Empirical Environmental Economics: Paper replication

By Tsubasa Takaya (222617)*

I replicate the paper by Herrnstadt et al. (2021), "Air pollution and criminal activity: microgeographic evidence from Chicago", published in American Economic Journal: Applied Economics, 13(4): 70-100.

I. Summary

This paper investigates yet another potential negative impact of air pollution on society by testing the hypothesis that air pollution induces violent criminal activity. Since violent crimes pose significant social costs, overlooking such adverse effects of air pollution would lead to suboptimal externality-correcting regulation or Pigouvian taxes. The authors find that air pollution indeed causes violent crimes to increase by about 2% but not property crimes, a consistent result from both time-series and microgeographic analyses. While small in percentage terms, a rough conservative extrapolation of this increase suggests that the annual costs of crimes associated with local air pollution range from several hundred million to a few billion dollars.

For the identification of the causal effects, the authors exploit temporal and geographical variations in air pollution and crimes in the city of Chicago as a quasi-experiment. For the former, they compare crime occurrences at the city level on days with higher PM_{10} measures against days with lower PM_{10} measures using regression. Observed confounders that are controlled for in this strategy include calendar fixed effects and weather conditions such as temperature and precipitation (see below). To further address unobserved economic activity, they instrument for daily PM_{10} measures using wind direction. The assumed relevance is based on the idea that the wind blowing from the southeast and southwest of Chicago, where major industries are located, will cause elevated PM_{10} levels (see Table 2 for F statistics as a measure of relevance). To the extent that weather controls eliminate other confounders that could cause wind direction, pollution, and crimes, the exogeneity assumption is also valid. The exclusion restriction can be potentially invalid, however, if we believe that the wind could directly affect crimes independently of air pollution (the authors mention winds off of Lake Michigan in the north as a hypothetical example). The monotonicity assumption is not discussed but remains as a matter of plausibility[1].

* Takaya: Hertie School. My submission also replicates the AEJ-style.

[1]In this context, defiers would mean days that experience *lower* pollution when the wind blows from the southeast or southwest and higher pollution when the wind comes from the north. This is not testable, as it involves an evaluation of the counterfactuals of such days. The credibility of this assumption is not

To bypass these underlying theoretical and empirical challenges of time-series analysis, the authors further compare local neighborhoods on opposite sides of interstate highways on days when the wind hits the direction of the interstate orthogonally. Seeing interstates as a major fixed source of air pollution, they argue that the "downwind" region would experience a higher level of air pollution due to pollutants from vehicles when the wind intersects the highway orthogonally. Since the neighborhoods around a given interstate on a given day likely face identical weather conditions and unobserved economic activity, the upwind side of the interstate can be considered as a credible counterfactual of the downwind side absent the wind blowing in that direction. The authors include interstate-side fixed effects and their interaction with weather covariates, as well as interstate-date fixed effects to address remaining unobserved heterogeneity.

An ideal specification would regress local crime outcomes on local measures of air pollution directly, without relying on air pollution induced by interstates and the wind. In fact, the "treatment" is defined only in terms of sides of an interstate and the wind direction because such local pollution readings are not available. Hence, there is no guarantee that air pollution is actually the primary cause of the variation in observed crimes. To address this, the authors conduct a set of robustness checks, including different thresholds of the wind direction and distance to interstates, leads and lags of pollution, placebo tests, and alternative specifications. In the placebo tests, the authors iteratively estimate the effect of being in the downwind side on violent crime for each latitude across the city and find that the treatment effect becomes statistically and substantively significant exactly at the latitude of an interstate that cuts east-west through Chicago (See Figure 7 of the paper). This result provides compelling evidence for a causal effect of local air pollution on violent crimes[2].

Overall, this research has high internal validity, and the estimated associations can be interpreted causally. In the following section, I replicate the main findings from the city-level time-series analysis and the microgeographic analysis using the replication package available on the openICPSR platform[3]. In the final section, I examine the external validity of the findings.

II. Replication

The replication package contains a MASTER.do file written in Stata, a code folder, and a raw data folder. Relevant Stata and R scripts are listed in sequence in the MASTER.do file to reproduce the data sets, tables, and figures used in the paper. For the city-level analysis, I reconstruct a data set (henceforth "city data

discussed in the paper.

[2]Strictly speaking, other disamenities than air pollution could be brought about by the intersection of the wind and the interstate and in turn increase crimes. The authors mention traffic noise as an example. However, they cite studies that find no difference in noise dispersion between the upwind and downwind sides of roads and conclude that this is not a concern when using the upwind side as a control.

[3]https://www.openicpsr.org/openicpsr/project/119403/version/V1/view

Table 1—Summary statistics

	Mean	Standard deviation
Citywide sample:		
Number of dates	3,641	
Daily citywide violent crime	57.4	18.7
Daily citywide property crime	420.1	68.4
Precipitation (mm)	2.75	7.74
Maximum temperature (°C)	15.5	11.6
Daily average CO (ppm)	0.59	0.27
Daily average NO2 (ppm)	0.027	0.0085
Daily average ozone (ppm)	0.023	0.012
Daily average PM10 (mug/m3)	27.7	14.4
Wind speed (km/h)	12.3	4.40
Dew point (°C)	4.44	10.1
Air pressure (hpa)	1,016.6	7.09
Cloud cover sunrise to sunset (percent)	63.8	27.7
Interstate sample:		
Interstate-side days	41,730	
Daily interstate-side violent crimes	1.1	1.4
Daily interstate-side property crimes	7.3	5.2

Notes: This table replicates Table 1 of the paper. The reproduced city data set is missing one day compared to the original paper. Otherwise, the mean and standard deviation of all variables are identical except for a difference of 0.0001 in the standard deviation of the daily average NO_2.

set") from raw crime data, weather data, and pollution data covering the study period of 2001-2012 using Python. For the microgeographic analysis, I simply reuse the data set (henceforth "micro data set") already created by the authors[4]. For the replication of models and tables, I use R.

Table 1 shows summary statistics of relevant variables, which is a replication of Table 1 of the paper. There are a few minor differences between my table and the original table. First, the city data set I create is missing one sample. This missingness most likely arises from the procedure of dropping rows with missing values. Specifically, the authors filter out rows with missing values in any of the variables shown in the table. While this is supposed to result in dropping only one day, my code drops two days: April 21, 2002, which is missing air pressure, and March 9, 2007, which is missing daily maximum temperature. I cannot tell which of the two days the original data set is missing. Nevertheless, the lack of one sample does not affect the summary statistics and the replication of the main result of the city-level analysis. There is also a difference of 0.0001 in the standard deviation of the daily average NO_2, but this is as small as some rounding error.

[4] I did write a function for constructing micro data set in Python as well (see `create_micro_dataset()`). But the reproduced data set diverges when creating a treatment variable, and so do the reproduced results. Although the results are still fairly similar, I use the provided data set here for completeness.

Table 2—PM_{10} impact on daily part 1 crime, 2001-2012

	Violent crimes			Property crimes		
	OLS		IV	OLS		IV
	(1)	(2)	(3)	(4)	(5)	(6)
Standardized PM_{10} reading	0.0614	0.0135	0.0291	0.0041	-0.0015	0.0004
	(0.0037)	(0.0043)	(0.0128)	(0.0014)	(0.0019)	(0.0058)
First stage F-statistic			25.7			25.7
Calendar fixed effects	X	X	X	X	X	X
Weather controls		X	X		X	X
Historical mean temp		X	X		X	X
Observations	3,643	3,641	3,641	3,643	3,641	3,641
R^2	0.69	0.75	0.75	0.79	0.81	0.81

Notes: This table replicates Table 2 of the paper. While there are a few differences in the sample size, the coefficient in Column 6, and the first stage F-statistic, the estimated coefficients and standard errors, as well as R^2 match those in the paper.

For the interstate sample, there is no difference between my table and the original table, since I simply use the data set that is preprocessed by the authors and provided in the replication package. Overall, Table 1 replicates Table 1 of the paper almost perfectly.

Next, I present the replicated result of the city-level analysis in Table 2, which corresponds to Table 2 of the paper. For all specifications, I use the `feols` function from the `fixest` package because it provides a richer set of options for robust standard errors. Here I report Newey-West robust standard errors with a lag of one day, emulating the Stata implementation. We can observe mainly three discrepancies. First, similar to Table 1 above, the number of observations used for the estimation is different. Columns 2, 3, 5, and 6 are missing one more sample compared to the original because two days have a missing value in the weather covariates in my data set. For Columns 1 and 4, I use the complete data set, as there is no variable that has missing values. On the other hand, the authors explicitly drop a row with missing values before fitting any model and thus use the same number of observations across all the specifications. With that said, the coefficients, standard errors, and R^2 in Columns 1 and 4 are completely the same. Second, the coefficient in Column 6 of the original result (0.0013) is three times larger than that of the replicated result (0.0004), although both share the same standard error. I cannot pin down the exact cause of this discrepancy, but given the relatively large uncertainty, it might be plausible that the estimate is unstable. Finally, the first stage F-statistic is different by 0.7. This is likely because of different implementations of the first stage regression between Stata and R.

More importantly, despite these few differences, the main IV specification in Column 3 yields an identical causal effect estimate with an identical standard

Table 3—crime downwind of interstates

	(1)	(2)	(3)	(4)
Panel A: Violent crime				
Treatment (downwind)	0.0558	0.0186	0.0186	0.0188
	(0.0108)	(0.0094)	(0.0089)	(0.0091)
Route × side fixed effects		X	X	X
Route × date fixed effects			X	X
Route × side weather interaction				X
Observations	41,730	41,730	41,730	41,720
R^2	0.001	0.274	0.678	0.68
Panel B: Property crime				
Treatment (downwind)	0.0015	-0.0014	-0.0014	-0.0007
	(0.0071)	(0.0046)	(0.0041)	(0.0043)
Route × side fixed effects		X	X	X
Route × date fixed effects			X	X
Route × side weather interaction				X
Observations	41,730	41,730	41,730	41,720
R^2	0	0.609	0.841	0.843

Notes: This table replicates Table 4 of the paper. The original micro data set provided by the authors is used. All values are identical to those in the paper.

error: one standard deviation increase in PM_{10} level causes violent crimes to increase by 2.9%, and the effect is statistically significant. Other values such as R^2 are also the same. Overall, the original results are well replicated.

Lastly, I reproduce the main findings of this paper from the microgeographic analysis and show the results in Table 3, which corresponds to Table 4 of the paper. The implementation is straightforward, as I use the same data set the authors use, and all model specifications are based on fixed effects OLS. Hence, identical results are obtained in the replication for all the values including coefficients, standard errors, sample size, and R^2. Focusing on Column 4, which includes all the fixed effects and interaction with temperature and precipitation, violent crime increases by 1.9% on the downwind side of the interstate on days when the wind blows orthogonally to the interstate. As with the city-level results, there is no effect on property crime.

It is worth mentioning a few details on the data set used for this microgeographic regression, as the extension builds on it. First, since the unit of measurement is interstate-side-day (i.e. for each day between 2001 and 2012, there are seven interstates in the sample, for each of which there are two sides) but most of the variables in the data set vary only at the daily level, we cannot control for other confounders such as local temperature and local wind direction. As a result, the possible set of controls is restricted to fixed effects at different hierarchies of interstates, sides, and days, and their interaction with weather covariates. We can see that after controlling for interstate-side fixed effects (which is the most granular possible covariate), the inclusion of other terms does not lead to meaningfully different estimates, as the signals tend to be sparse. Second, the weather variables considered here are daily maximum temperature and precipitation. While the latter is the same variable as previously used, the temperature variable is different from the one in the summary statistics or the city-level analysis.

III. Extension

In this section, I attempt to extend the external validity of the microgeographic evidence by shedding light on effect heterogeneity. In the main specification in Table 3, the authors estimate an average treatment effect of being in the downwind side of an interstate on violent crime when the wind blows orthogonally and, in doing so, pool all observations from all seven interstates in the sample. While the identification strategy likely has high internal validity, whether the estimated causal effect generalizes to a new interstate on another day with different weather conditions (a "test" interstate-side-day) is not addressed in the paper. In the following, I explore how the treatment effect varies by different interstates and is moderated by other covariates. To this end, I employ the generalized random forest algorithm developed by Athey, Tibshirani and Wager (2019), which is a non-parametric forest-based method that can flexibly capture heterogeneity in a key parameter of interest (i.e. conditional treatment effect).

The generalized random forest algorithm (Athey, Tibshirani and Wager 2019) recursively splits training samples into smaller leaf nodes based on covariates (features). We can think of this recursive partitioning as a process of generating adaptive local weights for neighbors of a specific test point x. The algorithm then estimates the target function, that is conditional average treatment effect (CATE) function $\tau(x)$ in the present context, as a solution to a weighted set of an "estimation equation" (also known as moment condition). This data-driven, forest-based approach allows for a more flexible and accurate prediction of such potentially complex and nonlinear functions as the CATE function compared to parametric methods (e.g. interactive regression) and scales well to a high-dimensional covariate space unlike conventional nearest neighbor estimators such as a kernel weighting function.

To allow for inference with valid confidence intervals centering around the target $\tau(x)$, the generalized random forest algorithm implements random subsampling of the training data and subsample splitting called honesty (Wager and Athey 2018). At a high level, the latter technique ensures that only a subset of the subsample is used for growing a tree, and the other portion is held out for estimating the target parameter. The sampling distribution of the estimate produced by honest forests is asymptotically Gaussian and unbiased.

In this extension, I restrict the outcome variable to a mean-scaled count of violent crimes only. Upon applying the generalized random forests method to the microgeographic analysis of the paper, I heavily rely on the implementation explained in Athey and Wager (2019). First, I fit generalized random forests with

Table 4—Heterogeneous treatment effect downwind interstate

	Non-clustered	Clustered
Average treatment effect for the overlap (ATO)		
Treatment (downwind)	0.0227	0.0569
	(0.0103)	(0.0472)
Conditional average treatment effect (CATE)		
Daily maximum temperature	0.0022	0.0027
	(0.0009)	(0.0023)
Daily precipitation	-0.0010	-0.0024
	(0.0014)	(0.0026)
Daily average wind speed	0.0023	0.0246
	(0.0059)	(0.0236)
Route × side fixed effects	X	X
Route × side weather interaction	X	X
Route fixed effects		X
Observations	41,716	41,716

Notes: Treatment effects estimated by generalized random forests. The dependent variable is the number of violent crimes within one mile of one side of the interstate normalized by the mean number of crimes. The top panel shows the average treatment effect for the overlap (ATO), which gives smaller weights to samples with propensity scores close to 0 and 1. The bottom panel shows the best linear projection of conditional treatment effects (CATE) along the respective covariates. The clustered model clusters samples at the interstate level.

a similar specification to Column 4 of Table 3, along with other variables. Specifically, I include interstate-side fixed effects and their interaction with weather covariates (maximum temperature and precipitation) but omit interstate-date fixed effects because adding these does not change results significantly. Instead of binning moderation variables like the authors do, I estimate average treatment effects conditional on continuous measures of maximum temperature, wind speed[5], and precipitation. For the interactive regression in Figure 5 of the paper, the authors use the other daily temperature variable than the one used as a fixed effect (remember there are two maximum temperature variables). Consequently, their specification contains two different maximum temperature variables. I use only one for both fixed effects and the moderation variable[6].

Second, I run the forest algorithm with interstate-level clusters. The main specification of the paper treats each of the interstate-side-days exchangeable when estimating the effect of downwind on violent crime. While the authors take into account interstate-side fixed effects and interstate-date fixed effects, the level of pollution when the wind hits a given interstate orthogonally would likely vary at the interstate level due to unobserved traffic volumes, car density, highway

[5]I scale by 10 following the authors' implementation in microregs_interactions.do.
[6]This temperature variable is the same as the one the city-level analysis.

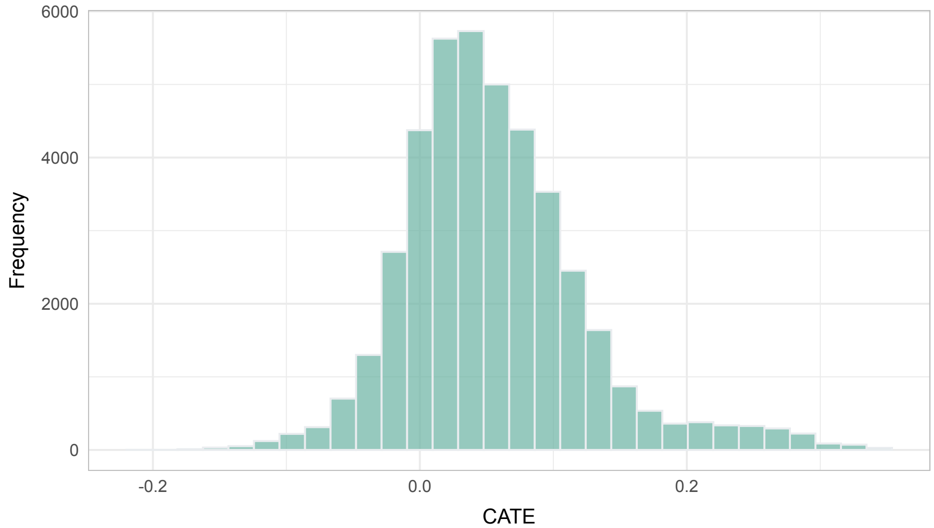


Figure 1. Histogram of out-of-bag CATE estimates

Notes: Distribution of CATE estimates produced by a causal forest with clustered samples. For each sample, a prediction is made using only those trees that are grown without the cluster the unit belongs to (out-of-bag prediction for clustered data).

structure, and so on. These potential unobserved factors at the interstate level may in fact be treatment effect modifiers and could have important implications when we extrapolate the obtained estimates beyond the seven interstates in the sample (Athey and Wager 2019). In order to conduct a cluster-robust conservative inference on the treatment effect, the second model controls for interstate fixed effects and allows for spatial correlations of errors at the interstate level.

Table 4 presents results of the two generalized random forests, with and without clustering. The top panel reports the overall effect of downwind on violent crimes. Following the recommendation by Li, Morgan and Zaslavsky (2018), here I show the average treatment effect *for the overlap* because some of the samples have poor common support with estimated treatment propensities being very close to 0 and 1. We can see that overlooking potential spatial correlations at the interstate level leads to an underestimation of standard errors. In fact, once the samples are clustered at the interstate level, the treatment effect is no longer statistically significant[7]. The wide distribution of estimated CATE confirms that the average treatment effect is statistically indistinguishable from zero (Figure 1).

The bottom panel of Table 4 shows the best linear projection of conditional average treatment effects moderated by daily maximum temperature, precipitation, and average wind speed. While higher temperature seems to amplify the treat-

[7]The effect size becomes also similar to that of Column 1 of Table 3, which does not include any controls. It is probably because all the fixed effect dummies receive variable importance of zero once the data are clustered at the interstate level.

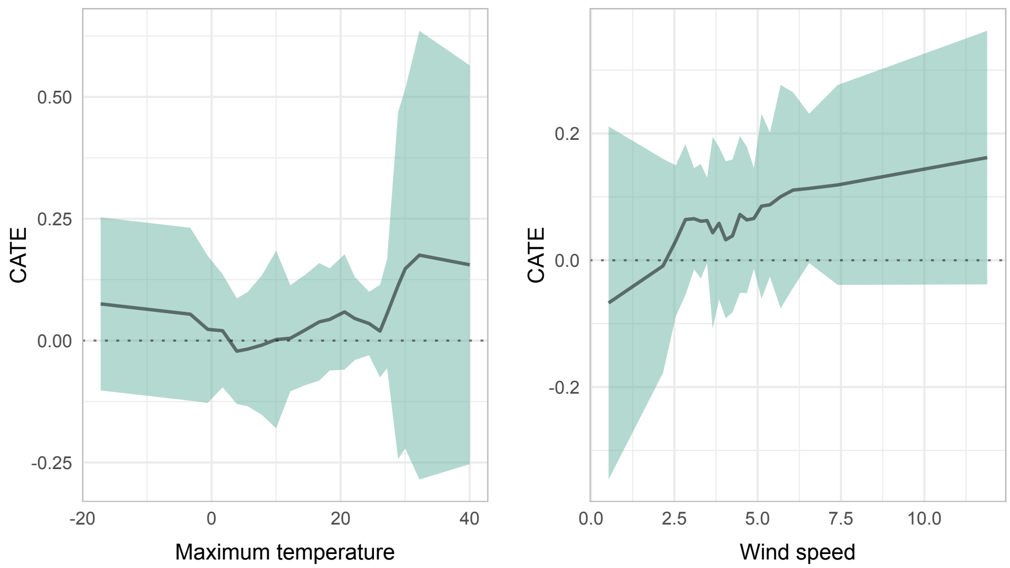


FIGURE 2. ESTIMATED CONDITIONAL AVERAGE TREATMENT EFFECTS BY TEMPERATURE AND WIND

Notes: Figure plots CATE estimates and pointwise 95% confidence intervals produced by a causal forest with clustered samples. A positive value indicates that the treatment effect increases as an x-axis variable increases. The right plot shows treatment effects conditional on daily maximum temperature while holding other variables at their median values. The left plot shows treatment effects conditional on daily average wind speed while holding other variable at their median values. All fixed effects (a set of dummy variables) are set to zero, as they do not affect the estimates.

ment effect according to the non-clustered model, none of the weather covariates has a statistically significant coefficient once we cluster the data, implying that there is no evidence of heterogeneity along these weather covariates (a calibration test based on the best linear predictor method also indicates poor quality of the estimates from the clustered casual forest. The results are shown in Table A1). Although this "omnibus" test does not mean that there is no heterogeneity present, individual t-tests do not find strong evidence of effect heterogeneity by temperature ([-0.11, 0.26]), precipitation ([-0.12, 0.25]), or wind speed ([-0.10, 0.27]) either. Beyond these linear tests, Figure 2 visualizes more granular CATE estimates along daily maximum temperature and daily average wind speed (See Figure A1 for precipitation). Although there is substantial uncertainty, the effect of downwind on violent crime seems to be more pronounced on very cold and warm days. Also, the CATE estimates move upward with wind speed in an almost linear fashion, although signals in the high wind region seem sparse.

Lastly, I present the (marginal) distribution of estimated CATEs for each of the seven interstates across the city of Chicago, produced by generalized random forests with clustered samples. Overall, these interstates share a similar central tendency in the treatment effect estimates, although the density at the peak and the tails slightly differs among them. In particular, I90B (the middle portion of

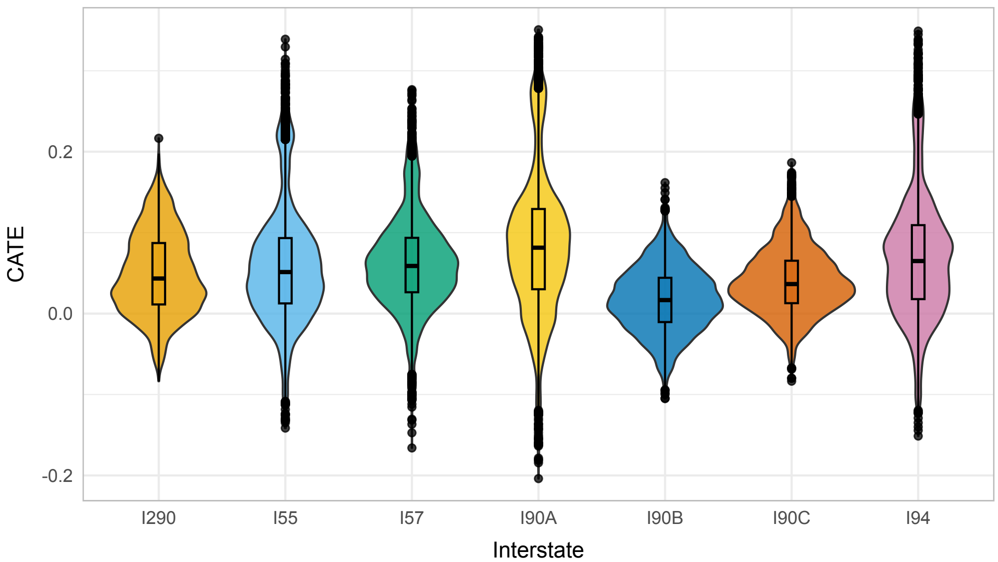


Figure 3. Distribution of estimated treatment effects by interstate

Notes: Figure plots the distribution of CATE estimates for each of the seven interstates produced by a causal forest with clustered samples. I90A, I90B, and I90C refer to the northwestern, middle, and southeastern portion of I90 respectively, as shown in Figure 3 of the paper.

I90) and I90C (the southeastern portion of I90) seem to have slightly more modest expected treatment effects with a larger peak density near zero. Nevertheless, these interstates are indistinguishable in terms of treatment heterogeneity, and none of the treatment effects of individual interstates is statistically significant in the clustered model, which is in line with the overall results.

To conclude, I do not find evidence of weather-driven heterogeneity in the causal effect of being in the downwind side of an interstate on violent crime when the wind blows orthogonally. Once I adopt the cluster-robust casual forest algorithm, not only the conditional average treatment effect estimates but also the average treatment effect for the overlap (ATO) is no longer statistically significant. According to Athey and Wager (2019), this clustering approach can be thought of as fitting a random effects model where treatment effect functions are estimated for each cluster while enabling cluster-robust inference. Hence, to the extent that we believe there might be unobserved per-interstate factors that may affect the sampling variability of the outcome (e.g. spatial correlations), the null findings from the per-cluster approach challenge the generalizability of the main micro-geographic evidence to an out-of-sample interstate.

REFERENCES

Athey, Susan, and Stefan Wager. 2019. "Estimating Treatment Effects with Causal Forests: An Application." Version Number: 1.

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. "Generalized random forests." *The Annals of Statistics*, 47(2).

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.

Li, Fan, Kari Lock Morgan, and Alan M. Zaslavsky. 2018. "Balancing Covariates via Propensity Score Weighting." *Journal of the American Statistical Association*, 113(521): 390–400.

Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association*, 113(523): 1228–1242.

Appendix

Supplementary tables and figures

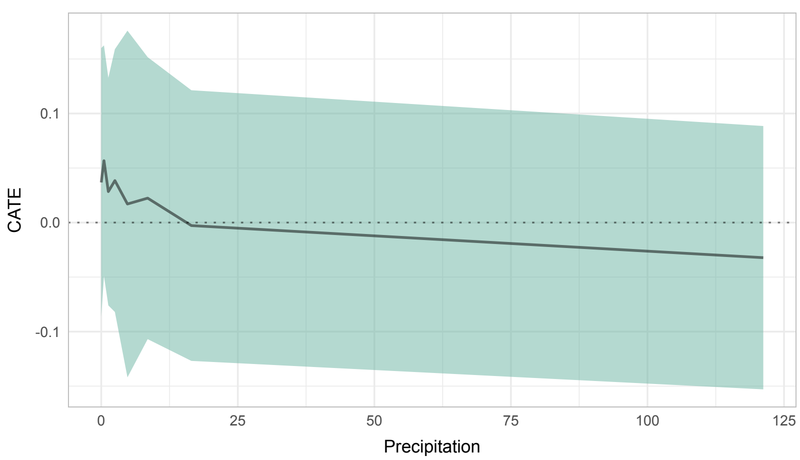
Table A1—heterogeneity calibration test

	Non-clustered	Clustered
Mean forest prediction	1.073	1.033
	(0.495)	(0.748)
Differential forest prediction	0.236	-0.708
	(0.123)	(0.793)
Route × side fixed effects	X	X
Route × side weather interaction	X	X
Route fixed effects		X
Observations	41,716	41,716

Notes: This table reports the results of a calibration test for the presence of heterogeneity. A coefficient for the mean forest prediction absorbs the average treatment effect, whereas a coefficient for the differential forest prediction measure the quality of the estimates of treatment heterogeneity. If a coefficient for the latter is 1, heterogeneity estimates are well calibrated. If it is positive and statistically significant, we can at least reject the null of no heterogeneity. In the clustered model, no evidence of heterogeneity is found. See the vignette of the `grf` R package[a] and Athey and Wager (2019) for more detailed descriptions of the test.

[a]https://grf-labs.github.io/grf/index.html

Figure A1. estimated conditional average treatment effects by precipitation



Notes: Figure plots CATE estimates and pointwise 95% confidence intervals produced by a causal forest with clustered samples. A positive value indicates that the treatment effect increases as an x-axis variable increases. The figure shows treatment effects conditional on precipitation while holding other variables at their median values. All fixed effects (a set of dummy variables) are set to zero.

Code

All the replication code to generate the data set, tables, and figures can be found in my GitHub repository (https://github.com/tsubasatakaya/emp-env-econ-replication-project)