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 heterogeneity in mortality, health care and medical costs caused by pollution. Specifically, this paper focuses on heterogeneity in age, suggesting that a small subset of the elderly population is most vulnerable to pollution. Henceforth, 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. Essentially this paper combines various econometric methods. This paper instruments for air pollution by leveraging changes in wind direction to estimate the causal impact of PM 2.5 on health. Using this instrument, they conduct a quasi-experimental investigation. Their research ensures robustness by looking at fixed effects, climate variables, changes in country's daily wind direction. Rather than focusing on certain geospatial areas, this paper looks at large scale variation independent of geographic features. Further the authors argue that one key consideration for potential bias is mortality displacement, people who would have died regardless of treatment. Here, this paper ensures that there is no mortality displacement by using other time windows (5,14,28 days).

To account for mortality displacement outside of this window, this paper also estimates number of life years lost due to pollution exposure. The authors use various survival models, also looking at two machine learning methods namely a cox-lasso model and for reference a random survival forest. Individual level estimates are aggregate to get a country level measure.

Based on these estimates, cost caused by exposure to pollution are estimates. Subsequently, the authors suggest that mortality is not linear in age. Further, they argue that life expectancy is a better indicator for vulnerability to pollution. The argument suggests that the 'mortality burden' is predominately placed within the elderly population with five to ten or two to five years remaining. Using 40 billion people observations, this 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 population, but allows us to see that in top 5 percent has an substantial effect.

2 Data

This section looked at the different dataset. The authors looked at (1) air pollution, (2) atmospheric conditions and (3) mortality, morbidity and medical costs. (1) Air pollution is measured looking at PM 2.5 given in the Air Quality System Database, provided by the EPA starting in 1999. Other measures are also included in the dataset such as sulfur dioxide, ozone, nitrogen dioxide and carbon monoxide. (2) For atmospheric conditions, we use information on wind speed and wind direction for the years 1999-2013 given within the North American Regional Reanalysis (NARR) daily reanalysis data. Information on Wind is reported at the level of a 32 by 32 kilometer grid with two vector pairs. The authors use 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). Again, measures are averaged to get county-day level measures. (3) Mortality, morbidity and medical costs are measured by using medicare administrative data. The sample focuses on beneficiaries between 65 and 100 years old. To calculate medical costs, the authors 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 only looking at age and sex.

3 Methods (Empirical Strategies)

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

Deryugina et al. suggest the following equation:

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

Here, Y is the outcome variable,

3.1 Mortality and health care use

The authors use health care spending as a measure for medical costs and also look at health care. These measures are corrected for confounding factors. 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. Subsequently, the authors take the estimates and aggregate them to the country level, weighting by population per captia as the dependent variable.

Further the authors cluster air pollution monitors to ensure that pollution not dependent on the location of the monitor using k-means algorithm and creating 100 spatial groups.

- Robustness to different clustering choices. (good)
- IV- Strategy: Daily wind direction varying by geography
- Looking at population of 65 and above

3.2 Life-Years Lost

Rather than using statistical value of Life in this paper we are using an estimate for life years lost. To measure life years lost usually one needs information on counterfactual life expectancy, but this information is unobserved. The previous studies use either counterfactual life expectancy via population life tables. Using estimation change in cause and age specific mortality over time (here argue all prior studies only use age and sex but here we are using preconditions). Here, we want to disentangle the effect of pollution as opposed to average life years. The solution proposed in this paper is using rich set of different health and non-health characteristics. There are various way to define statistical value of life. Examples include: Excess death, hazard fraction, relative risk, premature deaths, attributable deaths. These various notions of statistical value in itself offer variety to create an estimate for exposure. Here, we are only looking at very few comparisons, without actually even comparing actual performance of these measures. Henceforth it is somewhat speculative to argue one way or another. Irrespective, it is not unlikely that there is also variation in these different estimates. This angle is not explored in this paper. Solely based on the research presented in this paper, one cannot exclude that the chosen measure was cherry picked to ensure significance. One cannot know by merely looking at the values whether it would have changed anything but it surely would have added considerable robustness, allowing for different values of statistical life measures rather than focusing entirely on estimates prediction scores without ever evaluating their performance.

The problem with using lags is determining the lag structure of our data. Here, we simple use a single year lag which is somewhat naive, because we don't know whether the medicare data would actually translate in that manner. This is perhaps also a source of mis-specification.

3.3 Treatment Effect Heterogeneity

To disentangle heterogeneity in our treatment this paper uses the CDDF approach proposed by Chernozhukov et al. (2018). The authors focus on estimating the best linear predictor of conditional average treatment effect under general conditions.

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

The authors argue that computational constraints make it unable to include country, state-by month and month by year fixed effects in this regression. Instead the authors propose the usuage of month, year and division fixed effects. Further, they use subgroups to estimate the equations and averages (using 250 subsets). Subsequently, the authors use a gradient boosted decision tree.

The contribution to the literature of this paper is suggested to be twofold. Firstly, this paper spends a sizable amount of defining a suitable estimate for life years left, partially by introducing machine learning for time to event studies for a economics study, looking beyond age and sex as determinants for life expectancy. Specifically, this paper uses two machine learning methods namely survival random forest and a cox-lasso model. Secondly, this paper uses this proposed generic machine learning for inference on heterogeneous treatment effects to a quasi-experimental study.

3.3.1 Review

The review is structured as follows: Firstly, there is a brief outlook towards the biggest constraints in this paper. Arguably, there are various constriants, but the emphasis will be on the machine learning tools used. Therefore, the following section allocates time towards the evaluation of the machine learning method for the life years lost estimate using the cox-ph lasso regression and the survival random forest. Thereafter, I will analyse the contribution by the different proposed tools also undertake a brief discourse to the novel tools, namely the CDDF approach. Subsequently follows a brief analysis of computational shortcomings mentioned in the paper. Lastly, there is a brief conclusion with an emphasis on open questions for further research.

Open Points:

autocorrelation:

- Check for autocorrelation but dont account for autocorrelation plots.
- It would have been nice to include these plots to actually get a bigger picture of the true lags rather than assuming a lag of 1 for both instruments.
- This would have allowed for more transparency.

fixed effects geography

- Authors include FE for geography but dont include it as a source of heterogeneity in treatment (same would be interesting for demographic features)

Performance of gradient boosted decision tree:

- Rather than introducing a random forest for surivival analysis, here one could have actually used these methods to see if there are any difference in generic ml methods
- Would have been a more compelling addition as opposed to the survival random forest which was not in the center stage of the analysis
- Further, again no prediction accuracy scores are reported which would have been interesting (even if not of central importance) to get a gist of actual predictive performance to put the entire analysis into proportion
- Suppose the prediction would have been at either one of the extremes such an extremely high or low MSE, it would have been interesting to see and would have allowed some more insights into the relationship of different predictors.
- 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 even if so marginal, giving the machine learning methods more attention and existential justification.

3.3.2 Model evaluation for ML

Traditional model evaluation is omitted in this paper, justified by the argument that if there is no variation in the results, it does not matter how precise our estimates are. In my opinion, this argument is very weak and to ensure scientific rigor, one should have at least reported some sort of model performance for the machine learning methods. Irrespective,model evaluation would have been a substantial improvement to this paper because it would have allowed to assess the credibility of these different estimates allowing the reader to make his own judgment about the contribution of using these estimates apart from statistical significance of the results. Given the similarity of the estimates, it is not unlikely that these estimates would have hinted towards an only very marginal improvement, if any.

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 don't die within the time frame), we would use methods employing some sort of 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 (or in short c-statistics) or a brier score here to evaluate the performance. For more robustness, usally one compares the different thresholds of the data (25, 50, 75 quantile percentage).

3.3.3 Contribution of the CDDF Approach

This appraoch is very new has not been implemented by many studies. Essentially, the argument states that one can use machine learning tools to disentangle heterogeneity without violating any important assumptions within economics. Henceforth, machine learning methods allow Economists to undertake feature selection despite inducing bias because we are not interested in the prediction itself.

3.3.4 Usage of OLS regression

- Usage of regularized method for machine learning method but not for main regression
- PLS for regular regression for transparent feature selection rather than p-value hacking

3.3.5 Importance ML method for survival analysis

- Here generic machine learning inference is used for subsequent HTE analysis.
- Inevitably, one should consider the actual contribution of this method.
- Novel machine learning method here lead to lower expected prediction of life expectancy.

The difference between the cox-ph model and the survival random forest or 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 really a 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 actual survival, or whether these score are simply lower. Further the argument that you don't need a concise estimate for identification is somewhat confusing. The entire point of building the different survival models is to get a more precise survival estimate, accounting for further conditions. If we disregard precision and don't 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? If it is not for the more accurate prediction, there is no justification for using the survival random forest or the cox-lasso model. Further the argument that it only matters if the results play out differently is unscientific. Generally speaking, this does not imply that machine learning tools are not applicable here, solely that the contribution made here is marginal as compared to what it could have been

3.3.6 Computational Shortcomings

Due to the vast number of observations and data, it is computationally unfeasible to apply advanced tools from atmospheric science models here. Henceforth one can only speculate how these models would have performed.

Unable to include country, state-by month, and month by year FE. All replaced by month, year, and division fixed effects.

Only use random 5 percent of FFS medicare sample, still amounting to 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. Little to nothing is known about this random sample and maybe by looking other specific estimates, accounting for factors like geography would have unfolded further variation in our treatment or at least influence our life expectancy estimates. Especially given how sparse information on performance of these estimates is in this paper, this might have significantly altered the estimates.

4 Conclusion

In conclusion, this paper does spark an interesting discourse about pollution exposure, shifting the narrative from focusing on 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 are more affected by pollution. Irrespective, there are various things that could have been done to further improve this study and enhanced their scientific rigor.

The biggest shortcoming was the implementation of machine learning method. While it did not harm the results, it contributed near to nothing to this study. There was no significant improvement or variation by using these tools as compared to the extensive cox-model. What is even worse is the disregard for actual model evaluation, reducing their credibility and justification even further. The argument that precision is only important for identification when results differ is not very scientific and misplaced.

A second shortcoming of this paper is 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 important factors to disentangle. Especially from a policy perspective, this study should have explored

4.0.1 Unanswered Questions

As mentioned, this paper truely addresses a broader field of research. One of the many interesting questions that remain unanswered is the effect of pollution in other shares of the populations. One interesting addition would be looking at infant mortality and the impact of pollution on younger age groups. 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 in the long run and perhaps accounting for heterogeneity over exposure time to pollution. 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. Lastly, there remains an ethical question. 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 suscept to pollution exposure? Even looking at the new research paper published by the mostly similar research team (apart from one coauthor), 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.

5 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.