ÉCOLE DES HAUTES ÉTUDES EN SCIENCES SOCIALES

ÉCOLE DOCTORALE: ED 465 – ÉCONOMIE PANTHÉON SORBONNE
PARIS SCHOOL OF ECONOMICS

DOCTORAT

DISCIPLINE: SCIENCES ÉCONOMIQUES

Léo ZABROCKI

Improving the Design of Studies on the Acute Health Effects of Air Pollution

---

Thèse dirigée par Hélène OLLIVIER

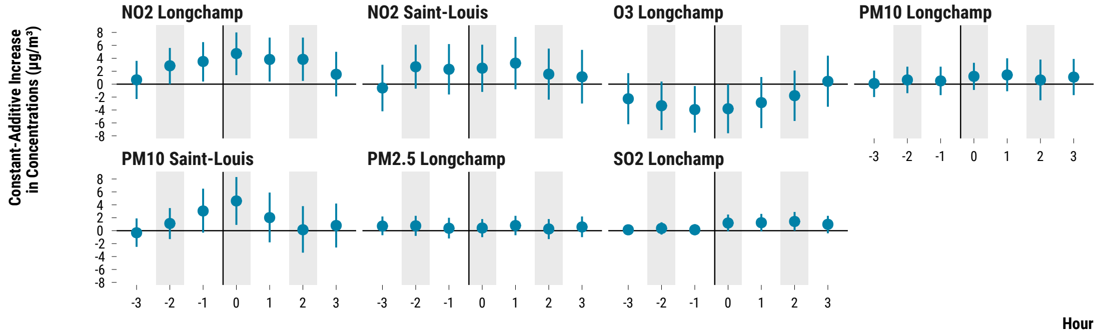
Date de soutenance : le 17 juin 2022

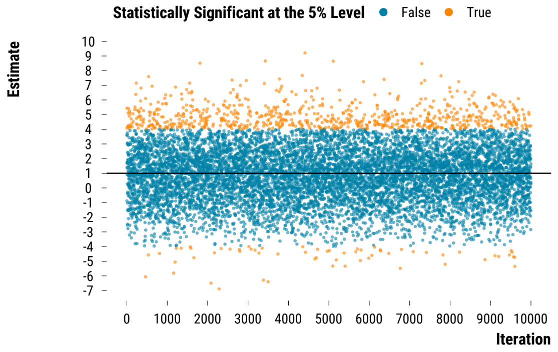
Rapporteurs	1	Michela BACCINI	Professeure à l'Université de Florence
	2	Tatyana DERYUGINA	Professeure à l'Université de l'Illinois à Urbana-Champaign
Jury	1	Sylvain CHABÉ-FERRET	Chargé de recherche à l'INRAE et à TSE
	2	Clément DE CHAISEMARTIN	Professeur à Sciences Po
	3	Hélène OLLIVIER	Chargée de recherche au CNRS et à PSE
	4	Katheline SCHUBERT	Professeure à l'Université Paris 1 et PSE

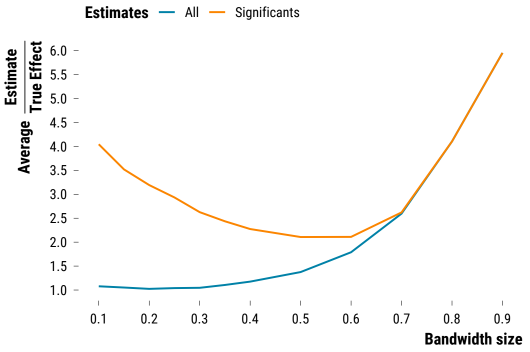
LÉO ZABROCKI

PHD THESIS

IMPROVING THE DESIGN OF STUDIES ON THE ACUTE HEALTH EFFECTS OF AIR POLLUTION







PARIS SCHOOL OF ECONOMICS - EHESS

Contents

I can live with doubt and uncertainty and not knowing. I think it is much more interesting to live not knowing than to have answers that might be wrong. If we will only allow that, as we progress, we remain unsure, we will leave opportunities for alternatives. We will not become enthusiastic for the fact, the knowledge, the absolute truth of the day, but remain always uncertain ... In order to make progress, one must leave the door to the unknown ajar.

— Richard Phillips Feynman, physicist

Whatever else is done about null-hypothesis tests, let us stop viewing statistical analysis as a sanctification process. We are awash in a sea of uncertainty, caused by a flood tide of sampling and measurement errors, and there are no objective procedures that avoid human judgment and guarantee correct interpretations of results.

— Robert Paul Abelson, psychologist

You can't stand on the beach of the sea of uncertainty with the waves lapping at your ankles. You have to jump into the sea and stick your head underwater and blow some bubbles.

— Andrew Gelman, statistician

Remerciements

En cette fin de thèse, je peux enfin remercier publiquement Hélène Ollivier d'avoir été une brillante et excellente directrice de thèse. Merci en particulier à Sylvie Lambert de m'avoir incité à rencontrer Hélène. Cette thèse te doit beaucoup Hélène: tes nombreux commentaires et critiques ont grandement contribué à améliorer sa qualité. Merci également pour ton soutien pendant le job market qui me permet ainsi de commencer ma carrière dans de bonnes conditions à Milan.

I feel very privileged to have gathered in my PhD jury people whose work I greatly admire. I would like to warmly thank Michela Baccini and Tatyana Deryugina for accepting to be the referees of my PhD thesis. Your remarks during the pre-defense really helped improve all chapters of the dissertation. Je tiens également à remercier Sylvain Chabé-Ferret, Clément de Chaisemartin et Katheline Schubert d'être membres de mon jury de thèse. Un très grand merci Sylvain d'avoir tout de suite répondu présent pour discuter de nos projets avec Vincent et pour ton soutien au moment du job market. Merci également à Katheline d'avoir suivi l'avancée de cette thèse et de m'avoir encouragé au cours de ces quatre années. Enfin, merci à Clément de Chaisemartin dont je suis impatient d'avoir les retours.

Cette thèse n'aurait pas pu aboutir sans le soutien de mes co-auteurs. Merci tout d'abord à Marie-Abèle Bind qui a été constamment à mes côtés depuis mon master 2 jusqu'à ma recherche d'un poste pour l'après thèse. Ma vision de la recherche sur le matching et la randomization inference a été en grande partie influencée par nos échanges. J'espère aussi que l'enthousiasme que tu m'as apporté tout au long de la thèse se ressent à la lecture ces pages. Merci également à Tarik Benmarhnia qui me suit depuis le début de ma thèse et qui, lui aussi, a toujours répondu présent. Ton éclectisme sur les méthodes d'inférence causale, ta productivité et ta bienveillance constante sont une source d'inspiration pour moi. Je tiens aussi à remercier Anna Alari qui a été présente aux bons moments et qui a toujours su me remotiver. Mes co-auteurs doctorants m'auront aussi beaucoup appris. Grâce à Marion Leroutier, je suis devenu plus rigoureux, efficace et j'ai compris qu'il ne fallait pas perdre l'intérêt d'un projet de recherche au-delà de considérations méthodologiques. Merci d'avoir permis à ce projet marseillais de garder le cap et d'avoir été

présente pendant les périodes de confinement. Enfin, il me faut remercier tout particulièrement Vincent Bagilet qui est devenu un ami cher après deux années de recherche en commun. Vincent m'aura beaucoup influencé en termes de programmation, de reproductibilité et aussi de création de memes. J'ai hâte de savoir ce que nous réserve l'avenir pour nos collaborations futures.

Cette thèse s'est déroulée dans de très bonnes conditions matérielles et financières grâce à PSE et l'EHESS, ainsi qu'à l'ENS qui a financé mon contrat doctoral. Merci aussi à Marc Gurgand de m'avoir permis de financer ma quatrième année de thèse en étant ATER au département d'économie de l'École Normale Supérieure. Laurence Vincent aura été un rayon de ce soleil durant cette dernière année intense. Je tiens également à remercier Pascale Combemale qui m'a permis d'enseigner au CPES de PSL des cours aux contenus pédagogiques novateurs. Ces 200 heures d'enseignement auront fait de moi un meilleur chercheur en m'apprenant qu'il y a une grande différence entre connaître le nom d'une technique, la comprendre et savoir l'appliquer.

Je tiens aussi à remercier Adrien, Benjamin, Celia, Martin, Milena, Nicolas, Paul et Thiago pour avoir rendu la vie à PSE plus agréable et drôle. Je remercie tout particulièrement Quentin Lippmann et Georgia Thebault. Quentin a été un véritable mentor et ami pendant ces 4 ans. Merci d'avoir été toujours disponible pour moi. Je n'aurais certainement pas continué en thèse à PSE sans Georgia, qui m'accompagne dans la vie depuis 2014. Je sais la chance que j'ai de t'avoir à mes côtés.

Mes derniers remerciements vont à ma belle-famille et à ma famille. Merci à Christine et Bruno de m'avoir soutenu matériellement et émotionnellement depuis tant d'années. Leur présence au quotidien m'a apporté beaucoup de maturité et de bonheur, tout comme celles de leurs enfants et compagnons de vie respectifs : Chloé et Alain, Alice et Haixia, Benjamin et Aurélie, Samuel et Gretchen. Mes deux petits frères Simon et Merlin m'ont apporté également beaucoup de joie: je suis fier de leurs parcours respectifs et des adultes qu'ils sont devenus. Merci à mes parents, Carole et Stephan, pour leur soutien inconditionnel et les valeurs qu'ils m'ont transmises. Après dix années d'études supérieures, cette thèse se doit de leur être dédiée.

Enfin, merci à Camille pour tout son amour.

Léo Zabrocki, Wittenheim, mai 2022

A Very Short Introduction to My Thesis

In the first section of this introduction, I explain how my thesis fits into the literature on the acute health effects of air pollution and how it could help improve the design of studies. In the two following sections, I briefly summarize the research questions, method and results of the four chapters of the thesis.

CONTACT:
Léo Zabrocki
PhD candidate
PSE - EHESS
leo.zabrocki@psemail.eu
https://lzabrocki.github.io/

Design Trumps Analysis

From extreme events such as the London Fog of 1952 to the development of sophisticated time-series analyses, a vast scientific literature in epidemiology has established that air pollution induces adverse health effects on the very *short-term* (Schwartz 1994, Samet et al. 2000, Le Tertre et al. 2002, Bell and Davis 2001, Bell et al. 2004, Samoli et al. 2008). Increases in the concentration of several ambient air pollutants such as fine particulate matter ($PM_{2.5}$) or nitrogen dioxide (NO_2) have been found to be associated with small relative increases in daily mortality and emergency admissions for respiratory and cardiovascular causes (Samet et al. 2000, Shah et al. 2015, Orellano et al. 2020). In the most recent large-scale study based on data from 625 studies around the world, Liu et al. (2019) find that an increase of 10 $\mu g/m^3$ in the 2-day moving average of $PM_{2.5}$ concentration is associated with a 0.74% relative increase in daily respiratory mortality (95% CI, 0.53 to 0.95)[1]. The results of the literature and their replications have played a crucial role in strengthening air pollution regulations (Bell et al. 2004).

Despite this success, researchers have been very careful to restrain from qualifying the estimated dose-responses as causal relationships (Wang et al. 2012, Gutman et al. 2012). Most of the literature in epidemiology is not based on a well-defined causal framework and has relied on time-series Poisson generalized additive models and case-crossover designs (Jaakkola 2003, Lu and Zeger 2007, Peng and Dominici 2008, Bhaskaran et al. 2013). The situation however changed in the last decade when researchers in economics and epidemiology have revisited the question with causal inference methods based on the framework of the Neyman-Rubin Causal Model (Holland 1986, Zigler and Dominici 2014, Dominici and Zigler 2017, Bind 2019). Compared to the previous literature, these new studies aim to overcome the biases due to unmeasured confounding and measurement errors in air pollution exposure. To do so, a wide range of natural

---

[1] To make sense of these figures, we can take Paris as an example. A 10 $\mu g/m^3$ increase in $PM_{2.5}$ is equivalent to one standard deviation increase in the concentration of this air pollutant, whose average daily concentration is equal to 16 $\mu g/m^3$. On average, 12 individuals die each day from circularly-respiratory causes in Paris *intra-muros*. If a 0.74% increase in daily mortality could seem to be a small effect, it would represent about 320 deaths attributable to air pollution over a decade. This is equivalent to the number of people dying in road accidents in Paris over a decade. Besides, if we extrapolate these figures to an entire country, the burden due to the short-term effects of air pollution are worrisome.

experiments have been exploited, from meteorological phenomena such as changes in wind patterns or thermal inversions (Schwartz et al. 2015; 2018, Arceo et al. 2016, Jans et al. 2018, Deryugina et al. 2019), extreme events like forest fires or volcanic eruptions (Sheldon and Sankaran 2017, Halliday et al. 2019) to variations in the intensity of modes of transports (Moretti and Neidell 2011, Schlenker and Walker 2016, Bauernschuster et al. 2017, Godzinski et al. 2019, Giaccherini et al. 2021). Newly obtained results confirm the acute health effects of air pollution. In economics, the literature is now moving to investigate other morbidity outcomes such as sickness leaves (Holub et al. 2020), cognitive abilities (Ebenstein et al. 2016), workers' productivity (Chang et al. 2016; 2019, He et al. 2019), ... and even criminal activities (Burkhardt et al. 2019, Bondy et al. 2020, Herrnstadt et al. 2021).

If these studies clearly shine by finding very credible sources of exogenous variation in air pollution, there could still be room for their designs to be further improved. King and Zeng (2006) provide a very convenient decomposition of bias in observational studies that help reflect on potential areas of improvement. The difference between an estimate and the true value of the causal estimand can be decomposed into four types of biases:

$$\begin{aligned}\text{Bias} = \;& \text{Omitted Variable Bias} \quad + \\ & \text{Post-Treatment Bias} \quad + \\ & \text{Interpolation Bias} \quad + \\ & \text{Extrapolation Bias}\end{aligned}$$

The new literature based on causal inference methods has made great efforts to reduce the first component of bias caused by unmeasured confounding. The second term, *Post-Treatment Bias*, happens if we adjust for variables that are also influenced by the treatment of interest. It could happen if we adjust for an health outcome that is also affected by air pollution. For most the literature, it does not seem to be an issue. The two other terms, *Interpolation Bias* and *Extrapolation Bias* are the two areas where the design of studies could be improved. These two types of bias have been overlooked since researchers, especially in economics, have relied heavily on parametric multivariate regression models (Angrist and Pischke 2008). If the source of exogenous variation in air pollution is really quasi-random, the use of simple multivariate regression models should not be problematic. Yet, in many studies, additional covariates adjustments are required to make the as-if random assumption more plausible. The interpolation bias arises when we fail to adjust for confounding variables with the correct functional forms. This is a key issue in the literature since it can be hard to guess how to adjust for weather parameters or seasonal effects. The imbalance in covariates can therefore make results depend on the specification of the model (Koop and Tole 2004). Even if several models are reported in articles, the full universe of plausible specifications can never be reported. The issue of covariate imbalance becomes more worrying when there is a lack of overlap in their distributions. If non-similar units are not discarded, results rely on the ability of the model to correctly impute missing potential outcomes of units without empirical counterfactuals. If that is not the case, an extrapolation bias occurs. This is also an important issue for the literature since many studies are based on rare exogenous shocks. A small number of treated units are then compared to a large number of control units which often have very different covariates values.

To overcome these two biases, leading statisticians have advocated using matching methods to pre-process the data for a long time (Rosenbaum and Rubin 1983, Rosenbaum 2002, Rubin 2008, Rosenbaum 2010, Imbens 2015, Imbens and Rubin 2015). First, matching adjusts nonparametrically for observed covariates, which reduces the interpolation bias. Second, it reveals the common support of the data and by discarding units without empirical counterfactuals, limit the extrapolation bias. Once a balanced sample is found, the sensitivity of the analysis to the specification of the statistical model is reduced (Ho et al. 2007). On top of these advantages, matching is a principled approached to split the study between a design stage where we do not look at the outcomes of interest and an analysis stage where we run the statistical model on a balanced sample. The imputation of missing potential outcomes is also clearer than a regression-based approach. There are currently few studies based on matching methods in the literature but they provide a clear template to implement them (Baccini et al. 2017, Forastiere et al. 2020, Sommer et al. 2021). The first two chapters of my thesis revisit with matching two empirical strategies found in the literature: the use of wind patterns and maritime traffic as exogenous sources of variation in air pollution. We show that the common support of the data required to overcome interpolation and extrapolation biases is actually very small. We also complement the robustness of our results with underused techniques to quantify the bias due to an unmeasured confounder (Rosenbaum 2010, Fogarty 2020) and compute the uncertainty of estimates with alternative modes of inference (Neyman 1923, Fisher et al. 1937, Rubin 1991).

If the bias decomposition of King and Zeng (2006) is useful to reflect on the strengths and weaknesses of study designs in the literature, I have often found that it could not fully explained the discrepancy in effect size magnitudes between studies from the traditional epidemiology literature and new ones based on causal inference methods. Causal estimates are often larger. Even for studies based on the same causal inference method, effect sizes seem to be vary a lot. Three explanations could be advanced for this observed difference. First, non-causal estimates could suffer from omitted variable bias and attenuation bias due to measurement error in air pollution exposure. Second, large shocks in air pollution sometimes happen in natural experiments leading to significant effects on health outcomes. Third, causal studies exploit different exogenous shocks in different places for different health outcomes. Replications of the same design across similar contexts are still rare, making it harder to draw comparisons. In my thesis, I propose an alternative and complementary reason based on the seminal works of Ioannidis (2008) and Gelman and Carlin (2014). The causal inference literature, which is mostly published in economics, relies very strongly on the null hypothesis significance testing framework (NHST) (Fisher et al. 1937). Studies focus on rejecting the null hypothesis of no effect and interpreting the statistical significance and the effect size of estimates. This research practice is however pernicious when studies are under-powered and a publication bias favors statistically significant estimates. The consequence is that estimates of under-powered studies must be large to pass the statistical significance filter: they exaggerate the true effect sizes of the causal estimands of interest and therefore are misleading. As the signal-to-noise ratio for estimates on the acute health effects is known to be low (Bell et al. 2004, Peng and Dominici 2008), it seemed necessary to me to explore this underrated argument. In the third chapter of the thesis, we provide the first evidence that low statistical power and the inflation of estimates are real issues in the literature. We also show which parameters of a research design affect statistical power, leading to concrete recommendations for improving studies. The fourth chapter generalizes the findings of the previous chapter. All causal inference methods reduce the variation in the treatment to overcome unmeasured confounding. In some cases, it could lead to a loss in statistical power, and thereby an inflation of statistically significant estimates. Given the rising concerns on research practices based on the NHST (Ziliak and McCloskey 2008, Brodeur et al. 2016; 2020, McShane et al. 2019, Romer 2020), this trade-off between statistical power and confounding should be better taken into account in applied research.

Unmasking Interpolation & Extrapolation Biases

The first chapter of the thesis is entitled *Improving the Design Stage of Air Pollution Studies Based on Wind Patterns* and co-authored with Anna Alari (ISGlobal) and Tarik Benmarhnia (UCSD). A growing literature in economics and epidemiology has exploited changes in wind patterns as a source of exogenous variation to better measure the acute health effects of air pollution. As an alternative to current practices and to gauge the extent of these issues, we propose to implement a causal inference pipeline to embed this type of observational study within an hypothetical randomized experiment. We illustrate this approach using daily data from Paris, France, over the 2008-2018 period. Using the Neyman-Rubin potential outcomes framework, we first define the treatment of interest as the effect of North-East winds on particulate matter concentrations compared to the effects of other wind directions. We then implement a matching algorithm to approximate a pairwise randomized experiment. It adjusts nonparametrically for observed confounders while avoiding model extrapolation by discarding treated days without similar control days. We find that the effective sample size for which treated and control units are comparable is surprisingly small. The precision of estimates is therefore traded for a reduction in bias. It is however reassuring that results on the matched sample are consistent with a standard regression analysis of the initial data, even if the two samples are not directly comparable. We finally carry out a quantitative bias analysis to check whether our results could be altered by an unmeasured confounder: estimated effects seem robust to a relatively large hidden bias. Th approach we developed in this chapter could be relevant for similar strategies based on binary instruments such as thermal inversions or public transport strikes.

The second chapter of the thesis is entitled why *Estimating the Local Air Pollution Impacts of Cruise Traffic: A Principled Approach for Observational Data* and co-authored with Marion Leroutier (Misum, Stockholm School of Economics) and Marie-Abèle Bind (Biostatistics Center, Massachusetts General Hospital). The air pollution and health effects of cruise vessel traffic is a growing concern in the Mediterranean area. We propose a novel methodology based on high-frequency observational data to estimate the causal effects of maritime traffic on air pollution. We apply this method to cruise traffic in Marseille, a large Mediterranean port city. Using a new pair-matching algorithm designed for time series data, we create hypothetical randomized experiments and estimate the change in air pollution caused by a short-term increase in cruise traffic. We carry out a randomization-based approach to quantify uncertainty and compute nonparametric 95% Fisherian intervals (FI). At the hourly level, cruise vessels' arrivals have relatively large impacts on city-level hourly concentrations of nitrogen dioxide, particulate matter and sulfured dioxide. At the daily level, we do not observe any clear effects. Our results suggest that well-designed hypothetical randomized experiments provide a principled approach to evaluate identification challenges in such time series data but also help credibly estimate the negative externalities of maritime traffic.

Tackling Low Statistical Power Issues

The third chapter of the thesis is entitled *Why Acute Health Effects of Air Pollution Could Be Inflated* and is co-authored with Vincent Bagilet (Columbia University). Accurate and precise measurements of the short-term effects of air pollution on health play a key role in setting air quality standards. Yet, statistical power calculations are rarely—if ever—carried out. We first collect estimates and standard errors of all available articles found in the epidemiology and economics literatures. We find that nearly half of them may suffer from a low statistical power and could thereby produce statistically significant estimates that are actually inflated. We then run simulations based on real data to identify which parameters of research designs affect statistical power. Despite their large sample sizes, we show that studies exploiting rare exogenous shocks such as transport strikes or thermal inversions could have a very low statistical power, even if effect sizes are large. Our simulation results indicate that the observed discrepancy in the literature between instrumental variable estimates and non-causal ones could be partly explained by the inherent imprecision of the two-stage least-squares estimator. We also provide evidence that subgroup analysis on the elderly or children should be implemented with caution since the average number of events for an health outcome is a major driver of power. Based on these findings, we build a series of recommendations for researchers to evaluate the design of their study with respect to statistical power issues.

The fourth chapter is entitled with *Unconfounded but Inflated Causal Estimates* and is also co-authored with Vincent Bagilet (Columbia University). It can been considered as an extension and generalization of the third chapter. To avoid confounding, quasi-experimental studies focus on specific sources of variation. This often leads to a reduction in statistical power. Yet, published estimates can overestimate true effects sizes when power is low. Using fake data simulations, we first show that for all causal inference methods, there can be a trade-off between confounding and exaggerating true effect sizes due to a loss in power. We then discuss solutions to assess whether statistically significant estimates from observational studies could be inflated. A more systematic reporting of prospective and retrospective power calculations could help limit this issue.

Bibliography

Angrist, Joshua D and Jörn-Steffen Pischke (2008) *Mostly harmless econometrics*: Princeton university press.

Arceo, Eva, Rema Hanna, and Paulina Oliva (2016) "Does the effect of pollution on infant mortality differ between developing and developed countries? Evidence from Mexico City," *The Economic Journal*, 126 (591), 257–280.

Baccini, Michela, Alessandra Mattei, Fabrizia Mealli, Pier Alberto Bertazzi, and Michele Carugno (2017) "Assessing the short term impact of air pollution on mortality: a matching approach," *Environmental Health*, 16 (1), 1–12.

Bauernschuster, Stefan, Timo Hener, and Helmut Rainer (2017) "When labor disputes bring cities to a standstill: The impact of public transit strikes on traffic, accidents, air pollution, and health," *American Economic Journal: Economic Policy*, 9 (1), 1–37.

Bell, Michelle L and Devra Lee Davis (2001) "Reassessment of the lethal London fog of 1952: novel indicators of acute and chronic consequences of acute exposure to air pollution.," *Environmental health perspectives*, 109 (suppl 3), 389–394.

Bell, Michelle L, Jonathan M Samet, and Francesca Dominici (2004) "Time-series studies of particulate matter," *Annu. Rev. Public Health*, 25, 247–280.

Bhaskaran, Krishnan, Antonio Gasparrini, Shakoor Hajat, Liam Smeeth, and Ben Armstrong (2013) "Time series regression studies in environmental epidemiology," *International journal of epidemiology*, 42 (4), 1187–1195.

Bind, Marie-Abèle (2019) "Causal modeling in environmental health," *Annual review of public health*, 40, 23–43.

Bondy, Malvina, Sefi Roth, and Lutz Sager (2020) "Crime is in the air: The contemporaneous relationship between air pollution and crime," *Journal of the Association of Environmental and Resource Economists*, 7 (3), 555–585.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes (2020) "Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics," *American Economic Review*, 110 (11), 3634–3660, 10.1257/aer.20190687.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg (2016) "Star wars: The empirics strike back," *American Economic Journal: Applied Economics*, 8 (1), 1–32.

Burkhardt, Jesse, Jude Bayham, Ander Wilson et al. (2019) "The effect of pollution on crime: Evidence from data on particulate matter and ozone," *Journal of Environmental Economics and Management*, 98, 102267.

Chang, Tom, Joshua Graff Zivin, Tal Gross, and Matthew Neidell (2016) "Particulate pollution and the productivity of pear packers," *American Economic Journal: Economic Policy*, 8 (3), 141–69.

Chang, Tom Y, Joshua Graff Zivin, Tal Gross, and Matthew Neidell (2019) "The effect of pollution on worker productivity: evidence from call center workers in China," *American Economic Journal: Applied Economics*, 11 (1), 151–72.

Deryugina, Tatyana, Garth Heutel, Nolan H Miller, David Molitor, and Julian Reif (2019) "The mortality and medical costs of air pollution: Evidence from changes in wind direction," *American Economic Review*, 109 (12), 4178–4219.

Dominici, Francesca and Corwin Zigler (2017) "Best practices for gauging evidence of causality in air pollution epidemiology," *American journal of epidemiology*, 186 (12), 1303–1309.

Ebenstein, Avraham, Victor Lavy, and Sefi Roth (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.

Fisher, Ronald Aylmer et al. (1937) "The design of experiments.," *The design of experiments.* (2nd Ed).

Fogarty, Colin B (2020) "Studentized sensitivity analysis for the sample average treatment effect in paired observational studies," *Journal of the American Statistical Association*, 115 (531), 1518–1530.

Forastiere, Laura, Michele Carugno, and Michela Baccini (2020) "Assessing short-term impact of PM 10 on mortality using a semiparametric generalized propensity score approach," *Environmental Health*, 19 (1), 1–13.

Gelman, Andrew and John Carlin (2014) "Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors," *Perspectives on Psychological Science*, 9 (6), 641–651.

Giaccherini, Matilde, Joanna Kopinska, and Alessandro Palma (2021) "When particulate matter strikes cities: Social disparities and health costs of air pollution," *Journal of Health Economics*, 78, 102478.

Godzinski, Alexandre, M Suarez Castillo et al. (2019) "Short-term health effects of public transport disruptions: air pollution and viral spread channels,"Technical report, Institut National de la Statistique et des Etudes Economiques.

Gutman, R, DB Rubin, and Stijn Vansteelandt (2012) "Analyses that Inform Policy Decisions [with Discussions]," *Biometrics*, 68 (3), 671–678.

Halliday, Timothy J, John Lynham, and Aureo de Paula (2019) "Vog: Using volcanic eruptions to estimate the health costs of particulates," *The Economic Journal*, 129 (620), 1782–1816.

He, Jiaxiu, Haoming Liu, and Alberto Salvo (2019) "Severe air pollution and labor productivity: Evidence from industrial towns in China," *American Economic Journal: Applied Economics*, 11 (1), 173–201.

Herrnstadt, Evan, Anthony Heyes, Erich Muehlegger, and Soodeh Saberian (2021) "Air pollution and criminal activity: Microgeographic evidence from Chicago," *American Economic Journal: Applied Economics*, 13 (4), 70–100.

Ho, Daniel E, Kosuke Imai, Gary King, and Elizabeth A Stuart (2007) "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference," *Political analysis*, 15 (3), 199–236.

Holland, Paul W (1986) "Statistics and causal inference," *Journal of the American statistical Association*, 81 (396), 945–960.

Holub, Felix, Laura Hospido, and Ulrich J Wagner (2020) "Urban air pollution and sick leaves: Evidence from social security data."

Imbens, Guido W (2015) "Matching methods in practice: Three examples," *Journal of Human Resources*, 50 (2), 373–419.

Imbens, Guido W and Donald B Rubin (2015) *Causal inference in statistics, social, and biomedical sciences*: Cambridge University Press.

Ioannidis, John P. A. (2008) "Why Most Discovered True Associations Are Inflated," *Epidemiology*, 19 (5), 640–648.

Jaakkola, JJK (2003) "Case-crossover design in air pollution epidemiology," *European Respiratory Journal*, 21 (40 suppl), 81s–85s.

Jans, Jenny, Per Johansson, and J. Peter Nilsson (2018) "Economic Status, Air Quality, and Child Health: Evidence from Inversion Episodes," *Journal of Health Economics*, 61, 220–232, 10.1016/j.jhealeco.2018.08.002.

King, Gary and Langche Zeng (2006) "The dangers of extreme counterfactuals," *Political analysis*, 14 (2), 131–159.

Koop, Gary and Lise Tole (2004) "Measuring the health effects of air pollution: to what extent can we really say that people are dying from bad air?" *Journal of Environmental Economics and Management*, 47 (1), 30–54.

Le Tertre, A, S Medina, E Samoli 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.

Liu, Cong, Renjie Chen, Francesco Sera et al. (2019) "Ambient Particulate Air Pollution and Daily Mortality in 652 Cities," *New England Journal of Medicine*, 381 (8), 705–715, 10.1056/NEJMoa1817364.

Lu, Yun and Scott L Zeger (2007) "On the equivalence of case-crossover and time series methods in environmental epidemiology," *Biostatistics*, 8 (2), 337–344.

McShane, Blakeley B., David Gal, Andrew Gelman, Christian Robert, and Jennifer L. Tackett (2019) "Abandon Statistical Significance," *The American Statistician*, 73 (sup1), 235–245, 10.1080/00031305.2018.1527253.

Moretti, Enrico and Matthew Neidell (2011) "Pollution, health, and avoidance behavior evidence from the ports of Los Angeles," *Journal of Human Resources*, 46 (1), 154–175.

Neyman, Jersey (1923) "Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes," *Roczniki Nauk Rolniczych*, 10, 1–51.

Orellano, Pablo, Julieta Reynoso, Nancy Quaranta, Ariel Bardach, and Agustin Ciapponi (2020) "Short-term exposure to particulate matter (PM10 and PM2. 5), nitrogen dioxide (NO2), and ozone (O3) and all-cause and cause-specific mortality: Systematic review and meta-analysis," *Environment international*, 142, 105876.

Peng, Roger D and Francesca Dominici (2008) "Statistical methods for environmental epidemiology with R," *R: a case study in air pollution and health*.

Romer, David (2020) "In Praise of Confidence Intervals," *AEA Papers and Proceedings*, 110, 55–60, 10.1257/pandp.20201059.

Rosenbaum, Paul R. (2002) *Observational Studies*, Springer Series in Statistics, New York, NY: Springer New York, 10.1007/978-1-4757-3692-2.

Rosenbaum, Paul R (2010) *Design of observational studies*: Springer.

Rosenbaum, Paul R and Donald B Rubin (1983) "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70 (1), 41–55.

Rubin, Donald B (1991) "Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism," *Biometrics*, 1213–1234.

——— (2008) "For objective causal inference, design trumps analysis," *The Annals of Applied Statistics*, 2 (3), 808–840.

Samet, Jonathan M, Scott L Zeger, Francesca Dominici, Frank Curriero, Ivan Coursac, Douglas W Dockery, Joel Schwartz, and Antonella Zanobetti (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.

Samoli, Evangelia, Roger Peng, Tim Ramsay et al. (2008) "Acute effects of ambient particulate matter on mortality in Europe and North America: results from the APHENA study," *Environmental health perspectives*, 116 (11), 1480–1486.

Schlenker, Wolfram and W Reed Walker (2016) "Airports, air pollution, and contemporaneous health," *The Review of Economic Studies*, 83 (2), 768–809.

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

Schwartz, Joel, Elena Austin, Marie-Abele Bind, Antonella Zanobetti, and Petros Koutrakis (2015) "Estimating causal associations of fine particles with daily deaths in Boston," *American journal of epidemiology*, 182 (7), 644–650.

Schwartz, Joel, Kelvin Fong, and Antonella Zanobetti (2018) "A national multicity analysis of the causal effect of local pollution, NO 2, and PM 2.5 on mortality," *Environmental health perspectives*, 126 (8), 087004.

Shah, Anoop SV, Kuan Ken Lee, David A McAllister et al. (2015) "Short term exposure to air pollution and stroke: systematic review and meta-analysis," *bmj*, 350.

Sheldon, Tamara L. and Chandini Sankaran (2017) "The Impact of Indonesian Forest Fires on Singaporean Pollution and Health," *American Economic Review*, 107 (5), 526–529, 10.1257/aer.p20171134.

Sommer, Alice J, Emmanuelle Leray, Young Lee, and Marie-Abèle C Bind (2021) "Assessing environmental epidemiology questions in practice with a causal inference pipeline: An investigation of the air pollution-multiple sclerosis relapses relationship," *Statistics in Medicine*, 40 (6), 1321–1335.

Wang, Chi, Giovanni Parmigiani, and Francesca Dominici (2012) "Bayesian effect estimation accounting for adjustment uncertainty," *Biometrics*, 68 (3), 661–671.

Zigler, Corwin Matthew and Francesca Dominici (2014) "Point: clarifying policy evidence with potential-outcomes thinking—beyond exposure-response estimation in air pollution epidemiology," *American journal of epidemiology*, 180 (12), 1133–1140.

Ziliak, Stephen Thomas and Deirdre N. McCloskey (2008) *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, Economics, Cognition, and Society, Ann Arbor: University of Michigan Press.

Improving the Design Stage of Air Pollution Studies Based on Wind Patterns

AUTHORS:
Léo Zabrocki
PSE - EHESS
leo.zabrocki@psemail.eu

Anna Alari
ISGlobal
anna.alari@isglobal.org

Tarik Benmarhnia
UCSD - Scripps
tbenmarhnia@ucsd.edu

A growing literature in economics and epidemiology has exploited changes in wind patterns as a source of exogenous variation to better measure the acute health effects of air pollution. Since the distribution of wind components is not randomly distributed over time and related to other weather parameters, multivariate regression models are used to adjust for these confounding factors. However, this type of analysis relies on its ability to correctly adjust for all confounding factors and extrapolate to units without empirical counterfactuals. As an alternative to current practices and to gauge the extent of these issues, we propose to implement a causal inference pipeline to embed this type of observational study within an hypothetical randomized experiment. We illustrate this approach using daily data from Paris, France, over the 2008-2018 period. Using the Neyman-Rubin potential outcomes framework, we first define the treatment of interest as the effect of North-East winds on particulate matter concentrations compared to the effects of other wind directions. We then implement a matching algorithm to approximate a pairwise randomized experiment. It adjusts non-parametrically for observed confounders while avoiding model extrapolation by discarding treated days without similar control days. We find that the effective sample size for which treated and control units are comparable is surprisingly small. It is however reassuring that results on the matched sample are consistent with a standard regression analysis of the initial data. We finally carry out a quantitative bias analysis to check whether our results could be altered by an unmeasured confounder: estimated effects seem robust to a relatively large hidden bias. Our causal inference pipeline is a principled approach to improve the design of air pollution studies based on wind patterns.

Introduction

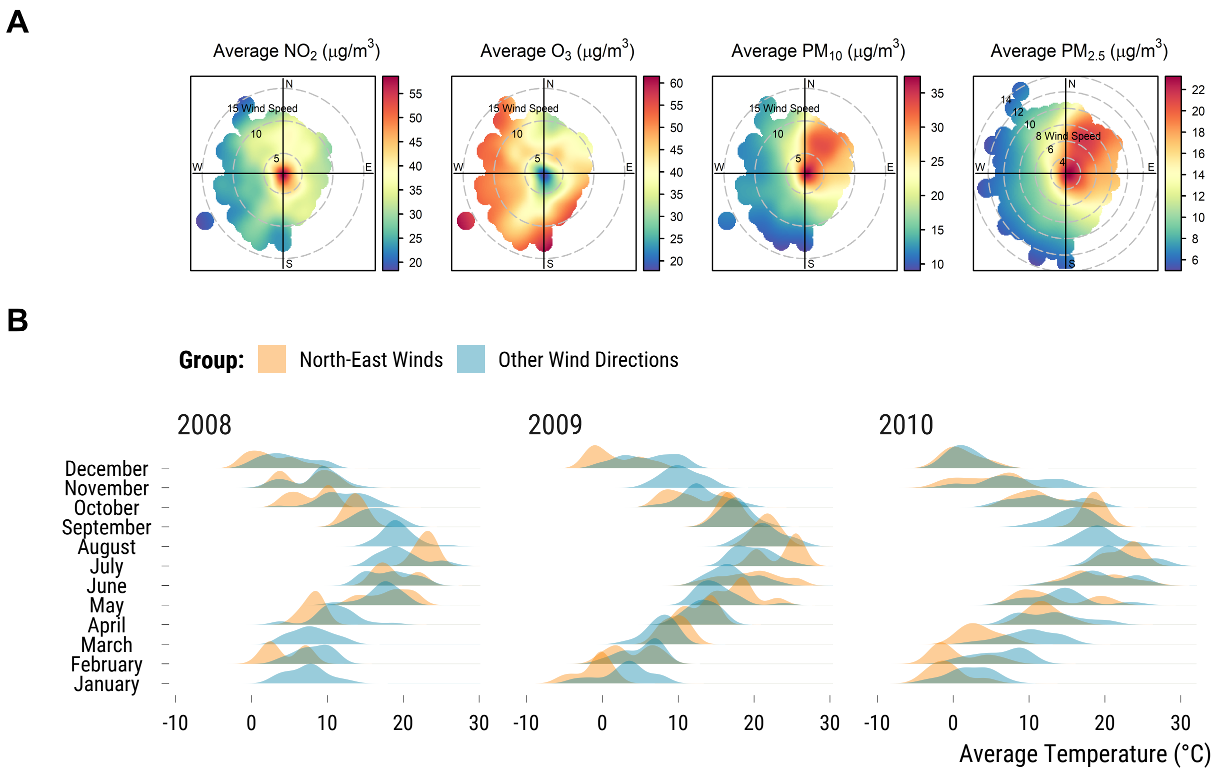
WEBSITE:
https://lzabrocki.github.io/design_stage_wind_air_pollution/

REPLICATION DATA:
https://osf.io/7x23u/

A growing literature in economics and epidemiology has recently re-examined the short-term effects of air pollution on mortality and emergency admissions using causal inference methods. Among these techniques, instrumental variable strategies have been very popular since they can overcome the biases caused by unmeasured confounders and measurement errors in air pollution exposure (Schlenker and Walker 2016, Arceo et al. 2016, Schwartz et al. 2017; 2018, Halliday et al. 2019, Deryugina et al. 2019). Daily changes in wind directions are such instrumental variables since they arguably meet two of the three main requirements for the method to be valid: they can strongly affect air pollutant concentrations while having no direct effects on health outcomes (Angrist et al. 1996, Angrist and Pischke 2008, Baiocchi et al. 2014). This strategy however rests on the remaining assumption that changes in wind directions occur randomly, which is often not credible without further statistical adjustments. One could unfortunately fear that the resulting analysis would depend on the quality of the model (King and Zeng 2006, Stuart and Rubin 2008). Does the model take into account all relevant confounding factors, and if so, are they adjusted for with the correct functional forms? Is the model also able to extrapolate when there is little overlap in covariate distributions?

To illustrate these issues, imagine that we are interested in estimating the influence of particulate matters on daily mortality in Paris, France, over the 2008-2018 period. Research in atmospheric science has shown that winds blowing from the North-East could transport particulate matters due to wood burning in the region but also from other sources located in North-Eastern Europe (Bressi et al. 2014, Petetin et al. 2014, Stirnberg et al. 2021). We could therefore use the comparison of winds blowing from the North-East to those from other directions as an instrumental variable for particulate matters.

Figure 1: Polar Plots of Air Pollutant Concentrations Predicted by Wind Components and Average Temperature Imbalance of Wind Directions by Year and Month. *Notes:* In panel A, each plot represents the concentrations (in µg/m^3) of an air pollutant that were predicted using a generalized additive model based on a smooth isotropic function of the two wind components u and v (Carslaw and Ropkins 2012). The direction from which the wind blows is described on a 360° compass rose and wind speed (in m/s) is represented by a series of increasing circles starting from the intersection of the two cardinal directions axes where wind speed is null: the farther the circle is away from the intersection, the faster the wind speed is. In panel B, the density distribution of the average temperature (in °C) is drawn for North-East winds (orange colour) and other wind directions (blue colour). The figure is divided into subplots by month and year (2008-2010).



In Panel A of Figure 1, we display polar plots of air pollutant concentrations that were predicted using a Generalized Additive Model

(GAM) and wind components as inputs (Carslaw and Ropkins 2012). We clearly see that winds blowing from the North-East are associated with higher PM_{10} and $PM_{2.5}$ concentrations. These patterns could however be confounded by other variables such as the weather parameters or a shared seasonality in air pollution and wind patterns. For instance, in Panel B of Figure 1, the density distribution of the average temperature (°C) is not similar for the groups of wind directions. We must take into account this confounding variable if we want to make the as-if random distribution of North-East wind more credible. Multivariate linear regression have been the standard approach to help achieve this goal but more flexible methods such as generalized additive models and machine learning algorithms could also be used (Grange et al. 2018, Grange and Carslaw 2019). Yet, even a very flexible model will not overcome the second issue visible in Panel B of Figure 1: as for January 2008, the model will sometimes depend on extrapolation since there are no empirical counterfactuals to estimate what would have happened had the wind blown from the North-East. Finally, it could be argued that we fail to adjust for a confounding variable which we have not measured. In addition to explaining with qualitative arguments why it is not likely the case, we should also try to quantify the bias induced by an unmeasured confounder.

In this paper, we show how we can evaluate the extent to which studies exploiting wind directions as instrumental variables could be prone to the issues raised above. To achieve this goal, we follow the four consecutive stages of the causal inference pipeline proposed by (Bind and Rubin 2019; 2021) that explicitly embed the design of this type of observational study within an hypothetical randomized experiment (Rubin 2008, Rosenbaum 2010, Imbens and Rubin 2015, Hernán and Robins 2016).

First, in a *conceptual stage*, we clearly state the causal question of interest using the Neyman-Rubin potential outcomes framework (Neyman 1923, Rubin 1974). Our treatment of interest is the effect of North-East winds on air pollution compared to other wind directions. To estimate this effect, for treated days with winds blowing from the North-East, we need to impute the concentrations that would have been observed had winds blown from other directions. The issue is that wind patterns are not randomly assigned: control days with wind blowing from other directions are not similar to treated days.

We therefore implement a *design stage* where we approximate a pairwise randomized experiment using a matching algorithm recently designed for air pollution studies (Sommer et al. 2021). Matching is a transparent method to adjust for confounders without making parametric assumption and directly looking at observed outcomes (Ho et al. 2007, Stuart 2010). Given a set of chosen covariate distances, each treated day is matched to its closet control day. This method also avoids model extrapolation since treated days for which no control days exist in the data are discarded from the analysis.

The third step is an *analysis stage* where we estimate the influence of North-East winds on air pollutant concentrations. We simply compute the average difference in concentrations between matched treated and control days and rely on Neymanian inference to compute an estimate of the sampling variability (Imbens and Rubin 2015). The last and fourth step is to carry out a *sensitivity analysis*. Throughout the previous steps, we must make the strong assumption that no unmeasured variables could be related both to wind patterns and air pollutant concentrations. Quantitative bias analysis was initially proposed by Cornfield et al. (1959) to assess which magnitude of hidden bias would be required to alter observed results. We follow here the method developed by Rosenbaum (2010) and Fogarty (2020).

With this study, we aim to bring two contributions to the causal inference literature on the acute health effects of air pollution. First, we show that using wind directions as instrumental variables requires more caution to make the assumption that they are "as-if" randomly distributed according to observed covariates convincing. The effective sample size where treated and control units are similar on a set of observed covariates is actually small. The standard approach used in the literature based on multivariate regression models will therefore rely on its ability to adjust correctly for the functional forms of covariates and extrapolate to units without empirical counterfactuals. Second, our quantitative bias analysis reveals that the estimated increase in particulate matter concentrations due to North-East winds is relatively robust to the presence of hidden bias. Even if an unobserved confounding factor is twice more common among days with winds blowing from the North-East than among days with winds from other directions, the large range of estimates consistent with the data remains positive.

We also hope that the approach we propose in this paper could be of interest to atmospheric scientists. The fact that wind patterns play a key role in the variation of air pollution concentrations is obviously not new (Wilson and Suh 1997, Hoek et al. 2008, Tai et al. 2010, Aguilera et al. 2020). Yet, causal inference methods have rarely been implemented in atmospheric science to estimate the influence of weather parameters on air pollution. We believe that mimicking a randomized experiment corresponds to an intuitive approach and could complement source apportionment and emission inventory approaches. While wind is non manipulable, emission sources are and our framework could also serve as a stepping-stone to evaluate potential interventions to control emissions—if a source is shut-down in the North-East of Paris, would wind blowing from this direction influence less specific air pollutant concentrations?

We took great care to make our work fully reproducible to help researchers implement but also improve and criticize our approach. Data and detailed **R** codes are available at `https://lzabrocki.github.io/design_stage_wind_air_pollution/` and backed-up in an Open Science Framework repository (Zabrocki 2022).

Methods

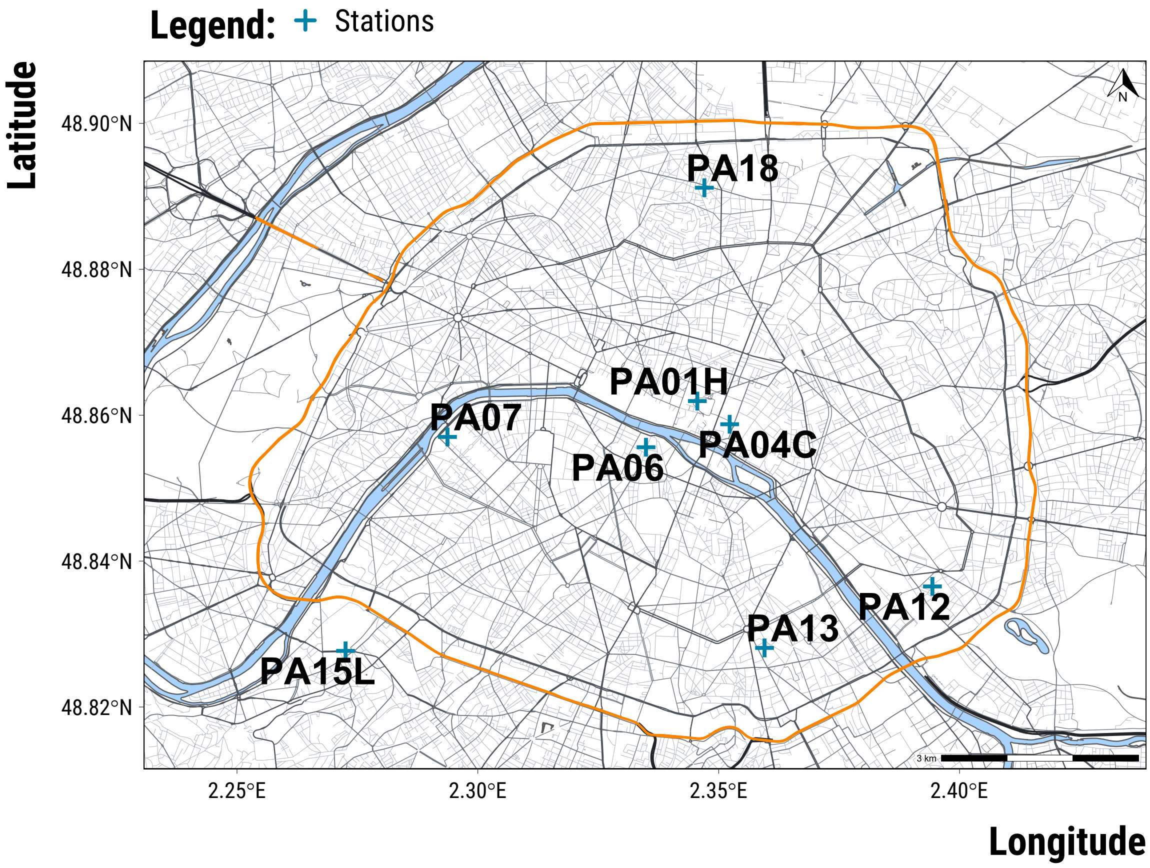


Figure 2: Map of road network and location of air pollution measuring stations in Paris, France. *Notes:* Grey lines represent the road network. The orange line is the orbital ring surrounding Paris. Blue crosses are the locations of air pollution measuring stations. NO_2 concentrations are measured at stations PA07, PA12, PA13, PA18; O_3 concentrations at PA13, PA18; PM_{10} at PA18; $PM_{2.5}$ at PA01H and PA04C. The map was created with the R programming language (version 4.1.0) (R Core Team 2021), data were provided by OpenStreetMap (OpenStreetMap contributors 2017) and retrieved with the `osmdata` package (Padgham et al. 2017).

Data

We built a dataset combining daily time series of air pollutant concentrations and weather parameters in Paris over the 2008-2018 period. We chose to carry out an analysis at the daily level as done in studies on the acute health effects of air pollution (Schwartz et al. 2017; 2018, Deryugina et al. 2019).

First, we obtained hourly air quality data from AirParif, the local air quality monitoring agency. Figure 2 displays the location of the selected measuring stations. Using a 2.5% trimmed mean, we first averaged at the daily level the concentrations ($\mu g/m^3$) of background measuring stations for NO_2, O_3 and PM_{10}. For a given day, if more than 3 hourly readings were missing, the average daily concentration was set to missing. The proportion of missing values for stations ranged from 2.8% up to 9.1%. We also computed the average daily concentrations of $PM_{2.5}$ but 25% of the recordings were missing: the air pollutant was not measured by Airparif between 2009/09/22 and 2010/06/23. It is important to note that we did not retrieve data from traffic monitors but only from background monitors as they are used to assess the residential exposure of a city population in epidemiological studies.

We then retrieved meteorological data from the single monitoring station located in the South of the city and ran by the French national meteorological service Météo-France. We extracted daily observations on wind speed (m/s), wind direction (measured on a 360° wind rose where 0° is the true North), the average temperature (°C), and the rainfall duration (min). Weather parameters had very few missing values (e.g., at most 2.5% of observations were missing for the rainfall duration).

Finally, to avoid working with a reduced sample size, we imputed missing values for all variables but $PM_{2.5}$. There were no clear patterns in the missingness of NO_2, O_3 and PM_{10} concentrations. We used the chained random forest algorithm implemented by the **R** package missRanger (Mayer 2019). A small simulation exercise showed that it had good performance for imputing NO_2 concentrations (the absolute difference between observed and imputed values was equal to 3.2 $\mu g/m^3$ for an average concentration of 37.6 $\mu g/m^3$) but was much less effective for imputing PM_{10} concentrations (the absolute difference between observed and imputed values was equal to 6.1 $\mu g/m^3$ for an average concentration of 23.4 $\mu g/m^3$). Once the data were imputed, we averaged the air pollutant concentrations at the city level as it is the spatial level of analysis used in Schwartz et al. (2017; 2018).

Further details on data wrangling and an exploratory analysis of the data can be found in the supplementary materials. We were not allowed to share weather data from Météo-France so we added some noise to the weather parameters.

A Causal Inference Pipeline

We present below the four stages of the causal inference pipeline we advocate to use for improving the design of air pollution studies based on wind patterns. Its implementation was done with the R programming language (version 4.1.0) (R Core Team 2021).

Stage 1: Defining the Treatment of Interest The first step of our causal inference approach is to clearly state the question we are trying to answer: *What is the effect of North-East winds on particulate matter in Paris over the 2008-2018 period?* This question is motivated by the exploratory analysis of Figure 1 and research in atmospheric science on the sources of particulate matter located in the North-East of the city. Our treatment of interest is therefore defined as the comparison of air pollutant concentrations when winds are blowing from the North-East (10°-90°) with concentrations when wind come from other directions. We frame this question in the Rubin-Neyman causal framework (Neyman 1923, Rubin 1974). Our units are 4,018 days indexed by i (i=1,..., I). For each day, we define our treatment indicator W_i which takes two values. It is equal to 1 if the unit is treated (the wind blows from the North-East), and 0 if the unit belongs to the control group (the wind is blowing from another direction). Under the Stable Unit Treatment Value Assumption (STUVA), we assume that each day can have two potential concentrations in $\mu g/m^3$ for an air pollutant: $Y_i(1)$ if the wind blows from the North-East and $Y_i(0)$ if the wind blows from another direction.

The fundamental problem of causal inference states that we can only observe for each day one of these two potential outcomes: it is a missing data problem (Holland 1986, Ding and Li 2018). The observed concentration of an air pollutant Y^{obs} is defined as $Y^{obs} = (1\text{-}W_i) \times Y_i(0) + W_i \times Y_i(1)$. If the unit is treated, we observe $Y_i(1)$. If it is a control, we observe $Y_i(0)$. To estimate the effect of North-East winds on air pollutant concentrations, we therefore need to impute the missing potential outcomes of treated units—what would have been the air pollutant concentrations if the wind had blown from another direction?

Stage 2: Designing the Hypothetical Randomized Experiment The second stage of our causal inference pipeline is to embed our non-randomized study within an hypothetical randomized experiment. We are dealing with an observational study where North-East winds are not randomly distributed through a year and are correlated with other weather parameters influencing air pollutant concentrations. In Figure 3, we plot, for each month, the absolute standardized mean differences between treated and control units for the average temperature, relative humidity and wind speed: most differences are superior to 0.1, which is often considered as a threshold to assess the imbalance of covariates.

To better approximate a randomized experiment, we must therefore find the subset of treated units which are similar to control units. Formally, we want to make plausible for this subset of units the assumption that the treatment assignment is independent from the potential outcomes of units given their covariates $\mathbf{X}$: $\Pr(\mathbf{W} \mid \mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = \Pr(\mathbf{W} \mid \mathbf{X})$. The issue is that some units' covariates are observed while other are not. Unlike a randomized experiment where both observed and unobserved covariates will be, on average, balanced across treatment and control groups, we must assume that no unobserved covariates affect the treatment assignment.

Matching methods are particularly convenient to design hypothetical randomized experiments. Contrary to standard regression approaches, matching is a non-parametric way to adjust for observed covariates while avoiding model extrapolation since units without counterfactuals in the data are discarded from the analysis. Specifically, we use a constrained matching algorithm to design a pairwise randomized experiment where, for each pair, the probability of receiving the treatment is equal to 0.5 (see (Sommer et al. 2021) for further details on the algorithm). Each treated unit is matched to its closest unit given a set of covariate constraints which represent the maximum distance, for each covariate, allowed between treated and

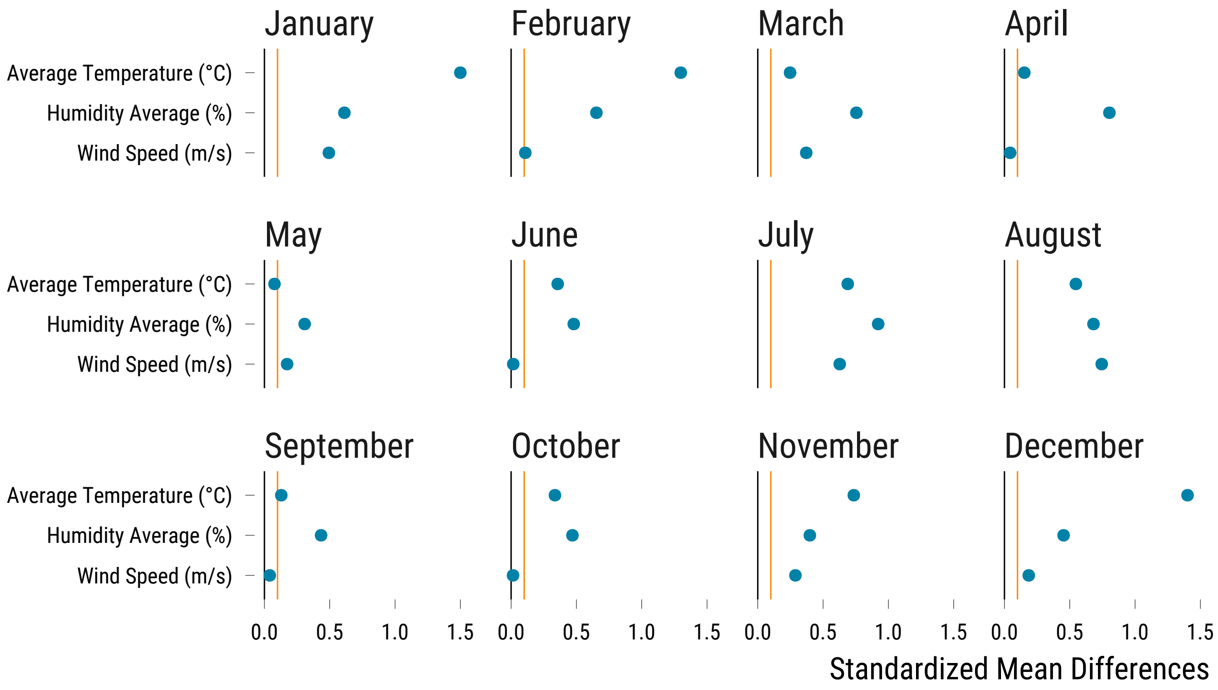


Figure 3: Evidence of Imbalance for Weather Covariates. *Notes:* For each month, we compute the absolute standardized differences for continuous weather covariates between treated and control groups. These differences are represented as blue points. The vertical orange line is the 0.1 threshold which is used in the matching literature to spot covariates imbalance. The vertical black line is at 0.

control units. We match on the two sets of covariates influencing both wind directions and air pollutant concentrations.

First, we match on calendar variables such as the Julian date, weekend, holidays and bank days indicators. A treated unit could be matched up to a control unit with a maximum distance of 60 days. If we extend this distance, it would be easier to match treated units to control units but the treatment effect could be biased by seasonal variation in air pollutant concentrations. We match exactly treated and control units for the other calendar indicators.

Second, we match on weather variables. The average temperature between treated and control units could not differ by more than 5°. The difference in wind speed must be less than 0.5 m/s. The rainfall duration (divided in four ordinal categories) needs to be the same and the absolute difference in average humidity could be up to 12 percentage points. We also force the absolute difference in PM_{10} concentrations in the previous day to be less or equal to 8 $\mu g/m^3$. The thresholds we set up were chosen through an iterative process were we checked (i) that they led to balanced sample of treated and control units and (ii) that there were enough matched pairs to draw our inference upon.

Finally, the Stable Unit Treatment Value Assumption (SUTVA) requires that there is no interference between units and no hidden variation of the treatment. To make this assumption more plausible, we discard from the analysis the matched pairs for which the distance in days is inferior to 4 days and make sure that the first lag of the treatment indicator for treated and control units.

Stage 3: Analyzing the Experiment using Neymanian Inference In the third stage, we proceed to the analysis of our hypothetical pairwise randomized experiment. Several modes of statistical inference such as Fisherian, Neymanian or Bayesian could be implemented (Rubin 1991). Here, we take a Neymanian perspective where the potential outcomes are assumed to be fixed and the treatment assignment is the basis of inference. Our goal is to measure the average causal effect for the sample of matched units. We assume that each of the two units of a matched pair j has two potential concentrations for an air pollutant. If we were able to observe these potential outcomes, we could simply measure the effect of North-East winds on air pollutant concentrations by computing the finite-sample average treatment effect for matched treated units τ_{fs}. We would first compute for each pair the mean difference in concentrations and then average the differences over the J pairs. While we only observe one potential outcome for each unit, we can nonetheless estimate τ_{fs} with the average of observed pair differences $\hat{\tau}$:

$$\hat{\tau} = \frac{1}{J}\sum_{j=1}^{J}(Y^{\text{obs}}_{\text{t},j} - Y^{\text{obs}}_{\text{c},j}) = \overline{Y}^{\text{obs}}_{\text{t}} - \overline{Y}^{\text{obs}}_{\text{c}}$$

Here, the subscripts t and c respectively indicate if the unit in a given pair is treated or not. Since there are only one treated and one control unit within each pair, the standard estimate for the sampling variance of the average of pair differences is not defined. We can however compute a conservative estimate of the variance (Imbens and Rubin 2015):

$$\hat{\mathbb{V}}(\hat{\tau}) = \frac{1}{J(J-1)}\sum_{j=1}^{J}(Y^{\text{obs}}_{\text{t},j} - Y^{\text{obs}}_{\text{c},j} - \hat{\tau})^2$$

We finally compute an asymptotic 95% confidence interval using a Gaussian distribution approximation:

$$\text{CI}_{0.95}(\tau_{\text{fs}}) = \left(\hat{\tau} - 1.96 \times \sqrt{\hat{\mathbb{V}}(\hat{\tau})},\ \hat{\tau} + 1.96 \times \sqrt{\hat{\mathbb{V}}(\hat{\tau})}\right)$$

The obtained 95% confidence interval gives the set of effect sizes compatible with our data (Amrhein et al. 2019).

Stage 4: Sensitivity Analysis The fourth step of our causal inference pipeline is to explore how sensitive our analysis is to violation of the assumptions it relies upon. We carry out three types of robustness checks.

First, we make the strong assumption that the treatment assignment is as-if random: winds blowing from the North-East occur randomly conditional on a set of measured covariates. Other researchers could however argue that we fail to adjust for unmeasured variables influencing both the occurrence of North-East winds and air pollutant concentrations. Within matched pairs, these unobserved counfounders could make the treated day more likely to have wind blowing from the North-East than the control day. We therefore implement the quantitative bias analysis, also called sensitivity analysis, that was developed by Rosenbaum (2010) and Fogarty (2020). It allows us to explore how our results would be altered by the effect of an unobserved confounder on the treatment odds, denoted by Γ. In our matched pairwise experiment, we assume that within each pair, control and treated days have the odds to see the wind blowing from the North-East: the odds of treatment is such that $\Gamma = 1$. The quantitative bias analysis allows to compute the 95% confidence intervals obtained for different values of bias the unmeasured confounder has on the treatment assignment. For instance, if we assume that an unmeasured confounder has a small effect on the odds of treatment (i.e., for a $\Gamma > 1$ and close to 1) but the resulting 95% confidence interval becomes completely uninformative, it would imply that our results are highly sensitive to hidden bias. Conversely, if we assume that an unmeasured confounder has a strong effect on the odds of treatment (i.e., for a large Γ) and we find that the resulting 95% confidence interval remains similar, it would imply that our results are very robust to hidden bias. In a complementary manner, we also check whether unmeasured biases could be present by using the first daily lags of air pollutant concentrations as control outcomes (Rosenbaum 2018). If our matched pairs are indeed similar in terms of unobserved covariates, the treatment occurring in t should not influence concentration of air pollutants in $t - 1$.

Second, for many matched pairs, air pollutant concentrations were imputed using the chained random forest algorithm (Mayer 2019). We check whether the results are sensitive to the imputation by rerunning the analysis for the non-missing concentrations.

Third, we make sure that the treatment assignment within pairs was effective to increase the precision of estimates. We compare the estimate of the sampling variance of a pairwise randomized experiment to the one of a completely randomized experiment. If the estimate of sampling variability for the pairwise experiment is smaller than the estimate of sampling variability for a complete experiment, it means that our matching procedure was successful to match similar units within pairs compared to randomly selected units (Imbens and Rubin 2015).

Results

Performance of the Matching Procedure

Our initial dataset consists in 4,018 daily observations, divided into 912 treated units and 3,106 control units. The matching procedure results in 121 pairs of matched treated-control units—only 13% of treated units could be matched to similar control units given the constraints we set. In the supplementary materials, we show that the matched sample has different characteristics from the initial sample: observations belong more to the period ranging from May to October, their average temperature is higher and their relative humidity is lower.

In Figure 4, we display how the balance of continuous and categorical covariates improves after the matching procedure. Blue dots represent either the absolute mean differences between treated and control units for continuous variables or the absolute differences in percentage points for categorical variables. For continuous covariates, the average standardized mean differences between treated and control days is 0.26 before matching and reduces to 0.07 after the procedure. For categorical covariates, the average difference in percentage points diminishes from 6.2 to 1.8 after matching. Our matching procedure therefore leads to a consequent reduction of our sample size but allows us to compare treated units that are more similar to control units. A complete analysis of the balance improvement for each covariate is available in the supplementary materials.

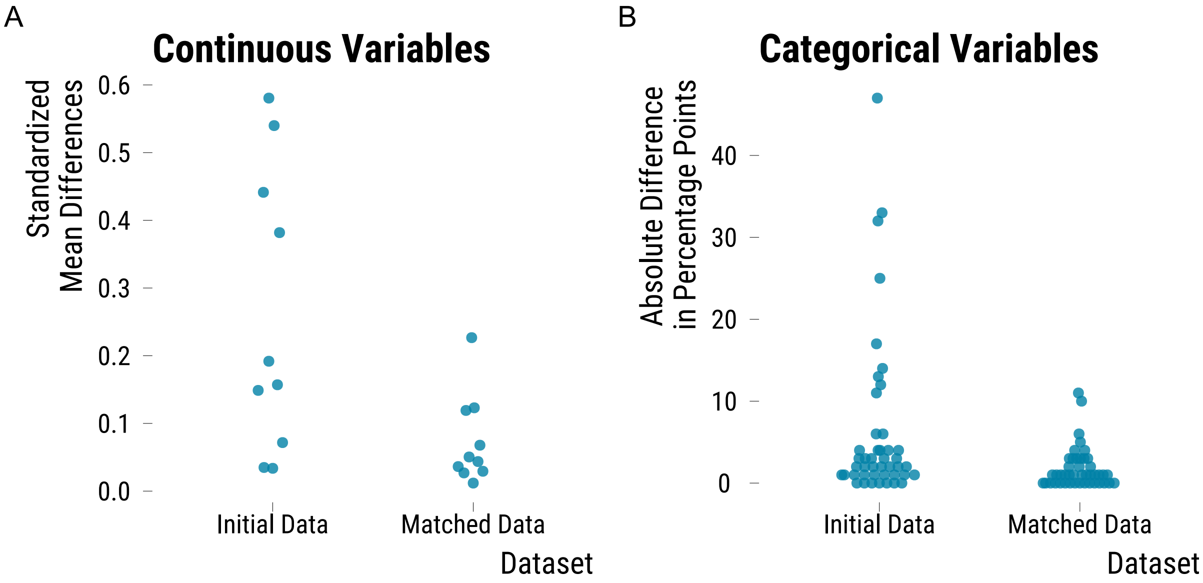


Figure 4: Overall Balance Improvement in Continuous and Categorical covariates. *Notes:* In Panel A, we plot, before and after matching, the absolute standardized differences in continuous covariates between treated and control groups. Each blue dot represents an absolute mean difference for a given covariate. In panel B, we plot, before and after matching, the absolute difference in percentage points for categorical covariates.

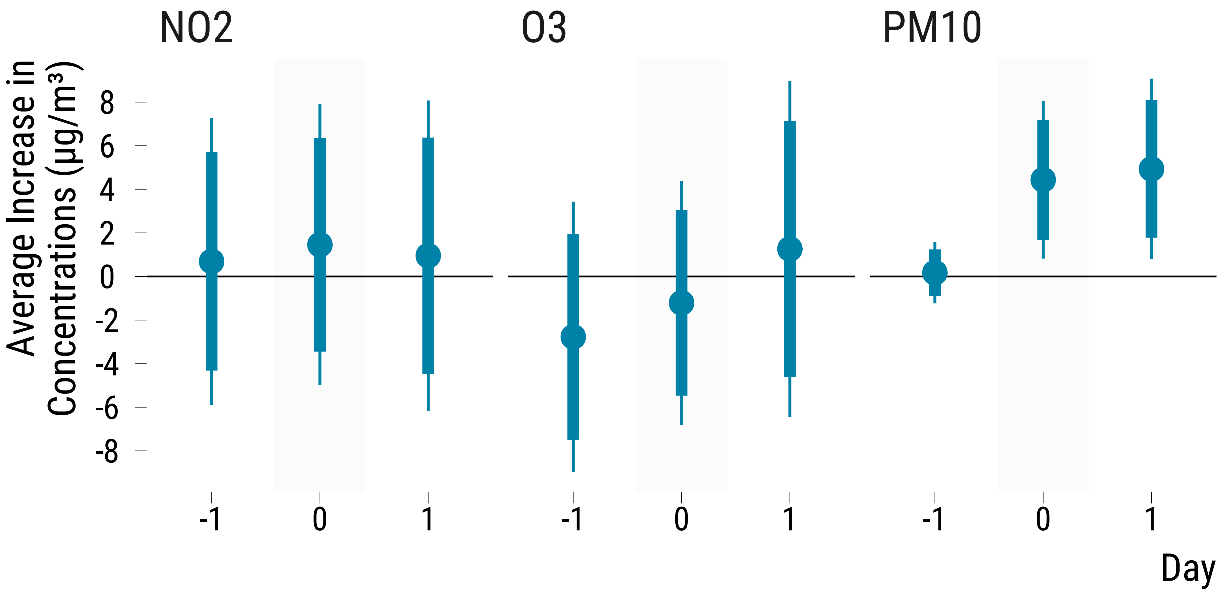
North-East Wind Effects on Air Pollutant Concentrations

For each air pollutant, we plot in Figure 5 the estimated average difference in concentration (µg/m^3) between North-East winds and other wind directions. We also display the estimated differences for the previous day and the following day. Thick lines represent the 95% confidence intervals while thin lines are the 99% confidence intervals. The third panel of Figure 5 confirms the exploratory analysis of the polar plot. When wind blows from the North-East, PM_{10} concentrations increase by 4.4 µg/m^3, with the lower and upper bounds of the 95% confidence being respectively equal to an increase by 1.7 µg/m^3 and 7.2 µg/m^3. The estimated difference represents an 18% increase in the average concentration of PM_{10}. We also observe a positive difference of 25% in PM_{10} concentrations the following day (point estimate of 4.9; 95% CI: 1.8, 8.1).

North-East winds do not seem to influence NO_2 (point estimate of 1.5; 95% CI: -3.4, 6.4), and O_3 (point estimate of -1.2; 95% CI: -5.5, 3.1) concentrations on the current day. This is also the case for the concentrations of these two air pollutants on the following day.

Regarding the effects of North-East winds on $PM_{2.5}$, we restrain our analysis to pairs without missing concentrations. For the current and following days, we respectively find an average increase of 1.4 µg/m^3 (95% CI: -0.6, 3.4) and 2.7 µg/m^3 (95% CI: 0.8, 4.5). These point estimates respectively represent a 8.8% and a 17% relative increases in $PM_{2.5}$ concentrations.

Figure 5: Effects of North-East Winds on Air Pollutant Concentrations. *Notes:* In each panel, we plot the estimated effects of North-East winds on air pollutant concentrations for the previous, current and following days. Point estimates are depicted by blue points; blue thick lines are 95% confidence intervals and thin lines are 99% confidence intervals. The 95% and 99% confidence intervals associated with the estimated average difference in PM_{10} in the first lag are smaller than other intervals for the following days since we added a constraint in the matching procedure for this lag of the air pollutant.



Sensitivity Analysis

Our quantitative bias analysis reveals that if we have failed to adjust for an unobserved confounder twice more common among treated days, the resulting 95% confidence intervals for the estimated effects of North-East winds on PM_{10} would be equal to (0.5, 9) for the current day and to (-0.2, 10) for the the following day. Confidence intervals are still consistent with mostly positive effects but are relatively wide. As a complementary test for unobserved confounders, we also check that the occurrence of North-East winds on the current day does not have any effect on concentrations measured in the previous day. Reassuringly, for NO_2 and O_3, 95% confidence intervals do not suggest clear negative or positive average differences in concentrations as shown in Figure 5 (for $PM_{2.5}$, the estimated average difference is -0.1 $\mu g/m^3$ (95% CI: -1.2, 1)).

In the supplementary materials, we check whether the imputation of missing air pollutant concentrations did not drive our results. For NO_2, O_3 and PM_{10}, 13%, 8% and 7% of concentrations were respectively imputed. We replicate our analysis on the subset of pairs without missing observations: point estimates remain very similar but confidence intervals are a bit larger due to the sample size loss. This robustness check implies that our imputation did not bias our estimates.

Finally, the pairwise design of our hypothetical experiment does not help increase the precision of the estimated differences in PM_{10} concentrations. The standard error under a completely randomized assignment is equal to 1.35 while the one of a pairwise randomized assignment is 1.4. The pairwise design however increases the precision estimates for O_3 by 23% for O_3 but decreases the precision by 42% for NO_2.

Discussion

In our study, we follow a causal inference pipeline to craft a hypothetical experiment for measuring the effects of North-East winds on daily particulate matter concentrations in Paris. Our constrained pair matching algorithm enables us to find the subset of treated days that were similar to control days for a set of calendar and weather confounding factors. Compared to a statistical adjustment based on a multivariate regression model, matching is non-parametric and avoids to extrapolate to units without empirical counterfactuals. At the very heart of this method, graphical displays of covariates balance allow to check in a transparent manner whether the as-if random distribution of the treatment was achieved conditional on observed confounders. We were surprised that covariates balance could only be achieved for 13% of treated units. It would be an interesting question for future research to see if alternative methods such as cardinality matching or bayesian additive regression trees lead to similar results (Hill 2011, Hill and Su 2013, Visconti and Zubizarreta 2018). The relevant structure of the hypothetical experiment to target should also be of interest since our pair matching algorithm failed to increase the precision of estimates compared to a completely randomized assignment of the treatment.

The difficulty to find similar treated and control units could lead researchers interested in the acute health effects of air pollution to worry that instrumental variable strategies exploiting wind patterns and based on multivariate regression models might suffer from extrapolation bias (King and Zeng 2006, Ho et al. 2007). In the supplementary materials, we show that results based on an outcome regression approach, even if they are based on the entire sample, are consistent with those found with the matched data. This may increase the confidence in the capability of a multivariate regression model to correctly extrapolate. Matching estimates are however much less precise. Further research is therefore needed to better understand if improving the design stage of instrument variable studies with matching methods is feasible given the small sample size it entails (Small and Rosenbaum 2008, Baiocchi et al. 2012, Kang et al. 2016, Keele and Morgan 2016). If it is the case, could matching methods actually lead to different results (Schwartz et al. 2015, Baccini et al. 2017, Forastiere et al. 2020)?

In addition to providing evidence on the effective sample size for which covariates balance was achievable, our study was the occasion to assess whether the estimated effects of North-East wind on particulate matters were robust hidden bias. It would require an unmeasured confounder twice more common among treated days to raise doubt on the direction of the estimated effects. This raises our confidence in the assumption that North-East wind are also randomly distributed according to unobserved variables. To the best of our knowledge, this assumption was waiting to be quantitatively evaluated. This could be explained by the fact that the sensitivity analysis we rely on was developed for pairwise matched data (Fogarty 2020). As an alternative, researchers wishing to keep working with a regression approach could implement the new method developed by Cinelli and Hazlett (2020a;b).

Finally, our study presents two main limits regarding the improvement of the design stage of air pollution studies based on wind directions. The first limit concerns the definition of the contrast of interest, that is to say the difference of air pollutant concentrations between North-East winds and other wind directions. If this comparison is easy to understand, the treatment we defined is not manipulable contrary to those found in randomized controlled trials. It might lack a certain appeal to policy-makers as our estimates only indicate whether North-East winds lead to higher particulate matter concentrations than other wind directions (Zigler and Dominici 2014, Dominici and Zigler 2017), without determining the origin of the sources emitting the air pollutant. To overcome this limit, a study exploiting variations in wind directions should be combined with a clear shock on one of the sources emitting an air pollutant. For instance, in a recent paper in Southern California (Aguilera et al. 2020), it was shown that Santa Ana winds have a predominant ventilation effect on $PM_{2.5}$ but when inland wildfires occur, Santa Ana winds are instead increasing $PM_{2.5}$ levels on the coast.

The second limit revolves around the assumption that, for wind direction to be a valid instrument, its effects on a health outcome must be fully mediated by a single air pollutant (Angrist et al. 1996, Angrist and Pischke 2008, Baiocchi et al. 2014). As recognized by researchers, studies exploiting wind patterns could violate this assumption if changes in wind direction affect simultaneously several air pollutants. In our study, once the data are matched, it seems that North-East winds only influence particulate matter, which could reinforce the credibility of the assumption. Yet, this should not be always the case as it would be highly dependent on the city and air pollutant investigated. Methodological work is much needed to understand in which cases the air pollutants co-variance structure could lead to biased dose-response. In a recent work, Godzinski and Castillo (2021) propose to run a multi-pollutant model where each air pollutant concentration is predicted by selecting the optimal set of instrumental variables using least absolute shrinkage and selection operator (lasso). The authors show that results of an instrumented multi-pollutant model can be very different from those found by single-pollutant models. It remains to be studied if matching could also help limit this well-known issue.

Acknowledgments

Airparif and Metéo-France helped us access the data we used in our analysis. Map data copyrighted OpenStreetMap contributors and available from `https://www.openstreetmap.org`. Grégory Abbou provided valuable feedback on an early and different draft of the paper. We especially thank Stéphane Shao for developing the matching algorithm and Marie-Abèle Bind and Alice Sommers for formulating and advocating in several papers the causal inference pipeline we rely on. Comments from Michela Baccini, Tatyana Deryugina, Hélène Ollivier, as well from two reviewers greatly improved the article.

Bibliography

Aguilera, Rosana, Alexander Gershunov, Sindana D Ilango, Janin Guzman-Morales, and Tarik Benmarhnia (2020) "Santa Ana winds of Southern California impact PM2. 5 with and without smoke from wildfires," *GeoHealth*, 4 (1), e2019GH000225.

Amrhein, Valentin, David Trafimow, and Sander Greenland (2019) "Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication," *The American Statistician*, 73 (sup1), 262–270.

Angrist, Joshua D, Guido W Imbens, and Donald B Rubin (1996) "Identification of causal effects using instrumental variables," *Journal of the American statistical Association*, 91 (434), 444–455.

Angrist, Joshua D and Jörn-Steffen Pischke (2008) *Mostly harmless econometrics*: Princeton university press.

Arceo, Eva, Rema Hanna, and Paulina Oliva (2016) "Does the effect of pollution on infant mortality differ between developing and developed countries? Evidence from Mexico City," *The Economic Journal*, 126 (591), 257–280.

Baccini, Michela, Alessandra Mattei, Fabrizia Mealli, Pier Alberto Bertazzi, and Michele Carugno (2017) "Assessing the short term impact of air pollution on mortality: a matching approach," *Environmental Health*, 16 (1), 1–12.

Baiocchi, Michael, Jing Cheng, and Dylan S Small (2014) "Instrumental variable methods for causal inference," *Statistics in medicine*, 33 (13), 2297–2340.

Baiocchi, Mike, Dylan S Small, Lin Yang, Daniel Polsky, and Peter W Groeneveld (2012) "Near/far matching: a study design approach to instrumental variables," *Health Services and Outcomes Research Methodology*, 12 (4), 237–253.

Bind, Marie-Abèle C. and DB Rubin (2021) "The importance of having a conceptual stage when reporting non-randomized studies," *Biostatistics & Epidemiology*, 5 (1), 9–18.

Bind, Marie-Abèle C. and Donald B. Rubin (2019) "Bridging observational studies and randomized experiments by embedding the former in the latter," *Statistical Methods in Medical Research*, 28 (7), 1958–1978.

Bressi, M, Jean Sciare, Véronique Ghersi et al. (2014) "Sources and geographical origins of fine aerosols in Paris (France)," *Atmospheric Chemistry and Physics*, 14 (16), 8813–8839.

Carslaw, David C and Karl Ropkins (2012) "Openair—an R package for air quality data analysis," *Environmental Modelling & Software*, 27, 52–61.

Cinelli, Carlos and Chad Hazlett (2020a) "Making sense of sensitivity: Extending omitted variable bias," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82 (1), 39–67.

——— (2020b) "An omitted variable bias framework for sensitivity analysis of instrumental variables," *Work. Pap.*

Cornfield, Jerome, William Haenszel, E Cuyler Hammond, Abraham M Lilienfeld, Michael B Shimkin, and Ernst L Wynder (1959) "Smoking and lung cancer: recent evidence and a discussion of some questions," *Journal of the National Cancer institute*, 22 (1), 173–203.

Deryugina, Tatyana, Garth Heutel, Nolan H Miller, David Molitor, and Julian Reif (2019) "The mortality and medical costs of air pollution: Evidence from changes in wind direction," *American Economic Review*, 109 (12), 4178–4219.

Ding, Peng and Fan Li (2018) "Causal inference: A missing data perspective," *Statistical Science*, 33 (2), 214–237.

Dominici, Francesca and Corwin Zigler (2017) "Best practices for gauging evidence of causality in air pollution epidemiology," *American journal of epidemiology*, 186 (12), 1303–1309.

Fogarty, Colin B (2020) "Studentized sensitivity analysis for the sample average treatment effect in paired observational studies," *Journal of the American Statistical Association*, 115 (531), 1518–1530.

Forastiere, Laura, Michele Carugno, and Michela Baccini (2020) "Assessing short-term impact of PM 10 on mortality using a semiparametric generalized propensity score approach," *Environmental Health*, 19 (1), 1–13.

Godzinski, Alexandre and Milena Suarez Castillo (2021) "Disentangling the effects of air pollutants with many instruments," *Journal of Environmental Economics and Management*, 102489.

Grange, Stuart K and David C Carslaw (2019) "Using meteorological normalisation to detect interventions in air quality time series," *Science of The Total Environment*, 653, 578–588.

Grange, Stuart K, David C Carslaw, Alastair C Lewis, Eirini Boleti, and Christoph Hueglin (2018) "Random forest meteorological normalisation models for Swiss PM 10 trend analysis," *Atmospheric Chemistry and Physics*, 18 (9), 6223–6239.

Halliday, Timothy J, John Lynham, and Aureo de Paula (2019) "Vog: Using volcanic eruptions to estimate the health costs of particulates," *The Economic Journal*, 129 (620), 1782–1816.

Hernán, Miguel A and James M Robins (2016) "Using big data to emulate a target trial when a randomized trial is not available," *American journal of epidemiology*, 183 (8), 758–764.

Hill, Jennifer L (2011) "Bayesian nonparametric modeling for causal inference," *Journal of Computational and Graphical Statistics*, 20 (1), 217–240.

Hill, Jennifer and Yu-Sung Su (2013) "Assessing lack of common support in causal inference using Bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes," *The Annals of Applied Statistics*, 1386–1420.

Ho, Daniel E, Kosuke Imai, Gary King, and Elizabeth A Stuart (2007) "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference," *Political analysis*, 15 (3), 199–236.

Hoek, Gerard, Rob Beelen, Kees De Hoogh, Danielle Vienneau, John Gulliver, Paul Fischer, and David Briggs (2008) "A review of land-use regression models to assess spatial variation of outdoor air pollution," *Atmospheric environment*, 42 (33), 7561–7578.

Holland, Paul W (1986) "Statistics and causal inference," *Journal of the American statistical Association*, 81 (396), 945–960.

Imbens, Guido W and Donald B Rubin (2015) *Causal inference in statistics, social, and biomedical sciences*: Cambridge University Press.

Kang, Hyunseung, Benno Kreuels, Jürgen May, and Dylan S Small (2016) "Full matching approach to instrumental variables estimation with application to the effect of malaria on stunting," *The Annals of Applied Statistics*, 10 (1), 335–364.

Keele, Luke and Jason W Morgan (2016) "How strong is strong enough? Strengthening instruments through matching and weak instrument tests," *The Annals of Applied Statistics*, 10 (2), 1086–1106.

King, Gary and Langche Zeng (2006) "The dangers of extreme counterfactuals," *Political analysis*, 14 (2), 131–159.

Mayer, Michael (2019) *missRanger: Fast Imputation of Missing Values*, `https://cran.r-project.org/package=missRanger`, R package version 2.1.0.

Neyman, Jersey (1923) "Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes," *Roczniki Nauk Rolniczych*, 10, 1–51.

OpenStreetMap contributors (2017) "Planet dump retrieved from https://planet.osm.org ," `https://www.openstreetmap.org`.

Padgham, Mark, Bob Rudis, Robin Lovelace, and Maëlle Salmon (2017) "osmdata," *The Journal of Open Source Software*, 2 (14), 10.21105/joss.00305.

Petetin, H, M Beekmann, J Sciare, M Bressi, A Rosso, O Sanchez, and V Ghersi (2014) "A novel model evaluation approach focusing on local and advected contributions to urban PM 2.5 levels–application to Paris, France," *Geoscientific Model Development*, 7 (4), 1483–1505.

R Core Team (2021) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, `https://www.R-project.org/`.

Rosenbaum, Paul (2018) *Observation and experiment*: Harvard University Press.

Rosenbaum, Paul R (2010) *Design of observational studies*: Springer.

Rubin, Donald B (1974) "Estimating causal effects of treatments in randomized and nonrandomized studies.," *Journal of Educational Psychology*, 66 (5), 688.

——— (1991) "Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism," *Biometrics*, 1213–1234.

——— (2008) "For objective causal inference, design trumps analysis," *The Annals of Applied Statistics*, 2 (3), 808–840.

Schlenker, Wolfram and W Reed Walker (2016) "Airports, air pollution, and contemporaneous health," *The Review of Economic Studies*, 83 (2), 768–809.

Schwartz, Joel, Elena Austin, Marie-Abele Bind, Antonella Zanobetti, and Petros Koutrakis (2015) "Estimating causal associations of fine particles with daily deaths in Boston," *American journal of epidemiology*, 182 (7), 644–650.

Schwartz, Joel, Marie-Abele Bind, and Petros Koutrakis (2017) "Estimating causal effects of local air pollution on daily deaths: effect of low levels," *Environmental health perspectives*, 125 (1), 23–29.

Schwartz, Joel, Kelvin Fong, and Antonella Zanobetti (2018) "A national multicity analysis of the causal effect of local pollution, NO 2, and PM 2.5 on mortality," *Environmental health perspectives*, 126 (8), 087004.

Small, Dylan S and Paul R Rosenbaum (2008) "War and wages: the strength of instrumental variables and their sensitivity to unobserved biases," *Journal of the American Statistical Association*, 103 (483), 924–933.

Sommer, Alice J, Emmanuelle Leray, Young Lee, and Marie-Abèle C Bind (2021) "Assessing environmental epidemiology questions in practice with a causal inference pipeline: An investigation of the air pollution-multiple sclerosis relapses relationship," *Statistics in Medicine*, 40 (6), 1321–1335.

Stirnberg, Roland, Jan Cermak, Simone Kotthaus et al. (2021) "Meteorology-driven variability of air pollution (PM 1) revealed with explainable machine learning," *Atmospheric Chemistry and Physics*, 21 (5), 3919–3948.

Stuart, Elizabeth A (2010) "Matching methods for causal inference: A review and a look forward," *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25 (1), 1.

Stuart, Elizabeth A and Donald B Rubin (2008) "Best practices in quasi-experimental designs," *Best practices in quantitative methods*, 155–176.

Tai, Amos PK, Loretta J Mickley, and Daniel J Jacob (2010) "Correlations between fine particulate matter (PM2. 5) and meteorological variables in the United States: Implications for the sensitivity of PM2. 5 to climate change," *Atmospheric environment*, 44 (32), 3976–3984.

Visconti, Giancarlo and José R Zubizarreta (2018) "Handling limited overlap in observational studies with cardinality matching," *Observational Studies*, 4 (1), 217–249.

Wilson, William E and Helen H Suh (1997) "Fine particles and coarse particles: concentration relationships relevant to epidemiologic studies," *Journal of the Air & Waste Management Association*, 47 (12), 1238–1249.

Zabrocki, Léo (2022) "Improving the Design Stage of Air Pollution Studies Based On Wind Patterns," `https://osf.io/7x23u/`.

Zigler, Corwin Matthew and Francesca Dominici (2014) "Point: clarifying policy evidence with potential-outcomes thinking—beyond exposure-response estimation in air pollution epidemiology," *American journal of epidemiology*, 180 (12), 1133–1140.

Estimating the Local Air Pollution Impacts of Cruise Traffic: A Principled Approach for Observational Data

The air pollution and health effects of cruise vessel traffic is a growing concern in the Mediterranean area. We propose a novel methodology based on high-frequency observational data to estimate the causal effects of maritime traffic on air pollution. We apply this method to cruise traffic in Marseille, a large Mediterranean port city. Using a new pair-matching algorithm designed for time series data, we create hypothetical randomized experiments and estimate the change in air pollution caused by a short-term increase in cruise traffic. We carry out a randomization-based approach to quantify uncertainty and compute nonparametric 95% Fisherian intervals (FI). At the hourly level, cruise vessels' arrivals have relatively large impacts on city-level hourly concentrations of nitrogen dioxide, particulate matter and sulfured dioxide. At the daily level, we do not observe any clear effects. Our results suggest that well-designed hypothetical randomized experiments provide a principled approach to evaluate identification challenges in such time series data but also help credibly estimate the negative externalities of maritime traffic.

AUTHORS:
Léo Zabrocki
PSE - EHESS
`leo.zabrocki@psemail.eu`

Marion Leroutier
MISUM - SSE
`marion.leroutier@hhs.se`

Marie-Abèle Bind
Biostatistics Center - MGH
`ma.bind@mail.harvard.edu`

Introduction

Particulate matter pollution induced by maritime traffic was estimated to cause 60,000 premature deaths worldwide in 2007, with the highest burden in the Mediterranean area ctcorbett2007mortality. In the past few years, local environmental organizations and media have raised concerns over air pollution induced by cruise vessel traffic (Friedrich 2017-07-31, Chrisafis 2018-07-6), which peaked in 2018 with four million cruise passengers in the Mediterranean region (Cruise Lines International Association 2019). Although cruise tourism brings with it economic benefits, it could also harm the health of local residents. Due to historical urban planning, many Mediterranean cities have often their port located in the city center and a large fraction of their population could be exposed to vessels' emissions. Besides, the Mediterranean region is not yet part of an Emission Control Area (ECA), unlike US coasts, where stringent reg-

WEBSITE:
`https://lzabrocki.github.io/cruise_air_pollution/`

REPLICATION DATA:
`https://osf.io/v8aps/`

ulations on fuel sulfur content have been implemented. In this context, estimating the impact of cruise traffic on ambient air pollution is required to address public health concerns.

Our study focuses on the city of Marseille, France, which is a perfect example of the air pollution issue at stake in Mediterranean port cities. It is the second largest city of the country, with 870,000 inhabitants, and its second largest port, with 3 million passengers in 2019 (INSEE 2020, GPMM 2020). Pollution levels are high relative to the World Health Organization's recommendations and European legal standards. It is estimated that 1.7% of total annual mortality in Marseille could be avoided if annual $PM_{2.5}$ levels decreased to the WHO recommended thresholds (Khomenko et al. 2021). According to emission inventories, maritime traffic contributes up to 18% of local fine particulate matter emissions (AtmoSud 2020). If cruise traffic contributes to the pollution exposure of residents in similar proportions to its emissions, it could be a key sector to target for improving ambient air quality.

Yet, isolating the contribution of vessel emissions to observed air pollutant concentrations is known to be challenging. Complex meteorological patterns can prevail along coastal sites and ports are often located near major roads and industrial complexes, making it difficult to disentangle the specific amount of air pollution induced by maritime traffic. Atmospheric scientists have typically relied on two complementary approaches to estimate the contribution of vessel traffic to city-level pollution (Mueller et al. 2011, Damien Piga and Salameh 2013, Viana et al. 2014, Liu et al. 2016, Merico et al. 2016, Atmosud 2019, Murena et al. 2018, Liu et al. 2019, Sorte et al. 2020). The first approach is a model-based method. It starts from establishing an emission inventory based on activity data such as the type of engines of the ships arriving in the port, and then infers how emissions turn into concentrations using a dispersion model. The second approach is based on source apportionment methods which require dedicated measurement campaigns with sensors deployed in the city at different seasons. The samples are then analyzed in the laboratory to detect chemical signatures and trace back the likely origin of the particles. The first approach rests on the quality of the emission inventory and the validity of the dispersion model while the second approach is often limited by measurement campaigns of short duration.

The method we develop in this study here differs in several respects and we see it as an alternative to existing approaches found in the atmospheric science literature. It should be more familiar to researchers willing to evaluate the subsequent impact of vessel emissions on various health outcomes. Similarly to Contini et al. (2011) and Moretti and Neidell (2011), we start by combining high-frequency time series data on cruise traffic, weather parameters and air pollutant concentrations over the 2008-2018 period. In contrast to these studies, we then explicitly framed our study within the Neyman-Rubin Causal Model, which enables us to separate the design phase of the observational study from its statistical analysis (Rubin 1974, Holland 1986, Rubin 2005). Using the natural variation in vessel traffic, we try to emulate hypothetical randomized experiments targeted for estimating the impact of a short-term increase in cruise traffic on air pollutants.

To better capture the temporal chemistry of air pollutants reaction, we carry out two types of analysis: one at the hourly level and one at the daily level. We construct pairs of comparable periods of three hours or two days, with and without an increase in cruise traffic, using a recent constrained pair-matching algorithm designed for time series data (Sommer et al. 2018). This new algorithm was developed since other matching approaches such as propensity score can fail to balance the lags of covariates within matched pairs. Our algorithm enables us to set in a flexible manner the maximum distance allowed between treated and control units for each covariate. The lags of covariates will be more balanced within each matched pair by design. Besides, compared to a more standard approach based on a multivariate regression model, matching has several major advantages. First, it adjusts in a nonparametric way for observed covariates such as weather parameters. Second, it helps better evaluate the imbalance and the lack of overlap in observed covariates. As cruise traffic has a strong seasonality, it is important to prune control units which do not belong to the common support of the data to avoid model extrapolation (King and Zeng 2006, Ho et al. 2007, Stuart 2010, Rosenbaum 2010, Imbens 2015, Imbens and Rubin 2015, Rosenbaum 2018). Third, matching is more transparent than a regression approach to understand which observations are used as counterfactuals for treated units. Once we obtain pairs of similar treated and control time series, we assume that the increase in cruise traffic is as-if randomized conditional on a set of observed weather parameters and calendar indicators.

As we shall see, our matching procedures drastically reduces the initial sample sizes. Given the small sample sizes of matched data, we therefore decided to rely on randomization-inference to quantify the uncertainty of the estimated causal effects. We build 95% Fisherian intervals that give the range of constant effects supported by the data. While this mode of inference relies on the unrealistic assumption that the causal effect of an increase of cruise traffic is the same for all matched pairs, it does not make any large-sample approximation and is distribution free for the test statistic of interest (Fisher et al. 1937, Rubin 1991, Ho and Imai 2006, Rosenbaum 2010, Imbens and Rubin 2015, Dasgupta and Rubin 2021). Since the constant unit-level treatment effect assumption is arguably unrealistic in our study, we also provide results using Neyman's mode of inference which focuses on the average treatment effect in matched pairs (Neyman 1923).

Our results show that estimating the causal effects of cruise vessels on local air pollutant concentration is challenging. First, only 4% of hourly treated units and 8% of daily treated units could be

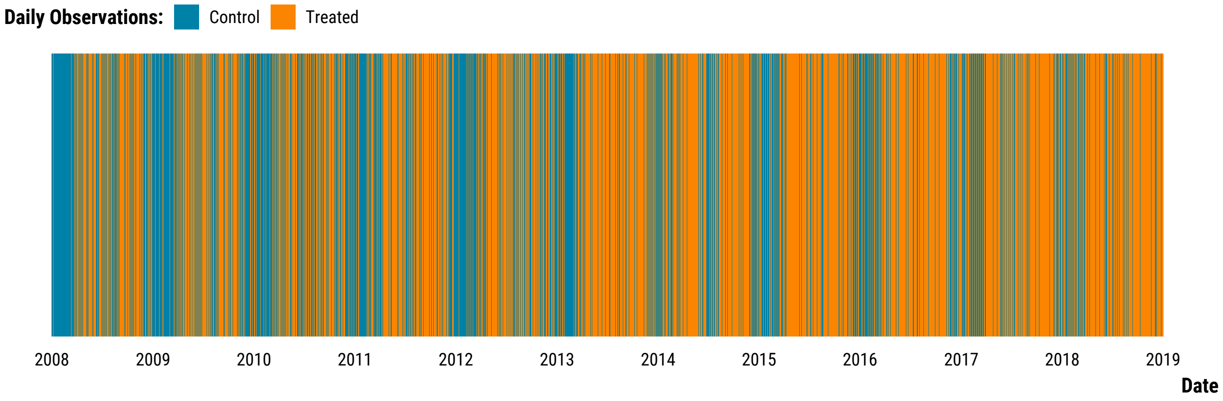


Figure 6: Regularity in Daily Cruise Vessel Traffic. *Notes:* Each blue line is a day without cruise traffic and each orange line is a day with cruise traffic.

matched to similar control units without cruise traffic. This is due to fact that cruise vessel is very regular as we can see in Figure 6, which plots the distribution of days with and without cruise traffic over time. Besides, up to 20% of matched pairs could suffer from spillovers effects since they could be temporally too close from each others. Despite these drawbacks, our matching procedure was successfull to create well-balanced pairs of treated and controls with similar covariate values.

At the hourly level, we find that the arrival of cruise vessel could increase nitrogen dioxide (NO_2) concentrations between 5% up to 25%, coarse particulate matter (PM_{10}) between 3% up 27%, and sulfur dioxide (SO_2) between 4% up to 109%. Ozone concentrations (O_3) simultaneously could decrease up to 14%, which could be consistent with the titration of this air pollutant due to an increase in nitrogen oxide (Diesch et al. 2013, Eckhardt et al. 2013, Merico et al. 2016). Since we have few matched pairs, our 95% Fisherian intervals are imprecise. Contrary to hourly level results, we do not observe any clear impact of cruise traffic at the daily level. This lack of observable effects at the daily level agrees with a measurement campaign carried out by the local air quality monitoring agency in Marseille (Atmosud 2019). At the daily level, it seems that the impact of road traffic on air pollutants largely emitted by cars such as NO_2 is much more visible. The higher salience of plumes emitted by cruise vessels and the potential larger concerns over this source of air pollution is an interesting area for future research. An extensive set of robustness checks on the sensitivity of our results to hidden bias, outliers, missing data and low statistical power complements our main analyses.

The approach we lay out in this paper is the principal contribution to the small literature studying the impact of vessel traffic on air pollution with *observational* data (Moretti and Neidell 2011, Contini et al. 2011, Merico et al. 2016, Sorte et al. 2020). The closest paper to ours is by Zhu and Wang (2021) who study the effects of fuel content regulation on air pollution in four Chinese ports. The implementation of this policy leads to a convincing source of identification based on a time-series regression discontinuity design and a difference-in-discontinuity strategy. However, many port cities around the world do not belong to emission control areas and such research design cannot be implemented to inform future regulation policies. We instead try to make the most of high-frequency time series data on port calls and city air pollution by crafting hypothetical experiments. Our approach should be widely applicable as observational data on weather, air pollution, and port call statistics are easy to access in several port cities and over a long period of time.

Besides, we rely on procedures underused by economists to make the design and analysis stages of observational studies more credible. As reminded by Imbens (2015), compared to an outcome regression approach, matching is a more principled approach to adjust for observed counfouders. The matching algorithm developed in Sommer et al. (2021) has the great advantage to have been specifically designed for time series data since we can flexibly choose the maximum distance between a treated and a control unit for each covariate and its lags. On top of the matching procedure, we also show how to carry out randomization-based inference when the sample size could be deemed too small for traditional inference methods. Even if this mode of inference has recently been the subject of a renewed interest in social sciences (Ho and Imai 2006, Cohen and Dupas 2010, Bowers and Panagopoulos 2011, Gerber and Green 2012, Athey and Imbens 2017, Heß 2017, Bowers and Leavitt 2020) and statistics (Cattaneo et al. 2015, Ding et al. 2016, Keele and Miratrix 2019, MacKinnon and Webb 2020, Caughey et al. 2021, Wu and Ding 2021, Zhao and Ding 2021), it is yet to be adopted by environmental economists. We make great efforts to clearly explain the advantages and drawbacks of this mode of inference but also how to concretely implement it.

Finally, our study could be used as a template to help strengthen the design and analysis stages of the growing literature exploiting exogenous transport shocks to estimate the acute health effects of air pollution (Moretti and Neidell 2011, Schlenker and Walker 2016, Knittel et al. 2016, Bauernschuster et al. 2017, Zhong et al. 2017, Simeonova et al. 2021, Godzinski et al. 2019, Giaccherini et al. 2021). Matching and randomization inference have already proven to be beneficial in air pollution studies on health outcomes which do not rely on natural experiments (Baccini et al. 2017, Forastiere et al. 2020, Sommer et al. 2021, Lee et al. 2021). Even if researchers find credible source of identification, it is important to reveal the common support of the data and clearly lay out the mode of inference (Gutman et al. 2012, Zigler and Dominici 2014, Bind and Rubin 2019, Bind 2019, Bind and Rubin 2021).

The rest of our paper is organized as follows. In the second and third sections, we present our data and describe the research design we rely on. In the fourth section, we present the results and their robustness checks. In the last section, we discuss the advantages but also the limits of our approach and reflect on future paths for research on this topic. We strive to make our analysis easily and fully reproducible. Annotated codes and supplementary materials are available on this website. Our data are archived on a Open Science Framework repository.

Data

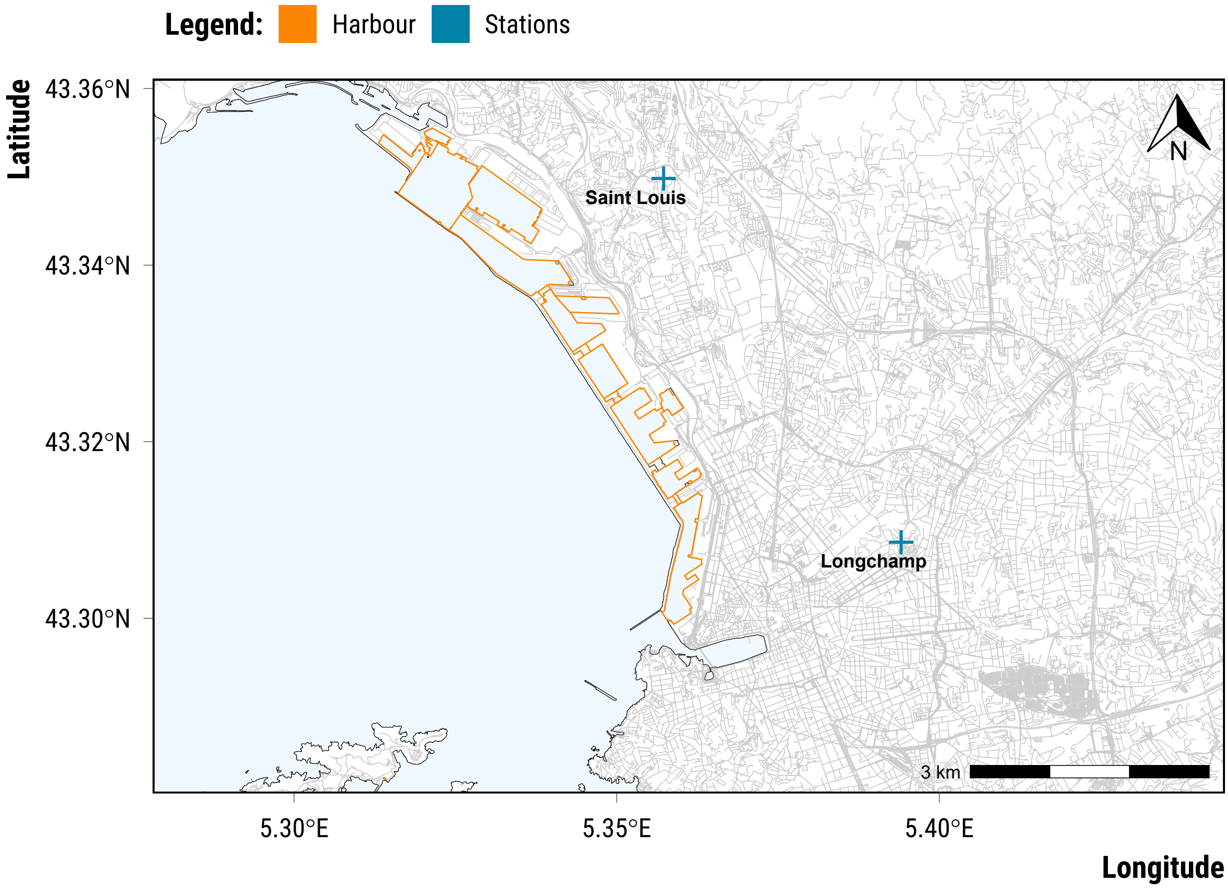


Figure 7: Map of Marseille's Port surrounding Area. *Notes:* This map of Marseille city with its port and the two air quality monitoring stations located in Lonchamp and Saint-Louis neighborhoods. Grey lines represent the road network of the city.

We built two datasets for the 2008-2018 period, one at the hourly level with 96,432 observations, and one at the daily level with 4,018 observations. Below we detail the data sources and variables used. In the Data section of our website, we report additional information on the data wrangling procedure and carry out a full exploratory data analysis (Tufte 1985, Cleveland 1993).

Vessel Traffic Data

We obtained data on 41,015 port calls from the Marseille Port authority. They represent the universe of all port calls between 2008 and 2018. For each vessel docking at the port, we know the exact date and hour of arrival and departure, as well as its name, its type, and its gross tonnage, which is a nonlinear and unitless measure of a vessel's overall internal volume. This measure of a vessel's volume

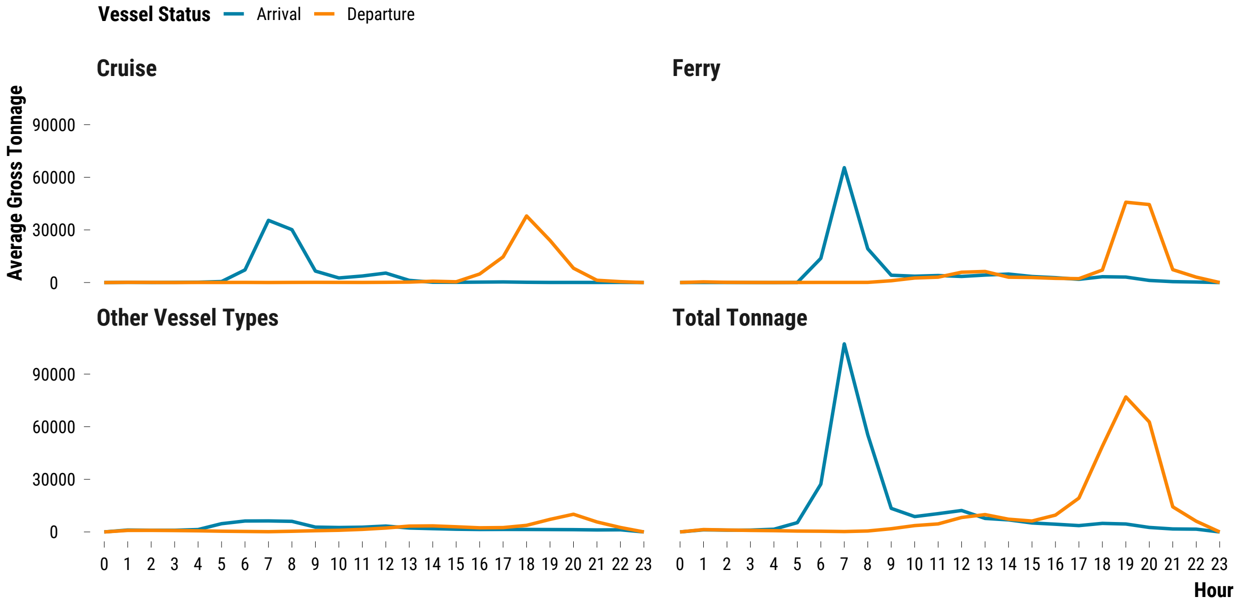


Figure 8: Hourly Vessel Traffic Variation. *Notes:* This figure plots the average hourly variation in the gross tonnage of vessel arriving and departing the port. Gross tonnage is a unitless measure of the volume of a ship.

can be related to its emissions of air pollutants and has been used in other studies as a proxy for the intensity of vessel traffic (Contini et al. 2011, Moretti and Neidell 2011). Using information on vessel characteristics, we defined three broad categories: cruise, ferry, and other types of ships. We then calculated, for each vessel type, the total number of vessels and the sum of gross tonnage entering and leaving the port at the hourly and daily levels. As shown in Figure 8, vessel traffic is regular: most vessels dock in the port in the morning and leave in the evening.

Air Pollution and Weather Data

We retrieved air pollution data from the two background monitoring stations managed by AtmoSud, the local air quality agency. The first station, Saint-Louis, is the closest to the cruise terminal. It is located two kilometers away from the cruise terminal (North-Western extremity of the port) and six kilometers away from the ferry terminal (South-Eastern extremity of the port) (See Figure 7). It only monitors NO_2 and PM_{10}. The second station, Longchamp, is located six kilometers away from the cruise terminal and three kilometers away from the ferry terminal (See Figure 7). The Longchamp station monitors NO_2, SO_2, ozone (O_3), $PM_{2.5}$ and PM_{10}. Sulphur oxides (SOx), nitrogen oxides (NOx), and fine particulate matter are emitted to the atmosphere as a direct result of the combustion of maritime fuel (Sorte et al. 2020). SOx and NOx emissions directly produce NO_2 and SO_2, and contribute to the formation of secondary pollutants such as particulate matter of a larger size (i.e., $PM_{2.5}$ and PM_{10}), and O_3 (Viana et al. 2014).

Weather data come from Météo-France, the French national meteorological service. We obtained data from the closest weather station, located 25 kilometers away from the city center, at Marseille airport. We calculated hourly and daily values for weather variables: rainfall height (mm), average temperature (°C), humidity (%), wind speed (m/s), and wind direction measured on a 360 degrees compass rose where 0° is North.

To avoid losing statistical power, we imputed missing values of cruise gross tonnage, air pollutant concentration and weather parameters. We relied on the chained random forest algorithm provided by the **R** package missRanger (Mayer 2019). There were no clear missingness patterns for these variables and we checked with simulation exercises that this algorithm had a relatively good performance for imputing missing values, even if it could sometimes result in large discrepancies.

Road Traffic Data

We obtained hourly data on the average flow of vehicles and road occupancy rates over the 2011-2016 period from the *Direction Interdépartementale des Routes*, a decentralized state administration in charge of managing, maintaining, and operating roads. We selected hourly data for the six traffic monitoring stations with the best available recordings, two located North and four located East of the city. As measures of road traffic, we focus on the hourly flow of vehicles (number of vehicles) and the occupation rate of the road (%).

Research Design

We conceptualize plausible but hypothetical randomized experiments to estimate the short-term effects of an increase in vessel traffic on air pollutant concentrations in Marseille. We follow a causal inference pipeline conceived to analyze observational data in a rigorous and transparent manner (Rosenbaum 2010, Sommer et al. 2018, Bind and Rubin 2019, Sommer et al. 2021).

Stage 1: Formulating Plausible Interventions on Vessel Traffic

We are interested in the following causal question: *Does cruise vessel traffic contribute to background air pollutant concentrations in Marseille?* The “ideal" experiment would randomly allocate hours or days to high versus low cruise vessel traffic. We could then confidently attribute the resulting differences in pollutant concentrations to vessel emissions. In the absence of such randomized experiment in Marseille, we try to approximate an experimental setting by comparing pairs of short time series that are as similar as possible on a set of *observed* covariates but differ in their level of vessel traffic. We define below our hypothetical randomized experiments using the framework of the Neyman-Rubin Causal Model (Rubin 1974, Holland 1986, Rubin 2005). We conceive two hypothetical experiments:

one experiment at the hourly level to test if an increase in cruise traffic affects hourly air pollutant concentrations in the very short-run; and one experiment at the daily level to examine if an increase in cruise traffic affects daily average concentrations.

The units, which we index by t (t = 1, ..., T), are either hours or days spanning over the 2008-2018 period, depending on the time scale of the experiment considered. At the hourly level, V_t is the sum of the gross tonnage of cruise vessels docking in the port during hour t. We focus on the pollution impact of cruise vessels arriving in the port rather than aggregating arrivals and departures, because the pollution impact of traffic is likely to depend on the direction of the flow. For example, cruise vessels entering the port may take time to finish maneuvering and generate emissions while they are docked. In contrast, cruise vessels leaving the port may start running their engines a few hours before effectively leaving, and therefore generate pollution over a long period of time. Here, we focus on cruise vessels' arrivals. Our treatment indicator is W_t and takes two values:

$$W_t = \begin{cases} 1 & \text{if } V_t > 0 \\ 0 & \text{if } V_t\text{=0} \end{cases} \tag{1}$$

Hourly units with W_t equal to one are considered as "treated" while units with W_t equal to zero belong to the control group. A treated hour is an hour with some cruise vessel arrivals—in practice, no more than two vessels enter the port at a given hour, and more often there is only one. A control hour is an hour with no cruise vessel arriving.

At the daily level, we create a hypothetical randomized experiment easily understandable from a policy point of view. We define N_t as the number of cruise vessels entering Marseille port on day t. Our treatment indicator is W_t and takes two values:

$$W_t = \begin{cases} 1 & \text{if } \mathrm{N}_t = 1 \\ 0 & \text{if } \mathrm{N}_t\text{=0} \end{cases} \tag{2}$$

Daily units with W_t equal to one are considered as "treated" while units with W_t equal to zero belong to the control group. A treated day is a day with one cruise vessel arriving at the port. A control day is a day without any cruise vessel arriving. On average, there is around one cruise vessel entering the port each day in the initial sample. Therefore, the results of this hypothetical experiment can be interpreted as reflecting the contribution of cruise vessel traffic on an average day of the year.

In our setting, each hourly and daily unit has two continuous potential outcomes whose values range in the set of plausible pollutant concentrations in $\mu g/m^3$, $\mathrm{Y}_t(0)$ if W_t=0 and $\mathrm{Y}_t(1)$ if W_t=1. It is important to note that we are working with a multivariate time series. Further assumptions are required to properly define causal effects.

As explained in the following section, our matching algorithm approximates a pairwise randomized experiment by finding similar pairs of short-time series. First, we should check that pairs are well-balanced in terms of interventions occurring in the pre-treatment period. Second, we should make the Stable Unit Treatment Value Assumption (STUVA) plausible (Rubin 1974, Baccini et al. 2017, Forastiere et al. 2020)[2] . In the context of our hypothetical experiments, there must be no spillovers effects within and across matched pairs. Within a matched pair, a treated unit should be temporally far away from a control unit. Across pairs, the first lead outcome of a treated unit in one pair should not be used a control in another pair. This assumption could be harder to make for the hourly experiment since we do not have clear priors on when the treatment would actually occur. For instance, during the maneuvering phase, cruise vessels could already impact air pollutant concentrations before being docked. Once they are docked, they keep their engines on and could emit air pollution in the following hours. In the time series of the treated unit, it is therefore difficult to precisely define which lags and leads of the concentration of an air pollutant is not affected by vessel emission.

Finally, given the definition of our two hypothetical experiments, control days of the daily experiments are made of hours that could serve as control hours in the hourly experiment; and treated days could contain one hour that is treated in the hourly experiment. However, both experiments should be seen as independent from each other, as they aim at testing the pollution impact of cruise traffic at two different time frames. Whether the impact should be stronger for specific air pollutants at the hourly or daily level is ambiguous: a study conducted in Marseille found that maritime traffic's impact on pollution is only detectable at the hourly level and near the port (Atmosud 2019). It could still be argued that cruise traffic may also impact daily concentrations more than hourly concentrations of secondary pollutants due to the lag in their formation.

Stage 2: Designing the Hypothetical Randomized Experiments

At the design stage, our goal is to obtain a sample of similar units for which the assignment to the treatment and control groups can be assumed to be unconfounded (Rubin 1991). Formally, this unconfoundedness assumption states that the assignment to treatment is independent from the potential outcomes given a set of *observed* confounders. Instead of adjusting for confounding variables with a multivariate regression model, we use a novel pair-matching algorithm to obtain treated and control units with similar values for observed covariates (Sommer et al. 2021). Matching is a nonparametric method which prunes the observations to limit the imbalance between treated and control units (Ho et al. 2007, Rubin 2006, Stuart 2010, Imbens 2015). By revealing the common support available in the data, matching avoids the statistical model to extrapolate to units without empirical counterfactual.

Concretely, let $\mathbf{X}_t$ be the vector of observed covariates for each unit, with t the time indicator and $X_t^{(k)}$ the kth covariate. Our algo-

[2] We are currently working to formalize mathematically these assumptions in regards of recent works on the Neyman-Rubin Causal Model applied to time series data (Bojinov and Shephard 2019, Menchetti et al. 2021).

rithm matches a treated unit to a control unit only if the componentwise distances between their covariate vectors ($X_t^{(1)}$, $X_t^{(2)}$, ..., $X_t^{(K)}$) are lower than pre-defined thresholds (δ_1, δ_2, ..., δ_K). For a pair of covariate vectors $\mathbf{X}_t$ and $\mathbf{X}_{t'}$, we use the following distance:

$$\Delta_{\mathbf{X}_t,\mathbf{X}_{t'}} = \begin{cases} 0 & \text{if } |X_t^{(k)} - X_{t'}^{(k)}| < \delta_k \text{ for all k} \\ +\infty & \text{otherwise} \end{cases} \tag{3}$$

Compared to a propensity score approach, we can make sure with this algorithm that observed confounders and their lags are balanced within pairs (Greifer and Stuart 2021). To limit confounding, we select two sets of covariates. First, calendar variables (i.e., hour of the day, day of the week, bank day, holidays, month, and year) are related to both vessel traffic and air pollution. Second, weather covariates (i.e., average temperature, rainfall indicator, average humidity, wind direction blowing either from the East or West, and wind speed) could also influence both vessel traffic and air pollution. We use lags of these variables to ensure that treated and control units are as similar as possible before the treatment occurs. We define matching thresholds noting that they should be strict enough to make treated and control units comparable with each other, but not too strict to avoid reducing the sample size too much. Given this trade-off, the thresholds are stricter for the hourly experiment for which the sample size is 24 times larger. Table 1 displays all threshold values used in our matching procedure.

At the hourly level, we match exactly on calendar variables (hour of the day, day of the week, bank days, holidays) over the current and two previous hours before the treatment occurred (i.e., 0, 1, 2 lags) and allow a maximum distance of 30 days between treated and control units. For weather parameters, we carried out an iterative process, for which we tried different discrepancy values and kept the ones that led to balanced treated and control groups while resulting in enough matched pairs. We found that a maximum discrepancy of around half a standard deviation often yields a good balance. We match exactly for the East and West wind directions because they play an important role in the dispersion of air pollutants.

At the daily level, we create similar pairs of treated and control units over the current and previous day before the treatment occurred (i.e., 0 and 1 lags). We relax some of the constraints from the hourly level to have enough matched pairs. We strictly match on the day of the week, bank days, and holidays over the two days of the series. We allow treated and control units to have up to three years of difference, but they should belong to the same month. For weather parameters, we match exactly on the rainfall indicator and the wind direction on days *t* and *t-1*, and we allow a small discrepancy threshold for temperature and wind speed on *t* and *t-1*.

Based on these thresholds, each treated unit is matched to its closest control unit using a maximum bipartite matching algorithm (Micali and Vazirani 1980). If no control unit is available to match a treated unit, it is discarded. We thus approximate the design of a pairwise randomized experiment where the assignment mechanism is a Bernoulli trial with a treatment probability of 0.5. Given this design, for each hypothetical experiment, the number of possible permutations is 2^P, with P being the number of matched pairs.

Table 1: Maximum Discrepancies allowed for each Covariate between Treated and Control Units, Hourly and Daily Experiments.

	Hourly Experiment	**Daily Experiment**
Calendar Indicators		
Distance in days	30	1095
Hour of the day in *t*	0	
Weekday, Bank Days and Holidays in *t*	0	0
Weekday, Bank Days and Holidays in *t-1*	0	0
Weekday, Bank Days and Holidays in *t-2*	0	
Month in *t*		0
Weather Parameters		
Average Temperature (°C) in *t*	4	4
Average Temperature (°C) in *t-1*	4	4
Average Temperature (°C) in *t-2*	4	
Rainfall Dummy in *t*	0	0
Rainfall Dummy in *t-1*	0	0
Rainfall Dummy in *t-2*	0	
Average Humidity (%) in *t*	9	
Average Humidy (%) in *t-1*	9	
Average Humidity (%) in *t-2*	9	
Wind direction in 2 categories (East/West) *t*	0	0
Wind direction in 2 categories (East/West) *t-1*	0	0
Wind direction in 2 categories (East/West) *t-2*	0	
Wind speed (m/s) in *t*	1.8	2
Wind speed (m/s) in *t-1*	1.8	2
Wind speed (m/s) in *t-2*	1.8	

Notes: This table displays the maximum distance allowed for each covariate in the pair matching algorithm, for each experiment. For example, it means that, for each matched pair, treated and control units must have the same values for weekday, bank days and holidays indicators in *t*. If a discrepancy value is missing in one of the two column, it means that the associated covariate was not used for matching for the corresponding experiment.

Stage 3: Analyzing the Experiments using Randomization-based Inference

Once we obtained a balance sample of matched pairs, we implement a randomization-based inference procedure to analyze the effects of cruise vessels on air pollutant concentrations. Given that we have a low number of matched pairs, we rely on this particular mode of inference since it avoids large-sample approximation and is distribution-free.

Point estimate for the unit-level treatment effect size. We assume a constant additive unit-level treatment effect τ:

$$Y_t(1) = Y_t(0) + \tau \ \forall t = 1, \dots, T \tag{4}$$

Under such assumption, the average pair difference in pollutant concentrations across treated and control units is an unbiased estimator for τ (Keele et al. 2012). Thus, for an experiment with I_{Pairs} matched pairs, where $Y_{1,i}^{\text{obs}}$ is the observed pollutant concentration for the treated unit of pair i and $Y_{0,i}^{\text{obs}}$ is the observed pollutant concentration for the control unit of pair i, we take as a point estimate the observed value of the average pair differences:

$$\hat{\tau} = \frac{1}{I_{\text{Pairs}}} \sum_{i=1}^{N_{\text{Pairs}}} (Y_{1,i}^{\text{obs}} - Y_{0,i}^{\text{obs}}) \tag{5}$$

Randomization-based quantification of uncertainty. We carry out a test-inversion procedure to build 95% Fisherian (also called "Fiducial") Intervals (FI) for the constant unit-level treatment effect. We closely follow the procedure detailed by T. Dasgupta and D.B. Rubin in their forthcoming book (Dasgupta and Rubin 2021). On our website, we provide a very detailed toy example to explain this mode of inference. Instead of gauging a null effect for all units, we test J sharp null hypotheses H_0^j: Y$_t$(1) = Y$_t$(0) + τ_j for j =1,. . ., J, where τ_j represents a constant unit-level treatment effect size. We test a sequence of sharp null hypotheses of constant treatment effects ranging from -10 µg/m^3 to +10 µg/m^3 with an increment of 0.1 µg/m^3. As a test-statistic, we use the observed value of the average of pair differences, which is commonly used in randomization-based inference (Keele et al. 2012, Imbens and Rubin 2015). For each constant treatment effect j, we calculate the upper p-value associated with the hypothesis H_0^j: Y$_t$(1) - Y$_t$(0) $>$ τ_j and the lower p-value for H_0^j: Y$_t$(1) - Y$_t$(0) $<$ τ_j. We run 10,000 permutations for each hypothesis to approximate the null distribution of the test statistic. Running the exact number of possible allocations is computationally too intensive given the number of matched pairs we found. The results of testing the sequence of J hypotheses H_0^j: Y$_t$(1) - Y$_t$(0) $>$ τ_j forms an upper p-value function of τ, $p^+(\tau)$, while the sequence of alternative hypotheses H_0^j: Y$_t$(1) - Y$_t$(0) $<$ τ_j makes a lower p-value function of τ, $p^-(\tau)$. To calculate the bounds of the 100(1-α)% Fisherian interval, we solve $p^+(\tau) = \frac{\alpha}{2}$

for τ to get the lower limit and $p^{-}(\tau) = \frac{\alpha}{2}$ for the upper limit. We set our α significance level to 0.05, and thus calculate two-sided 95% Fisherian intervals. This procedure allows us to get the range of constant treatment effects consistent with our data, and the hypothetical assignment mechanism we posit (Rosenbaum 2010, Dasgupta and Rubin Fall 2015).

Drawbacks of randomization inference. Many researchers restrain from using randomization inference as a mode of inference since it assumes that treatment effects are constant across units. In our study, this is arguably an unrealistic assumption since it would imply that the effect of cruise vessel on air pollutant concentrations is the same for all units. To overcome this limit, we carry out two other quantification of treatment effects uncertainty.

First, we can compare the results of the randomization inference procedure with the ones we would obtain with Neyman's approach (Neyman 1923). In that case, the inference procedure is built to target the average causal effect and the source of inference is both the randomization of the treatment and the sampling from a population. We can estimate the finite sample average effect, τ_{fs}, with the average of observed pair differences $\hat{\tau}$:

$$\hat{\tau} = \frac{1}{I} \sum_{i=1}^{J} (Y_{\text{t},i}^{\text{obs}} - Y_{\text{c},i}^{\text{obs}}) = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}$$

Here, the subscripts t and c respectively indicate if the unit in a given pair is treated or not. I is the number of pairs. Since there are only one treated and one control unit within each pair, the standard estimate for the sampling variance of the average of pair differences is not defined. We can however compute a conservative estimate of the variance, as explained in chapter 10 of Imbens and Rubin (2015):

$$\hat{\mathbb{V}}(\hat{\tau}) = \frac{1}{I(I-1)} \sum_{I=1}^{I} (Y_{\text{t},i}^{\text{obs}} - Y_{\text{c},i}^{\text{obs}} - \hat{\tau})^2$$

We finally compute an asymptotic 95% confidence interval using a Gaussian distribution approximation:

$$\text{CI}_{0.95}(\tau_{\text{fs}}) = \left(\hat{\tau} - 1.96 \times \sqrt{\hat{\mathbb{V}}(\hat{\tau})},\ \hat{\tau} + 1.96 \times \sqrt{\hat{\mathbb{V}}(\hat{\tau})}\right)$$

Second, Wu and Ding (2021) recently propose to adopt a studentized test statistic that is finite-sample exact under sharp null hypotheses but also asymptotically conservative for weak null hypotheses (i.e., average treatment effects). In our case, this studentized test statistic is equal to the observed average of pair differences divided by Neyman's standard error of a pairwise experiment. We therefore follow the same previous procedure for computing Fisherian intervals but use the studentized statistic.

Stage 4: Robustness Checks

We carry out several robustness checks to evaluate different aspects of the design and results of our study.

Randomization check for overall balance. During the matching procedure, we assess the balance with Love plots that display for each covariate the standardized difference in means between treated and control units before and after matching. To better assess the overall balance, we implement the randomization inference method developed by Branson (2021) to evaluate if the treatment indicator is as-if randomized according the pairwise structure in the matched data. As a test statistic, Branson (2021) proposes to use the Mahalanobis distance which summarizes the imbalance in the means of all covariates but also takes into account their joint relationships. The randomization inference procedure consists in permuting the treatment indicator many times, computing the Mahalanobis distance for each iteration and plotting the resulting null distribution of the test statistic. If the observed value of the Mahalanobis distance is far away from the distribution, it means that the treatment indicator is not as-if randomized according to observed covariates.

Sensitivity to hidden bias. The causal interpretation of our results is based on the plausibility of the hypothetical experiment and the unconfoundedness assumption (Rubin 1991). This is a strong assumption as it states that the treatment assignment probability is not a function of potential outcomes given observed and *unobserved* confounding factors (Sekhon 2009). To evaluate the consequence of hidden bias, we rely on the sensitivity analysis framework developed by (Rosenbaum 1987; 2010). The goal of this method is to quantify how the treatment estimates would be altered by the effect of an unobserved confounder on the treatment odds, denoted by Γ. In our matched pairwise experiments, we make the assumption that within each pair, control and treated units have the same probability of 0.5 to be treated, that is to say to have a positive shock on cruise traffic. The odds of treatment is such that $\Gamma = (0.5/(1-0.5))/(0.5/(1-0.5)) = 1$. As explained in Rosenbaum (2010), we can implement a randomization inference procedure to compute the 95% Fisherian intervals obtained for a given value of bias that the unmeasured confounder has on the treatment assignment. For instance, if we assume that an unmeasured confounder has a small effect on the odds of treatment (i.e., for a $\Gamma > 1$ and close to 1) but the resulting 95% Fisherian interval is consistent with negative, null and positive effects, it would imply that our results are highly sensitive to hidden bias. Conversely, if we assume that an unmeasured confounder has a strong effect on the odds of treatment (i.e., for a large Γ) and we find that the resulting 95% Fisherian interval remains similar, it would strength our view that results do not suffer from hidden bias. Again, the method of Rosenbaum (2010) relies on the assumption of constant additive treatment effects, which is unrealistic in our study. To overcome this limit, we implement the new method proposed by Fogarty (2020) which extends the sensitivity analysis for sample average treatment effects. In a complementary evaluation of hidden bias, we also check whether unmeasured biases could be present by using the lags of air pollutant concentrations as placebo/control outcomes (Imbens and Rubin 2015). If our matched pairs are indeed similar in terms of unobserved covariates, the treatment occurring in t should not influence concentration of air pollutants in the first lag at the daily level and concentrations for further lags at the hourly level.

Sensitivity of results to outliers and missing observations. In the matched data, the observed concentration of air pollutants are sometimes very high. To make sure that our results are not influenced by outliers, we run again our randomization inference procedure with the Wilcoxon signed-rank statistic. If D_i is the observed difference in concentrations between the treated and control unit of pair i for a given pollution outcome, the Wilcoxon signed-rank statistic is defined as $T = \sum_{i=1}^{I} sgn(D_i) \times q_i$, where $sgn(D_i) = 1$ if $D_i > 0$ and $sgn(D_i) = 0$ if $D_i \leq 0$, and q_i is the rank of $|D_i|$ (Rosenbaum 2010). Besides, we imputed missing values and we could fear that their imputations affect the results. For instance, at the hourly, up to 25% of the pairs have missing values for an air pollutant. Our simulation exercise also shows that large imputation errors sometimes occur. We therefore run again our randomization inference procedure for pairs with observed air pollutant concentrations.

Low statistical power and inflation of statistically significant estimates. In the hourly and daily experiments, our matching procedure resulted in few matched pairs, which decreases the precision of our treatment effect estimates. Moreover, if our statistical power is low and we obtain a "statistically significant" effect, we have a higher chance that this estimate is of the wrong sign (Type S error) and/or overestimates the true effect of vessel traffic on air pollutant concentrations (Type M error) (Gelman and Carlin 2014, Gelman et al. 2020). Here we carry out retrospective power calculations to evaluate the risks of making type-S and type-M errors. While it is impossible to know what the true effect of cruise vessels on an air pollutant is, we can calculate the statistical power and the risks to make type S and M errors under a set of plausible effect sizes using the closed-form expression derived by Lu et al. (2019) and implemented in the `retrodesign` R package by Timm (2019).

Indirect treatment effect of cruise traffic. One issue of our design could be the presence of an indirect treatment effect due to the increase in road traffic induced by cruise vessel passengers and its subsequent effects on air pollution. This is part of the causal effect that we want to capture but it is not the proper causal effect of cruise vessel emissions. We therefore check if road traffic measures are balanced before and after the treatment occurs.

Strictness of the matching procedure. Our matching procedure is strict and result in a small number of matched pairs, both at the hourly and daily levels. This is partly due to the regularity in vessel traffic which makes it hard to find control units that are temporally close to treated units and with similar covariate values. To relax the stringency of our matching procedure, we implement a propensity score matching procedure where each treated unit is matched to its closest control unit if their distance is less than 0.01 of the standard deviation of the propensity score distribution.

Comparison with an outcome regression approach. Finally, we compare our results to estimates found using a simple multivariate regression model on the initial hourly and daily datasets. It is however important to keep in mind that the matched datsets are sub-samples of initial datasets with different covariate values. Estimated effects are therefore not directly comparable. For each experiment, we run the following model:

$$p_{t+j} = \alpha + \beta W_t + \mathbf{X}_t \gamma + \mathbf{C}_t \theta + \epsilon_t$$

where j is the index of the lag or lead, t is either the hour (for the two hourly experiment) or the day index (for the daily experiment), p_{t+j} the concentration of an air pollutant p at date $t+j$, W_t the binary treatment indicator, $\mathbf{X}_t$ the vector of weather covariates, which include the average temperature, the squared of the average temperature, an indicator for the occurrence of rainfall, the average humidity, the wind speed, the wind direction divided in the four principal directions (North-East, South-East, South-West, North-West), $\mathbf{C}_t$ the vector of calendar variables, which are indicators for the hour of the day (for the hourly experiment), the day of the week, bank days, holidays, month, year and the interaction of these last two variables, and ϵ_t an error term. We run this simple model from lag 3 to lead 3 of an air pollutant for the hourly experiment on vessels' arrivals and from lag 1 to lead 1 for the daily experiment.

Results

In this section, we first present covariate balance diagnostics on the performance of our matching procedure. We then display and interpret the results for the effects of hourly and daily cruise vessel traffic on air pollutant concentrations. We end the section with our set of robustness checks.

Matching Results

Hourly matching diagnostics. As shown in Table 2, our matching procedure at the hourly level results in few matched treated units, with less than 4% of treated units matched to similar control units. Two main reasons explain this result. First, cruise vessel traffic is regular over time, so that it is hard to find similar control and treated hours which are not temporally too far away from each other. Second, even if we relax our matching constraints, it is difficult to find treated and control units with similar weather covariates. We check that within pairs spillovers are not likely to occur since within a pair, treated and control units are at least 7 days away. However, there could be spillovers across pairs. For instance, for 16% of treated units, the minimum distance with a control unit in another pair is inferior or equal to 5 hours. Dropping these pairs or modifying the matching algorithm to avoid having pairs too close temporally of each others would be required to avoid spillover effects.

Table 2: Number of Matched Pairs by Experiment.

	Hourly Cruise Experiment	**Daily Cruise Experiment**
N_{Total}	96,432	4,018
N_{Treated}	4,034	2,485
N_{Control}	92,396	1,532
N_{Pairs}	138	189

Notes: This table displays the total number of observations, N_{Total} for each experiment, the number of potential treated and controls units before matching, N_{Treated} and N_{Control}, and the number of matched pairs, N_{Pairs}.

In Figure 9, Panel A displays the average increase in cruise vessel arrivals at hour 0. The average difference in gross tonnage between treated and control units is about 65,000 for the hourly cruise experiment, which is the average gross tonnage of one cruise vessel. Panel B shows that, on average, treated and control units have similar vessel traffic for other vessel types and flows. The matching procedure at the hourly level improves the overall balance of covariates as shown in Figure 10. The balance of covariates is also confirmed by the randomization inference procedure advocated by Branson (2021). Further diagnostics on covariates balance are available at the hourly level on our website.

Finally, compared to the initial data, matched hours are more likely to fall on spring and summer days, which are hotter on average. They fall disproportionately around 7 am, the time where cruise vessels tend to arrive in Marseille port.

Daily matching diagnostics. At the daily level, we found 189 matched pairs, which means that 8% of the treated units were matched to similar control units. There should not be within pair spillovers since treated and control units are at least 7 days away. However, as in the hourly experiment, there could be spillovers across pairs if cruise vessel emissions impact the first lead of air pollutant concentrations. For 22% of treated units, the minimum distance with a control unit in another pair is equal to one day.

In Panel A of Figure 11, the average difference in gross tonnage between treated and control units is around 150,000, which corre-

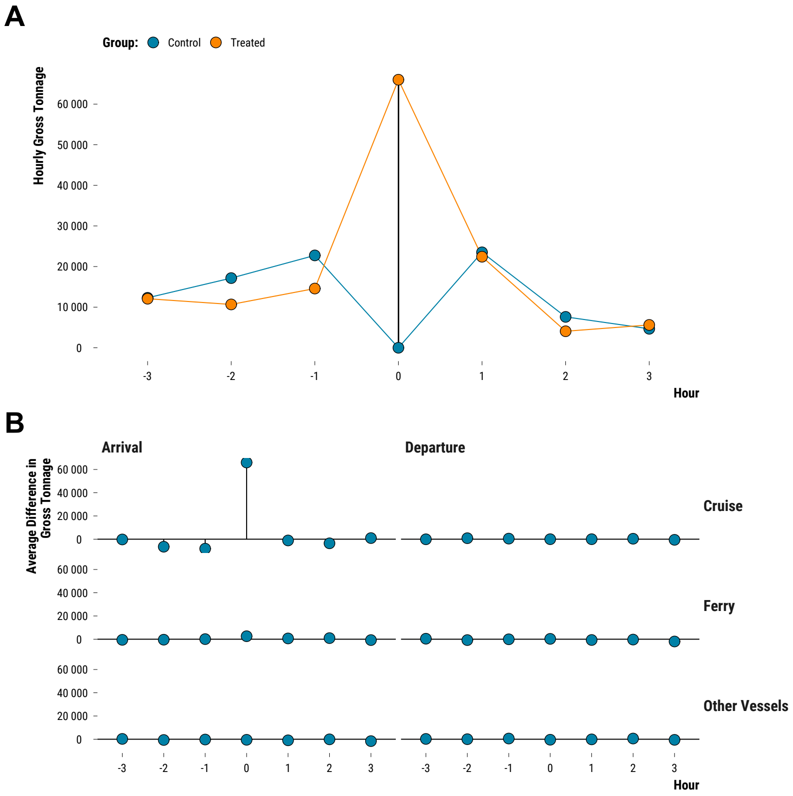


Figure 9: Intervention Diagnostics for the Hourly Experiment. *Notes:* Panel A shows the average hourly total gross tonnage for matched treated and control units. Panel B plots the average difference in total gross tonnage between treated and control units by vessel type and flow.

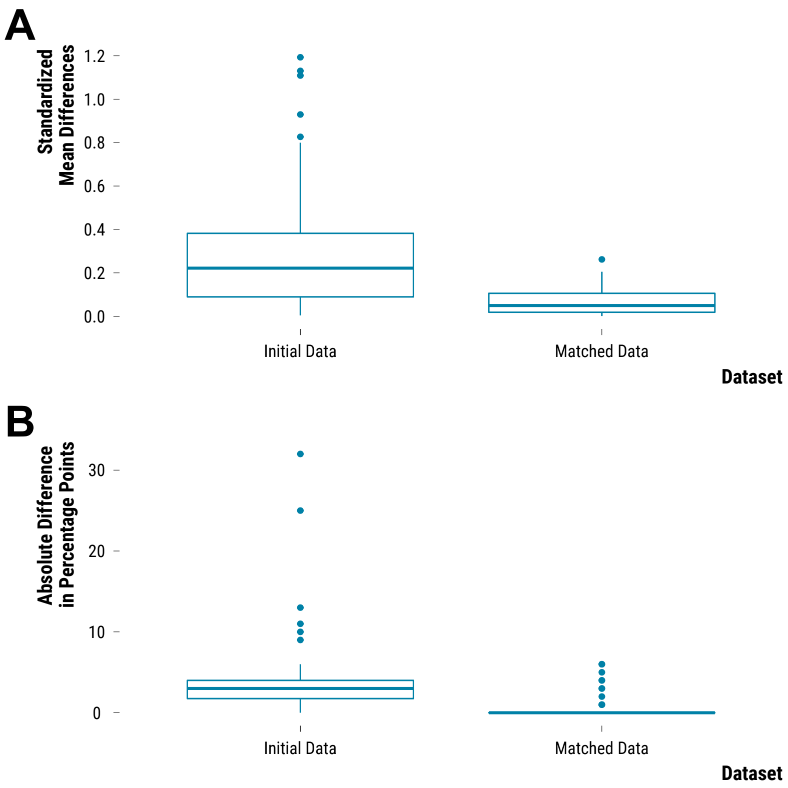


Figure 10: Improvement in Covariates Balance for the Hourly Experiment. *Notes:* Panel A shows the boxplot distribution of the absolute standardized mean differences in continuous covariates before and after matching. Panel B shows the boxplot distribution of the absolute mean differences in categorical covariates before and after matching.

sponds to the tonnage of two vessels. The cruise vessel entering the port in the morning most likely leaves the port in the evening after docking at the port during the day. The variation in gross tonnage for other vessel types is similar across treated and control units, as shown in Panel B of Figure 11. Similarly to the hourly experiment, the matching procedure improved the balance of covariates, as shown in Figure 12. The randomization check for the overall balance also supports the as-if randomization of the treatment. Further diagnostics on covariates balance are available at the daily level on our website.

As for the hourly experiment, days in the matched sample are more often in summer so that they are hotter, less rainy and with a lower wind speed than the average day from the initial sample.

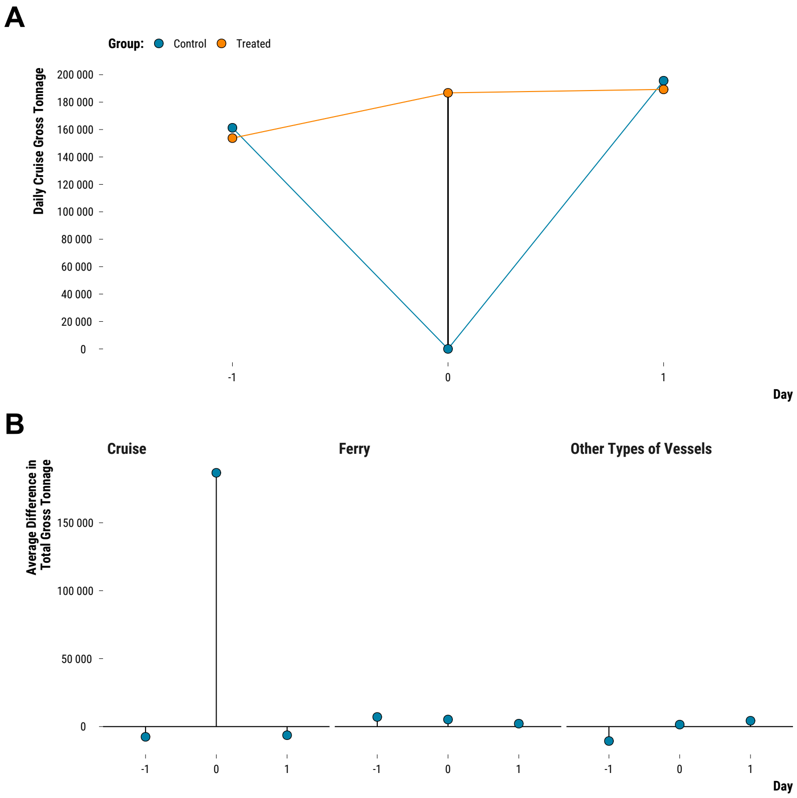


Figure 11: Intervention Diagnostics for the Daily Experiment. *Notes:* Panel A shows the average daily total gross tonnage for matched treated and control units. Panel B plots the average difference in total gross tonnage between treated and control units by vessel type and flow.

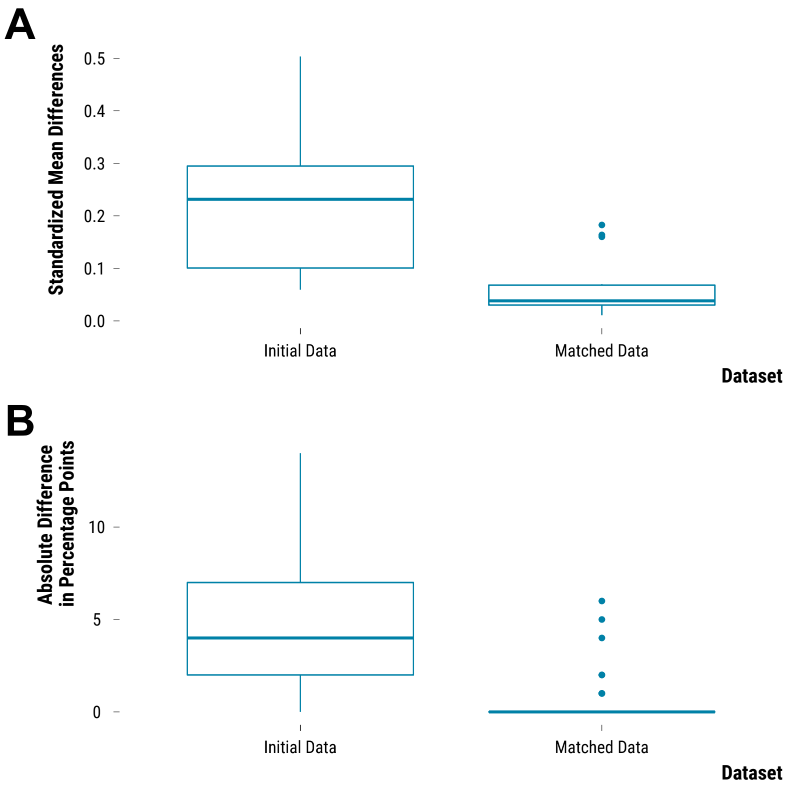


Figure 12: Improvement in Covariates Balance for the Daily Experiment. *Notes:* Panel A shows the boxplot distribution of the absolute standardized mean differences in continuous covariates before and after matching. Panel B shows the boxplot distribution of the absolute mean differences in categorical covariates before and after matching.

The Effects of Cruise Vessel Traffic on Air Pollutants

Hourly Effects. In Figure 13, we plot the point estimates and the 95% Fisherian intervals of the constant treatment effects on air pollutant concentrations that are consistent with our data. We compute these effects for the three previous hours before the treatment occurs up to the three following hours in order to capture the impacts of emissions during the maneuvering phase of cruise vessels but also while they are docked with their engines on.

For NO_2, we observe an increase in concentration from the second previous hour up to the second following hour. The pattern is clearer for the Longchamp station than the Saint-Louis station where the signal is more noisy. At hour 0, concentrations are higher by 4.7 µg/m^3 (95% FI: [1.4, 8.0]). In relative terms, this represents a 16% increase in the average hourly concentration of NO_2 measured at Longchamp station. The 95% Fisherian are relatively wide since the data are consistent with constant effects ranging from a 5% increase up to a 27% increase in concentration.

For O_3, we see the opposite relationship since there seems to be a decrease in concentration in the three previous hours, followed by an increase in the following hours. In hour 0, there is a constant decrease of O_3 concentrations by 3.8 µg/m^3 (95% FI: [-7.6, 0.0]). This is equivalent to a 7% decrease in the average hourly concentration of the air pollutant. Again, the 95% Fisherian intervals are wide: the data are consistent with null effects up to a 14% decrease.

For SO_2, we observe an increase in concentrations of 1.2 µg/m^3 at hour 0 (95% FI: [-0.1, 2.5]), which persists over the two following hours. The constant increase is equivalent to a very large relative increase in concentration by 52%. The 95% Fisherian interval is also wide since the data are consistent with a relative decrease of 4% up to a relative increase of 109%.

For particulate matter, there are no very clear patterns for PM_{10} and $PM_{2.5}$ concentrations measured at Longchamp station. However, we observe an increase in PM_{10} concentrations measured at Saint-Louis that is followed by a decrease. At hour 0, the constant increase is equal to 4.6 µg/m^3 (95% FI: [0.9, 8.3]). This is equivalent to a 15%: the data are consistent with relative increase from 3% up to 27%.

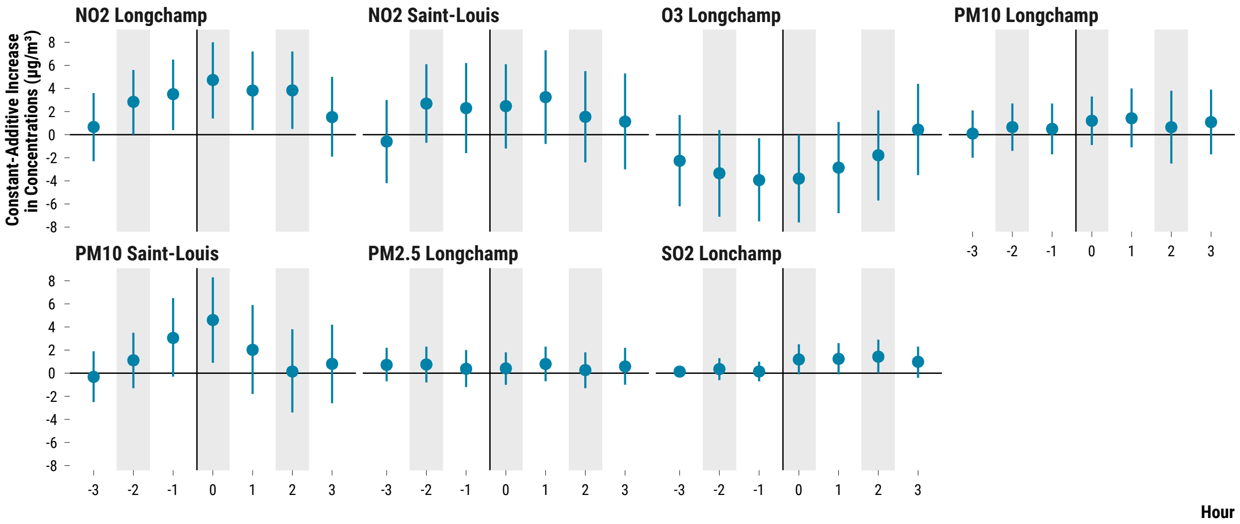


Figure 13: Effects of Cruise Vessel Traffic on Pollutant Concentrations at the Hourly Level. *Notes:* The treatment occurs at hour 0. Dots represent the point estimate of the unit-level treatment effect on a pollutant concentration. Lines are 95% Fisherian intervals of constant treatment effects consistent with the data. The effects are plotted from the third lag to the third lead.

Daily Effects. Figure 14 shows the results for the daily experiment. We can see that are no clear patterns for all air pollutants. The 95% Fisherian intervals are relatively large. For instance, if the point estimate for the constant effect on NO_2 in Longchamp is nearly null, the data are consistent with effects ranging from a 6% decrease up to a 5% increase.

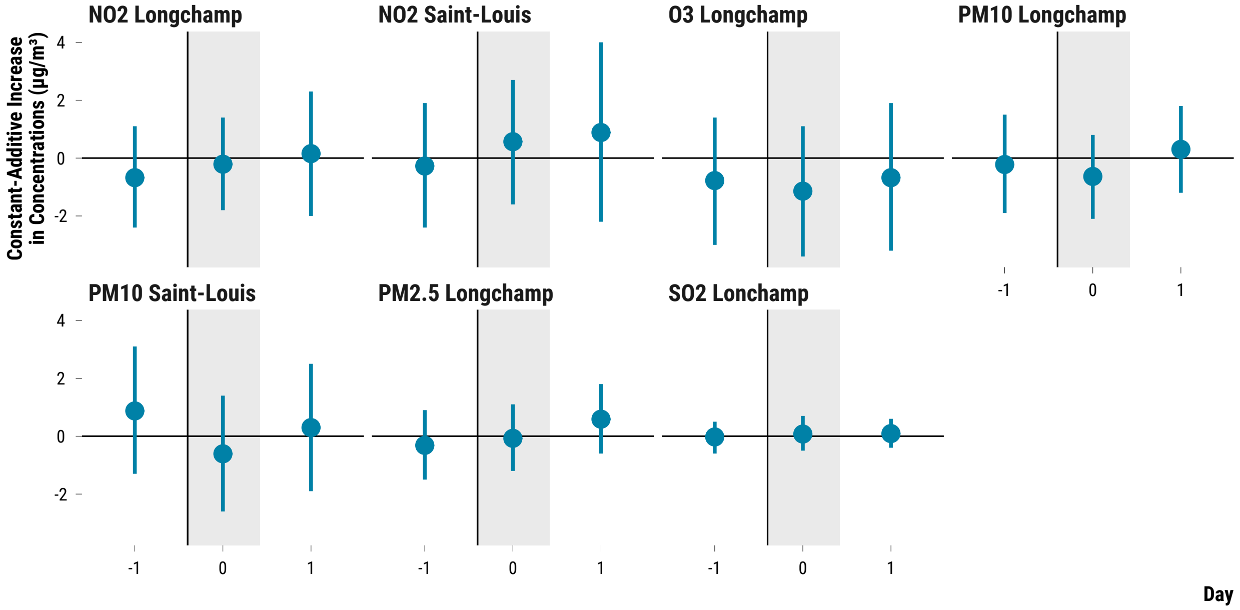


Figure 14: Effects of Cruise Vessel Traffic on Pollutant Concentrations at the Daily Level. *Notes:* The treatment occurs at day 0. Dots represent the point estimate of the unit-level treatment effect on a pollutant concentration. Lines are 95% Fisherian intervals of constant treatment effects consistent with the data. The effects are plotted from the first lag to the first lead.

Neyman's approach and randomization inference for average treatment effects. For the hourly and daily experiments, the 95% Fisherian intervals for constant treatment effects are very similar to the intervals for the average treatment effects computed with Neyman's approach. They are also very similar to those found with the studentized randomization inference that is conservative for weak null hypotheses. With these two alternative mode of inference, we can also confidently interpret the previous 95% Fisherian intervals as the range of *average* treatment effects consistent with the data. These results are available on our website.

Heterogeneity Analysis. We carry out two heterogeneity analyses for the hourly and daily experiments. First, depending on the wind direction, the effects of cruise vessel emissions on air pollutant concentrations are likely to be attenuated or increased. At the hourly level, we observe stronger differences in concentrations for all air pollutant when the wind is blowing from the West, that is to say when vessel emissions are more likely to spread over the city. At the daily level, there are no clear patterns.

Second, we also visually explore the relationship between pair differences in air pollutant concentrations against the pair differences in gross tonnage. Ideally, we should see a positive relationship since the higher the pair difference in gross tonnage (i.e., the higher the treatment shock is), the larger the pair difference in concentrations should be. We do not see any clear patterns, both for the hourly and daily experiments.

Robustness Checks

Sensitivity to hidden bias. Our sensitivity analysis reveals that the estimated effects of cruise vessels emissions on NO_2 concentrations at Longchamp station and PM_{10} concentrations at Saint-Louis station could be affected by a relatively weak hidden bias. Concretely, if we fail in our matching procedure to adjust for an unobserved confounder which is 1.5 times more common among treated units, the resulting 95% Fisherian interval for the effects on NO_2 ranges from -1.5 $\mu g/m^3$ to 11.4 $\mu g/m^3$ and the intervals for the effects on PM_{10} ranges from -1.9 $\mu g/m^3$ to 12.2 $\mu g/m^3$. Our data would be still consistent with mostly positive effects of cruise vessel emissions on these two air pollutants but they could be null and even negative. It is however hard to think about an unobserved confounder which would change the odds of treatment by 50%. To complement this sensitivity analysis, we also note that there are no differences in the first lag of air pollutant concentrations for the daily experiment. For the hourly experiment, we also see that for further lags and leads, estimated differences in air pollutant concentration are mostly null.

Sensitivity of results to outliers and missing observations. First, because the pair differences in pollutant concentrations were particularly disperse, we use the Wilcoxon signed-rank test statistic, known to be less sensitive to outliers. The 95% Fisherian intervals obtained with this test statistic are similar to those obtained with the average of pair differences. Second, we reproduce the analysis on non-missing concentrations because up to 20% of pollutant concentrations were imputed in our matched data. We find similar results with slightly wider 95% Fisherian intervals.

Low statistical power and inflation of statistically significant estimates. Our matching procedure resulted in few matched treated units: we might therefore have a low statistical power to detect the effect of cruise vessels on air pollutant concentrations. More worryingly, when a study is under-powered, we have a higher chance to obtain a "statistically significant" estimate of the opposite sign of the true effect (Type S error). "Statistically significant" estimates also tend to exaggerate the true effect size (Type-M error). While we do not know what the true effect of cruise on air pollutants is, we can explore our statistical power and the risk that "statistically significant" estimates are misleading. For instance, we observe a 4.7 $\mu g/m^3$ increase in NO_2 concentrations in Longchamp due to cruise vessel arrivals. If other researchers think that this effect size is too large, we can retrospectively compute the power of our study according to a range of alternative true effect size. In Figure 15, if we assume that the true effect is equal to +2.35 µg/m^3 (dashed line), our study would have a power of 30% and "statistically significant" estimates would be on average 1.8 times too large. However, the probability that a "statistically significant" estimate is of the opposite sign is nearly null. For other air pollutants for which 95% Fisherian intervals are wider, this risk could be high. With the few number of matched pairs found in our hypothetical experiments, there is clearly a chance that "statistically significant" estimates could be misleading: as we did, we should rather interpret the width of the 95% Fisherian intervals.

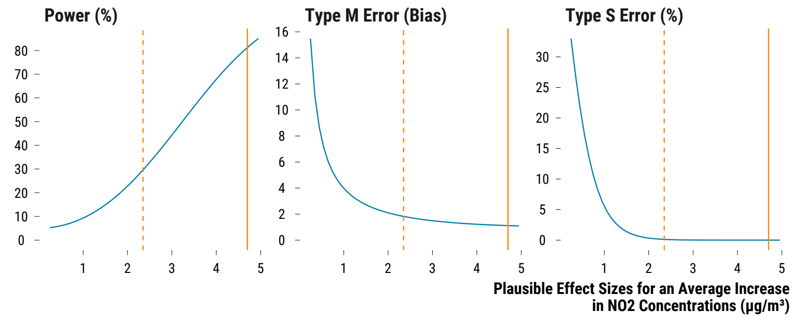


Figure 15: Statistical Power, Type M and S Errors for Hourly Experiment Effect on NO_2. *Notes:* In the first left panel, the statistical power of the hourly experiment on the effect of cruise vessel on NO_2 is plotted against hypothetical true effect sizes. In the middle panel, the inflation/exaggeration ratio of statistically significant estimates is plotted against against hypothetical true effect sizes. In the right panel, the probability to get a negative statistically estimate is plotted against against hypothetical true effect sizes. The solid line is the observed estimate of the average treatment effect. The dashed line is half the value of the observed estimate.

Indirect treatment effect of cruise traffic. For the hourly and daily experiments, we observe that road traffic flow and road occupancy rate appear relatively balanced across treated and control units in the matched samples of the two experiments (see hourly balance checks and daily balance checks). It is the case before and after the treatment occurs: when we observe an increase in air pollutant concentrations, this is unlikely to be due to an increase in road traffic.

Strictness of the matching procedure. We relax the strictness of our matching approach by running a more flexible procedure base on one-to-one nearest-neighbor propensity score matching (see hourly propensity score results and daily propensity score results). At the hourly level, 6,710 pairs were matched. The Love plot indicates that covariates balance has increased but the randomization balance check suggests that the treatment is not as-if randomized in the matched data. Estimates found with the propensity score approach are more precise but of smaller magnitudes and often of opposite sites. At the daily level, 1,846 pairs were matched: again, the randomization balance check indicates that the treatment is not as-if randomized in the matched data. Estimates are more precise and relatively consistent with those found with our approach. It is very important to remind that when we compare the results of our constrained pair matching procedure with the propensity score approach, we are comparing two different subsamples of the initial dataset.

Comparison with an outcome regression approach. Finally, we compare our results with those found with a multivariate regression model applied to the initial dataset (see hourly regression results and daily regression results). At the daily level, estimates found with the regression approach are relatively similar but much more precise to those obtained with our matching procedure. However, at the hourly level, as for the comparison with the propensity score approach, regression estimates are of smaller magnitudes and even of opposite signs for some air pollutants (see Figure 15). Again, results are not directly comparable as they are based on different samples.

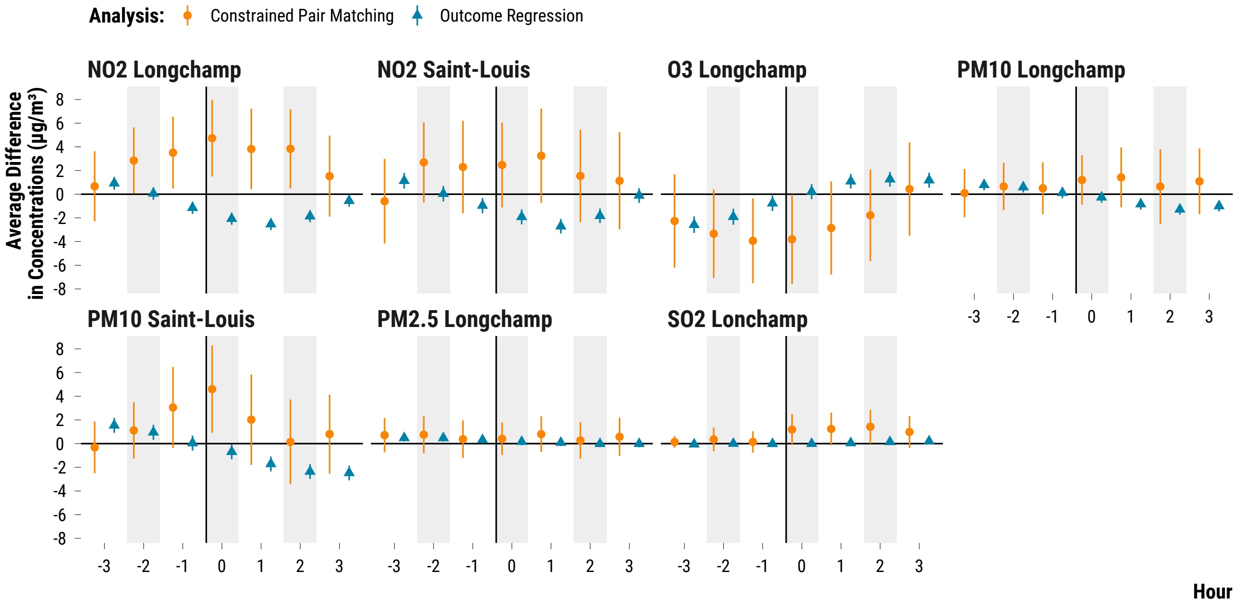


Figure 16: Comparison of Matching and Regression Results for the Hourly Experiment. *Notes:* The treatment occurs at hour 0. The orange color represents the results of our matching procedure while the blue color represents results from the regression approach. Orange dots and blue triangles represent the point estimate of the average treatment effect on a pollutant concentration. Lines are 95% confidence intervals of average treatment effects consistent with the data. The effects are plotted from the third lag to the third lead.

Discussion

In this section, we start by discussing our results in view of the environmental science literature. We then reflect on the new statistical methods used for our analyses. Finally, we suggest paths for future research assessing the causal impact of maritime traffic on air pollution and health.

Putting our Results into Perspective

Our results point to a potential short-term effect of cruise traffic on the concentrations of NO_2, O_3, SO_2, and PM_{10} at the hourly level. At the daily level, we do not observe an impact of cruise vessel on all air pollutant concentrations. However, for both experiments our 95% Fisherian intervals are often wide, and the implied degree of randomization-based uncertainty can be quite large relative to the average concentration of these air pollutants.

Directly comparing our results to those found in the atmospheric science literature is difficult for several reasons. First, they are based on other methods—either source apportionment techniques or dispersion modeling—and usually only report average effects without comparable measures of uncertainty. Second, they often consider the entire traffic of vessels rather than isolating the impact of a predefined treatment, as we do. Third, recent literature reviews have shown that the contribution of vessel emissions to local air pollution depends highly on the port-city considered and the procedure carried out by researchers (Viana et al. 2014, Murena et al. 2018). We can nonetheless assess whether our causal estimates are of the same order of magnitude as estimates from the atmospheric sciences literature.

For gaseous pollutants such as NO_2 and SO_2, the atmospheric science literature has mostly used emissions inventories combined with dispersion modeling (Viana et al. 2014). The few studies on ports from the Mediterranean area find different contributions of maritime traffic to city-level concentrations depending on the size of the city, the location of the monitoring stations, the prevailing wind patterns, the type of boat considered and the assumptions used in the emissions inventory (Murena et al. 2018, Mocerino et al. 2020). These estimates typically take into account all the phases where a vessel may contribute to pollution, in particular the hotelling phase, while we do not have information of the duration of the different phases. For NO_2, estimates range from 1.2-3.5% for the contribution of cruise ships in summer in Naples (Murena et al. 2018), a city three times more populated than Marseille, to 32.5% for the contribution of all types of ships in the Italian city of Brindisi (Merico et al. 2017), much smaller than Marseille. Our estimated contribution of cruise traffic to NO_2 concentrations at the hourly level is equal to an increase of 16%. The estimates for SO_2 range from 1.5% for Naples in winter (Murena et al. 2018) to 46% for Brindisi in summer (Merico et al. 2017). At the hourly level, we observe an increase of 52% in SO_2 concentrations.

For particulate matter, source apportionment methods are commonly used (Sorte et al. 2020, Viana et al. 2008). Estimates for the contribution of vessels to PM_{10} concentrations range from 1.1% for Rijeka in Croatia up to 11% for Genoa in Italy (Merico et al. 2016, Bove et al. 2014), while we do not observe an effect on particulate matter in our daily experiment. This is however consistent with a measurement campaign carried out by Marseille's air quality monitoring agency (Atmosud 2019).

The media and non-governmental organizations have insisted on the high contribution of vessel traffic, and in particular cruise vessel traffic, to city-level emissions as measured by emission inventories. Our hourly experiment confirms that cruise traffic can increase air pollutant concentration on relatively short-time scale. Yet, the results of our daily experiment fail to suggest an impact of cruise vessels to air pollution on a longer time scale. We can contrast the results of our daily experiment on NO_2 concentrations with the contribution of road traffic, which can be inferred from a simple comparison between weekdays and weekends (see our simple road traffic analysis). Because they are balanced in terms of weather covariates, the difference in observed concentrations between weekdays and weekends can be attributed to differences in economic activity only, and in particular to differences in road traffic. Road traffic decreases by 20% on average on Saturdays and Sundays. In parallel, NO_2 concentrations decrease by 20% of their average level at the Saint-Louis station. Although other sources of pollution may be less intense on weekends, the road traffic and NO_2 time series follow an extremely similar pattern, suggesting a strong contribution of road traffic to ambient concentrations compared to maritime traffic. Besides, cruise traffic tends to be higher on week-end and this positive flow of vessels does not offset the likely effect of road traffic on NO_2 concentrations. Beyond emission inventories informing on the relative contribution of different sectors to emissions, more systematic assessments based on observational studies are needed to understand the relative contribution of different sources to ambient concentrations. It would help better evaluate the benefits of abatement in each sector and prioritize policies.

Reflection on the Methods

The causal inference pipeline we follow helps to clearly distinguish the design stage of our study—where we create hypothetical experim ents—from its statistical analysis. Our pair-matching procedure has two notable advantages. First, it prunes treated units for which we cannot find a similar control unit, and thereby avoids extrapolating treatment effects for units without any empirical counterfactuals. In a way, a matching procedure reveals the common support available in the data from which we can draw our statistical inference upon. Second, our approach adjusts for covariates in a nonparametric way and achieves balance between treated and control units on observed covariates. This is another advantage, as it is often hard to guess what functional forms are needed to adjust for confounding factors (Cochran and Rubin 1973, Ho et al. 2007, Imbens 2015).

Yet, matching applied to high-frequency and regular vessel traffic data also poses difficulties. Finding comparable treated and control units is challenging. At the hourly level, it is difficult to match a treated unit with a control unit because vessel traffic is very regular within different periods of the year. For instance, cruise vessels nearly always dock in the port at particular hours and days of the week—leaving few control hours without any cruise traffic. In addition, obtaining days with close weather patterns over several consecutive days is extremely difficult: at the hourly level, it was nearly impossible to find similar pairs over three lags of covariates.

Surprisingly, even if we strive to find similar pairs of treated and control units, we observe a wide heterogeneity in pair differences in pollutant concentrations, which makes it difficult to precisely estimate the potential contribution of vessel emissions. In our study, we are therefore confronted with a trade-off between the comparability of units within pairs and the sample size on which we base our statistical analysis.

Analyzing the full sample using a multivariate regression model delivers more precisely estimated treatment effects. At the daily level, regression results are relatively similar to matching results but are often of the opposite sign at the hourly level. This could be due to the fact that we are comparing two different samples. The alternative reason to explain this discrepancy could be due the multivariate regression model failure to correctly adjust for the functional forms of confounders and to its inherent tendency to extrapolate treatment effects outside the support of the data. Hourly results on the impact of vessel emission on air pollution are also more consistent with what has been observed in previous observational studies on the impact of cruise traffic on air pollutant concentrations (Diesch et al. 2013, Eckhardt et al. 2013, Merico et al. 2016).

Regarding the statistical inference procedure, randomization-based inference allows us to avoid large-sample approximations and makes no assumption on the distribution of our test statistic under the sharp null hypothesis (Rosenbaum 2010, Bind and Rubin 2020). Given that we deal with small sample sizes and provided that our treatment effect assumptions are correct (e.g., constant and additive causal effect, unconfoundedness of the treatment), we believe that our procedure offers a more appropriate quantification of uncertainty in our estimates. However, randomization-based inference, as any inference mode, does not overcome issues implied by having a low statistical power to detect plausible effect sizes of cruise traffic on air pollution. For the hourly hypothetical experiment, we would have a low statistical power if the true effect of cruise vessel traffic on pollutant concentrations was lower than the observed point estimates: estimated effects that are “statistically significant" would overestimate the true effect of vessel traffic on air pollutant concentrations. We therefore recommend to interpret the full width of the uncertainty intervals (Amrhein et al. 2019).

Last, our randomization-based inference procedure relies on the stringent assumption that the treatment is constant. This is arguably an unrealistic assumption. We therefore provide results from a Neymanian inference perspective (Neyman 1923, Imbens and Rubin 2015), which considers average treatment effects rather than unit-level treatment effects. Although based on a different interpretation of the data, results from Fisherian and Neymanian inference are very similar. The recent approach proposed by Wu and Ding (2021) to make a randomization inference procedure conservative for average treatment effects also give the same results. As an alternative to Fisherian and Neyman modes of inference, we could also have implemented a Bayesian model-based approach, which explicitly imputes the missing potential outcomes given the observed data and can target a larger variety of causal estimands (Rubin 1978, Imbens and Rubin

2015, Bind and Rubin 2019).

Potential Paths for Future Research

We see at least three main improvements for future research based on observational data on the effects of maritime traffic on air pollution. First, it would be useful to exploit data on the duration vessels keep their engines running while docked at the port. Several studies indicate that a large share of air pollutant emissions occur during this phase (CAIMAN 2015, Murena et al. 2018). Second, monitoring stations in Marseille only measure some air pollutants and are located relatively far away from the port. It would be useful to carry out similar analyses as ours in a port city where pollutants such as ultrafine particles are monitored and with receptors located in the port at different heights (Viana et al. 2014, Mocerino et al. 2020). Besides, the weather data we exploit are located 25km away from the city, which adds noise. It would also be useful to have a monitoring station located within the city. Third, several areas have implemented regulations to decrease the sulfur content of vessel fuel. This type of policy is particularly well-suited for causal inference methods such as interrupted-time series, difference-in-differences, and synthetic control (Kotchenruther 2017, Grange and Carslaw 2019, Zhu and Wang 2021). They are arguably easier to implement than finding hypothetical experiments within very regular time series data on vessel traffic.

Concluding Remarks

Our study proposes a complementary approach to current source-apportionment and dispersion modeling methods. It should be more familiar to researchers willing to estimate the health effects of the air pollution induced by maritime traffic. We provide detailed replication materials in the hope that researchers could implement and improve our method for other ports which are not part of emission control areas. Even if there remains challenges with regards to potential spillover effects and the imprecision of estimates, we believe that well-designed observational studies relying on the proposed causal inference pipeline could bring new insights on the environmental and health consequences of maritime traffic.

Acknowledgments

We thank the Marseille Port officials for sharing their port call data, Méteo-France for providing us their weather data, AtmoSud for openly sharing their air pollution data on their website, and Milena Suarez-Castillo for providing road traffic data from the DIR Méditerranée. We are grateful to Tirthankar Dasgupta for his guidance on the computation of Fisherian intervals and to Stéphane Shao for developing the matching algorithm. We thank Michela Baccini, Geoffrey Barrows, Olivier Chanel, Augustin Colette, Tatyana Deryugina, Francois Libois, Mirabelle Muûls, Helene Ollivier, Thomas Piketty, Laure de Preux, Quentin Lippmann, Philippe Quirion, Adam Rosenberg, Thiago Scarelli, Katheline Schubert, Georgia Thebault, Ulrich Wagner, as well as participants to the PSE Applied Economics, PSIPSE and REM seminars and conference participants from EAERE and FAERE for their feedback. Leo Zabrocki and Marion Leroutier acknowledge the support of the EUR grant ANR-17-EURE-0001.

Bibliography

Amrhein, Valentin, David Trafimow, and Sander Greenland (2019) “Inferential statistics as descriptive statistics: There is no replication crisis if we don’t expect replication,” *The American Statistician*, 73 (sup1), 262–270.

Athey, Susan and Guido W Imbens (2017) “The econometrics of randomized experiments,” in *Handbook of economic field experiments*, 1, 73–140: Elsevier.

Atmosud (2019) “Quelle qualité de l’air pour les riverains des ports de Nice et Marseille? Campagnes de mesure 2018,”Technical report, `https://www.atmosud.org/sites/paca/files/atoms/files/200511_synthese_travaux_ports_2018_0.pdf`.

AtmoSud (2020) “CIGALE: Consultation d’Inventaires Géolocalisés Air CLimat Energie,” `https://cigale.atmosud.org/`.

Baccini, Michela, Alessandra Mattei, Fabrizia Mealli, Pier Alberto Bertazzi, and Michele Carugno (2017) “Assessing the short term impact of air pollution on mortality: a matching approach,” *Environmental Health*, 16 (1), 1–12.

Bauernschuster, Stefan, Timo Hener, and Helmut Rainer (2017) “When labor disputes bring cities to a standstill: The impact of public transit strikes on traffic, accidents, air pollution, and health,” *American Economic Journal: Economic Policy*, 9 (1), 1–37.

Bind, Marie-Abèle (2019) “Causal modeling in environmental health,” *Annual review of public health*, 40, 23–43.

Bind, Marie-Abèle C. and DB Rubin (2021) “The importance of having a conceptual stage when reporting non-randomized studies,” *Biostatistics & Epidemiology*, 5 (1), 9–18.

Bind, Marie-Abèle C. and Donald B. Rubin (2019) “Bridging observational studies and randomized experiments by embedding the former in the latter,” *Statistical Methods in Medical Research*, 28 (7), 1958–1978.

——— (2020) "When possible, report a Fisher-exact P value and display its underlying null randomization distribution," *Proceedings of the National Academy of Sciences*, 117 (32), 19151–19158.

Bojinov, Iavor and Neil Shephard (2019) "Time series experiments and causal estimands: exact randomization tests and trading," *Journal of the American Statistical Association*, 114 (528), 1665–1682.

Bove, MC, P Brotto, F Cassola, E Cuccia, D Massabò, A Mazzino, A Piazzalunga, and P Prati (2014) "An integrated PM2.5 source apportionment study: Positive Matrix Factorisation vs. the chemical transport model CAMx," *Atmospheric Environment*, 94, 274–286.

Bowers, Jake and Thomas Leavitt (2020) "Causality and Design-Based Inference," in *The SAGE Handbook of Research Methods in Political Science and International Relations*, 769–804: SAGE Publications Ltd.

Bowers, Jake and Costas Panagopoulos (2011) "Fisher's randomizationmode of statistical inference, then and now.."

Branson, Zach (2021) "Randomization Tests to Assess Covariate Balance When Designing and Analyzing Matched Datasets," *Observational Studies*, 7 (2), 1–36.

CAIMAN (2015) "Air quality impact and greenhouse gases assessment for cruise and passenger ships," , Technical Report.

Cattaneo, Matias D, Brigham R Frandsen, and Rocio Titiunik (2015) "Randomization inference in the regression discontinuity design: An application to party advantages in the US Senate," *Journal of Causal Inference*, 3 (1), 1–24.

Caughey, Devin, Allan Dafoe, Xinran Li, and Luke Miratrix (2021) "Randomization Inference beyond the Sharp Null: Bounded Null Hypotheses and Quantiles of Individual Treatment Effects," *arXiv preprint arXiv:2101.09195*.

Chrisafis, Angelique (2018-07-6) "'I don't want ships to kill me': Marseille fights cruise liner pollution," *The Guardian*, https://www.theguardian.com/environment/2017/jul/31/heading-to-venice-dont-forget-your-pollution-mask.

Cleveland, William S (1993) *Visualizing data*: Hobart press.

Cochran, William G and Donald B Rubin (1973) "Controlling bias in observational studies: A review," *Sankhyā: The Indian Journal of Statistics, Series A*, 417–446.

Cohen, Jessica and Pascaline Dupas (2010) "Free distribution or cost-sharing? Evidence from a randomized malaria prevention experiment," *The Quarterly Journal of Economics*, 1–45.

Contini, D, A Gambaro, F Belosi, S De Pieri, WRL Cairns, A Donateo, E Zanotto, and M Citron (2011) "The direct influence of ship traffic on atmospheric PM2. 5, PM10 and PAH in Venice," *Journal of Environmental Management*, 92 (9), 2119–2129.

Cruise Lines International Association (2019) "CLIA Reveals Growth in Global and North American Passenger Numbers and Insights," `https://cruising.org/news-and-research/press-room/2019/april/clia-reveals-growth`.

Damien Piga, Magali Devèze Michael Parra Nicolas Marchand Anaïs Detournay, Alexandre Armengaud and Dalia Salameh (2013) "Synthèse du projet APICE - Marseille," , Technical Report.

Dasgupta, Tirthankar and Donald B. Rubin (2021) *Experimental Design: A Randomization-Based Perspective*: Unpublished Textbook.

Dasgupta, Tirthankar and Donald B Rubin (Fall 2015) *STAT 240: Matched Sampling and Study Design*: Harvard university.

Diesch, J-M, F Drewnick, T Klimach, and S Borrmann (2013) "Investigation of gaseous and particulate emissions from various marine vessel types measured on the banks of the Elbe in Northern Germany," *Atmospheric Chemistry and Physics*, 13 (7), 3603–3618.

Ding, Peng, Avi Feller, and Luke Miratrix (2016) "Randomization inference for treatment effect variation," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78 (3), 655–671.

Eckhardt, Sabine, Ove Hermansen, Henrik Grythe, Markus Fiebig, Kerstin Stebel, Massimo Cassiani, Are Bäcklund, and Andreas Stohl (2013) "The influence of cruise ship emissions on air pollution in Svalbard–a harbinger of a more polluted Arctic?" *Atmospheric Chemistry and Physics*, 13 (16), 8401–8409.

Fisher, Ronald Aylmer et al. (1937) "The design of experiments.," *The design of experiments.* (2nd Ed).

Fogarty, Colin B (2020) "Studentized sensitivity analysis for the sample average treatment effect in paired observational studies," *Journal of the American Statistical Association*, 115 (531), 1518–1530.

Forastiere, Laura, Michele Carugno, and Michela Baccini (2020) "Assessing short-term impact of PM 10 on mortality using a semiparametric generalized propensity score approach," *Environmental Health*, 19 (1), 1–13.

Friedrich, Axel (2017-07-31) "Heading to Venice? Don't forget your pollution mask," *The Guardian*, `https://www.theguardian.com/environment/2017/jul/31/heading-to-venice-dont-forget-your-pollution-mask`.

Gelman, Andrew and John Carlin (2014) "Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors," *Perspectives on Psychological Science*, 9 (6), 641–651.

Gelman, Andrew, Jennifer Hill, and Aki Vehtari (2020) *Regression and other stories*: Cambridge University Press.

Gerber, Alan S and Donald P Green (2012) *Field experiments: Design, analysis, and interpretation*: WW Norton.

Giaccherini, Matilde, Joanna Kopinska, and Alessandro Palma (2021) "When particulate matter strikes cities: Social disparities and health costs of air pollution," *Journal of Health Economics*, 78, 102478.

Godzinski, Alexandre, M Suarez Castillo et al. (2019) "Short-term health effects of public transport disruptions: air pollution and viral spread channels,"Technical report, Institut National de la Statistique et des Etudes Economiques.

GPMM (2020) "Port de Marseille Fos - Grand Port Maritime de Marseille: Yearly Figures," , Technical Report.

Grange, Stuart K and David C Carslaw (2019) "Using meteorological normalisation to detect interventions in air quality time series," *Science of The Total Environment*, 653, 578–588.

Greifer, Noah and Elizabeth A Stuart (2021) "Matching methods for confounder adjustment: an addition to the epidemiologist's toolbox," *Epidemiologic reviews*, 43 (1), 118–129.

Gutman, R, DB Rubin, and Stijn Vansteelandt (2012) "Analyses that Inform Policy Decisions [with Discussions]," *Biometrics*, 68 (3), 671–678.

Heß, Simon (2017) "Randomization inference with Stata: A guide and software," *The Stata Journal*, 17 (3), 630–651.

Ho, Daniel E and Kosuke Imai (2006) "Randomization inference with natural experiments: An analysis of ballot effects in the 2003 California recall election," *Journal of the American statistical association*, 101 (475), 888–900.

Ho, Daniel E, Kosuke Imai, Gary King, and Elizabeth A Stuart (2007) "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference," *Political analysis*, 15 (3), 199–236.

Holland, Paul W (1986) "Statistics and causal inference," *Journal of the American statistical Association*, 81 (396), 945–960.

Imbens, Guido W (2015) "Matching methods in practice: Three examples," *Journal of Human Resources*, 50 (2), 373–419.

Imbens, Guido W and Donald B Rubin (2015) *Causal inference in statistics, social, and biomedical sciences*: Cambridge University Press.

INSEE (2020) "Commune de Marseille (13055)," , Populations légales - 2017.

Keele, Luke, Corrine McConnaughy, and Ismail White (2012) "Strengthening the experimenter's toolbox: Statistical estimation of internal validity," *American Journal of Political Science*, 56 (2), 484–499.

Keele, Luke and Luke Miratrix (2019) "Randomization inference for outcomes with clumping at zero," *The American Statistician*, 73 (2), 141–150.

Khomenko, Sasha, Marta Cirach, Evelise Pereira-Barboza et al. (2021) "Premature mortality due to air pollution in European cities: a health impact assessment," *The Lancet Planetary Health*, 5 (3), e121–e134, 10.1016/S2542-5196(20)30272-2, Publisher: Elsevier.

King, Gary and Langche Zeng (2006) "The dangers of extreme counterfactuals," *Political analysis*, 14 (2), 131–159.

Knittel, Christopher R, Douglas L Miller, and Nicholas J Sanders (2016) "Caution, drivers! Children present: Traffic, pollution, and infant health," *Review of Economics and Statistics*, 98 (2), 350–366.

Kotchenruther, Robert A (2017) "The effects of marine vessel fuel sulfur regulations on ambient PM2.5 at coastal and near coastal monitoring sites in the US," *Atmospheric Environment*, 151, 52–61.

Lee, Kwonsang, Dylan S Small, and Francesca Dominici (2021) "Discovering heterogeneous exposure effects using randomization inference in air pollution studies," *Journal of the American Statistical Association*, 116 (534), 569–580.

Liu, Huan, Mingliang Fu, Xinxin Jin, Yi Shang, Drew Shindell, Greg Faluvegi, Cary Shindell, and Kebin He (2016) "Health and climate impacts of ocean-going vessels in East Asia," *Nature Climate Change*, 6 (11), 10.1038/nclimate3083.

Liu, Huan, Zhi-Hang Meng, Zhao-Feng Lv et al. (2019) "Emissions and health impacts from global shipping embodied in US–China bilateral trade," *Nature Sustainability*, 2 (11), 10.1038/s41893-019-0414-z.

Lu, Jiannan, Yixuan Qiu, and Alex Deng (2019) "A note on Type S/M errors in hypothesis testing," *British Journal of Mathematical and Statistical Psychology*, 72 (1), 1–17.

MacKinnon, James G and Matthew D Webb (2020) "Randomization inference for difference-in-differences with few treated clusters," *Journal of Econometrics*, 218 (2), 435–450.

Mayer, Michael (2019) *missRanger: Fast Imputation of Missing Values*, `https://cran.r-project.org/package=missRanger`, R package version 2.1.0.

Menchetti, Fiammetta, Fabrizio Cipollini, and Fabrizia Mealli (2021) "Estimating the causal effect of an intervention in a time series setting: the C-ARIMA approach," *arXiv preprint arXiv:2103.06740*.

Merico, E, A Donateo, A Gambaro et al. (2016) "Influence of in-port ships emissions to gaseous atmospheric pollutants and to particulate matter of different sizes in a Mediterranean harbour in Italy," *Atmospheric Environment*, 139, 1–10.

Merico, Eva, Andrea Gambaro, A Argiriou et al. (2017) "Atmospheric impact of ship traffic in four Adriatic-Ionian port-cities: Comparison and harmonization of different approaches," *Transportation Research Part D: Transport and Environment*, 50, 431–445.

Micali, Silvio and Vijay V Vazirani (1980) "An O (v| v| c| E|) algorithm for finding maximum matching in general graphs," in *21st Annual Symposium on Foundations of Computer Science (sfcs 1980)*, 17–27, IEEE.

Mocerino, Luigia, Fabio Murena, Franco Quaranta, and Domenico Toscano (2020) "A methodology for the design of an effective air quality monitoring network in port areas," *Scientific Reports*, 10 (1), 1–10.

Moretti, Enrico and Matthew Neidell (2011) "Pollution, health, and avoidance behavior evidence from the ports of Los Angeles," *Journal of Human Resources*, 46 (1), 154–175.

Mueller, Daniel, Stefanie Uibel, Masaya Takemura, Doris Klingelhoefer, and David A Groneberg (2011) "Ships, ports and particulate air pollution-an analysis of recent studies," *Journal of Occupational Medicine and Toxicology*, 6 (1), 1–6.

Murena, F, L Mocerino, F Quaranta, and D Toscano (2018) "Impact on air quality of cruise ship emissions in Naples, Italy," *Atmospheric Environment*, 187, 70–83.

Neyman, Jersey (1923) "Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes," *Roczniki Nauk Rolniczych*, 10, 1–51.

Rosenbaum, Paul (2018) *Observation and experiment*: Harvard University Press.

Rosenbaum, Paul R (1987) "Sensitivity analysis for certain permutation inferences in matched observational studies," *Biometrika*, 74 (1), 13–26.

——— (2010) *Design of observational studies*: Springer.

Rubin, Donald B (1974) "Estimating causal effects of treatments in randomized and nonrandomized studies.," *Journal of Educational Psychology*, 66 (5), 688.

——— (1978) "Bayesian inference for causal effects: The role of randomization," *The Annals of Statistics*, 34–58.

——— (1991) "Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism," *Biometrics*, 1213–1234.

——— (2005) "Causal inference using potential outcomes: Design, modeling, decisions," *Journal of the American Statistical Association*, 100 (469), 322–331.

——— (2006) *Matched sampling for causal effects*: Cambridge University Press.

Schlenker, Wolfram and W Reed Walker (2016) "Airports, air pollution, and contemporaneous health," *The Review of Economic Studies*, 83 (2), 768–809.

Sekhon, Jasjeet S (2009) "Opiates for the matches: Matching methods for causal inference," *Annual Review of Political Science*, 12, 487–508.

Simeonova, Emilia, Janet Currie, Peter Nilsson, and Reed Walker (2021) "Congestion pricing, air pollution, and children's health," *Journal of Human Resources*, 56 (4), 971–996.

Sommer, Alice J, Mihye Lee, and Marie-Abèle C Bind (2018) "Comparing apples to apples: an environmental criminology analysis of the effects of heat and rain on violent crimes in Boston," *Palgrave communications*, 4 (1), 1–10.

Sommer, Alice J, Emmanuelle Leray, Young Lee, and Marie-Abèle C Bind (2021) "Assessing environmental epidemiology questions in practice with a causal inference pipeline: An investigation of the air pollution-multiple sclerosis relapses relationship," *Statistics in Medicine*, 40 (6), 1321–1335.

Sorte, Sandra, Vera Rodrigues, Carlos Borrego, and Alexandra Monteiro (2020) "Impact of harbour activities on local air quality: A review," *Environmental Pollution*, 257, 113542.

Stuart, Elizabeth A (2010) "Matching methods for causal inference: A review and a look forward," *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25 (1), 1.

Timm, Andrew (2019) *retrodesign: Tools for Type S (Sign) and Type M (Magnitude) Errors*, https://CRAN.R-project.org/package=retrodesign, R package version 0.1.0.

Tufte, Edward R (1985) "The visual display of quantitative information," *The Journal for Healthcare Quality (JHQ)*, 7 (3), 15.

Viana, Mar, Pieter Hammingh, Augustin Colette, Xavier Querol, Bart Degraeuwe, Ina de Vlieger, and John van Aardenne (2014) "Impact of maritime transport emissions on coastal air quality in Europe," *Atmospheric Environment*, 90, 96–105.

Viana, Mar, Thomas AJ Kuhlbusch, Xavier Querol et al. (2008) "Source apportionment of particulate matter in Europe: a review of methods and results," *Journal of Aerosol Science*, 39 (10), 827–849.

Wu, Jason and Peng Ding (2021) "Randomization tests for weak null hypotheses in randomized experiments," *Journal of the American Statistical Association*, 116 (536), 1898–1913.

Zhao, Anqi and Peng Ding (2021) "Covariate-adjusted Fisher randomization tests for the average treatment effect," *Journal of Econometrics*, 225 (2), 278–294.

Zhong, Nan, Jing Cao, and Yuzhu Wang (2017) "Traffic congestion, ambient air pollution, and health: Evidence from driving restrictions in Beijing," *Journal of the Association of Environmental and Resource Economists*, 4 (3), 821–856.

Zhu, Junming and Jiali Wang (2021) "The effects of fuel content regulation at ports on regional pollution and shipping industry," *Journal of Environmental Economics and Management*, 106, 102424.

Zigler, Corwin Matthew and Francesca Dominici (2014) "Point: clarifying policy evidence with potential-outcomes thinking—beyond exposure-response estimation in air pollution epidemiology," *American journal of epidemiology*, 180 (12), 1133–1140.

Why Acute Health Effects of Air Pollution Could Be Inflated

Accurate and precise measurements of the short-term effects of air pollution on health play a key role in setting air quality standards. Yet, statistical power calculations are rarely—if ever—carried out. We first collect estimates and standard errors of all available articles found in the epidemiology and economics literatures. We find that nearly half of them may suffer from a low statistical power and could thereby produce statistically significant estimates that are actually inflated. We then run simulations based on real data to identify which parameters of research designs affect statistical power. Despite their large sample sizes, we show that studies exploiting rare exogenous shocks such as transport strikes or thermal inversions could have a very low statistical power, even if effect sizes are large. Our simulation results indicate that the observed discrepancy in the literature between instrumental variable estimates and non-causal ones could be partly explained by the inherent imprecision of the two-stage least-squares estimator. We also provide evidence that subgroup analysis on the elderly or children should be implemented with caution since the average number of events for an health outcome is a major driver of power. Based on these findings, we build a series of recommendations for researchers to evaluate the design of their study with respect to statistical power issues.

AUTHORS:
Léo Zabrocki
PSE - EHESS
`leo.zabrocki@psemail.eu`

Vincent Bagilet
Columbia University - SIPA
`vincent.bagilet@columbia.edu`

Introduction

From extreme events such as the London Fog of 1952 to the development of sophisticated time-series analyses, a vast scientific literature in epidemiology has established that air pollution induces adverse health effects on the very short-term (Schwartz 1994, Le Tertre et al. 2002, Bell et al. 2004, Di et al. 2017, Liu et al. 2019). Increases in the concentration of several ambient air pollutants have been found to be associated with small relative increases in the daily mortality and emergency admissions for respiratory and cardiovascular causes (Samet et al. 2000, Shah et al. 2015, Orellano et al. 2020). All this evidence led to the implementation of public policies such as air quality alerts to mitigate the acute effects of air pollutants. Accurate estimates of the short-term health effects of air pollution are therefore crucial as they directly inform public health policies.

WEBSITE:
`https://vincentbagilet.github.io/inference_pollution/`

With this objective in mind, researchers in economics and epidemiology have addressed the issue of unmeasured confounding variables with causal inference methods in the last decade (Dominici and Zigler 2017, Bind 2019). Newly obtained results confirm the acute health effects of air pollution (Schwartz et al. 2015; 2018, Deryugina et al. 2019). Yet, causal estimates are often larger than what would have been predicted by the standard epidemiology literature. For instance, in Panel A of Figure 17, we see that instrumental variable estimates of 9 studies in the causal inference literature based on this method are always larger than naive ordinary least squares estimates. This could arguably be explained by the fact that instrumental variable strategies remove omitted variable bias and reduce attenuation bias coming from classical measurement error in air pollution exposure. Panel B of Figure 17 however suggests an alternative and complementary explanation. For the 29 papers using causal inference methods found in this literature, we plot the standardized estimates against the inverse of their standard errors, which is a proxy for a study's precision. Large effect sizes are only found in imprecise studies and the more precise the study, the smaller the effect size. The negative relationship between effect sizes and studies' precision has also been observed in fields such as medicine, psychology and economics (Button et al. 2013, Camerer et al. 2016, Schäfer and Schwarz 2019).

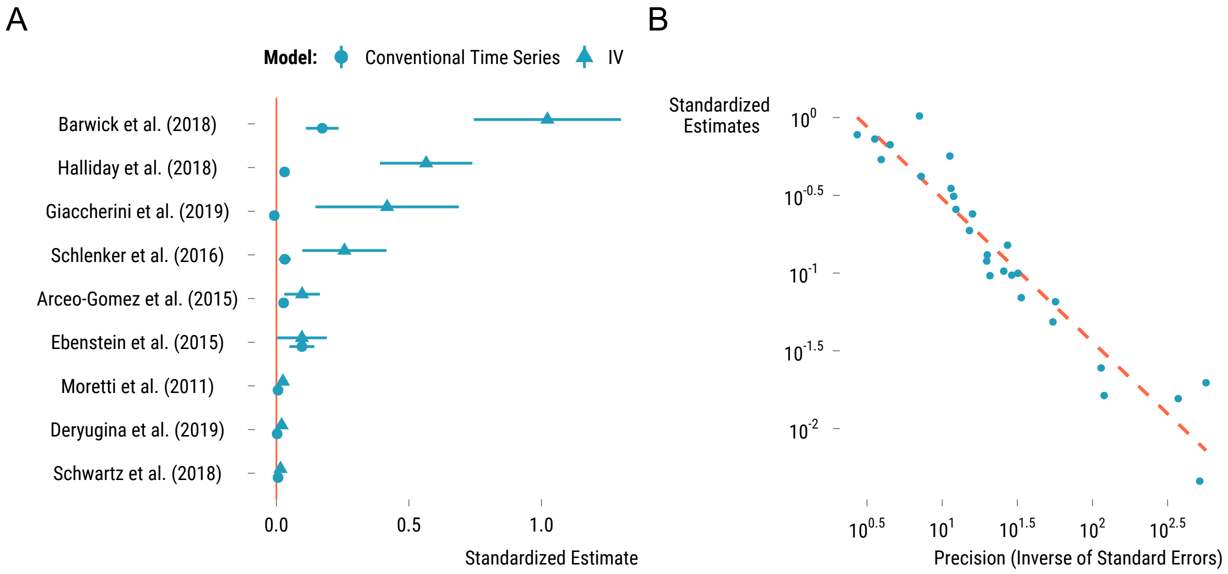


Figure 17: Naive versus Causal Estimates and the Deflation of Effect Sizes as Precision Increases. *Notes*: In Panel A, standardized estimates and their associated 95% confidence intervals are displayed for the 9 articles of the causal inference literature based on instrumental variable strategies and for which estimates from naive regressions are available. Triangles represent instrumental variable estimates with dots are naive regression estimates. In panel B, standardized estimates of the 29 articles of the causal inference literature are plotted against the inverse of the standard errors, which can been considered as a measure of precision. Both axes are on a log10 scale.

The variation in studies' statistical power could explain the origin of this negative relationship but also help understand why causal estimates are larger than those found in the epidemiology literature. Simply put, studies with low precision result in larger effect sizes (Ioannidis 2008, Gelman and Carlin 2014). Their statistical power is low and, to be statistically significant, their estimates need to be large enough, at least 2 standard errors away from 0 at the 5% significance level. Since statistically significant results are more likely to be published, some estimates found in the literature may be inflated as they would come from a non-representative sample of the estimates, those large enough to be statistically significant (Brodeur et al. 2016; 2020). The consequences of low statistical power are not specific to studies on short-term health effects of air pollution but may be particularly salient in this literature where the signal-to-noise ratio is often low (Peng et al. 2006, Peng and Dominici 2008).

In this paper, we undertake the first empirical investigation to determine if studies on the short-term health effects of air pollution could be under-powered and thereby produce inflated estimates. We start tackling this question by gathering a unique corpus of about 600 studies based on associations and 29 articles that rely on causal inference. For each of these papers, we run statistical power calculations to assess whether the design of the study would be robust enough to confidently detect an effect size smaller than the observed estimate (Gelman and Carlin 2014, Ioannidis et al. 2017, Lu et al. 2019, Timm 2019). Using real data from the US National Morbidity, Mortality, and Air Pollution Study (Samet et al. 2000), we then implement simulations to identify the characteristics of research designs that drive their statistical power and the inflation of statistically significant estimates (Black et al. 2021, Gelman et al. 2020, Altoè et al. 2020).

The results of our statistical power calculations show that research designs based on associations and causal inference methods are similarly prone to statistical power issues. Half of the studies in the two strands of the literature have a statistical power below 80% to detect effect sizes that are only 25% smaller than their observed estimates. In the standard epidemiology literature, under-powered studies could produce statistically significant estimates 2 times larger than true effect sizes. Our retrospective power calculations also highlight a wide heterogeneity in the robustness of articles with respect to statistical power issues. For example, if the true effect sizes are equal to the ones predicted by the standard epidemiology literature, the statistical power of studies using instrumental variable designs would range from 5% to 64%. In some studies, statistically significant estimates would be just 1.3 times larger than the true effect sizes, while in others, the inflation factor could be as high as 41.

Our simulation results help understand why some research designs face statistical power issues. We first show that a very large number of observations is needed for all causal inference methods to reach a sufficient statistical power. Regression discontinuity designs based on air quality alerts rely on sample sizes that may be too small for statistically significant estimates not to be inflated. Second, we show that the use of public transport strikes or thermal inversions as exogenous shocks on air pollution could be problematic. These studies are based on rare events, which in some cases represent less than 1% of the observations. The resulting statistical power is very low, around 15%, and statistically significant estimates can exaggerate even large true effect sizes by a factor of 2.7. Third, we find that the average daily count of cases of an health outcome is a key driver of statistical power for all empirical strategies. Statistically significant estimates of the effects of air pollution on the elderly or children can be very inflated since health outcomes for these groups often have few daily cases.

Our article makes two contributions to the literature on the acute health effects of air pollution. First, as highlighted by the replication crises in medicine, psychology and experimental economics (Button et al. 2013, Open Science Collaboration 2015, Camerer et al. 2016), there is a crucial need to evaluate the deficiencies of current statistical practices grounded in the null hypothesis significance testing framework (Ziliak and McCloskey 2008, Simonsohn et al. 2014, Smaldino and McElreath 2016, Greenland 2017, Christensen et al. 2019, Amrhein et al. 2019). Our paper participates in the growing literature that uses retrospective power calculations to evaluate the plausibility of published findings (Ioannidis 2008, Gelman and Carlin 2014, Smaldino and McElreath 2016, Ioannidis et al. 2017, Ferraro and Shukla 2020, Stommes et al. 2021). To the best of our knowledge, this paper is the first to show how to carry out retrospective statistical power calculations for studies on air pollution and human health. We also provide the first evidence that under-powered studies are a real issue in this field.

Second, except for standard models used in the epidemiology literature (Winquist et al. 2012), few statistical power analyses exist to help researchers improve their study designs (Bhaskaran et al. 2013). This paper is the first to give concrete recommendations to avoid statistical power issues for several research designs estimating the acute health effects of air pollution. Statisticians have long advocated the use of fake-data simulations to flexibly evaluate the inference properties of statistical models (Gelman and Carlin 2014, Vasishth and Gelman 2019, Altoè et al. 2020, Gelman et al. 2020). In our paper, we follow this advice but rely instead on real data since it is very complex to correctly simulate the relationships between ambient air pollution, weather parameters, calendar indicators and health outcomes. Our article is more closely connected to three recent articles evaluating the type I error rate and the lack of statistical power of several panel data models used to estimate the impacts of public policies on mortality outcomes (Schell et al. 2018, Black et al. 2021, Griffin et al. 2021). These simulations focus on event study designs and treatment effects happening on a medium to long time scale. On the contrary, our simulations gauge the capacity of reduced-form, instrumental variable and regression discontinuity designs to estimate very short-run effects in the context of high-frequency data.

Finally, we strive to make our analyses fully and easily reproducible to help researchers implement retrospective power calculations and power simulations in their own studies. We use state-of-the-art literate programming to explain and render all coding procedures in nicely formatted HTML documents (Allaire et al. 2018). All replication and supplementary materials are available on this website.

In the following section, we implement a simple simulation exercise to show why statistically significant estimates exaggerate true effect sizes when studies have a low statistical power. In the second section, we present our retrospective analysis of the literature. In the third section, we detail our simulation procedure to replicate empirical strategies. We display the simulation results in the fourth section and provide specific guidance on study design in the fifth section.

Background on Statistical Power, Type M and S errors

In a seminal paper, (Gelman and Carlin 2014) point out that researchers working in the null hypothesis significance testing framework are often unaware that "statistically significant" estimates suffer from a winner's curse in under-powered studies: these estimates can largely overestimate true effect sizes and can even be of the opposite sign. In this section, we implement a simple simulation exercise to illustrate these two counter-intuitive issues and explain why they could matter in studies on the acute health effects of ambient air pollutants.

Illustrative Example

Imagine that a mad scientist is able to implement a randomized experiment to measure the short-term effects of air pollution on daily non-accidental mortality. The experiment takes place in a major city over the 366 days of a leap year. The scientist is able to increase concentration of particulate matter with a diameter below 2.5 µm ($PM_{2.5}$) by 10 µg/m^3—a large shock equivalent to one standard deviation increase in the concentration of $PM_{2.5}$. Concretely, the scientist implements a complete experiment where they randomly allocate half of the days to the treatment group and the other half to the control group. They then measure the treatment effect of the intervention by computing the average difference in means between treated and control outcomes: the estimate for the treatment effect is equal to 4 additional deaths and is "statistically significant" at the 5% level, with a p-value of 0.04. The statistical significance of the estimate fulfills the scientist expectations, who immediately starts writing their paper. Had they not obtained a statistically significant estimate, they might not have submitted their result.

Unfortunately for the scientist, we know what the true effect of the experiment is since we created the data. In Table 3, we display the Science table where we observe the pair of potential outcomes of each day, $Y_i(W_i = 0)$ and $Y_i(W_i = 1)$ (Rubin 1974). Y_i represents a daily count of non-accidental deaths and W_i the treatment assignment, which is equal to 1 for treated units and 0 otherwise. We first simulated the daily non-accidental mortality counts in the absence of treatment (i.e., the $Y(0)$ column of Table 3), by drawing 366 observations from a negative binomial distribution with a mean of 106 and a variance of 402. We chose the parameters to approximate the distribution of non-accidental mortality counts in a large European city. We then defined the counterfactual distribution of mortality by adding, on average, 1 extra death (i.e., the $Y(1)$ column of Table 3).

Table 3: Science Table of the Experiment.

Day Index	$Y_i(0)$	$Y_i(1)$	τ_i	W_i	Y_i^{obs}
1	122	124	+2	1	124
2	94	96	+2	1	96
3	96	98	+2	0	96
⋮	⋮	⋮	⋮	⋮	
364	96	97	+1	0	96
365	98	98	+0	0	98
366	143	144	+1	1	144

Notes: This table displays the potential outcomes, the unit-level treatment effect, the treatment status and the observed outcomes for 6 of the 366 daily units in the scientist's experiment.

This treatment effect size represents approximately a 1% increase in the mean of the outcome[1]. Following the fundamental problem of causal inference, the daily count of deaths the scientist observes is given by the equation: $Y_i^{\text{obs}} = W_i \times Y_i(1) + (1 - W_i) \times Y_i(0)$. Treated units express their $Y_i(1)$ values and control units their $Y_i(0)$ values.

With a random assignment of the treatment, how is it possible that the statistically significant estimate found by the scientist can be 4 times larger than the true treatment effect size? Replicating many times the experiment can help understand why.

Defining Statistical Power, Type M and S errors

In Figure 18, we plot the estimates of 10,000 iterations of the experiment. If there is a large variation in the effect size of estimates, the average is reassuringly equal to the true treatment effect of 1 additional death. We can however see that estimates close to the true effect size would not be statistically significant at the 5% level. In a world without publication bias, we could be confident that several replications of this experiment would recover the true treatment effect. Unfortunately, researchers are—despite recent changes in scientific practices and editorial policies—not incited enough to publish replication exercises and not statistically significant estimates. In a world with publication bias, only statistically significant estimates would be made public. Out of the 10,000 estimates, about 800 are statistically significant at the 5% level. The *statistical power* of the experiment, which can be defined as the probability to reject the null hypothesis when there is actually an effect, is therefore equal to 8%. The scientist was therefore very lucky to get a statistically significant estimate.

But with such a low statistical power, statistically significant es-

---

[1] Note that the magnitude of this hypothetical effect is higher than what has been found in a recent and large-scale study based on 625 cities. (Liu et al. 2019) found that a 10 μg/m^3 increase in PM$_{2.5}$ concentration was associated with a 0.68% (95% CI, 0.59 to 0.77) relative increase in daily all-causes mortality.

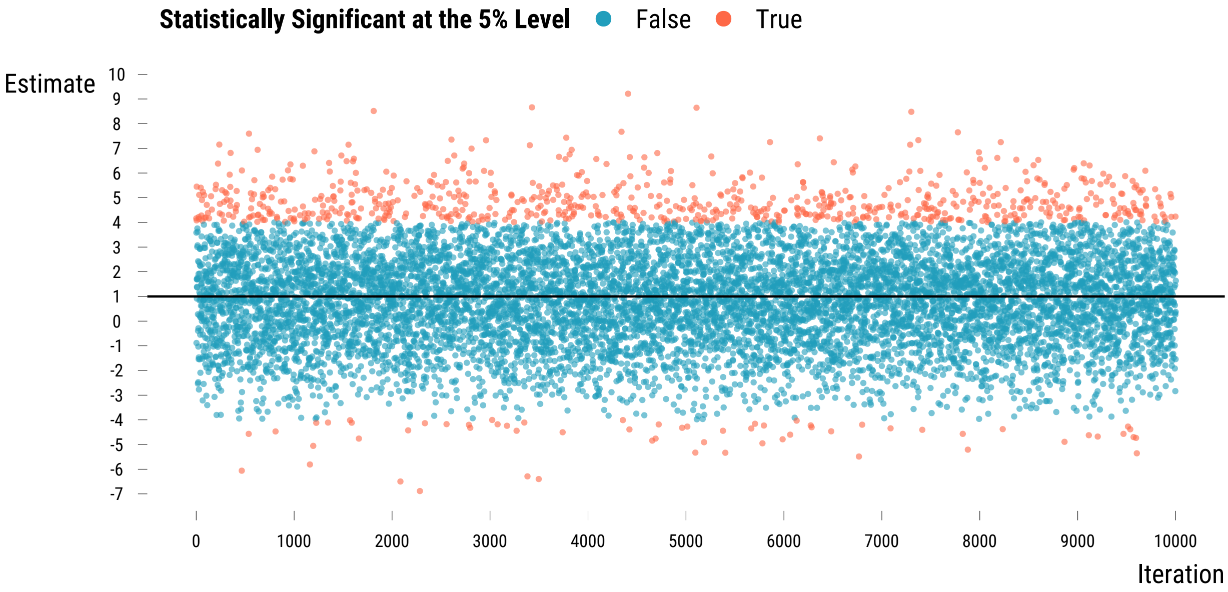


Figure 18: Estimates of the 10,000 Simulations. *Notes*: In Panel A, blue and red dots represent the point estimates of the 10,000 iterations of the randomized experiment ran by the mad scientist. Red dots are statistically significant at the 5% level while blue dots are not. The black solid line represents the true average effect of 1 additional death.

timates cannot be trusted anymore. Two metrics, the average type M (for magnitude) error and the probability to make a type S (for sign) error are useful to assess the negative consequences of lacking statistical power. First, we can evaluate by how much statistically significant estimates are inflated compared to the true treatment effect size by computing the average ratio of the absolute values of the statistically significant estimates over the true effect size (Gelman and Carlin 2014). With a statistical power of 8%, the scientist would on average make a type M error equal to 5! Second, we can notice that a non-negligible fraction of statistically significant estimates are of the wrong sign in Figure 18: this proportion is the probability of making a type S error (Gelman and Tuerlinckx 2000). For this experiment, a statistically significant estimate has a 8% probability of being of the wrong sign!

Thus, if the scientist would like to estimate the effect of the experiment through the prism of the statistical significance, they would need a larger number of observations: statistical power would then rise and conversely type M and S error would shrink.

Relevance for Studies on Acute Health Effects of Air Pollution

Type M and S errors are two concepts that highlight the danger of having too much confidence in statistically significant estimates when studies are under-powered. This issue is virtually absent from the literature but studies on the acute health effect of air pollution could be under-powered for several reasons. First, researchers work with observational data and can often not control the sample size of their studies due to data availability. Very few guidance on the drivers of studies' statistical power actually exists (Winquist et al. 2012, Bhaskaran et al. 2013). Moreover, reaching a large statistical power could be challenging since estimated effect sizes are remarkably small and the modeling of high-frequency variations in daily mortality or emergency admission is difficult (Peng et al. 2006, Peng and Dominici 2008). Finally, we observe both in the standard epidemiology and the causal inference literatures a negative relationship between estimated effect sizes and studies' precision. It is important to investigate if this pattern could be explained by imprecise studies making type M errors (Ioannidis 2008, Gelman and Carlin 2014, Ioannidis et al. 2017, Ferraro and Shukla 2020).

Retrospective Analysis of the Literature

In this section, we first explain how to implement a retrospective analysis of a study. Using different scenarios about the true effect sizes of studies found in the standard epidemiology and causal inference literatures, we then assess to what extent they could suffer from low statistical power issues.

How to Run a Retrospective Analysis

Based on the notations of Zwet and Cator (2021), the statistical framework for a retrospective power analysis can be formalized as follows. We take the general case where we want to estimate, for a given research design, the causal effect β of an air pollutant on an health outcome. We assume that we have a normally unbiased estimate b of β with a standard error s. If we knew the true value of β, we could first compute the statistical power of our study for rejecting the null hypothesis $H_0 : \beta = 0$. It is defined as $\Phi(-1.96 - \frac{\beta}{s}) + 1 - \Phi(1.96 - \frac{\beta}{s})$, where Φ is the cumulative function of the standard normal distribution. Given the power of our study, we could then compute the average magnitude of a type M error, that is to say the average inflation of statistically significant estimates at the 5% level for the study. It can be expressed as $\mathbb{E}(\frac{|b|}{|\beta|}|\beta, s, |b|/s > 1.96)$. Finally, we could compute the probability to make a type S error, that is to say the probability that the estimate is of the opposite sign of β. It is given by $\frac{\Phi(-1.96-\frac{\beta}{s})}{1-\Phi(1.96-\frac{\beta}{s})+\Phi(-1.96-\frac{\beta}{s})}$. Unfortunately, we never observe the true value of the causal estimand of interest. We therefore need to make guesses about the value of β to compute the three previous metrics. Gelman and Carlin (2014) propose to run simulations to calculte them. In our project, we rely on the closed-form expressions derived by Lu et al. (2019) and their implementation in the **R** package `retrodesign` developed by Timm (2019).

It is very important to keep in mind that the usefulness of a retrospective power analysis relies entirely on a credible guess of the true effect size a study is trying to estimate. As the true effect is never observed, researchers can have very different priors on its magnitude.

They could therefore assess differently the extent to which a study risks to suffer from statistical power issues. To illustrate this tension, we provide below a case study showing how a scientific discussion about effect sizes arises with a retrospective analysis.

In a flagship publication, Deryugina et al. (2019) instrument $PM_{2.5}$ concentrations with wind directions to estimate its effect on mortality, health care use, and medical costs among the US elderly. They gathered 1,980,549 daily observations at the county-level over the 1999–2013 period; it is one of the biggest sample sizes in the literature. When the authors instrument $PM_{2.5}$ with wind direction, they find that "a 1 μg/m^3 (about 10 percent of the mean) increase in $PM_{2.5}$ exposure for one day causes 0.69 additional deaths per million elderly individuals over the three-day window that spans the day of the increase and the following two days". The estimate's standard error is equal to 0.061. In Figure 19, we plot the statistical power, the inflation factor of statistically significant estimates and the probability that they are of the wrong sign as a function of hypothetical true effect sizes.

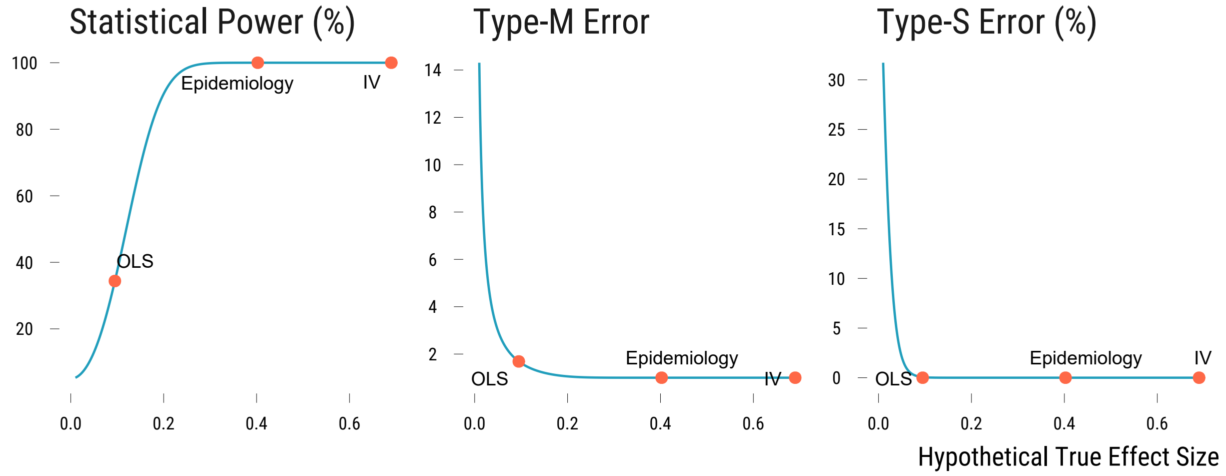


Figure 19: Power, Type M and S Errors Curves for Deryugina et al. (2019). *Notes*: In each panel, a metric, such as the statistical power, the exaggeration ratio or the probability to make a type S error, is plotted against the range of hypothetical effect sizes. The "IV" label represents the value of the corresponding metric for an effect size equal to Deryugina et al. (2019)'s two-stage least square estimate. The "Epidemiology" label stands for the estimate found in Di et al. (2017), which is the epidemiology article most similar to Deryugina et al. (2019). The "OLS" label corresponds to the estimate found by Deryugina et al. (2019) when the air pollutant is not instrumented.

The estimate found by Deryugina et al. (2019) represents a relative increase of 0.18% in mortality. We labeled it as "IV" in Figure 19. Is this estimated effect size large compared to those reported in the standard epidemiology literature? We found a similar article to draw a comparison. Using a case-crossover design and conditional logistic regression, Di et al. (2017) find that a 1 μg/m^3 increase in $PM_{2.5}$ is associated with a 0.105% relative increase in all-cause mortality in the Medicare population from 2000 to 2012. The effect size found by Deryugina et al. (2019) is larger than this estimate labeled as "Epidemiology" in Figure 19. If the estimate found by Di et al. (2017) was actually the true effect size of $PM_{2.5}$ on elderly mortality, the study of Deryugina et al. (2019) would have enough statistical power to perfectly avoid type M and S errors. Now, suppose that the true effect of the increase in $PM_{2.5}$ was 0.095 additional deaths per million elderly individuals—the estimate the authors found with a "naive" multivariate regression model. The statistical power would be 34%, the probability to make a type S error could be null but the overestimation factor would be on average equal to 1.7. Even with a sample size of nearly 2 million observations, Deryugina et al. (2019) could make a non-negligible type M error if the true effect size was the naive ordinary least square estimate. Yet, the authors could argue that their instrumental variable strategy leads to a higher effect size as it overcomes unmeasured counfounding bias and measurement error. Besides, for effect sizes down to 0.182 additional deaths per million elderly individuals (a 0.05% relative increase), their study has a very high statistical power and would not run into substantial type M error. A retrospective analysis is thus a very convenient way to think about the statistical power of a study to accurately detect alternative effect sizes.

Standard Epidemiology Literature

Hundreds of papers have been published on the short-term health effects of air pollution in epidemiology, medicine and public health journals. A large fraction of articles are based on Poisson generalized additive models, which allow to flexibly adjust for the temporal trend of health outcomes and for non-linear effects of weather parameters. This literature spans over 20 years and has replicated analyses in a large number of settings, providing crucial insights on the acute health effect of air pollution. Advocates of causal methods would surely argue that these articles could suffer from omitted variable biases. Even if they may be more biased, we find it valuable to assess their potential statistical power issues and compare them with causal inference papers.

To gather a corpus of relevant articles, we use the following search query on PubMed and Scopus to select studies on the short-term health effects of air pollution:

```
’TITLE((”air pollution” OR ”air quality” OR ”particulate matter” OR
”ozone”’, ’OR ”nitrogen dioxide” OR ”sulfur dioxide” OR ”PM10” OR ”PM2.5”
OR’, ’ ”carbon dioxide” OR ”carbon monoxide”)’, ’AND (”emergency” OR
”mortality” OR ”stroke” OR ”cerebrovascular” OR’, ’”cardiovascular” OR
”death” OR ”hospitalization”)’, ’AND NOT (”long term” OR ”long-term”))
AND ”short term”’
```

We retrieve the abstracts of 1834 articles. We then extract estimates and confidence intervals from these abstracts using regular expressions (regex). We illustrate this procedure using one sentence of a randomly selected article from this literature review (Vichit-Vadakan et al. 2008):

> ”The excess risk for non-accidental mortality was **1.3% [95% confidence interval (CI), 0.8–1.7]** per 10 $\mu g/m^3$ of PM10, with higher excess risks for cardiovascular and above age 65 mortality of **1.9% (95% CI, 0.8–3.0)** and **1.5% (95% CI, 0.9–2.1)**, respectively.”

Our algorithm detects phrases such as “95% confidence interval (CI)” or “95% CI” and looks for numbers directly before this phrase or after and in a confidence interval-like format. Using this method, we retrieve 2666 estimates from 784 abstracts. We then read these abstracts and filter out articles whose topic falls outside of the scope of our literature review. Our corpus is thus composed of 668 articles for which we detect 2155 estimates. Importantly, the set of articles considered is limited to those displaying confidence intervals and point estimates in their abstracts.

Based on this subset of articles, we implement a retrospective analysis in which we check the overall sensitivity of studies for true effect sizes expressed as fraction of observed estimates. Without carefully reading each article, we cannot make more informed guesses about true effect sizes since estimates are expressed for different increases in air pollution concentration. We think that our rough approach is still valuable since a well-designed study should be able to detect effect sizes smaller than the estimated one. For instance, if we find that a study has a statistical power of 30% when we assume that the true effect is 25% lower than the measured estimate, it is likely that the study is not very robust to statistical power issues.

Our results for the standard epidemiology literature are at first sight reassuring. If the true effect sizes of the studies were equal to 75% of estimated coefficients, the median statistical power would be equal to 85% and the median exaggeration factor would be only 1.1. At least 50% of this literature does not seem to suffer from substantial statistical power issues since their power would be above 80%. Type S error is not an issue for most articles. Yet, even if the measured effect was close to the true effect, a non-negligible proportion of articles would display low statistical power and presents a substantial risk of making a type M error. About 47% of estimates would not reach the conventional 80% statistical power threshold if the true effect was 75% the size of the measured effect. For these under-powered studies, the average type M is 1.9 and the median 1.5. We also observe that the proportion of under-powered studies has been stagnating since the 1990s, revealing that practices regarding statistical power have not evolved over time.

Finally, skeptic researchers could rightly complain that assuming for each study a true effect size equal to 75% of the estimate is arbitrary. To overcome this criticism, we expand our review of the standard epidemiology literature by running statistical power calculations based on two recent meta-analyses: one by Shah et al. (2015) on mortality and emergency admission for stroke, and the other one by Orellano et al. (2020) on broader causes of mortality. We use the meta-analysis estimates as true effect sizes for the 290 studies gathered by Shah et al. (2015) and . This is the approach recommended by Gelman and Carlin (2014) and Ioannidis et al. (2017) to make more informed guesses about true effect sizes. 60% of studies in Orellano et al. (2020) have a statistical power below 80%. The median exaggeration ratio of statistically significant estimate is equal to 2. The proportion of under-powered studies is similar in Shah et al. (2015) but the median type M error is equal to 3. With more informed guesses about true effect sizes, we clearly see that under-powered studies are an issue in the standard epidemiology literature.

Causal Inference Literature

Using Google Scholar and PubMed, we search papers using causal inference methods and investigating the short-term effects of air pollution on mortality or emergency admission outcomes. Specifically, we only consider articles that exploit short-run exogenous shocks such as air pollution alerts, public transport strikes, changes in wind direction, thermal inversions, to name but a few. For instance, we did not select articles on the impact of low emission or congestion pricing zones as they evaluate health effects over several months or years. In Figure 20, we display the 29 articles that match our search criteria. We read each article and retrieve the estimates and standard errors for the main results: for simplicity, we only select one of the main results discussed by the researchers. We also record the numbers of observations and summary statistics on the outcome and independent variables to compare studies by standardizing the estimated effect sizes.

Figure 20: Our Corpus of Papers from the Causal Inference Literature.

Article	**Location**	**Health Outcome**	**Independent Variables**	**Study Design**
Arceo et al. (2016)	Mexico City, Mexico	Infant Mortality	PM10, Thermal Inversion (IV)	Instrumental Variable
Austin et al. (2020)	Counties, USA	Rates of Confirmed COVID-19 Deaths	PM2.5 (air pollutant), Wind Direction (IV)	Instrumental Variable
Baccini et al. (2017)	Milan, Italy	Non-Accidental Mortality	Dummy for PM10 Concentration >To 40 $\mu g/m^3$	Propensity Score Matching
Barwick et al. (2018)	All Cities, China	Number of Health Spending Transactions	PM2.5, Spatial Spillovers of PM2.5 (IV)	Instrumental Variable
Bauernschuster et al. (2017)	5 Largest Cities, Germany	Admissions for Abnormalities of Breathing (age below 5)	PM10, Public Transport Strikes Dummy	Difference in Differences
Beard et al. (2012)	Salt Lake County, USA	Emergency Visits For Asthma	Thermal Inversions	Time-stratified case-crossover design
Chen et al. (2018)	Toronto, Canada	Asthma-Related Emergency Department Visits	Air Quality Eligibility, Air Quality Altert	Fuzzy Regression Discontinuity
Deryugina et al. (2019)	Counties, USA	All Causes of Mortality (Age 65+)	PM2.5, Wind Direction (IV)	Instrumental Variable
Ebenstein et al. (2015)	2 Cities, Israel	Hospital Admissions Due To Lung Illnesses	PM10 (air pollutant), Sandstorms (IV)	Instrumental Variable
Forastiere et al. (2020)	Milan, Italy	Non-Accidental Mortality	Setting PM10 Daily Exposure Levels >To 40 $\mu g/m^3$ To 40	Generalized Propensity Score
Giaccherini et al. (2021)	Municipalities, Italy	Respiratory Hospital Admission	PM10, Public Transport Strikes	Difference in Differences
Godzinski and Suarez Castillo (2019)	10 Cities, France	Emergency Admissions for Upper Respiratory System (Age 0-4)	CO, Public Transport Strikes	Difference in Differences
Halliday et al. (2019)	Hawaii, USA	ER Admission for Pulmonary Outcomes	PM2.5, SO2 Emissions From Kilauea Volcano and Wind Direction (IV)	Instrumental Variable
He et al. (2016)	34 Urban Districts, China	Monthly Standardized Mortality Rate	PM10, Regulation and Traffic Control Status (IV)	Instrumental Variable
He et al. (2020)	China	Monthly Number of Deaths for All-Causes	PM2.5, Straw Burning (IV)	Instrumental Variable
Isphording and Pestel (2021)	Counties, Germany	Mortality of Covid-19 Positive Male Patients (Age 80+)	PM10, Wind direction (IV)	Instrumental Variable
Jans et al. (2018)	Sweden	Children Health Care Visits for Respiratory Illness	PM10, Thermal Inversion (IV)	Reduced-Form
Jia and Ku (2019)	South Korea	Mortality Rates for Respiratory and Cardiovascular Diseases	Dusty Days Times China's AQI	Reduced-Form
Kim et al. (2013)	South Korea	Hospital Admissions for Respiratory Illnesses	PM10 (air pollutant), Average PM10 Level By Date (IV)	Instrumental Variable
Knittel et al. (2016)	California, USA	Infant Mortality	PM10, Road Traffic Flow and Weather variables (IV)	Instrumental Variable
Moretti and Neidell (2011)	South California, USA	Hospital Admissions for Respiratory Illnesses	O3, Vessel Traffic (IV)	Instrumental Variable
Mullins and Bharadwaj (2015)	Santiago Metropole, Chile	Cumulative Deaths (age >64)	PM10, Air quality Alerts	Matching + Difference in Differences
Schlenker and Walker (2016)	California, USA	Acute Respiratory Hospitalization	CO, Planes Taxi Time (IV)	Instrumental Variable
Schwartz et al. (2015)a	Boston, USA	Non-Accidental Mortality	PM2.5, Back Trajectories of PM2.5 (IV)	Instrumental Variable
Schwartz et al. (2017)	Boston, USA	Non-Accidental Mortality	PM2.5, Height Of Planetary Boundary Layer and Wind Speed (IV)	Instrumental Variable
Schwartz et al. (2018)	135 Cities, USA	Non-Accidental Mortality	PM2.5, Planetary Boundary Layer, Wind Speed, and Air Pressure (IV)	Instrumental Variable
Sheldon and Sankaran (2017)	Singapore	Acute Upper Respiratory Tract Infections	Pollutant Index, Indonesian Fire Radiative Power (IV)	Instrumental Variable
Williams et al. (2019)	USA	Asthma Rescue Event	PM2.5	Poisson fixed-effects models
Zhong et al. (2017)	Beijing, China	Ambulance Call Rate for Coronary Heart Problem	NO2, Number 4 Day (IV)	Instrumental Variable

Notes: For each study, we report its location, one of the health outcome analyzed, the independent variables (the air pollutant and in the case of an instrumental variable strategy, the instrument) and the study design.

To evaluate potential statistical power issues in this literature, we first proceed exactly as for the standard epidemiology literature. We compute the statistical power, the exaggeration factor and the probability to get an estimate of the wrong sign for all studies based on hypothetical true effect sizes expressed as decreasing fraction of observed estimates. In Figure 21, each gray line represent the statistical power and average type M error curves of an article. The blue lines represent the average power and exaggeration factor of all causal inference studies.

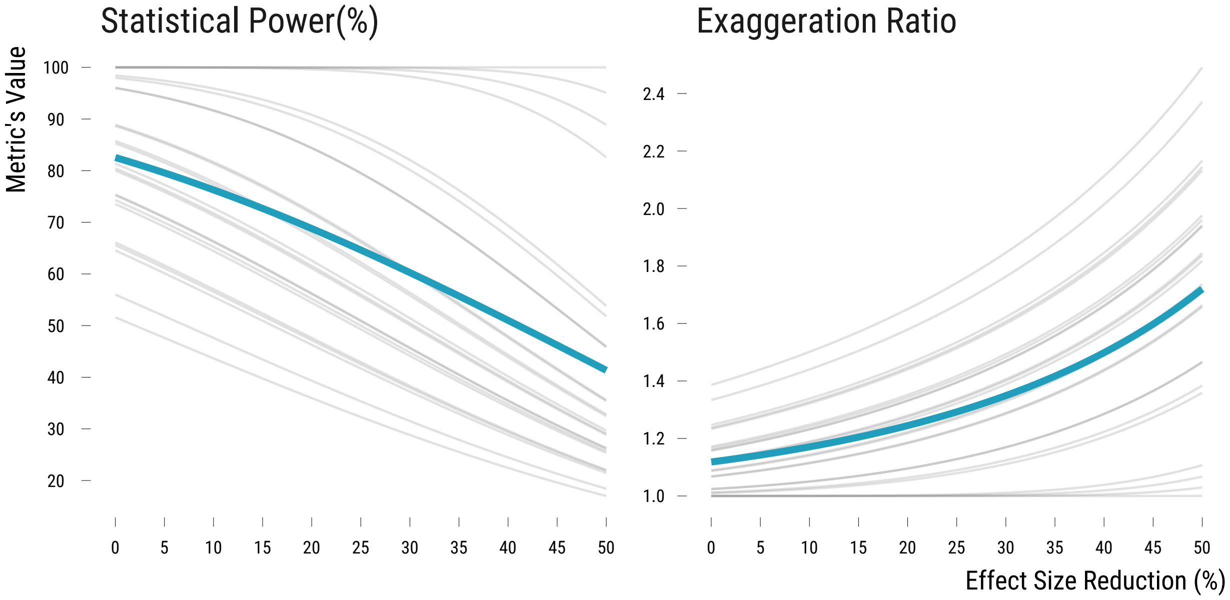


Figure 21: Statistical Power and Type M Error of Causal Inference Studies. *Notes*: For each causal inference paper, we compute its statistical power and the average type M error for decreasing effect sizes expressed as percentage reduction in observed estimates. Each gray line represents a specific causal inference paper. The blue lines are the average of a metric for all causal inference papers.

If the true effect size of each study was equal to 75% of the estimate, the median statistical power would be about to 60% and the median Type M error would be 1.3. In the causal inference literature, at least half of studies have enough statistical power so that statistically significant estimates are not inflated. In Figure 21, we can however see that there is a wide heterogeneity in the robustness of studies to statistical power issues—some of them are relatively well powered while others run quickly into Type M error. A large share of studies in the literature would not have designs with enough statistical power to detect effects of half the size of their observed estimates. In that scenario, the median statistical power would be about 40% and the median type M error would be 1.8. Overall, this comprehensive retrospective analysis of the literature reveals that some studies are under-powered and could run into type M error. It may help explain why there is a large heterogeneity in effect sizes across articles.

Again, expressing true effect sizes as decreasing fraction of observed estimates is arbitrary. We also carry out another retrospective analysis where we take as true effect sizes the estimates that would be predicted using non-causal inference methods. We do so for the subset of the 9 instrumental variables that also display estimates in the case when the air pollutant concentration is not instrumented. Two reasons are often advanced in the causal literature to explain the discrepancy between instrumented and non-instrumented estimates: (i) instrumental variables help overcoming omitted variable bias and (ii) if the air pollution is measured with classical error, instrumental variables also reduce the resulting attenuation bias. We think that, for some studies, statistical power issues could also partly explain the observed difference between causal and non-causal methods. In Table 4, we display the statistical power, the average type M error and the probability to make a type S error for instrumental variable studies. For some studies, the statistical power of the instrumental variable strategy could be extremely low. This results in large type M errors, which magnitude partially close the gap between instrumented and non-instrumented estimates. Given this possibility, future research should carry out quantitative bias analysis to explore the trade-off between using an instrumental variable strategy to overcome omitted variable and attenuation biases and running into a type M error due to low statistical power (Rosenbaum 2010, Dorie et al. 2016, VanderWeele and Ding 2017, Cinelli and Hazlett 2020).

Table 4: Retrospective Analysis of Instrumental Variable Papers Where Naive OLS Estimates are Assumed to be True Effect Sizes.

Paper	**Power (%)**	**Type S Error (%)**	**Type M Error**
Giaccherini et al. (2021)	5	43.3	40.7
Halliday et al. (2019)	6	16.6	6.9
Schlenker and Walker (2016)	7	13.8	6.1
Moretti and Neidell (2011)	11	3.7	3.5
Arceo et al. (2016)	12	2.4	3.1
Barwick et al. (2018)	23	0.3	2.1
Deryugina et al. (2019)	34	0.1	1.7
Ebenstein et al. (2015)	52	0	1.4
Schwartz et al. (2018)	64	0	1.3

Notes: For each study based on an instrumental variable strategy, we computed the statistical power, the average type M error and the probability to make a type S error using the non-instrumented estimate as a guess for the true effect size.

Prospective Analysis of Causal Inference Methods

The review of the standard epidemiology and causal literatures shows that some articles could have produced inflated estimates on the short-term health effects of air pollution. This analysis however does not allow us to clearly identify which parameters of a study influence its statistical power. We therefore implement a prospective analysis to overcome this limitation (Gelman and Carlin 2014, Altoè et al. 2020). We run simulations based on real-data to emulate the main empirical strategies found in the literature. Using real data frees us from the difficult task to model the long-term and seasonal variations in health outcomes but also the specific effects of weather variables such as temperature. In this section, we describe how we implement these simulations. We start by presenting the research designs we wish to simulate, then briefly describe the data we rely on and finally detail how we actually simulate the research designs.

Research Designs Simulated

Several empirical strategies have been implemented to estimate the short-term health effects of air pollution. In our simulations, we try to simulate the main ones found in the literature. We assume below that we are working with data consisting in daily time series of various health outcomes, air pollutant concentrations and weather parameters for a sample of cities.

Standard regression approach. The standard strategy consists in directly estimating the dose-response between an air pollutant and an health outcome. In the epidemiology literature, researchers often rely on Poisson generalize additive models where the daily count of an health outcome is regressed on the concentration of an air pollutant, while flexibly adjusting for weather parameters, seasonal and long-term variations. Because most causal methods are estimated with linear regression, our simulations are instead based on ordinary least square estimation to approximate the warhorse model used by epidemiologists. We can summarize this approach with the following model:

$$Y_{c,t} = \alpha + \beta P_{c,t} + \mathbf{W}_{c,t}\phi + \mathbf{C}_{c,t}\gamma + \epsilon_{c,t}$$

where c is the city index and t the daily time index. $Y_{c,t}$ is the daily count of cases of an health outcome and $P_{c,t}$ the average daily concentration of an air pollutant. The coefficient β measures the short-term effect of an increase in the air pollutant concentration on the health outcome. To address confounding issues, the model adjusts for a set of weather covariates, $\mathbf{W}_{c,t}$, and calendar indicators $\mathbf{C}_{c,t}$. The error term is denoted by $\epsilon_{c,t}$.

Instrumental variable approach. The standard strategy could be prone to omitted variable bias and measurement error. A growing number of articles therefore exploit exogenous variations in air pollution. Most causal inference papers rely on instrumental variable designs where the concentration of an air pollutant is instrumented by thermal inversions (Arceo et al. 2016), wind patterns (Schwartz et al. 2017, Deryugina et al. 2019, Isphording and Pestel 2021), extreme natural events such as sandstorms or volcano eruptions (Ebenstein et al. 2015, Halliday et al. 2019), or variations in transport traffic (Moretti and Neidell 2011, Knittel et al. 2016, Schlenker and Walker 2016). This approach can be summarized with a two-stage model where the first stage is:

$$P_{c,t} = \delta + \theta Z_{c,t} + \mathbf{W}_{c,t}\phi + \mathbf{C}_{c,t}\gamma + e_{c,t}$$

where $Z_{c,t}$ is the instrumental variable. The second stage is then:

$$Y_{c,t} = \alpha + \beta \widehat{P}_{c,t} + \mathbf{W}_{c,t}\psi + \mathbf{C}_{c,t}\lambda + \epsilon_{c,t}$$

where $\widehat{P}_{c,t}$ is the exogenous variation in an air pollutant predicted by the instrument. The causal effect measured by this approach is a weighted average of per-unit causal responses to an increase in the concentration of an air pollutant (Angrist and Imbens 1995).

Reduced-form approach. A subset of articles however restrain from instrumenting the concentrations of an air pollutant with exogenous shocks, but instead choose to directly estimate the relationship between the health outcome and the shocks. This why we call this approach a reduced-form analysis. The articles concerned by this approach focus on public transport strikes (Bauernschuster et al. 2017, Godzinski et al. 2019, Giaccherini et al. 2021). Their empirical strategy can be summarized with the following model:

$$Y_{c,t} = \alpha + \beta Z_{c,t} + \mathbf{W}_{c,t}\phi + \mathbf{C}_{c,t}\gamma + \epsilon_{c,t}$$

The coefficient β captures a type of intention-to-treat effect.

Regression-discontinuity design approach. The final empirical strategy found in the literature consists in measuring the effects of air quality alerts with a regression-discontinuity design (Chen et al. 2018). The approach is summarized with the model:

$$g\{E(Y_{c,t})\} = \beta_0 + \beta_1(I_{c,t} - l) + \beta_2 E_{c,t} + \beta_3(I_{c,t} - l) \times E_{c,t} + \epsilon_{c,t}$$

where $g(.)$ is a generic link function, $I_{c,t}$ is the daily air quality index of a city, $E_{c,t}$ is the indicator for the eligibility to issue an air quality alert whose threshold is l. This approach estimates the intention-to-treat effect of air quality alerts. It is important to note that it can capture both the effect due to a subsequent decrease in air pollution due to traffic restriction policies and the effect caused by inhabitants' avoidance behavior.

Data

Our simulation exercises are based on a subset of the US National Morbidity, Mortality, and Air Pollution Study (NMMAPS). The dataset has been exploited in several major studies of the early 2000s to measure the short-term effects of ambient air pollutants on mortality outcomes (Peng and Dominici 2008). It is publicly available and allows us to work with increasing sample sizes for our simulations. Specifically, we extracted daily data on 68 cities over the 1987-1997 period, which represent 4,018 observations per city, for a total sample size of 273,224 observations. For each city, the average temperature (C°), the standardized concentration of carbon monoxide (CO), and mortality counts for several causes are recorded. We choose to work with CO as it is the air pollutant measured in most cities over the period.

Less than 5% of carbon monoxide concentrations and average temperature readings are missing in the initial data set and we impute them using the chained random forest algorithm provided by the `missRanger` package (Mayer 2019).

Simulations Set-Up

Simplifying assumptions. To only capture the specific issues arising due to low statistical power, we make several simplifications to make sure that all the assumption of empirical strategies are met. First, in all research designs, there is no bias due to unmeasured confounders or measurement errors. All models retrieve on average the true value of the treatment effect we set in the data. Second, for instrumental variable and reduced-form strategies, we only simulate exogenous shocks that are binary (e.g. the occurrence of a thermal inversion or not). They are randomly allocated. Third, for the regression discontinuity approach, we only model sharp designs where an air quality is always activated above a randomly chosen threshold.

Two approaches for simulating research designs. We take two different approaches to simulate the research designs. For the reduced-form and regression discontinuity designs, we follow the Neyman-Rubin causal framework by creating a Science table (Rubin 1974). The recorded value of a health outcome in the dataset represent the potential outcome $Y_{c,t}(0)$ of that day t in city c when it is not exposed to the treatment denoted by $W_{c,t}$. It is equal to 1 when a treatment occurs and 0 otherwise. To create the counterfactuals $Y_{c,t}(1)$, we add a treatment effect drawn from a Poisson distribution whose parameter correspond to the magnitude of the treatment. We then randomly draw the treatment indicators $W_{t,c}$ for exogenous shocks or air quality alerts. For reduced-form strategies, the treatment status of each day is drawn from a Bernoulli distribution with parameter equal to the proportion of exogenous shocks desired. For air pollution alerts, we randomly draw a threshold from a uniform distribution and select a bandwidth such that it yields the correct proportion of treated observations. We finally express the observed values Y^{obs} of potential outcomes according to the treatment assignment: $Y^{obs}_{c,t}$ = (1-$W_{c,t}$)$\times Y_{c,t}(0)$ + $W_{c,t} \times Y_{c,t}(1)$.

For the standard regression and the instrument variable strategies, we rely on a model-based approach. For the standard regression strategy, we first estimate the following statistical model on our data:

$$Y_{c,t} = \alpha + \beta Z_{c,t} + \mathbf{W}_{c,t}\phi + \mathbf{C}_{c,t}\gamma + \epsilon_{c,t}$$

We then predict new observations of a $Y_{c,t}$ using the estimated coefficients of the model ($\hat{\beta}$, $\hat{\phi}$, and $\hat{\gamma}$) and by adding noise drawn from the residuals distribution $\widehat{\epsilon_{c,t}}$ (Peng et al. 2006). We modify the slope of the dose-response relationship by changing the value of the air pollution coefficient β. For the instrumental variable strategy, we use the same method as for the standard regression approach but first modify observed air pollutant concentrations $P_{c,t}$ according to the desired effect size θ of the randomly allocated instrument:

$$P_{c,t} = P_{c,t} + \theta Z_{c,t}$$

The allocation of each day to an exogenous shock is drawn from a Bernoulli distribution with parameter equal to the proportion of exogenous shocks desired. We then estimate a two-stage least squares model, modify the coefficient for the effect of the air pollutant on an health outcome, and finally generate the fake observations of the health outcome using coefficients of the two-stage least squares model and noise drawn from the residuals of the estimated model.

General procedure. Our simulation procedure therefore follows 7 main steps:

1. Randomly draw a study period and a sample of cities.
2. For instrumental variable, reduced-form and regression-discontinuity designs, randomly allocate days to binary exogenous shocks/air quality alerts.
3. Add the treatment effect size of interest.
4. Run the model of the empirical strategy.
5. Store the point estimate of interest and its standard error.
6. Repeat the procedure 1000 times.
7. Finally compute the statistical power, the exaggeration ratio of statistically significant estimates and the probability that they are of the wrong sign.

Varying parameters. To understand which parameters affect statistical power issues, we can change one aspect the research design while keeping other parameters values constant. We consider in our simulations four main parameters. First, by drawing a different number of cities and changing the length of the period, we vary the sample size. Second, we change the effect size of air pollution or an exogenous shock on an health outcome. Third, we also allocate increasing proportions of exogenous shocks/air quality alters. Fourth, we can run similar simulations but for different health outcomes with small or large number of cases per day.

Simulating Flagship Studies.

Our simulations based on real data help explore the consequences of varying each parameter on statistical power issues. We could however make the results of our simulations even more informative by setting realistic values for the four parameters of a research design. We therefore also try to reproduce three flagship studies using our own dataset.

Results

In this section, we first explore how statistical power is related to the treatment effect size, the number of observations, the proportion of treated units and the distribution of the health outcome. We then try to replicate the design of flagship publications to highlight their potential weaknesses with respect to low statistical power issues.

Evolution of Power, Type M and S Errors with Study Parameters

First, we analyze how statistical power, type M and S errors are affected by the value of different study parameters. To do so, we set baseline values for these parameters and vary the value of each of them one by one. This enables us to get a sense of the impact of each parameter, other things being equal. The baseline parameters are such that:

- The sample size is equal to 100,000 observations (2500 days × 40 cities).
- The effect size of air pollution or an exogenous shock is equal to a 1% relative increase in an health outcome.
- The proportion of exogenous shocks represents 50% of observations. For air pollution alerts analyzed with regression discontinuity designs, we choose a smaller proportion of treated units: 10%.
- The health outcome is the total daily number of non-accidental deaths. It is the health outcome with the largest average number of counts—the average daily mean is equal to 23 cases.

For all statistical models, we adjust for temperature, temperature squared, city and calendar (weekday, month, year, month×year) fixed effects. We also repeat the simulations for a smaller sample size of 10,000 observations.

Sample Size As shown in Figure 22, we unsurprisingly find that, for all identification methods, statistical power increases and type M error decreases with the number of observations.

Yet, statistical power and type M error issues arise even for a large number of observations. For a sample size of 40,000 observations, an instrumental variable strategy would only have a statistical power of 54% and would overestimate the true effect by a factor of 1.4. On the contrary, a standard regression strategy is much less prone to power issues than the instrumental variable strategy. This is explained by the fact that the variance of the two stage least-square estimator is larger than the variance of the ordinary least square estimator. In our simulations, we also note that, for all identification method, Type S error is not a problem for any sample sizes.

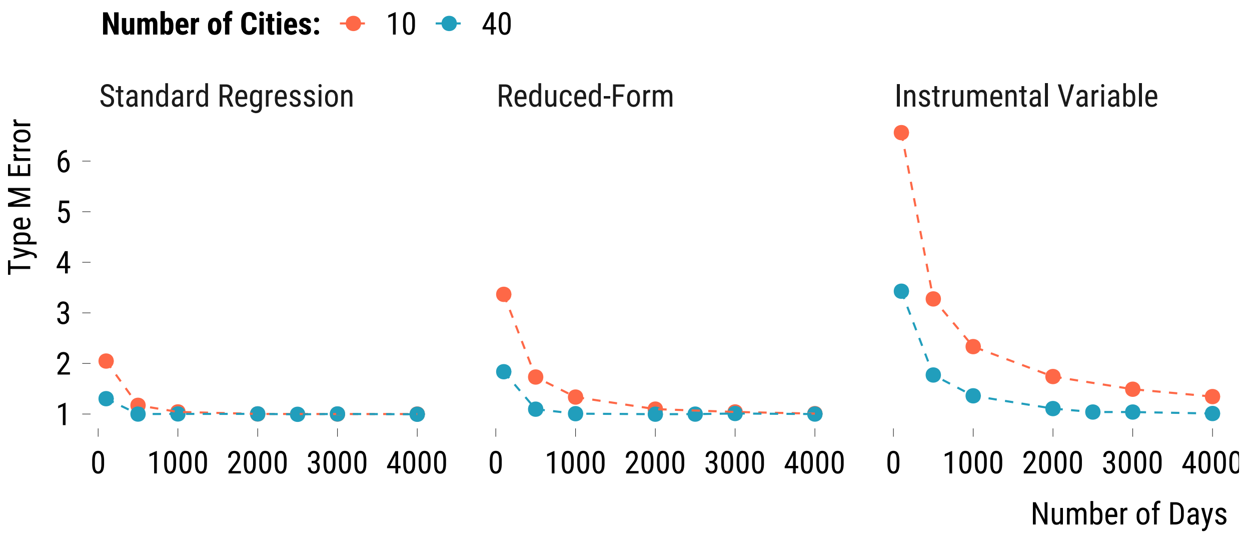


Figure 22: Evolution of Type M Error against Sample Size. *Notes*: The true effect size is a 1% relative increase in the health outcome. The health outcome used in the simulations is the total number of non-accidental deaths. The proportion of exogenous units is 50% for instrumental variable and reduced-form designs.

Effect Size The second unsurprising result of our simulations is that the larger the effect size, the larger the power and the lower type M and S errors are. With our advantageous baseline parameters, statistical power issues however start to appear in instrumental variable and regression discontinuity designs for effect sizes below 1%. For instance, for an effect of 0.5%, the average type M error is about 1.7. Such effect sizes are similar to those sometimes found in the standard epidemiology literature. As for results on sample sizes, standard regression and reduced-form strategies suffer less from power issues, even for small effects.

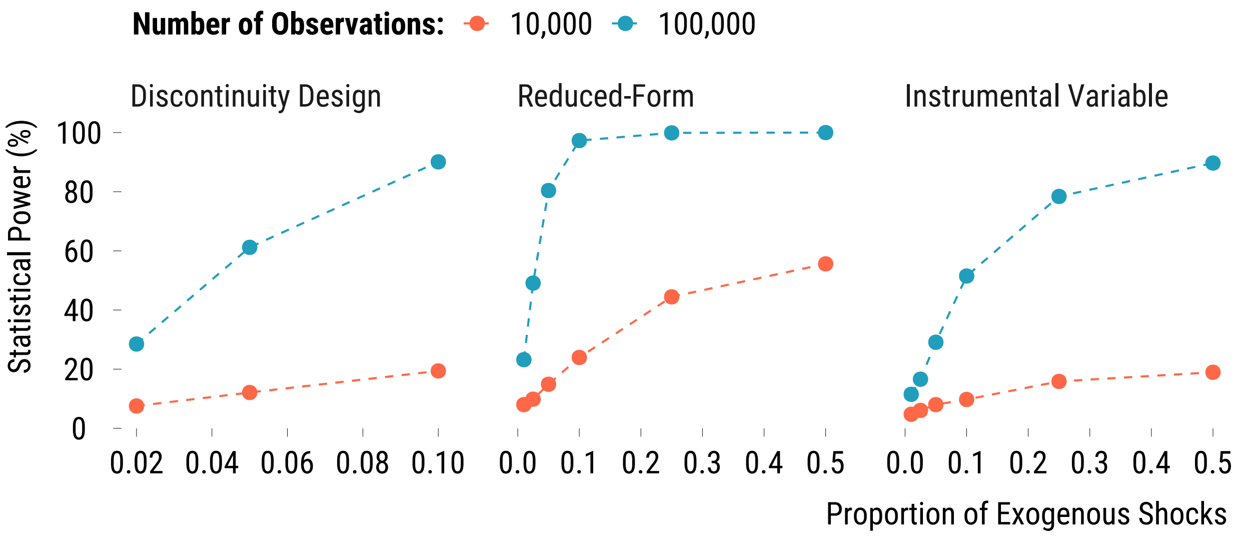


Figure 23: Evolution of Statistical Power with the Proportion of Exogenous Shocks. *Notes*: The true effect size is a 1% relative increase in the health outcome. The health outcome used in the simulations is the total number of non-accidental deaths.

Proportion of Exogenous Shocks The link between the proportion of exogenous shocks and statistical power might be less known to researchers. In Figure 23, we see that the statistical power increases with larger proportions of treated units for instrumental variable, regression discontinuity and reduced-form designs. As in the case of randomized controlled trials, the precision of studies will be maximized when half of the observations are exposed to the treatment of interest.

Conversely, as shown in Figure 24, the average Type M error increases as the proportion of exogenous shocks decreases.

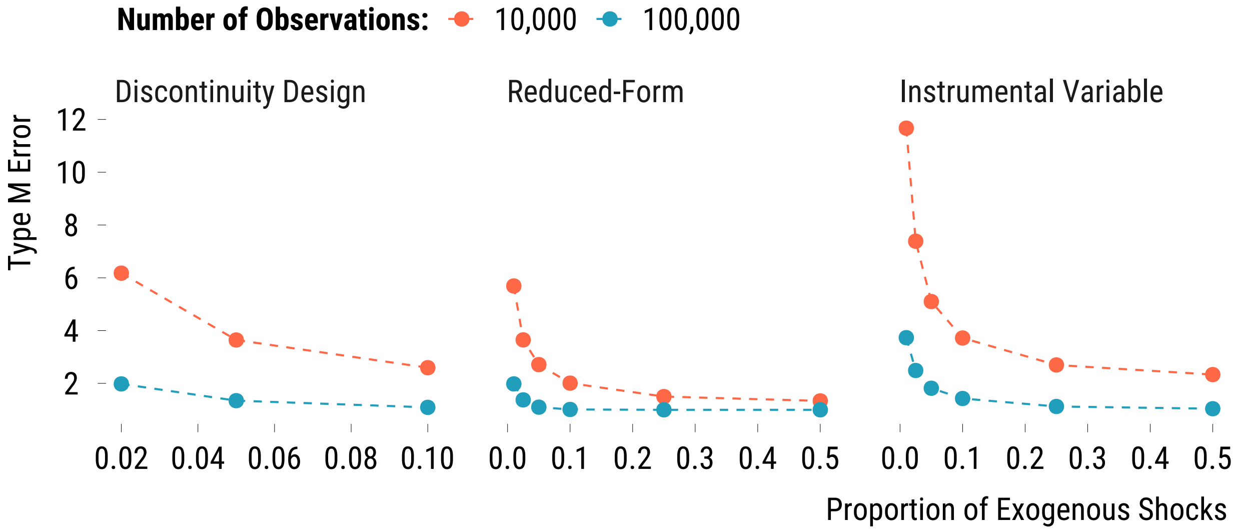


Figure 24: Evolution of Type M Error with the Proportion of Exogenous Shocks. *Notes*: The true effect size is a 1% relative increase in the health outcome. The health outcome used in the simulations is the total number of non-accidental deaths.

Air pollution alerts, thermal inversion or transportation strikes are however rare events. They can represent less than 5% of the observations in some studies. With a dataset of 10,000 observations, the average type M error is 2.7 for reduced-form strategies. The causal inference literature might therefore be particularly prone to type M error due to a very low proportion of treated units, even though sample sizes are often large.

Average Count of Cases of the Health Outcome Perhaps less known to economists than the influence of sample and effect sizes, the average count of cases also critically affects statistical power. For instance, a 1% increase in the number of deaths in a setting where there are only 2 deaths per day corresponds to rare additional deaths that might therefore be more difficult to detect. To simulate situations with various number of cases, we consider three different outcome variables, with different counts of cases: the total number of non-accidental deaths (daily mean $\simeq$ 23), the total number of respiratory deaths (daily mean $\simeq$ 2) and the number of chronic obstructive pulmonary disease cases for people aged between 65 and 75 (daily mean $\simeq$ 0.3). Using baseline parameters and in the case of the large dataset, we find that statistical power is close to 100% when empirical strategies target a 1% increase in the total number of non-accidental deaths. However, statistical power quickly drops when the average count of cases decreases. For instance, an instrumental variable strategy has only 16% of statistical power to detect an increase by 1% in respiratory deaths. The average type M error is then equal to 2.4. For chronic obstructive pulmonary deaths, the situation is even worse, with an average type M error of 5.9. Studies with a small count of cases may therefore lead to extreme statistical power issues.

Issues Specific to the Instrumental Variable Design For instrumental variable strategies, we also analyze how the statistical power is affected by the strength of the instrument. In our simulations, we define the strength of the instrument as the standardized effect size on the air pollutant concentration. A strength equals to 0.2 means that the instrument increases the concentration by 0.2 standard deviation.

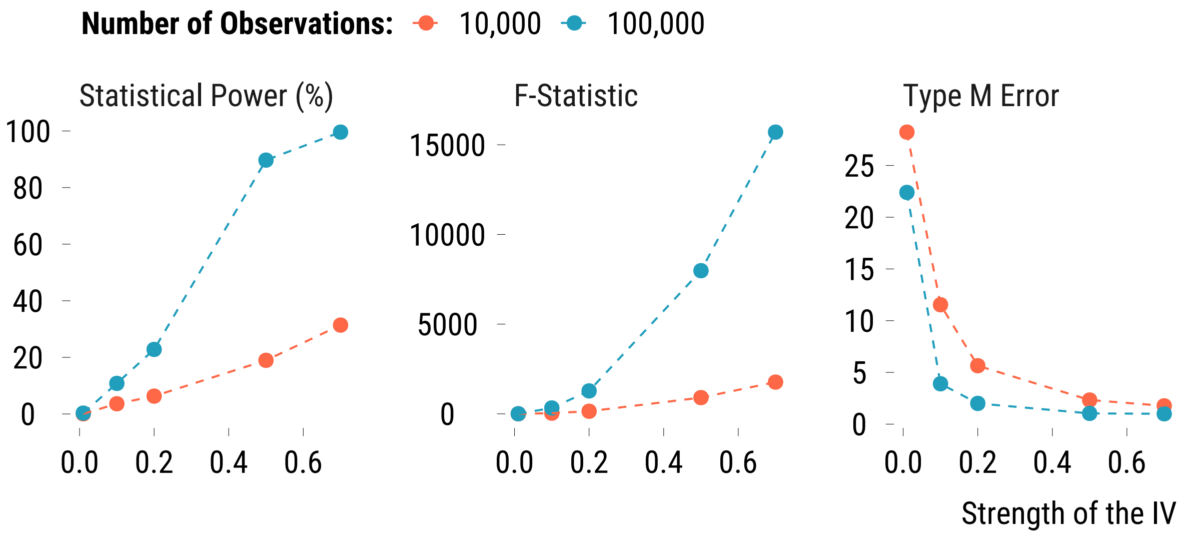


Figure 25: Evolution of Type M Error with the Strength of the Instrumental Variable. *Notes*: The true effect size is a 1% relative increase in the health outcome. The health outcome used in the simulations is the total number of non-accidental deaths. Half of the observations are exposed to exogenous shocks. The strength of the instrumental variable is its effect in standard deviation on the air pollutant concentration.

As shown in Figure 25, we find that statistical power collapses and type M error soars when the instrument's strength decreases. Importantly, this issue arises for rather large instrument's strengths. Even in the case of the large data set with 100,000 observations, an instrumental variable's strength of 0.2, and effect size of a 1% increase in the health outcome, statistical power is only 23% and the average type M error is 2. This statistical power issue arises for a *F*-statistics of 1278! A large *F*-statistic could therefore hide a weak instrumental variable that results in a low statistical power.

Simulating Flagship Studies

The simulation results of the previous section help build the intuition for the parameters influencing the statistical power of studies. Yet, they represent an ideal setting, with relatively large sample size, proportion of treated units, outcome counts and instrumental variable strength. These parameters may not perfectly represent actual studies. For each causal inference method, we therefore consider a realistic set of parameters based on examples from the literature. We then vary the value of key parameters one by one in order to see what could be changed in each study to avoid running into power issues.

Public Transport strikes Public transport strikes are unique but rare exogenous events where air pollution increases. Even in a large data set, with several cities and a long study period, the proportion of treated days might be very small. For instance, Bauernschuster et al. (2017) investigate the effect of public transport strikes on air pollution and emergency admission in the five biggest German cities over a period of 6 years. The sample size of the study is equal to 11,000 observations but there are only 45 1-day strikes. This study could be prone to statistical power issues since the proportion of treated units is 0.4%. We thus try to simulate with our data a similar design. In our baseline simulation, we set as the true effect size the point estimate found by Bauernschuster et al. (2017): days with strikes see an 11% relative increase in the health outcome of interest. The average count of cases for our health outcome—the total number of respiratory deaths–is however 3 times larger than the one in their study, which is equal to 0.69.

In the baseline scenario, we find that the statistical power is only 15% and the average type M error is 2.7. If the researchers had looked at the effect for an health outcome with an average of 23 cases per day, there would however be no statistical power issues. The effect size found by the authors could nonetheless be argued to be a very large increase in an health outcome. If the true effect was only 5% and the average count of the health outcome was 23, there would still be a substantial risk to overestimate statistically significant estimates by a factor of 1.8. Estimating the effects of rare exogenous events on health outcomes with few cases could be therefore difficult.

Air pollution Alerts Air pollution alerts are also rare events. Contrary to public transport strikes or thermal inversions, their effects are estimated using regression discontinuity design. Only observations closed to the air quality threshold are included in the analysis. As a consequence, the effective sample size may end up being particularly small. For instance, in Chen et al. (2018), while the initial sample size is equal to 3652 observations, the effective sample size is only of 143 (100 control observations and 43 treated ones). The proportion of treated observation is 1.2%. With our data, we try to approximate the setting of Chen et al. (2018). In the baseline scenario, we sample one city with a time period of 3652 days and randomly allocate the treatment to 1.2% of observations. We also consider a true effect size of 12%, as found in the study. The average number of cases of their health outcome is 26 cases per day. In our simulations, we use the total number of non-accidental deaths as our outcome variable since the daily mean is equal to 23 deaths.

In the baseline scenario, we find that the statistical power is only 10% and the average type M error is 4.6. If we consider smaller true effect sizes, type M error shoots up and power collapses. As a consequence, care must be taken when interpreting estimates from studies with such a small sample size and few air quality alerts. Of course, one could always argue that air quality alerts have large and protective effects on health outcomes if individuals adopt avoidance behavior.

Instrumenting Air Pollution Finally, we investigate the most common strategies used in the causal inference literature, which are based on instrumental variables. These papers often rely on very large datasets. For instance Schwartz et al. (2018) gathered 591,570 observations (135 cities with a length of study of approximately 4382 days). In this study, air pollution is instrumented with a complex mix of variables and we cannot easily observe the proportion of treated units. The effect size found by the authors is equal to a 1.5% relative increase in an health outcome with an average daily number of cases equal to 23. In our simulations, we therefore assume that half of the observations are exposed to exogenous shocks. We only vary the strength of the instrument and use the total number of non-accidental deaths as the outcome variable. Our data set being smaller than the one used in the study, we only consider 2500 days and 40 cities.

If the instrumental variable increases air pollution concentration by 0.5 standard deviation, we find a statistical power of nearly 100% and an average type M error of 1. Yet, for smaller values of the instrument's strength, statistical power rapidly decreases. For an instrument's strength of 0.2, the statistical power is 48% and the average type M error is 1.4. For a strength of 0.1, power is only 16% and the average type M error is 2.6. In these two scenarios, the values of the F-statistic remain extremely large, with respective values equal to 1287 and 320. A large F-statistic could be a poor indicator of statistical power issues.

Discussion

> *"I think that when we know that we actually do live in uncertainty, then we ought to admit it."*
> — Richard P. Feynman

Our findings should make us worried about statistical power issues when we are trying to estimate the acute health effects of air pollution. Our retrospective analysis of the literature suggests that under-powered studies with inflated effect sizes could be a real issue both in the standard epidemiology and the causal inference literatures. We thus recommend to adopt retrospective calculations since they are very easy to implement and force us to reflect on the range of plausible effect sizes we are trying to estimate.

Unfortunately, a retrospective analysis will not help us understand which parameters of the research design that influence the statistical power of our studies. Our prospective analysis, using simulations based on real-data, fills this gap and leads to issue four warnings.

First, sample size matters for all causal inference methods but especially for the regression-discontinuity design applied to air pollution alerts. Given the sample size its entails, we advise researchers to interpret findings with extra care as the inflation of statistically significant estimates can be extremely large, even when we assume that true effect sizes are large. Second, despite their large sample sizes, when we exploit rare exogenous shocks such as transport strikes, we should be aware that the small proportion of exogenous shocks observed in our studies can lead to a dramatically low statistical power. Third, although it is well-known that two-stage least square estimates are inherently less precise than ordinary least square estimates, it also makes instrumental variable strategies more prone to power issues. If one thinks that omitted variable and attenuation biases are small, the benefits of using an instrumental variable strategy could be questioned. The trade-off between targeting an unbiased estimate with causal inference methods and the risk of running into a type M error could be a fruitful area of research for quantitative bias analysis (Rosenbaum 2010, Dorie et al. 2016, VanderWeele and Ding 2017, Cinelli and Hazlett 2020). Fourth, the power of all research designs in the literature is driven by the average count of the health outcome. Many articles investigate the acute effects of air pollution for specific groups such as children and the elderly. In such settings, there is potentially a huge risk to make a type M error, even with large sample sizes. While they are more involved than a retrospective analysis, simulating the research design we want to implement is the best way to assess if it could suffer from statistical power issues. Our simulation codes in the replication material provide a template to run such prospective analysis.

On top of these specific considerations, we think that the literature would benefit from adopting a different view towards statistically insignificant results (Ziliak and McCloskey 2008, Wasserstein and Lazar 2016, Wasserstein et al. 2019, McShane et al. 2019). The null hypothesis testing framework remains very strong in the field, especially for causal inference papers, since nearly all of them dichotomize evidence using the 5% significance threshold (Greenland 2017). This statistical significance filter leads to publication bias and is at the very heart of the inflation of statistically significant estimates in under-powered studies (Amrhein et al. 2019, Gelman et al. 2020, Romer 2020). Even if we could not improve the statistical power of our studies, the distribution of the acute health effects of air pollution could be more accurate if statistically insignificant results were not kept in the file drawer (Hernán and Robins 2020).

Finally, our results show that a credible identification strategy does not necessarily lead to a correct estimation of the actual true effect (Young 2019). When we qualify estimates as "statistically significant", there is often much more uncertainty lying behind, an uncertainty that should be computed and embraced to better help policymakers evaluate the adverse effects of air pollution. We are convinced that prospective and retrospective power analyses can help us design better studies and improve the interpretation of their results.

Acknowledgements

We are very grateful to Hélène Ollivier and Jeffrey Shrader for their fantastic guidance on this project. Many thanks to Geoffrey Barrows, Tarik Benmarhnia, Marie-Abèle Bind, Sylvain Chabé-Ferret, Marion Leroutier, Quentin Lippmann, Thomas Piketty, as well as seminars participants at the SusDev Colloquium at Columbia, IPWSD, Benmarhnia's Lab, M&A's Lab, and FAERE for their feedbacks.

Bibliography

Allaire, JJ, Rich Iannone, Alison Presmanes Hill, and Yihui Xie (2018) "Distill for R Markdown," `https://rstudio.github.io/distill`.

Altoè, Gianmarco, Giulia Bertoldo, Claudio Zandonella Callegher, Enrico Toffalini, Antonio Calcagnì, Livio Finos, and Massimiliano Pastore (2020) "Enhancing Statistical Inference in Psychological Research via Prospective and Retrospective Design Analysis," *Frontiers in Psychology*, 10, 2893, 10.3389/fpsyg.2019.02893.

Amrhein, Valentin, David Trafimow, and Sander Greenland (2019) "Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication," *The American Statistician*, 73 (sup1), 262–270.

Angrist, Joshua D and Guido W Imbens (1995) "Two-stage least squares estimation of average causal effects in models with variable treatment intensity," *Journal of the American statistical Association*, 90 (430), 431–442.

Arceo, Eva, Rema Hanna, and Paulina Oliva (2016) "Does the effect of pollution on infant mortality differ between developing and developed countries? Evidence from Mexico City," *The Economic Journal*, 126 (591), 257–280.

Barwick, Panle Jia, Shanjun Li, Deyu Rao, and Nahim Bin Zahur (2018) "The Morbidity Cost of Air Pollution: Evidence from Consumer Spending in China,"Technical Report w24688, National Bureau of Economic Research, Cambridge, MA, 10.3386/w24688.

Bauernschuster, Stefan, Timo Hener, and Helmut Rainer (2017) "When labor disputes bring cities to a standstill: The impact of public transit strikes on traffic, accidents, air pollution, and health," *American Economic Journal: Economic Policy*, 9 (1), 1–37.

Bell, Michelle L, Jonathan M Samet, and Francesca Dominici (2004) "Time-series studies of particulate matter," *Annu. Rev. Public Health*, 25, 247–280.

Bhaskaran, Krishnan, Antonio Gasparrini, Shakoor Hajat, Liam Smeeth, and Ben Armstrong (2013) "Time series regression studies in environmental epidemiology," *International journal of epidemiology*, 42 (4), 1187–1195.

Bind, Marie-Abèle (2019) "Causal modeling in environmental health," *Annual review of public health*, 40, 23–43.

Black, Bernard S., Alex Hollingsworth, Leticia Nunes, and Kosali Ilayperuma Simon (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, 10.2139/ssrn.3368187.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes (2020) "Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics," *American Economic Review*, 110 (11), 3634–3660, 10.1257/aer.20190687.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg (2016) "Star wars: The empirics strike back," *American Economic Journal: Applied Economics*, 8 (1), 1–32.

Button, Katherine S., John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafò (2013) "Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience," *Nature Reviews Neuroscience*, 14 (5), 365–376, 10.1038/nrn3475.

Camerer, Colin F., Anna Dreber, Eskil Forsell et al. (2016) "Evaluating Replicability of Laboratory Experiments in Economics," *Science*, 351 (6280), 1433–1436, 10.1126/science.aaf0918.

Chen, Hong, Qiongsi Li, Jay S Kaufman, Jun Wang, Ray Copes, Yushan Su, and Tarik Benmarhnia (2018) "Effect of Air Quality Alerts on Human Health: A Regression Discontinuity Analysis in Toronto, Canada," *The Lancet Planetary Health*, 2 (1), e19–e26, 10.1016/S2542-5196(17)30185-7.

Christensen, Garret, Jeremy Freese, and Edward Miguel (2019) *Transparent and reproducible social science research*: University of California Press.

Cinelli, Carlos and Chad Hazlett (2020) "Making sense of sensitivity: Extending omitted variable bias," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82 (1), 39–67.

Deryugina, Tatyana, Garth Heutel, Nolan H Miller, David Molitor, and Julian Reif (2019) "The mortality and medical costs of air pollution: Evidence from changes in wind direction," *American Economic Review*, 109 (12), 4178–4219.

Di, Qian, Lingzhen Dai, Yun Wang, Antonella Zanobetti, Christine Choirat, Joel D. Schwartz, and Francesca Dominici (2017) "Association of Short-Term Exposure to Air Pollution With Mortality in Older Adults," *JAMA*, 318 (24), 2446, 10.1001/jama.2017.17923.

Dominici, Francesca and Corwin Zigler (2017) "Best practices for gauging evidence of causality in air pollution epidemiology," *American journal of epidemiology*, 186 (12), 1303–1309.

Dorie, Vincent, Masataka Harada, Nicole Bohme Carnegie, and Jennifer Hill (2016) "A flexible, interpretable framework for assessing sensitivity to unmeasured confounding," *Statistics in medicine*, 35 (20), 3453–3470.

Ebenstein, Avraham, Eyal Frank, and Yaniv Reingewertz (2015) "Particulate Matter Concentrations, Sandstorms and Respiratory Hospital Admissions in Israel," 17, 6.

Ferraro, Paul J. and Pallavi Shukla (2020) "Feature—Is a Replicability Crisis on the Horizon for Environmental and Resource Economics?" *Review of Environmental Economics and Policy*, 14 (2), 339–351, 10.1093/reep/reaa011.

Gelman, Andrew and John Carlin (2014) "Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors," *Perspectives on Psychological Science*, 9 (6), 641–651.

Gelman, Andrew, Jennifer Hill, and Aki Vehtari (2020) *Regression and other stories*: Cambridge University Press.

Gelman, Andrew and Francis Tuerlinckx (2000) "Type S Error Rates for Classical and Bayesian Single and Multiple Comparison Procedures," *Computational Statistics*, 15 (3), 373–390, 10.1007/s001800000040.

Giaccherini, Matilde, Joanna Kopinska, and Alessandro Palma (2021) "When particulate matter strikes cities: Social disparities and health costs of air pollution," *Journal of Health Economics*, 78, 102478.

Godzinski, Alexandre, M Suarez Castillo et al. (2019) "Short-term health effects of public transport disruptions: air pollution and viral spread channels,"Technical report, Institut National de la Statistique et des Etudes Economiques.

Greenland, Sander (2017) "Invited commentary: The need for cognitive science in methodology," *American journal of epidemiology*, 186 (6), 639–645.

Griffin, Beth Ann, Megan S. Schuler, Elizabeth A. Stuart et al. (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]*.

Halliday, Timothy J, John Lynham, and Aureo de Paula (2019) “Vog: Using volcanic eruptions to estimate the health costs of particulates,” *The Economic Journal*, 129 (620), 1782–1816.

Hernán, Miguel A and James M Robins (2020) *Causal Inference: What If*, boca raton: chapman & hall/crc edition.

Ioannidis, John P. A. (2008) “Why Most Discovered True Associations Are Inflated,” *Epidemiology*, 19 (5), 640–648.

Ioannidis, John P. A., T. D. Stanley, and Hristos Doucouliagos (2017) “The Power of Bias in Economics Research,” *The Economic Journal*, 127 (605), F236–F265, 10.1111/ecoj.12461.

Isphording, Ingo E and Nico Pestel (2021) “Pandemic meets pollution: poor air quality increases deaths by COVID-19,” *Journal of Environmental Economics and Management*, 108, 102448.

Knittel, Christopher R, Douglas L Miller, and Nicholas J Sanders (2016) “Caution, drivers! Children present: Traffic, pollution, and infant health,” *Review of Economics and Statistics*, 98 (2), 350–366.

Le Tertre, A, S Medina, E Samoli 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.

Liu, Cong, Renjie Chen, Francesco Sera et al. (2019) “Ambient Particulate Air Pollution and Daily Mortality in 652 Cities,” *New England Journal of Medicine*, 381 (8), 705–715, 10.1056/NEJMoa1817364.

Lu, Jiannan, Yixuan Qiu, and Alex Deng (2019) “A note on Type S/M errors in hypothesis testing,” *British Journal of Mathematical and Statistical Psychology*, 72 (1), 1–17.

Mayer, Michael (2019) *missRanger: Fast Imputation of Missing Values*, `https://cran.r-project.org/package=missRanger`, R package version 2.1.0.

McShane, Blakeley B., David Gal, Andrew Gelman, Christian Robert, and Jennifer L. Tackett (2019) “Abandon Statistical Significance,” *The American Statistician*, 73 (sup1), 235–245, 10.1080/00031305.2018.1527253.

Moretti, Enrico and Matthew Neidell (2011) “Pollution, health, and avoidance behavior evidence from the ports of Los Angeles,” *Journal of Human Resources*, 46 (1), 154–175.

Open Science Collaboration (2015) “Estimating the Reproducibility of Psychological Science,” *Science*, 349 (6251), aac4716, 10.1126/science.aac4716.

Orellano, Pablo, Julieta Reynoso, Nancy Quaranta, Ariel Bardach, and Agustin Ciapponi (2020) “Short-term exposure to particulate matter (PM10 and PM2. 5), nitrogen dioxide (NO2), and ozone

(O3) and all-cause and cause-specific mortality: Systematic review and meta-analysis," *Environment international*, 142, 105876.

Peng, Roger D and Francesca Dominici (2008) "Statistical methods for environmental epidemiology with R," *R: a case study in air pollution and health*.

Peng, Roger D, Francesca Dominici, and Thomas A Louis (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.

Romer, David (2020) "In Praise of Confidence Intervals," *AEA Papers and Proceedings*, 110, 55–60, 10.1257/pandp.20201059.

Rosenbaum, Paul R (2010) *Design of observational studies*: Springer.

Rubin, Donald B (1974) "Estimating causal effects of treatments in randomized and nonrandomized studies.," *Journal of Educational Psychology*, 66 (5), 688.

Samet, Jonathan M, Scott L Zeger, Francesca Dominici, Frank Curriero, Ivan Coursac, Douglas W Dockery, Joel Schwartz, and Antonella Zanobetti (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.

Schäfer, Thomas and Marcus A Schwarz (2019) "The meaningfulness of effect sizes in psychological research: Differences between subdisciplines and the impact of potential biases," *Frontiers in Psychology*, 10, 813.

Schell, Terry L., Beth Ann Griffin, and Andrew R. Morral (2018) "Evaluating Methods to Estimate the Effect of State Laws on Firearm Deaths: A Simulation Study,"Technical report, RAND Corporation.

Schlenker, Wolfram and W Reed Walker (2016) "Airports, air pollution, and contemporaneous health," *The Review of Economic Studies*, 83 (2), 768–809.

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

Schwartz, Joel, Elena Austin, Marie-Abele Bind, Antonella Zanobetti, and Petros Koutrakis (2015) "Estimating causal associations of fine particles with daily deaths in Boston," *American journal of epidemiology*, 182 (7), 644–650.

Schwartz, Joel, Marie-Abele Bind, and Petros Koutrakis (2017) "Estimating causal effects of local air pollution on daily deaths: effect of low levels," *Environmental health perspectives*, 125 (1), 23–29.

Schwartz, Joel, Kelvin Fong, and Antonella Zanobetti (2018) "A national multicity analysis of the causal effect of local pollution, NO 2, and PM 2.5 on mortality," *Environmental health perspectives*, 126 (8), 087004.

Shah, Anoop SV, Kuan Ken Lee, David A McAllister et al. (2015) "Short term exposure to air pollution and stroke: systematic review and meta-analysis," *bmj*, 350.

Simonsohn, Uri, Leif D Nelson, and Joseph P Simmons (2014) "P-curve: a key to the file-drawer.," *Journal of experimental psychology: General*, 143 (2), 534.

Smaldino, Paul E and Richard McElreath (2016) "The natural selection of bad science," *Royal Society open science*, 3 (9), 160384.

Stommes, Drew, P. M. Aronow, and Fredrik Sävje (2021) "On the Reliability of Published Findings Using the Regression Discontinuity Design in Political Science," *arXiv:2109.14526 [stat]*.

Timm, Andrew (2019) *retrodesign: Tools for Type S (Sign) and Type M (Magnitude) Errors*, `https://CRAN.R-project.org/package=retrodesign`, R package version 0.1.0.

VanderWeele, Tyler J and Peng Ding (2017) "Sensitivity analysis in observational research: introducing the E-value," *Annals of internal medicine*, 167 (4), 268–274.

Vasishth, Shravan and Andrew Gelman (2019) "How to embrace variation and accept uncertainty in linguistic and psycholinguistic data."

Vichit-Vadakan, Nuntavarn, Nitaya Vajanapoom, and Bart Ostro (2008) "The Public Health and Air Pollution in Asia (PAPA) Project: estimating the mortality effects of particulate matter in Bangkok, Thailand," *Environmental Health Perspectives*, 116 (9), 1179–1182.

Wasserstein, Ronald L. and Nicole A. Lazar (2016) "The ASA Statement on p -Values: Context, Process, and Purpose," *The American Statistician*, 70 (2), 129–133, 10.1080/00031305.2016.1154108.

Wasserstein, Ronald L., Allen L. Schirm, and Nicole A. Lazar (2019) "Moving to a World Beyond " $p < 0.05$"," *The American Statistician*, 73 (sup1), 1–19, 10.1080/00031305.2019.1583913.

Winquist, Andrea, Mitchel Klein, Paige Tolbert, and Stefanie Ebelt Sarnat (2012) "Power estimation using simulations for air pollution time-series studies," *Environmental Health*, 11 (1), 1–12.

Young, Alwyn (2019) "Consistency without inference: Instrumental variables in practical application."

Ziliak, Stephen Thomas and Deirdre N. McCloskey (2008) *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, Economics, Cognition, and Society, Ann Arbor: University of Michigan Press.

Zwet, Erik W. and Eric A. Cator (2021) "The Significance Filter, the Winner's Curse and the Need to Shrink," *Statistica Neerlandica*, 75 (4), 437–452, 10.1111/stan.12241.

Unconfounded but Inflated Causal Estimates

Convincing research designs make empirical economics credible. To avoid confounding, quasi-experimental studies focus on specific sources of variation. This often leads to a reduction in statistical power. Yet, published estimates can overestimate true effects sizes when power is low. Using fake data simulations, we first show that for all causal inference methods, there can be a trade-off between confounding and exaggerating true effect sizes due to a loss in power. We then discuss how reporting statistical power calculations could help address this issue.

AUTHORS:
Léo Zabrocki
PSE - EHESS
leo.zabrocki@psemail.eu

Vincent Bagilet
Columbia University - SIPA
vincent.bagilet@columbia.edu

Introduction

WEBSITE:
https://vincentbagilet.github.io/causal_inflation/

One of the main challenges in empirical economics is to reduce confounding to identify causal effects. Identifications strategies based on Regression Discontinuity (RDD), Instrumental Variable (IV) and Difference-in-Differences (DID) can help achieve this goal. To do so, these strategies only use part of the variation in the data. They exploit the exogenous variation in the treatment or decrease the sample size by only considering observations for which the as-if random assumption is credible. By reducing variation, these methods could however decrease statistical power, that is to say the probability of rejecting the null hypothesis of no effect when it is false. The resulting tension between maintaining enough statistical power and reducing confounding could be problematic for observational studies.

When statistical power is low, statistically significant estimates not only become imprecise but also exaggerate the true effect size of the treatment of interest (Ioannidis 2008, Gelman and Carlin 2014, Lu et al. 2019, Zwet and Cator 2021). This is true even if all the assumptions of a causal inference method are satisfied. Recall that to be statistically significant at the 5% level, estimates must be at least two standard errors away from zero. When statistical power is high, most estimates will be two standard errors away from zero and their distribution will be centered around the true value of the causal estimand. As power decreases, statistically significant estimates start being located in the tails of the distribution of all possible estimates and automatically far away from the true value of the effect.

This counter-intuitive consequence of low statistical power would not be 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). Published estimates from under-powered studies could make-up a very biased sample of the true distributions of causal estimands and 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). Quasi-experimental studies could be more prone to this issue since statistical power is not central to the analysis in current practices. 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. Understanding the determinants of low power is key to avoid the inflation of published estimates.

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 in the treatment. It can lead to a reduction in statistical power. Due to the statistical significance filter, the resulting published estimates could be inflated and thus misleading.

In the first section of this paper, we illustrate the existence and consequences of this trade-off using fake-data simulations based on examples drawn from education, labor, environmental and political economics. We consider separately the main causal inference methods used in the economics literature: selection on observables through matching, RDD, IV and. For each identification strategy we discuss the key factors affecting the confounding / exaggeration trade-off. When assuming that all confounders are measured, matching prunes treated units that cannot be matched to untreated ones. 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. In an IV setting, we only use the treatment variation explained by the instrument. In DID event studies, the variation used to identify an effect sometimes only comes from a limited number of treated observations.

In the second section of the article, we discuss solutions to assess whether statistically significant estimates from observational studies could be inflated. We advocate reporting power calculations. They can be computed before and after the analysis is carried out. By approximating the data generating process, prospective power simulations help identify the design parameters affecting power (Gelman et al. 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 et al. 2021). Our companion website describes in details how such solutions can be implemented.

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 the well-known bias/variance tradeoff (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 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.

Second, recent studies discussing the inflation of statistically significant estimates due to low power focus on specific causal identification methods separately (Schell et al. 2018, Black et al. 2021, Stommes et al. 2021, Young 2021). We show that using causal identification methods may in itself cause power issues. 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 replication failures in empirical economics, despite the widespread use of convincing causal identification methods.

Simulations

To illustrate the trade-off between avoiding confounding and overestimating true effect sizes due to low statistical power, we rely on the conceptual framework developed by Gelman and Carlin (2014) and recently formalized by Lu et al. (2019) and Zwet and Cator (2021). Based on the notation of Zwet and Cator (2021), imagine that we trying to estimate a treatment effect β with a causal inference method. We assume that we have a normally unbiased estimate b of β with a standard error s. When carrying out our research project, we would like first to be able to compare $\mathbb{E}(\frac{|b|}{|\beta|}|\beta, s, |b|/s > 1.96)$, the inflation of statistically significant estimates, with $\mathbb{E}(\frac{|b|}{|\beta|}|\beta, s)$, the inflation of all estimates regardless of their statistical significance. But we would also want to understand how this comparison changes according to our statistical power for rejecting the null hypothesis $H_0 : \beta = 0$, which is given by $\Phi(-1.96 - \frac{\beta}{s}) + 1 - \Phi(1.96 - \frac{\beta}{s})$ where Φ is the cumulative function of the standard normal distribution. Using these three metrics, we could evaluate by how much statistically significant estimates overestimate the true effect size depending on the statistical power. Unfortunately, these two metrics can only be computed if the true effect value of β is known, which is never the case in real world settings without making guesses. Besides, we would also like to be able to compare $\mathbb{E}(\frac{|b|}{|\beta|}|\beta, s, |b|/s > 1.96)$ to $\mathbb{E}(\frac{|b|}{|\beta|}|\beta, s)$ depending on the parameter of causal inference allowign to overcome confounding (e.g. the bandwidth size in RDD). Again, we cannot measure the bias arising due to unobserved confounding in a study[2]. We therefore turn to fake data Monte-Carlo simulations for which we know the true value of the causal estimand of interest.

For clarity, we split the analysis by identification strategy. 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. We build simulations that reproduce real world examples from labor economics for matching, economics of education for RDD, political economy for IV and environmental economics for DID event studies. Real world settings enable to clearly grasp the relationships between the different variables and to set realistic parameter values. Since our simulations have an illustrative purpose only, we intentionally restrict our simulation exercise to settings in which statistical power can be low. All our models are correctly specified and accurately represent the data generating process, except for matching and RDD where a bias arises due to unobserved confounding.

For each identification strategy, we start by laying out the intuition behind the method and how it enables to estimate causal effects. It naturally points to the key parameter through which the confounding/exaggeration trade-off is mediated. We then briefly describe the case-studies considered and our simulation assumptions. We finally display the simulation outputs and discuss the implications of the trade-off that are specific to the identification strategy considered. Very detailed codes for simulation procedures are available on the project's website.

Matching

Intuition for the trade-off. We first focus on the ideal case for which all confounders are assumed to be observed. Under this assumption, one can use matching to estimate the causal effect specific to matched treated units. Contrary to multivariate regression models, this method makes the common support of the data explicit, avoids model extrapolation and non-parametrically adjusts for observed confounders (Ho et al. 2007). In the case of propensity score matching, observed confounders are adjusted for by predicting the probability of units to take the treatment, which is often done with a logistic model where the treatment indicator is regressed on relevant covariates. Treated units are then matched to control units for which their differences in propensity scores are less or equal than the value of a distance metric called the caliper. It is expressed in standard deviation of the propensity score distribution. The smaller the caliper,

---

[2] We are currently working on extending the framework formalized by Lu et al. (2019) and Zwet and Cator (2021) for a causal inference setting by adding the issue of bias arising due to confounding.

the more comparable units are and therefore the lower the risk of confounding is. Yet, with a stringent caliper, some units may not be matched, decreasing the sample size. This could lead to a loss in statistical power and procedure statistically significant estimates that are inflated. In the case of matching, the confounding / exaggeration trade-off is therefore mediated by the value of the caliper.

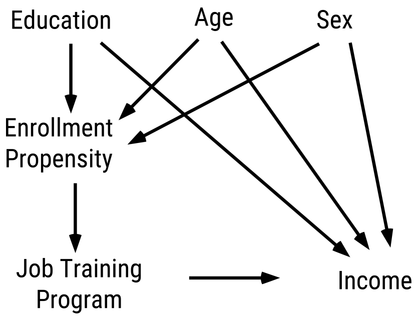


Figure 26: Directed Acyclic Graph for Matching Simulations. *Notes*: The representation of the directed acyclic graph for propensity score matching comes from Huntington-Klein (2021).

Case-study and simulation procedure. We illustrate this issue by simulating a labor training program where the treatment is not randomly allocated (Dehejia and Wahba 1999). As shown in the directed acylic graph in Figure 26, individuals self-select into the training program and may therefore have different characteristics from individuals who do not choose to enroll. We simulate this type of studies by taking a short-cut. Many simulations found in the applied statistics literature test the performance on matching algorithms by first simulating covariates and then simulating the true but unknown probability of units to be treated. Our goal here is different as we do not want to test the performance of various matching algorithms but rather illustrate how a lack of common overlap in propensity scores can result in a loss of statistical power. We therefore first assign a fraction of individual to the treatment and then simulate the true propensity score variable for treated and control units. For treated units, we draw the propensity scores from a normal distribution $N(\mu_T, \sigma_T)$ and for control units, from $N(\mu_C, \sigma_C)$. Once the true propensity scores are created, we define the potential outcomes of each individual. Here, potential outcomes represent the monthly income (in euros) of the individuals if they undertake the training program or not. The potential outcome of each individual i without treatment adoption, $Y_i(0)$, is simulated using the following equation: $Y_i(0) = \text{Wage} \times \text{PS}_i + N(\mu_\text{N}, \sigma_\text{N})$. Wage is the baseline wage, PS_i the propensity score of individual i and some noise is drawn from $N(\mu_\text{N}, \sigma_\text{N})$. This equation makes the potential outcomes Y(0) partly different for treated and control units, creating the required common support issue. We then simulate the potential outcomes $Y_i(1)$ by adding a constant treatment effect of the training program. The constant treatment effect assumption is made to simplify the illustration of the issue we are interested in. In our simulations, when we make the propensity score matching more stringent, not all treated units can be matched to similar control units. The causal estimand would no longer be the average treatment on the treated if the causal effect was not constant across units.

Based on this simulation framework, we generate 1000 datasets for each propensity score matching procedure with caliper values ranging from 0 to 1. Parameters values of the simulation are set to make them realistic and can be found here. Once units are matched, we simply regress the observed revenue on the treatment indicator.

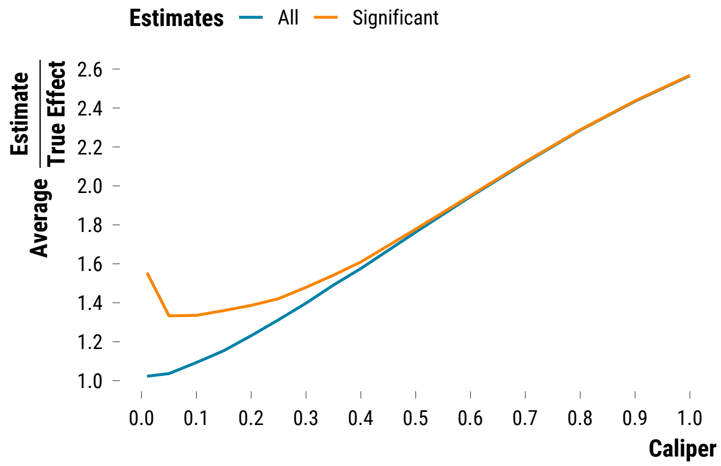


Figure 27: Evolution of Bias with the Caliper in Propensity Score Matching, Conditional on Statistical Significance. *Notes*: The blue line indicates the average bias for all estimates, regardless of their statistical significance. The orange line represents the inflation of statistically significant estimates at the 5% level. The caliper is expressed in standard deviation of the propensity score distribution. Details on the simulation are available at this link.

Results. Figure 27 indicates that the average bias of estimates, regardless of their statistical significance, decreases with the value of the caliper as units become more comparable. As the caliper decreases, statistically significant estimates start being more inflated than the entire sample of estimates. For large caliper values, units are not comparable enough and confounding bias the effect. For small caliper values, the sample size becomes too small to be able to precisely estimate the treatment effect and exaggeration arises. It is well-known that matching procedure can result in imprecise estimates since it does not use information on outcomes but rather focuses on reducing bias arising from covariates imbalance (Rubin 2001). Yet, in a context of publication bias favoring statistically significant estimates, it may make the method produce misleading claims on treatment effect sizes.

Regression Discontinuity Design

Intuition for the trade-off. To identify a causal effect, a regression discontinuity approach relies on the assumption that for values close to the threshold, treatment assignment is quasi-random. Under this assumption, individuals just below and just above the threshold would be comparable in terms of observed and unobserved covariates, and only differ in their treatment status. To avoid confounding, the RDD focuses on observations within a certain bandwidth around the threshold and discards observations further away. The effective sample size where the identification of causal effect of the treatment is the most credible differs from the total sample size. For this method, the confounding / exaggeration trade-off is therefore mediated by the size of the bandwidth.

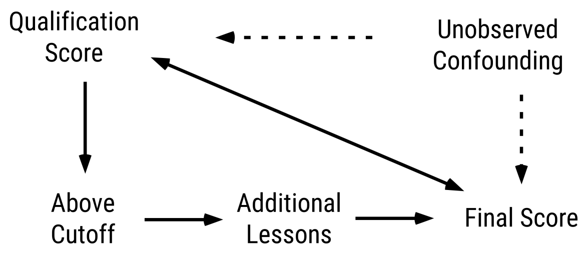


Figure 28: Directed Acyclic Graph for Regression Discontinuity Design Simulations. *Notes*: The representation of the directed acyclic graph for the RDD comes from Huntington-Klein (2021).

Case-study and simulation procedure. To illustrate this trade-off, we consider a standard application of the sharp RD design in economics of education in which students are offered additional lessons based on the score they obtained on a standardized test (Thistlethwaite and Campbell 1960). As shown in the directed acyclic graph of Figure 28, students with initial test scores below a given threshold follow additional lessons while those above do not. Since students far above and far below the threshold may differ along unobserved characteristics such as ability, a RDD estimates the effect of the treatment on final test scores by comparing outcomes of students whose initial test scores are just below and just above this threshold.

Our simulation framework for RDD is as follows. We assume that if a student i has an initial scores $Qual_i$ below a cutoff C, she must take additional lessons. The allocation of the treatment T_i is sharp: $T_i = \mathbb{I}[Qual_i < C]$. The final scores of students $Final_i$ are correlated with their qualification score $Qual_i$. We further assume that both qualification and final test scores are affected by students' unobserved ability U_i in a non-linear (cubic) way. A large ability has a strong positive impact on test scores. Similarly a particularly low ability strongly impacts test scores negatively. An average ability does not have much impact on test scores. Such a functional form seems realistic. The final test scores $Final_i$ are thus defined as follows: $Final_i = \alpha^f + \beta T_i + \gamma Qual_i + \delta^f f(U_i) + \epsilon_i$, where α^f is a constant, f a non linear function and ϵ_i random noise drawn from $\mathcal{N}(0, \sigma_e)$ noise. The causal parameter of interest is β.

To make our simulations realistic, we set the parameters of our simulations based on our reading of Kraft (2020). They can be found here. Given these parameters values, we then generate 1000 datasets with 10,000 observations. For each dataset, we finally estimate the treatment effect by regressing the final score on the treatment status and the qualifying score for different bandwidth sizes.

Results. Figure 28 shows that, as the bandwidth around the threshold decreases, the bias of all estimates decreases. This is due to the fact that for large bandwidths, omitted variable arises, while for small bandwidths, treated and control units are very similar in terms of observed and unobserved covariates. However, we observe a U-

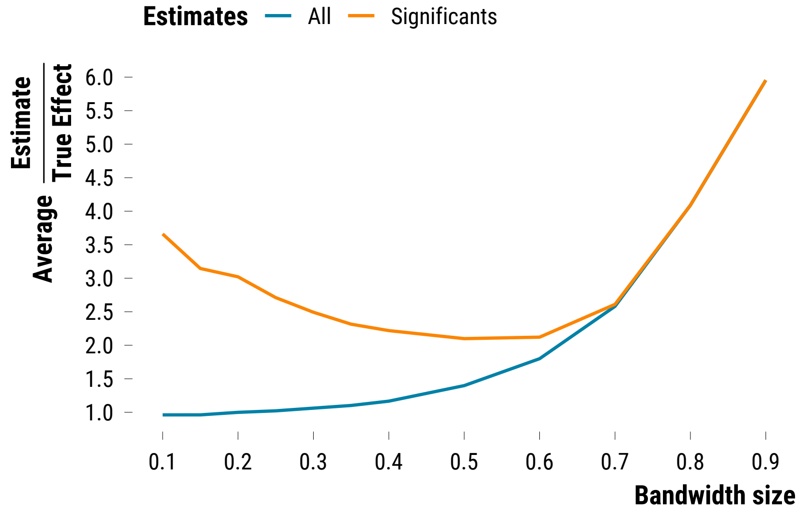


Figure 29: Evolution of Bias with Bandwidth Size in Regression Discontinuity Design, Conditional on Statistical Significance. *Notes*: The blue line indicates the average bias for all estimates, regardless of their statistical significance. The orange line represents the inflation of statistically significant estimates at the 5% level. In this simulation, N = 10,000. The bandwidth size is expressed as the proportion of the total number of observations of the entire sample. Details on the simulation are available at this link.

shape relationship for statistically significant estimates as the bandwidth decreases. For small values of the bandwidth, the statistical power shrinks and statistically significant estimates become inflated. The optimal bandwidth literature describes a similar trade-off but from a different perspective (Imbens and Kalyanaraman 2012). They consider a bias/precision trade-off while we consider a omitted variable bias / exaggeration bias trade-off due to publication bias favoring statistical significance.

Instrumental Variable Strategy

Intuition for the trade-off. Instrumental variable overcomes the issue of unobserved confounding by only considering the exogenous variation in the treatment. When this exogenous fraction of the variation is limited, the instrument can still successfully eliminate confounding on average. However, the IV estimator will be imprecise and statistical power low. In the case of the IV, the confounding / exaggeration trade-off is therefore mediated by the strength of the instrument considered. The weaker the instrument, the more inflated statistically significant estimates will be.

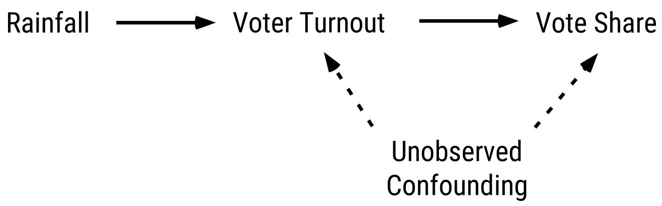


Figure 30: Directed Acyclic Graph for Instrument Variable Simulations.

Case-study and simulation procedure. To illustrate this trade-off, we take as an example the studies estimating the impact of voter turnout on election results (Gomez et al. 2007, Fujiwara et al. 2016, Cooperman 2017). As shown in Fig. 30, to avoid the threat of confounding, researchers have taken advantage of exogenous factors that affect voter turnout such as rainfall. In our simulations, we assume that the vote share of a party in location i, $Share_i$, can defined such that: $Share_i = \alpha + \beta Turnout_i + \delta u_i + e_i^{(S)}$, where α is a constant, u represents an unobserved variable and $e^{(S)} \sim \mathcal{N}(0, \sigma_{e_S})$ noise. The causal parameter of interest is β. Turnout observations $Turnout_i$ are given by the following model: $Turnout_i = \gamma + \lambda Rain_i + \eta u_i + e_i^{(T)}$, where $Rain_i$ is either a continuous variable (amount of rain in location i on the day of the election) or a dummy variable (whether it rained or not) and $e^{(T)}$ is random noise drawn form $\mathcal{N}(0, \sigma_{e_T})$. We refer to λ as the strength of the instrumental variable.

We discuss in great details how we choose realistic parameters for these two models here. For each value of the IV strength considered, we create 1000 datasets. We run both a naive ordinary least squares model and a two-stage least squares model to estimate the impact of voter turnout on the vote share of a party.

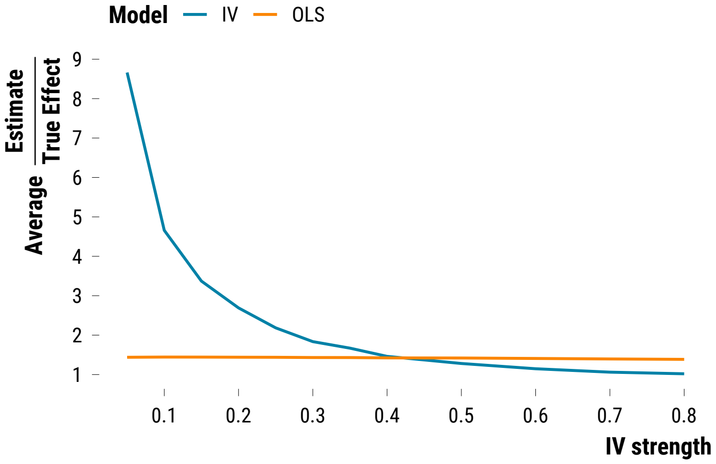


Figure 31: Evolution of Bias for Statistically Significant Estimates Against Intensity of the Instrument in Instrumental Variable Design. *Notes*: The blue line indicates the average bias for statistically significant IV estimates at the 5%. The orange line represents the bias of statistically significant OLS estimates at the 5% level. The strength of the instrumental variable is expressed as the value of the linear parameter linking rainfall to turnout. In this simulation, N = 10,000. Details on the simulation are available at this link.

Results. Figure 31 displays, for different IV strengths, the average of statistically significant estimates scaled by the true effect size for both the IV and the naive regression model. When the instrument is strong, the IV will recover the true effect, contrarily to the the naive regression model. Yet, when the IV strength decreases, the exaggeration of statistical significant estimates skyrockets. Even if the intensity of the omitted variable bias is large, for limited IV strengths, the exaggeration ratio can become larger than the omitted variable bias. When the only available instrument is weak, using the naive regression model would, on average, produce statistically significant estimates that are closer to the true effect size than the IV. Of interest for applied research, a large F-statistic does not necessarily attenuate this problem. For the parameter values considered here, this phenomenon arises even in cases for which the F-statistic is substantially larger than the usually recommended threshold of 10, as illustrated in our supplementary materials.

Difference-in-Differences Event Study Design

Intuition for the trade-off. To avoid confounding, DID event studies take advantage of situations for which we can adjust how the outcome evolved before and after an intervention in a treated group by using the same comparison for an untreated group. The credibility of the method rests on the assumption that the outcomes of the two groups should have evolved similarly had the intervention not occurred. In some cases, while the number of observations may be large, the proportion of units affected by the intervention might be limited. As a consequence, the number of treated observations is small and the variation available to identify the treatment is limited. In studies using discrete exogenous shocks, a confounding / exaggeration trade-off is thus mediated by the number of observations treated. It does not only concern DID event studies but is particularly salient in this case.

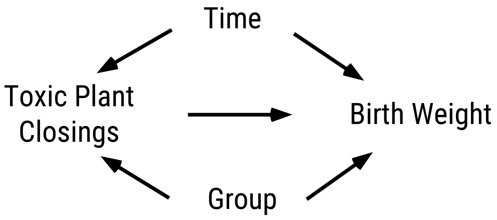


Figure 32: Directed Acyclic Graph for Difference in Differences Simulations. *Notes*: The representation of the directed acyclic graph for the RDD comes from (Huntington-Klein 2021).

Case-study and simulation procedure. To illustrate this trade-off, we simulate a study on the impact of air pollution reduction on newborn weight of babies. To avoid confounding, one can exploit exogenous shocks to air pollution such as plant closures, creation of a low emission zone or of an urban toll. We simulate our analysis at the zip code and monthly levels and focus on the example of toxic plant closures (Currie et al. 2015). We consider that the average birth weight in zip code z at time period t, bw_{zt}, depends on a zip code fixed effect ζ_z, a time fixed effect τ_t, and the treatment status $T_{z,t}$, which is equal to one if a plant closed in this period and 0 otherwise. The average birth weights bw_{zt} is defined as follows: $bw_{z,t} = \alpha + \beta T_{z,t} + \zeta_z + \tau_t + \epsilon_{z,t}$. To simplify further the simulations, we assume that the treatment allocation is not staggered and its effect is constant in time and homogeneous across zip codes. We vary in the simulations the proportion of zip codes affected by toxic plant closings.

The parameters values of our simulations are inspired by Currie et al. (2015) and can be found here. For a fixed sample size of 120,000 observations, we generate 1000 datasets for an increasing number of treated observations and run our two-way fixed effects model.

Results. Even though the actual sample size is extremely large in our example, if the number of treated observations is small, the ex-

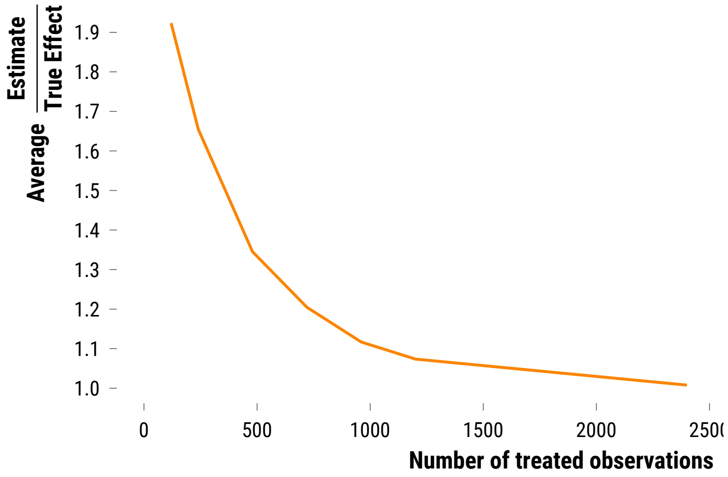


Figure 33: Evolution of Bias for Statistically Significant Estimates against the Number of Treated Observations in Difference-in-Differences Event Study Design. *Notes*: The blue line indicates the average bias for statistically significant estimates at the 5%. In this simulation, N = 120,000. Details on the simulation are available at this link.

aggeration can be important, as shown in Figure 33. A very large number of observations does not necessarily prevent the exaggeration issue to arise. The intuition behind this result can be compared to the case of a complete randomized controlled trial where, for fixed sample size, the statistical power is maximized when units are equally allocated to treatment and control groups.

Practical Recommendations

Once we realize that causal methods can produce inflated estimates when the study is under-powered, how can we address this problem? Even though it does not produce uninflated estimates, reporting power calculations enables to evaluate the risk of exaggeration for a study. In this section, we present a workflow to evaluate and report the power of a study before and after its implementation. We then discuss how changing our attitude towards statistical significance and replicating studies can help limit this issue.

Before Analyzing the Data

In randomized controlled trials, presenting statistical power calculations before running the experiment is not only an established practice but also a requirement (Duflo et al. 2007, McConnell and Vera-Hernandez 2015, Athey and Imbens 2016). Power is however rarely reported in observational studies, despite the availability of specific power formulas for some causal inference methods (Freeman et al. 2013, Cattaneo et al. 2019). Two main reasons could explain this limited reporting of power calculations. First, we do not directly control the data collection process in our research projects. Second, we may fear that available power formulas are not flexible enough to capture the complexity of their design. On top of these reasons, there is a lack of guidance on how to design well-powered observational studies. In causal inference textbooks, very few pages are devoted to the topic (Angrist and Pischke 2008; 2014, Imbens and Rubin 2015, Cunningham 2021). To the best of our knowledge, only two textbooks discuss the matter in depth (Shadish et al. 2002, Huntington-Klein 2021).

Simulating the design of an observational study is a solution to overcome these limits (Hill 2011, Gelman et al. 2020, Black et al. 2021). Similarly to what we did in the previous section, the goal of this approach is to simulate the data generating process of the study from scratch. It requires thinking about the distribution of the variables and their relationships. External information found in previous studies can help guide the simulation process to make it more realistic. If the relationships among covariates are too complex to emulate, a second approach starts from an existing dataset to which a simulated treatment and potential outcomes are added[3].

When simulations indicate that statistical power is low, additional data could be collected or the statistical model could be expanded to increase precision. In any case, it should not stop from carrying out a research project. Simulation results rest on the way the data generation process was modeled and it can be difficult to gauge the amount of noise present in data before actually analyzing them. The two actual benefits of a prospective simulation procedure are to think about factors that affect power and not to be mislead by statistically significant estimates if power is low.

Once the Main Analysis is Completed

Once we have obtained a statistically significant estimate for the treatment of interest, we still need to think about the statistical power of the study to check whether the magnitude of our estimate is trustworthy. A *retrospective* power analysis helps evaluate whether the design of the study would produce uninflated statistically significant estimates if the true effect was smaller than the observed estimate (Gelman and Carlin 2014, Ioannidis et al. 2017, Stommes et al. 2021).

We illustrate how a retrospective analysis works by taking the example of Card (1993) on the relationship between human capital and income. He finds that an additional year of education, instrumented by the distance of growing near a four-year college, causes a 13.2% average increase in wage. The associated standard error is 5.5%. As noted by the author himself, the estimate is very imprecise: if the true causal effect was slightly smaller than the observed estimate, the study would very likely be under-powered. For instance, imagine that prior evidence suggests that the true effect could be to closer a 10% increase in wage. Computing the statistical power of the study only requires to draw many estimates from a normal distribution centered around the hypothesized true effect of 10% and with a standard deviation equal to the 5.5% standard error obtained in Card (1993). Concretely, one proceeds as if they were able to repli-

---

[3] In the future version of the paper, we will explain how we can easily simulate from scratch the study by Card (1993). For now, examples of simple simulations are available on our companion website.

cate the study many times under the assumption that the true effect is different from the observed estimate. The proportion of sampled estimates that are statistically significant at the 5% level, 44% in this case, is the statistical power. The inflation of significant estimates is then computed as the average ratio of the values of statistically significant estimates over the assumed true effect size: these estimates would be 1.5 times too large on average.

For a retrospective power analysis to be useful, it is therefore necessary to make informed guesses about the range of plausible effect sizes. Such guesses can be based on results from meta-analyses or previous studies with a convincing design (e.g., a large randomized controlled trial). When such information is not available, power calculations can be run for a range of smaller but credible effect sizes[4].

Results from power and exaggeration calculations would not only be highly informative but could also be reported very concisely in the robustness section of articles. `R` and `Stata` packages have been developed (Timm 2019, Linden 2019) to easily implement retrospective power analyses.

Attitude Towards Statistical Significance and Replication

General changes in scientific practices could also limit the inflation of statistically significant estimates in under-powered studies. As shown in our simulations, if estimates were not filtered by their statistical significance, even under-powered studies would on average recover the true effect. The publication bias arising from dichotomizing evidence according to p-values has long been criticized in many disciplines but has seen a revival with the recent replication crises in psychology, medicine and social sciences. Many researchers advocate abandoning statistical significance as a measure of a study's quality (McShane et al. 2019). This would essentially eliminate the trade-off described in this paper.

To be effective, this change in attitude towards statistical significance should be paired with an effort to replicate studies (Christensen and Miguel 2018). Replications, even of low powered studies, would eventually enable to build the distribution of the causal estimand of interest. Meta-analyzes would then reduce the uncertainty around the true value of the causal estimand by pooling estimates (Hernán 2021).

Finally, the inflation of statistically significant estimates can be limited by considering confidence intervals as compatibility intervals (Shadish et al. 2002, Amrhein et al. 2019, Romer 2020). The width of these intervals gives a range of effect sizes compatible with the data. Confidence intervals will be wide in under-powered studies signaling that point estimates should not be taken at face value, even if statistically significant.

---

[4] We are currently working on the adoption of two existing approaches that could help address the potential inflation of statistically significant estimates. The first approach is based on the work by Zwet and Gelman (2021), who propose to use a Bayesian procedure to shrink statistically significant estimates based on a corpus of estimates from prior studies. The second approach consists in carrying out quantitative bias analyses to evaluate whether the threat of unobserved confounding requires a restrictive causal approach. Rosenbaum (2002), Oster (2019) and Cinelli and Hazlett (2020) have developed different methods to run sensitivity analyses. They could be paired with power calculations to better evaluate the hidden bias / power tradeoff of competing research designs. For instance, if a sensitivity analysis reveals that a simple selection on observables strategy is very robust to omitted variable bias, one may avoid using an IV model since it has a higher chance to produce inflated estimates.

Conclusion

Causal identification strategies have undoubtedly participated in making empirical analyses more credible (Angrist and Pischke 2010). To avoid confounding, they only exploit the exogenous part of treatment variation. In this paper, we argue that the same aspect that makes causal identification strategies credible can create another type of bias. Not only the lack of precision makes it more difficult to precisely get a sense of the magnitude of the actual effect but it also increases the probability of published estimates to be inflated. The confounding / exaggeration trade-off we highlight in this paper manifests itself along different dimensions for each identification strategy. A systematic reporting of statistical power calculations in observational studies could help gauge the risk of falling into this low power trap.

Bibliography

Abadie, Alberto (2020) "Statistical Nonsignificance in Empirical Economics," *American Economic Review: Insights*, 2 (2), 193–208, 10.1257/aeri.20190252.

Amrhein, Valentin, David Trafimow, and Sander Greenland (2019) "Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication," *The American Statistician*, 73 (sup1), 262–270.

Andrews, Isaiah and Maximilian Kasy (2019) "Identification of and Correction for Publication Bias," *American Economic Review*, 109 (8), 2766–2794, 10.1257/aer.20180310.

Angrist, Joshua D and Jörn-Steffen Pischke (2008) *Mostly harmless econometrics*: Princeton university press.

Angrist, Joshua D. and Jörn-Steffen Pischke (2010) "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics," *Journal of Economic Perspectives*, 24 (2), 3–30, 10.1257/jep.24.2.3.

——— (2014) *Mastering 'Metrics: The Path from Cause to Effect*: Princeton University Press.

Athey, Susan and Guido Imbens (2016) "The Econometrics of Randomized Experiments," *arXiv:1607.00698 [econ, stat]*.

Black, Bernard S., Alex Hollingsworth, Leticia Nunes, and Kosali Ilayperuma Simon (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, 10.2139/ssrn.3368187.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes (2020) "Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics," *American Economic Review*, 110 (11), 3634–3660, 10.1257/aer.20190687.

Button, Katherine S., John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafò (2013) "Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience," *Nature Reviews Neuroscience*, 14 (5), 365–376, 10.1038/nrn3475.

Camerer, Colin F., Anna Dreber, Eskil Forsell et al. (2016) "Evaluating Replicability of Laboratory Experiments in Economics," *Science*, 351 (6280), 1433–1436, 10.1126/science.aaf0918.

Card, David (1993) "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," Working Paper 4483, National Bureau of Economic Research, 10.3386/w4483.

Cattaneo, Matias D., Rocío Titiunik, and Gonzalo Vazquez-Bare (2019) "Power Calculations for Regression-Discontinuity Designs," *The Stata Journal: Promoting communications on statistics and Stata*, 19 (1), 210–245, 10.1177/1536867X19830919.

Chang, Andrew C. and Phillip Li (2022) "Is Economics Research Replicable? Sixty Published Papers From Thirteen Journals Say "Often Not"," *Critical Finance Review*, 11, 10.1561/104.00000053.

Christensen, Garret and Edward Miguel (2018) "Transparency, Reproducibility, and the Credibility of Economics Research," *Journal of Economic Literature*, 56 (3), 920–980, 10.1257/jel.20171350.

Cinelli, Carlos and Chad Hazlett (2020) "Making sense of sensitivity: Extending omitted variable bias," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82 (1), 39–67.

Cooperman, Alicia Dailey (2017) "Randomization inference with rainfall data: Using historical weather patterns for variance estimation," *Political Analysis*, 25 (3), 277–288.

Cunningham, Scott (2021) *Causal Inference: The Mixtape*: Yale University Press, 10.2307/j.ctv1c29t27.

Currie, Janet, Lucas Davis, Michael Greenstone, and Reed Walker (2015) "Environmental health risks and housing values: evidence from 1,600 toxic plant openings and closings," *American Economic Review*, 105 (2), 678–709.

Deaton, Angus and Nancy Cartwright (2018) "Understanding and Misunderstanding Randomized Controlled Trials," *Social Science & Medicine*, 210, 2–21, 10.1016/j.socscimed.2017.12.005.

Dehejia, Rajeev H. and Sadek Wahba (1999) "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94 (448), 1053–1062, 10.2307/2669919.

Duflo, Esther, Rachel Glennerster, and Michael Kremer (2007) "Using Randomization in Development Economics Research: A Toolkit," in Schultz, T. Paul and John A. Strauss eds. *Handbook of Development Economics*, 4, 3895–3962: Elsevier, 10.1016/S1573-4471(07)04061-2.

Freeman, G., B. J. Cowling, and C. M. Schooling (2013) "Power and Sample Size Calculations for Mendelian Randomization Studies Using One Genetic Instrument," *International Journal of Epidemiology*, 42 (4), 1157–1163, 10.1093/ije/dyt110.

Fujiwara, Thomas, Kyle Meng, and Tom Vogl (2016) "Habit formation in voting: Evidence from rainy elections," *American Economic Journal: Applied Economics*, 8 (4), 160–88.

Gelman, Andrew and John Carlin (2014) "Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors," *Perspectives on Psychological Science*, 9 (6), 641–651.

Gelman, Andrew, Jennifer Hill, and Aki Vehtari (2020) *Regression and other stories*: Cambridge University Press.

Gomez, Brad T, Thomas G Hansford, and George A Krause (2007) "The Republicans should pray for rain: Weather, turnout, and voting in US presidential elections," *The Journal of Politics*, 69 (3), 649–663.

Hernán, Miguel A. (2021) "Causal Analyses of Existing Databases: No Power Calculations Required," *Journal of Clinical Epidemiology*, S0895435621002730, 10.1016/j.jclinepi.2021.08.028.

Hernán, Miguel A and James M Robins (2020) *Causal Inference: What If*, boca raton: chapman & hall/crc edition.

Hill, Jennifer L (2011) "Bayesian nonparametric modeling for causal inference," *Journal of Computational and Graphical Statistics*, 20 (1), 217–240.

Ho, Daniel E, Kosuke Imai, Gary King, and Elizabeth A Stuart (2007) "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference," *Political analysis*, 15 (3), 199–236.

Huntington-Klein, Nick (2021) *The Effect: An Introduction to Research Design and Causality*, Boca Raton: Chapman and Hall/CRC, 1st edition, 10.1201/9781003226055.

Imbens, Guido and Karthik Kalyanaraman (2012) "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," *The Review of Economic Studies*, 79 (3), 933–959.

Imbens, Guido W and Donald B Rubin (2015) *Causal inference in statistics, social, and biomedical sciences*: Cambridge University Press.

Ioannidis, John P. A. (2008) "Why Most Discovered True Associations Are Inflated," *Epidemiology*, 19 (5), 640–648.

Ioannidis, John P. A., T. D. Stanley, and Hristos Doucouliagos (2017) "The Power of Bias in Economics Research," *The Economic Journal*, 127 (605), F236–F265, 10.1111/ecoj.12461.

Kasy, Maximilian (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, 10.1257/jep.35.3.175.

Kraft, Matthew A (2020) "Interpreting effect sizes of education interventions," *Educational Researcher*, 49 (4), 241–253.

Linden, Ariel (2019) "RETRODESIGN: Stata Module to Compute Type-S (Sign) and Type-M (Magnitude) Errors," Boston College Department of Economics, October.

Lu, Jiannan, Yixuan Qiu, and Alex Deng (2019) "A note on Type S/M errors in hypothesis testing," *British Journal of Mathematical and Statistical Psychology*, 72 (1), 1–17.

McConnell, Brendon and Marcos Vera-Hernandez (2015) "Going beyond Simple Sample Size Calculations: A Practitioner's Guide,"Technical report, Institute for Fiscal Studies, 10.1920/wp.ifs.2015.1517.

McShane, Blakeley B., David Gal, Andrew Gelman, Christian Robert, and Jennifer L. Tackett (2019) "Abandon Statistical Significance," *The American Statistician*, 73 (sup1), 235–245, 10.1080/00031305.2018.1527253.

Open Science Collaboration (2015) "Estimating the Reproducibility of Psychological Science," *Science*, 349 (6251), aac4716, 10.1126/science.aac4716.

Oster, Emily (2019) "Unobservable Selection and Coefficient Stability: Theory and Evidence," *Journal of Business & Economic Statistics*, 37 (2), 187–204, 10.1080/07350015.2016.1227711.

Ravallion, Martin (2020) "Should the Randomistas (Continue to) Rule?" Working Paper 27554, National Bureau of Economic Research, 10.3386/w27554.

Romer, David (2020) "In Praise of Confidence Intervals," *AEA Papers and Proceedings*, 110, 55–60, 10.1257/pandp.20201059.

Rosenbaum, Paul R. (2002) *Observational Studies*, Springer Series in Statistics, New York, NY: Springer New York, 10.1007/978-1-4757-3692-2.

Rosenthal, Robert (1979) "The File Drawer Problem and Tolerance for Null Results," *Psychological Bulletin*, 86 (3), 638–641, 10.1037/0033-2909.86.3.638.

Rubin, Donald B (2001) "Using propensity scores to help design observational studies: application to the tobacco litigation," *Health Services and Outcomes Research Methodology*, 2 (3), 169–188.

Schell, Terry L., Beth Ann Griffin, and Andrew R. Morral (2018) "Evaluating Methods to Estimate the Effect of State Laws on Firearm Deaths: A Simulation Study,"Technical report, RAND Corporation.

Shadish, William R., Thomas D. Cook, and Donald Thomas Campbell (2002) *Experimental and Quasi-experimental Designs for Generalized Causal Inference*: Houghton Mifflin.

Stommes, Drew, P. M. Aronow, and Fredrik Sävje (2021) "On the Reliability of Published Findings Using the Regression Discontinuity Design in Political Science," *arXiv:2109.14526 [stat]*.

Thistlethwaite, Donald L. and Donald T. Campbell (1960) "Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment," *Journal of Educational Psychology*, 51 (6), 309–317, 10.1037/h0044319.

Timm, Andrew (2019) *retrodesign: Tools for Type S (Sign) and Type M (Magnitude) Errors*, `https://CRAN.R-project.org/package=retrodesign`, R package version 0.1.0.

Young, Alwyn (2021) "Leverage, Heteroskedasticity and Instrumental Variables in Practical Application," 43.

Zwet, Erik and Andrew Gelman (2021) "A Proposal for Informative Default Priors Scaled by the Standard Error of Estimates," *The American Statistician*, 1–9, 10.1080/00031305.2021.1938225.

Zwet, Erik W. and Eric A. Cator (2021) "The Significance Filter, the Winner's Curse and the Need to Shrink," *Statistica Neerlandica*, 75 (4), 437–452, 10.1111/stan.12241.

Conclusion

The reading of the four chapters of this dissertation could rightly leave the impression that improving the design of studies on the acute health effects of air pollution is not easy. This is correct but some progress has been made and new research questions have been raised.

Lessons learnt from Matching, Sensitivity Analysis and Randomization-inference

The two initial chapters show that the common support in air pollution studies based on wind patterns and changes in maritime traffic can be relatively small. Adjusting nonparametrically for observed covariates with the constrained pair matching algorithm developed in Sommer et al. (2021) is difficult. Should we therefore absolutely focus on removing bias at the cost of obtaining very imprecise estimates? Future research could focus on testing different matching algorithms like cardinality matching (Visconti and Zubizarreta 2018) or near-far matching (Baiocchi et al. 2010; 2012), but also alternative methods such bayesian additive regressions trees (Hill 2011, Hill and Su 2013). Based on recent works in statistics (Bojinov and Shephard 2019, Menchetti et al. 2021), it is also necessary to better define the STUVA in this type of time series studies and understand how this assumption could be made more credible. It remains that matching and the visual inspections that go with it are a principled approach to check the balance of observed covariates. Future research is however needed to strengthen the design of more continuous instrumental variables with the relevant matching methods (Lopez and Gutman 2017, Fong et al. 2018, Bennett et al. 2020, Forastiere et al. 2020).

These two chapters were also the occasion to implement underused techniques in the literature. First, the sensitivity analysis framework developed by Rosenbaum (1987) and extended to average causal effects by Fogarty (2020) effectively complements informal arguments on the issue of unmeasured confounding. A comprehensive investigation of the sensitivity of the literature to unmeasured confounding could bring interesting results. The sensitivity analysis method recently developed by Cinelli and Hazlett (2020a;b) in a multivariate regression framework should make this task more easy to carry out. Second, it is often not easy to quantify the source of uncertainty of estimates in observational studies. Once a balanced sample is obtained, it is relatively intuitive to analyse the data as an hypothetical experiment. Fisherian inference and Neymaniam inference are two elegant modes of inference to understand how the range of effect sizes consistent with the data is computed (Rubin 1991, Rosenbaum 2010, Imbens and Rubin 2015). If randomization inference avoids large-sample approximation and is distribution free, it relies on the very unrealistic assumption that unit-level causal effects are constant. Yet, in the context of the first two chapters, Neyman's intervals were similar to those found with randomization-based inference procedure. Besides, the recent studentized test statistic proposed by Wu and Ding (2021) for making randomization inference conservative for weak nulls also lead to similar intervals. In future research, randomization inference procedures could be useful to better quantify the uncertainty due to the spatial auto-correlation of observations since conventional standard errors have been found to underestimate the sampling variability (Barrios et al. 2012, Cooperman 2017, Kelly 2021). Compared to the early literature based on few city-level data, this issue is now more important because many studies rely on zipcode or county levels data where the air pollution exposure is correlated across spatial units. Randomization inference procedure could be also used an additional robustness check when instrumental variables are feared to be weak (Imbens and Rosenbaum 2005).

Better Designs and Attitudes for Under-Powered Studies

The last two chapters of the dissertation take seriously the issue of working with under-powered studies when there is a publication bias favoring statistical significance. Both in the standard epidemiology literature and the causal inference literature, a large fraction of studies are arguably under-powered and could produce statistically significant estimates that are too large. Two easy solutions exist for practitioners to not be misled when this is issue could arise. The simplest solution is to avoid focusing on statistical significance (McShane et al. 2019, Amrhein et al. 2019) and to start interpreting 95% confidence intervals. They indicate the full range of effect sizes consistent with the data: if a study has a low power, the interval will be large. It brings more scientific information to fully display the uncertainty of the results than putting all trust in a single point estimate that could be very inflated. The second solution is to carry out a retrospective power analysis (Gelman and Carlin 2014, Gelman et al. 2020, Stommes et al. 2021). This type of analysis is very easy to run and can help evaluate the robustness of a study to low power issues. There exists a third solution but it is more difficult to implement. Running a prospective power analysis is a very powerful way to evaluate, in advance, the parameters of a research design that are likely to affect the power. The third chapter of the thesis show several parameters (e.g. the number of exogenous shocks or the average daily count of cases of an health outcome) that are important for observational studies on the acute health effects of air pollution. Yet, researchers might fear that it is too cumbersome and time consuming to set up an entire simulation procedure. The codes available on the website of the project should decrease this cost but, to be really useful to the research community, a dedicated R package may be required.

The last and fourth chapter tries to generalize some of the findings from all the chapter of the dissertation. For each causal inference method, there is a trade-off between maintaining statistical power to avoid producing statistically significant estimates that are inflated and reducing the confounding bias. This trade-off is not only valid for studies on the acute health effects of air pollution but all research questions. Showing that this trade-off exists with simulations is arguably not sufficient to better take it into account: its statistical formalization should be derived and strategies to evaluate when its pernicious consequences could happen should be developed. This thesis ends therefore on much needed but exciting methodological questions.

Bibliography

Amrhein, Valentin, David Trafimow, and Sander Greenland (2019) “Inferential statistics as descriptive statistics: There is no replication crisis if we don’t expect replication,” *The American Statistician*, 73 (sup1), 262–270.

Baiocchi, Mike, Dylan S Small, Scott Lorch, and Paul R Rosenbaum (2010) “Building a stronger instrument in an observational study of perinatal care for premature infants,” *Journal of the American Statistical Association*, 105 (492), 1285–1296.

Baiocchi, Mike, Dylan S Small, Lin Yang, Daniel Polsky, and Peter W Groeneveld (2012) “Near/far matching: a study design approach to instrumental variables,” *Health Services and Outcomes Research Methodology*, 12 (4), 237–253.

Barrios, Thomas, Rebecca Diamond, Guido W Imbens, and Michal Kolesár (2012) “Clustering, spatial correlations, and randomization inference,” *Journal of the American Statistical Association*, 107 (498), 578–591.

Bennett, Magdalena, Juan Pablo Vielma, and José R Zubizarreta (2020) “Building representative matched samples with multi-valued treatments in large observational studies,” *Journal of Computational and Graphical Statistics*, 29 (4), 744–757.

Bojinov, Iavor and Neil Shephard (2019) “Time series experiments and causal estimands: exact randomization tests and trading,” *Journal of the American Statistical Association*, 114 (528), 1665–1682.

Cinelli, Carlos and Chad Hazlett (2020a) “Making sense of sensitivity: Extending omitted variable bias,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82 (1), 39–67.

——— (2020b) "An omitted variable bias framework for sensitivity analysis of instrumental variables," *Work. Pap.*

Cooperman, Alicia Dailey (2017) "Randomization Inference with Rainfall Data: Using Historical Weather Patterns for Variance Estimation," *Political Analysis*, 25 (3), 277–288, 10.1017/pan.2017.17.

Fogarty, Colin B (2020) "Studentized sensitivity analysis for the sample average treatment effect in paired observational studies," *Journal of the American Statistical Association*, 115 (531), 1518–1530.

Fong, Christian, Chad Hazlett, and Kosuke Imai (2018) "Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements," *The Annals of Applied Statistics*, 12 (1), 156–177.

Forastiere, Laura, Michele Carugno, and Michela Baccini (2020) "Assessing short-term impact of PM 10 on mortality using a semiparametric generalized propensity score approach," *Environmental Health*, 19 (1), 1–13.

Gelman, Andrew and John Carlin (2014) "Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors," *Perspectives on Psychological Science*, 9 (6), 641–651.

Gelman, Andrew, Jennifer Hill, and Aki Vehtari (2020) *Regression and other stories*: Cambridge University Press.

Hill, Jennifer L (2011) "Bayesian nonparametric modeling for causal inference," *Journal of Computational and Graphical Statistics*, 20 (1), 217–240.

Hill, Jennifer and Yu-Sung Su (2013) "Assessing lack of common support in causal inference using Bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes," *The Annals of Applied Statistics*, 1386–1420.

Imbens, Guido W and Paul R Rosenbaum (2005) "Robust, accurate confidence intervals with a weak instrument: quarter of birth and education," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168 (1), 109–126.

Imbens, Guido W and Donald B Rubin (2015) *Causal inference in statistics, social, and biomedical sciences*: Cambridge University Press.

Kelly, Morgan (2021) "Persistence, randomization, and spatial noise."

Lopez, Michael J and Roee Gutman (2017) "Estimation of causal effects with multiple treatments: a review and new ideas," *Statistical Science*, 432–454.

McShane, Blakeley B., David Gal, Andrew Gelman, Christian Robert, and Jennifer L. Tackett (2019) "Abandon Statistical Significance," *The American Statistician*, 73 (sup1), 235–245, 10.1080/00031305.2018.1527253.

Menchetti, Fiammetta, Fabrizio Cipollini, and Fabrizia Mealli (2021) "Estimating the causal effect of an intervention in a time series setting: the C-ARIMA approach," *arXiv preprint arXiv:2103.06740*.

Rosenbaum, Paul R (1987) "Sensitivity analysis for certain permutation inferences in matched observational studies," *Biometrika*, 74 (1), 13–26.

——— (2010) *Design of observational studies*: Springer.

Rubin, Donald B (1991) "Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism," *Biometrics*, 1213–1234.

Sommer, Alice J, Emmanuelle Leray, Young Lee, and Marie-Abèle C Bind (2021) "Assessing environmental epidemiology questions in practice with a causal inference pipeline: An investigation of the air pollution-multiple sclerosis relapses relationship," *Statistics in Medicine*, 40 (6), 1321–1335.

Stommes, Drew, P. M. Aronow, and Fredrik Sävje (2021) "On the Reliability of Published Findings Using the Regression Discontinuity Design in Political Science," *arXiv:2109.14526 [stat]*.

Visconti, Giancarlo and José R Zubizarreta (2018) "Handling limited overlap in observational studies with cardinality matching," *Observational Studies*, 4 (1), 217–249.

Wu, Jason and Peng Ding (2021) "Randomization tests for weak null hypotheses in randomized experiments," *Journal of the American Statistical Association*, 116 (536), 1898–1913.