Causal Inference Workshop

Week 10 - Fixed effects, climate regressions, remote sensing

Causal Inference Workshop

April 12, 2024

Workshop outline

A. Causal inference fundamentals
   - Modeling assumptions matter too
   - Conceptual framework (potential outcomes framework)

B. Design stage: common identification strategies
   - IV + RDD [coding]
   - DiD, DiDiD, Event Studies, New TWFE Lit [coding]
   - Synthetic Control / Synthetic DiD [coding]

C. Analysis stage: strengthening inferences
   - Limitations of identification strategies, pre-estimation steps
   - Estimation [controls] and post-estimation steps [supporting assumptions]

D. Other topics in causal inference and sustainable development
   - Inference (randomization inference, bootstrapping, etc.)
   - Fixed effects, climate regressions, remote sensing data
   - Intro to text analysis, other topics (?), and wrap up!

Outline

Workshop outline

Fixed effects

Climate regressions

Remote sensing

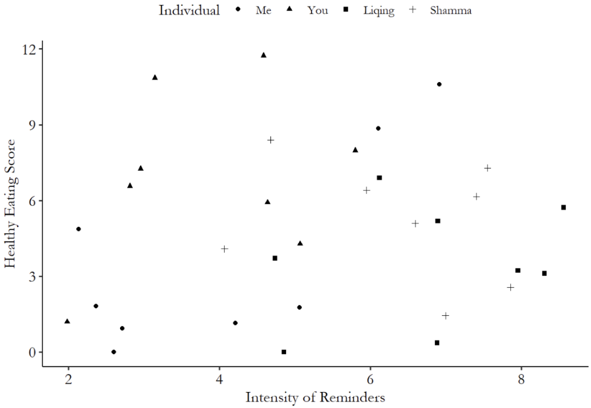
Fixed effects

- Want to control for all confounders. But what if we can't measure all the variables that we need to control for?

- Fixed effects control for *all* variables (observed or not), as long as they are constant *within* some larger category
  - Control for larger category → control for everything constant within that category
  - This means getting rid of variation *between* categories (e.g., taking all the variation in the data explained by the category and getting rid of it)
  - Examples:
    - Effect of rural towns getting electricity on productivity, controlling for geography → town FEs
    - Controlling for "person's background" → individual FEs

Fixed effects: Individual, unit, group fixed effects

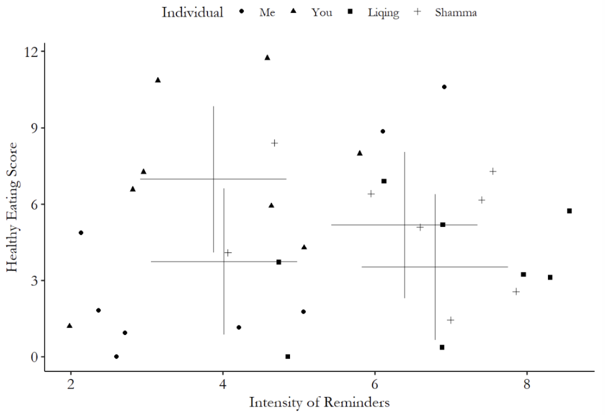
- Use *within* variation; getting rid of variation *between* units or categories



Source: The Effect Book (theeffectbook.net)

Fixed effects: Individual, unit, group fixed effects

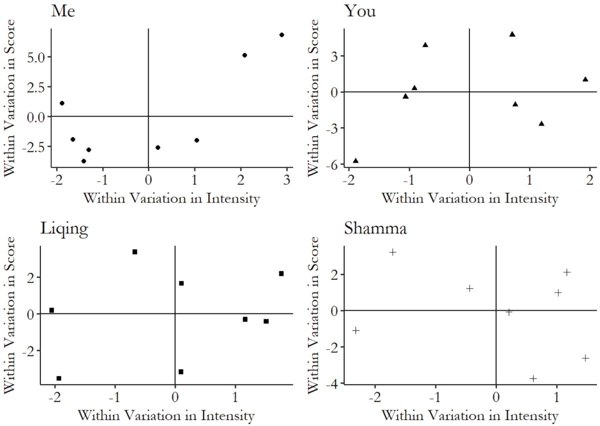
- Use *within* variation; getting rid of variation *between* units or categories



Source: The Effect Book (theeffectbook.net)

Fixed effects: Individual, unit, group fixed effects

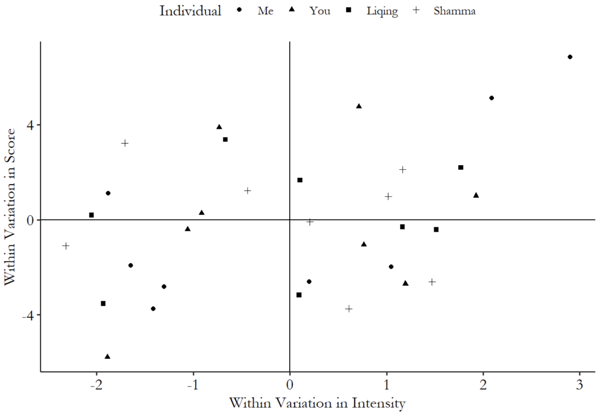
- Use *within* variation; getting rid of variation *between* units or categories



Source: The Effect Book (theeffectbook.net)

Fixed effects: Individual, unit, group fixed effects

- Use *within* variation; getting rid of variation *between* units or categories



Source: The Effect Book (theeffectbook.net)

Fixed effects: Multiple fixed effects

- Gets more complicated with multiple sets of fixed effects

- Example 1: Two non-time fixed effects, e.g., individual and city
  - Isolating variation *within* individual and *within* city
  - City dummies will be based on individuals who move [see coding example]

Fixed effects: Multiple fixed effects

- Gets more complicated with multiple sets of fixed effects

- Example 1: Two non-time fixed effects, e.g., individual and city
  - Isolating variation *within* individual and *within* city
  - City dummies will be based on individuals who move [see coding example]

- Example 2: Unit and time (two-way fixed effects)
  - Isolating variation *within* individual as well as *within* year
    → variation *relative to what we'd expect given that individual, and given that year*
  - Example:

Year	**Person A**	**All People**
2008:	$120,000	$40,000
2009:	$116,000	$30,000
2010:	$120,000	$40,000

  - Variation that's used to estimate treatment effect focuses more heavily on individuals that have a lot of variation over time (Chaisemartin and D'Haultfœuille 2020)

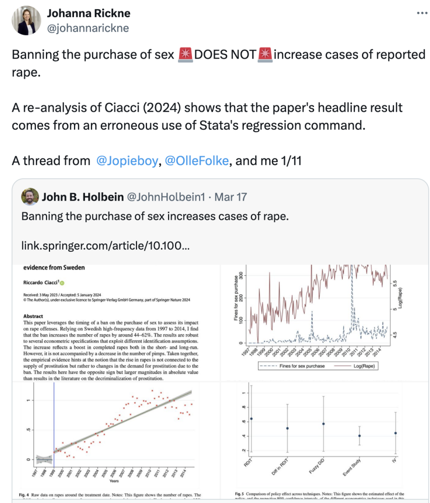
Fixed effects: Takeaways

- **Always, always think about what variation is left after fixed effects!**
  - Don't just throw in a bunch of fixed effects without thinking about them
  - Helpful to say in words (in paper or presentation) where the variation is coming from

- For *unit* fixed effects → treatment variable should vary within the unit
  - If treatment is fixed for all units, treatment dummy will be perfectly collinear with unit FE

- For *time* fixed effects → treatment variable should vary within the time dimension
  - If all units are treated at the same time, treatment dummy will be perfectly collinear with time FE

Fixed effects: A cautionary tale!



Johanna Rickne
@johannarickne

Banning the purchase of sex 🚨 DOES NOT 🚨 increase cases of reported rape.

A re-analysis of Ciacci (2024) shows that the paper's headline result comes from an erroneous use of Stata's regression command.

A thread from @Jopieboy, @OlleFolke, and me 1/11

John B. Holbein @JohnHolbein1 · Mar 17

Banning the purchase of sex increases cases of rape.

link.springer.com/article/10.100...

evidence from Sweden

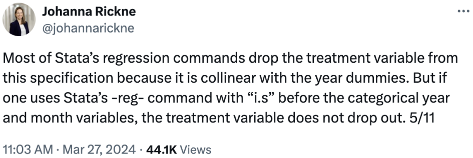
Riccardo Ciacci[1]

Received: 3 May 2023 / Accepted: 5 January 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

Abstract
This paper leverages the timing of a ban on the purchase of sex to assess its impact on rape offenses. Relying on Swedish high-frequency data from 1997 to 2014, I find that the ban increases the number of rapes by around 44–62%. The results are robust to several econometric specifications that exploit different identification assumptions. The increase reflects a boost in completed rapes both in the short- and long-run. However, it is not accompanied by a decrease in the number of pimps. Taken together, the empirical evidence hints at the notion that the rise in rapes is not connected to the supply of prostitution but rather to changes in the demand for prostitution due to the ban. The results here have the opposite sign but larger magnitudes in absolute value than results in the literature on the decriminalization of prostitution.

Fig. 4 Raw data on rapes around the treatment date. Notes: This figure shows the number of rapes. The

Fig. 5 Comparison of policy effect across techniques. Notes: This figure shows the estimated effect of the policy, and the respective 90% confidence intervals, of the different econometrics techniques used in this



Johanna Rickne
@johannarickne

Most of Stata's regression commands drop the treatment variable from this specification because it is collinear with the year dummies. But if one uses Stata's -reg- command with "i.s" before the categorical year and month variables, the treatment variable does not drop out. 5/11

11:03 AM · Mar 27, 2024 · **44.1K** Views

- Country-level ban on purchase of sex in Sweden (1999)
- Regression has year, month, and county fixed effects
- But no variation of treatment within year, treatment dummy (post-1999) collinear with year FE

Fixed effects: A cautionary tale!

- No variation in time dimension, but if you use Stata's `reg` or R's `lm`, and include treatment variable *before* the fixed effects, treatment won't drop out but will essentially just be a year FE

- See `01_fe.R` as an example

Outline

Workshop outline

Fixed effects

Climate regressions

Remote sensing

Climate regressions: Brief intro to two aspects of climate analyses

1. Spatiotemporal aggregation of climate data for use in climate impacts analysis
   - Independent variable: high-resolution gridded climate data (e.g., ERA-5)
   - Dependent variable: spatially irregular economic data (e.g., mortality at ADM2 level)

2. Regressions with this data
   - Want to use plausibly exogenous variation in temperatures
   - Want to capture non-linearities

Climate regressions: Spatiotemporal aggregation

1. Spatiotemporal aggregation of climate data for use in climate impacts analysis
   - Independent variable: high-resolution gridded climate data (e.g., ERA-5)
   - Dependent variable: spatially irregular economic data (e.g., mortality at ADM2 level)

- Want to preserve nonlinearities (e.g., very high temperatures in city with large population, even if it was cooler elsewhere in the administrative region)

Climate regressions: Spatiotemporal aggregation

1. Spatiotemporal aggregation of climate data for use in climate impacts analysis
    - Independent variable: high-resolution gridded climate data (e.g., ERA-5)
    - Dependent variable: spatially irregular economic data (e.g., mortality at ADM2 level)

- Want to preserve nonlinearities (e.g., very high temperatures in city with large population, even if it was cooler elsewhere in the administrative region)

- Steps to follow:
    1. Temporal averaging (e.g., calculate daily max from hourly data)
    2. Transformations at grid-cell level (e.g., square, spline, etc. of daily max temperature)
    3. Weighted spatial average of grids overlapping with polygon (e.g., population weighted)

- Useful resources and packages:
    - `stagg`: R package for spatiotemporal aggregation of ERA-5 / other gridded climate data
        - see demo poster for an illustration of what package does
    - Google Earth Engine: useful for processing ERA-5, CHIRPS, etc.
        - easy to calculate long-term averages (example script) or export ERA-5 data (example script)

Climate regressions: Regressions with climate data

2. Regressions with this data
   - Want to use plausibly exogenous variation in temperatures
   - Want to capture non-linearities

- Usually some version of the following regression

$$y_{it} = f(T_{it}) + g(P_{it}) + \alpha_i + \delta_t + \epsilon_{it}$$

- What function of temperature?
  - Bins, polynomials, or cubic spline

Climate regressions: Regressions with climate data

2. Regressions with this data

$$y_{it} = f(T_{it}) + g(P_{it}) + \alpha_i + \delta_t + \epsilon_{it}$$

- What fixed effects to include?
  - **monthly mortality rate** in US in Barreca et al. (2016): state-by-month (seasonality, seasonal employment, etc.) + year-by-month (changes in mortality common across states, e.g., changes in Medicare) + quadratic time trend
  - **annual all-cause mortality rate** in Carleton et al. (2022): ADM2 (within-location year-to-year variation in temperature and rainfall exposure) + country × year (time-varying trends or shocks to mortality rates)
  - **minutes worked on date** in Rode et al. (2022): subnational location (within-location variation in temperature) + country-by-year (long-term trends) + country-by-week-of-year (seasonality)
  - Show robustness to alternative spatiotemporal controls in the appendix

Climate regressions: Regressions with climate data

2. Regressions with this data

$$y_{it} = f(T_{it}) + g(P_{it}) + \alpha_i + \delta_t + \epsilon_{it}$$

- Main results:

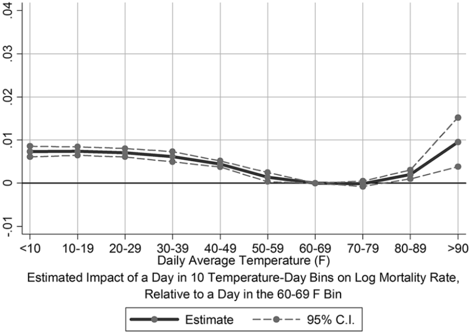


Figure: Barreca et al. (2016)

Climate regressions: Regressions with climate data

2. Regressions with this data

$$y_{it} = f(T_{it}) + g(P_{it}) + \alpha_i + \delta_t + \epsilon_{it}$$

- Main results:

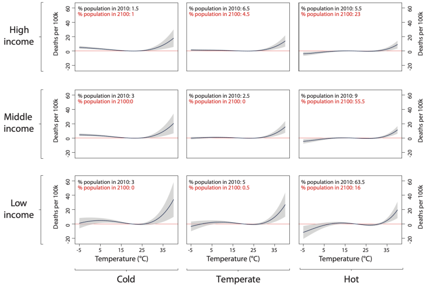


FIGURE I

Heterogeneity in the Mortality-Temperature Relationship (Age > 64 Mortality Rate)

Figure: Carleton et al. (2022)

Climate regressions: Regressions with climate data

2. Regressions with this data

$$y_{it} = f(T_{it}) + g(P_{it}) + \alpha_i + \delta_t + \epsilon_{it}$$

- Main results:

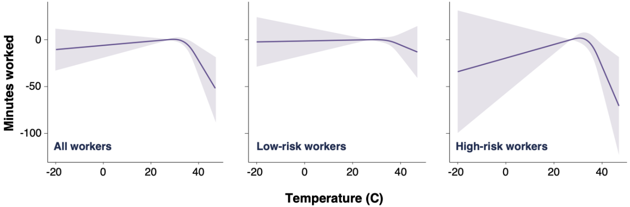


Figure 3: Changes in weekly minutes worked per person due to daily temperature. Labor supply-temperature response functions are estimated for all workers (left), low-risk workers (middle), and high-risk workers (right), corresponding to Columns 1, 2, and 3 in Table 1, respectively. Points along each curve represent the effect on weekly labor supply of a single day at the daily maximum temperature value shown on the x-axis, relative to a day with a maximum temperature of 27°C (81°F). Shaded areas indicate 95% confidence intervals.

Figure: Rode et al. (2022)

Outline

Workshop outline

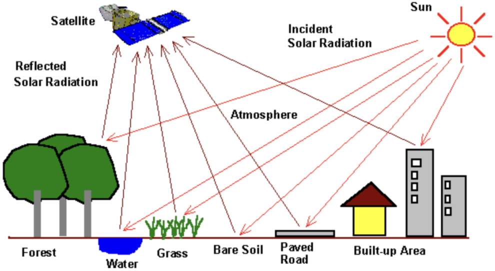
Fixed effects

Climate regressions

Remote sensing

Remote sensing: Super brief introduction

- **Remote sensing**: science of obtaining information about objects or areas from a distance, usually using satellites or aircraft
  - *Passive remote sensing*: Respond to external stimuli and record natural energy that is reflected or emitted from the Earth's surface (e.g., sunlight)
    - *Optical remote sensing*: Use visible, near infrared and short-waveinfrared sensors to form images of the Earth's surface → solar radiation is reflected from targets on the ground → different materials reflect and absorb differently at different wavelengths
  - *Active remote sensing*: Use internal stimuli to collect data (e.g., laser beams)



Remote sensing: (Some) available data and resources

- (Some) public satellite data used in social sciences:
  - **MODIS** (2000-): 250m, 500m, 1km spatial resolution, 1-2 day revisit, 36 spectral bands, many uses and products, including vegetation, ocean color, burned area, etc.
  - **Landsat** (1975-): 30m spatial resolution, 16 day revisit, multispectral
  - **Sentinel-2** (2016-): 10m spatial resolution, 3-5 day revisit, multispectral
  - **TROPOMI**, Sentinel-5P (2018-): $\sim 7 \times 3.5$km spatial resolution, daily revisit, pollution data, including CO, formaldehyde, NO2, ozone, SO2, methane, aerosol layer
  - **VIIRS** (2000/2012-): 375m-1km, $\sim$2x a day, active fire detections (link)
  - **NAIP** (2002-): High resolution (0.6m); growing season in cont. US; RGB + near-IR

- **Google Earth Engine** makes processing of satellite data much much easier; in my opinion worth learning if you're interested in satellite data (ChatGPT pretty helpful!)
  - Geo For Good (annual conference) has very helpful videos and scripts for getting started (e.g., Introduction to Google Earth Engine in Python)
  - `geemap` very helpful for working in Python
  - Datasets available in GEE fun to browse (and get ideas?)

Remote sensing: Causal inference

- Remote sensing measures are indirect measures (e.g., forest cover, illegal mining, land use) that often exhibit substantial (non-classical) measurement error
  - → Bias when used in downstream regressions
  - Despite this, common to use satellite-based measures without any correction as dependent or independent variable

- Recent and growing literature documenting non-random measurement error in machine learning / remotely sensed data (Balboni et al. 2023; Bluhm and McCord 2022; Fowlie et al. 2019)

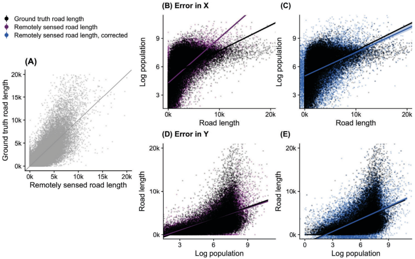
- Today focus three working or recent papers that address this issue

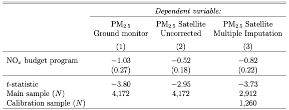
Remote sensing: Causal inference

1. "Remotely Incorrect? Accounting for Nonclassical Measurement Error in Satellite Data on Deforestation" by Alix-García and Millimet (2023), *JAERE*
   - Binary measurement of deforestation (land use change)
   - May suffer from nonclassical measurement error
     - Finds that deforestation measures may under-report true extent of deforestation (but false positive rate is low)
   - Compares two satellite-based deforestation measures in Mexico in the same area
     - Binary measures diverge for 18% of the samples and are correlated with environmental and sensor attributes
   - Proposes misclassification model: researchers specify the variables that may cause measurement error (e.g., cloud cover/reflectance angle) → estimate coefficients of interest & error rates simultaneously using a maximum-likelihood estimator
     - Requires that researchers take a stand on variables that determine measurement error

Remote sensing: Causal inference

2. “Parameter Recovery Using Remotely Sensed Variables” by Proctor et al. (2023), *WP*
   - Shows substantial bias in point estimates using satellite data
     - → Bias mostly due to differential measurement error (correlations between errors in one variable and levels of another variable)
   - Proposes **multiple imputation** procedure to correct for this
     - Need some ground truth data (calibration dataset), which allows researcher to estimate the structure of measurement error
   - Use-case: Using remotely sensed data (not creating it) & access to some ground truth data





	Dependent variable:		
	$PM_{2.5}$ Ground monitor	$PM_{2.5}$ Satellite Uncorrected	$PM_{2.5}$ Satellite Multiple Imputation
	(1)	(2)	(3)
NO_x budget program	−1.03 (0.27)	−0.52 (0.18)	−0.82 (0.22)
t-statistic	−3.80	−2.95	−3.73
Main sample (N)	4,172	4,172	2,912
Calibration sample (N)			1,260

Table 1: Replication of Deschenes, Greenstone and Shapiro (2017) using remotely sensed air pollution, with and without correction via multiple imputation. Column (1) shows the paper’s original estimate and standard errors of the effect of the NO_x budget program on ambient $PM_{2.5}$. Standard errors and t-statistic are calculated from the full original dataset, where standard errors are clustered at the state-season level. Column (2) shows the same estimate using uncorrected satellite $PM_{2.5}$ data from Van Donkelaar et al. (2021) in place of ground monitor data. Standard errors and t-statistic are computed identically to column (1). Column (3) shows point estimates and standard errors corrected using multiple imputation, where 30% of the data was used as a calibration dataset. The 70/30 split of the data sample is done with 200 bootstrap samples and parameters shown (including standard errors and the t-statistic) reflect means across this distribution of bootstrap samples (median estimates are nearly identical).

Remote sensing: Causal inference

3. "Remote Control: Debiasing Remote Sensing Predictions for Causal Inference" by Gordon et al. (2023)
   - Machine learning predictions used on remotely sensed data can produce biased estimates when used for causal analysis
     - Consider the case where remotely sensed data is used as the outcome variable
     - Bias when the measurement error in *outcome* variable is correlated with the *treatment* variable
   - Use **adversarial debiasing algorithms** to correct for bias when generating ML predictions
     - Two models working together (primary model attempts to minimize prediction error), while adversary tries to minimize predicting treatment status
   - Use-case: Address measurement error while making the machine learning estimates, *prior* to estimation steps

Remote sensing: Causal inference

3. "Remote Control: Debiasing Remote Sensing Predictions for Causal Inference" by Gordon et al. (2023)

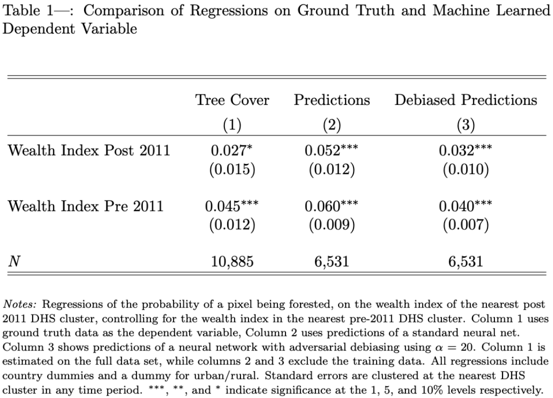


Table 1—: Comparison of Regressions on Ground Truth and Machine Learned Dependent Variable

	Tree Cover	Predictions	Debiased Predictions
	(1)	(2)	(3)
Wealth Index Post 2011	0.027*	0.052***	0.032***
	(0.015)	(0.012)	(0.010)
Wealth Index Pre 2011	0.045***	0.060***	0.040***
	(0.012)	(0.009)	(0.007)
N	10,885	6,531	6,531

Notes: Regressions of the probability of a pixel being forested, on the wealth index of the nearest post 2011 DHS cluster, controlling for the wealth index in the nearest pre-2011 DHS cluster. Column 1 uses ground truth data as the dependent variable, Column 2 uses predictions of a standard neural net. Column 3 shows predictions of a neural network with adversarial debiasing using $\alpha = 20$. Column 1 is estimated on the full data set, while columns 2 and 3 exclude the training data. All regressions include country dummies and a dummy for urban/rural. Standard errors are clustered at the nearest DHS cluster in any time period. ***, **, and * indicate significance at the 1, 5, and 10% levels respectively.

Questions? Comments?

Thank you!

References I

Alix-García, Jennifer, and Daniel L. Millimet. 2023. "Remotely Incorrect? Accounting for Nonclassical Measurement Error in Satellite Data on Deforestation." *Journal of the Association of Environmental and Resource Economists* 10 (5): 1335–1367. https://doi.org/10.1086/723723. eprint: https://doi.org/10.1086/723723. https://doi.org/10.1086/723723.

Balboni, Clare, Aaron Berman, Robin Burgess, and Benjamin A. Olken. 2023. "The Economics of Tropical Deforestation." *Annual Review of Economics* 15 (Volume 15, 2023): 723–754. ISSN: 1941-1391. https://doi.org/https://doi.org/10.1146/annurev-economics-090622-024705. https://www.annualreviews.org/content/journals/10.1146/annurev-economics-090622-024705.

Barreca, Alan, Karen Clay, Olivier Deschenes, Michael Greenstone, and Joseph S. Shapiro. 2016. "Adapting to Climate Change: The Remarkable Decline in the US Temperature-Mortality Relationship over the Twentieth Century." *Journal of Political Economy* 124 (1): 105–159. ISSN: 0022-3808, 1537-534X, accessed April 3, 2024. https://doi.org/10.1086/684582. https://www.journals.uchicago.edu/doi/10.1086/684582.

Bluhm, Richard, and Gordon C. McCord. 2022. "What Can We Learn from Nighttime Lights for Small Geographies? Measurement Errors and Heterogeneous Elasticities." *Remote Sensing* 14 (5). ISSN: 2072-4292. https://doi.org/10.3390/rs14051190. https://www.mdpi.com/2072-4292/14/5/1190.

Carleton, Tamma, Amir Jina, Michael Delgado, Michael Greenstone, Trevor Houser, Solomon Hsiang, Andrew Hultgren, et al. 2022. "Valuing the Global Mortality Consequences of Climate Change Accounting for Adaptation Costs and Benefits." *The Quarterly Journal of Economics* 137 (4): 2037–2105. ISSN: 0033-5533. https://doi.org/10.1093/qje/qjac020. eprint: https://academic.oup.com/qje/article-pdf/137/4/2037/51054116/qjac020.pdf. https://doi.org/10.1093/qje/qjac020.

References II

Chaisemartin, Clément de, and Xavier D'Haultfœuille. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." *The American Economic Review* 110 (9): pp. 2964–2996. ISSN: 00028282, 19447981, accessed February 15, 2024. https://www.jstor.org/stable/26966322.

Fowlie, Meredith, Edward Rubin, and Reed Walker. 2019. "Bringing Satellite-Based Air Quality Estimates Down to Earth." *AEA Papers and Proceedings* 109:283–88. https://doi.org/10.1257/pandp.20191064. https://www.aeaweb.org/articles?id=10.1257/pandp.20191064.

Gordon, Matthew, Megan Ayers, Eliana Stone, and Luke Sanford. 2023. *Remote Control: Debiasing Remote Sensing Predictions for Causal Inference.* Working Paper.

Proctor, Jonathan, Tamma Carleton, and Sandy Sum. 2023. *Parameter Recovery Using Remotely Sensed Variables.* Working Paper, Working Paper Series 30861. National Bureau of Economic Research. https://doi.org/10.3386/w30861. http://www.nber.org/papers/w30861.

Rode, Ashwin, Rachel E. Baker, Tamma Carleton, Anthony D'Agostino, Michael Delgado, Timothy Foreman, Diana R. Gergel, et al. 2022. *Labor Disutility in a Warmer World: The Impact of Climate Change on the Global Workforce.* Working Paper, SSRN Working Paper 4221478.