Dissertation Prospectus

-

Vincent Bagilet

Sustainable Development PhD Program, Columbia University

January 19, 2023

Chapter 1

The environment in the public sphere

I Notes

This project is the one I am the most enthusiastic about and the one I would like to work on for my Job Market Paper. It is also the one in its most preliminary stage. Despite having thought about this topic in the background for quite some time, while working on other projects, I have not started actual data work. I would like to devise a clear plan before hand. Unfortunately, I have not been able to narrow the topic down to a simple causal question yet. This might be due to the fact that I am actually interested in a broad question and for which effects are "causally diffuse". I am aware that a massive amount of work still needs to be done in order to frame and answer a relevant question.

This is also the project I am looking for the most feedback on. I am wondering to what extent only having the two first sections (the descriptive/diffusion part and a simple theoretical model) as my job market paper would be detrimental on the market. That would be a step back from the usual precise but arguably narrow causal studies that focus only on a couple of actors. It would enable me to build a broad picture on how environmental questions might be gaining a more central position in the public sphere. If this type of study is not sufficient, I am also looking for feedback on the type of subquestions worth addressing, in particular the actors it would be interesting for me to focus on.

II Quick introduction to the project

II. 1 Motivation and broad questions

The motivation for this project stems from the fact that environmental questions are more and more present in the public sphere, in particular in the media (Boykoff 2007, Schmidt et al. 2013). I am wondering how this came about and how environmental questions diffused in the public sphere, across actors and social strata. Was it first scientists that mentioned these questions? Were they then vocalized by activists, by the cultural sector, directly by the press, by radio broadcasts or some other actor? Did they directly affect the whole population or just a specific social stratum? When did public opinion react? When did brands and marketing agencies considered that their customers cared about environmental questions? When did it lead to policy changes?

Basically, I am trying to capture the environmental *zeitgeist* and to get a sense of how it is forming. Ultimately, I would also like to study its impact on individual attitudes or behavior, in particular in terms of support for environmental public policy or environmental advocacy. Understanding the determinants and consequences of a potential environmental *zeitgeist* is paramount to me. The current environmental "crises" are multidimensional and require vast changes that would not stem from nowhere. For such large changes to arise, environmental questions have to be at the forefront of the public debate and to occupy a central position in the public sphere[1]. There would also need to be some public support and demand for such large policy changes. I believe, and need to test that in the data, that public support is facilitated by the omnipresence of the environment in the public debate. The position of the environment in the public sphere could indeed affect individual attitudes and behaviors in various ways. It could provide individual with information regarding both the consequences of decreased environmental quality and the impact of anthropic actions on environment quality. It could also affect perceived social norms and second order beliefs.

II. 2 Descriptive section

Capturing the environmental *zeitgeist* and getting a sense of how it is forming would constitute the first part of this project. The goal of this section would be to describe how environmental concerns and discussions diffuse across actors and social strata. The steps to do so would be:

1. Gather data. It would mostly be text data. I am thinking of considering a rather wide array of actors and sources, including academic research, medias (newspapers, radio, TV), ads (as a proxy for what firms think consumers value), culture (movies, books, songs), politics (policy texts, speeches), public opinion, finance, education (school programs), activism, etc. This list being rather long, I could first focus on academic journals, newspapers articles, ads and public policies. I have already identified precise data sets for many of these actors and sources.
2. Run classification algorithms to determine whether each text is about the environment. I could apply the same algorithm to each source. That would make the cost of considering one additional source rather low. This low marginal cost of adding one source is the reason why I am thinking of considering a wide range of sources. I plan to compare an array of algorithms, ranging from simple co-occurence to more complex supervised machine learning classification methods. Also I potentially could consider more complex sentiment analyses and feed my data to a deep-learning language model like GPT-3.
3. Build graphs to describe the evolution of the environment in the public sphere by year and by "actor" or sources. That should enable me to visualize which actors and which

1. They could of course be intertwined with other key questions such as inequalities.

social strata first start making the environment a central issue. I could also explore whether there are potential shocks in environmental "coverage" that I could leverage in a causal section.

4. Estimate some simple SVARs models, mostly for descriptive purposes.

II. 3 Causal section

I assume this paper should also contain a more causal section. I can think of two types of causal analyses I could carry out on this topic. I would like to tie both of them to questions of culture and social norms. I (non-scientifically) believe that to address the current environmental crises, there needs to be a structural change. Usual and marginal policy changes would be insufficient to address these crises. Such a large change could thus only happen if we, as a society, reconsider our value system. I am therefore interested in studying how cultural and social constructs form and affect what people, individually and collectively, aspire to. These cultural and social norm aspects are diffuse, making them in essence particularly challenging to capture.

In the first type of analysis, I could study an example of how representations of the environment in the public sphere affect individual aspirations, attitudes or behaviors. To me, a good example of the impact of a cultural representation on individual behavior and aspirations is Ferrara et al. (2012). In this paper, the authors study how the depiction of small nuclear families in soap operas affected fertility in Brazil. My interpretation of it is that a different cultural representation of what family is affected what households aspire their families and lives to be like. Fertility choices are not benign decisions and I find remarkable (but not surprising) that a cultural object such as *telenovelas* affects them. I would love to carry out a somehow similar analysis but for an environmental question. Basically, I would like to measure whether and how an increased or a change in the depiction of the environment in the public sphere affects individual attitudes and behavior towards the environment. I would thus need to first focus on a particular outcome variable, measurable at an individual level. I would much rather consider revealed than stated preferences. I would also need a shock to environmental representations. I believe that looking at the data from the descriptive section could help me identify such a shock.

Otherwise, and more linked to the descriptive section of this paper, I could causaly study the diffusion of shocks to environmental representations between actors and social strata. I would need to focus on a specific environmental aspect and find exogenous shocks to its depiction by some actors and look at how it affects other actors. For instance, old movies directed by Hayao Miyazaki with strong environmental themes such as *Nausicaä of the Valley of the Wind* gained exposure after the commercial success of a later Miyazaki movie with no strong environmental component, *Spirited away*. It might have not affected all countries similarly, depending on the local success of *Spirited away*. If so I could instrument new exposure to *Nausicaä* and look at its impact on depiction of pollution in newspapers or on TV. This example is hypothetical and far from being perfect but it hopefully gives an intuition of the type of analysis I could carry out. I could consider a small set of shocks and impacts (about 3), *e.g.*, fiction on newspapers, newspapers on ads, school programs on public policies. Identifying a set of shocks would be more challenging than only using one but with a wider range, I could afford to have less polished analyses and could keep the topic broad. If I focus on the effect of one actor on another, that would necessarily make my paper very specific to these two actors. I would also need to develop a strong institutional knowledge of these two actors. Instead, I would rather try to build a broad but coarse picture, including various actors. However, I may need to be pragmatic and resort myself to study only one shock as it might be challenging enough. I am hoping that the descriptive section will enable me to identify interesting shocks to look into. I also hope that the discussion in my orals will help me clarify the type of interesting relationships to study. To think about actual causal identification strategies, I would need to narrow the set of actors and impacts down.

II. 4 Theoretical section

In between the descriptive and causal sections, I could add a simple theoretical model to serve as a conceptual framework. Based on observations from the descriptive section and the existing literature, I could build a simple theoretical model of diffusion of environmental matters among actors and social strata. This model could describe cases where some actors and social groups lead the way in making the environment a central issue in the public sphere. It could also describe some diffusion mechanisms. I however do not know what form this model would take yet.

If I consider the other type of causal question, *i.e.* the impact of the position of the environment in the public sphere on individual action, I could build a simple toy model to introduce the causal section. I could for instance adapt a very simple game theory model of individual environmental action I have built based on Heal and Kunreuther (2010) and make it more specific to the questions I address in this project. A description of this model is available here.

II. 5 Context

As the effects I am trying to capture are likely small, I would need to focus on a context where I expect environmental concerns to be relatively important. To have have enough statistical power, I would need to give up on some external validity. As I would be working with cultural aspects, I think I would like to work on a context in which I would be an insider. Based on these two requirements, I am currently thinking of working on France.

For the descriptive section, I could however consider a small set of countries, for instance

France, the US and an additional country with high environmental concerns, maybe a nordic one.

II. 6 Reasons to work on this project for my JMP

I really want to work on this project for my JMP despite the fact that I have not concretely started working on this project, for several reasons. First, so far, I have spent most of my PhD learning more about causal inference. This resulted in the papers described in Chapters 3 and 4 of this prospectus. While I enjoyed learning all this and will be able to apply it to other questions, I would consider finishing my PhD without having worked on culture and the environment a personal failure. I would like to dedicate the remainder of my PhD to topics that I am truly passionate about.

Then, I think that this project would be a good opportunity to work with nice and original data. I think my comparative advantage on the market may lie in the fact that I am willing to gather and wrangle a large number of "unconventional" datasets. When my objectives are well defined, I believe that I can crunch quite a substantial amount of coding rather quickly. Incidentally, it would also be an opportunity for me to learn new techniques and methods that may be in high demand (mostly Natural Language Processing and Machine Learning techniques).

Next, even though I have not properly defined the question I want to address in this project yet, I think the general topic is of paramount importance. The scope of this project might be ambitious but I think that it might be helpful to build a broad but imprecise picture of the position of the environmental in the public sphere. My impression is that most existing economic studies are causal and narrow and do not provide a sense of the broad picture. Similarly, it seems to me that other social sciences could but have not yet leveraged the new tools available for big data and text analysis to build this broad picture. It seems that most studies focus on environmental attitudes as measured via public opinion surveys. These are helpful but may exacerbate the weight put on individual action and not fully acknowledge the interlocking of households attitudes and behavior into larger social and cultural structures.

Finally, I am aware that the bulk of this project relies on descriptive methods that are not in the usual standards of what is expected for an empirical job market paper. That is why I would like to build the small causal case study to showcase that I am capable of carrying out causal analyses, even if I mostly use descriptive methods in this project.

III Next steps and pending questions

1. **Build a clear research plan**, prioritizing the various tasks I have to implement.
2. **Define a clear research question** and thus frame this question and define a context. Is the question of the formation of an environmental *zeitgeist* and its diffusion between

actors interesting and sufficient? And thus $\hookrightarrow$

3. **Is the descriptive section worth carrying?** If so, I should maybe refine a bit the question before starting to run the analysis. And if it is worth carrying $\hookrightarrow$
4. **Do I also need a causal section?** If so, I would first need to choose between the two types of potential causal analyses described in Section II. Then, I would need to identify which actors I would like to consider and then think about convincing identification strategies.
5. **Organize my literature review**. I have read or skimmed through a pretty large number of papers on the broad topic of culture and environment, attitudes, preferences, social norms. In my opinion, these readings are too broad for a review to be actually useful. When I have a clearer and narrower idea of the type of question I want to tackle, I will need to organize all these readings into a clearer literature review. That may also help me refine my research question. Most of the papers I have read were from economics journals, I should expand my readings to other disciplines.

Chapter 2

Public acceptability of environmental regulations:

gasoline tax pass-through and the yellow vests crisis

I Notes

I included this project in my prospectus mostly to decide whether it is worth pursuing and in order to receive feedbacks on the potential future developments I could undertake. In my opinion, this project would only be worth continuing if I can formulate an interesting question. To do so, I would need to convincingly tie the pass-through analysis to the yellow vest crisis.

This project would be sort of a back-up project that I would work only sparsely on for the remainder of my PhD. I only aim to have a first set of results and to potentially use this project to take "breaks" from the one I will focus on more and that I described in Chapter 1. I worked on the pass-through part of this project for my master thesis at PSE. I updated it for Wolfram's class. An updated draft version can be accessed here.

II Introduction to the project

The French yellow vests movement emerged suddenly during the fall of 2018, initially in response to increases in the tax component of petrol prices resulting both from an increase in a carbon tax component in 2014 and a catch-up of the diesel tax system towards gasoline's. Protests raised due to distributional impacts of the tax as it was criticized for strongly affecting the budget of low income and rural households with no access to public transports. Protesters denounced both the design of the tax and the lack of alternatives to cars in these areas.

In this project, I aim to investigate whether the local intensity of mobilization was directly linked to the local tax rate or to more structural factors [1]. The acceptability of this tax has partly been studied quantitatively using surveys and descriptive methods. However, to my knowledge, a clear empirical link has not been established with the actual burden of the tax. Using a representative online survey, Douenne and Fabre (2022) study acceptability of carbon taxation in France, in the wake of the yellow vest crisis. They show that respondents form

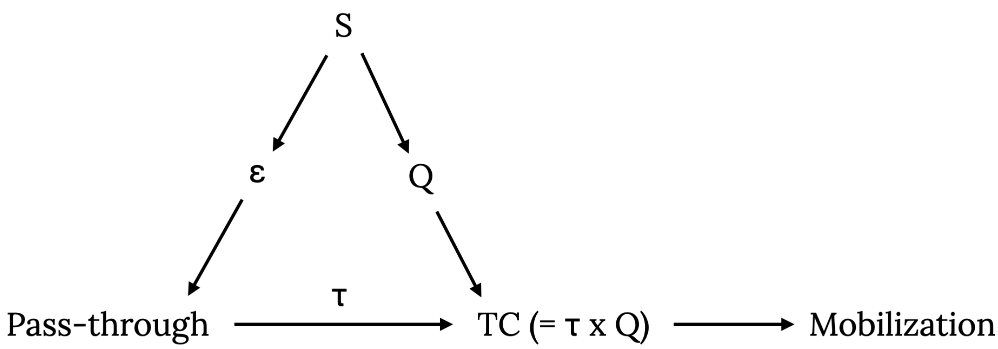
1. The exact question to address in this paper might however still be to refine in order to make this paper interesting.

incorrect beliefs about the impacts of the tax and largely reject carbon taxation. Boyer et al. (2020a) explored the determinants of yellow vest mobilization through descriptive analyses. Their results suggest a link between mobilization and mobility. They show that local mobilization intensity is positively correlated with the average distance between workplace and home. They also find that it is positively correlated with inequality, unemployment rates and abstention and negatively correlated with vote for Emmanuel Macron.

Disparities in the pass-through rate may have also played a role in the blooming of the yellow vests movement. Following the protests, the French government froze any further tax increases in the transportation fuel tax rates. If the direct effect of the tax was not the actual

main cause of protests, the government's response did not address the actual problem and postponed climate change mitigation efforts. Discussing the link between pass-through and mobilization might be key due for environmental policy acceptability. Yet, describing a simple correlation (or its absence) between the local pass-through and mobilization would be insufficient. Structural factors (S) such as spatial integration and thus mobility patterns or availability of public transport alternatives might indeed affect both pass-through and mobilization. Households leaving in regions with less alternatives to car might use more their cars. In such area, the average distance traveled by car by households (Q) would thus be larger. The demand would also be less elastic ϵ, leading to higher local pass-through rates and higher local effective tax rates faced by households τ, *i.e.* after the tax has been passed on to consumers. The impact on households and therefore the initial local level of mobilization would thus be linked to both Q and τ, the total cost of the tax faced by consumers being $\tau \times Q$. Mobilization levels would be affected by the local pass-through through τ and by Q. S would create an open back-door path in the following Directed Acyclic Graph (DAG), even assuming that mobilization is solely determined by the local impact of the tax:

Figure 2.1 – DAG for pass-through and mobilization



In the aggregated pass-through analysis, these structural factors should be absorbed by the station fixed effects. In the local analysis, I could try to control for/match along structural factors S at the petrol station level, such as the population density in the municipality, the distance to the closest city, the average distance between work place and home in the municipality and the share of inhabitants commuting by car. However, this strategy would not yield credible causal effects. I could also consider plausibly exogenous shocks to the pass-through rate and study their impact on mobilization. For instance, pass-through rates might vary across gas station brands. Different brands might use different formula to set their prices, leading pass-through to vary across brands. If so, I could instrument the pass-through rate with proximity to station from a brand with higher (lower) pass-through rate. Similarly, some brands, associated with supermarkets, occasionally run marketing plans in which they sell gasoline with zero markup to attract customers to their shops. For instance, E.Leclerc, launched such an initiative a few weeks before the burst of the yellow vest crisis, in a subsample of its stations, leading some areas to have access to relatively "randomly" low prices. This marketing plan would not be directly linked to pass-through but to gasoline prices themselves. Some plans might however also have happened around carbon tax increases. I will need to investigate these further, if the question of the link between pass-through and mobilization is actually worth exploring.

Studying the pass-through of fuel taxes to gasoline prices is of interest in itself, even without explicit links to acceptability. Any thorough impact evaluation of a tax change on households consumption requires an assumption on the magnitude of its pass-through as any further analysis would rely on the actual price change implied by the tax increase. Measuring the pass-though of fuel and fuel taxes is thus paramount for their economic analysis. In the literature, most papers studying petrol consumption of households assume a full pass-through of fuel taxes to the consumer. This assumption stems from standard economic theory which shows that the most price inelastic factor bares the burden of the tax. Short run price elasticities of demand for transportation fuel are often estimated to be relatively small, leading tax increases to fall entirely on consumers (Hughes et al. 2008).

These theoretical findings have been confronted to empirical analyses in the US, mostly based on monthly US data at the state level and relying on differences in fuel tax systems across states (Chouinard and Perloff 2004; 2007, Alm et al. 2009, Marion and Muehlegger 2011, Lade and Bushnell 2019, Kaufmann 2019). The automotive transportation sector might however largely differ between the US and other countries. For instance, per capita consumption of transportation fuel is about 2.5 times higher in the US than in France (in 2015, 1225 Mt/capita in the US, 489 Mt/capita in France (IEA 2015)). Until recently, empirical evidence from other parts of the world remained scarce. Closely related to the US, Erutku (2019) studies the case of two Canadian provinces, Ontario and Quebec, and finds evidence of over-shifting in Ontario. To my knowledge, only Harju et al. (2022) and Stolper (2018) assess empirically the incidence of fuel taxes outside North America. To do so, they both leverage daily-station level data. Harju et al. (2022) uses data from Finland and take advantage of a large increase in diesel price in 2012 through a difference in differences (DiD) strategy, using gasoline as a control. They find a pass-through of approximately 80%. They also document spatial heterogeneity, finding larger pass-through in low income and rural areas.[2] Stolper (2018) investigates spatial heterogeneity in the pass-through of Spanish fuel taxes.[3] The author finds an average pass-through of 1 but provides evidence of important heterogeneity with pass-though at the station level ranging from 0.7 to 1.2. He also shows that higher pass-throughs are associated with greater market power. My study closely relates to these two papers as it also focuses on heterogeneity and leverages daily-station level data in an European context. However, it aims to to explicitly tie this heterogeneity to public acceptability concerns.

France implemented a carbon tax in 2014 at the level of $7€/t_{CO_2}$. It is scheduled to reach $100€/t_{CO_2}$ in 2030 and, as of 2018, its value is of $44.6€/t_{CO_2}$ (MTES 2017).[4] Due to the large carbon content of transportation fuels, the carbon tax affects their prices through a new carbon component in fuel taxes. The fuel tax increased each 1^{st} of January because of increases in its carbon component but also, in 2016, 2017 and 2018, because of a catch up procedure of the diesel tax system towards gasoline's. Using daily data at the gasoline station-level, I assess the extent of the pass-through of the French fuel tax to transportation fuel prices. To do so, I implement a DiD strategy where Spain is used as a control group. Spain makes a suitable control group for transportation fuels prices as both countries present similar characteristics. I then investigate geographic heterogeneity in pass-through rates between rural and urban regions, between poor and rich ones, depending on the level of competition and the location on an highway or not. I also try to link these variations to local intensity of yellow vests protests. In the current version of the paper, I only consider simple interactions. In a subsequent version of this work, I plan to run a matching or synthetic control algorithm in order to compute a pass-through at the station level.

I find that increases in the French fuel tax are more than fully passed on to consumers. A 1c€/L tax increase leads to an increase of 1.2c€/L in diesel prices and of 1.3c€/L for gasoline prices. This result can be explained by market power and strategic behaviour of transportation fuel retailers. Moreover, there seems to be some heterogeneity among years with a pass-through rate close to 1 in 2015. I should however recover ownership data, and potentially data on volume sold to confirm these findings when using a better set of controls. While I still need to carry out further research on this topic, I find some evidence of spatial heterogeneity in pass-through rates. I also show that there might be some anticipatory behaviour or stickiness of prices. Yet, the magnitude of these effects remains limited.

The contribution to the literature of this paper would be threefold. First, it is the first paper to provide evidence about the magnitude of the pass-through of the carbon tax in France. Most

2. This heterogeneity analysis was not present in the version of this paper available when I was working on my master thesis.

3. I was not aware of the existence of this working paper when I started working on this project. I learned about its existence only a year after the submission of my master thesis.

4. MTES refers to the Ministry for the Ecological and Inclusive Transition (*Ministère de la Transition Ecologique et Solidaire*)

existing studies focus on the pass-through of fuel taxes in the US and empirical evidence in other contexts is still needed. Second, most published papers find a pass-through of transportation fuel taxes smaller or equal to one. This paper provides evidence of a significant over-shifting. This further underlines the specificity of the French transportation fuel sector. Finally, to my knowledge, it would be the first paper trying to relate public acceptance issues and protests to questions questions of pass-through of gasoline taxes.

III Next steps

1. **Is this project worth continuing?** I first need to answer this question. Considering the recently published papers (in particular Harju et al. (2022)), the pass-through section would not be enough in itself, even after adding a good heterogeneity section. The interest of this paper would depend on how I could make the yellow vests part relevant.
2. **Frame an interesting question I could answer**. I could carry out the analysis to see if there is a correlation between the local pass-through and yellow vest mobilization intensity. However, this would require some substantial work and I should only do it if I can frame a reasoning such that both a positive and a negative answer to this question would be interesting. In the mean time, I could tried to see if there is a correlation between the mobilization intensity and a crude measure of the pass-through (the difference in gasoline prices one week before the tax change and one week after). I computed how this price change deviates from the mean for each petrol station in 2018 and plotted an average in Figure 2.2. For comparison, this figure also displays a map of in person yellow vest mobilization provided by Boyer et al. (2020b). There seems to be some correlation with larger mobilization in regions with higher pass-through. The existence of this correlation is however debatable and could be an artifact of some computational choices.
3. **Recover usable yellow vest data**. A research team already mapped the yellow vest mobilization, both online and offline Boyer et al. (2020a;b). I guess I could recover their data. I studied at PSE with one of the authors, Germain Gauthier, we worked on a few class projects together. Since he knows this topic well, I could also work with him on this project, if he is interested.

   Otherwise, to get protest participation at a finer resolution, I was thinking I could use cell phone data at the tower level to measure the number of people present at each meeting point (the protests were initially scattered all over the country, around roundabouts). I have not looked into how to get this data yet. Tower level data would be sufficient for my purpose. I assume it would be more easily accessible than disaggregated data and regardless, I do not want to use such individual data because of the privacy concerns associated with their use.

Figure 2.2 – Comparison between local 'pass-through' and mobilization

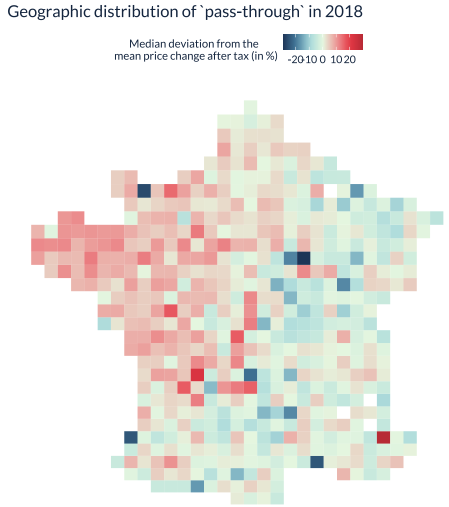
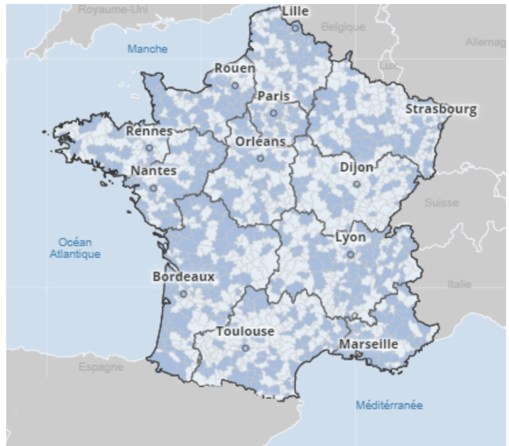


Figure A.4 – Blocking half of France at first try



Notes: Darker areas are consolidated municipalities affected by a blockade on 11/17.
Sources: Blockade map.

4. **Retrieve station ownership data**. These data are not available in the historical dataset I use but are available on files for the current day. Until recently, the website had some complex `javascript` code but has now been migrated to a more recent interface. Recovering the data might thus be easier. Also, the French Ministry of the Environment, who is making the price data available, have access to this data directly. Otherwise, I could use the Internet Archive or something similar to scrape this information.
5. **Recover data on the volumes of gas sold per station**. When I was a master student, I reached out to the French Ministry of the Environment who has these data (if I understood correctly). They were not willing to give me this data. They may be more willing to give me access to it now that I am a PhD student.
6. **Gasoline price data for Italy**. I have started recovering them. I can clean my code and recover all of these data rather easily now. However, due to regional variations in the gasoline tax level, recovering it is time consuming and I should only do it if I actually need data from another country.

Chapter 3

Unconfounded but Inflated Causal Estimates

I Notes

I initially worked on this project with Léo Zabrocki-Hallak during is PhD at the Paris School of Economics. Léo is however leaving academia. He is dropping all his research projects and wants me to turn this project into a single authored one, only mentioning him in the acknowledgements section[1]. A version of this paper has been published as a CEEP working paper in March 2022. As the point we make in this paper is rather straightforward once underlined, I plan to submit this short paper to the *American Economic Review: Insights* in the coming weeks. However, since the CEEP version of the working paper, I added a key components to the paper: a mathematical proof of the trade-off we underline. I also restructured part of the argument, to center it more on the trade-off and not only exaggeration. With this added material, I could also consider a "usual length" publication. I could also flesh out tools to visualize the variation used for identification but I am struggling to define a cohesive framework that would enable to apply this tool to all causal identification strategies.

The following summary is a modified version of the new introduction of the paper. Details of our work are described on the project's website. The mathematical proof, absent from the CEEP working paper is available here. I am mostly looking for feedback on whether the argumentation in the introduction is convincing enough and whether the structure and content of the mathematical section is good enough.

II Abstract

Quasi-experimental studies make empirical economics credible. To avoid confounding, causal identification strategies focus on a subset of the variation in the data, the plausibly exogenous part. In this paper, we argue that it can reduce statistical power and lead published estimates to exaggerate true effects sizes. Causal strategies avoid a bias at the cost of generating another one. We show that using causal inference methods creates a trade-off between confounding and exaggerating true effect sizes using realistic fake data simulations and a mathematical derivation. We then discuss potential avenues to address this issue.

1. We built the idea and the structure of the argumentation together. Léo took care of the matching simulations and I did the RDD, IV, event-study and cross-sectional ones. I derived the mathematical section and I am working on the visualization section.

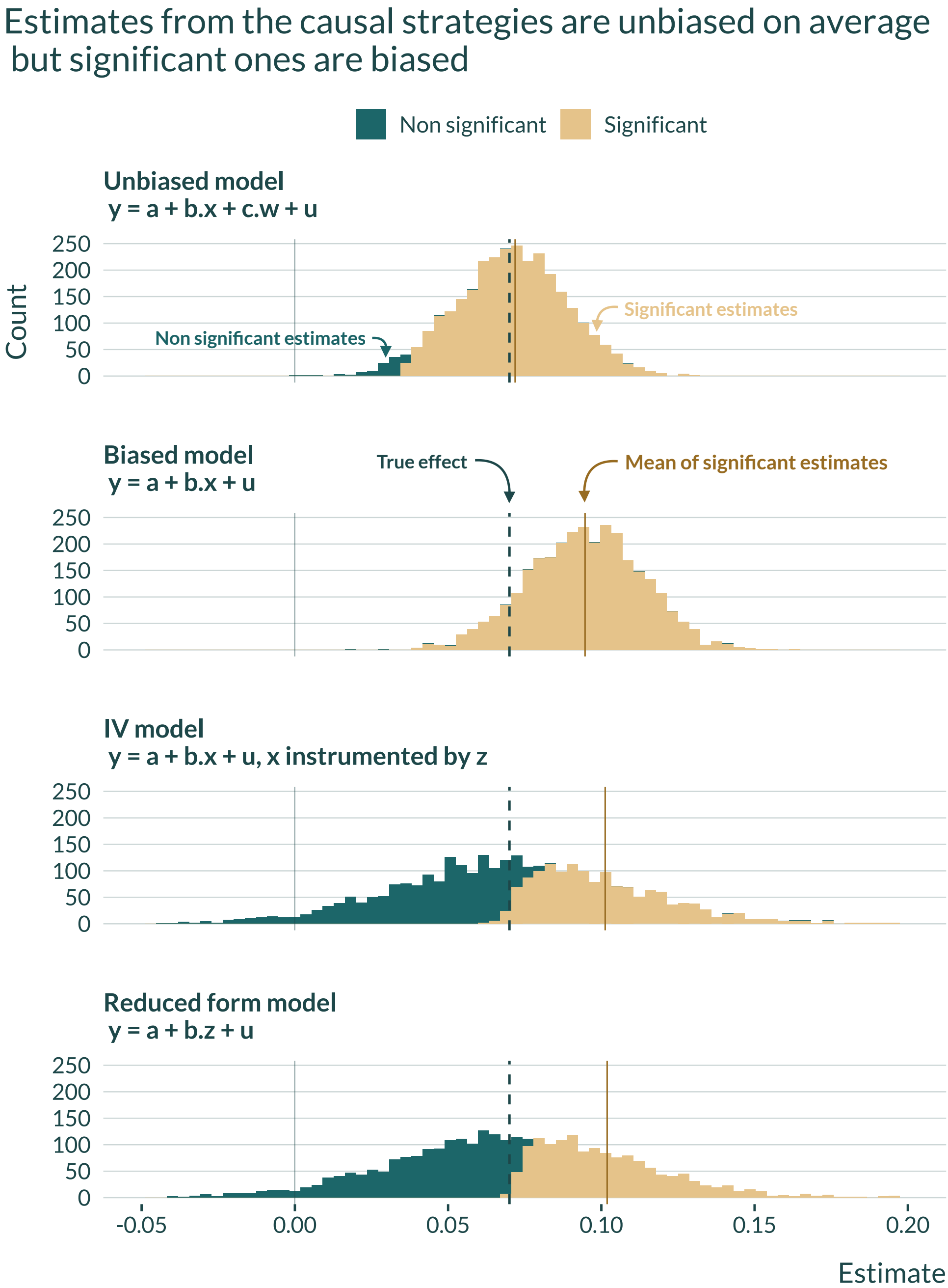
III An introduction to the project

One of the main challenges of empirical economics is to reduce confounding to identify causal effects. Identifications strategies such as Regression Discontinuity (RD), Instrumental Variable (IV), Difference-in-Differences (DID) and event studies help achieve this goal. To do so, these strategies only use part of the variation in the data. They exploit the exogenous part of the variation in the treatment or decrease the sample size by only considering observations for which the as-if random assignment assumption is credible. Reducing the variation used decreases precision and thus statistical power, that is to say the probability of rejecting the null hypothesis of no effect when it is false. It often creates a tension between statistical power and reducing confounding.

When statistical power is low, not only the estimator is imprecise but statistically significant estimates exaggerate the true effect size of the treatment of interest (Ioannidis 2008, Gelman and Carlin 2014, Lu et al. 2019, Zwet and Cator 2021). Only estimates at least two standard errors away from zero will be statistically significant at the 5% level. In under-powered studies, these estimates make up a selected sub-sample of all estimates, located in the tails of the distribution of all possible estimates. The average of these statistically significant estimates will differ from the true effect, located at the center of the distribution if the estimator is unbiased. When power is low, obtaining a statistically significant estimate from an unbiased estimator does not guarantee that it will be close to the true effect. An estimate $\hat{\beta}$ of β might be unbiased in the traditional sense but conditionally biased: $\mathbb{E}[\hat{\beta}] = \beta$ but $\mathbb{E}[\hat{\beta}| \text{ Significant}] \neq \beta$. Figure 3.1 illustrates this trade-off through the estimation of different models on simple simulated datasets.

This consequence of low statistical power could be non-problematic if a large literature had not underlined the existence of a publication bias favoring statistically significant results (Rosenthal 1979, Andrews and Kasy 2019, Abadie 2020, Brodeur et al. 2020, for instance). Published estimates from under-powered studies thus form a biased sample of the distribution of causal estimates and can greatly exaggerate their true effect sizes. This participates to the current replication crisis affecting various fields such as economics, epidemiology, medicine or psychology (Button et al. 2013, Open Science Collaboration 2015, Camerer et al. 2016, Chang and Li 2022). Even in experimental economics, with a high level of control and an arguable absence of confounders, estimates published in top economic journals have failed to replicate (Camerer et al. 2016). Original studies were on average inflated by a factor of at least 1.5. Quasi-experimental studies are likely even more exposed to this exaggeration issue as, in current practices, statistical power is not central to the analysis. Despite usually large sample sizes, Ioannidis et al. (2017) concernedly finds that the median statistical power in a wide range of economic studies is no more than 18% and that nearly 80% of estimates may be exaggerated by a factor of two. The magnitude of these exaggerations is thus considerable and in some situations, it could be on par with that of bias caused by confounders. It is thus essential to take exaggeration risks into account.

Figure 3.1 – Distribution of estimates across models on simple simulated datasets



Notes: y, x, w and u are respectively the outcome variable, the explanatory variable of interest, an unobserved variable and an error term. Each model is estimated on 3000 simulated datasets. Parameter values are chosen to produce designs with low power. The plotted distributions are the resulting estimates b, the parameter of interest. Parameter values as set to make the IV and reduced form estimands comparable (an increase in z yields the same increase in x).

In this paper, we show that design choices in quasi-experimental studies can be seen as a trade-off between avoiding confounding and overestimating true effect sizes due to a resulting loss in power. To limit the threat of confounding, causal inference methods discard variation thus reducing statistical power. This can result in inflated significant and published estimates. In some settings, reaching unbiasedness can actually be impossible as getting rid of bias caused by confounders creates another type of bias.

We analyze the key factors affecting the confounding / exaggeration trade-off for a wide range of identification strategies : RD, IV, event studies, strategies such as DiD relying on fixed effects (FEs) or generic controls and matching. While the general idea that causal inference methods discard variation to identify effects is shared across strategies, the confounding / exaggeration trade-off is mediated through a distinctive channel for each of them. In RD designs, while the initial sample size may be large, we discard part of the variation by only considering observations within the bandwidth, decreasing the effective sample size and thus precision. In an IV setting, we only use the subset of the variation in the treatment that is explained by the instrument. In event studies, the variation used to identify an effect sometimes only comes from a limited number of treated observations. Generic controls or fixed effects can increase the variance of the estimate if they absorb more of the variation in the treatment than in the outcome variable. Assuming that all confounders are measured, matching prunes treated units that cannot be matched to untreated ones.

In the first section of this paper, we illustrate the existence of this "causal exaggeration" using examples drawn from education, labor, environmental and political economics. The inflation of statistically significant estimates can be defined as the ratio of the estimated effect over the true effect, which is never known in real world setting. We therefore have to turn to fake-data simulations. Since our Monte-Carlo simulations have an illustrative purpose only, we intentionally focus on settings in which statistical power can be low. All other simulation assumptions are chosent to make it easier to retrieve the effect of interest. We consider simple linear models with constant treatment effects and all our models are correctly specified and accurately represent the data generating process, except for the omitted variable.

In the second section of the article, we derive a formal proof of the existence of this trade-off in a simple linear homoskedastic setting. Specifically, we show that the bias caused by exaggeration can be larger than the one caused by confounders. We also study the drivers of exaggeration and show that it increases as the strength of the instrument decreases, the number of exogenous shocks decreases or when adding covariates that absorb more of the variation in the treatment than in the outcome.

In the third section, we discuss avenues to address this causal exaggeration. First, we advocate for the use of tools to evaluate the potential magnitude and risk of both confounding and exaggeration issues. Sensitivity analyses help with the former while power calculations help with the latter. By approximating the data generating process, prospective power simulations help identify the design parameters affecting power and exaggeration (Gelman 2020, Black et al. 2021). Retrospective power calculations allow to evaluate whether a study would have enough power to confidently estimate a range of smaller but credible effect sizes (Gelman and Carlin 2014, Stommes, Aronow and Sävje 2021). Our companion website describes in details how such solutions can be implemented. Focusing more specifically on the trade-off and its drivers, we present tools to visualize the variation actually used for identification when using causal identification strategies. Finally, we briefly discuss potential solutions to mitigate this trade-off.

Our paper contributes to three strands of the literature. First, the idea that causal identification estimators, while unbiased, may be imprecise is not new; this is part of the well-known bias /variance trade-off (Imbens and Kalyanaraman 2012, Deaton and Cartwright 2018, Hernán and Robins 2020, Ravallion 2020). In under-powered studies, resulting estimates have large confidence intervals, suggesting that a wide range of effects are consistent with the data. We approach this literature from a different angle: through the prism of statistical power and publication bias. Not only the limited precision resulting from the use of causal identification methods could make it difficult to draw clear conclusions regarding the exact magnitude of the effect but we argue that it might also inherently lead to inflated published effect sizes, creating another "bias".

Second, recent studies discussing the inflation of statistically significant estimates due to low power focused on specific causal identification methods separately and usually do not investigate the determinants of this exaggeration (Schell et al. 2018, Black et al. 2021, Stommes, Aronow and Sävje 2021, Young 2021). this paper, we suggest an overarching mechanism, inherent to causal identification strategies as a whole, that can explain these issues. Although each strategy does so through a different means, their essence is to discard part of the variation and can create exaggeration. This connection could be exacerbated by the fact that, as noted by Brodeur et al. (2020), publication bias is more prevalent for some methods such as the IV.

Third, our study contributes to the literature on reproducibility in economics (Camerer et al. 2016, Ioannidis et al. 2017, Christensen and Miguel 2018, Kasy 2021). The trade-off presented in this paper may be an additional explanation for observing replication failures in empirical economics, despite the widespread use of convincing causal identification methods.

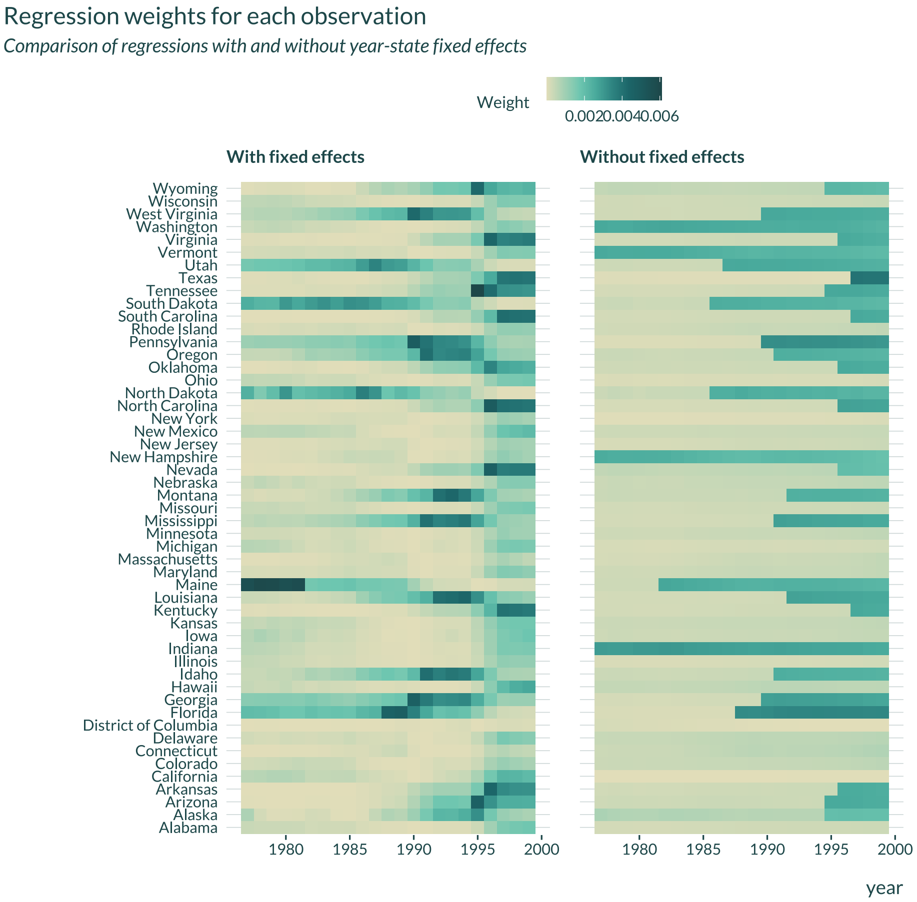
IV Next steps

1. **Make sure the structure of the introduction is convincing**. I see this paper as mostly an argumentation challenge. I think the point we make is valid and important; the main challenge is to make sure it is clear to readers, editors and referees.
2. **Write a short version of the math section**. For now, I have a long and detailed version (basically the appendix) available here. I need to write the section in the paper.
3. **Decide whether to develop a data visualization tool**. This tool should enable to visualize where the variation in the treatment used for identification actually comes from and thus to gauge the risk of exaggeration. However, I have trouble finding a cohesive framework that fits all identification strategies. The difficulty comes from the fact that

often various identification strategies are used together (*e.g.* RD and event-study in studies of air pollution alerts for instance etc) and that the control/fixed effect affects most analyses.

The tool I have in mind at the moment is a heatmap of the contribution of each observation to the estimate of the treatment, with time on the x-axis and individual identifier on the y-axis (or choropleth maps faceted by time period for spatial data). For matching, RD and event study it is basically a visualisation of the treatment allocation. For matching and RD, it makes explicit which observations are actually present in the final data set and which are dropped. For event studies, it makes explicit which units are treated and which are not. Regarding generic controls and FEs, it creates a heatmap of the regression weights defined by Aronow and Samii (2016), as illustrated in Figure 3.2. The more variation in the treatment is explained by controls, the smaller these weights. Hence, observations whose treatment status is very well explained by controls have a small weight. As a result, only a small subsample of the observations may actually contribute to the estimations and precision might be low and thus exaggeration high. For IV with a continuous instrument, the visualization I should build is not obvious to me.

Figure 3.2 – Example of the visualization with regression weights from Aronow and Samii (2016)



If I actually build such a tool, I assume I should wrap it into a `R` package. Making this package should only add about one day of work, once the tool works and its goal clear, mostly to write the documentation.

4. **Compute an effective sample size?** These visualizations may enable me to compute an effective sample size but I am not clear on how to proceed. I was thinking of implementing a "leave-one-out" approach, removing one observation (or a set of observations) after the other, starting with those with smaller weights and rerunning the regression, extracting the variance of the resulting estimate and find the sample size for which the variance starts to substantially differ from the variance in the case of the full sample.
5. **A concrete example for diagnostic tests?** Decide whether to add one. I was considering adding a concrete (and brief) example of how to run the set of diagnostics I mention in the last section of the paper. I was thinking of running the battery of diagnosticcs for a study that has been published several decades ago, for instance Card (1993).
6. **Submit the paper** to the *American Economic Review: Insights*. Get rejected. Try somewhere else.

Chapter 4

Why Some Acute Health Effects of Air Pollution Could Be Inflated

I Notes

This project was joint with Léo Zabrocki-Hallak (RFF-CMC EIEE) [1]. I used a previous version of this paper as my MA thesis. Léo used a subsequent one as his job market paper. The project is now in its final stages. We submitted it to the Review of Economic Studies and got rejected. The editor, Katrine Løken, said: "While adding interesting evidence for the literature on health effects of air pollution, the literature on publication bias, weak instruments and selection bias is large and I do not see a novel and broad enough contribution in your study in more generally understanding new economic principles or stimulate new directions of research". The summary in Section III is a slightly modified version of the introduction of the paper. The paper can be found here. Details of our work, including our code, are described on the project's website.

II Abstract

Hundreds of studies show that air pollution affects health in the immediate short-run, and play a key role in setting air quality standards. Yet, estimated effect sizes sometimes vary widely across studies. Analyzing the results published in epidemiology and economics, we first find that a substantial share of estimates are likely to be inflated due publication bias and a lack of statistical power. Second, we run real data simulations to identify the design parameters causing these issues. We show that this exaggeration may be driven by the small number of exogenous shocks leveraged, by the limited strength of the instruments used or by sparse outcomes. These concerns and their determinants could extend to studies in other fields relying on comparable research designs. Our paper provides a principled workflow to evaluate and avoid the risk of exaggeration when conducting an observational study.

1. As mentioned previously, Léo left academia and from now on, I will work on this project alone. We both contributed to most tasks on this project. Léo however took care of the causal inference literature review while I did the epidemiology one.

III An introduction to the project

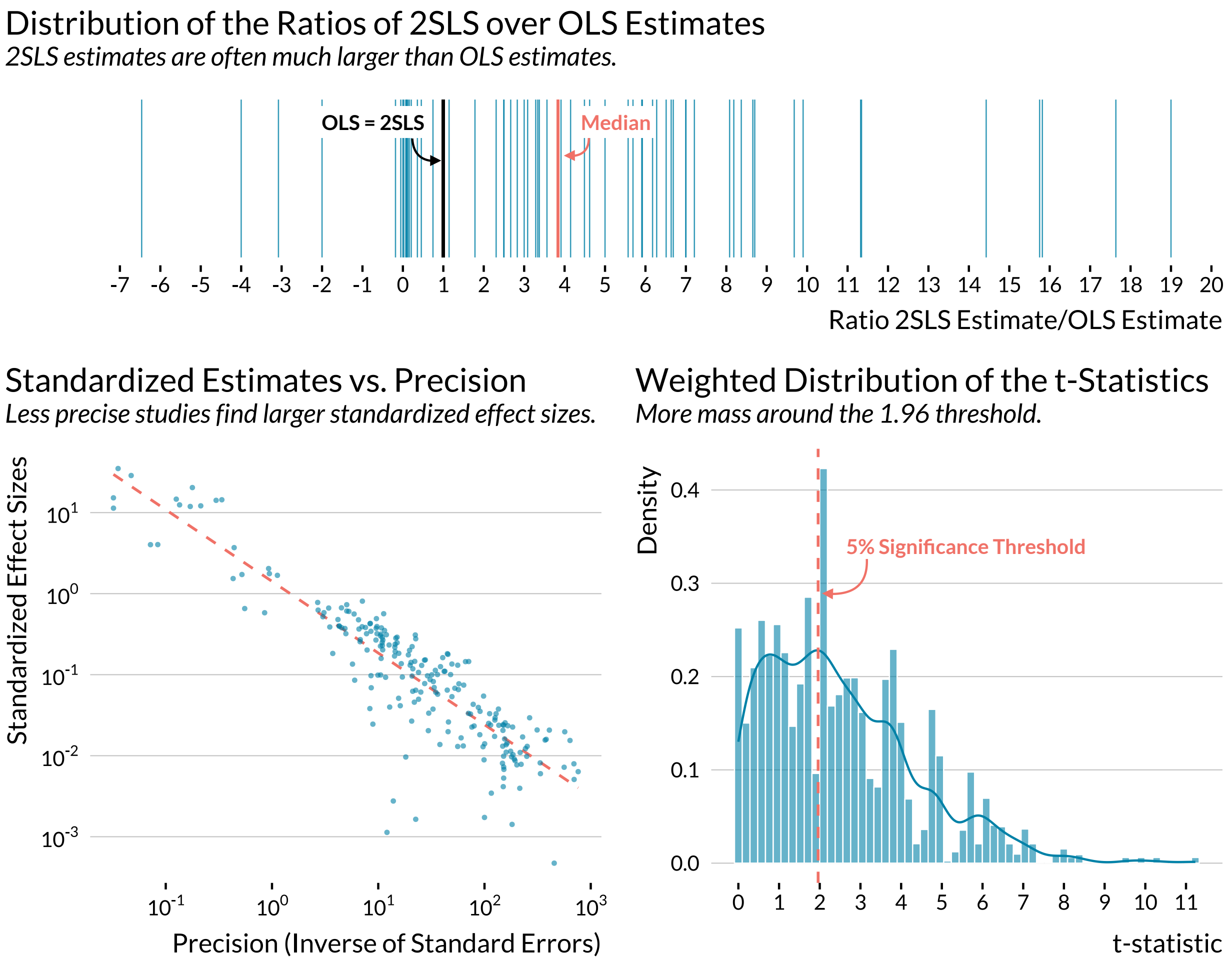
From extreme events such as the London Fog of 1952 to the development of sophisticated time-series analyses, a vast epidemiology literature of more than 600 studies has established that air pollution induces adverse health effects on the very short-term. Increases in the concentration of several ambient air pollutants have been found to be associated with small increases in daily mortality and emergency admissions for respiratory and cardiovascular causes (Schwartz 1994, Samet et al. 2000, Le Tertre et al. 2002, Bell et al. 2004, Liu et al. 2019). Based on these results, environmental protection and public health agencies have designed policies such as air quality alerts to mitigate the burden of air pollution. Obtaining accurate estimates is therefore crucial as they are directly used to implement and update policies.

With this objective in mind, researchers in economics and epidemiology have recently used causal inference methods to improve on the standard epidemiology literature that relied on associations (Dominici and Zigler 2017, Bind 2019). Newly obtained results confirm the short-term health effects of air pollution (Schwartz et al. 2015; 2018, Deryugina et al. 2019). Yet, causal estimates substantially differ from what would have been predicted by the standard epidemiology literature: they can be up to one order of magnitude larger. Reviewing the causal inference literature, we find that the median of the ratio of the obtained Two-Stage Least-Squares (2SLS) to the "naive" Ordinary Least-Squares (OLS) estimates is 3.8, as shown in the top panel of Figure 4.1. This discrepancy could arguably be explained by the fact that instrumental variable strategies remove omitted variable bias, reduce attenuation bias caused by classical measurement error in air pollution exposure or target a different causal estimand.

Our review of the literature however suggests an alternative but complementary explanation based on publication bias and low statistical power. The bottom left panel of Figure 4.1 reveals that large standardized effect sizes are only found in imprecise studies. This pattern is often indicative of publication bias. Among imprecise studies, those that found an effect size large enough to be statistically significant—at least 2 standard errors away from 0 at the 5% significance level—were more likely to be published. Studies with low precision therefore produce inflated estimates in the presence of publication bias (Ioannidis 2008, Gelman and Carlin 2014). The bottom right panel of Figure 4.1 confirms the presence of a publication bias in this literature. The mass of the t-statistics distribution is larger at the 5% statistical significance threshold (Brodeur et al. 2016; 2020).

The existence of power issues, publication bias and weak instrument problems has been underlined in many settings. However, exaggeration, one of their most dire consequence, has been much less commented on in the economic literature, despite the fact that it can lead published estimates to be far of true effect sizes. The determinants of these issues also remain understudied. While many other literatures do suffer from exaggeration, this issue is particularly salient in studies on the short-term health effects of air pollution since their signal-to-noise ratio is often low (Peng et al. 2006, Peng and Dominici 2008).

Figure 4.1 – Suggestive Evidence of Publication Bias, Power and Exaggeration Issues in the Causal Inference Literature.



Notes: In the top panel, we plot the ratios of the 2SLS estimate to the corresponding "naive" OLS one (same health outcome and air pollutant). Out of the 72 ratios obtained, we exclude 7 outliers with extremely large ratios that distort the graph. The orange line represents the median and is equal to 3.8. In the bottom left panel, we display 382 standardized effect sizes against the inverse of their standard error, a measure of precision. Both axes are on a log10 scale. In the bottom right panel, following Brodeur et al. (2020), we plot the weighted distribution of the 537 t-statistics. The weights are equal to the inverse of the number of tests displayed in the same table multiplied by the inverse of the number of tables in the article. The dashed orange line represents the 5% significance threshold.

In this paper, we analyze the consequences and determinants of low statistical power in studies on the short-term health effects of air pollution. We first tackle this question by gathering 2692 estimates from a unique corpus of 668 articles based on associations and 36 articles that rely on causal inference methods. For each of these studies, we run statistical power calculations to assess whether the design of the study would be robust enough to capture the true effect if it was smaller than the observed estimate (Gelman and Carlin 2014, Ioannidis et al. 2017, Lu et al. 2019, Timm et al. 2019). Yet, these calculations rely on hypotheses about the true effect of the treatment and do not enable to understand the causes of low power. Using real data from the US National Morbidity, Mortality, and Air Pollution Study (Samet et al. 2000), we therefore implement simulations to identify the characteristics of research designs that drive their statistical power and the inflation of statistically significant estimates. We finally provide a principled workflow to evaluate the risk of exaggeration along with a list of concrete recommendations to improve studies designs.

Our analysis suggests that a substantial share of estimates published in the epidemiology and causal inference literature could be inflated. Since Ioannidis et al. (2017) and Ferraro and Shukla (2020) find that half of the estimates published in economics and environmental economics are inflated by a factor of at least two, we evaluate the ability of the studies in our review to retrieve effects that would be twice as small as the obtained estimates. Reassuringly, a reasonable share of studies might not be prone to exaggeration. However, for a quarter of studies, estimated effect sizes could be at least inflated by a factor of 1.9. With better informed guesses of true effect sizes, we confirm that exaggeration issues occur in subsets of the two literatures.

Our simulation results help understand why some published estimates can be inflated. We first show that, regardless of the identification strategy used, even a large number of observations can cause important exaggeration issues. Regression discontinuity designs exploiting air quality alerts are particularly prone to sample size concerns. They are bound to produce inflated estimates since observations close to the air pollution threshold are often scarce. Second, we find that using rare exogenous shocks also produce inflated estimates. Even if studies leveraging public transportation strikes or thermal inversions as exogenous shocks can have large sample sizes, the number of shocks sometimes represents less than 1% of the observations in these studies. The variation available for identification is therefore small, leading to exaggeration, even for large true effect sizes. Third, we show that the count of cases of the outcome is a key driver of exaggeration. Estimated effects of air pollution on the elderly or children can be exaggerated since there are few daily hospital admissions or deaths for these groups.

By quantifying the respective influence of parameters affecting the power of studies, we fill an important gap in the literature on the acute health effects of air pollution. There was a lack of guidance on how to design an observational study to avoid low power issues, except for generalized additive models used in the standard epidemiology literature (Winquist et al. 2012). While our simulations focus on health effects, our conclusions could be extended to studies with similar designs but investigating the impacts of air pollution on different outcomes such as criminality, cognitive skills and productivity (Herrnstadt et al. 2021, Ebenstein et al. 2016, Adhvaryu et al. 2022). More broadly, we expect studies focusing on settings with small effect sizes, a limited number of exogenous shocks or of cases in the count outcome to also be subject to power and exaggeration issues.

Our paper makes three main contributions. First, it contributes to a growing literature assessing power issues in various fields (Ioannidis 2008, Gelman and Carlin 2014, Ioannidis et al. 2017, Ferraro and Shukla 2020, Stommes, Aronow and Sävje 2021, Arel-Bundock et al. 2022).

These meta-analyses help understand the recent replication crises in medicine, psychology and experimental economics (Button et al. 2013, Open Science Collaboration 2015, Camerer et al. 2018). Our analysis complements the literature by showing the existence of such issues for a major branch of health and environmental economics. The algorithm we developed to automatically review the epidemiology literature is readily available to evaluate power issues in other fields reporting point estimates and confidence intervals in plain text.

Second, existing meta-analyses do not usually discuss the determinants of the lack of power they describe. We overcome this key limitation by coupling our literature review with simulations. More generally, the drivers of low statistical power remain understudied in observational studies. To our knowledge, only three articles thoroughly address this critical question (Schell et al. 2018, Griffin et al. 2021, Black et al. 2022). These studies focus on event-study designs and treatment effects happening over medium to long time scales. We consider a different setting. Our simulations focus on shot-run effects in the context of high-frequency data. We analyze all research designs used in our literature: standard regression, reduced-form, instrumental variable and regression discontinuity designs.

Third, our study provides a reproducible workflow to evaluate and address power issues when running an observational study. Compared to psychology (Altoè et al. 2020), researchers in economics lack concrete recommendations to evaluate and understand the causes of low power issues. We suggest to build simulations using existing datasets before carrying out a study to evaluate whether it is likely to suffer from exaggeration issues. Once the analysis is completed, we recommend to run and report a retrospective power analysis to assess whether the design used would have recovered the true effect if it was in fact smaller than the one estimated. To ease the adoption of these tools, we make all replication and supplementary materials available on the project's website.

On top of these specific recommendations, we should not forget that published estimates only suffer from exaggeration in the presence of publication bias. The causal inference literature would therefore benefit from adopting a different view towards statistically insignificant results (Ziliak and McCloskey 2008, Wasserstein and Lazar 2016, McShane et al. 2019). It currently dichotomizes evidence according to the 5% significance threshold, disregarding non-significant results (Greenland 2017). Instead, if results were published regardless of their significance, the resulting distribution would be centered around the true effect (Hernán 2022). To replace the null hypothesis testing framework, we recommend to focus on confidence intervals and to interpret the range of effect sizes supported by the data (Amrhein et al. 2019, Romer 2020).

IV Next steps

1. **Make some arguments more general**, in particular in the introduction.
2. **Resubmit the paper**. We first need to decide where to resubmit it.

Abadie, A. (2020), 'Statistical Nonsignificance in Empirical Economics', American Economic Review: Insights **2**(2), 193–208. 14

Adhvaryu, A., Kala, N. and Nyshadham, A. (2022), 'Management and shocks to worker productivity', Journal of Political Economy **130**(1), 1–47. 23

Alm, J., Sennoga, E. and Skidmore, M. (2009), 'Perfect Competition, Urbanization, and Tax Incidence in the Retail Gasoline Market', Economic Inquiry **47**(1), 118–134. 9

Altoè, G., Bertoldo, G., Zandonella Callegher, C., Toffalini, E., Calcagnì, A., Finos, L. and Pastore, M. (2020), 'Enhancing Statistical Inference in Psychological Research via Prospective and Retrospective Design Analysis', Frontiers in Psychology **10**, 2893.
URL: *https://www.frontiersin.org/article/10.3389/fpsyg.2019.02893/full* 24

Amrhein, V., Trafimow, D. and Greenland, S. (2019), 'Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don't Expect Replication', The American Statistician **73**(sup1), 262–270. 24

Andrews, I. and Kasy, M. (2019), 'Identification of and Correction for Publication Bias', American Economic Review **109**(8), 2766–2794. 14

Arel-Bundock, V., Briggs, R. C., Doucouliagos, H., Mendoza Aviña, M. and Stanley, T. (2022), Quantitative Political Science Research is Greatly Underpowered, preprint, Open Science Framework.
URL: *https://osf.io/7vy2f* 23

Aronow, P. M. and Samii, C. (2016), 'Does Regression Produce Representative Estimates of Causal Effects?', American Journal of Political Science **60**(1), 250–267. 18

Bell, M. L., Samet, J. M. and Dominici, F. (2004), 'Time-series studies of particulate matter', Annu. Rev. Public Health **25**, 247–280. 21

Bind, M.-A. (2019), 'Causal Modeling in Environmental Health', Annual Review of Public Health **40**(1), 23–43.
URL: *https://www.annualreviews.org/doi/10.1146/annurev-publhealth-040218-044048* 21

Black, B., Hollingsworth, A., Nunes, L. and Simon, K. (2022), 'Simulated power analyses for observational studies: An application to the Affordable Care Act Medicaid expansion', Journal of Public Economics **213**, 104713.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0047272722001153* 24

Black, B. S., Hollingsworth, A., Nunes, L. and Simon, K. I. (2021), Simulated Power Analyses for Observational Studies: An Application to the Affordable Care Act Medicaid Expansion, SSRN Scholarly Paper ID 3368187, Social Science Research Network, Rochester, NY. 16, 17

Boyer, P. C., Delemotte, T., Gauthier, G., Rollet, V. and Schmutz, B. (2020a), 'Les déterminants de la mobilisation des Gilets jaunes', Revue economique **Vol. 71**(1), 109–138. 8, 11

Boyer, P., Delemotte, T., Gauthier, G., Rollet, V. and Schmutz, B. (2020b), The Gilets Jaunes: Offline and Online, SSRN Scholarly Paper ID 3612849, Social Science Research Network, Rochester, NY. 11

Boykoff, M. (2007), 'Media coverage of climate change: Current trends, strengths, weaknesses', Human Development Report p. 54. 1

Brodeur, A., Cook, N. and Heyes, A. (2020), 'Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics', American Economic Review **110**(11), 3634–3660. 14, 17, 21, 22

Brodeur, A., Lé, M., Sangnier, M. and Zylberberg, Y. (2016), 'Star Wars: The Empirics Strike Back', American Economic Journal: Applied Economics **8**(1), 1–32.
URL: *https://pubs.aeaweb.org/doi/10.1257/app.20150044* 21

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J. and Munafò, M. R. (2013), 'Power failure: Why small sample size undermines the reliability of neuroscience', Nature Reviews Neuroscience **14**(5), 365–376. 14, 24

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M. and Wu, H. (2016), 'Evaluating replicability of laboratory experiments in economics', Science **351**(6280), 1433–1436. 14, 17

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E.-J. and Wu, H. (2018), 'Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015', Nature Human Behaviour **2**(9), 637–644.
URL: *http://www.nature.com/articles/s41562-018-0399-z* 24

Card, D. (1993), Using Geographic Variation in College Proximity to Estimate the Return to Schooling, Working Paper 4483, National Bureau of Economic Research. 19

Chang, A. C. and Li, P. (2022), 'Is Economics Research Replicable? Sixty Published Papers From Thirteen Journals Say "Often Not"', Critical Finance Review **11**. 14

Chouinard, H. and Perloff, J. M. (2004), 'Incidence of federal and state gasoline taxes', Economics Letters **83**(1), 55–60. 9

Chouinard, H. and Perloff, J. M. (2007), 'Gasoline Price Differences: Taxes, Pollution Regulations, Mergers, Market Power, and Market Conditions', The B.E. Journal of Economic Analysis & Policy **7**(1). 9

Christensen, G. and Miguel, E. (2018), 'Transparency, Reproducibility, and the Credibility of Economics Research', Journal of Economic Literature **56**(3), 920–980. 17

Deaton, A. and Cartwright, N. (2018), 'Understanding and misunderstanding randomized controlled trials', Social Science & Medicine **210**, 2–21. 17

Deryugina, T., Heutel, G., Miller, N. H., Molitor, D. and Reif, J. (2019), 'The Mortality and Medical Costs of Air Pollution: Evidence from Changes in Wind Direction', American Economic Review **109**(12), 4178–4219.
URL: *https://pubs.aeaweb.org/doi/10.1257/aer.20180279* 21

Dominici, F. and Zigler, C. (2017), 'Best Practices for Gauging Evidence of Causality in Air Pollution Epidemiology', American Journal of Epidemiology **186**(12), 1303–1309.
URL: *https://academic.oup.com/aje/article/186/12/1303/4105921* 21

Douenne, T. and Fabre, A. (2022), 'Yellow Vests, Pessimistic Beliefs, and Carbon Tax Aversion', American Economic Journal: Economic Policy **14**(1), 81–110. 7

Ebenstein, A., Lavy, V. and Roth, S. (2016), 'The Long-Run Economic Consequences of High-Stakes Examinations: Evidence from Transitory Variation in Pollution', American Economic Journal: Applied Economics **8**(4), 36–65.
URL: *https://pubs.aeaweb.org/doi/10.1257/app.20150213* 23

Erutku, C. (2019), 'Carbon pricing pass-through: Evidence from Ontario and Quebec's wholesale gasoline markets', Energy Policy **132**, 106–112. 9

Ferrara, E. L., Chong, A. and Duryea, S. (2012), 'Soap Operas and Fertility: Evidence from Brazil', American Economic Journal: Applied Economics **4**(4), 1–31. 3

Ferraro, P. J. and Shukla, P. (2020), 'Feature—Is a Replicability Crisis on the Horizon for Environmental and Resource Economics?', Review of Environmental Economics and Policy **14**(2), 339–351.
URL: *https://www.journals.uchicago.edu/doi/10.1093/reep/reaa011* 23

Gelman, A. (2020), Regression and Other Stories, Cambridge University Press, Cambridge New York, NY Port Melbourne, VIC New Delhi Singapore. 16

Gelman, A. and Carlin, J. (2014), 'Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors', Perspectives on Psychological Science **9**(6), 641–651. 14, 16, 21, 22, 23

Greenland, S. (2017), 'Invited Commentary: The Need for Cognitive Science in Methodology', American Journal of Epidemiology **186**(6), 639–645.
URL: *https://academic.oup.com/aje/article/186/6/639/3886035* 24

Griffin, B. A., Schuler, M. S., Stuart, E. A., Patrick, S., McNeer, E., Smart, R., Powell, D., Stein, B. D., Schell, T. and Pacula, R. L. (2021), 'Moving beyond the classic difference-in-differences model: A simulation study comparing statistical methods for estimating effectiveness of state-level policies', arXiv:2003.12008 [stat] . 24

Harju, J., Kosonen, T., Laukkanen, M. and Palanne, K. (2022), 'The heterogeneous incidence of fuel carbon taxes: Evidence from station-level data', Journal of Environmental Economics and Management p. 102607. 9, 11

Heal, G. and Kunreuther, H. (2010), 'Social Reinforcement: Cascades, Entrapment, and Tipping', American Economic Journal: Microeconomics **2**(1), 86–99. 4

Hernán, M. A. and Robins, J. M. (2020), Causal Inference: What If, boca raton: chapman & hall/crc edn. 17

Hernán, M. A. (2022), 'Causal analyses of existing databases: no power calculations required', Journal of Clinical Epidemiology **144**, 203–205.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0895435621002730* 24

Herrnstadt, E., Heyes, A., Muehlegger, E. and Saberian, S. (2021), 'Air Pollution and Criminal Activity: Microgeographic Evidence from Chicago', American Economic Journal: Applied Economics **13**(4), 70–100.
URL: *https://pubs.aeaweb.org/doi/10.1257/app.20190091* 23

Hughes, J. E., Knittel, C. R. and Sperling, D. (2008), 'Evidence of a Shift in the Short-Run Price Elasticity of Gasoline Demand', The Energy Journal **29**(1), 113–134. 9

IEA (2015), 'Statistics by country', http://www.iea.org/statistics/statisticssearch/. 9

Imbens, G. and Kalyanaraman, K. (2012), 'Optimal Bandwidth Choice for the Regression Discontinuity Estimator', The Review of Economic Studies **79**(3), 933–959. 17

Ioannidis, J. P. A. (2008), 'Why Most Discovered True Associations Are Inflated', Epidemiology **19**(5), 640–648. 14, 21, 23

Ioannidis, J. P. A., Stanley, T. D. and Doucouliagos, H. (2017), 'The Power of Bias in Economics Research', The Economic Journal **127**(605), F236–F265. 14, 17, 22, 23

Kasy, M. (2021), 'Of Forking Paths and Tied Hands: Selective Publication of Findings, and What Economists Should Do about It', Journal of Economic Perspectives **35**(3), 175–192. 17

Kaufmann, R. K. (2019), 'Pass-through of motor gasoline taxes: Efficiency and efficacy of environmental taxes', Energy Policy **125**, 207–215. 9

Lade, G. E. and Bushnell, J. (2019), 'Fuel Subsidy Pass-Through and Market Structure: Evidence from the Renewable Fuel Standard', Journal of the Association of Environmental and Resource Economists **6**(3), 563–592. 9

Le Tertre, A., Medina, S., Samoli, E., Forsberg, B., Michelozzi, P., Boumghar, A., Vonk, J., Bellini, A., Atkinson, R., Ayres, J. et al. (2002), 'Short-term effects of particulate air pollution on cardiovascular diseases in eight european cities', Journal of Epidemiology & Community Health **56**(10), 773–779. 21

Liu, C., Chen, R., Sera, F., Vicedo-Cabrera, A. M., Guo, Y., Tong, S., Coelho, M. S., Saldiva, P. H., Lavigne, E., Matus, P., Valdes Ortega, N., Osorio Garcia, S., Pascal, M., Stafoggia, M., Scortichini, M., Hashizume, M., Honda, Y., Hurtado-Díaz, M., Cruz, J., Nunes, B., Teixeira, J. P., Kim, H., Tobias, A., Íñiguez, C., Forsberg, B., Åström, C., Ragettli, M. S., Guo, Y.-L., Chen, B.-Y., Bell, M. L., Wright, C. Y., Scovronick, N., Garland, R. M., Milojevic, A., Kyselý, J., Urban, A., Orru, H., Indermitte, E., Jaakkola, J. J., Ryti, N. R., Katsouyanni, K., Analitis, A., Zanobetti, A., Schwartz, J., Chen, J., Wu, T., Cohen, A., Gasparrini, A. and Kan, H. (2019), 'Ambient Particulate Air Pollution and Daily Mortality in 652 Cities', New England Journal of Medicine **381**(8), 705–715.
URL: *http://www.nejm.org/doi/10.1056/NEJMoa1817364* 21

Lu, J., Qiu, Y. and Deng, A. (2019), 'A note on Type S/M errors in hypothesis testing', British Journal of Mathematical and Statistical Psychology **72**(1), 1–17. 14, 22

Marion, J. and Muehlegger, E. (2011), 'Fuel tax incidence and supply conditions', Journal of Public Economics **95**(9), 1202–1212. 9

McShane, B. B., Gal, D., Gelman, A., Robert, C. and Tackett, J. L. (2019), 'Abandon Statistical Significance', The American Statistician **73**(sup1), 235–245. 24

MTES (2017), 'Fiscalité carbone', https://www.ecologique-solidaire.gouv.fr/fiscalite-carbone. 10

Open Science Collaboration (2015), 'Estimating the reproducibility of psychological science', Science **349**(6251), aac4716. 14, 24

Peng, R. D. and Dominici, F. (2008), 'Statistical methods for environmental epidemiology with r', R: a case study in air pollution and health . 21

Peng, R. D., Dominici, F. and Louis, T. A. (2006), 'Model choice in time series studies of air pollution and mortality', Journal of the Royal Statistical Society: Series A (Statistics in Society) **169**(2), 179–203. 21

Ravallion, M. (2020), Should the Randomistas (Continue to) Rule?, Working Paper 27554, National Bureau of Economic Research. 17

Romer, D. (2020), In praise of confidence intervals, in 'AEA Papers and Proceedings', Vol. 110, pp. 55–60. 24

Rosenthal, R. (1979), 'The file drawer problem and tolerance for null results', Psychological Bulletin **86**(3), 638–641. 14

Samet, J. M., Zeger, S. L., Dominici, F., Curriero, F., Coursac, I., Dockery, D. W., Schwartz, J. and Zanobetti, A. (2000), 'The national morbidity, mortality, and air pollution study', Part II: morbidity and mortality from air pollution in the United States Res Rep Health Eff Inst **94**(pt 2), 5–79. 21, 23

Schell, T. L., Griffin, B. A. and Morral, A. R. (2018), Evaluating Methods to Estimate the Effect of State Laws on Firearm Deaths: A Simulation Study, Technical report, RAND Corporation. 17, 24

Schmidt, A., Ivanova, A. and Schäfer, M. S. (2013), 'Media attention for climate change around the world: A comparative analysis of newspaper coverage in 27 countries', Global Environmental Change **23**(5), 1233–1248. 1

Schwartz, J. (1994), 'What are people dying of on high air pollution days?', Environmental research **64**(1), 26–35. 21

Schwartz, J., Austin, E., Bind, M.-A., Zanobetti, A. and Koutrakis, P. (2015), 'Estimating Causal Associations of Fine Particles With Daily Deaths in Boston: Table 1.', American Journal of Epidemiology **182**(7), 644–650.
URL: *https://academic.oup.com/aje/article-lookup/doi/10.1093/aje/kwv101* 21

Schwartz, J., Fong, K. and Zanobetti, A. (2018), 'A National Multicity Analysis of the Causal Effect of Local Pollution, NO2, and PM2.5 on Mortality', Environmental Health Perspectives **126**(8), 087004.
URL: *https://ehp.niehs.nih.gov/doi/10.1289/EHP2732* 21

Stolper, S. (2018), 'Local Pass-Through and the Regressivity of Taxes: Evidence from Automotive Fuel Markets', p. 51. 9, 10

Stommes, D., Aronow, P. M. and Sävje, F. (2021), 'On the reliability of published findings using the regression discontinuity design in political science', arXiv:2109.14526 [stat] . 16, 17

Stommes, D., Aronow, P. M. and Sävje, F. (2021), 'On the reliability of published findings using the regression discontinuity design in political science'. 23

Timm, A., Gelman, A. and Carlin, J. (2019), 'Retrodesign: Tools for Type S (Sign) and Type M (Magnitude) Errors'. 22

Wasserstein, R. L. and Lazar, N. A. (2016), 'The ASA Statement on p -Values: Context, Process, and Purpose', The American Statistician **70**(2), 129–133. 24

Winquist, A., Klein, M., Tolbert, P. and Sarnat, S. E. (2012), 'Power estimation using simulations for air pollution time-series studies', Environmental Health **11**(1), 1–12. 23

Young, A. (2021), 'Leverage, Heteroskedasticity and Instrumental Variables in Practical Application', p. 43. 17

Ziliak, S. T. and McCloskey, D. N. (2008), The cult of statistical significance: how the standard error costs us jobs, justice, and lives, Economics, cognition, and society, University of Michigan Press, Ann Arbor. OCLC: ocn168717577. 24

Zwet, E. W. and Cator, E. A. (2021), 'The significance filter, the winner's curse and the need to shrink', Statistica Neerlandica **75**(4), 437–452. 14