Evaluation of pollution containment policies in the US and the role of machine learning algorithms

Margherita Gerolimetto[a] and Stefano Magrini[a]

[a]Ca' Foscari Univesity Venice; marco.dicataldo@unive.it, margherita.gerolimetto@unive.it, stefano.magrini@unive.it, alessandro.spiganti@unive.it

Abstract

The aim of this study is to analyse policy actions and institutional changes in local governance structures as determinants of air pollutant reductions in US urban areas. First, we construct a dataset on traffic-related air pollution and socio-economic characteristics across urbanized areas of the US. Some of these data are available through Google Earth engine, others are instead provided by institutional sources. In general, raw data come from application of machine learning techniques to either satellite images or monitoring station records and are available at different temporal and spatial resolutions. Then we adopt regression discontinuity design techniques for the evaluations of pollution reduction policies, exploiting the designation of US Transport Management Areas as a quasi-experimental framework.

Keywords: air pollution, policy evaluation, regression discontinuity design, machine learning

1. Introduction

Road traffic is one of the main contributors to air pollution and greenhouse gasses around the world (among others McDuffie et al., 2021). The mixture of vehicle exhausts from fuel combustion and non-exhaust from engine, brake, tire, and road surface wear and re-suspended street dust materials significantly contributes to particulate matter (PM), nitrogen oxides (NOx), and carbon dioxide (CO2) emissions (European Environment Agency, 2016). These emissions disperse into the ambient air as traffic-related air pollution (TRAP), which degrades ambient air quality.

Humans exposed to TRAP are at a higher risk of developing a wide range of adverse health effects, from premature mortality and cardiovascular illness to cognitive and metabolic effects (Fu et al., 2021; Khreis, 2020). On top of these direct effects, there are additional societal burdens: medical costs, missed school days and workdays (among others Nurmagambetov et al. 2018), reduced workers' productivity (Chang et al. 2016), and brain drain (Xue et al. 2021).

Most human exposures to TRAP happen in urban areas (Kura et al., 2013). Even if air quality in western countries has improved enormously in the past decades, most cities around the world struggle to meet air quality standards and guidelines (World Health Organization). Since the number of urban residents is expected to grow rapidly around the world (United Nations and Department of Economic and Social Affairs Population Division, 2022), a greater quantity of people will soon be at risk of TRAP exposure.

With this in mind, the aim of this work is to analyse policy actions and institutional changes in local governance structures as determinants of air pollutant reductions in US urban areas over the last decade.

We exploit the designation of Transport Management Areas (TMAs) as a quasi-experimental framework. TMAs are designated by the US Secretary of Transportation for urbanized areas that overcome the population threshold of 200,000 as defined by the Bureau of Census, in recognition of the complexity of transportation issues. They are subject to several transportation planning requirements among which a Congestion Management Process and an Air Quality Plan.

From the methodological point of view, we rely on Regression Discontinuity Design (RDD) techniques, a long-standing way to obtain credible causal estimates that is gaining increasing popularity in recent times (among others, Cattaneo and Titiunik, 2022). Like other causal inference approaches, the RDD can benefit from the combination with machine learning methods, both to carry out supplementary analyses enhancing the credibility of the results and to handle specific computational issues, such as bandwidth determination (Athey and Imbens, 2017).

As for the data, we construct a dataset on TRAP and socio-economic characteristics across urbanized areas of the US. Some of these data are available through Google Earth engine, others are instead provided by institutional sources like NASA or Environment Protection Agency (EPA). In general, raw data come from application of machine learning techniques to either satellite images or monitoring station records and are available at different temporal and spatial resolutions. The preliminary aim of this work is to gather and merge data on specific variables from different sources and harmonize them to produce a novel dataset to be used for the main objective of the paper that is evaluating via RDD the effectiveness of policy actions implemented in US urban areas with the objective of containing TRAP.

The structure of the paper is as follows. In the second section we present our methodological framework. In the third section we describe the model and the data set. In the last section we illustrate some preliminary results.

2. Methods

In this section we will present our methodological framework.

2.1 Regression Discontinuity Design

A large literature on causal inference for policy evaluations has focused on methods for statistical estimation to answer a question about the counterfactual impact of change in a policy (or treatment). The policy change has not necessarily being observed before or may have been observed for a subset of the population. The goal is then to estimate the impact of small set of treatments using data from randomized experiments or, more commonly, observational studies (that is non-experimental data). The literature identifies a variety of assumptions that, when satisfied, allow the researcher to draw the same type of conclusions that would be available from a randomized experiment.

Drawing inference about the causal effect of a policy from observational data is rather challenging. The main issue is that there are factors, which may be unobserved that are said confounders in the sense that they induce correlation that is not indicative of what would happened if the policy had been changed. In this sense, identification strategies are central to causal inference and in economics the RDD is among the most credible, because it relies on weak and easy to implement non parametric identifying assumptions which permit flexible and robust identification and inference for local treatment. The key feature of RDD is the existence of a score, or running variable, for each unit of the sample, which determines treatment assignment via hard-thresholding: all units whose score is above a known cutoff are treated, while all units below the cutoff are not treated. Identification, estimation, and inference proceed by comparing the outcomes of units near the cutoff taking those below (control group) as counterfactuals to those above (treatment group). For extensive literature reviews, see, among others, Lee and Lemieux (2010) and Cattaneo and Titiunik (2022).

In this work we adopt the potential outcome (or continuity) approach[1], introduced by Hahn et al.

[1]As a complement to the potential outcome approach, Cattaneo et al. 2015, introduced the local randomization framework that is built on the idea that near the cutoff the RD design can be interpreted as a randomized experiment or more precisely as a natural experiment.

(2001), where potential outcomes are taken as random variables, with the n units of analysis forming a random sample from an underlying population and the running variable, or score, X is assumed to be continuously distributed. Let $Y_i(0), Y_i(1)$ denote the pair of potential outcomes for unit i and let $T_i \in \{0, 1\}$ denote the treatment[2]. The realized outcome is $Y_i = Y_i(T_i)$. The treatment received is a function of the running (pretreatment) variable X_i, more specifically $T_i = I_{X_i \geq c}$, where c denotes the threshold or cutoff, that must be exogenous. Formally, the treatment effect is defined from the following identity

$$\tau = E\left(Y_i(1) - Y_i(0)|X = c\right) \tag{1}$$

The two key assumptions for identifications are: a) the regression functions $E\left(Y_i(0)|X_i = x\right)$ and $E\left(Y_i(1)|X_i = x\right)$ are continuous in x at c and b) the density of the running variable near the cutoff is positive. These assumptions capture the idea that units that are barely above and below the cutoff c would exhibit the same average response if their treatment status did not change. Then by implication, any difference between the average response of treated and control units at the cutoff can be attributed to the treatment and can be interpreted as the causal average effect, estimated as the discontinuity in the conditional expectation of Y_i as a function of the running variable at the cutoff:

$$\tau = lim_{X \to c^-} E\left(Y_i|X_i\right) - lim_{X \to c^+} E\left(Y_i|X_i\right) \tag{2}$$

In practice, we have

$$\begin{array}{c|c} -h \leq X_i < c & c \leq X_i < h \\ Y_i = \beta_{0-} + (X_i - c)\beta_{1-} + \epsilon_{i-} & Y_i = \beta_{0+} + (X_i - c)\beta_{1+} + \epsilon_{i+} \end{array}$$

where $\hat{\tau}_{RD} = \hat{\beta}_{0+} - \hat{\beta}_{0-}$ is the vertical distance between the two estimated expectations, h is bandwidth that guarantees that only units that are close to the cutoff c are involved, ϵ_{i-} and ϵ_{i+} are error terms.

The idea is to estimate regression functions for control and treatment group locally and this, in its most basic fashion, can be done by estimating

$$Y_i = \alpha + \tau_{RD} T_i + (X_i - c)\beta_1 + \epsilon_i, \quad -h \leq X_i \leq h \tag{3}$$

where $\hat{\tau}_{RD}$ is the desired estimated discontinuity and ϵ_i is the error term. As mentioned above, the regression is estimated on a subsample of the data that is statistically optimally close to the cutoff so that units are most comparable each other and this should reduce the influence of the confounding factors. However, this reduces also the number of available observations and thus makes estimates and causal inference increasingly imprecise, limiting the ability to access policies. Thus one needs to identify an optimal bandwidth around the cut-off that optimally balances variance and bias. There are several methods to find the optimal bandwidth (e.g. Imbens and Kalyanaraman, 2012). A very interesting recent proposal is a machine learning based method developed by Long and Rooklyn (2020)

Typically researches tend to estimate model (3) adopting local polynomial methods tailored to flexibly approximate, above and below the cutoff, the unknown conditional mean function of the outcome variable given the running variable. In practise, researchers often choose a local linear polynomial and perform the estimation using weighted linear least squares, giving higher weights to observations close to the cutoff. If present, this discontinuity is interpreted as some average response to the treatment at the cutoff, depending on the assumptions and the setting under examination.

2.2 Machine Learning Algorithms

The machine learning literature has traditionally focused on discovering pattern and on prediction, using data-driven approaches to build rich models and relying on cross-validation as powerful tool for model selection.

Supervised machine learning tools could be useful for causal inference given that, similarly to regressions, can summarise linear and non-linear relationships in the data and make predictions (Varian,

[2]See for example Imbens and Rubin (2015) for more details on this set-up.

2014). However, given the lack of interpretable coefficients for some of the algorithms and the lack of standard errors of the obtained coefficients, predictions tools from these literature cannot be readily used for causal inference. Nevertheless, one particularly mature strand of literature includes approaches that incorporate supervised machine learning techniques in the so-called supplementary analyses to improve the credibility of the policy evaluations, such as placebo analysis, internal validity, external validity for RDD methods (Athey and Imbens, 2017). Moreover, in causal inference many estimators involve the specification of parameters, such as the optimal bandwidth in RDD, which are not interest per se, but are necessary to estimate the target parameter, and their determination can be done via machine learning (Long and Rooklyn, 2020). Overall, while most machine learning methods cannot be used to infer causal effects, they can surely help the process.

From a different perspective, machine learning methods can have a fundamental role in the data processing that might be preliminary to subsequent causal inference analyses. This comes from the literature on machine learning methods to estimate variables investigated in geo- and environmental sciences (Wuepper and Finger, 2023). For example, in case of greenhouse gas emissions, data might not be available and researchers must rely on proxies. Most natural phenomena are non-linear, multivariate, highly variable and correlated at many spatio-temporal scales. The analysis and treatment of such complex data and their integration/assimilation with science-based models is a difficult problem that is addressed by contemporary machine learning approaches, e.g. random forest algorithms to generate new variables even directly in Google Earth Engine (2022). Given the growing abundance and improving resolution (spatial, temporal, and spectral) of satellite imagery, in a recent review of the field, Burke et al. (2021) discuss the rapidly increasing literature regarding satellite imagery and measurements of different human outcomes, with specific attention to approaches that combine imagery with machine learning. Researchers in economics also increasingly use satellite imagery, and particularly nightlights imagery, for a variety of applications (among others, Henderson et al., 2012).

3. Model and Data

In order to evaluate the impact of policy actions and institutional changes in local governance structures on air pollutant reductions in US urban areas we will consider the TMAs designation as quasi-natural framework. TMAs are designated by the US Secretary of Transportation for urbanized areas that overcome the population threshold of 200,000 as defined by the Bureau of Census, in recognition of the complexity of transportation issues. When an urban area is designated as a TMA, the Metropolitan Planning Organization (MPO) responsible for that urban area (an MPO is mandatory for urban areas with population over 50,000) is subject to several transportation planning requirements among which a Congestion Management Process

Here, with reference to TMA designation after 2010 Census, we estimate a (very preliminary) RDD model for year 2015. The statistical units are the urbanized areas and the running variable is the population. Treated units are those that, overtaking the threshold $c = 200,000$, have been designed as TMAs. The RDD model we have in mind is:

$$Y_i = \alpha + T_i\tau_{RD} + (X_i - c)\beta + \mathbf{Z}_i\gamma + \epsilon_i, \quad -h \leq X_i \leq h \tag{4}$$

where for each statistical unit i, Y is the level of traffic related pollutant C02, X is the population (running variables), $\mathbf{Z}$ is a vector of covariates, specifically income (proxied by the nightlights) and meteorological variables, such as wind and precipitations, ϵ_i is the error term. The regression is run considering urbanized areas whose value of the running variable is close to the cutoff $c = 200,000$, according to the bandwidth h.

To conduct this analysis, we build a novel dataset, from the following data sources:

Population, TMA designation (source US Census Bureau and US Federal Register): these sources provide information also on the shapefiles for the administrative boundaries of the urbanized areas.

Traffic-related CO2 (source NASA): DARTE (Database of Road Transportation Emissions) data set provides a 38-year, 1-km resolution inventory of annual on-road CO2 emissions for the contermi-

nous United States based on roadway-level vehicle traffic data and state-specific emissions factors for multiple vehicle types on urban and rural roads.

Nightlights (source Google Earth Catalog): VIIRS Nighttime Day/Night Band Composites Version 1 provides monthly average radiance composite images using nighttime data from the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB).

Wind and precipitations (source Google Earth Catalog): TerraClimate is a dataset of monthly climate and climatic water balance for global terrestrial surfaces.

4. Some results

In the presentation of some preliminary results, we show the estimation of the basic version of model (4), that is the RDD regression without covariates:

$$Y_i = \alpha + T_i\tau_{RD} + (X_i - c)\beta + \epsilon_i, \quad -h \leq X_i \leq h \tag{5}$$

Figure 1 shows the local estimation of model (5). It is a local polynomial estimation of degree 2, with triangular kernel where the bandwidth has been optimally selected by MSE minimization, as suggested in Calonico et al. (2014), leaving for a more advanced version of the paper the estimation with a bandwidth selected with machine learning criteria (see section Sect. 2.). In the graph it is evident the discontinuity at the cutoff of the conditional expectation of Y as a function of the running variable: this represents the local effect of the treatment.

Moving now to Table 1, we can see that the treatment effect is significant at 10% level and it has a negative sign, as expected. This confirms the idea of a possible effect of the TMAs designation on the levels of traffic related C02, that will be further explored in a more advanced version of the paper through the estimation of model (4), i.e. in the version including covariates. However, given that the effect is not so strongly significant, we shall also consider other case studies of pollution containment policies.

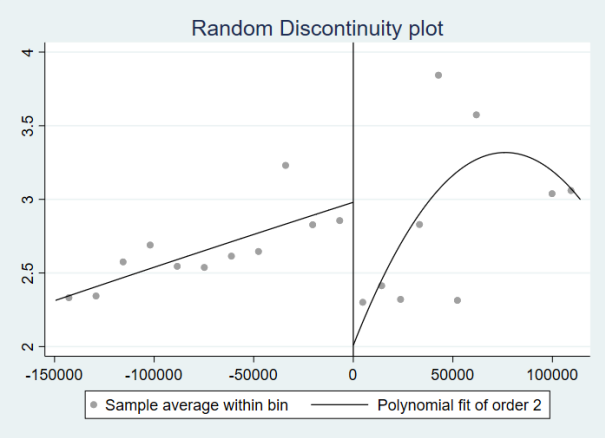

Figure 1: Random discontinuity plot for basic RDD with local polynomial regression

Table 1: Basic RDD estimation with local polynomial regression

Method	Coeff	Std. Error	z	$P > \|z\|$
Conventional	-1.2167	.68582	-1.7741	0.076
Robust	-	-	-1.7200	0.085

References

1. Athey, S., Imbens, G.W.: The State of Applied Econometrics: Causality and Policy Evaluations. Journal of Economics Perspectives **31**, 3−32 (2017)
2. Burke, M., Driscoll, A., Lobell, D.B., and Ermon, S. . Using satellite imagery to understand and promote sustainable development. Science **371**, 1−12 (2021)
3. Calonico, S., Cattaneo, M., Titiunik, R.: Robust non parametric confindence intervals for regression discontinuity designs. Econometrica. **82**, 2295−2326 (2014)
4. Cattaneo, M. D.,Frandsen, B., Titiunik,R.: Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. Senate. Journal of Causal Inference **3**, 1−24 (2015)
5. Cattaneo, M. D., Titiunik, R.: Regression Discontinuity Designs. Annual Review of Economics, **14**, 821−851 (2022)
6. Chang, T., Graff Zivin J., Gross, T., Neidell, M.: Particulate Pollution and the Productivity of Pear Packers. American Economic Journal: Economic Policy, **8** 141−69. (2016)
7. European Environment Agency: European Environment Agency: Explaining Road Emissions. 10.2800/71804 (2016)
8. Fu, S., Viard, V. B., Zhang, P.: Air Pollution and Manufacturing Firm Productivity: Nationwide Estimates for China. The Economic Journal **131**, 241−3273 (2021)
9. Google Earth Engine (2022) `https://earthengine.google.com/`
10. Hahn, J., Todd, P., Van Der Klaauw,W.: Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design. Econometrica, **69**, 201−209 (2001)
11. Henderson J., Soreygard, A. Weil D.: Measuring Economic Growth from Outer Space, American Economic Review **102**, 994−1028 (2012).
12. Imbens, G., Kalyanaraman, K.: Optimal Bandwidth Choice for the Regression Discontinuity Estimator. The Review of Economic Studies, **79**, 933−959 (2012)
13. Imbens, G., Rubin, D.: Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge: Cambridge University Press (2015)
14. Khreis, H.: Traffic, air pollution, and health. In: Nieuwenhuijsen, M.J., Khreis, H. (eds.), Advances in Transportation and Health. Elsevier (2020)
15. Kura, B., Verma, S., Ajdari, E., Iyer, A.: Growing Public Health Concerns from Poor Urban Air Quality: Strategies for Sustainable Urban Living. Comput. Water, Energy. Environ. Eng. **02**, 1−9 (2013)
16. Lee, D. S., Lemieux, T.: Regression Discontinuity Designs in Economics. Journal of Economic Literature, **48**, 281−355 (2010)
17. Long, M.C., Rooklyn, J.: NEXT: Stata module to perform regression discontinuity. Statistical Software Components S458238, Boston College Department of Economics, revised 13 Oct 2020 (2020)
18. McDuffie, E.E., Martin, R.V., Spadaro, J.V., Burnett, R., Smith, S.J., O Rourke, P., Hammer, M.S., van Donkelaar, A., Bindle, L., Shah, V., Jaegle, L.: Source sector and fuel contributions to ambient PM2. 5 and attributable mortality across multiple spatial scales. Nat. Commun. **12**, 1−12 (2021)
19. Nurmagambetov, T., Kuwahara, R., Garbe, P.: The Economic Burden of Asthma in the United States, 2008−2013. Ann. Am. Thorac. Soc. **15**, 348−356. (2018)
20. United Nations, Department of Economic and Social Affairs Population Division. World Population Prospects 2022: Summary of Results. UN DESA/POP/2022/TR/ NO. 3. (2022)
21. Varian, H.R.: Big data: new tricks for econometrics. Journal of Economics Perspectives **23**, 317−320 (2014)
22. Wuepper, D., Finger, R.: Regression discontinuity designs in agricultural and environmental economics. European Review of Agricultural Economics, **50**, 1−28 (2023)
23. Xue, S., Zhang, B., Zhao, X.: Brain drain: The impact of air pollution on firm performance. Journal of Environmental Economics and Management. **110**, 102546 (2021)