounts for the number of NUTS2 regions) and the income quintiles (so that q takes values from 1 to 5.), using microdata from the EU Statistics on Income and Living Conditions (EU-SILC) and the EU Labor Force Survey (EU-LFS) to derive $\phi_j(q)$.[9] In particular, we will make use of EU-SILC and EU-LFS for 2021.[10] On the other side, we will use our results of the CACR for the GDP per capita from above to estimate the $\tilde{\tau}_j$ for each NUTS2 level (i.e., $j \equiv NUTS2$) for 2021, so that $\tilde{\tau}_j = \hat{\tau}(\bar{x}_{j,t=2021}, NUTS2 = j, t = 2021)$, where $\hat{\tau}(\cdot)$ is the function for the predicted effect from the IVforest, $\bar{x}_{j,t=2021}$ correspond to the average (for each of the NUTS2 regions by 2021) of the heterogeneity covariates.

Our results for the estimations of the ACR, the CACR, and the proposed vulnerability indicators are discussed in the next section.

3 Results

This section presents the empirical estimates of the effects of pollution on regional GDP, following the strategy outlined in the previous section. We proceed sequentially. First, Table 1 reports the

[8] Note that, given the definition for $GDP_j(q)$, $\frac{\partial GDP_j(q)}{\partial\ pollution} = \frac{\partial GDP_j(q)}{\partial pollution}\ \phi_j(q) + \frac{\partial \phi_j(q)}{\partial pollution}\ GDP_j(q)$. Therefore, the approximation behind Equation 9 will be closer to equality under the assumption that $\frac{\partial \phi_j(q)}{\partial pollution}$ is either zero or small enough to be insignificant. In practice, this will mean that, given a pollution shock, the percentage change in each group's income will be the same. Therefore, under this assumption, the observed differences among income groups will arise solely from the disproportionate distribution of income groups across Europe. In particular, the indicator will capture whether a given income group is disproportionately located in highly affected regions. We will return to this discussion in the Results section.

[9] Access to the microdata was granted through the Eurostat data request RPP 162/2024-LFS-EU-SILC (the responsibility for all conclusions drawn from the data lies entirely with the authors).

[10] Although we acknowledged there are other years in EU-SILC and EU-LFS, the estimation for multiple years of the $\phi_j(q)$ proportion is out of the scope of this paper. We leave the estimation of multiple years for future work. Details on the construction of the participation share $\phi_j(q)$ are left to the Appendix A.2' data section.

estimates from a naive OLS specification (Column 1) and from a standard 2SLS model that addresses the endogeneity of pollution (Column 2), abstracting from any heterogeneity. These results facilitate a comparison with the existing literature, which primarily estimates the Average Causal Response (ACR). We then present results from the Flexible Instrumental Variables (FIV) approach, which identifies the (local) ACR while allowing for heterogeneity in the first stage (Column 3). All of these estimates use the same set of fixed effects and cluster the standard errors at the NUTS3 level. After presenting some robustness checks, Figure 3 report estimates from our preferred model, the IV Causal Forest, which additionally accommodates and measure heterogeneity in the second stage and provides a Conditional (Local) Average Causal Response (CACR). Finally, we finish with the estimation of the proposed vulnerability indicator by income quintiles.

Table 1: Effect of $PM_{2.5}$ on Regional GDP

	(1) OLS	(2) 2SLS	(3) FIV
Point estimate	-0.00146*	-0.01654**	-0.00316**
	(0.00083)	(0.007123)	(0.001275)
Weak IV − Robust pVal		0.00152	
Relevance test		31.6795	6874.7623
Cluster	NUTS3	NUTS3	NUTS3
FE	NUTS3, Ctry×Year	NUTS3, Ctry×Year	NUTS3, Ctry×Year
Obs.	24,792	24,792	24,792

Notes: The table presents the point estimate of the regression of the logarithm of GDP per capita on $PM_{2.5}$ concentration. Each column correspond to a different estimator. Our preferred specification correspond to the FIV estimator in column (3). The third row presents the p-value for the Anderson-Rubin Wald test under weak IV. The fourth row presents the F-statistic for the instrument's relevance test. Standard errors, clustered at the NUTS3 level, in parentheses. $* \sim 10\%$, $** \sim 5\%$, $*** \sim 1\%$ significance level.

We find a negative and statistically significant effect of pollution on regional GDP across the three specifications. As it has been found in previous implementations, correcting for potential endogeneity with an IV-setting, using the 2SLS estimator, yields larger estimates than the corresponding naive OLS results. In our case, the 2SLS model suggests that a one μg per cubic meter increase in $PM_{2.5}$ is associated with an average decrease in regional GDP of 1.6%. For comparison, Dechezleprêtre et al. (2020b) and Dechezleprêtre et al. (2020a) found point estimates of 0.81 and 1.06 percent, respectively.

Nevertheless, our main interest is to allow the model to account for regional heterogeneity in the first stage. As discussed in previous sections, we are concern that the atmospheric phenomena

we used as instruments (thermal inversion, vertical and horizontal ventilation) may affect $PM_{2.5}$ differently across space, given that their effect depends on the topography, surface energy balance, land-sea interactions, that in contrast are relative invariant within bioregions. Not accounting for such heterogeneity can potentially increase noise in the estimation of the ACR, and the extension of the traditional 2SLS interacting each instrument with bioregions indicators is prompt to bias in finite samples. The Flexible Partially Linear IV (FIV) model provides an alternative by using ML to regularize the instruments, limiting the risk for the many-IV bias, while allowing complex fitting in the first stage (Column 3). The FIV coefficient is negative and statistically significant, albeit of lower magnitude than the 2SLS coefficient. In particular, our estimate suggests that an increase of one $\mu g/\mathrm{m}^3$ is associated with a 0.3% decrease in regional GDP, reducing the point estimate from the one found with a traditional 2SLS. This estimate is higher than the effect predicted by an ordinary OLS estimator.[11]

In order to test the robustness of our results, Table 2 presents results when key components of the FIV specification are changed, such as the functional form of the outcome variable, the pollution endogenous variable, and the fixed-effects specification. In particular, Columns (1) and (3) of Table 2 present the results of changing from logarithms to levels in the specification of the per capita GDP. Columns (2) and (3) present the results when using an alternative measure for $PM_{2.5}$ concentration, coming from the Copernicus Global Air Quality Validated Reanalysis.[12] Finally, column (4) presents the results using our main specification (Table 1, column (3)) but changing the form of the fixed effects. Overall, we find that the key messages of our preferred specification (column 3 of Table 1) remain unchanged under alternative specifications. After accounting for the interpretation of the coefficients under the change in functional form, the results are qualitatively the same, placing the FIV estimator between the OLS and 2SLS estimators.

We would also like to see that our results are not driven by any particular NUTS3 region, country, or year. Therefore, in Figure 2, the first three sets of points identify how our FIV estimator would change when excluding each of these groups of observations one at a time. For example, the first set of points in Figure 2 displays the FIV point estimate (vertical axis) after excluding each of the 1,171 NUTS3 regions, in the sample used to estimate the FIV, one at a time. As a reference,

[11]The Wu-Hausman test ($F_s = 7.01731 \sim F_{1,24789}$, $p - value < 0.01$) rejects the hypothesis of exogeneity of pollution at a 1% significance level.

[12]Retrieved from ADS by December 2024. The data at the NUTS3-yearly level is built by extracting annual averaged for the data's grid and using population weighted averages within NUTS3 regions (details on the data construction can be found in Appendix A.1). Some regions are lost when using Copernicus data, and the time span is different between the two datasets (2003-2022). Our preferred specification make use of the Donkelaar et al. (2021)' data so it iscomparable to the previous estimations for Europe (Dechezleprêtre et al., 2020b,a)

Table 2: Alternative Specification for the FIV Estimator for the Average Causal Response

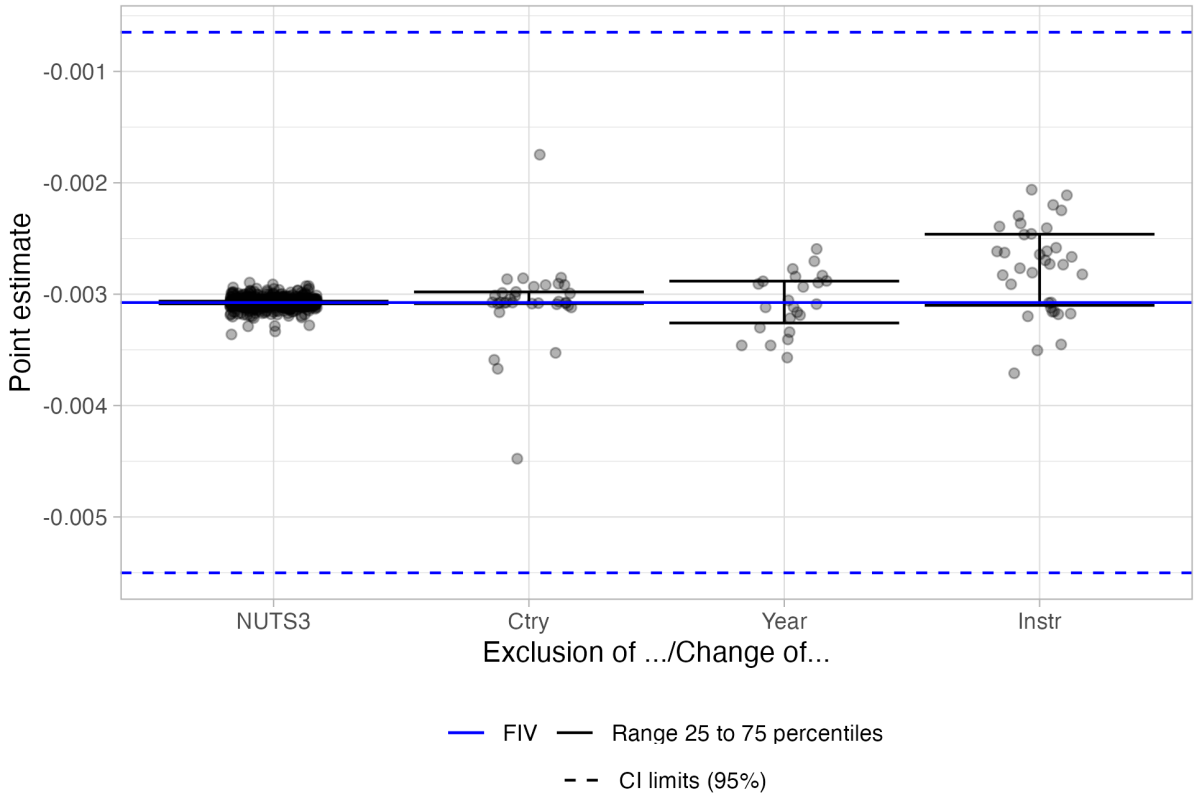
	(1)	(2)	(3)	(4)
Point estimate	-136.877***	-0.00186*	-238.257***	-0.00865***
	(48.516)	(0.001066)	(45.877)	(0.001875)
Mean dep. var.	24,750	9.64	20,155	9.71
Form dep. var.	Levels	Log	Levels	Log
Pollution var.	VD	COP	COP	VD
Cluster	NUTS3	NUTS3	NUTS3	NUTS3
FE	NUTS3, Ctry×Year	NUTS3, Ctry×Year	NUTS3, Ctry×Year	NUTS3, Year
Obs.	24,792	14,546	14,546	23,253

Notes: The dependent variable is GDP per capita in columns (1) and (3), and the logarithm of GDP per capita in columns (2) and (4). Columns (1) and (4) use $PM_{2.5}$ concentration data coming from Washington University in Saint Louis (VD), while columns (2) and (3) use data from Copernicus Global Air Quality Validated Reanalysis (COP). Finally, columns (1) through (3) use NUTS3 and country-year fixed effects, while column (4) uses NUTS3 and year fixed effects. $* \sim 10\%$, $** \sim 5\%$, $*** \sim 1\%$ significance level.

the Figure also reports the point estimate of the FIV of column (3) of Table 1 (solid blue line), and its 95% confidence interval (dashed lines). We see that no particular region is driving the results, as all the points are well concentrated around the solid blue line. Similar conclusions can be drawn from excluding one country at a time (second set of points) or excluding one year at a time (third set). Notice that, although the variability in these cases could be higher as each exclusion implies a bigger and more diverse number of excluded observations, all the point estimates still fall within a reasonable distance of our preferred estimation.

The last set of points shown in Figure 2 takes an extra step by exploring different ways to specify the Instrumental Variables. We re-estimate the FIV, allowing for (1) the exclusion of some of the three selected instruments and (2) various methods of aggregating the instruments annually. For instance, regarding thermal inversion, we consider a setup that counts the number of TIs days separately for summer and winter (as in Dechezleprêtre et al. (2020b)). For the BLH and Ventilation instruments, we explore different annual aggregations, using the mode instead of skewness for BLH, or the 20th percentile instead of the mean for the ventilation coefficient. We found that these alternative approaches do not significantly alter our main results. Although, as expected, changing the set of instruments creates higher variability, all estimates remained within the confidence intervals of our primary point estimator (represented by the blue lines).

Although the first-stage component of the FIV estimator is not as directly interpretable as the first stage of a conventional linear 2SLS specification, it can still be examined to provide



Notes: The graph plots the FIV estimator resulting from either eliminating a set of groups from the sample (columns (1) through (3)) or when using a different set of instruments (column(4)). All the estimations use the same specification as that of our main FIV estimator in column (3) of Table 1. Each point corresponds to a point estimate, the black solid error lines mark the 25th and 75th percentiles of the set of points in each column, and the blue lines correspond to the main FIV estimator (solid line) and its corresponding 95% confidence interval (dashed lines).

Figure 2: Sensitivity of the FIV's Average Causal Response to Observations and Instruments

diagnostic evidence on instrument relevance. The intuition behind the implementation in Ahrens et al. (2024) is closely related to using a flexible prediction of the endogenous variable, $PM2.5_{i,t}$, based on the excluded instruments and exogenous covariates. In our setting, this corresponds to estimating an optimal-instrument-type nuisance component of the form $\tilde{Z}_{i,t} = E[PM2.5_{i,t} \mid Z_{i,t}, W_{i,t}]$, where $Z_{i,t}$ includes the atmospheric instruments and their bioregion interactions, and $W_{i,t}$ includes the exogenous controls, such as temperature and precipitation. The FIV framework allows this component to be estimated flexibly using machine-learning methods. Following Ahrens et al. (2024), this predicted instrument is then orthogonalized with respect to the covariates through an additional nuisance component, $E[\tilde{Z}_{i,t} \mid W_{i,t}]$, yielding a residualized instrument ($Z^*_{i,t} = \tilde{Z}_{i,t} - E[\tilde{Z}_{i,t} \mid W_{i,t}]$) to be used in the IV estimation, as proposed by Ahrens et al. (2024).

The learned instrument should retain the expected relationship with the underlying atmospheric instruments. We therefore examine whether each individual instrument is related in the expected direction to the predicted pollution component, $\tilde{Z}_{i,t}$, and to its residualized counterpart, $Z^*_{i,t}$. To do so, we use a Best Linear Projection (BLP), a common diagnostic in the machine-learning interpretability literature, which linearly projects the prediction onto the original predictors/features. The BLP is intended to provide a low-dimensional descriptive summary of how the learned prediction relates to each variable in the average. Although the coefficients' magnitudes will not be interpretable, their signs indicate whether the learned prediction aligns with the expected direction of the underlying first-stage relationships.

Table 3 presents the BLP results for both $\tilde{Z}_{i,t}$ in Column (1) and $Z^*_{i,t}$ in Column (2). Both regressions use the cross-fitted predictions from the FIV procedure. In addition to the instruments, each regression includes the same set of covariates, $W_{i,t}$, and fixed effects as the main specification (NUTS3 and country-year fixed effects). Standard errors are clustered at the NUTS3 level, and standardized coefficients are reported. The estimated signs are consistent with the expected first-stage relationships discussed in Section 2.

A higher frequency of thermal inversion events and a positive skew in boundary layer height are associated with higher predicted pollution, while higher vertical and horizontal ventilation are associated with lower predicted $PM_{2.5}$ pollution. To assess whether the learned first-stage component contains systematic bioregion-specific variation, in the last row of Table 3 we report the joint-relevance test of the bioregions' interactions in a non-reported regression that includes interactions between the instruments and bioregion indicators. In both cases, the interaction terms are jointly significant, suggesting that allowing for first-stage heterogeneity across bioregions is empirically relevant.

We now present the results of our preferred model which, by coupling the FIV estimator with the IV Causal Forest (IVforest), allows flexibility in the first stage while accounting for possible heterogeneity in the second. The IVforest is an estimator of conditional averages, by estimating local moment equations for the traditional 2SLS estimator. The estimator will provide an estimate for each NUTS3-year value, modeled based on a set of pre-defined heterogeneity covariates X^{het}. The weights of the moment equations vary at the region-year level, so that, in this context, the conditional averages could be interpreted as the individualized causal response of GDP to changes in pollution. Athey et al. (2019) derive the limiting distribution for the point estimates, so that confidence intervals and significance tests could be done over the derived coefficients.

Table 3: Best Linear Projection of the FIV's First Stage Equivalent

	(1) $\tilde{Z}_{i,t}$	(2) $Z^*_{i,t}$
Num. TIs	0.26546***	0.29371***
	(0.164406)	(0.113401)
Ske. BLH	0.17417***	0.2266***
	(0.012164)	(0.00915)
Mean Vent. coeff.	-1.76869***	-2.33213***
	(3.7e-05)	(2.5e-05)
F-stat instruments	121.91	164
FE	NUTS3, Ctry×Year	NUTS3, Ctry×Year
SE Cluster lev.	NUTS3	NUTS3
F-stat of bioregions interactions when included	21.44	24.81

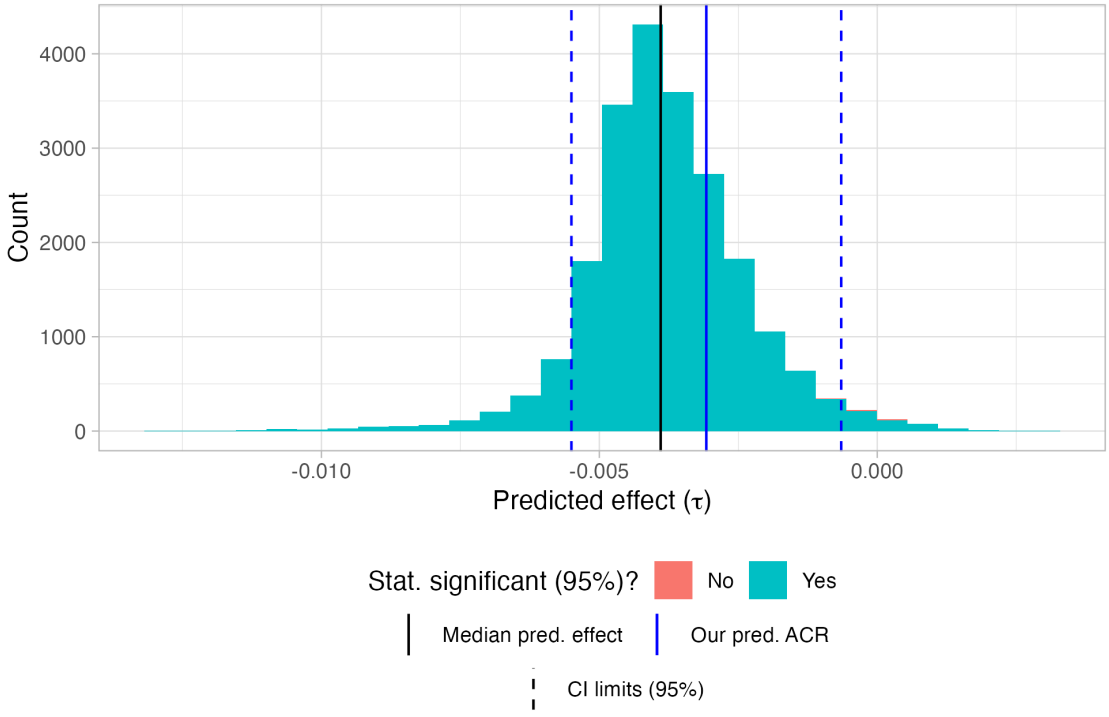
Notes: $\tilde{Z}_{i,t}$ correspond to the predicted pollution given the exogenous instruments (frequency of thermal inversion, annual skewness of the Boundary Layer Height, and ventilation coefficient), and the model covariates (i.e., $W_{i,t}$), while $Z^*_{i,t}$ is its residualized version with respect covariates. The regression in the table correspond then to the Best Linear Projection (BLP) to the instruments, estimates through OLS. We are interested the sign (not the magnitudes) to be consistent on how we expect each individual instrument relates with $PM_{2.5}$ air pollution. Regression includes the IV covariates ($W_{i,t}$) and the fixed effects of our main specification. Standard errors are clustered at the NUTS3 level. Coefficients are standardized. Table also includes the F-stat of the joint relevance of the instruments, and the F-stat of the bio-regions interactions with the instruments in an alternative regression where they are included. $* \sim 10\%$, $** \sim 5\%$, $*** \sim 1\%$ significance level.

As described in Section 2, the set of heterogeneous covariates X^{het} used in the IVforest includes the shares of employment and gross value added in the primary and secondary sectors, population growth, and relative pollution level. In addition, we include year and NUTS2 region indicators in all of our specifications for heterogeneity modeling. All variables are expressed as five-year rolling averages excluding the current year, to prevent contemporaneous endogeneity. Again, the main innovation in adapting the IVforest to this context would be to account for multiple sources of heterogeneity simultaneously, while preventing confounded interpretations of each source. Given that the X^{het} are changing in time and NUTS3 region, the IVforest and its estimation of CACR will provide estimates of the individual effect $\hat{\tau}_{i,t} = \tau(X^{het} = x^{het}_{i,t})$, as represented in Equation 5.

The raw results of the coupling (FIV+IVforest) are presented in Figure 3, showing a histogram for the estimated τs. The vertical black line corresponds to their median, while the blue line corresponds to the estimated ACR (in solid line) and its 95% confidence interval (taken from Table 1, column (3)). The color of the bar indicates the signifcance of the individual τs, using clustered errors at the NUTS3 region level to be consistent with the panel structure of the data.

We found in general significant estimators. Notice first that almost all of the point estimates

lie below zero, suggesting that the negative average derived in the previous section is not just the result of the aggregation. In terms of the general distribution, we see a flat left tail, suggesting asymmetry in impacts, with a few regions-year with significantly higher impacts, reaching almost a negative effect of 10% per unit of $PM_{2.5}$ concentration. Notice that our results show that relying solely on the ACR can be misleading about the heterogeneity underlying it. In this case, assigning each region the average effect could misrepresents its effect by up to a factor of three.



Notes: Density function for the predicted individual (NUTS3-year) effects ($\hat{\tau}_{i,t,}$). The predicted effect significance account for clustered standard errors at the NUTS3 level. Bar's color identifies the statistical significance of the estimator, at a confidence level of 95%. Black vertical line corresponds to the median predicted effect. Blue line correspond to the predicted ACR (solid line), under our preferred specification (column (3), Table 1), and its corresponding confidence interval (CI) limits at a 95% (in dashed lines).

Figure 3: Density Function for Individual Causal Effects (τs) Estimates

Our first interest is to found statistical evidence for heterogeneity, and that the dispersion observed in Figure 3 is not just the result of pure noise. A test for treatment-heterogeneity has been suggested by Chernozhukov et al. (2025), and has become the standard test in the few existing applications and software implementations (e.g., StataCorp (2025)). The test rejects the null hypothesis of homogeneous effects at 1% significance level. This test, in its current implementation, has only been discussed in the context of unconfoundedness. We propose a revised test in Appendix

A.3, but the conclusions are the same, still supporting the existence of heterogeneity.[13]

We now aim to understand the contribution of each of the heterogeneity covariates (X^{het}) to the estimated individualized τ. Notice that the random forest behind the IVforest algorithm allows for a flexible relationship between the individualized treatment effect (τ) and the covariates (X^{het}), likely resulting in complex functional forms. We again follow the limited literature and use the two most common approximations to detect a possible correlation. The first strategy is to use the Best Linear Projection (BLP) of the estimated effects (Golub Capital Social Impact Lab, 2023), that is, the linear projection of the estimated treatment effects onto the heterogeneity covariates via the classic linear regression model, as we did above for the prediction of pollution in the ACR IV-estimation. In particular, we use the regular OLS estimator with the same set of fixed effects as the main specification (i.e., NUTS3 regions and country-year), and errors clustered at the NUTS3 level. As discussed above, while the magnitudes of the coefficients are not fully interpretable, the signs provide useful information about the direction of the average effect of each of the covariates. Table 4 presents the results of the BLP for each of the covariates.

The second strategy relies on constructing Partial Dependence Plots (PDPs). PDPs summarize the relationship between a covariate and the estimated treatment effect by evaluating how the predicted effect changes with changes in the covariate of interest, as the remaining covariates are left constant, usually set to their average. PDPs allow capturing potentially nonlinear relationships to be implied by the underlying machine learning estimator. The PDPs for our estimations are shown in Figure 4. The y-axis indicates the predicted causal response of τ, while the x-axis, the values of the covariates. While the BLP provides a convenient statistical test and interpretable summary of heterogeneity along specified dimensions, PDPs complement this analysis by illustrating the functional form of these relationships and highlighting nonlinear patterns that a linear projection may miss.

Starting with sectoral composition, both the BLP estimates (Table 4) and the partial dependence plots (Figure 4) indicate that a greater dependence on the agricultural sector is associated with larger (more negative) predicted effects of pollution on regional GDP. This suggests that regions with a higher share of agricultural production (at the expense of the services sector, excluded from the specification) are more vulnerable to air pollution shocks. This finding is consistent with

[13]The standard implementation identifies heterogeneity from a regression involving the residualized treatment (D^*). However, in an endogenous-treatment setting, D^* is not expected to be exogenous with respect to the regression error, so the test's identification argument breaks down. We propose replacing with the residualized instrument (Z^*) which is exogenous by the exclusion restriction—thereby restoring valid identification. The discussion and full set of the test results are presented in Appendix A.3.

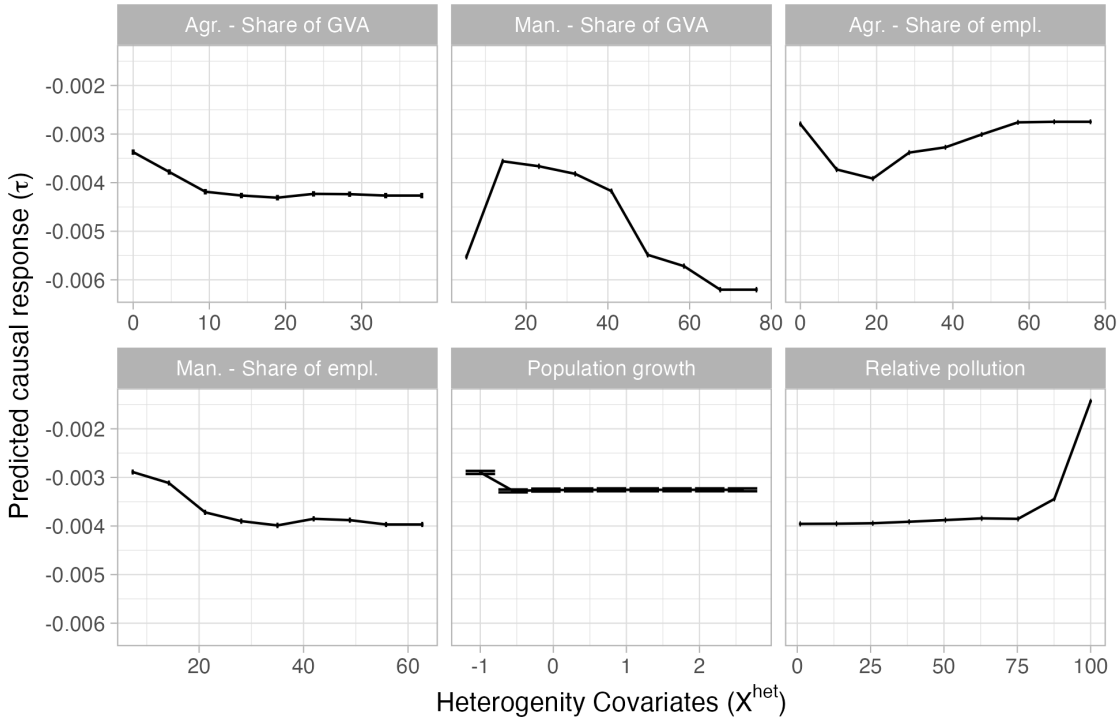
Table 4: Best Linear Projection of the Predicted τ on the Heterogeneity Covariates

	(1)
	$\hat{\tau}$
Agr. - Share of GVA	-5.2e-05***
	(1.549e-05)
Man. - Share of GVA	-4.1e-05***
	(6.53e-06)
Agr. - Share of empl.	4.8e-05***
	(8.61e-06)
Man. - Share of empl.	-6e-06
	(9.35e-06)
Population growth	0.000312
	(0.00070414)
Relative pollution	4e-06**
	(1.89e-06)
F-stat	6.98
FE	NUTS3, Ctry×Year
SE Cluster lev.	NUTS3

Notes: The Best Linear Projection for the ML estimator for $\hat{\tau}_{i,t}$ in the heterogeneity covariates. Regressions estimated by OLS, including fixed effects (NUTS3 regions and country-year) and using clustered standard errors. $* \sim 10\%$, $** \sim 5\%$, $*** \sim 1\%$ significance level.

the literature, which documents that agricultural activity could be particularly sensitive to environmental conditions and pollution exposure (Borgschulte et al., 2022; Hill et al., 2023; Zivin and Neidell, 2012; Chang et al., 2016). A similar pattern emerges for the share of the manufacturing sector. The BLP results indicate that the negative impact of pollution tends to increase with the share of manufacturing in regional production, holding the size of the primary sector constant, and this relationship is also visible in the corresponding partial dependence plot. If anything, the PDP suggests that the manufacturing sector's effect is more important than that of the agricultural sector.

Results for the manufacturing employment suggest a potential mechanism. The PDP indicates that regions with higher employment shares in manufacturing exhibit more negative predicted effects. However, the BLP coefficient for manufacturing employment is not statistically significant, suggesting substantial variability across regions and years. An alternative explanation is that manufacturing activity may indirectly reflect the propagation of pollution-related shocks through supply chains linked to the primary sector, in the same way that Leroutier and Ollivier (2024) suggest that the final sectors are more likely to accumulate the negative consequences.



Notes Partial dependence plot for each of the heterogeneity covariates on the estimation of the individualized treatment effects ($\hat{\tau}(\cdot)$)). For each covariate, we graph how the predicted effect changes with increasing values of that covariate, holding the continuous covariates at their average. Shares of employment, GVA, and population growth are measured in percentage. Relative pollution is measured as the annual percentile across the European pollution distribution. :

Figure 4: Partial Dependence Plots for the Predicted Effect ($\hat{\tau}$)

As for the employment in the primary sector, the positive BLP coefficient suggests that the negative impact of pollution on GDP becomes smaller as agricultural employment increases. The corresponding partial dependence plot displays a similar pattern, initially decreasing but subsequently increasing, suggesting that the relationship may be nonlinear and non-monotonic. One possible interpretation is that regions with larger agricultural labor forces may have greater incentives to recognize and respond to the productivity consequences of pollution, potentially adopting practices that reduce vulnerability.

A related pattern appears when examining relative pollution levels. The positive coefficient of BLP suggests that the negative effects of pollution become less pronounced at higher levels of historical exposure to pollution. However, the PDP highlights that this is true mainly for those regions with high historical concentration of pollutants. In this sense, these results could validate the findings of Borgschulte et al. (2022) and Behrer et al. (2023) on the lack of heterogeneity,

at least in historically clean regions, and, at the same time, support the claim of Dechezleprêtre et al. (2020a) on non-linearities and diminishing marginal damages. In epidemiology, several PM2.5 concentration-response functions imply that marginal health risks are steeper at lower concentrations and flatter at higher concentrations. Gong et al. (2023) find larger marginal mortality impacts at lower PM2.5 concentrations. Additionally, this pattern may reflect adaptation or mitigation efforts in regions that have historically faced higher pollution levels, where the pollution problem is more likely to be recognized for its severity relative to peers.[14] Taken together, these findings suggest that the magnitude of pollution damage may depend not only on exposure and sectoral composition but also on regions' and sectors' capacity to recognize and adapt to environmental risks.

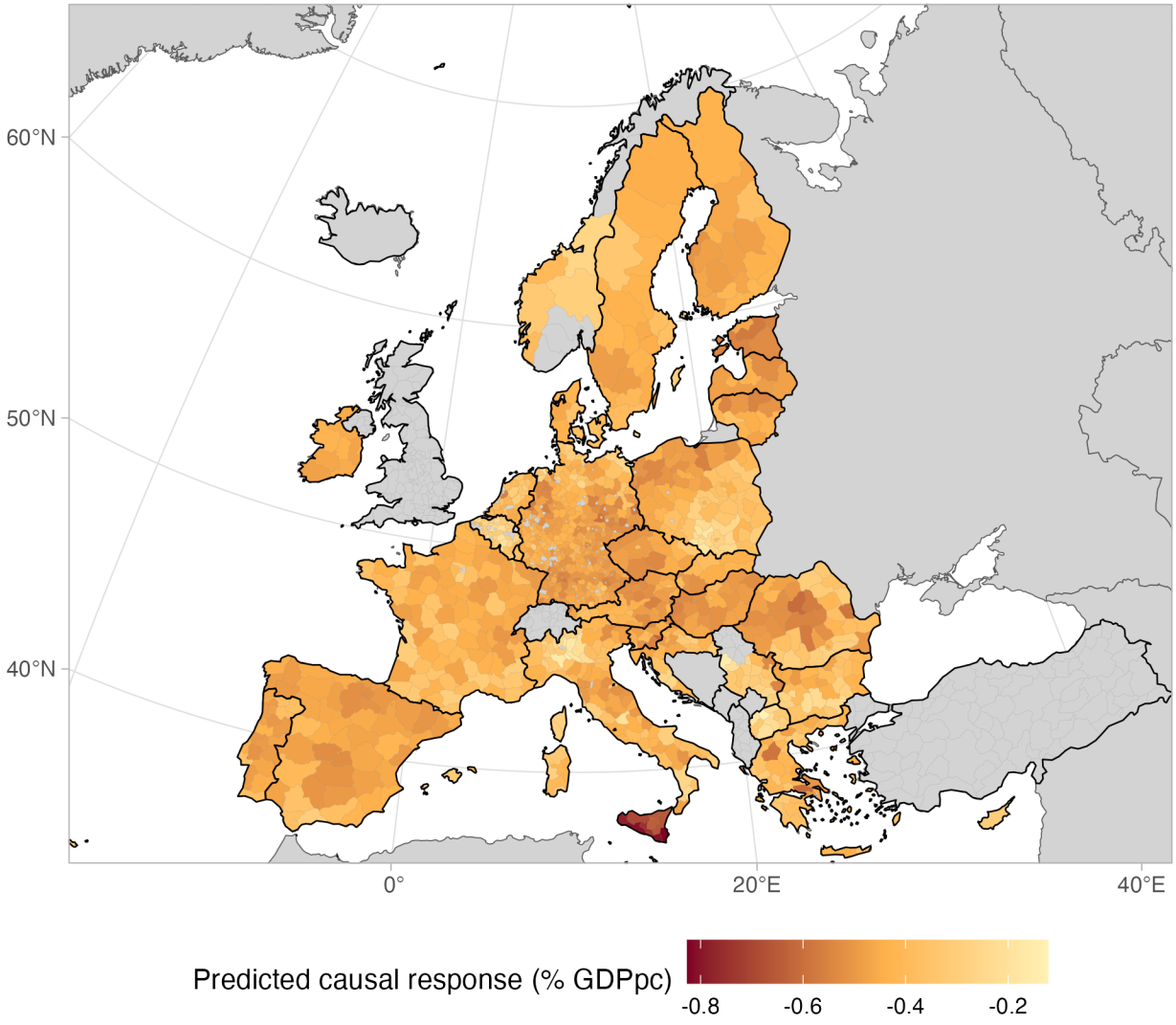
Finally, notice that the BLP (on Table 4) and the PDP (in Figure 4) coincide in suggesting that the rate of population growth does not seem to have a significant participation in the estimated effect.

An additional set of results is obtained by exploring the spatial dispersion of the effects. To do this, we aggregate the NUTS3-year raw coefficients from the IVforest at the NUTS3 level. Our results are shown in the map in Figure 5, where each NUTS3 region is compared against the average effect reported in Column (3) of Figure 1. We can see a spatial concentration of both extremes of the distribution (below and above the median— i.e., more and less negative), with both more likely to occur in the south-central and eastern areas of Europe. There is a high concentration of effects above the median (darker colors) around eastern Germany, the Czech Republic, northern Austria, and in southern Italy. Meanwhile, high concentrations of areas with predicted effects below the median (lighter color) are located in northern Italy, southern Poland, Bulgaria, Serbia, and North Macedonia.

We next examine whether any visual patterns emerge in the relationship between the heterogeneity covariates and the predicted regional causal effects. As discussed above, we predict that each variable relates (*ceteris paribus*) differently to the predicted effect. We see a strong overlap between points above the mean (less negative) of the predicted causal effects in Figure 5 and those with high historical pollution levels (bottom-right panel of Figure 6). This is consistent with the positive value of BLPs presented in Table 4. For example, this is the case for the South West of

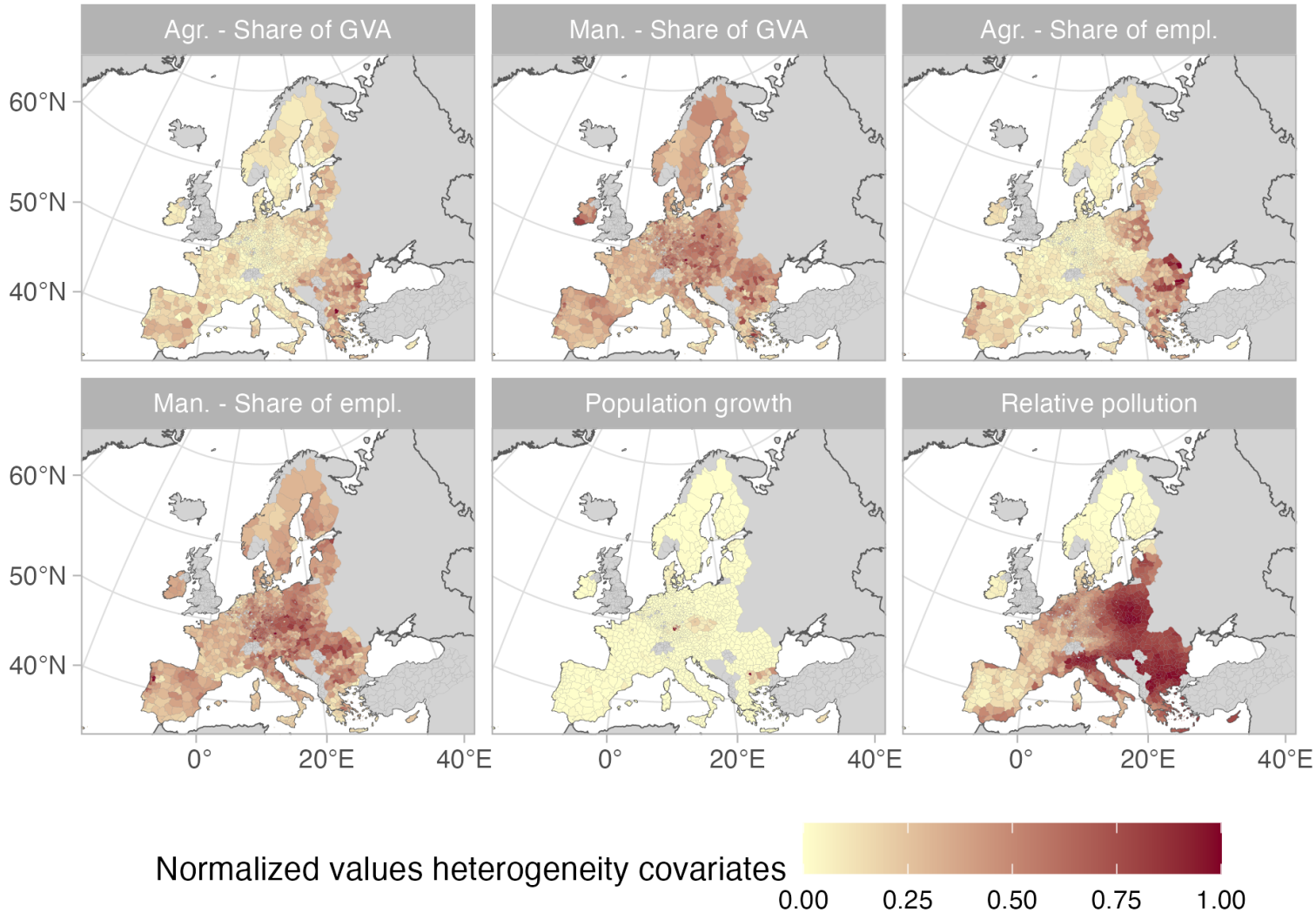
[14]For example, Cheung et al. (2020) find that the mortality effect of air pollution in Hong Kong declined substantially over two decades, and they attribute this mainly to improvements in the healthcare system rather than to lower pollution levels. Similarly, Barwick et al. (2024) have shown that pollution information is a crucial determinant for avoidance behavior and defensive spending.

Italy (the provinces of Lazio and Campania). Meanwhile, areas consistently below the median effect (i.e., more negative) are highly correlated with areas of manufacturing sector predominance. For example, notice the case for South Ireland (Munster province). However, notice that for others, the final effect will be the result of the interaction of two or more sources. This is the case, for example, of Lombardia, which, although it has a significant share of production in the manufacturing sector, is still located among the less affected regions. This is one of the advantages of the proposed estimation of CACR through IVforest: not only do we see how each variable creates heterogeneity, but we can now make this comparison while keeping other possible sources of heterogeneity constant and, in cases of conflicting directions, determine which source is more determinant.



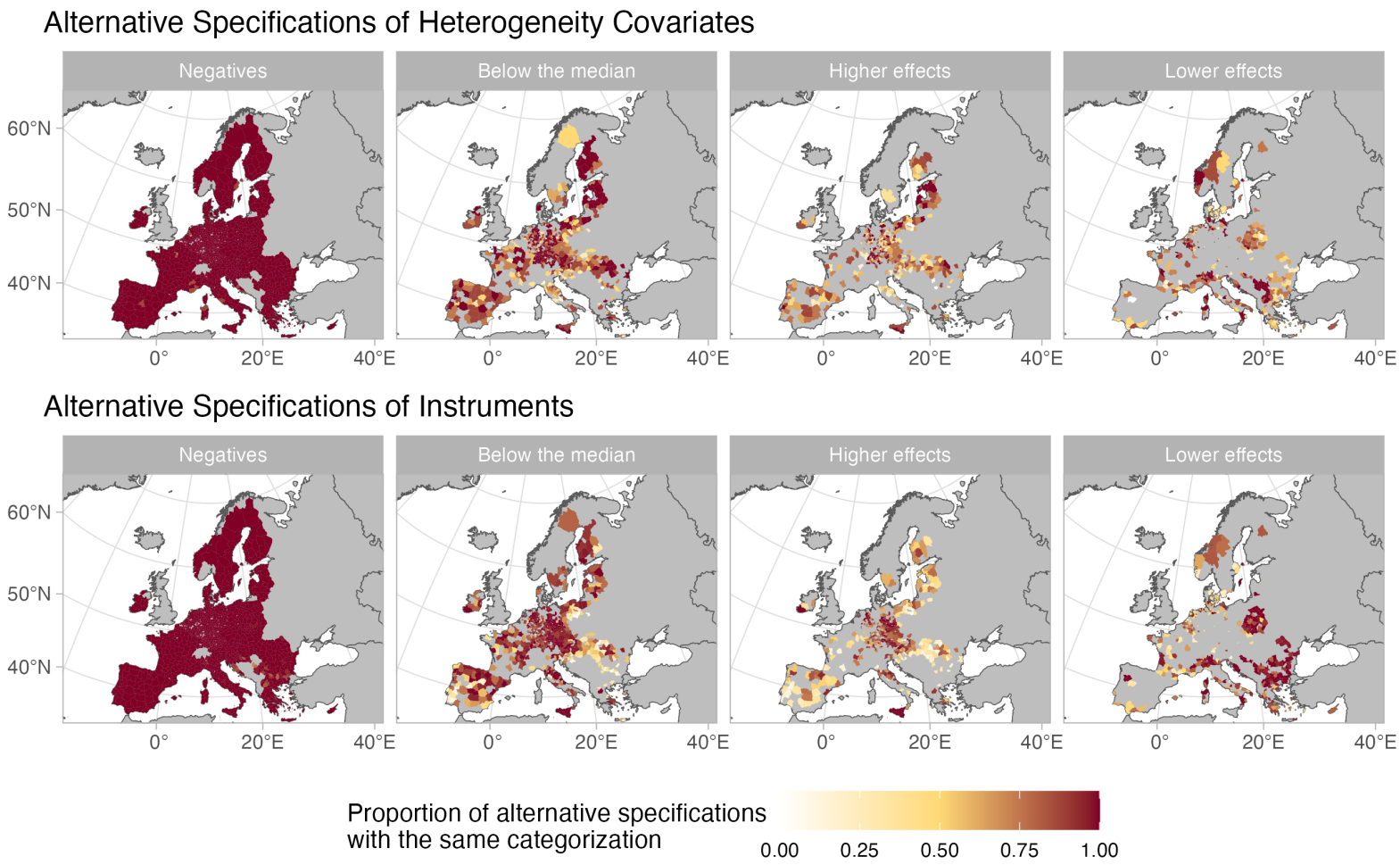
Notes: Predicted causal effect of an increase of 1 $\mu g/m^3$ on the GDP per capita ($GDPpc$) by NUTS3 region. Map elaborated from the prediction of the average causal response from the IVforest, feeding the learner of the effect with the averages of the heterogeneity covariates across time. Areas with predicted (absolute) effects below the median are indicated with a lighter color; areas with effects above the median are indicated with darker colors.

Figure 5: Predicted Causal Response by NUTS3 Regions



Notes: Spatial distribution of the heterogeneity covariates, normalized to the interval $[0, 1]$ (i.e, $(x - min(x))/(max(x) - min(x))$), so that regions with higher average values for a given covariate appear in darker colors.

Figure 6: Spatial Distribution of the Heterogeneity Covariates' Averages



Notes: Proportion of alternative scenarios where the predicted effect of pollution on regional GDP, from left to right, (1) remains a negative value, (2) remains below the median, (3) it remains in the top-third (regions more affected), and (4) it remains in the bottom-third (regions less affected). A darker color in a region indicates that a higher proportion of alternative specifications agree on the classification assigned to that region. Two forms of alternative specifications are considered: first, changing the set of heterogeneity covariates used in the IVforest (top panel); and second, changing the specification of the instrumental variables (bottom panel).

Figure 7: Changes in the Effects Distribution's Attributes Across Alternative Specifications of the IV Causal Forest (IVforest)

To test the consistency of our results, we proceed as before with some robustness checks to changes in the model specification. In particular, note that there are multiple attributes we could evaluate for the CACR estimates, as they are high-dimensional. We will focus, however, on testing whether changing the specification affects key elements of the ordering behind the distribution of the regional CACR in Figure 5. We would argue for robustness if key characteristics of the distribution of our results are preserved under changes in the specification. In particular, we are testing for four attributes of the CACR distribution: (1) does the prediction for each region remains negative under alternative specifications, (2) do the regions predicted to have an effect below the median (i.e., more affected) are still predicted to be so under the new specification, (3) are the regions classified into the most-affected tercile still being classified as such in the other specifications, and (4) the same as (3) but with the least affected tercile.

We evaluate two types of changes in the specification. First, we assess how the ordering of the CACR (and the four particular attributes described above) changes under alternative specifications of the heterogeneity covariates. We evaluate eight alternative specifications for X^{het}: six resulting from taking one covariate at a time, and two from adding two possible alternative covariates (the proportion of employment from small firms, and the risk of poverty index) one at a time. [15]. The second type of changes corresponds to changes in the set of Instrumental Variables used in our model, in the same way as described in the robustness of the ACR in Figure 2 from above.

The results of the robustness test proposed above are shown in Figure 7. The maps show the proportion of the alternative specification in which the distribution's corresponding attributes match those in our main specification. The scale could therefore be interpreted as the level of classification stability. Regarding the effect's negativity, we found that all alternative specifications also predict negative estimates of τ in almost all NUTS3 regions. For the remaining distribution attributes, although we start to see disagreement among the models, a significant portion of them will predict the same attribute for their own estimates. Furthermore, the conclusion for key regions remain the same. For example, the classification of regions in Germany, the Czech Republic, northern Austria, and in southern Italy as highly affected regions can be replicated by several of the alternative models, as shown in the third column of graph in Figure 7. Similarly, it also holds for the characterization of regions in northern Italy, southern Poland, Bulgaria, Serbia, and

[15]The proportion of employment from small firms, and the risk of poverty index, are averaged over the five previous years, excluding the current period, in the same way as the rest of the heterogeneity covariates. Both covariates were also extracted from Eurostat, retrieved in July 2025 from here and here. Although we consider both of them likely variables for describing a region's vulnerability to pollution and its ability to adapt, they do not have full spatial coverage and therefore were not included in our main specification of the IVforest.

North Macedonia as less affected among other NUTS3 regions, as shown in the fourth column. We therefore found that our main results and conclusions will be similar even under alternative specifications.

Distributional Economic Effects of Air Pollution Across Income Groups

As discussed in Section 2, we are also interested in understanding possible distributional consequences of pollution. Given the vulnerability measure $V(q)$ for income group q, as described in Equation 7, we define a normalization of the vulnerability as:

$$\tilde{V}(q) = \frac{V(q)}{GDP(q)} \tag{10}$$

This normalization allows us to estimate the effect of pollution as a proportion of the production of each income group across Europe ($GDP(q)$) and make them comparable between income groups. As the proposed estimator corresponds to a series of non-linear transformations of our estimates of $\tau_{i,t}$, we approximate the statistical properties of $\tilde{V}(q)$ using empirical p-values obtained by bootstrapping the IVforest estimation at the NUTS3 level, under our main specification.

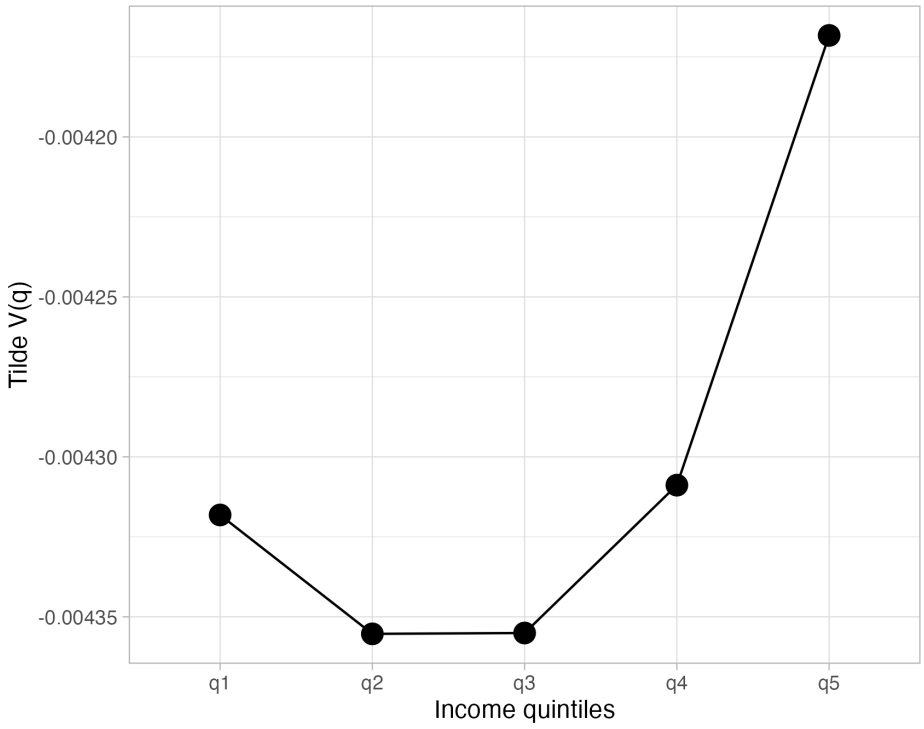
Figure 8 presents the point estimates across income groups, while the empirical p-values for a set of alternative hypotheses of interest are shown in Table 5. The estimates indicate that air pollution has negative effects on economic activity across all income groups. However, we find a non-monotonic relationship in the distribution of damages. Specifically, the proposed measure of vulnerability across income groups shows a U-shaped pattern, with middle-income groups experiencing the largest losses, while higher-income groups are the least affected.

Table 5: Empirical p-values for Comparison of the Vulnerability Indicator among Income Quintiles

Ha	$q1 < q5$	$q2 < q5$	$q3 < q5$	$q2 < q1$
p-value	0.18	0.09	0.07	0.95

Notes: .

This pattern is consistent with a productivity-based mechanism. Higher-income groups are more likely to work in indoor environments, benefit from better health protection, and have greater capacity to adapt to pollution exposure, thereby mitigating its economic impact. At the other end of the distribution, the lowest-income quintile does not bear the largest losses. This may reflect their relatively smaller contribution to aggregate GDP and concentration in lower-productivity or more constrained economic activities, which limits the marginal impact of pollution on measured



Notes: Normalized vulnerability indicator (Equation 10) for each of the income quintiles, using the results for the regional effect of pollution from the IVforest from previous section.

Figure 8: Normalized Vulnerability Indicator by Income Quintile

output.

In contrast, mid-income quintiles (q2-q3) seem to have the largest negative effects. These groups constitute a core share of the active labor force and are more likely to be employed in sectors that are both pollution-exposed and less able to fully offset productivity losses. These differences are also supported by the empirical p-values. Notice that we can reject the null hypothesis of q5 being more affected than q2 and q3, with a significance level of 10%.

Importantly, our approach should be interpreted with a limitation. The distributional indicators we construct reflect the allocation of aggregate regional damages across income groups, based on their share in the economic activity. As such, they do not capture individual-level causal effects of pollution exposure, nor do they account for within-region heterogeneity in exposure, behavior, or adaptive capacity. Instead, they provide a consistent and policy-relevant approximation of how aggregate economic losses are distributed across groups.

4 Conclusions

This paper revisits the estimation of pollution damage functions at the regional level by explicitly allowing for heterogeneity in both the first and second stages of an instrumental-variables framework. Using a rich panel of European NUTS3 regions and high-resolution measures of $PM_{2.5}$, we show that standard homogeneous IV specifications mask substantial variation in how pollution affects regional economic activity. By combining flexible machine-learning-based first-stage estimation with an instrumental causal forest approach for the second stage, we obtain localized causal estimates that remain robust to functional-form misspecification and weak-instrument concerns. Our estimates suggest that higher levels of particulate matter have a statistically significant negative effect on regional GDP, although the magnitude is smaller than that implied by conventional linear IV models.

Beyond the average effect, our results reveal a clear and systematic pattern of heterogeneity across regions. The economic costs of air pollution are larger in areas with greater dependence on the agricultural and manufacturing sectors. Conversely, the predicted effects are smaller in context of greater exposure, such as regions with historically higher levels of pollution or larger dependence on labor in agriculture. This evidence may result from the combination of vulnerability to pollution and strong adaptation efforts in these regions. Our main contribution is to show that, when accounting for heterogeneity in the IV estimates of the effect of $PM_{2.5}$ pollution on regional economic aggregates, there is not only potential to improve estimates for the average effect, but also the opportunity to uncover substantial variation.

Importantly, the heterogeneity we uncover is not merely a statistical artifact; it has direct policy relevance. It identifies the regions that are most vulnerable to air-quality shocks and highlighting where environmental regulation is likely to yield the greatest economic gains, or is most needed to avoid aggravating existing inequalities. Furthermore, this variation has a policy value by allowing us to identify potential drivers of heterogeneity, identify regions with higher vulnerability, and develop new indicators that yield insights into the unequal burden of air pollution. Uniform policy approaches would overlook substantial variation in local gains, pointing to the value of targeted interventions. The non-monotonic distribution of damages across income groups indicates that the economic costs of pollution are not concentrated solely among the poorest, but instead fall heavily on middle-income populations that sustain economic activity. This pattern is reinforces the view that air quality improvements can generate economic returns trough productivity gains.

Our framework offers a tractable and scalable way to integrate distributional considerations into empirical and modeling exercises. In particular, the combination of heterogeneous causal estimates with group-level economic shares can be readily incorporated into Integrated Assessment Models (IAMs) and other ex-ante policy tools, allowing for a nuanced evaluation of the distributional consequences of environmental policies. This sets the stage for policy analysis that explicitly accounts for how the costs and benefits of pollution mitigation are distributed across populations.

References

Abadie, A., J. Gu, and S. Shen (2024, 3). Instrumental variable estimation with first-stage heterogeneity. *Journal of Econometrics 240*.

Adhvaryu, A., N. Kala, and A. Nyshadham (2019). Management and shocks to worker productivity.

Aguilar-Gomez, S., H. Dwyer, J. G. Zivin, and M. Neidell (2022). This is air: The "nonhealth" effects of air pollution.

Ahrens, A., V. Chernozhukov, C. Hansen, D. Kozbur, M. Schaffer, and T. Wiemann (2025, 4). An introduction to double/debiased machine learning.

Ahrens, A., C. B. Hansen, M. E. Schaffer, and T. Wiemann (2024, 3). ddml: Double/debiased machine learning in stata. *The Stata Journal: Promoting Communications on Statistics and Stata 24*.

Angrist, J. D., K. Graddy, and G. W. Imbens (2000, 7). The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. *The Review of Economic Studies 67*, 499–527.

Angrist, J. D. and G. W. Imbens (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity.

Aragón, F. M., J. J. Miranda, and P. Oliva (2017, 11). Particulate matter and labor supply: The role of caregiving and non-linearities. *Journal of Environmental Economics and Management 86*, 295–309.

Archsmith, J., A. Heyes, and S. Saberian (2018, 10). Air quality and error quantity: Pollution and performance in a high-skilled, quality-focused occupation. *Journal of the Association of Environmental and Resource Economists 5*, 827–863.

Athey, S., K. Inoue, and Y. Tsugawa (2025). Targeted treatment assignment using data from randomized experiments with noncompliance.

Athey, S., J. Tibshirani, and S. Wager (2019, 4). Generalized random forests. *https://doi.org/10.1214/18-AOS1709 47*, 1148–1178.

Avila-Uribe, A. (2023). The effect of air pollution on us aggregate production.

Avnery, S., D. L. Mauzerall, J. Liu, and L. W. Horowitz (2011, 4). Global crop yield reductions due to surface ozone exposure: 1. year 2000 crop production losses and economic damage. *Atmospheric Environment 45*, 2284–2296.

Barwick, P. J., S. Li, L. Lin, and E. Y. Zou (2024, 5). From fog to smog: The value of pollution information†. *American Economic Review 114*, 1338–1381.

Basu, D. (2024, 10). Frisch–waugh–lovell theorem-type results for the k-class and 2sgmm estimators. *Statistics & Probability Letters 213*, 110188.

Batheja, D., S. Hirshleifer, J. T. Mullins, and E. Org (2023). Uc berkeley cega working papers title more than particulates matter: Multiple pollutants and productivity in indian call centers publication date.

Behrer, A. P., R. Choudhary, and D. Sharma (2023). Air pollution reduces economic activity: Evidence from india.

Borgschulte, M., D. Molitor, and E. Zou (2022). Air pollution and the labor market: Evidence from wildfire smoke.

Bracht, F. and D. Verhoeven (2022). Air pollution and innovation.

Chang, T., J. G. Zivin, T. Gross, and M. Neidell (2016, 8). Particulate pollution and the productivity of pear packers. *American Economic Journal: Economic Policy 8*, 141–169.

Chang, T. Y., J. G. Zivin, T. Gross, and M. Neidell (2019). The effect of pollution on worker productivity: Evidence from call center workers in china. *American Economic Journal: Applied Economics 11*, 151–172.

Chen, H., S. Liu, X. Liu, and J. Wang (2023). Local air pollution and corporate innovation. *Applied Economics Letters*.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018, 2). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal 21*, C1–C68.

Chernozhukov, V., M. Demirer, E. Duflo, and I. Fernández-Val (2025). Fisher–schultz lecture: Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. *Econometrica 93*, 1121–1164.

Cheung, C. W., G. He, and Y. Pan (2020). Mitigating the air pollution effect? the remarkable decline in the pollution-mortality relationship in hong kong. *Journal of Environmental Economics and Management 101*, 102316.

Dechezleprêtre, A., N. Rivers, and B. Stadler (2020a, 11). The economic cost of air pollution: Evidence from europe.

Dechezleprêtre, A., N. Rivers, and B. Stadler (2020b, 8). The economic cost of air pollution: Evidence from europe. Technical report.

Dong, R., R. Fisman, Y. Wang, and N. Xu (2021, 3). Air pollution, affect, and forecasting bias: Evidence from chinese financial analysts. *Journal of Financial Economics 139*, 971–984.

Donkelaar, A. V., M. S. Hammer, L. Bindle, M. Brauer, J. R. Brook, M. J. Garay, N. C. Hsu, O. V. Kalashnikova, R. A. Kahn, C. Lee, R. C. Levy, A. Lyapustin, A. M. Sayer, and R. V. Martin (2021, 11). Monthly global estimates of fine particulate matter and their uncertainty. *Environmental Science and Technology 55*, 15287–15300.

European Environmental Agency, E. (2018). Unequal exposure and unequal impacts: Social vulnerability to air pollution, noise and extreme temperature in europe. Technical report, European Environment Agency.

Fan, M. and C. Grainger (2023, 9). The impact of air pollution on labor supply in china. *Sustainability (Switzerland) 15*.

Fu, S., V. B. Viard, and P. Zhang (2021, 11). Air pollution and manufacturing firm productivity: Nationwide estimates for china. *Economic Journal 131*, 3241–3273.

Gan, T., H. Yang, and W. Liang (2021, 2). How do urban haze pollution and economic development affect each other? empirical evidence from 287 chinese cities during 2000–2016. *Sustainable Cities and Society 65*.

Golub Capital Social Impact Lab (2023, 5). Machine learning-based causal inference tutorial.

Gong, Y., S. Li, N. J. Sanders, and G. Shi (2023). The mortality impact of fine particulate matter in china: Evidence from trade shocks. *Journal of Environmental Economics and Management 117*, 102759.

Hanna, R. and P. Oliva (2015, 2). The effect of pollution on labor supply: Evidence from a natural experiment in mexico city. *Journal of Public Economics 122*, 68–79.

Health Effects Institute (2025). State of global air 2025: A report on air pollution and its role in the world's leading causes of death. Technical report, Health Effects Institute.

Hill, A. E., J. Burkhardt, J. Bayham, K. O'Dell, B. Ford, E. V. Fischer, and J. R. Pierce (2023). Air pollution, weather, and agricultural worker productivity. *American Journal of Agricultural Economics*.

Holub, F., L. Hospido, U. J. Wagner, J. Baarck, L. C. Fernandez, M. Römmele, J. Salinas, A. Santonja, and S. Werner (2016). We thank conference participants at aere.

Kim, M. J. (2024, 11). Air pollution and solar photovoltaic power generation: Evidence from south korea. *Energy Economics 139*, 107924.

Leroutier, M. and H. E. Ollivier (2024). The cost of air pollution for workers and firms.

Li, S., A. D. Leakey, C. A. Moller, C. M. Montes, E. J. Sacks, D. K. Lee, and E. A. Ainsworth (2023, 11). Similar photosynthetic but different yield responses of c3 and c4 crops to elevated o3. *Proceedings of the National Academy of Sciences of the United States of America 120*, e2313591120.

Li, X., F. Wagner, W. Peng, J. Yang, and D. L. Mauzerall (2017, 11). Reduction of solar photovoltaic resources due to air pollution in china. *Proceedings of the National Academy of Sciences of the United States of America 114*, 11867–11872.

Lichter, A., N. Pestel, and E. Sommer (2017, 10). Productivity effects of air pollution: Evidence from professional soccer. *Labour Economics 48*, 54–66.

Lin, J., H. Wan, and Y. Yu (2024, 2). What you breathe makes you poor: The effect of air pollution on income. *China Economic Review 83*.

Manono, B. O., F. K. Sadiq, A. A. Sadiq, T. A. Matsika, and F. Tanko (2025, 9). Impacts of air quality on global crop yields and food security: An integrative review and future outlook. *Air 2025, Vol. 3, Page 24 3*, 24.

Montt, G. (2018, 12). Too polluted to work? the gendered correlates of air pollution on hours worked. *IZA Journal of Labor Economics 7*.

Rai, P. K. (2016, 7). Impacts of particulate matter pollution on plants: Implications for environmental biomonitoring. *Ecotoxicology and Environmental Safety 129*, 120–136.

Rentschler, J. and N. Leonova (2023, 7). Global air pollution exposure and poverty. *Nature Communications 2023 14:1 14*, 4432–.

Sarmiento, L. (2022, 1). Air pollution and the productivity of high-skill labor: evidence from court hearings. *Scandinavian Journal of Economics 124*, 301–332.

Sarmiento, L. and A. Nowakowski (2023, 11). Court decisions and air pollution: Evidence from ten million penal cases in india. *Environmental and Resource Economics 86*, 605–644.

StataCorp (2025). *cate postestimation - Postestimation tools for cate*.

Tan, Z. and L. Yan (2021, 11). Does air pollution impede corporate innovation? *International Review of Economics and Finance 76*, 937–951.

Wager, S. and S. Athey (2018, 7). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association 113*, 1228–1242.

Wang, C., Q. Lin, and Y. Qiu (2022, 3). Productivity loss amid invisible pollution. *Journal of Environmental Economics and Management 112*.

Wu, Z., C. Song, Y. Yang, and X. Qie (2025, 4). The impact of surface ozone and particulate matter 2.5 on rice yield in china: An econometric approach. *Sustainability 2025, Vol. 17, Page 3357 17*, 3357.

Yuval, Y. Levi, U. Dayan, I. Levy, and D. M. Broday (2020, 1). On the association between characteristics of the atmospheric boundary layer and air pollution concentrations. *Atmospheric Research 231*.

Zivin, J. G. and M. Neidell (2012, 12). The impact of pollution on worker productivity. *American Economic Review 102*, 3652–3673.

Zivin, J. G. and M. Neidell (2018, 1). Air pollution's hidden impacts. *Science 359*, 39–40.

Appendix

A.1 Data

A.1.1 NUTS3-Year Panel data

This section describes the data used to estimate the effect of $PM_{2.5}$ on regional GDP and its heterogeneity. The analysis is built on a NUTS3-year panel covering the outcome variable (regional GDP per capita), the treatment variable ($PM_{2.5}$ pollution), IV covariates, and heterogeneity covariates. We focus on the European continental area (approximately longitude $-25°$ to $45°$, latitude $30°$ to $72°$) from 1998 to 2022, although effective coverage depends on data availability across variables. The data structure follows, as closely as possible, the strategy adopted by previous studies for Europe — in particular Dechezleprêtre et al. (2020a) and Bracht and Verhoeven (2022) — so that our estimates remain comparable with the existing literature. We adopt the NUTS3 classification under the 2021 revision, extracted from the Eurostat GISCO portal. The data sources are divided into three broad categories: socio-economic data, pollution data, and weather data.

The main source of socio-economic data is Eurostat, from which we retrieve key variables at the NUTS3 level by year. The set of extracted variables includes regional GDP ($nama_10r_3gdp$), regional GVA ($nama_10r_3gva$) and employment ($nama_10r_3empers$) by sector, and regional population ($nama_10r_3popgdp$), among others.[16] Since Eurostat already reports these data at the NUTS3 level, no major spatial aggregation is required. For specific analyses, we also draw on NUTS2-level data, including GDP ($nama_10r_2gdp$) and risk-of-poverty indicators (ilc_li41), assigned to the corresponding NUTS3 regions in the relevant year. All data are downloaded using the `eurostat` R package.

Our main pollution indicator is $PM_{2.5}$ concentration.[17] Following the approach of previous studies for Europe, we rely primarily on data from the Atmospheric Composition Analysis Group (ACAG) at Washington University in St. Louis (WUSTL), based on Donkelaar et al. (2021). This dataset combines satellite-, simulation-, and monitor-based sources to produce annual average $PM_{2.5}$ concentrations with global coverage; although we restrict it to the European area of in-

[16] Other variables were downloaded but only used marginally, for example for robustness exercises or alternative specifications (e.g., data on employment by size of the enterprise - bd_size_r3) or unit standardisations (e.g., prices deflator - prc_hicp_aind)

[17] As stated in the main document, $PM_{2.5}$ is one of the most commonly used indicators of ambient air pollution in the literature, given that its health consequences are especially well established and severe. In addition, it is positively correlated with other pollutants, as they share emission sources and are jointly shaped by atmospheric dispersion conditions (Bracht and Verhoeven, 2022; Dechezleprêtre et al., 2020a). Also, $PM_{2.5}$ is usually the pollutant measured with the broadest coverage.

terest. We use version V5.GL.04, retrieved in July 2024, which provides data at a $0.01° \times 0.01°$ spatial resolution; for the temporal dimension, we use the authors' pre-processed annual-level data. To aggregate to the NUTS3 level, we combine the pollution grid with gridded population data from the Global Human Settlement Layer (GHSL), produced by the Joint Research Center (JRC, European Commission) under release 2023A.[18] Since GHSL data are produced every five years, we use the nearest past quinquennial benchmark for each analysis year. The NUTS3-level pollution value for a given year is then a weighted average across all grid cells overlapping the region, with each cell weighted by its population and the share of the region's area it contributes. For robustness exercises, we additionally extract $PM_{2.5}$ data from Copernicus: specifically, the CAMS Global Reanalysis EAC4 dataset from the Copernicus Atmosphere Data Store.[19] This product is available at 3-hourly temporal resolution from January 2003, at a $0.75° \times 0.75°$ spatial resolution; we restrict extraction to the European continental area.[20] The 3-hourly observations are averaged to the annual level at the grid and then aggregated to the NUTS3 region using the same population-weighted procedure applied to the WUSTL data.

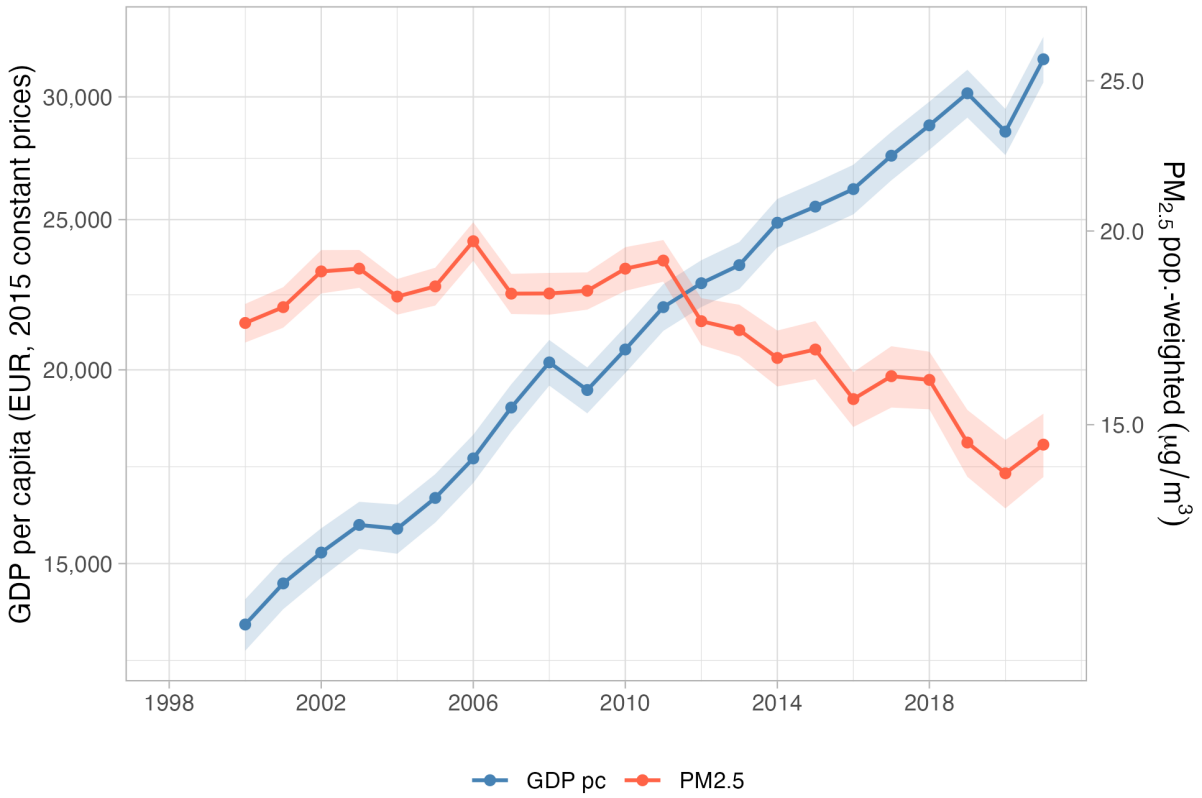
Figures A1 and A2 present snapshots of the behavior of the two main variables of interest.

Figure A1 tracks the evolution of the European average of regional GDP per capita and population-weighted $PM_{2.5}$ over the sample period. The data reflect the well-known trends in each variable: rising GDP per capita alongside declining population-weighted pollution. As discussed in the literature, the simultaneity of the two variables prevents a causal interpretation of the observed negative correlation, which persists even after accounting for weather covariates and time- and region-specific fixed effects. In fact, both previous literature and our results suggest that the causal effect of pollution reduction is even larger than the Within estimator implies. On the other side, Figure A2 presents the spatial distribution of both variables for three years of the sample (2011, 2016, and 2021), expressed as within-year percentile ranks. The relative position of each region across the distribution is broadly stable over time. High GDP per capita is concentrated in north-central European countries, while eastern European regions consistently record lower income and production per capita. In the case of pollution, high concentrations are found in eastern and southern areas, while northern countries exhibit the lowest levels.

[18] The combination is done using the `raster::resample` function in R for resampling the population grid of the GHSL into the grid of the WUSTL data. GHSL was retrieved on July 2024 from here.

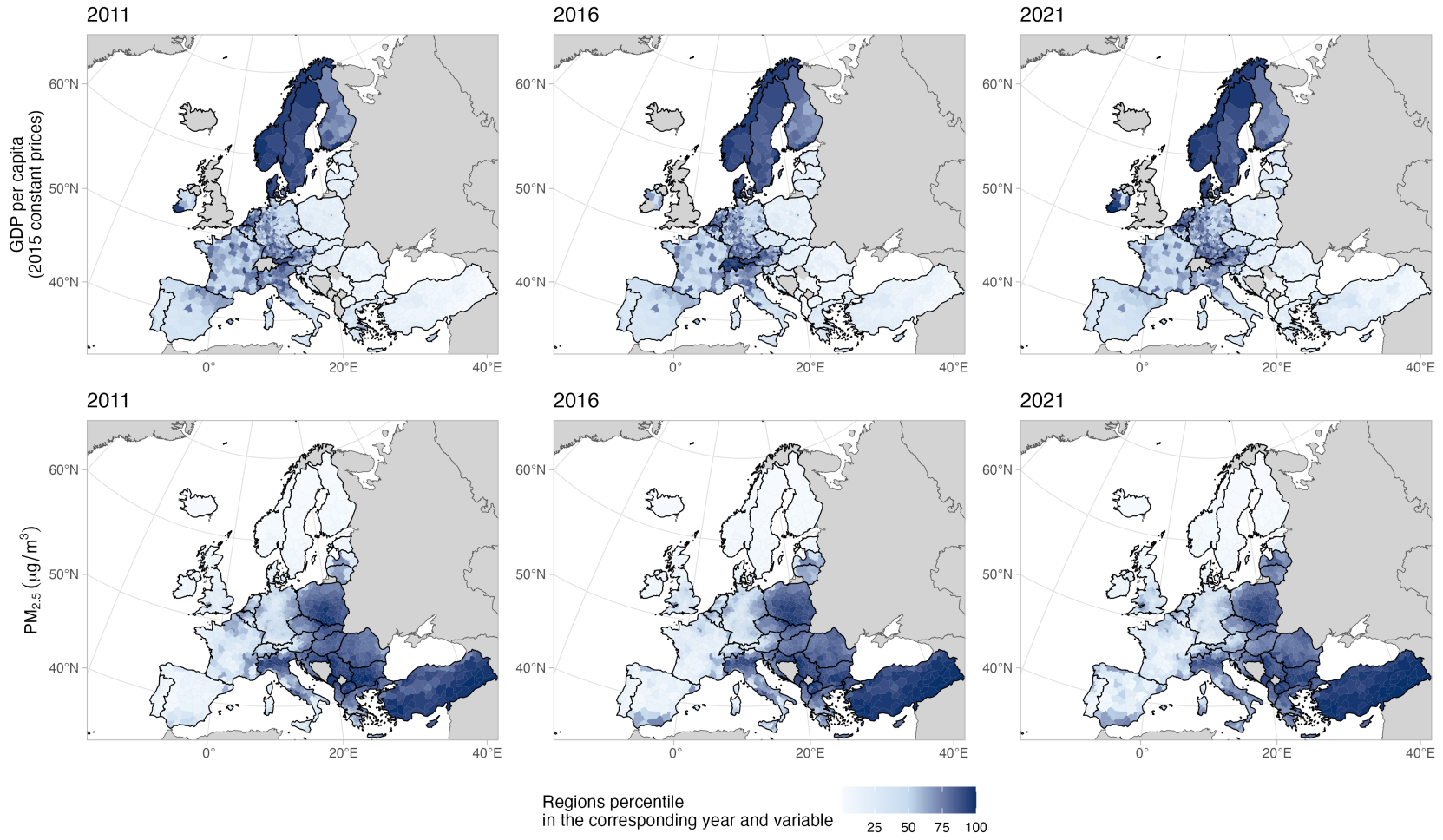
[19] Variable *particulate_matter_2.5um* retrieved by December 2024, using the ECMWFR API.

[20] Although the CAMS European air quality reanalyses data, from the ADS, has a higher spatial resolution (0.1° vs 0.75°), it has an even smaller time frame than the Global Copernicus (starting only from 2013), and therefore not as comparable with Donkelaar et al. (2021).



Notes: Cross-sectional means across all NUTS3 regions in the sample, by year, for regional GDP (left axis) and $PM_{2.5}$ (right axis). $PM_{2.5}$ is averaged using regional population (pop) as weights, so the series reflects the pollution exposure. Axes use a logarithmic scale.

Figure A1: GDP per capita and population-weighted PM2.5, European average



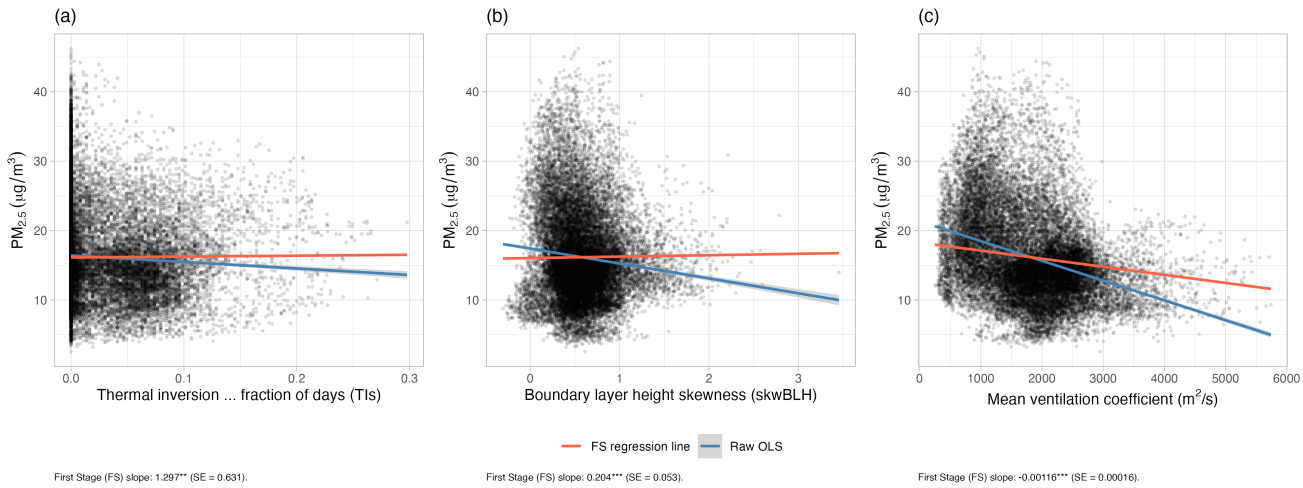
Notes: Spatial concentration of regional GDP and $PM_{2.5}$ concentration across Europe for 2011, 2016 and 2021. The top row shows GDP per capita percentile rank; the bottom row shows PM2.5 concentration percentile rank. The scale represents the percentile for each region in its corresponding year, in the distribution of the corresponding variable.

Figure A2: GDP per capita and PM2.5 concentration across NUTS3 regions, 2011, 2016, 2021

The weather data come from two ERA5 datasets from the Copernicus Climate Data Store, accessed via the ECMWFR API. Both share a raw spatial resolution of $0.25° \times 0.25°$ and hourly temporal resolution. From the ERA5 reanalysis of hourly data on single levels, we extract 2-metre temperature ($2m_temperature$), boundary layer height (BLH, $boundary_layer_height$), mean sea level pressure ($mean_sea_level_pressure$), total precipitation ($total_precipitation$), and wind speed — derived from the eastward ($10m_u_component_of_wind$) and northward ($10m_v_component_of_wind$) components at 10 metres. Aggregation proceeds in three steps: hourly observations are first averaged — or, for precipitation, summed — to the daily level within each grid cell; each grid cell is then assigned to a NUTS3 region, whose value is the population-weighted average of the overlapping cells; finally, daily NUTS3 values are aggregated to the annual level, either as the proportion of days falling within pre-defined bins (twenty bins for temperature and precipitation; thirteen bins for wind speed, based on the Beaufort scale) or as a simple annual average (for sea level pressure). A similar procedure is applied to the ERA5 reanalysis of hourly data on pressure levels, from which we extract relative humidity ($relative_humidity$) at 1000 hPa and temperature at 1000 hPa and 975 hPa. Relative humidity is aggregated following the same procedure as mean sea level pressure. The pressure-level temperature series is aggregated to the daily NUTS3 level using the same approach as surface temperature, but is retained only at that intermediate stage to construct the daily thermal inversion indicators described below.

Two sets of variables require a deeper description. The first concerns the construction of the instrumental variables. As described in the main text, these exploit different atmospheric phenomena as sources of exogenous variation in $PM_{2.5}$ concentration that are unrelated to regional GDP.

Three families of instruments are used. The first is based on the characterization of thermal inversions (TI), following Dechezleprêtre et al. (2020b). At the daily NUTS3 level, a day is classified as experiencing a thermal inversion event if the daily average temperature at 975 hPa exceeds the daily average at 1000 hPa. Two annual aggregations of this indicator are constructed: the share of days with a TI event over the full year, and separate shares for summer (April 15^{th} to October 14^{th}) and winter days. Our main specification uses the annual share; the expected positive association with $PM_{2.5}$ is confirmed by the first-stage regression line in panel (a) of Figure A3. The second family is derived from the boundary layer height (BLH). After aggregation to the NUTS3-daily level, BLH is further summarised at the annual level in two forms: its annual skewness and its annual mode. Our preferred specification uses the annual BLH skewness: higher skewness reflects a greater frequency of low-BLH episodes over the year and, therefore, higher pollution concentrations. The



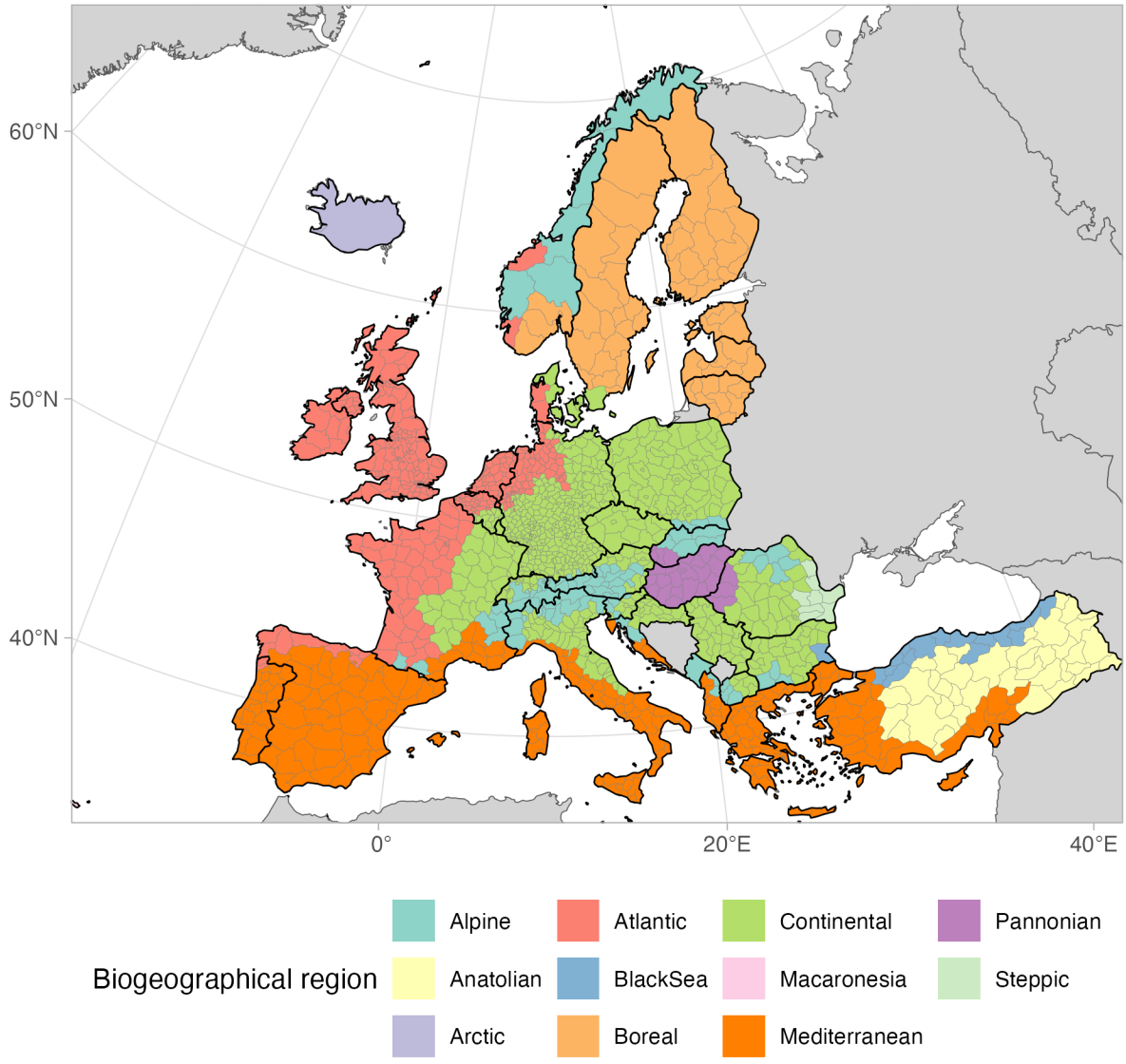
Notes: Each panel shows the raw scatter of the instrument against population-weighted $PM_{2.5}$ on the instrument-specific complete-cases sample. The blue line corresponds to the raw OLS with 95% CI. The red line (FS regression line) corresponds to the regression of $PM_{2.5}$ on the corresponding instrument, accounting for the same set of controls and fixed effects used in the 2SLS first-stage (FS) estimation. Panel captions report the corresponding slope for the FS with significance stars ($* \sim 10\%$, $** \sim 5\%$, $*** \sim 1\%$), using NUTS3 clustered standard errors. Notice that the regression line (and the significance of the reported coefficients) confirms the expected sign for each instrument (positive for thermal inversion and BLH skewness, and negative for the mean ventilation coefficient).

Figure A3: Instrument–PM2.5 relationships: Share of thermal inversion events (a), BLH annual skewness (b), Annual mean Ventilation Coefficient (c)

positive relationship is confirmed by the regression line in panel (b) of Figure A3. The third family corresponds to the Ventilation Coefficient (VC), proposed by Bracht and Verhoeven (2022), which measures the speed of pollutant dispersion by jointly accounting for vertical dispersion through BLH and horizontal dispersion through wind speed. The daily ventilation coefficient is defined as the product of average BLH and average wind speed. Two annual aggregations are constructed: a simple average and the 20th percentile, following Bracht and Verhoeven (2022). Our preferred specification uses the mean annual VC. As higher dispersion speed reduces $PM_{2.5}$ concentration, the expected relationship is negative, as confirmed by the regression slope in panel (c) of Figure A3.

The second set of variables consists of the biogeographical region indicators, required to allow heterogeneity in the first stage of the IV estimation. We use the European Environment Agency (EEA) biogeographical regions classification, which divides the continent into eleven bioregions: Alpine, Anatolian, Arctic, Atlantic, Black Sea, Boreal, Continental, Macaronesia, Mediterranean, Pannonian, and Steppic. Each NUTS3 region is assigned to the bioregion in which it falls; when a NUTS3 region spans more than one bioregion, it is assigned to the one covering the larger area.

Figure A4 presents the spatial distribution of these regions across Europe.



Notes: European Environment Agency (EEA) biogeographical regions.

Figure A4: EEA biogeographical regions

A.1.2 Distributional Indicator Inputs

The construction of distributional pollution-damage indicators requires, alongside the Conditional Average Causal Response (CACR) estimates derived from the Flexible IV and the IV Causal Forest, a measure of the share of aggregate household income in each region j attributable to households in income group q—hereafter $\phi_j(q)$. This section proposes and evaluates two strategies

for estimating these parameters from Eurostat microdata.[21] Both alternatives are constructed at the NUTS2 spatial resolution (so that j denotes a NUTS2 region), defined over income quintiles ($q \in \{1, 2, 3, 4, 5\}$), and calibrated for the year 2021.[22] The two alternatives draw on the Survey on Income and Living Conditions (EU-SILC) and the Labor Force Survey (EU-LFS), whose respective strengths and limitations motivate the hybrid approach described below.

The population parameter $\Phi_j(q)$ is defined as the share of total income in region j attributable to households in income group q:

$$\Phi_j(q) \equiv \frac{\sum_{i|r(i)=j,g(i)=q}^{N_j} y_i}{\sum_{i|r(i)=j}^{N_j} y_i}, \tag{A1}$$

where $r(\cdot)$ assigns each household to its region, $g(\cdot)$ maps it to its income group, N_j is the total number of units in region j, and y is the income measure of interest. By construction, $\sum_q \Phi_j(q) = 1$ for all regions j:

$$\sum_q \Phi_j(q) = 1 \text{ for all region } j \tag{A2}$$

so $\Phi_j(q)$ constitutes a proper income-share distribution across quintiles. Because collecting data for every household in a region is infeasible, the working parameter $\phi_j(q)$ approximates $\Phi_j(q)$ using a survey sample $s(j)$ of cardinality n_j:

$$\phi_j(q) \sim \phi_j^s(q) \equiv \frac{\sum_{i|r(i)=j,g(i)=q,i\in s(j)}^{n_j} y_i}{\sum_{i|r(i)=j,i\in s(j)}^{n_j} y_i}. \tag{A3}$$

The estimator ϕ_j^s is a valid approximation for $\Phi_j(q)$ under appropriate assumptions on the sample design, but its construction requires observing, for each unit i, the region $r(\cdot)$, the income group $g(\cdot)$, and a meaningful income measure y_i, data requirements that no single Eurostat survey fully satisfies, under a scientific access, at the NUTS2 level for all European countries.

We argue that reasonable estimates for $\phi_j(q)$ can be obtained at the NUTS2 level using Eurostat microdata. The ideal choice would be the NUTS3 resolution, to match the spatial granularity of the CACR estimates; however, even with scientific access to the microdata through Eurostat, individual's location at the NUTS3 level is not observable (i.e., $r(\cdot)$ is not identifiable below NUTS2),

[21] Access to the microdata was granted through the Eurostat data request RPP 162/2024-LFS-EU-SILC (the responsibility for all conclusions drawn from the data lies entirely with the authors).

[22] We decided to use 2021, given that, at the moment of the data request, it was the latest year for which surveys were available for most of the countries. Extraction from other years, before and after 2021, is out of the scope of this paper and is left for future work.

ruling out direct computation of ϕ_j^s at that scale. Working at the NUTS2 level is a workable alternative: the European area comprises approximately 200 regions at this resolution, which still affords substantial spatial variation for the distributional analysis. Two complementary surveys are combined for the estimation, since neither one alone provides all the inputs required for a clean computation of $\Phi_j(q)$ across all of Europe. The EU-SILC (European Union Statistics on Income and Living Conditions) is a harmonized annual survey coordinated by Eurostat that collects comparable micro-level data on household income, poverty, social exclusion, and living conditions across EU member states and several associated countries. It is the primary source of sub-national household income data in Europe and therefore the preferred basis for estimating income shares; however, the default spatial resolution is NUTS1, with NUTS2-level identifiers available only for a subset of countries. The EU-LFS (European Union Labor Force Survey) is a large-scale harmonized household survey that collects comparable micro-level data on the labor market situation of individuals, covering employment status, occupation, hours worked, educational attainment, and demographic characteristics. It provides NUTS2-level regional identifiers by default, offering finer spatial resolution than EU-SILC, but does not directly observe household income; it does, however, include an income decile variable (`INCDECIL`) for household members reported as employees.

For a subset of countries — Spain (ES), Portugal (PT), and France (FR) — EU-SILC reports income data at the NUTS2 level, providing all the inputs required to compute the sample estimator ϕ_j^s directly. We take advantage of this and propose two ways to approximate ϕ_j^s for the remaining European countries. The two proposed strategies differ in how they handle the remaining countries: the first exploits NUTS1-level SILC data; the second combines NUTS2-level LFS data with aggregate income information. We use the NUTS2 sample estimators ϕ_j^s for ES, PT, and FR as the ground truth against which both strategies are evaluated. Because any valid estimate of $\Phi_j(q)$ must satisfy the simplex constraint in Equation A2, we adopt an error measure that respects this structure. In particular, we use the Kullback-Leibler Divergence Score (KL score) to quantify the discrepancy between an estimated distribution $\hat{\phi}_j(q)$ and the reference $\phi_j^s(q)$:

$$D_{KL}\left(\phi_j^s \,\|\, \hat{\phi}_j\right) = \sum_q \phi_j^s(q)\, \ln\left(\frac{\phi_j^s(q)}{\hat{\phi}_j(q)}\right). \tag{A4}$$

As the score approaches zero, $\hat{\phi}_j$ grows closer to the reference distribution ϕ_j^s. Results are reported in Table A1. The two proposed estimation methods are described in turn below.

Method 1 — NUTS1 Extrapolation (ϕ^{Agg}). The first and simplest strategy extrapolates the

NUTS1 sample estimator $\phi^s_{k(j)}$ to every NUTS2 region within the same NUTS1 parent. Letting $k(\cdot)$ denote the function that maps each NUTS2 region j to its NUTS1 parent, the extrapolated estimator is

$$\phi^{Agg}_j(q) = \phi^s_{k(j)}(q). \tag{A5}$$

The NUTS1-level reference value $\phi^s_{k(j)}(q)$ is computed from EU-SILC data, exploiting the fact that at least the NUTS1 location is observable for every household. Using weighted quantiles, we estimate the quintile cutoffs for equivalized disposable income within each NUTS1 region and assign each household to one of the five income groups $q \in \{1, 2, 3, 4, 5\}$. The resulting income-share estimator at the NUTS1 level is

$$\phi^s_k(q) \equiv \frac{\sum^{n_j}_{i|k(r(i))=k,\, g(i)=q} yH_i}{\sum^{n_j}_{i|k(r(i))=k} yH_i}, \tag{A6}$$

where yH corresponds to total household income as reported by Eurostat.

Method 2 — LFS-based Hybrid Estimator (ϕ^{o1}). The second method combines individual-level data from EU-LFS with aggregate income data from EU-SILC. Letting $population(j)$ denote region j's population, $\bar{y}_{c(j)}(q)$ the average total household income in quintile q at the country level (estimated from EU-SILC), and $\hat{\psi}_i(q)$ the estimated probability that household i belongs to income group q, the estimator is defined as

$$\phi^{o1}_j(q) = \frac{\sum_{i|r(i)=j} \hat{\psi}_i(q) \cdot population(j) \cdot \bar{y}_{c(j)}(q)}{\sum_q \sum_{i|r(i)=j} \hat{\psi}_i(q) \cdot population(j) \cdot \bar{y}_{c(j)}(q)} \tag{A7}$$

If $\hat{\psi}_i(q)$ is a consistent estimate of the true group-membership probability, the numerator of Equation A7 approximates the total income in region j attributable to income group q households, and the denominator normalizes to total household income in the region. Regional population data are taken from Eurostat.[23] The national average income by quintile, $\bar{y}_{c(j)}(q)$, is derived from EU-SILC in the same way as the quintile cutoffs used for $\phi^{Agg}_j(q)$.

Although EU-LFS includes an income decile variable (`INCDECIL`), we opt for a probability-based assignment rather than a hard classification in order to mitigate noise in small regions with few observed households per income category: a continuous probability vector borrows statistical strength across quintiles and produces smoother regional estimates. The probabilities $\hat{\psi}_i(q)$ are

[23] Via `eurostat` package in `R`, variable *tgs00096*.

estimated with a Neural Network Classifier (NNC) trained separately for each country on the LFS microdata.[24] Features of the NNC include region, urbanization of the place of residence, household composition, labor market participation, hours of work, age structure, household education level, economic sector of employed household members, and citizenship.

Both strategies are estimated for all countries covered by Eurostat microdata. We evaluate their performance against the sample estimator ϕ_j^s for the three countries for which NUTS2-level EU-SILC data are available (ES, PT, and FR). Table A1 reports the KL score for each strategy relative to ϕ_j^s. The first strategy, ϕ^{Agg}, achieves the lowest KL scores in every validation country and overall, indicating that the NUTS1 extrapolation from EU-SILC provides a closer approximation to the ground-truth income shares than the LFS-based hybrid estimator.

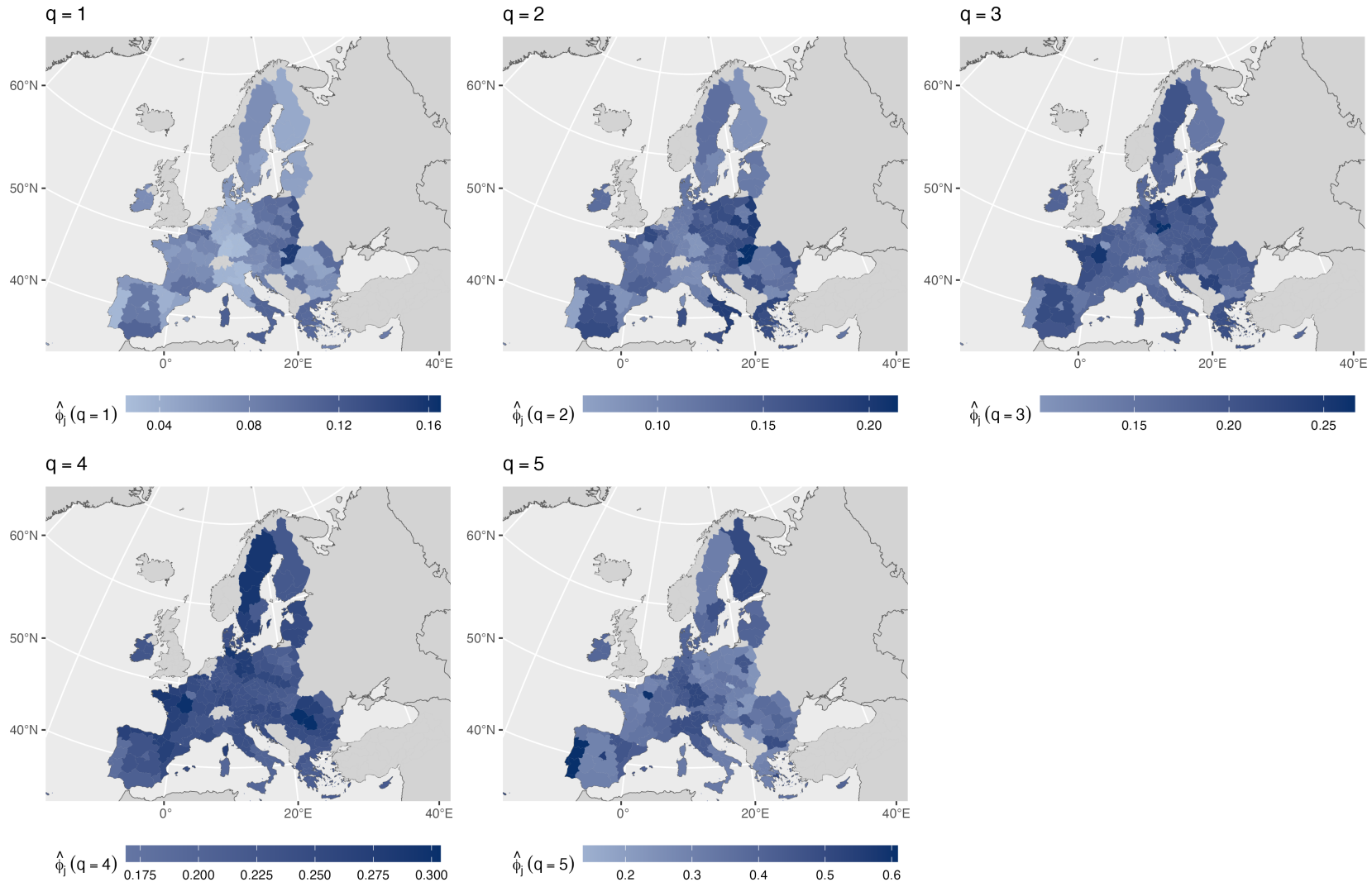
Table A1: Kullback-Leibler Divergence Score (KL score)

	ϕ^{Agg}	ϕ^{o1}
ES	0.027	0.0486
FR	0.0108	0.0309
PT	0.0124	0.0571
Overall	0.0175	0.0417

Notes: Kullback-Leibler Divergence Score (KL score) of the two proposed methods for estimating Φ at the NUTS2 level, for countries for which, using EU-SILC, the sample estimator was possible to estimate at this level (ES - Spain, PT - Portugal, and FR - France). The KL score is a measure of dispersion created by an incoming probability distribution $\hat{\phi}$ and a-priori distribution (ϕ). As the score goes to zero, the incoming probability is more similar to the a-priori.

On the basis of the KL-score comparison, we adopt ϕ^{Agg} as the preferred estimator of the income-share parameters for the full set of European countries. The resulting values are presented in Figure A5 and serve as the inputs for constructing the distributional indicators of pollution damages. From this point forward, the ϕ^{Agg} values are treated as fixed parameters in the distributional analysis; accounting for the uncertainty associated with their estimation is left for future work.

[24] NNC architecture: two hidden layers (64 $\rightarrow$ 32 units), ReLU activations, 30% dropout, softmax output (5 classes = quintiles).



Notes: Final set of values used as the proportion of the income in each NUTS2 region coming from a given income quintile $q \in \{1, 2, 3, 4, 5\}$.

Figure A5: Proportion of Income Coming from Income Quintile q by NUTS2 Region (Φ estimates)

A.2 FIV, IVforest and its coupling

In this appendix we further complement the discussion behind our empirical strategy and the adaptation of the Flexible IV (FIV) and the IV Causal Forest (IVForest) for the estimation

The FIV is implemented using Two Stage Least Square (2SLS) with the outcome ($Y^*_{i,t}$), endogenous ($D^*_{i,t}$) and instrumental ($Z^*_{i,t}$) variables transformed to accommodate the partially linear structure of the model (Ahrens et al., 2024). The transformations for the outcome and endogenous variables are computed using the original variables minus their prediction using the covariates:

$$Y^*_{i,t} = LogGDP_{i,t} - E[LogGDP_{i,t}|X_{i,t}] \tag{B8}$$

$$D^*_{i,t} = Pollution_{i,t} - E[\hat{D}_{i,t}|X_{i,t}] \tag{B9}$$

The transformation of the instrument is as follows:

$$Z^*_{i,t} = \hat{D}_{i,t} - E[\hat{D}_{i,t}|X_{i,t}]. \tag{B10}$$

$$\hat{D}_{i,t} = E[Pollution_{i,t}|Z_{i,t}, X_{i,t}] \tag{B11}$$

It corresponds to the predicted pollution when using the instruments and the covariates (Equation B11), minus the prediction using the covariates.

As suggested by (Chernozhukov et al., 2018) and (Ahrens et al., 2025), the multiple predictions could be done by virtually any ML algorithm, and the estimator would retain good properties as long as they are done using K-folding out-of-sample prediction in order to prevent overfitting.

Notice that the estimation of the FIV as proposed by Chernozhukov et al. (2018) and Ahrens et al. (2024) is somehow the combination of using ML in the first stage with a standard orthogonalization with respect to covariates, such as the one suggested by Basu (2024). In this case, the first stage will be to estimate $\hat{D}$ using ML to characterize the relationship between the instruments and the covariates. Note that this approach is well-suited to the set of instrumental variables and the specification proposed above. The Machine Learning algorithm will allow for flexibility in the effect of the instruments, while the Neyman Orthogonality Condition, behind the FIV and the DDML framework, will limit the effect of measurement error on the algorithms (i.e., the regularization bias (Ahrens et al., 2025)), and the out-of-sample cross-fitting will prevent overfitting bias. Finally, to account for regional and annual-country fixed effects, we residualized all variables (outcome,

endogenous, instruments, and covariates) before implementation. This will force the estimation to account for the fixed effects, as in a traditional linear setting, rather than having the ML algorithm do so.

The predictions or nuisance components for the FIV (i.e., $E[LogGDP_{i,t}|X_{i,t}]$, $E[Pollution_{i,t}|Z_{i,t}, X_{i,t}]$, and $E[\hat{D}_{i,t}|X_{i,t}]$) are done using an ensemble estimator that weights the prediction using elastic net regressions, random forest and gradient boosting. In each case, we use five folds (i.e., $K = 5$).

Our proposal is to nest the FIV with the IV Causal Forest so that the later received the transformed version of the outcome, endogenous and instrument variables (i.e., Y^*, D^*, and Z^*, respectively) as the 2SLS does in the implementation proposed by Ahrens et al. (2024). Our intention is to provide the IV Causal forest with a stronger instrument, resulting from a first stage that allows for regional heterogeneity and non-linear relationships. In this sense, the coupling makes IV Causal Forest less vulnerable to weak instruments.

A.3 Testing Heterogeneity of the CATE Estimates

Testing for heterogeneity of treatment effects has become a standard diagnostic in applied work using Conditional Average Treatment Effect models such as those described in the main document. The existing literature on formal tests, however, has focused almost exclusively on settings where identification rests on (conditional) unconfoundedness; no established guidance exists for the IV Causal Forest (IVforest) setting. This appendix explains why the standard test loses its identification argument when the treatment is endogenous, and proposes an IV-adapted implementation. For the application to air pollution and regional GDP, the conclusion of the test is unaffected by the choice of implementation: both the standard and the adapted versions reject the null hypothesis of homogeneous effects.

For clarity of exposition, the setup is presented for the cross-sectional case. Let Y_i be an outcome, D_i a treatment variable, and W_i a vector of covariates under the additive model

$$Y_i = \tau \cdot D_i + g(W_i) + e_i \tag{C12}$$

Under the exogeneity condition $E[e_i \mid W_i] = 0$, taking the conditional expectation of Equation C12 and subtracting yields the partialled-out representation

$$E\left[Y_i \mid W_i\right] = E\left[D_i \mid W_i\right] + g(W_i) \tag{C13}$$

$$Y_i - E\left[Y_i \mid W_i\right] = \tau \cdot \left(D_i - E\left[D_i \mid W_i\right]\right) + e_i \tag{C14}$$

Double Debiased Machine Learning (DDML) exploits the fact that the moment function underlying Equation C14 satisfies Neyman orthogonality with respect to the nuisance functions $E\left[Y_i \mid W_i\right]$ and $E\left[D_i \mid W_i\right]$. Defining $h(W_i)$ as a predictor for $E\left[Y_i \mid W_i\right]$ and $f(W_i)$ as a predictor for $E\left[D_i \mid W_i\right]$, the parameter τ is consistently estimated from the residualized regression

$$\begin{aligned} Y_i - h(W_i) &= \tau \cdot \left(D_i - f(W_i)\right) + \epsilon_i \\ Y_i^* &= \tau D_i^* + \epsilon_i, \end{aligned} \tag{C15}$$

where $Y_i^* \equiv Y_i - h(W_i)$ and $D_i^* \equiv D_i - f(W_i)$, even under prediction error in $f(\cdot)$ and $h(\cdot)$, provided those predictions remain consistent.

When the treatment effect is heterogeneous and takes the form $\tau(X_i)$, the same identification argument applies, provided $E\left[e_i \mid W_i, X_i\right] = 0$. Equation C15 then extends to

$$Y_i^* = \tau(X_i) D_i^* + \epsilon_i, \tag{C16}$$

where $B_i^* \equiv B_i - g(W_i, X_i)$ and $g(W_i, X_i)$ is a predictor for $E\left[B_i \mid W_i, X_i\right]$ for any variable B.

he heterogeneity test proposed by Chernozhukov et al. (2025), as described by StataCorp (2025), is based on parameterising the true heterogeneous effect as a linear combination of the mean effect and the deviation of the Causal Forest estimate from that mean:

$$\tau(X_i) = \gamma_1 \bar{\tau} + \gamma_2(\hat{\tau}(X_i) - \bar{\tau}) + \varepsilon_i, \tag{C17}$$

where $\hat{\tau}(X_i)$ is the Causal Forest estimate of $\tau(X_i)$ and $\bar{\tau} \equiv E[\hat{\tau}(X_i)]$. Substituting Equation C17 into Equation C16 and replacing $\bar{\tau}$ with the sample average $\breve{\tau}$ yields the regression model from which γ_1 and γ_2 can be identified:

$$Y_i^* = \gamma_1 \breve{\tau} D_i^* + \gamma_2(\hat{\tau}(X_i) - \breve{\tau}) D_i^* + u_i \tag{C18}$$

Under this parameterisation, $\gamma_2 = 0$ implies homogeneous treatment effects, while $\gamma_2 \neq 0$ signals heterogeneity. Since γ_2 is identified from the regression in Equation C18, the null hypothesis $H_0 : \gamma_2 = 0$ is directly testable. Column (1) of Table C2 reports the estimates of this regression for our application.

Table C2: Heterogeneity Test of the Predicted Effects of $PM_{2.5}$ on Regional GDP

	(1) $\tilde{D}$	(2) $\tilde{Z}$	(3) $\alpha \cdot \tilde{Z}$
γ_1	0.4799***	3.03397***	1.02132***
	(0.104918)	(0.67157)	(0.226088)
γ_2	1.0644***	4.99578***	1.6849***
	(0.236477)	(1.364832)	(0.459504)
Wald stat	4.29	8.168	8.196
Wald pVal	0.0384	0.0043	0.0042

Notes: Table presents the results of the regression in Equation C2 using the residualized treatment D^* (column (1)) or the residualized instrument Z^* in columns (2) and (3). All the specifications suggest the significance of the γ_2, suggesting that there is, in fact, heterogeneity in the predicted effect of $PM_{2.5}$ on the regional GDP. $* \sim 10\%$, $** \sim 5\%$, $*** \sim 1\%$ significance level.

The identification of γ_1 and γ_2 from Equation C18 requires D_i^* to be exogenous with respect to the error u_i, a condition that fails when the treatment is endogenous. Under the instrumental variables design — and in particular under the FIV model described in the main document — the first stage implies $D_i^* = \alpha Z_i^* + v_i$, where Z_i^* is the residualized instrument. Because Z_i^* satisfies the independence assumption by construction, replacing D_i^* with Z_i^* (or αZ_i^*) in Equation C18 restores identification of the γ coefficients. Columns (2) and (3) of Table C2 report the results of this IV-adapted implementation. In both cases, γ_2 remains statistically significant, confirming that the conclusion of heterogeneous effects is robust to the choice of identification strategy.