The Mortalility and Medical Costs of Air Pollution - Review

Author: Daniel Saggau — daniel.saggau@campus.lmu.de
Supervisor: Nadzeya Laurentsyeva

15/03/2021

1 Introduction

This research paper examines the impact of pollution exposure on in mortality, health care and medical costs. Specifically, this paper focuses on heterogeneity in age, suggesting that a small subset of the elderly population is most vulnerable to pollution. Firstly, Deryugina et al. (2019) use an instrumental variable approach. They instruments for air pollution by leveraging changes in wind direction to estimate the causal impact of fine particular matter PM 2.5 on mortality and health care usuage, undertaking a in the dataset with a sample of 40 billion observations. To complement these insights, the authors use survival models to estimate life expectancy, also looking at two machine learning models namely a cox-lasso model and a random survival forest. These estimates are subsequently used for further analysis and for the calculation of pollution exposure costs. Thereafter, the study disentangles variation in exposure using the generic machine learning approach by Chernozhukov et al. (2018). We can see that there is a near 0 effect among 75 percent of their sample looking at elderly beneficiaries. The CDDF approach allows us to disentangle heterogeneity, and shows that from the 5 percentile threshold of the elderly population onward, pollution exposure has a very severe effect. The study illustrates that the 'mortality burden' is predominately placed within the elderly population with five to ten and two to five years remaining.

2 Data

Deryugina et al. (2019) look at three factors namely (1) air pollution, (2) atmospheric conditions and (3) mortality, morbidity and medical costs.

Air pollution is measured looking at fine particular matter PM 2.5. The Air Quality System Database (starting in 1999) provides data on the PM 2.5 scores. Other air pollution measures such as sulfur dioxide, ozone, nitrogen dioxide and carbon monoxide are also include in the dataset. For atmospheric conditions, the authors employ information on wind speed and wind direction for the years 1999-2013. They use the North American Regional Reanalysis (NARR) daily reanalysis dataset. Information on Wind is reported at the level of a 32 by 32 kilometer grid with two vector pairs. This study uses interpolation between grid points to estimate the components entailing two different directions. Average measures are subsequently transformed into wind direction and speed at the county level. To supplement this information, the authors obtain information on temperature and precipitation data from Schlenker and Roberts (2009). Measures are averaged to get county-day level measures. Mortality, morbidity and medical costs are estimated by utilizing medicare administrative data. The sample focuses on beneficiaries between the age of 65 and 100 years. To calculate medical costs, Deryugina et al. (2019) use the Medicare Provider Analysis and Review File (MedPAR). The unit of observation for these costs is the individual patient level. For the subsequent survival models, the authors also look at individual chronic conditions, to add to existing studies that only look at age and sex.

3 Methods (Empirical Strategies)

The methods and results are divided into three subsections, firstly looking at (1) mortality and health care use, then looking at (2) life years lost and thirdly (3) looking at treatment effect heterogeneity.

3.1 Mortality and health care use

For mortality the authors look at death per million beneficiaries as a three day total. The paper ensures that there is no mortality displacement by using various time windows (5,14,28 days) for robustness. Deryugina et al. (2019) suggest the following equation:

$$Y_{cdmy} = \beta \times PM2.5 + X'_{cdmy} \times \gamma + \alpha_c + \alpha_{as} + \alpha_{my} + \epsilon_{cdmy} \tag{1}$$

Here, Y is health care usage/mortality. For health care use the authors look at emergency room visits. This study uses estimates for ER visits that lead to hospital submission and as a placebo also looking at planned admissions. This measure should be independent of pollution. The study accounts for various fixed effects such as country-level fixed effects, state-by-month level fixed effects and month-by-year level fixed effects. Fixed effects are notated as α_c for state by county, α_{sm} for state-by-month, and α_{my} for month-by-year fixed effects. X is their specification for the remaining control variables. X entails 28899 weather conditions of which approximately 9300 are included per day and 27900 per regression. These measures are corrected for confounding factors. Deryugina et al. (2019) capture geographic variation by using the interaction coefficient β. This research paper takes the estimates and aggregate them to the country level, weighting by population per capita as the dependent variable.

Sometimes, pollution monitors are purposefully placed in locations with less pollution exposure to lower exposure scores. The authors create clusters for air pollution monitors to ensure that pollution not dependent on the location of the monitor. They create 100 spatial groups, using the k-means algorithm. This approach allows for robustness against bias caused by selectively placing pollution monitors. As mentioned, the paper uses an IV-strategy using daily wind direction. The first stage equation is defined as:

$$PM2.5_{cdmy} = \sum_{g \in G} \sum_{b=0}^{2} \beta_b^g 1[G_c = g] \times WINDDIR_{cdmy}^{90b} + X'_{cdmy}\sigma + \alpha_c + \alpha_s m + \alpha_m y + \epsilon_{cdmy} \tag{2}$$

The excluded variables are defined within the indicator function. The indicator function for WINDDIR is equal to 1 if the direction in the country falls within the bandwidth of the 90 degree interval and 0 otherwise. The alpha measures are then defined in the same manner as for the first equation. The authors argue that non-local pollution will be the driving force of variation within the second equation because the non-local effects will be more likely to have similar effects as opposed to the local effects which are more susceptible to differ. The effect of wind direction is confined by using bins. The specification for the remaining control variables is X. (see equation 1) This specification allows Deryugina et al. (2019) to capture variation dependent on wind changes and their affect on local pollution.

3.2 Life-Years Lost

To complement the analysis on mortality within the window-bins, this study proposed an estimate of life years lost due to pollution exposure. To measure life years lost, one usually needs information on counterfactual life expectancy which is unobserved. Previous studies for instance estimate counterfactual life expectancy via population life tables. The solution proposed in this paper is using a rich set of different health and non-health characteristics to add to existing survival estimates only looking at sex and age. The authors also propose estimates based on machine learning methods which allow us to enable feature selection. As mentioned the authors use a non-parametric model, a survival forest, and a regularization based model, a least absolute shrinkage and selection operator (LASSO) with a cox proportional hazard model. Random forest methods allow for considerable freedom due to their non-parametric nature. The regularization method induces a penalization term which is guarded by the parameter lambda, weighting the trade off between bias and variance. The authors report all life expectancy estimates for the different methods (Medicare FFS average, cox, cox with more covariates, survival random forest, cox-lasso). One should note, that there is a substantial drop in expected life expectancy when accounting for these chronic conditions and other non-health characteristics. The machine learning methods propose the lowest predicted life expectancy.

3.3 Treatment Effect Heterogeneity

To disentangle heterogeneity in their treatment this paper uses the CDDF approach proposed by Chernozhukov et al. (2018) to look at heterogeneity in age. Deryugina et al. (2019) focus on estimating the best linear predictor of conditional average treatment effect under general conditions. The authors use a gradient boosted decision tree for identification of treatment heterogeneity to get a proxy predictor. Using this proxy predictor Deryugina et al. (2019) construct a weighted regression, generate by a pooled sample with control and treatment group observations. Their research paper introduces the following notation:

$$Died_{it} = \alpha + \beta_1(T_{ik} - \hat{p}(Z_{it}) + \beta_2(T_{ik} - \hat{p}(Z_{it}))(\hat{S}(Z_{it}) - \bar{\hat{S}}) + \theta \hat{Died}^C(Z_{it}) + \epsilon_{it} \quad (3)$$

Where $\hat{S}(Z_{it})$ is the proxy predictor for mortality. The outcome $Died_{it}$ is equal to 1 if the patient died (and 0 otherwise). $\hat{p}(Z_{it})$ is the propensity score of treatment. $\bar{\hat{S}}$ is the average of the proxy predictor. $\hat{Died}^C(Z_{it})$ are their control variables. Henceforth, the β_1 in the first part of the regression is a coefficient for the difference between treatment (T_it) and predicted treatment ($\hat{p}(Z_{it})$). The second part of the regression looks at the same relationship but multiplies this term with $(\hat{S}(Z_{it}) - \bar{\hat{S}})$. $\hat{S}(Z_{it})$ is the difference in estimated deaths between treatment and control group. $\bar{\hat{S}}$ is average across the entire sample. As suggested by the authors, CDDF show that those terms are the best linear predictor of the conditional average treatment effect, allowing us to use the proxy predictor for this analysis. Next, the Deryugina et al. (2019) introduce a sorted group average treatment effect equation. The notation is as follows:

$$Died_{it} = \alpha + \sum_{k=1}^{\gamma} \gamma \times k(T_{ik} - \hat{p}(Z_{it})) * 1(G_k) + \theta \times \hat{Died}^C(Z_{it}) + \epsilon_{it} \quad (4)$$

Here, the paper uses an indicator function $1(G_k)$ to see whether the prediction lies within the kth interval or not. The gamma parameter is of special interest. It is an estimate for the average treatment effect for the group k. Two further challenges are addressed. Firstly, Deryugina et al. (2019) down-sample because the probability of dying is small and fairs poorly without doing so. Secondly, the authors argue that computational constraints make it impossible to include country, state-by-month and month by year fixed effects in this regression and instead used a modified fixed effect structure. Their research employs subgroups to estimate the equations and averages, using 250 subsets.

4 Results

For the results on mortality by age group the authors present both OLS and IV estimates, while one should note that OLS indicate significant bias. The same relationship holds for the health care usage results. The results of PM 2.5 exposure on estimates for the life years lost indicate that there is significant heterogeneity when accounting for more covariates. Using a 'back of the envelop calculation' Deryugina et al. (2019) also propose a cost estimate. One should note that with the lower life years estimates, these costs also reduce (around $76 billion dollars lower). Moreover, the study introduces the estimates for the different age groups proposed by the CDDF approach. Subsequently Deryugina et al. (2019) also explore robustness by taking into account other pollution measures, weather controls and lags.

5 Conclusion for the Paper by Deryugina et al. (2019)

The contribution to the literature of this paper is suggested to be twofold. Firstly, the authors suggest a machine learning model for time to event studies, looking beyond age and sex as determinants for life expectancy to complement existing methods. Secondly, this paper uses the novel CDDF approach to disentangle heterogeneity in our treatment effect. Briefly summarized, their study suggests that policy makers looking at the effect of exposure to pollution needs to account for vulnerable groups rather than focusing on improving the most polluted areas. This study exposes substantial heterogeneity in the effect of pollution exposure.

6 Review

This section will evaluate some of the most important and debatable components of this analysis. First, I will discuss the model evaluation for the predicted life expectancy. Afterwards, I will focus on the added value of using different machine learning methods. Subsequently, I will discuss other sources of heterogeneity that are ignored in this study. Thereafter, I will talk about further measures to improve robustness and open questions.

6.1 Model Evaluation: Machine Learning for Survival Studies

Traditional machine learning tools for model evaluation are omitted in this paper. This choice is justified by the argument that if there are no substantial differences in the results, it does not matter whether our estimates are precise. They can according to this reasoning, still be used for econometric evaluations without violating any assumptions. To advance scientific rigor, one should have at least reported some sort of model performance evaluation for the machine learning methods. If we disregard precision and do not report any information on model evaluation, why do we build a machine learning model that is so marginally different from standard measures in the first place? Model evaluation would have been a substantial improvement to this paper because it would have allowed us to assess the credibility of these different estimates ourselves.

There are various model evaluation tools that could have added value here. Model evaluation for classical classification tasks and survival tasks differentiate. To accommodate the right censored nature of our data (50 % of the individuals in our sample do not die within the time frame), we would use methods employing an estimation for our right censored observations e.g. via inverse weighted propensity estimates (IWPC). Not going into too much detail, we could have used the concordance statistics (in short c-statistics) or a brier-score to evaluate the performance. For more robustness, usually one would compare performance for the different thresholds within the data (typically: 25, 50, 75 percentage quantiles). Suppose the prediction would have been at either one of the extremes, such as an extremely high or low score, one could have had transparency rather than the uncertainty.

Further, it is very common for survival studies to focus on subsets of the sample for evaluation. Clinicians and Statisticians usually have an idea of what part of their sample is of most value to their study and which miss-classifications would translate into the highest costs and need to be penalized the most. In this case, we are interested in the most vulnerable group and how well these estimates actually reflect their true exposure. Evidently, this is a substantial shortcoming and something to consider for future research.

6.2 Added Value of Survival Analysis via Machine Learning

The difference between the the survival random forest estimate and the cox-lasso is marginal (5.33 to 5.23). The difference between the cox-ph model and the cox-lasso model is slighter bigger (5.33 to 4.8) but still not a substantial jump as opposed to the jump between medicare FFS average and the cox model. Unfortunately, it is largely unclear whether these lower predictions are actually a more accurate representation of survival, or whether these score are simply lower. It would have been really interesting to look at the different percentile thresholds for our data, given the subsequent arguments in the paper and the focus on vulnerability and heterogeneity within our sample. Generally speaking, this does not imply that machine learning tools are not applicable here. This solely implies that the contribution made here is marginal as compared to what existing methods, such as the cox-ph model, propose. One should also note that there are various way to define statistical value of life. Examples include: Excess death, hazard fraction, relative risk, premature deaths, attributable deaths (Hammitt et al., 2020). These various notions of statistical value in itself offer variety to create an estimate for life years remaining. It is not unlikely that there is also variation in these different estimates. Solely based on the research presented in this paper, one cannot exclude that the chosen measure was cherry picked to confirm a prior hypothesis or illustrate the relationship between traditional methods and machine learning estimates.

6.3 Contribution of the CDDF Approach and considerations

Firstly, one central issue for the CDDF is that performance evaluation tools are missing. As mentioned above, the authors argue that due to the argument that prediction is not of central importance, one does not need to report performance evaluation metrics. These scores would have been interesting (even if not of central importance) to put the entire analysis into perspective. Additionally, a single proxy estimator for treatment heterogeneity is hard to contextualize and evaluate. Rather than introducing a random forest for survival analysis, one could have actually used these methods to see if there are any difference between generic machine learning methods for the treatment effect heterogeneity given this high dimensional dataset. Especially also looking at e.g. bagging (Random Forest) and boosting methods (used here), it would have been insightful for completeness to see actual difference between these machine learning methods. Looking at the comparatively vast share of other measures for the other approaches, this section still seems a bit underdeveloped. One should also note that one could consider the difference between using generic machine learning methods versus e.g. a causal random forest. One can use a causal random forests for lower dimensional data when estimating conditional average treatment effects for randomized experiments. Causal random forests are less prone to error caused by splitting uncertainty and can create valid confidence intervals. Unfortunately, the causal random forest is not applicable here because we are working with a high dimensional dataset and the algorithm does not fair well in such settings. Also, the CDDF approach shows that one can utilize generic machine learning methods without violating any assumptions, given the different trajectory. Irrespective, the uncertainty remains, although substantial differences are unlikely.

Nevertheless, one should also note that this approach is a substantial contribution. Not only do the results co-align with the results provided by the estimated life expectancy, but this method proposed novel opportunity to intertwine state of the art machine learning methods for economic research questions. Not many papers have actually utilized this novel approach, and this extensive research paper provides a further pathway of acceptance to use these statistical tools despite their shortcomings with respect to interpretation. Further, this also illustrates that there is room within economics for other machine learning tools that are concerned with similar trajectories. One example is interpretable machine learning, where the goal is to remove complexity from black box models and create models that are interpretable and usable for a broader audience, including policy researchers.

6.4 Heterogeneity in our Sample

By utilizing only 5 percent of FFS medicare data, one neglects some sample variation (the authors still use 1.2 million individuals). This subset is justified in the paper as a tool for computational ease. While they do run the regression multiple times, it might have been nice to see other medicare beneficiaries. By disentangling the sample even further, one could potential discover even more heterogeneity. Accounting for factors like demographic characteristics would have unfolded further variation in our treatment. We do not even know whether there are difference between e.g. the exposure on infants and on the elderly within this study.

6.5 Other Considerations

Feature Selection: Given that the paper uses so many control variables, it would have been nice to actually include a feature selection stage for the core regression. Here, Deryugina et al. (2019) focus on feature selection for the main features but not for the other control variables. Looking at the large number of control variables included in these regression models, this seems like a instrumental consideration. For the sake of transparency, this would have been a nice addition. **Autocorrelation:** The problem with using lags is determining the optimal lag structure of our data. Here, we simple use a single unit lag which is somewhat naive, because we do not know whether the medicare data would actually translate in that manner. This is a potential source of miss-specification. Including an analysis on this issue in the appendix, would have allowed for more transparency. **Atmospheric science models:** As suggested by the authors themselves, due to the vast number of observations and data, it is computationally not feasible to apply advanced tools from atmospheric science models here. Henceforth one can only speculate how these models would have performed but this would be an interesting addition to the existing analysis. **Other Heterogeneity:** Another form of heterogeneity is the heterogeneity in our pollution exposure variable. We do not have any information on long term exposure to pollution or whether very long periods would add more insight into variation within this variable.

7 Conclusion

In conclusion, this paper does spark an interesting discourse about pollution exposure, shifting the narrative from improving the most polluted areas to focusing on the most affected areas. Essentially, this study does provide systemic evidence for heterogeneity in treatment, suggesting that the most vulnerable groups within the elderly are more affected by pollution. Irrespective, there are various things that could have been done to enhance scientific rigor. The biggest shortcoming was the implementation of the various machine learning methods. There was no substantial improvement for the life expectancy estimates by using these tools as compared to the extensive cox-model for the estimated life expectancy estimates. One instrumental consideration is the disregard for actual model evaluation, reducing their credibility. The argument that precision is only important for identification when results differ is not very rigorous. With respect to the second machine learning method for treatment effect heterogeneity one should also consider the lack of scope in the treatment heterogeneity. We clearly see strong variation in age groups when it comes to treatment heterogeneity, but this is just one of many instrumental factors to consider. Nevertheless, one should also note that this research paper is a tremendous contribution to existing research. This study truly shows the versatility of machine learning tools for applications in a systematic manner. The CDDF approach illustrates great potential for future research to disentangle treatment heterogeneity for other factors such as (in this case) demographic and socio-economic variation.

Unanswered Questions

One of the many interesting questions that remain unanswered is the effect of pollution in other shares of the populations. Even looking at the new research paper published by the mostly same research team (apart from one co-author), there is no consideration for other subgroups. This is a pivotal shortcoming, given the strong policy statements given in this and the subsequent paper by these authors. One interesting addition would be looking at infant mortality and the impact of pollution on younger age groups. Is it ethical to ground policy decisions solely based on analyses using subgroups of the population? In other terms, is it fair to make pollution reduction policies without accounting for other subgroups susceptible to pollution exposure? Further, there are more convoluted aspects that remain hard to disentangle. It remains uncertain whether we will be able to causally disentangle the impact of pollution accounting for true long-run heterogeneity within exposure. Ultimately, the question remains whether these insights will really be causal or merely a mere approximation of an causal effect, capturing some other sort of variation due to omitted factors. Especially looking at the most vulnerable sub-groups within the elderly population, exposure to any mitigation seem prominent.

8 References

Chernozhukov, V., Demirer, M., Duflo, E., & Fernandez-Val, I. (2018). Generic machine learning inference on heterogenous treatment effects in randomized experiments (No. w24678). National Bureau of Economic Research.

Deryugina, T., Heutel, G., Miller, N. H., Molitor, D., & Reif, J. (2019). The mortality and medical costs of air pollution: Evidence from changes in wind direction. American Economic Review, 109(12), 4178-4219.

Deryugina, T., Miller, N., Molitor, D., & Reif, J. (2021). Geographic and socioeconomic heterogeneity in the benefits of reducing air pollution in the United States. Environmental and energy policy and the economy, 2(1), 157-189.

Hammitt, J. K., Morfeld, P., Tuomisto, J. T., & Erren, T. C. (2020). Premature deaths, statistical lives, and years of life lost: identification, quantification, and valuation of mortality risks. Risk Analysis, 40(4), 674-695.