

Essays in Environmental and Labor Economics

by

Andrew John Dickinson

A dissertation accepted and approved in partial fulfillment of the
requirements for the degree of
Ph.D.
in Economics

Dissertation Committee:

Glen Waddell, Co-Chair

Ed Rubin, Co-Chair

Grant McDermott, Core Member

Jake Searcy, Institutional Representative

University of Oregon

June 2025

© 2025 Andrew John Dickinson
All rights reserved.

DISSERTATION ABSTRACT

Andrew John Dickinson

Ph.D. in Economics

Title: Essays in Labor and Environmental Economics

This dissertation includes previously published and unpublished co-authored material. This dissertation consists of four essays that exploring two topics: a study on the effect of exogenous time shocks on labor productivity; and a study on the reliability of air pollution estimates from machine learning models, an empirical application this work using predicted fine particulate matter ($PM_{2.5}$), and a reproducible database of $PM_{2.5}$ concentrations across the contiguous United States (CONUS).

In the first substantive chapter, using data on GitHub users around the world, I estimate the effects of transitions to Daylight Saving Time (DST) on worker activity. I find that transitions appear short lived at the daily level, with evidence of two days of declines before activity returns to baseline levels. However, hourly analysis reveals a transition to Daylight Saving Time that is much longer, with losses appearing in the early working hours of work days into the second week following the initiation of Daylight Saving Time in the Spring.

In the second substantive chapter, I examine spatial predictions of fine particulate matter ($PM_{2.5}$). The accuracy and reliability of these predictions depend crucially on methodological choices. I examine these choices across the contiguous United States from 2002-2019. My analysis investigates how modeling and validation decisions affect (1) predictions' perceived accuracy for different tasks, (2) predictions' suitability throughout space, and (3) results in downstream

empirical analyses. I evaluate model performance across conventional and spatially explicit cross-validation procedures, demonstrating that the standard validation methods substantially overstate predictive accuracy due to data leakage. I estimate prediction uncertainty bounds to better understand confidence in the predictions and propose a novel monitor-presence probability model to subset predicted $\text{PM}_{2.5}$ to areas most similar to the training data, providing a more reliable set of estimates for empirical analyses. I conclude that there is no one-size-fits-all approach to modeling $\text{PM}_{2.5}$ and the methodological decisions to generate $\text{PM}_{2.5}$ predictions should be tailored to the specific application.

In my third substantive chapter, I revisit recent work evaluating the effect of the Clean Air Act's $\text{PM}_{2.5}$ nonattainment designations. Building on the methodological critiques from Chapter 2, I replicate and extend the Difference-in-Differences (DiD) analysis from a recent published paper on the subject, using both original and revised predicted $\text{PM}_{2.5}$ datasets. Using a binary classification model, I introduce a novel propensity-weighted DiD framework that reweights observations by their similarity to monitored locations, addressing monitor placement endogeneity. This approach reduces the size of estimated treatment effects and in one case, reverses the sign of the estimated effect. I conclude that the using predicted- $\text{PM}_{2.5}$ data to estimate the impact of regulation on pollution levels remains sensitive both measurement errors in the predictive data source and the endogenous placement of air quality monitors.

Following my second and third chapter, my fourth substantive chapter contributes to the growing $\text{PM}_{2.5}$ prediction literature by developing a reproducible, high-resolution database of $\text{PM}_{2.5}$ concentrations across the contiguous United States (CONUS) from 2002 to 2019. The code repository

includes all steps necessary to reproduce the results in this dissertation, including data collection, model training, and prediction generation. I provide a detailed description of the code repository's structure and functionality, allowing future researchers to adapt the PM_{2.5} predictions for their specific applications.

CURRICULUM VITAE

NAME OF AUTHOR: Andrew John Dickinson

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR, USA
San Diego State University, San Diego, CA, USA

DEGREES AWARDED:

Doctor of Philosophy, Economics, 2025 (expected), University of Oregon
Master of Science, Economics, 2021, University of Oregon
Master of Arts, Economics, 2019, San Diego State University
Bachelor of Arts, Economics, 2017, San Diego State University

AREAS OF SPECIAL INTEREST:

Environmental Economics
Applied Econometrics
Causal Inference and Machine Learning

PROFESSIONAL EXPERIENCE:

Research Assistant, Professor Alfredo Burlando, University of Oregon,
2023–2024
Instructor, Introduction to Econometrics I, University of Oregon, 2023–
2024
Instructor, Urban and Regional Economic Problems, University of Oregon,
2021–2022
Research Fellow, Center for Health Economics and Policy Studies, San
Diego State University, 2017–2019

GRANTS, AWARDS AND HONORS:

Kleinsorge Fellowship, University of Oregon, 2023

Core Exam Completion Award, University of Oregon, 2020
Graduate Teaching Fellowship, University of Oregon, 2019–present
McCuen Endowed Economics Scholarship, San Diego State University,
2017, 2018
Center for Public Economics Paper Award, San Diego State University, 2018

PUBLICATIONS:

Dickinson, A. J., & Waddell, G. R. (2024). Productivity losses in the transition to Daylight Saving Time: Evidence from hourly GitHub activity. *Journal of Economic Behavior & Organization*.

Churchill, B. F., Dickinson, A. J., Mackay, T., & Sabia, J. J. (2021). The Effect of E-Verify Laws on Crime. *Industrial and Labor Relations Review*.

TABLE OF CONTENTS

Chapter	Page
I. PRODUCTIVITY LOSSES IN THE TRANSITION TO DAYLIGHT SAVING TIME: EVIDENCE FROM HOURLY GITHUB ACTIVITY	14
1.1. Introduction	14
1.2. Data	18
1.2.1. Sample selection	18
1.2.2. Outcomes	20
1.2.3. Treatment	20
1.3. Results	21
1.3.1. Day-level analysis	22
1.3.2. Hour-level analysis	25
1.4. Conclusion	28
II. HOW GOOD ARE POLLUTION PREDICTIONS?	35
2.1. Introduction	35
2.2. Data and Methods	39
2.2.1. Point Observations	39
2.2.2. Model Inputs	40
2.2.3. Model Training	43
2.2.4. Assessing Model Performance	45
2.3. Empirical Application	49
2.4. Results	50
2.4.1. Model Performance	50
2.4.2. Uncertainty in Predictions	51

Chapter	Page
2.4.3. Compliance Assessment of Census Tracts	53
2.5. Discussion	54
2.6. Conclusion	56
2.7. Figures and Tables	57
III. EVALUATING THE IMPACT OF NONATTAINMENT DESIGNATIONS UNDER THE CLEAN AIR ACT	67
3.1. Introduction	67
3.2. Data and Methodology	69
3.2.1. Data	70
3.2.2. Empirical Methodology	71
3.2.3. Predicting Monitor Placement	72
3.3. Results	73
3.4. Conclusion	75
3.5. Figures and Tables	76
IV. HIGH RESOLUTION GEOSPATIAL DATABASE OF PM _{2.5} CONCENTRATIONS IN THE CONTIGUOUS UNITED STATES (2002–2019)	81
4.1. Introduction	81
4.2. Data and Methods	83
4.2.1. Monitoring Data	83
4.2.2. Covariates	83
4.2.3. Model Training	85
4.2.4. Model Validation	85
4.3. Results	86
4.3.1. Model Performance	86
4.3.2. Prediction Uncertainty	86

Chapter	Page
4.3.3. Dataset and Repository Overview	87
4.4. Conclusion	88
4.5. Figures	89
APPENDICES	
A. APPENDIX TO CHAPTER ONE	91
B. APPENDIX TO CHAPTER TWO	101
REFERENCES CITED	104

LIST OF FIGURES

Figure	Page
1. Mean number of GitHub events, by treatment status	30
2. Does DST induce a change in the total number of daily GitHub events?	31
3. Do we see different responses within working hours, or if we restrict events to “pushes” only?	32
4. Hour-specific estimates of the effect of DST and return to standard time on GitHub events	33
5. Does the pattern of hourly events change around DST?	34
6. PM _{2.5} monitoring sites across the CONUS in June 2011.	58
7. Cross validation in temporally repeated grids	59
8. Nested cross validation in temporally repeated grids	60
9. PM _{2.5} prediction accuracy declines steeply when spatially validated and/or restricted from using <i>close</i> spatial lags.	61
10. Out-of-sample PM _{2.5} prediction accuracy	62
11. Attainment Status by Design Value Rank	63
12. Pixel Design Values, uncertainty, and attainment.	64
13. PM _{2.5} Design Values and nonattainment counties	77
14. PM _{2.5}	78
15. Event Study of the CAA on PM _{2.5} Levels	79
16. Stacked Raster of Monthly Average PM _{2.5} Predictions	90
A.17. Constructing hourly difference-in-differences estimates	95
A.18. Do we see different responses across North America and Europe? . . .	96
A.19. Daily estimates of the effect of DST and the return to standard time on the total number of daily GitHub events?	97

Figure	Page
A.20.Hour-specific estimates of the effect of DST on GitHub events	98
A.21.Hour-specific estimates of the fall return to standard time	99
A.22.Additional GitHub events, and the declines in activity induced by DST	100
B.1. Changes in PM _{2.5} prediction accuracy as the number of spatial cross-validation splits increases.	102
B.2. In-sample PM _{2.5} prediction accuracy.	103

LIST OF TABLES

Table	Page
1. Recent papers that produce PM estimates largely report high cross-validated accuracy, <i>i.e.</i> , R^2 above .8.	65
2. Summary of Consumer Studies	66
3. Nonattainment Status and Changes in PM _{2.5}	80
A.4. GitHub events with proportional breakdowns	92
A.5. 50 most-active GitHub cities (ranked by total events) in the Northern Hemisphere in 2013–2019	93
A.6. Daylight Saving Time across treated locations, 2013–2019	94

CHAPTER I

PRODUCTIVITY LOSSES IN THE TRANSITION TO DAYLIGHT SAVING TIME: EVIDENCE FROM HOURLY GITHUB ACTIVITY

This dissertation includes both unpublished and published co-authored material. Chapter 1 is co-authored with Glen Waddell and has published in the *Journal of Economic Behavior and Organization*, we both contributed to statistical analysis and writing of the manuscript. Chapter 2 is unpublished, co-authored with Ed Rubin. We both contributed to the writing and statistical analysis of the manuscript.

1.1 Introduction

Motivated by potential energy savings, Daylight Saving Time (DST) was introduced in 1916 and widely adopted by many Western countries soon after. The shifting forward of the clock increased the availability of natural light in the evenings—at the time, this lessened the need for the carbon filaments and kerosene associated with indoor lighting (Nordhaus, 1997). Today, the consensus opinion with respect to energy consumption is that any savings associated with DST are likely small (Belzer, Hadley, & Chin, 2008; Kotchen & Grant, 2011)—they are on the order of plus-or-minus one percent. The merits of continuing these twice-yearly adjustments to the clocks of 1.6 billion people across 75 countries are therefore important contributors to the continuing debate, even as countries abandon the practice.¹ Most recently, the United States (US) Senate unanimously

¹ China first abandoned the practice in 1991, and several other countries following that path in the mid-to-late 2000s. Other countries discontinuing DST include Pakistan (2009), Russia (2010), Azerbaijan (2015), and Brazil (2019).

passed the Sunshine Protection Act in 2021, which would have the United States likewise abandoning the practice.²

While the Sunshine Protection Act remains in the US House of Representatives, of note is the House Energy and Commerce Committee’s interest in learning of the productivity effects of time changes prior to holding a vote. Our analysis contributes directly to the debate, then, and with immediate implication for policy. To that end, we document significant declines in worker activity following DST, with patterns that suggest a weeks-long period of transition through which workers are adjusting to the shifting of time relative to the cycle of daylight and darkness.

To measure the potential declines in activity induced by DST we exploit publicly available activity data from GitHub, a popular cloud based version-control platform used by collaborative programmers. Observations include detailed information regarding changes to a project’s codebase—including what was changed and who made the change—providing a direct measurement of labor activity. Of particular value, though, is that we observe GitHub user activity in second-level precision. This level of granularity makes GitHub a natural laboratory to consider the effects of DST. In our analysis, we consider activity aggregated to both the daily and hourly levels across cities that experience the Spring transitions to DST differently—those that experience it earlier than others, experience it later than others, or do not experience it at all.

In daily activity we find significant declines on the order of one or two days. While such declines are presumably associated with significant economic costs, this alone suggests that the productivity losses are relatively short-

² This national-level action follows several state-level initiatives that similarly moved toward the permanent adoption of DST. For example, Oregon (2019) and Washington (2019) have passed such legislation, and California’s Proposition 7 (2018) was approved by over 60 percent of voters.

lived. However, in hourly activity we find much longer periods of transition. In particular, in the two weeks following transitions to DST we find significant declines in worker activity in early work hours, between 8am and 10am. That daily measures of productivity evidence significant declines for such a short period of time suggests that the slow morning hours in the data are made up for with increases in activity elsewhere in the work day. However, increases in activity tend not to cluster in particular hours of the day. Rather, the data are consistent with workers recovering from morning productivity declines without a common strategy for dealing with the transition. In the end, it is clear from an hourly analysis that the belief in “one or two days of decline” likely fails to capture the full economic, personal, or potential social costs induced by DST.

One of the primary mechanisms highlighted by the literature to explain the productivity declines around DST changes is the sleep deprivation experienced by workers—and there is an established relationship between sleep and outcomes. For example, (Gibson & Shrader, 2018) use variation in sunset times (e.g., due to time-zone differences) to suggest that more natural light reduces sleep and earnings. (Giuntella & Mazzonna, 2019) also suggests that natural light is correlated with health outcomes that are related to circadian rhythm and per-capita income. (Costa-Font, Fleche, & Pagan, 2024a) suggests that weekly sleep is positively correlated with employment and weekly earnings. Among miners working in the United States, the one-hour time change induced by DST has been associated with decreases in sleep on the order of 40 minutes (Barnes & Wagner, 2009), making sleep loss a contending mechanism in explaining reductions in activity around DST.

Circadian rhythms—that is, the processes that synchronize and regulate the body’s sleep-wake cycle around the 24-hour day—are resilient to adjustments to

sleep-wake cycles, and the transition to a modified cycle of daylight and darkness is not immediate (Kantermann, Juda, Merrow, & Roenneberg, 2007). Thus, that we find long-lasting disruptions to activity are consistent with this resilience. This is also consistent with laboratory results suggest that transitions to DST are measured in days (Czeisler et al., 1999; Monk & Aplin, 1980), and associated with immediate losses of sleep that do not typically return to normal for roughly seven days (Kantermann et al., 2007; T. A. Lahti, Leppämäki, Lönnqvist, & Partonen, 2006). Workplace safety (Barnes & Wagner, 2009; T. Lahti, Sysi-Aho, Haukka, & Partonen, 2011), road safety (Bünnings & Schiele, 2021; Smith, 2016), and student performance (Gaski & Sagarin, 2011) have all been shown to decline following DST. Google searches for entertainment-related keywords and phrases in US cities that experience DST (i.e., those other than in Arizona and Hawaii) are differentially higher on the Monday that immediately follows DST transitions than on the Monday prior to DST (Wagner, Barnes, Lim, & Ferris, 2012). This has been interpreted as DST-induced increases in “cyberloafing” and suggests that declines in workplace productivity can result from DST. However, to our knowledge, documenting worker activity around DST transitions has thus far been absent in the literature.³

While high-frequency observations of activity available on GitHub make for a rare opportunity to learn about such behavior, GitHub users are not representative of the labor force of any country, and certainly not of the world.

³ Somewhat more removed from the research question, DST has also been associated with lower well-being and general life satisfaction (Kountouris & Remoundou, 2014; Kuehnle & Wunder, 2016), increases in myocardial infarction (Toro, Tigre, & Sampaio, 2015), reductions in crime (Doleac & Sanders, 2015), and increases in suicide (Berk et al., 2008). DST has been associated with movements in financial markets, though the evidence is mixed. For example, Kamstra, Kramer, and Levi (2000) suggests that there are negative financial returns following DST weekends, while Gregory-Allen, Jacobsen, and Marquering (2010) finds no evidence of a Daylight Saving Time anomaly.

However, with roughly 83 million users currently, GitHub does account for a large and growing sector of the global labor market. It is also arguable that our sample is representative of a much-larger group of similarly skilled workers who work across industry, government, academic, and policy environments. In many work environments, for example, similarly skilled individuals are working in proprietary tasks that preclude their formal participation in a public code-sharing environment, and in others we would find still more consumers of the public offerings of active GitHub users. In this way, it is reasonable to infer that the effects we identify extend naturally to a larger group of workers.

In Section 1.2 we describe the data we rely on in our analysis and provide the necessary context for the interpretation of GitHub user data as we consider whether there is evidence of a causal relationship between DST and worker activity. In Section 1.3 we discuss our methodology and report event-study analyses at both the daily and hourly level. We summarize and offer concluding remarks in Section 1.4.

1.2 Data

1.2.1 Sample selection. GitHub is a popular cloud based version-control and code-hosting platform. Widely adopted by software developers, engineers, and scientific coders for hosting and maintaining collaborative projects, GitHub is a web based extension to the most popular distributed version control system, Git.⁴ With a reported userbase of over 83 million, GitHub is the largest platform of its kind.

Consistent with its origin as a tool for collaboration with excellent version control, the primary feature of GitHub is the facilitation of code updates across

⁴ Of the 80,000 respondents in the 2021 Stack Overflow Annual Developer Survey, over 90 percent reported using Git.

multiple collaborative users. In essence, it provides broad access and version control to a remote copy of a project and all of its files, and as progress is made to a project hosted on the platform, contributions are tracked across time and contributor, which generates a record of labor activity metadata.

For the purpose of this study, we collect GitHub activity around the DST episodes from 2013 through 2019.⁵ To reflect “active” users we restrict each year’s sample to those we observe activity for in each month of that year. To limit the potential presence of “bot” activity, we discard all observations originating from accounts that ever have more than 30 events in one hour or 150 in one day (McDermott & Hansen, 2021). Following these restrictions, the total number of unique users in the sample is 174,505. Having matched each event in the raw data to the location identified in the user’s profile at the time of the event, we aggregate event counts to the city-by-day level (or city-by-hour level) and restrict the sample to the 50 most-active cities in the Northern Hemisphere. These cities are reported in Table A.5. As expected, the geographic distribution of these workers reflects major cities known for their robust presence in the tech industry. Our geographic restriction to the 50 most-active cities implies coverage of roughly 42 percent (174,505) of all active GitHub users with observable geographic location during the sample period (414,416). GitHub user activity also falls off precipitously. For example, the total activity of the 50 most-active cities represents 82.8 percent of the total activity of the 100 most-active cities. (All results are robust to the inclusion of activity originating in the 100 most-active cities.) Our

⁵ We draw our sample from two third-party sources (GH Archive and GHTorrent) that have organized the event data from GitHub’s public timeline for the purpose of being more accessible to researchers. Our analysis therefore does not include any activity on private repositories. For example, a business producing proprietary software, or a team of data scientists performing analysis on valuable or sensitive data, would host their repositories privately and would not be observed on the public timeline.

econometric models will consider three weeks before and after DST transitions. Thus, for “never-treated” cities we include all observations in the window of three weeks before the earliest treatment and three weeks after the latest treatment.

1.2.2 Outcomes. All activity on GitHub is recorded as an “event,” and categorized as one of 13 types (see Table A.4). In some way, they are each potential measures of incremental productivity on shared projects. However, given the inability to distinguish the relative productivities of event types, broadly, we will consider all events, and then separately consider the 46.6 percent of events that are classified as “pushes.” Push events are uploaded changes of a local file to the remote copy, which are likely stronger signals of productivity. For example, this analysis will exclude “pull requests,” which represent contributors’ requests to merge the code changes of others into their remote copy. In the end, we find comparable patterns in activity around transitions to DST in both measures.

1.2.3 Treatment. While many countries practice DST, there is variation across cities in both the practice of DST and in the timing of DST. Across the 50 cities in our sample, two important sources of treatment variation exist. In North America, clocks transition to DST on the second Sunday of March.⁶ In the European Union, clocks transition on the last Sunday of March. Both of these will contribute to identifying the effect of DST on GitHub activity—given the staggered timing of treatment, they each act as a control for the other, and can each be compared to the “never-treated” cities. We follow Sun and Abraham (2021) in the estimation of treatment effects.⁷

⁶ The exceptions to this in the 2013–2019 period are Hawaii, parts of Arizona, Puerto Rico, US territories in the Pacific Islands, Saskatchewan, and the Yukon. None of these are among the 50 most-active cities, and are therefore not in our sample.

⁷ In supplementary material (Figure A.18) we produce estimates separately for users in North American and European Union cities with users in “never-treated” cities as controls—results are robust to this distinction.

In Figure 1 we plot patterns of GitHub events around DST transitions. In Panel A we show daily events in the three weeks before and after DST and aggregate across sample years, separating users by whether they are in North America, in the EU, or in locations that do not experience DST. A large portion of GitHub activity originates in recognizable technology centers, many of which are located in the United States. Thus, activity in North America is higher across the sample period. This is also evident in Panel B, where we plot hourly events. As a general rule, activity is also higher during traditional working hours of 9am to 5pm, Monday through Friday—a dip in activity during traditional lunch hours is also apparent.

1.3 Results

To identify the effect of Daylight Saving Time on activity we employ an event-study design. We first consider specifications at the daily level, followed by similarly constructed models at the hourly level. Daily estimates provide insight into the size and duration of any level-decreases in labor activity associated with DST. Hourly estimates, on the other hand, can speak to within-day variation in user activity—this will prove important to understanding the full extent to which transitions to DST are disruptive to productivity, inducing a substitution of activity across hours of the day. In both environments, however, we are interested in identifying the potential changes in activity around the initiation of DST in treatment locations—their dynamics, in particular. In related literatures that assess impacts of DST transitions on outcomes, much of the identifying variation arises across time. For example, control groups often consist of only parts of Arizona and Hawaii. In such cases, treatment is often captured by estimating a regression discontinuity in time (RDiT) (Bünnings & Schiele, 2021; Doleac & Sanders, 2015; Smith, 2016; Toro et al., 2015). Given the worldwide nature of

GitHub, however, there is considerable cross-sectional variation available to us in estimating the change in activity induced by treatment—our modeling will include both variation in the timing of treatment among those who experience DST and and a “never treated” group where GitHub activity should be unperturbed by DST.

1.3.1 Day-level analysis. In estimating the effect of DST on daily activity we estimate

$$\ln(\text{Events}_{cyd}) = \alpha + \sum_{d=-21}^{-2} \beta_d + \sum_{d=0}^{21} \beta_d \text{DST}_{cyd} + \psi_c + \lambda_y + \delta_d + X_{cyd} + u_{cyd}, \quad (1.1)$$

where $\ln(\text{Events}_{cyd})$ is the log number of events in city c in year y on day d . With the day immediately before DST as the excluded category, this allows for the estimated $\hat{\beta}_d$ to be interpreted as percentage differences relative to the day before DST. In estimating (1.1) we also include city fixed effects (ψ_c) to absorb any level differences in average events across cities, day-of-year (δ_d) fixed effects to absorb differences over time and work-week seasonality, and year fixed effects (λ_y) to absorb any average difference in activity across years in our sample. In X_{cyd} we also include controls for day-of-the-week and for national holidays, capturing variation in activity that should not be attributed to DST when likely due to holidays that coincide with Sunday transitions.⁸ Errors are captured in u_{cyd} . Given the staggered timing of DST, we follow Sun and Abraham (2021), though the qualitative results do not vary with this consideration.

In Figure 2 we report estimates of (1.1) for a pooled model (i.e., across all days). However, patterns in Figure 1 suggest that day-of-week effects may

⁸ As Easter coincides with treatment in the European Union in 2013 and in 2016, we also differentiate Easter from other holidays. For example, Presidents’ Day falls within the three-week sample in six out of the seven years of the US cities. (In 2016, President’s Day is 27 days prior to DST.)

vary by city—some cities exhibit more or less weekend activity, for example. We are inclined, then, to allow day-of-week effects to vary by city. This is empirically challenging, however, as DST falls on Sundays, which precludes the estimation of separate day-of-week effects by city. To allow for the same flexibility, then, we estimate separate models for each day of the week, which we report in Panel B. (Stratifying the sample in this way allows the city fixed effect to absorb any city-specific day-of-week effects.) For ease of interpretation, we present each of the seven models together in Panel B. With each model using the same day in the week prior as the omitted category, the results are interpreted as percentage differences relative to the same day (e.g., Monday, Tuesday) the week before the initiation of DST.

In the end, in daily measures of activity we find a short-lived decline associated with DST. It is only the first Sunday of DST itself and the Monday following DST for which we can distinguish activity from baseline levels. That said, these effects are economically meaningful, suggesting reductions in the number of events on the public GitHub timeline on the order of 10.6 percent on the Sunday ($p < 0.001$, $\sigma = 0.11$) and 5.4 percent on the Monday following treatment ($p = 0.033$, $\sigma = 0.06$).^{9,10} While longer-lasting effects are statistically indistinguishable from zero, we note a distinct “Sunday effect” in the data following DST, which we interpret as a response to the extra hour of daylight that is coincident with DST inducing more of a weekend out of users in treated

⁹ To ensure this result is not merely reflecting the mechanical loss of one hour on treated Sundays, we impute the activity of this missing hour by adding the number of events that occurred during the same hour in the same city in the same year on the Sunday immediately prior to DST. Without this imputation, which induces a mechanical decline in treated Sundays, the measurable impact for the first treated Sunday is a reduction in activity by 12.0 percent (instead of 10.6 percent).

¹⁰ Results are robust to dropping all Easter Sundays from the sample.

cities relative to those in control cities. In terms of the hours of typical activity, these DST-induced decreases on Sunday and Monday correspond to the equivalent of losing 91.5 minutes of average Sunday activity and 35.9 minutes of average Monday activity.¹¹

In Figure 3 we restrict our sample in two ways, each having the potential to inform how we interpret the systematic relationship between DST and GitHub activity. In Panel A we consider events that occur between 9am and 5pm—these are the most active hours of the day, accounting for 58.8 percent of activity. Under this restriction we find a similar pattern—a 14.5-percent reduction in the number of events on Sunday ($p < 0.001$, $\sigma = 0.16$) and 7.1-percent reduction on the Monday following treatment ($p = 0.058$). In Panel B we restrict observations to only “pushes.”¹² Again, we find similar patterns in the number of events—declines of 13.3 percent on Sunday ($p < 0.001$, $\sigma = 0.14$) and 5.2 percent on the Monday following treatment ($p = 0.020$, $\sigma = 0.06$).¹³

¹¹ As another point of comparison, we also note that when controlling for city, year, and day of week, GitHub activity is 60.93-percent lower on Christmas day, suggesting that the estimated declines we find on Sunday and Monday correspond to roughly 17.36- and 8.84-percent of the declines experienced on this widely celebrated international holiday.

¹² Recall that “pushes” are among all events in the earlier analysis, but arguably represent the type of event type that we would think is most representative of productivity as they encompass all changes uploaded to a remote copy.

¹³ In Appendix Figure A.22 we report results of the same models applied to all other event types recorded by GitHub. Recall that the event most clearly interpretable with productivity is likely “pushes,” which are uploaded changes of a local file to the remote copy. However, across all thirteen categories (pushes, and the additional 12 reported in Figure A.22) we see similar directional responses to the DST transition. We interpret this as evidence against an appeal to a type of work simply being displaced for other types of work. That said, this inference should also be conditioned, as we do not have evidence of work activity outside of the GitHub environment, which may or may not change.

1.3.2 Hour-level analysis. To formally model *hourly* GitHub events around transitions to DST we estimate

$$\ln(\text{Events}_{cydh}) = \alpha + \sum_{d=-21}^{-2} \beta_d \text{DST}_{cydh} + \sum_{d=0}^{21} \beta_{dh} \text{DST}_{cydh} + \psi_c + \lambda_y + \delta_d + \sigma_h + X_{cyd} + u_{cydh}, \quad (1.2)$$

with $\ln(\text{Events}_{cydh})$ now defined at the city-by-year-by-day-by-hour (h). In daily activity, DST always falling on Sundays prohibited the estimating of separate day-of-week effects by city. To flexibly fit hour effects in (1.2) we model each hour of the day separately for each day of the week.¹⁴

In Panel A of Figure 4 we present six plots, each corresponding to an hour of the day between 7am and 12pm—each plot then allows for the across-time comparison of relative productivities in the same hour of the day. (For example, in the 9am panel, point estimates inform the question “What happens to productivity between 9:00am and 9:59am each day, before and after DST?”) Unlike daily estimates of the spring transition to DST, which identify one to two days of significant decreases in activity, the analysis of hourly data reveals longer-lived declines in activity following DST—this is particularly evident in the traditional morning work hours. Relative to baseline activity, the initiation of DST induces significant declines in the 8am, 9am, and 10am hours even into a second week. (Declines in the 8am hour are distinguishable from zero for 11 days, in the 9am hour for 11 days, and in the 10am hour for 8 days.)

¹⁴ In supplemental material (Figure A.17) we demonstrate the patterns in activity visually, focussing on one week before and after DST transitions. In Panel A we plot weekly mean event counts separately for each group (i.e., cities in North America, in the EU, and those never treated) across each hour of the day. In Panel B we plot the difference in means, subtracting hourly averages by treatment status separately for pre- and post-treatment period. In Panel C we plot the “second difference” in means, subtracting the post-treatment differences by pre-treatment differences (again, separately by group). This suggests declines in GitHub activity in the traditional morning work hours.

In magnitude, these estimated declines are between 32.0 and 46.8 percent in the 11 days following treatment. While there are no significant increases in activity elsewhere in the day, the overall patterns of daily activity are consistent with users adopting a variety of strategies to cope with the transition.¹⁵ Specifically identifying off of DST, (Costa-Font, Fleche, & Pagan, 2024b) suggests that DST increases the feeling of being pressed for time, corroborating the idea that if workers are less productive in the morning they may try to “catch up” the rest of the day.

We suggest that the differences evident in other hours of the day (i.e., outside of the morning hours), and the lack of significance in single hours, in particular, may reflect individual heterogeneity in how people best respond to the morning decline. To the extent the shock to productivity targets morning hours more precisely (where we identify significant departures from baseline) but the response is diffuse, we may not detect specific hourly estimates of the rebound. That said, we cannot rule out that worker responses are offsetting the morning declines, generally, to the extent that we are compelled to see the hourly estimates in the context offered earlier, in the day-level analysis. That is, overall productivity declines are on the order of two days, consistent with enough “offsetting” increases in activity in the afternoons, leaving overall productivity unchanged.

In Panel B of Figure 4 we report estimates of the same models, but on the fall return to standard time. Of note, activity *increases* in the early work hours on the Sunday of the return to standard time—this increase is evident into the 10am hour on the Mondays following the fall transition. While these estimated

¹⁵ In Appendix Figure A.19 we present the daily estimates for the return to standard time. In figures A.20 and A.21 we present the full 24-panel plots for spring and fall, each corresponding to an hour of the day.

effects do not persist in the same way as is evident with the loss of an hour of sleep in the spring, we interpret the absence of decreases in activity in the fall return to standard time as suggestive of sleep as a mechanism in explaining the spring results. For example, this is suggestive of more than merely a confusion or dissonance resulting from the change in calendar time, regardless of direction—this, we imagine, would result in decreased activity whether the hour was lost (spring) or gained (fall). This is also consistent with Jin and Ziebarth (2020), which finds that returns to standard time have been associated with increases in sleep.¹⁶

To synthesize the variation in hourly estimates and better understand the nature of users’ transitions to DST, in Figure 5 we consider whether there are discernible patterns in the spring point estimates from the hourly analysis. Separately for each week, we ask whether the point estimates of all 24 hourly models (in Figure A.20) exhibit any seasonality across hours of the day. On the left of Figure 5, then, we re-align those same point estimates to instead represent the 168 hours in each week leading into and out of DST. On the right, we then present the hourly variation that is identifiable in that week (i.e., one estimated parameter for each hour of the day), omitting 4am as the reference category. This context makes clear that the “shape” of what is normal variation in hourly activity changes with the arrival of DST. While no systematic patterns exists across hour of the day in the weeks prior to DST, with DST’s arrival the systematic declines in activity in the morning work hours are again evident. Immediately following DST, magnitudes of the decline are greatest in the 9am hour, with a 30.5-percent reduction. Further, this pattern attenuates in the second and third weeks following

¹⁶ Interestingly, Jin and Ziebarth (2020) also documents reductions in hospital admissions for cardiovascular diseases and other health issues.

DST, as workers arguably recover from the disruption to time. Reductions remain largest at 9am but reduce in magnitude to 11.0-percent and 6.8-percent in the second and third week following DST.

In considering how reasonable it is to extend to a larger group of workers, we note that to the extent GitHub workers are more time-flexible than average workers, being able to shift their workload more flexibly than may be typical, these measured responses may overestimate the effects in the wider population. That said, even in occupations where shifting the timing of work is more costly, leaving workers less apt to respond, the quality or efficiency of work may still be depressed. For example, (Wagner et al., 2012) suggests that cyberloafing increases with Daylight Saving Time, and (Costa-Font et al., 2024a) suggests that workers are less likely to report being a thorough worker when they experience less sleep. In general, we suggest caution with respect to external validity.

1.4 Conclusion

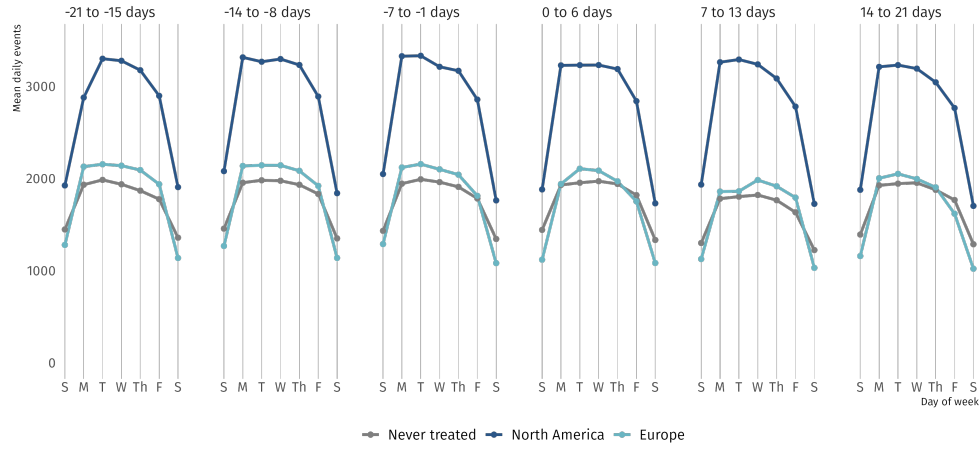
Even though Daylight Saving Time changes the social clocks of over one billion people each year, concerns over the associated economic and social welfare implications continue to be debated. By considering active users of a popular cloud based version-control platform, we exploit variation in the application of DST across technology centers around the world to estimate the productivity implications of shifting clock time this way.

At the daily level, we find significant level declines in the recorded activity of GitHub users who experience DST—on the Sunday of DST itself, and the Monday immediately following the transition to DST. While this suggests that the productivity declines are over quickly, there are two ways that the data support the belief that the transition is more costly to productivity. First, we find persistent “Sunday effects” in treated cities—while daily activity returns

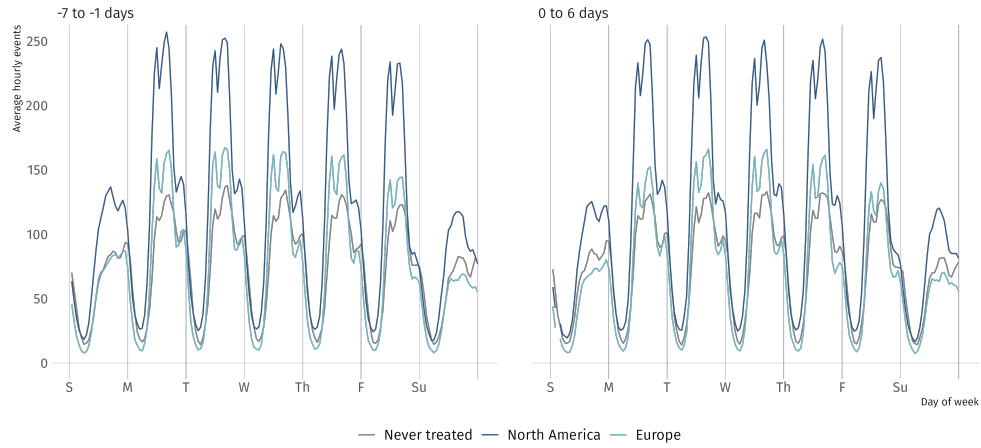
to baseline following two-day declines, GitHub activity is persistently lower on Sundays. Second, hourly analysis reveals much-longer-lasting disruptions—morning declines are evident for upwards of two weeks following the initiation of DST.

Figure 1. Mean number of GitHub events, by treatment status

A: Total events per day, ± 3 weeks to DST

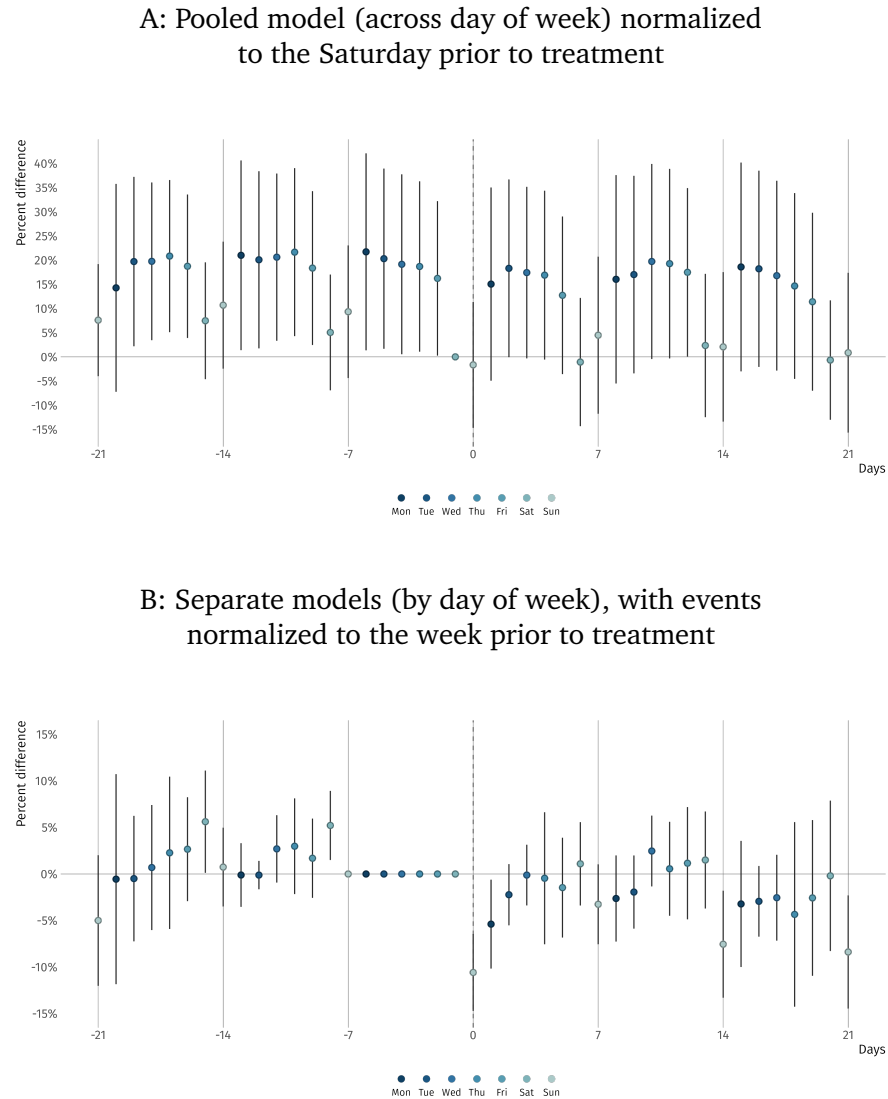


B: Total events per hour, ± 1 weeks to DST



Notes: The sample includes the observations of the total number of GitHub events in the 50 most-active cities, three weeks before and after DST ($n = 16,394$, representing 174,505 unique users). In Panel A we report day-level means. In Panel B we report hourly means (averaged across days) in the week before and after DST. North America includes the treated cities in the United States and Canada. Europe includes the treated cities in the European Union. See Table A.6 for specific dates associated with DST. (In the “-21 to -15” cell of Panel A, the lower level of activity on Monday in North America is attributable to President’s day.)

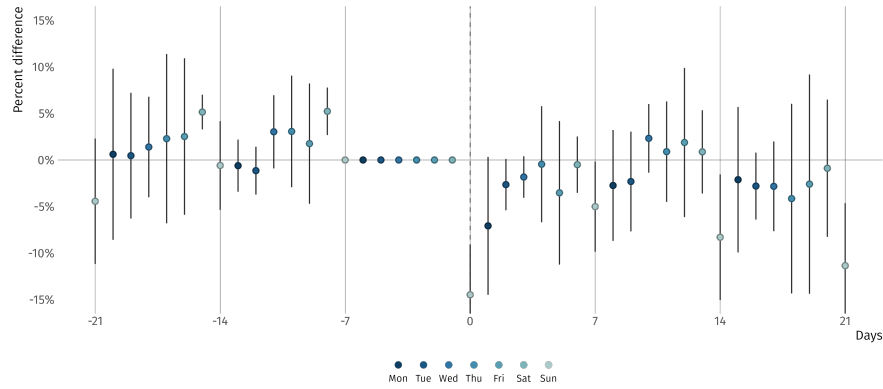
Figure 2. Does DST induce a change in the total number of daily GitHub events?



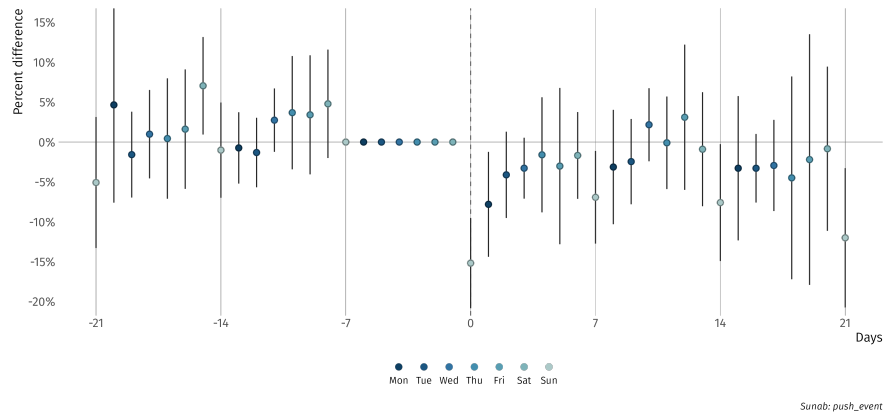
Notes: The sample includes the observations of the total (logged) number of GitHub events three weeks before and after DST, for the 50 most-active cities (representing 16,394 city-day observations of 174,505 unique users). In Panel A we estimate one pooled model of all days, with the Saturday prior to treatment as the omitted category. We estimate city, day-of-week, day-of-year, year, and holiday fixed effects. As day-of-week effects vary by city (e.g., some cities exhibit more or less weekend activity), we are inclined to allow day-of-week effects to vary by city. However, as treatment always falls on a Sunday we cannot estimate separate day-of-week effects by city. In Panel B we therefore estimate seven separate models (i.e., one model for each day of the week) where the city fixed effect then implicitly absorbs any day-of-week differences.

Figure 3. Do we see different responses within working hours, or if we restrict events to “pushes” only?

A: Total events during work hours (9a-7p)

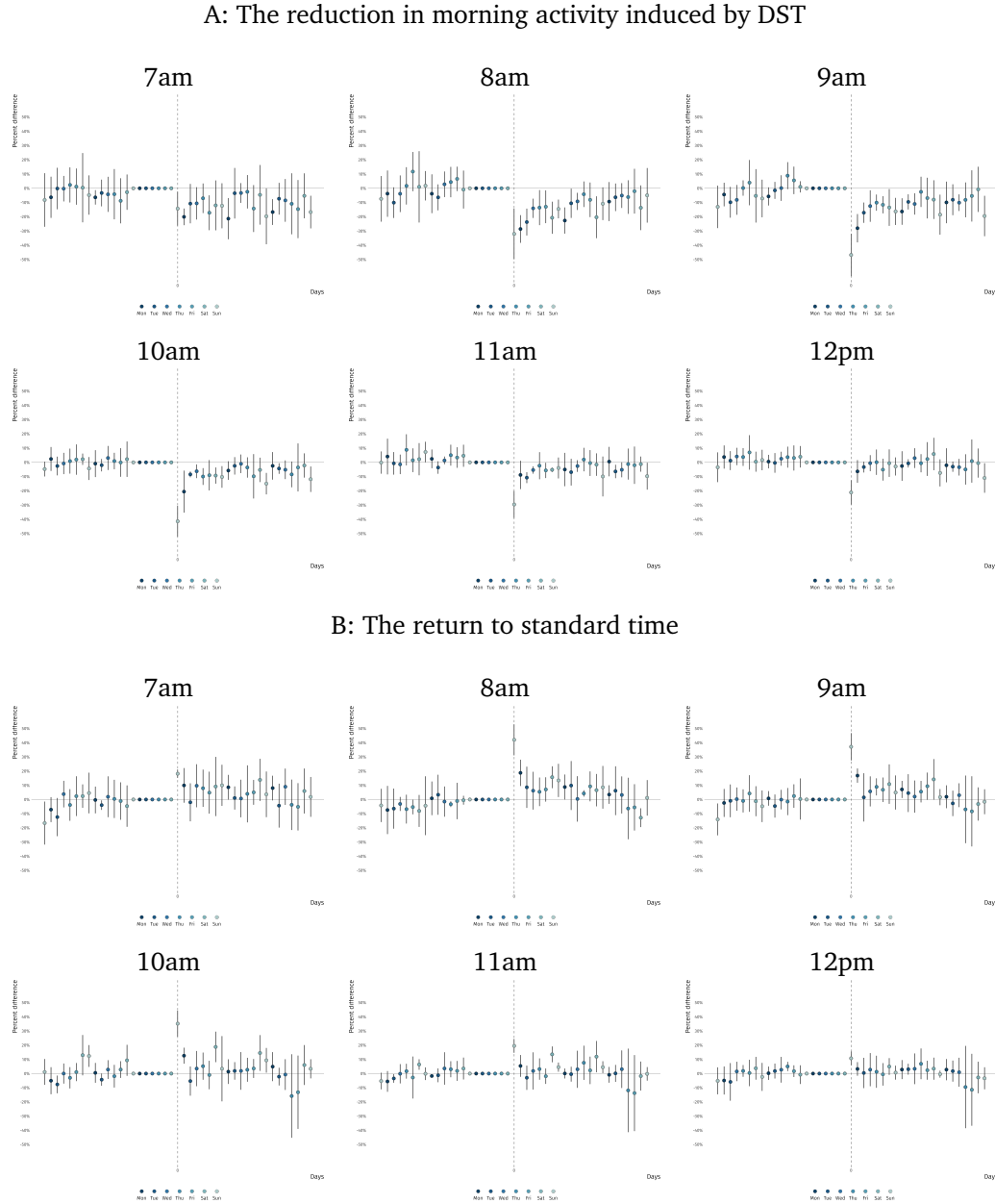


B: Pushes only (all hours)



Notes: The sample includes the observations of the total (logged) number of GitHub events three weeks before and after DST, for the 50 most-active cities (representing 16,394 city-day observations of 174,505 unique users). In both panels we estimate seven separate models (i.e., one model for each day of the week). In Panel A we model the (logged) total number of GitHub events that fall within the hours of 9am to 7pm. In Panel B we model the (logged) number of pushes.

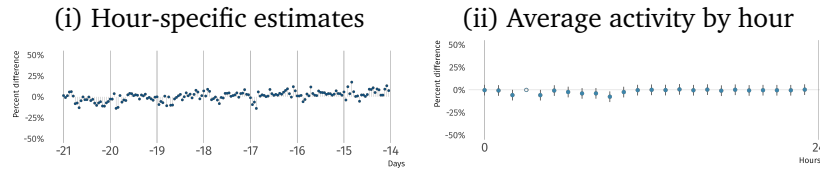
Figure 4. Hour-specific estimates of the effect of DST and return to standard time on GitHub events



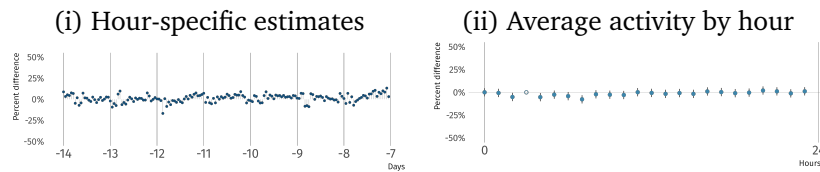
Notes: In both panels, the sample includes the observations of the total (logged) number of GitHub events three weeks before and after the transitions to DST and the return to standard time. In Panel A this represents 393,162 city-hour observations of 174,505 unique users. In Panel B, this represents 378,672 city-hour observations of 174,505 unique users. In both, we estimate 168 separate models (i.e., one model for each hour of the week) where the city fixed effect then implicitly absorbs any hour-of-week differences.

Figure 5. Does the pattern of hourly events change around DST?

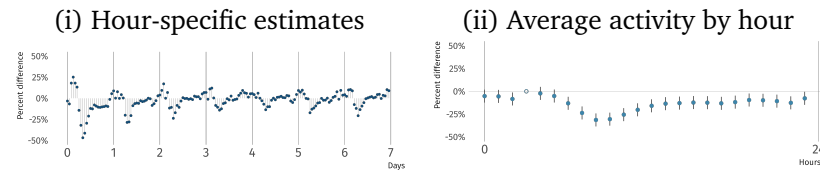
A: -21 to -15 days to DST



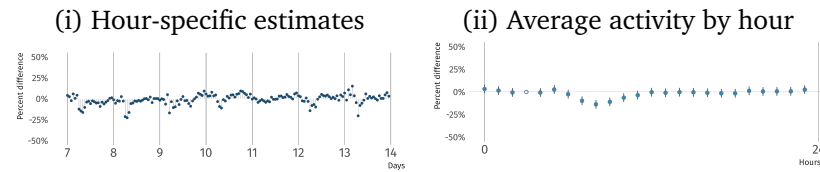
B: -14 to -8 days to DST



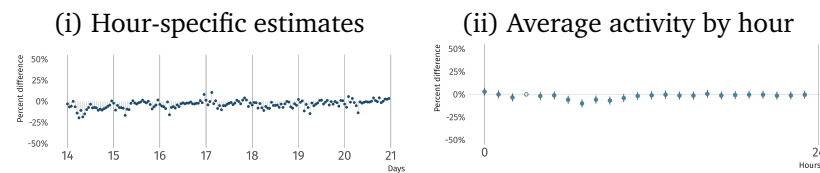
C: 0 to 6 days from DST



D: 7 to 13 days from DST



E: 14 to 20 days from DST



Notes: Point estimates in Column (i) are taken from corresponding hours in Figure A.20, and sorted by hour of the week. Point estimates in Column (ii) are estimates of the average hourly activity in the corresponding week, sorted by hour of the day.

CHAPTER II

HOW GOOD ARE POLLUTION PREDICTIONS?

This chapter is co-authored with Ed Rubin: we both contributed to the writing and statistical analysis of the manuscript.

2.1 Introduction

Recent developments in remote sensing and machine learning algorithms jointly spawned a multidisciplinary literature that predicts environmental features through time and space at high resolutions. These predicted datasets overcome limitations imposed by traditional *in situ* monitoring and offer extensive coverage at relatively low cost, leading to widespread adoption in empirical research.¹ Taken as truth, these predictions offer researchers, policymakers, and regulators an opportunity to observe a continuous measure of environmental quality, providing insights into previously unanswerable questions.

Yet these remotely sensed estimates, although promising, are not ground truths and are inherently uncertain. Relatively little is known about these estimates or the consequences of measurement error when we introduce these machine-learned predictions into more traditional causal inference frameworks.² Despite this fact—and likely reflective of researchers’ excitement for the data—the development and application of predicted data have greatly outpaced research

¹ For example, researchers have used these predictions to estimate forest cover (Hansen et al., 2013), map population density (Robinson, Hohman, & Dilkina, 2017; Xu et al., 2021), measure economic returns to electrification (Ratlidge, Cadamuro, De La Cuesta, Stigler, & Burke, 2022), map poverty (Jean et al., 2016), monitor water quality (Ubah et al., 2021), identify cropland extent (Potapov et al., 2021), and detect illegal mining activity (Swenson, Carter, Domec, & Delgado, 2011).

² Existing research highlights that these errors may introduce biases into subsequent analyses, particularly within traditional causal inference frameworks (Jain, 2020; Proctor, Carleton, & Sum, 2023).

on the empirical intricacies of how such predictions interact with causal inference assumptions (Alix-García & Millimet, 2023; Jain, 2020)

We focus on the use case of air quality—specifically, fine particulate matter smaller than 2.5 microns ($\text{PM}_{2.5}$). This case provides three benefits. First, it represents a large and rapidly growing application space.³ Table 1 summarizes selected studies popular in the literature. Second, $\text{PM}_{2.5}$ is a predicted environmental feature with high economic significance and policy relevance. Exposure to $\text{PM}_{2.5}$ is closely associated with many adverse health effects, including respiratory and cardiac diseases, lower birth weight, and stroke (“The Lancet Commission on pollution and health”, 2018). Third, regulation of $\text{PM}_{2.5}$ is costly for firms and governments. Accordingly, accurate estimation of local $\text{PM}_{2.5}$ is crucial. However, cost constraints of monitoring have limited $\text{PM}_{2.5}$ measurement to a sparse network of regulatory-grade air quality monitors—largely located in urban areas and near emissions inventories such as road ways and industrial sites. Lower-cost monitors are available but sacrifice reliability (Schulte, Li, Ghosh, Fine, & Epstein, 2020).

Given these limitations, public health and economics research has turned to prediction-based $\text{PM}_{2.5}$ data to fill gaps in monitoring, employing predicted $\text{PM}_{2.5}$ datasets in a diverse set of tasks.⁴ However, the reliability of these predictions are

³ More than 113,000 responses appear to a Google Scholar query for “PM2.5 prediction” as of March 2025.

⁴ Predicted $\text{PM}_{2.5}$ estimates have been used to estimate health burdens (Danesh Yazdi, Wang, Di, Zanobetti, & Schwartz, 2019; Di et al., 2017; Power et al., 2024; Rhee et al., 2019; Shi et al., 2023; Wu, Braun, Schwartz, Kioumourtzoglou, & Dominici, 2020), diagnose environmental justice and equity (Childs et al., 2022; Colmer, Hardman, Shimshack, & Voorheis, 2020; Currie, Voorheis, & Walker, 2023; Jbaily et al., 2022; Kelly et al., 2021), assess air pollution policy effectiveness (Currie et al., 2023; Wei et al., 2021), forecast pollution concentrations (Kelly et al., 2021), document gaps in monitoring and regulation (Considine et al., 2023; Van Donkelaar et al., 2021; Wang, Marshall, & Apte, 2024), and backcast missing data (Meng et al., 2019; Wei et al., 2020). For additional information, see Table 2.

crucially determined by modeling decisions and validation practices. Yet these datasets are often treated as universally reliable across diverse research and policy contexts, despite being susceptible nonclassical measurement error (Proctor et al., 2023). Furthermore, recent evidence suggests that conventional methods for evaluating model performance, such as independently resampled k -fold cross-validation, may significantly overstate predictive capabilities, especially when applied to predictions in geospatial contexts (Meyer, Reudenbach, Wöllauer, & Nauss, 2019). This mismatch between stated and realized predictive accuracy could distort empirical conclusions.

Despite the importance of these measurement errors, the literature has yet to comprehensively address several key questions. First, do the reported prediction performances accurately reflect actual performance for common uses—and if not, when does actual performance deviate from stated? Second, how do modeling choices—primarily validation approach and model structure—impact (1) the predictions and (2) empirical results for research that uses these predictions. Third, what are the sources and implications of measurement error, uncertainty, and bias in $\text{PM}_{2.5}$ predictions and how can we address these issues in future research and policy applications?

We contribute to filling these gaps by explicitly addressing these three dimensions. First, we assess the predictive accuracy of $\text{PM}_{2.5}$ estimates generated through standard and spatial cross-validation approaches, evaluating under what conditions actual predictive performance deviates from reported metrics. In the predictive $\text{PM}_{2.5}$ literature, we are the first to introduce a more robust resampling strategy, Nested Cross-Validation (Bischl et al., 2023), helping to reduce overfitting as well as ensuring unbiased performance evaluation. Second, we investigate how distinct modeling decisions—particularly validation methods and model

structure—impact empirical results using a policy scenario: assessing census tracts’ compliance with the Environmental Protection Agency’s (EPA) recently updated air quality standard for $\text{PM}_{2.5}$. Third, we quantify prediction uncertainty and explore practical methods to reduce these errors, thereby improving the reliability of $\text{PM}_{2.5}$ predictions for research and policy applications.

To answer these questions, we produce transparent, monthly $\text{PM}_{2.5}$ predictions across the contiguous United States (CONUS) from 2002 to 2019 at a 1 km by 1 km spatial resolution. We explore a series of modeling configurations, varying validation strategies and modeling structures, following an established predictive approach widely cited in the literature (Di et al., 2019, 2016). Our modeling process is fully transparent and reproducible. In a complementary study, we provide detailed documentation and open access to the full workflow—from accessing downloading raw input data through final model predictions—allowing future researchers to replicate or adapt our approach to their specific contexts.

Our findings provide novel insights into how predictive modeling choices critically shape predictive performance and empirical analyses using predicted $\text{PM}_{2.5}$. We document substantial discrepancies between conventional and spatially explicit validation strategies, demonstrating that traditional approaches often significantly overstate model performance. Nevertheless, we find that tuning model parameters with a contrasting validation strategy did not translate into a substantially different empirical policy conclusions regarding regulatory attainment standards. However, the uncertainty associated with these predictions remains substantial, highlighting the importance of explicitly quantifying uncertainty in analyses employing remotely sensed environmental predictions.

Ultimately, our paper contributes to a growing literature emphasizing rigorous and transparent practices in the development and application of

machine-learned environmental predictions. By rigorously exploring these dimensions, we aim to provide practical guidance for researchers to ensure that predictive environmental datasets effectively and reliably inform empirical analyses.

The remainder of the paper is structured as follows. In the following section, we review the literature on $\text{PM}_{2.5}$ prediction, focusing on the methods and data used in the literature. We then describe the data used in our analysis, including ground truth data, AOD measurements, CTM outputs, meteorological conditions, and land-use variables. Next, we discuss the methods used in our analysis, including performance evaluation, nested cross-validation, the base learner and models, and model prediction. We then present our results, including a comparison of prediction performance across different validation approaches, the impact of modeling choices on prediction estimates and empirical results, and the implications of measurement error, bias, and uncertainty for $\text{PM}_{2.5}$ predictions. Finally, we conclude with a discussion of our findings and their implications for research and policy.

2.2 Data and Methods

2.2.1 Point Observations. Our models are trained using highly accurate ground-based observations of $\text{PM}_{2.5}$ concentrations from the United States Environmental Protection Agency (EPA). Specifically, we obtained 24-hour average $\text{PM}_{2.5}$ concentration measurements covering the period from 2002 to 2019 from the EPA's Air Quality System (AQS), accessed via their publicly available Air Data platform.⁵ We aggregate daily observations to the monthly-level.

Due to costs and logistical constraints, these regulatory-grade monitors are geographically sparse and primarily clustered in urban areas or near industrial

⁵ Pre-Generated Data Files hosted on Air Data.

emissions sources. Figure 6 describes the spatial distribution of the monitoring sites across the CONUS at a point in time. Panel A indicates monitor locations that contribute to the training data during June 2011, the month with the highest number of active monitors. Importantly, not every monitoring site has data available throughout the study period. Throughout the sample periods, monitoring sites vary in their availability—new monitors are added, existing ones are relocated, and others are temporarily inactive or permanently disabled. The spatial distribution of these monitors, and their temporal coverage present important limitations for prediction tasks. On average, 1452.37 monitors are active across the CONUS each month.

Panel B further illustrates the spatial distribution of monitoring sites relative to each pixel in our sample area. The color of each pixel represents the distance to the nearest monitor, highlighting the gaps in monitoring coverage across the CONUS. Darker shaded areas indicate areas of the CONUS that are farthest away from monitoring sites, making them the most difficult to predict.

2.2.2 Model Inputs. To overcome limitations imposed by sparse monitoring networks, our predictive models integrate multiple spatially continuous, temporally varying environmental features widely use in $PM_{2.5}$ prediction research. These features capture a wide range of environmental and anthropogenic factors that influence $PM_{2.5}$ concentrations. The objective of our model is to learn the relationship between $PM_{2.5}$ concentrations and these features to predict $PM_{2.5}$ concentrations in regions where no monitoring is available.

We broadly follow the predictive approach established by two of the most widely cited papers in the literature (Di et al., 2019, 2016), though we our feature set is not as rich and we implement different imputation strategies. All features

are resampled to the 1 km by 1 km grid used by MODIS satellite products using bilinear interpolation. We report the full feature set used in our analysis below.

AOD Measurements. Aerosol Optical Depth (AOD) measurements are widely used as a predictor of surface $\text{PM}_{2.5}$. These measurements are derived from the Moderate Resolution Imaging Spectroradiometer (MODIS) instrument aboard the Aqua (2003-2019) and Terra (2002-2019) satellites using the Multi-Angle Implementation of Atmospheric Correction (MAIAC) algorithm.⁶ at a 1 km by 1 km spatial resolution. Missing AOD values frequently occur due to cloud cover or highly reflective surfaces (e.g., snow-covered regions). To address surface-related biases in AOD retrieval, we include MODIS surface reflectance data (MOD09A1, Band 3) as an additional predictor.

To mitigate missing data limitations inherent in satellite-derived AOD, we further incorporate aerosol components simulated by the Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). While subject to simulation errors, these model-based aerosol outputs are valuable complements to satellite-derived AOD, providing nearly continuous spatial and temporal coverage.

Chemical Transport Model Outputs. Chemical transport models (CTMs) like GEOS-Chem (geoschem.github.io) and CMAQ (epa.gov/cmaq) are a numerical simulation framework that computes the formulation, dispersion, and deposition of particulate matter based on atmospheric chemistry and known emission inventories. Although CTMs offered a more mechanistic perspective compared to purely statistical models, they were often computationally complex and had limited spatial resolution, which constrained their utility in high-resolution exposure assessments. This complexity limited the feasibility of using

⁶ We collect AOD measurements at both the 470 nm and 550 nm wavelengths.

CTMs for real-time applications and high-resolution exposure assessments, especially for fine spatial scales

However, the outputs of these models are valuable predictors of $\text{PM}_{2.5}$ concentrations. We obtain daily predictions of total $\text{PM}_{2.5}$ mass and several $\text{PM}_{2.5}$ components from the CMAQ model at 12-km resolution over the US. These predictions are aggregated to the 1 km by 1 km grid used in our analysis.

Meteorological Conditions. Meteorological features are an important predictor of $\text{PM}_{2.5}$ concentrations, as they influence the dispersion and transport of pollutants. We include 16 variables from NCEP North American Regional Reanalysis at daily resolution, including evaporation, precipitation, temperature, albedo, radiation flux, cloud cover, planetary boundary layer height, precipitable water, pressure, specific humidity, visibility, and wind speed.

Land-Use Variables. Land-use variables are valuable proxies for local emissions. For example, areas with high vehicular traffic or industrial activities tend to have elevated $\text{PM}_{2.5}$ levels due to emissions, while vegetation and water bodies can act as natural sinks, helping to reduce $\text{PM}_{2.5}$ concentrations. Due to high computation costs and limited data of fine-scale emissions, CTMs are unable to simulate $\text{PM}_{2.5}$ at fine spatial scales. Land-use variables are useful to approximate these local-level emissions.

The variables encompassed land-use coverage types obtained from the National Land Cover Database (NLCD), which provides 30 by 30-meter land cover data categorized by a 16-class legend-based system, with the modal category used to represent land cover type when aggregating across cells. Road density was calculated from county TIGER shapefiles, focusing on both primary and secondary roads to assess the impact of road traffic on $\text{PM}_{2.5}$ levels. Elevation data was sourced from the GMTED2010 product suite, specifically the 7.5 arc-second

raster elevation data (approximately 225 meters), which was aggregated up to 1 by 1 km grid cells, with metrics such as mean, minimum, maximum, and standard deviation of elevation calculated to capture topographic variability within each grid cell. Lastly, the Normalized Difference Vegetation Index (NDVI) was evaluated at 1 km by 1 km grid cells to assess vegetation health and vigor, elucidating the interaction between vegetation cover and $PM_{2.5}$ concentrations, as vegetation can act as a sink for pollutants.

Spatial Lags. Following previous studies (Di et al., 2019, 2016), we compute spatial lags using inverse distance weighting (IDW) to average $PM_{2.5}$ concentrations to neighboring monitors. We generate multiple spatial lags with buffer radii ranging from 0 to 500 km, explicitly excluding monitors within buffer distances to prevent local overfitting.

Imputation. Following the resampling into a common grid structure, two preprocessing techniques were applied to covariates: (1) the imputing of no-data values to fill raster pixels; and (2) the ordinal encoding of categorical covariates. The imputation of categorical covariates was done by filling no-data values with the most frequent value. For continuous variables, we impute missing values using a rolling average with varying windows: initially with a 3-month rolling average, and extending to 5, 7, and 9 months if the pixel remains missing.

2.2.3 Model Training. While more complex models like neural networks or CNNs have made recent advancements in spatial prediction tasks, and specifically for $PM_{2.5}$ prediction (Di et al., 2019, 2016; Park et al., 2020; Requia et al., 2020), simpler models like random forests and gradient boosting models are still widely used in the literature due to their interpretability and ease of implementation. Further, it has been shown that performance of these models are comparable to the latest deep learning algorithms (Di et al., 2019).

To predict the spatial distribution of PM_{2.5}, we employed LightGBM, a Histogram-based Gradient Boosting model which is a more computationally efficient implementation of the Gradient Boosting algorithm (Ke et al., 2017). All models were implemented in R using the ‘mlr3’ package (Lang et al., 2019).

The primary model is designed to predict monthly average PM_{2.5} concentrations, with optimization guided by minimization of the Mean Squared Error (MSE):

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (\text{PM}_{i,t} - \widehat{\text{PM}}_{i,t})^2$$

where $\text{PM}_{i,t}$ denotes observed PM_{2.5} concentrations, $\widehat{\text{PM}}_{i,t}$ are predicted PM_{2.5} concentrations, and n is the total number of observations. By minimizing MSE, the model provides predictions representing the conditional expectation of PM_{2.5} concentrations at each pixel, given observed covariates.

Recognizing that predictions include uncertainty, we further quantify this uncertainty using quantile regression methods. We trained additional LightGBM models, optimized with the pinball loss function, defined as:

$$f(\text{PM}_{i,t}, \widehat{\text{PM}}_{i,t}; \tau) := \begin{cases} \tau (\text{PM}_{i,t} - \widehat{\text{PM}}_{i,t}), & \text{if } \text{PM}_{i,t} > \widehat{\text{PM}}_{i,t}, \\ (\tau - 1) (\widehat{\text{PM}}_{i,t} - \text{PM}_{i,t}), & \text{if } \text{PM}_{i,t} \leq \widehat{\text{PM}}_{i,t} \end{cases}$$

where τ is the quantile of interest, $\text{PM}_{i,t}$ is the true concentration, and $\widehat{\text{PM}}_{i,t}$ is the predicted concentration. The pinball loss function, a quantile regression loss function, estimates the conditional quantiles of the response variable. Since quantile predictions may involve estimating values at the tails of a distribution, the pinball loss computes the difference between the predicted and observed values while adding bias to the error depending on how extreme the desired

quantile is, and whether the predicted value for that quantile is more or less extreme than the observed value.

Separate models were fitted to predict the 2.5th and the 97th percentiles of the response variable, providing a range in which we are 95 percent confident a new observation will fall. Using these separately trained models, we infer an estimate the prediction interval for each observation by taking the difference between the 97th and 2.5th percentile predictions for each observation. This prediction interval can be interpreted as the range in which we are 95 percent confident a new observation will fall given the model’s predictions.

2.2.4 Assessing Model Performance. Cross-validation is a widely used resampling technique helping to evaluate model generalizability and to preventing overfitting by tuning model hyperparameters for better accuracy (James, Witten, Hastie, & Tibshirani, 2013). The performances of machine learning algorithms are typically communicated via cross-validation estimates—where the cross-validation strategy should always be designed according to the purpose of the model. In the context of $\text{PM}_{2.5}$ prediction, we argue that current cross-validation approaches are not properly suited to the task of filling in gaps in monitoring networks.

The basic idea of cross-validation is to split the training data into a specified number of partitions, referred to as folds. At the observation-level, the training data are randomly sampled in into k folds. Each fold is held out once, while the model is trained on the remaining $k - 1$ folds, where the help out fold is intended to imitate unobserved observations. Each model is then evaluated on the held-out fold, and the process is repeated k times. The final model performance is averaged across the k folds to estimate the overall model accuracy.

In the $\text{PM}_{2.5}$ literature, independently resampled 10-fold cross-validation is the dominant approach, where the data is randomly partitioned at the

observation-level into ten folds (Di et al., 2019, 2016; van Donkelaar et al., 2016). Table 1 summarizes a selection of studies, including their validation strategy, and reported evaluation metrics (R^2). Using this approach, previous studies have reported high prediction accuracy, with R^2 values ranging from 0.7 to 0.9.

However, one significant challenge in modeling $\text{PM}_{2.5}$ is *data leakage*. The main idea of data leakage is that the model is evaluated on data that are not fully independent of the training data. Instead of the model learning the relationship between the features and the response vectors, the model is learning from underlying information that is leaked from the testing fold. The model's high performance is dependent on information that it does not have access to when predicting into unknown space, leading to an overinflated confidence of the model's predictive capability. Put simply, the model is cheating and we should not expect the same level of performance when predicting into new, unobserved regions.

For example, using independent resampling, two observations from the same air quality monitor may be split into two separate folds. In this scenario, the model has access to one observation before making a prediction on the other. If we assume that observations in time from the same monitor exhibit temporal autocorrelation, then the model would do well to predict the second observation based on the first. Similarly, if two monitors are located in close proximity, the model may learn the relationship between the two monitors, rather than the relationship between the features and the response vector.

This is particularly problematic in the context of predicting $\text{PM}_{2.5}$ concentrations, as the underlying structure of the data is highly spatially and temporally correlated. Furthermore, the presence of spatial lags in the model further complicate our understanding of the model's predictive capability. As

we predict into the darker regions of Panel B in Figure 6, the model will not have access to this information, leading to a decrease in ability to predict $\text{PM}_{2.5}$ concentrations in areas without a monitor.

We ask, do reported prediction performances accurately reflect actual performance? Instead of the model learning from the features, potentially the model is learning from the underlying structure of the data. To address this question, we use a spatial cross-validation approach, a resampling technique that ensures that training and test sets are spatially distinct, thereby mitigating the effect of autocorrelation and preventing data leakage (Meyer et al., 2019). Spatial cross-validation is particularly useful in scenarios where we might worry about data leakage. Previous studies have shown that spatial cross-validation can provide a more accurate estimate of model performance, particularly in the context of spatial prediction tasks (Schratz, Muenchow, Iturritxa, Richter, & Brenning, 2019). In the context of $\text{PM}_{2.5}$ prediction, this is not the first study to apply spatial cross-validation. Previous studies have implemented this approach and find reduction in model performance when compared to independent resampling (Boser, 2024; Reid, Considine, Maestas, & Li, 2021).

Figure 7 illustrates both independent resampling and spatial resampling strategies. Panel A shows the standard approach using 3-fold cross-validation. Each layer of pixel describes the sample across different points in time, and the color of each pixel describes the fold that the observation is assigned to. White folds indicate areas without monitors. Panel B shows the spatial resampling approach, where the data is clustered into 3 distinct spatial clusters. Each cluster is then used as a fold in the cross-validation process, effectively limiting the model to only learn from observations in the same cluster. This approach ensures that

the model is not indirectly learning from the test data, providing a more accurate assessment of spatial generalizability.

Nested resampling. We introduce a nested cross-validation, a rigorous resampling technique novel in the context of $\text{PM}_{2.5}$ predictions. This approach is increasingly applied in similar spatial context to estimate the performance of machine learning algorithms (Schratz et al., 2019). This approach ensures an unbiased estimate of the model’s generalization error. If the same data is used for determining the optimal configuration and the evaluation of the resulting model itself, the actual performance estimate might be biased (Bischi et al., 2023).

Nested cross-validation splits the data into an outer and an inner loop. The inner loop is used to tune the hyperparameters of the model, while the outer loop is used to evaluate the model’s performance. This approach prevents optimistic bias that arises when hyperparameter tuning is performed on the same dataset used for evaluation, thereby further avoiding overfitting to the validation set. This approach is particularly useful in scenarios where hyperparameter tuning can significantly impact performance, as it allows for a rigorous assessment of both the model’s predictive capability and its generalizability to new, unobserved regions.

Figure 8 illustrates the nested resampling approach using a 3-fold outer loop and a 3-fold inner loop. Layers of each raster plot corresponds to a different point in time and pixels are colored by the fold that the observation is assigned to. Only one outer fold is shown for clarity, colored in gray, but the process is repeated three times. Panels A-D illustrate the different combinations of the outer and inner loop, with the inner loop being either independent or spatially clustered. Specifically, Panel A shows the standard approach of independent resampling in both the inner and outer loop, Panel B shows spatial resampling in the inner loop and independent resampling in the outer loop, Panel C shows

independent resampling in the inner loop and spatial resampling in the outer loop, and Panel D shows spatial resampling in both the inner and outer loop.

2.3 Empirical Application

To evaluate the practical implications of different predictive modeling decisions, we conduct a related downstream empirical applications that relies on the monthly predicted $\text{PM}_{2.5}$ values aggregated into annual-average Design Values (DV) at the census-tract level. These DVs are constructed by average predicted monthly $\text{PM}_{2.5}$ concentrations from 2017-2019, matching the EPA's regulatory approach. Monthly estimates are aggregated to the census-tract level. DV is calculated as the quarterly average, aggregated to the annual level, and then averaged across the three years. This approach is consistent with the EPA's regulatory approach, which averages the annual average $\text{PM}_{2.5}$ concentrations over a three-year period to determine compliance with the standard. Below we describe each analysis separately.

We use predicted DVs to classify census tracts relative to the EPA's updated annual-average $\text{PM}_{2.5}$ air quality standard for $\text{PM}_{2.5}$ ($9 \mu\text{g}/\text{m}^3$). Following previous literature using $\text{PM}_{2.5}$ predictions examining similar policy classifications (Wang et al., 2024), each census tract is classified as either compliant or non-compliant based on whether the predicted DV exceeds the EPA threshold. A census tract is classified as compliant if:

$$DV_i \leq 9 \mu\text{g}/\text{m}^3$$

and non-complaint otherwise. We then evaluate how different modeling configurations—varying by validation strategies and model structures—affect

classification outcomes. This comparison directly addresses our research question about how modeling choices impact empirical results.

Further, we explicitly incorporate prediction uncertainty into our compliance assessment by using quantile regression prediction intervals. Specifically, we classify census tracts as confidently compliant, non-compliant, or uncertain by comparing the upper bound of the 95 percent prediction interval to the EPA threshold. Tracts whose upper prediction intervals (97.5th percentile) fall entirely below the EPA threshold are classified as confidently compliant, whereas tracts whose intervals contain the threshold are classified as uncertain. This approach allows us to assess the uncertainty in our classification and its implications for research and policy.

2.4 Results

2.4.1 Model Performance. To examine whether reported prediction performance accurately reflect actual performance, we evaluated $\text{PM}_{2.5}$ prediction models using 3- by 5-fold nested cross-validation with four distinct validation strategies: (1) independent resampling in both the inner and outer loops (IID-IID), (2) spatial resampling in the inner loop and independent resampling in the outer loop (SPCV-IID), (3) independent resampling in the inner loop and spatial resampling in the outer loop (IID-SPCV), and (4) spatial resampling in both the inner and outer loops (SPCV-SPCV). Model performance was measured using R^2 , and is summarized in Figure 9.

The matrix cells display R^2 values from the combination of the outer cross-validation strategy (row), inner cross-validation strategy (column), and available spatial lags (plot). The top row, IID CV, references the standard k -fold cross-validation approach of independent sampling at the observation level common in the literature for the outer loop. The bottom row, SPCV, uses a spatially clustered

cross-validation, again for the outer loop. Likewise, columns follow the same pattern, except describe the inner cross-validation strategy that is used to test for optimal model parameters prior to a final evaluation by the outer validation loop. The separate plots in the figure separate the models by the features available, specifically, whether spatial lags are included in the model, and whether the spatial lags include a buffer distance where only information beyond the buffer is included.

Our findings show significant discrepancies between traditionally independent resampling and spatial cross-validation strategies. Standard independent resampling strategies (IID-IID) yield consistently high R^2 values, exceeding 0.8, comparable to predictive accuracy widely reported in previous literature (see Table 1). However, when evaluated using spatial cross-validation (SPCV-SPCV), R^2 values drop considerably. For a model with 0km spatial lags, R^2 values fall from 0.84 to 0.63, a 25 percent decrease. When spatial lags are not present, the difference is even more stark, R^2 values fall from 0.83 to 0.35, a 58 percent decrease. This result suggests that high predictive performance observed in independent resampling context primarily reflects data leakage rather than genuine predictive capability that generalizes to new, unobserved regions.

While this can partially be attributed by the size of geospatial clusters used in the spatial cross-validation with only 3 outer folds, Figure B.1 shows that increasing the number of outer folds in the spatial cross-validation increases model performance. However, even with 5 outer folds, the model still underperforms the standard cross-validation model by a significant amount.

2.4.2 Uncertainty in Predictions. To quantify uncertainty, we derive 95 percent prediction intervals from quantile regression models. Figure 10 illustrates out-of-sample prediction accuracy for the four main validation strategies of PM_{2.5}

models, across baseline levels of $\text{PM}_{2.5}$ in pixels with a monitor. The predictions are estimated only on pixels with monitoring data and are generated using 3- by 5-fold nested cross-validation. Each prediction is the estimate $\text{PM}_{2.5}$ concentration in the held out fold, and the true value is the observed $\text{PM}_{2.5}$ concentration. Predictions are aggregated to 1 unit (μ/m^3) bins, winsorized at the 1st and 99th percentile, by the mean of the true $\text{PM}_{2.5}$ concentration. To compare the impact of spatial lags on prediction accuracy and uncertainty, the figure is split into two columns show alternative spatial lag specifications included all models—the left column without any spatial lags and the right column including 0km spatial lags.

We find that predictions across all models show a strong positive correlation with observed $\text{PM}_{2.5}$ concentrations. In panel A, we plot bins of observed $\text{PM}_{2.5}$ concentrations (x-axis) to predicted concentrations (y-axis). The 45-degree line represents perfect prediction, where the predicted value equals the true value. We find that at high levels of $\text{PM}_{2.5}$, the models tend to underestimate the true values, with spatially cross-validated models exhibiting larger degrees of bias. We attribute the underestimation of high $\text{PM}_{2.5}$ levels to the over-representation of moderate $\text{PM}_{2.5}$ levels in the training data, influencing the model to under predict high $\text{PM}_{2.5}$ levels.

In panel B, we show that 95 percent prediction intervals are generally wide across all baseline $\text{PM}_{2.5}$ levels and widest spatially cross-validated models, reflecting their more realistic assessment of uncertainty. Particularly at higher observed $\text{PM}_{2.5}$ levels. Panel C, shows the probability that the true value falls within the prediction interval for each model across true $\text{PM}_{2.5}$ concentrations. The lines plot the probability, calculated as the number of true values that fall within the prediction interval divided by the total number of true values. The histogram plots the distribution of baseline $\text{PM}_{2.5}$ across the training data.

Notably, the probabilities of all models is highest at the most common $\text{PM}_{2.5}$ levels, indicating that typical $\text{PM}_{2.5}$ values are more accurately predicted by the models and are well represented in the training data. At $\text{PM}_{2.5}$ concentrations on the tails of the distribution, the probability of the true value falling within the prediction interval decreases, underscoring the limitations of these models to accurately predict high, and low, pollution levels. Notably, the models with access to 0 km spatial lags manage much better at accurately containing the true value within the prediction interval, particularly at high $\text{PM}_{2.5}$ levels.⁷

2.4.3 Compliance Assessment of Census Tracts. We use predicted $\text{PM}_{2.5}$ Design Values to assess census-tract compliance with the EPA's updated annual average air quality standard. Figure 12 provides visualizations of this analysis. Each the four panels are generated using the inner independently resampled and outer independently resampled, 3- by 5-fold nested cross-validation model (IID-IID).

Panel A plots the DV for each pixel in the CONUS, between 2017-2019. To construct the DV, we aggregate the monthly predictions, first to the year-by-quarter level, then the the annual level, and finally to the 3-year level (Wang et al., 2024). The color value within each pixel indicates the predicted $\text{PM}_{2.5}$ concentration DV, where lighter values indicate higher levels of $\text{PM}_{2.5}$. In a similar fashion, panel B plots the width of the 95 percent prediction interval at the pixel level. The color value within each pixel indicates the width of the prediction interval, where lighter values indicate higher levels of uncertainty. Spatial variation in uncertainty is clear, with particularly wide intervals observed in the Western United States.

⁷ Figure B.2 reproduces the same figure using in-sample predictions.

Panel C aggregates predicted DVs to the census tract-level, classifying each tract according to the EPA's updated standard. Tracts with predicted DVs exceeding the regulatory threshold ($9 \mu\text{g}/\text{m}^3$) are indicated in black. Panel D further incorporates prediction uncertainty into regulatory classifications. Census tracts colored in black indicate locations where the 95 percent interval for $\text{PM}_{2.5}$ includes the EPA's standard, highlighting tracts for which compliance status cannot be confidently rejected.

Providing further detail, Figure 11 evaluates uncertainty at the census tract-level, limited to tracts containing EPA monitors. Panel A plots predicted DVs against their rank-order (from lowest to highest), along their associated prediction intervals. Vertical intervals show uncertainty around predicted DVs, with purple intervals indicating tracts confidently classified as compliant, and grey intervals indicating tracts where compliance status is uncertain.

Panel B of Figure 11 compares these results across all four validation strategies: (IID-IID, SPCV-IID, IID-SPCV, SPCV-SPCV). The y-axis again plots predicted $\text{PM}_{2.5}$ DVs with smoothed curves representing mean predictions, and their associated prediction intervals. Despite clear differences in predictive performance across these validation strategies, this comparison highlights relative consistency across validation approaches when predictions are aggregated at census tract levels. Differences in regulatory classifications across models, though present, are modest.

2.5 Discussion

This study provides a comprehensive analysis of methodological choices in predicting $\text{PM}_{2.5}$ concentrations using remotely sensed data, revealing significant implications for empirical research and policy decisions. Three central findings emerge from our results.

First, our analysis demonstrates substantial discrepancies between traditionally reported model accuracies and more realistic predictive performance obtained through spatial cross-validation methods. While previous studies commonly report high predictive accuracy metrics (typically R^2 values between 0.7 and 0.9) using independently resampled validation techniques. Our results show these metrics significantly decline when evaluated through spatially explicit cross-validation (SPCV). Using spatial validation reduced the estimated predictive accuracy by up to 60 percent relative to traditional cross-validation approaches. This finding strongly suggests that standard cross-validation approaches used widely in the literature systematically overstate actual model performance by implicitly leveraging underlying structures in the training data—a phenomenon known as data leakage. Consequently, the apparent accuracy of published predictions may not accurately represent real-world predictive capability, particularly in contexts requiring spatial extrapolation. This result has critical implications for non-classical measurement error and the reliability of downstream empirical analyses that use these predictions.

Further, we introduce nested cross-validation, novel within the $\text{PM}_{2.5}$ prediction literature, to more rigorously evaluate model performance. This approach substantially mitigates biases in traditional model evaluation strategies.

Second, by constructing a 95 percent prediction interval through quantile regression methods, we find substantial uncertainty in $\text{PM}_{2.5}$ predictions. These intervals are wide, particularly at high $\text{PM}_{2.5}$ levels. Fewer than 20 percent of census tracts can be confidently classified as compliant with the EPA's updated air quality standard. Importantly, the wide prediction intervals highlight that estimates alone inadequately convey uncertainty in model predictions and future research and policy decisions should carefully consider these uncertainties.

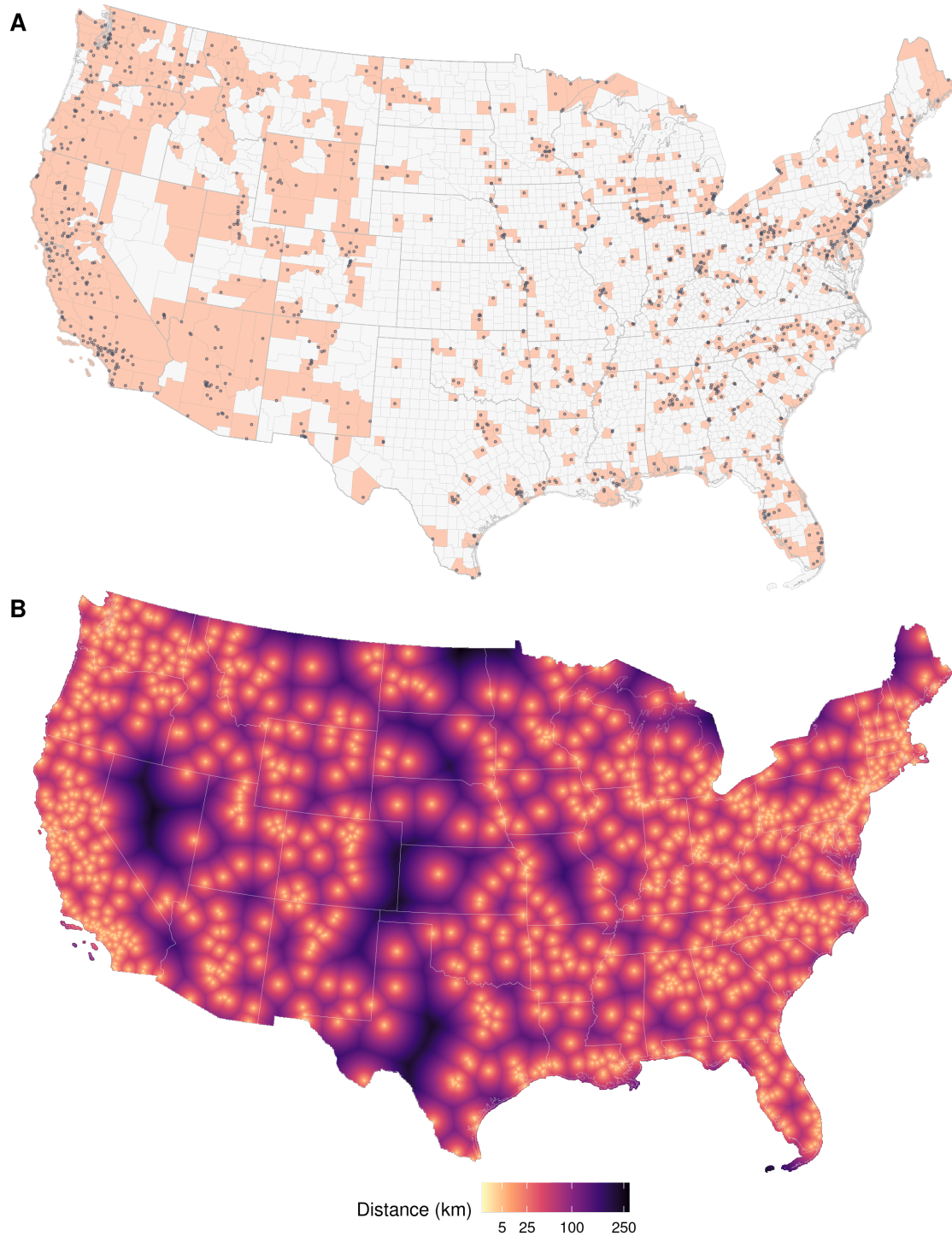
Taken together, our results have important methodological implications for both researchers and policymakers. First, they suggest that predictive modeling of environmental quality must incorporate validation procedures that match the intended application. Independently resampled cross-validation may serve well if the task is to fill in missing data during temporary monitoring gaps. However, if the goal is to predict $\text{PM}_{2.5}$ concentrations in unmonitored regions, spatially explicit validation is essential to ensure reliable model performance. Second, uncertainty quantification should become standard practice and should clearly be communicated alongside point predictions.

2.6 Conclusion

This paper contributes to a growing literature emphasizing methodological rigor and transparency in applying machine-learned environmental predictions. We demonstrate that traditional performance validation strategies can significantly overestimate predictive accuracy, masking substantial measurement error. By introducing spatial cross-validation and quantile-based uncertainty quantification, we provide more realistic assessments of model performance. Future research and policy decisions should carefully consider these methodological choices, incorporating spatially-aware validation methods, explicit uncertainty quantification, and representativeness checks to ensure that conclusions drawn from predicted environmental data are robust and reliable.

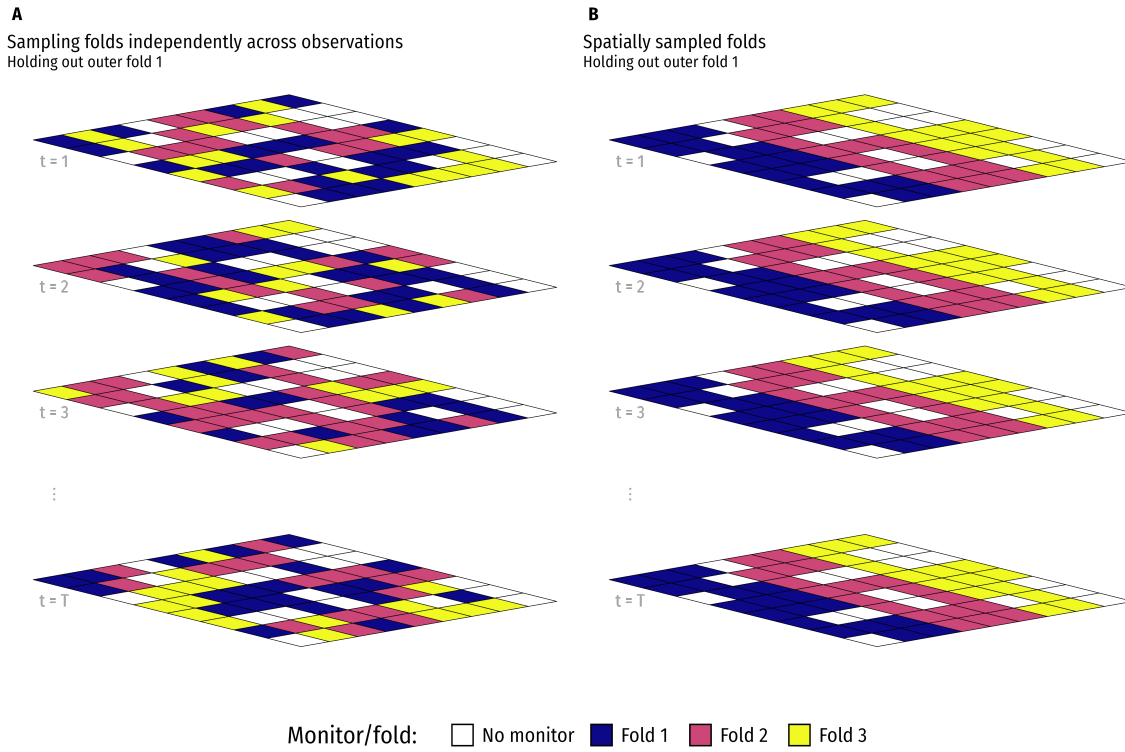
2.7 Figures and Tables

Figure 6. PM_{2.5} monitoring sites across the CONUS in June 2011.



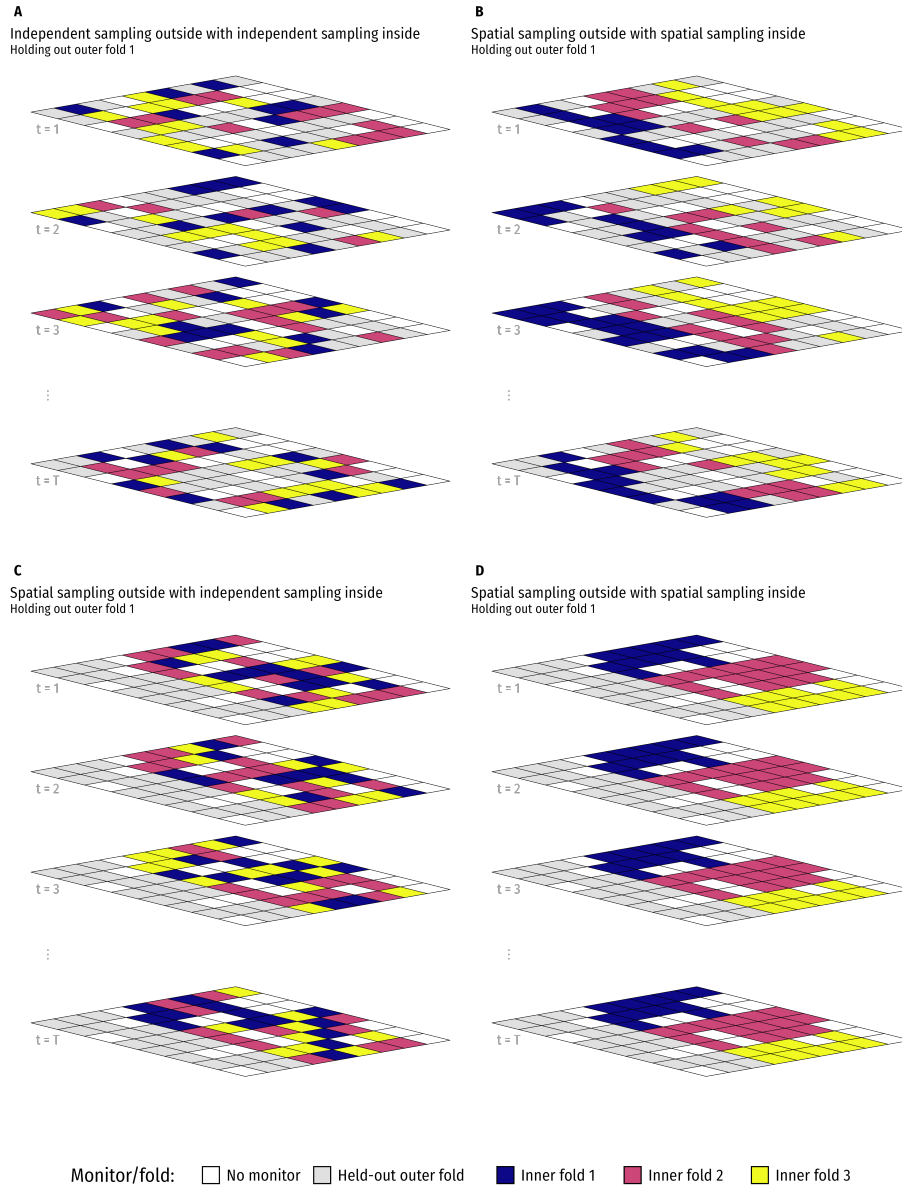
Panel A describes the spatial distribution of EPA monitoring sites across the CONUS at a point in time. Shaded counties represent the presence of a monitor within the county. Panel B describes the distance of each pixel to the nearest monitor. Pixel color describes the distance of each pixel to the nearest monitor, highlighting the gaps in monitoring coverage across the CONUS.

Figure 7. Cross validation in temporally repeated grids



Panel A shows the standard approach using 3-fold cross-validation. Each layer of pixel describes the sample across different points in time, and the color of each pixel describes the fold that the observation is assigned to. White folds indicate areas without monitors. Panel B shows the spatial resampling approach, where the data is clustered into 3 distinct spatial clusters. Each cluster is then used as a fold in the cross-validation process, effectively limiting the model to only learn from observations in the same cluster.

Figure 8. Nested cross validation in temporally repeated grids



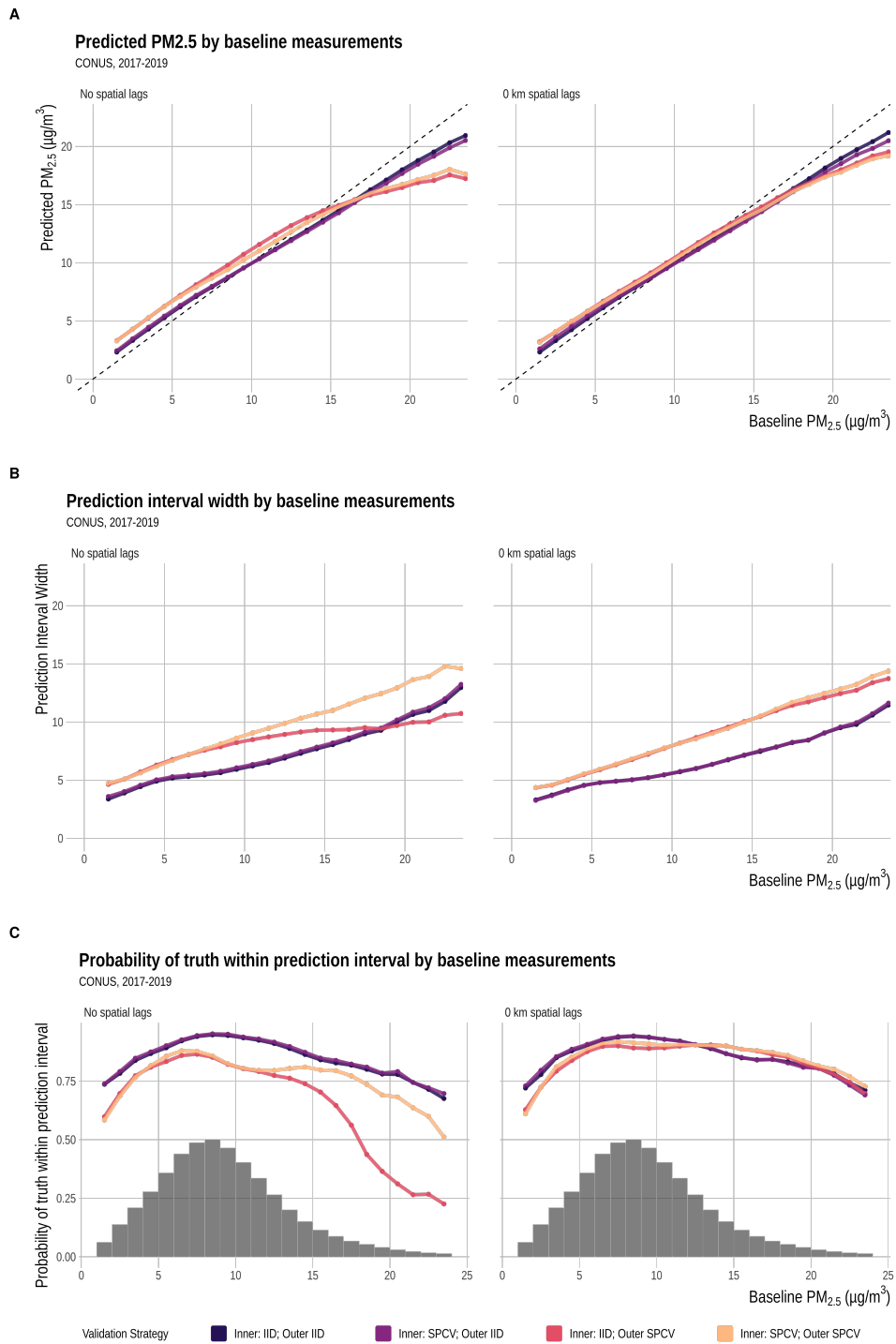
Panels A-D illustrate the different combinations of the outer and inner loop, with the inner loop being either independent or spatially clustered. Specifically, A-D represent the standard approach of independent resampling in both the inner and outer loop, spatial resampling in the inner loop and independent resampling in the outer loop, independent resampling in the inner loop and spatial resampling in the outer loop, and spatial resampling in both the inner and outer loop, respectively. Layers of each raster plot corresponds to a different point in time and pixels are colored by the fold that the observation is assigned to. Only one outer fold is shown for clarity, colored in gray, but the process is repeated three times.

Figure 9. PM_{2.5} prediction accuracy declines steeply when spatially validated and/or restricted from using *close* spatial lags.



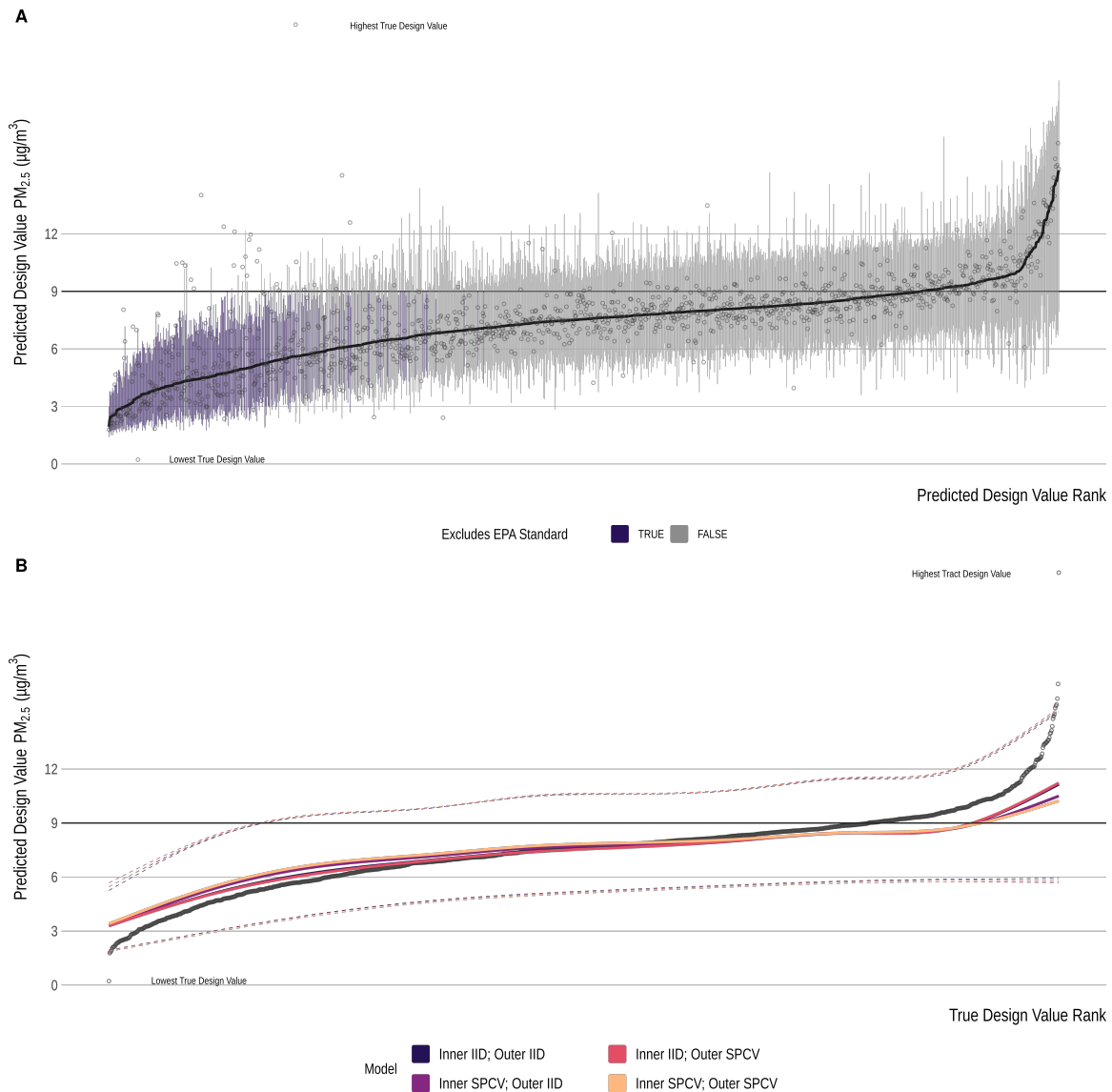
Matrix cells display (and are filled) by R^2 values from the combination of cross-validation approach (row) and available spatial lags (columns). The bottom row, *IID CV*, references the standard CV approach in the literature: independent sampling at the observation level. Remaining *SPCV* rows use spatially clustered CV with regions buffered according to the number. Columns describe the availability of spatial lags and whether they are buffered.

Figure 10. Out-of-sample PM_{2.5} prediction accuracy



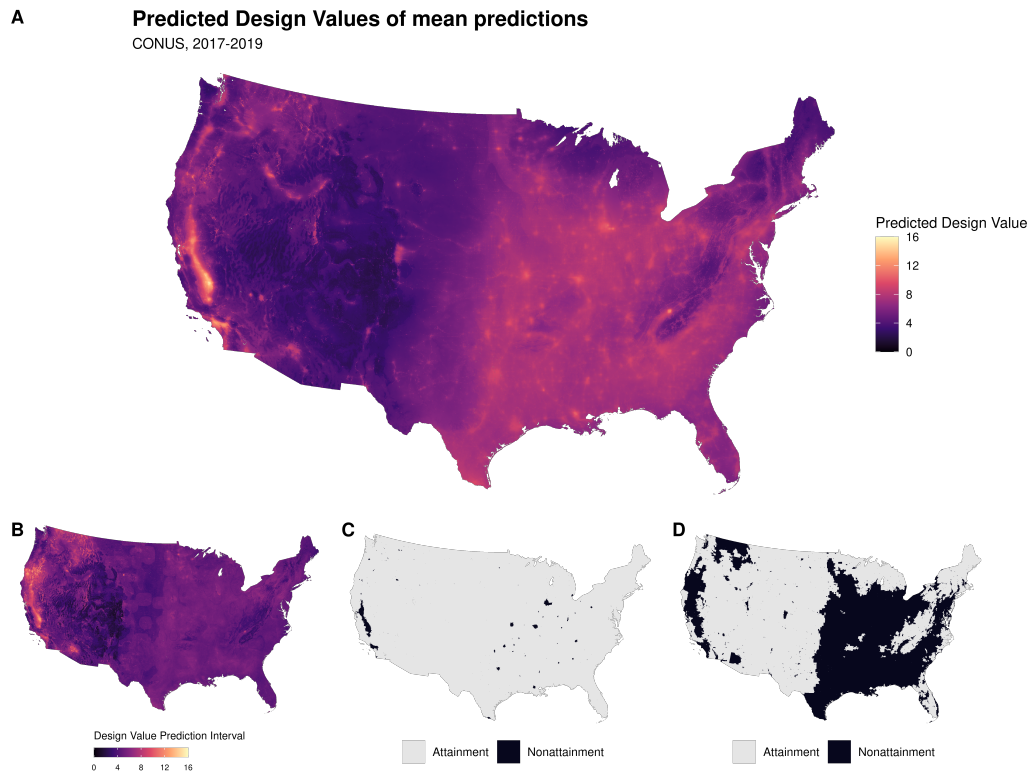
Panel A compares binned predicted PM_{2.5} values to binned true PM_{2.5} values for pixels with monitors. Panel B plots prediction intervals across the same binned truePM_{2.5} values. Panel C shows the probability that the true value falls within the prediction interval for each model.

Figure 11. Attainment Status by Design Value Rank



Panel A plots predicted Design Values against their rank-order (from lowest to highest), along with their associated prediction intervals. Vertical intervals show uncertainty around predicted Design Values, with purple intervals indicating tracts confidently classified as compliant, and grey intervals indicating tracts where compliance status is uncertain. Panel B compares these results across all four validation strategies. The y-axis plots predicted $PM_{2.5}$ Design Values with smoothed curves representing mean predictions, and their associated prediction intervals.

Figure 12. Pixel Design Values, uncertainty, and attainment.



(A) Predicted Design Values for each pixel (2017-2019). (B) Prediction intervals for pixel Design Values. (C) Census-tract attainment status: dark tracts do not meet the standard. (D) Tracts whose prediction interval includes the EPA standard (dark) cannot rule out exceedance.

Table 1. Recent papers that produce PM estimates largely report high cross-validated accuracy, *i.e.*, R^2 above .8.

Authors and year	Pred. years	Frequency	Extent	R^2	CV	Citations
van Donkelaar et al. (2016)	1998–2014	Yearly	Global	[.78, .81]	IID	1,204
Wei et al. (2021)	2000–2018	Monthly	China	[.80, .90]	IID	820
van Donkelaar, Martin, Li, and Burnett (2019)	1998–2019	Monthly	Global	[.75, .95]	IID	680
Di et al. (2019)	2000–2016	Daily	CONUS	[.73, .91]	IID	616
X. Hu et al. (2017)	2011	Daily	CONUS	[.64, .83]	IID	604
Van Donkelaar et al. (2021)	1998–2019	Monthly	Global	[.51, .86]	IID	542
Di et al. (2016)	2000–2012	Daily	CONUS	[.74, .88]	IID	506
Wei et al. (2020)	2018	Daily	China	[.88, .89]	IID, OOSCV, TCV	500
Reid et al. (2015)	2008	Daily	North Calif.	.80	IID	270
Requia et al. (2020)	2000–2016	Daily	CONUS	[.86, .93]	IID, OOSCV	195
Meng et al. (2019)	1981–2016	Yearly	North Am.	[.60, .85]	IID	101
Reid et al. (2021)	2008–2018	Daily	Western US	[.58, .73]	SPCV	48

This table provides selected popular papers in the PM-prediction literature. The first column provides linked citations with the lead author’s last name and the publication year. The following columns report the prediction years (Pred. years), temporal frequency, spatial extent, cross-validated performance (R^2), cross-validation strategy (CV), and citation count (via Google Scholar, as of 19 May 2025). Cross-validation strategies include: IID (independent resampling), OOSCV (out-of-station cross-validation), TCV (temporal cross-validation), and SPCV (spatial cross-validation).

Table 2. Summary of Consumer Studies

Authors & Year	Econ	Journal	Datasets Used	Years	Freq.
Polonik et al. (2023)	F	PNAS	Van Donkelaar et al. (2021)	2015–2019	Annual
Currie et al. (2023)	T	AER	Di et al. (2016), van Donkelaar et al. (2016)	2000–2015	Annual
Colmer et al. (2020)	F	Science	Meng et al. (2019)	1981–2016	Annual
Sager and Singer (2024)	T	AEJ Policy	Meng et al. (2019), Di et al. (2016), Van Donkelaar et al. (2021)	1990–2016	Annual
Shi et al. (2023)	F	Nature Comm.	Di et al. (2019)	2000–2016	Annual
Jbaily et al. (2022)	F	Nature Comm.	Meng et al. (2019)	2000–2016	Annual
Z. Hu et al. (2023)	F	Sustainability	Wei et al. (2020), Wei et al. (2021)	2015	Daily
Considine et al. (2023)	F	EST	Di et al. (2016), Reid et al. (2021)	2008–2016	Daily

This table provides selected papers that use PM_{2.5} prediction data. The first column provides linked citations with the lead author's last name and publication's year. The following columns provide T/F for papers published in Economics journals, the journal name, linked citations, datasets used, and temporal frequency of the predict PM_{2.5}, and years covered.

CHAPTER III

EVALUATING THE IMPACT OF NONATTAINMENT DESIGNATIONS UNDER THE CLEAN AIR ACT

3.1 Introduction

The Clean Air Act (CAA) and its subsequent amendments are the cornerstone of air quality regulation in the United States. The CAA establishes National Ambient Air Quality Standards (NAAQS) through which the Environmental Protection Agency (EPA) set limits on the concentration of six criteria pollutants. Responding to clear scientific consensus on the wide ranging health impacts of fine particulate matter (PM_{2.5}) on human health (“The Lancet Commission on pollution and health”, 2018), the EPA included PM_{2.5} in its criteria pollutants in 1997, but legal challenges delayed implementation of the new standard until 2005.

Given the substantial health and economic costs associated with PM_{2.5} exposure, recent literature seeks to quantify the causal effects of the CAA on pollution levels and its role in reducing racial disparities in exposure. Currie et al. (2023) leveraged the quasi-experimental implementation of the 2005 PM_{2.5} standard in a Difference-in-Differences (DiD) framework. They find that the CAA reduced PM_{2.5} levels by 1.52 $\mu\text{g}/\text{m}^3$ in non-attainment areas—suggesting it contributed 62 percent of the reduction in the pollution exposure gap between White and Black Americans. Sager and Singer (2024) refines these estimates, showing that the standard difference-in-differences framework overstate the impacts of the CAA on PM_{2.5} levels due to pre-existing trends. Their revised DiD with baseline controls (DiDwb) suggests a smaller reduction of roughly 0.49 $\mu\text{g}/\text{m}^3$.

There are two methodological challenges in these studies. First, monitoring of $PM_{2.5}$ levels is expensive, leading to sparse measurements of pollution across the country. Second, due to this sparsity, these monitors are often placed in urban, high-pollution areas, which are correlated with many demographic characteristics. Absent a proper accounting for this endogeneity, the estimated impact of the CAA on $PM_{2.5}$ levels would be suffer from selection bias, or overstate the impact of regulation by conflating it with pre-existing pollution trends in more polluted, demographically distinct areas. To address both of these limitations, Currie et al. (2023) and Sager and Singer (2024) use predicted- $PM_{2.5}$ data for a spatially complete source of pollution concentrations that fill in the gaps left by traditional, *in situ* monitoring networks. While Currie et al. (2023) and Sager and Singer (2024) acknowledge downward bias at the tails of the $PM_{2.5}$ distribution, they argue that prediction errors are likely small and if the resulting measurement error is classical, only attenuating the estimated coefficients. They support this argument citing these models perform well on average and match ground truth with very high, in-sample measures of fit and high agreement across datasets. (Di et al., 2019, 2016; Meng et al., 2019; Van Donkelaar et al., 2021) In the end, these studies effectively treat the predicted- $PM_{2.5}$ data as “truth”.¹

In the second chapter of this dissertation, I emphasized that these predicted- $PM_{2.5}$ data are not “truth” and that there is no “one-size-fits-all” dataset for $PM_{2.5}$. I showed two important limitations of these predicted- $PM_{2.5}$ data that are particularly relevant in this empirical setting. First, reported performance metrics of predicted- $PM_{2.5}$ models in the literature are likely overinflated due to improper validation strategies that allow data leakage across training and

¹ Sager and Singer (2024) does conduct robustness check using several alternative predicted- $PM_{2.5}$ datasets, but these datasets are all based on the same underlying model and suffer from the same limitations.

validation sets and are likely overfitting the training data. While this suggest performance in-sample (near monitors) may be high, performance out-of-sample (in areas without monitors) is much poorer. Second, these estimates are highly uncertain, and are likely to exhibit nonclassical measurement error. Together, these two limitations suggest that treating these predicted-PM_{2.5} data as “truth” is inappropriate.

In this chapter, I revisit the analysis using our tailored predicted-PM_{2.5} data. First, I replicate the analysis of Sager and Singer (2024) first using the original predicted-PM_{2.5} data, then using our re-evaluated predicted-PM_{2.5} data. There is little difference in the results of the two analyses. Second, to address the endogeneity inherent in the performance of the predicted-PM_{2.5} models, I produce propensity scores for each 1km grid cell in the United States that represent the probability of a monitor being placed in that area. I use these propensity scores to weight a DiD framework, placing more weight on monitored areas that look similar to areas without monitors and unmonitored areas that look similar to areas with monitors.

The remainder of this chapter proceeds as follows. Section 3.2, briefly reviews the data, regulatory context, and methodological details used in Currie et al. (2023) and Sager and Singer (2024), and including the binary classification model used to predict monitor placement. Section 3.3, presents the main empirical results, highlighting sensitivity to methodological choices. Section 3.4 concludes.

3.2 Data and Methodology

For more detail on the methodology and data used in these analyses, I refer the reader to these two studies. I briefly review them below.

3.2.1 Data. Regulatory Context The Clean Air Act (CAA) establishes National Ambient Air Quality Standards (NAAQS), enforced by the U.S. Environmental Protection Agency (EPA). In April 2005, the EPA began enforcement of the 1997 annual PM_{2.5} standard, which set the acceptable annual concentration of fine particulate matter (PM_{2.5}) at 15 µg/m³ based on the average of the previous three years (2001–2003). Areas exceeding this threshold were classified as “non-attainment,” triggering mandatory regulatory action, including state implementation plans (SIPs), stricter permitting processes for industrial facilities, and potential withholding of federal transportation funding. The designation was based solely on historical monitor data, making non-attainment status plausibly exogenous and suitable for quasi-experimental analysis.

Attainment Data: Non-attainment counties were identified using data from the EPA Green Book and Federal Register, which record EPA designations and registered PM_{2.5} “design values” (DVs) derived from ground-level monitors from 2002 to 2004. Counties without monitors but located within broader non-attainment zones were assigned the DV of the most polluted monitored county within their non-attainment area. Of approximately 3,000 counties in the contiguous United States, 208 counties were designated as non-attainment based on the annual PM_{2.5} standard. The map in 13 shows the design value for the baseline period in 1 km by 1 km cells as well as the 208 nonattainment counties.

Census Data: Population counts and demographics at the census-block level were obtained from the 2000 and 2010 Decennial Censuses. These data enabled population-weighted aggregation of pollution concentrations to census tracts and were used as regression weights. The data also facilitated examination of exposure disparities by race, specifically for Black and White populations.

3.2.2 Empirical Methodology. The primary analytic period includes pre-treatment averages (2002–2004) and post-treatment averages (2006–2008 and 2011–2013). This structure mitigates short-term fluctuations and measurement errors. The analysis uses publicly available data and focuses exclusively on evaluating the effectiveness of the CAA’s PM_{2.5} non-attainment designations on subsequent changes in pollution levels, without incorporating housing price data.

Following Sager and Singer (2024), I implement two primary empirical methods: Difference-in-Differences (DiD) and Difference-in-Differences with Baseline Controls (DiDwb). The standard DiD approach compares changes in PM_{2.5} concentrations from the pre-treatment to post-treatment periods between treated (non-attainment) and control (attainment) areas. Formally, the basic DiD specification is:

$$\Delta PM_{it} = \alpha + \beta \text{NonAttainment}_i + \Delta \varepsilon_{it} \quad (3.1)$$

where ΔPM_{it} is the change in pollution for area i , NonAttainment_i is an indicator for non-attainment designation, and β captures the causal effect of the CAA designation. Following this specification, I interact the non-attainment indicator with year dummies to estimate the effect of the CAA over time to generate an event study of the CAA on PM_{2.5} levels. This event study is shown in Figure 15, including both the original predicted-PM_{2.5} data from Meng et al. (2019) and our re-evaluated predicted-PM_{2.5} data generated in Chapter 4 of this dissertation (PM_{2.5} SPCV–SPCV).

The DiDwb method augments this by directly controlling for baseline PM_{2.5} levels, addressing potential confounding due to differential trends related to initial pollution levels. The specification for DiDwb is:

$$\Delta PM_{it} = \alpha + \beta \text{NonAttainment} * i + \gamma PM * i, pre + \Delta \varepsilon_{it} \quad (3.2)$$

where $PM_{i,pre}$ is the baseline pollution average for the period 2002–2004.

3.2.3 Predicting Monitor Placement. To highlight the limitations of the predicted-PM_{2.5} to address the endogeneity of monitor placement, I estimate propensity scores for each 1km grid cell in the United States. These propensity scores represent the probability of a monitor being placed in that area, given the characteristics of the area. Specifically, I construct a binary classification model that predicts whether a 1km grid cell hosts a monitor based on observable characteristics of the area—trained on the same set of features used in the predicted-PM_{2.5} models. This model captures the factors influencing monitor placement decisions in an attempt to account for the endogeneity of monitor placement.

The binary classification model employs a gradient boosting framework (LightGBM), chosen for its flexibility, predictive accuracy, and ability to handle large datasets in a computationally efficient manner. This model is trained on a set of features including distance to nearest monitor, road density, land-use classification, weather patterns, and outputs from chemical transport models, among others. These features are aggregated into a cross-sectional dataset between 2002 and 2019, with the target variable indicating whether a monitor was ever placed in the grid cell during this period.

To evaluate the model's predictive performance, I use a 5-fold cross-validation approach with an oversampling strategy to address class imbalance, as monitors are sparse across the landscape. This balancing technique trains the model by oversampling the minority class (monitored areas) giving it greater weight in the training process and improving the model's performance in predicting monitor placement.² Model hyperparameters are optimized using a bayesian optimization approach during cross-validation. Performance is measured using the Precision-Recall Area Under the Curve (PR AUC) metric which is suitable for imbalanced datasets. The PR AUC focusing on the precision and recall of the positive class (monitored areas). Following the training process, I removed unimportant features using a feature importance threshold of 0.005, retaining only the most influential predictors.

Figure 14 plots the predicted propensity scores for each 1km grid cell in the United States. The color gradient highlights the varying degree of similarity between monitored and unmonitored areas, closely following major population centers. The final model achieves a PR AUC of 0.53, indicating a modest predictive performance in distinguishing monitored from unmonitored areas.

3.3 Results

Table 3 presents estimates of the impact of nonattainment designation under the CAA on $PM_{2.5}$ concentrations at both the tract and pixel levels. Estimates are shown separately for two periods: a short-run period (2002–2004 to 2006–2008, Panel A) and a longer-run period (2002–2004 to 2011–2013, Panel B). Each panel compares results from the standard DiD approach, DiD with baseline controls (DiDwb), DiD with SPCV–SPCV predictions, DiDwb with

² In addition to the oversampling strategy, I also use an undersampling and Synthetic Minority Over-sampling Technique (SMOTE) strategy to balance the classes. The results of these models are similar, though the oversampling strategy performed best.

SPCV–SPCV, and DiD with propensity score weights (DiDp) using SPCV–SPCV predictions.

In panel A, we replicate the results of Table 1 in Sager and Singer (2024) using the original predicted-PM_{2.5} data from Meng et al. (2019). The results suggest that nonattainment designation reduced PM_{2.5} concentrations by 1.47 $\mu\text{g}/\text{m}^3$ (Column 1). Incorporating baseline controls (Column 2) reduces this estimate to 0.49 $\mu\text{g}/\text{m}^3$. Estimates using SPCV–SPCV predictions (Column 3) show a reduction of 0.75 $\mu\text{g}/\text{m}^3$, and when combined with baseline controls (Column 4), the estimate is 0.32 $\mu\text{g}/\text{m}^3$. The DiDp estimate (Column 5) using propensity score weights shows a reduction of 0.43 $\mu\text{g}/\text{m}^3$.

Pixel-level results, available only with SPCV–SPCV predictions, exhibit smaller estimated effects. This is likely due a few reasons: (1) Pixel-level data introduces more within county variation in PM_{2.5} that is uninformative for identification and adds noise to the estimates, attenuating the estimated effects; (2) Reduction in ecological fallacy, as pixel-level data allows for more precise identification of treatment effects at the local level; and (3) Large sample size. Standard DiD suggest modest reductions of 0.08 $\mu\text{g}/\text{m}^3$ (Column 3), which is reduced to 0.03 $\mu\text{g}/\text{m}^3$ with baseline controls (Column 4). The DiDp estimate (Column 5) shows a similar reduction of 0.03 $\mu\text{g}/\text{m}^3$.

Panel B, examining longer-term impacts, reveal similar but amplified patterns, Tract-level DiD estimates indicate a reduction of 2.35 $\mu\text{g}/\text{m}^3$ (Column 1), which is reduced to 0.56 $\mu\text{g}/\text{m}^3$ with baseline controls (Column 2). Using SPCV–SPCV predictions (Column 3) yields a reduction of 1.09 $\mu\text{g}/\text{m}^3$, and with baseline controls (Column 4), the estimate is 0.07 $\mu\text{g}/\text{m}^3$. The DiDp estimate (Column 5) are more moderate, at 1.00 $\mu\text{g}/\text{m}^3$.

Pixel-level estimates in Panel B continue to show smaller magnitude effects. The standard DiD estimate is $-0.16 \mu\text{g}/\text{m}^3$ (Column 3), reduced to $-0.08 \mu\text{g}/\text{m}^3$ with baseline controls (Column 4). Notably, propensity-weighted DiDp estimates (Column 5) are negligible or slightly positive at $0.02 \mu\text{g}/\text{m}^3$ (Column 5). This suggests that the true impact of the CAA on $\text{PM}_{2.5}$ levels may be smaller than previously estimated, particularly when accounting for the endogeneity of monitor placement. In general, the propensity-weighted DiDp are similar or smaller than the DiDwb estimates.

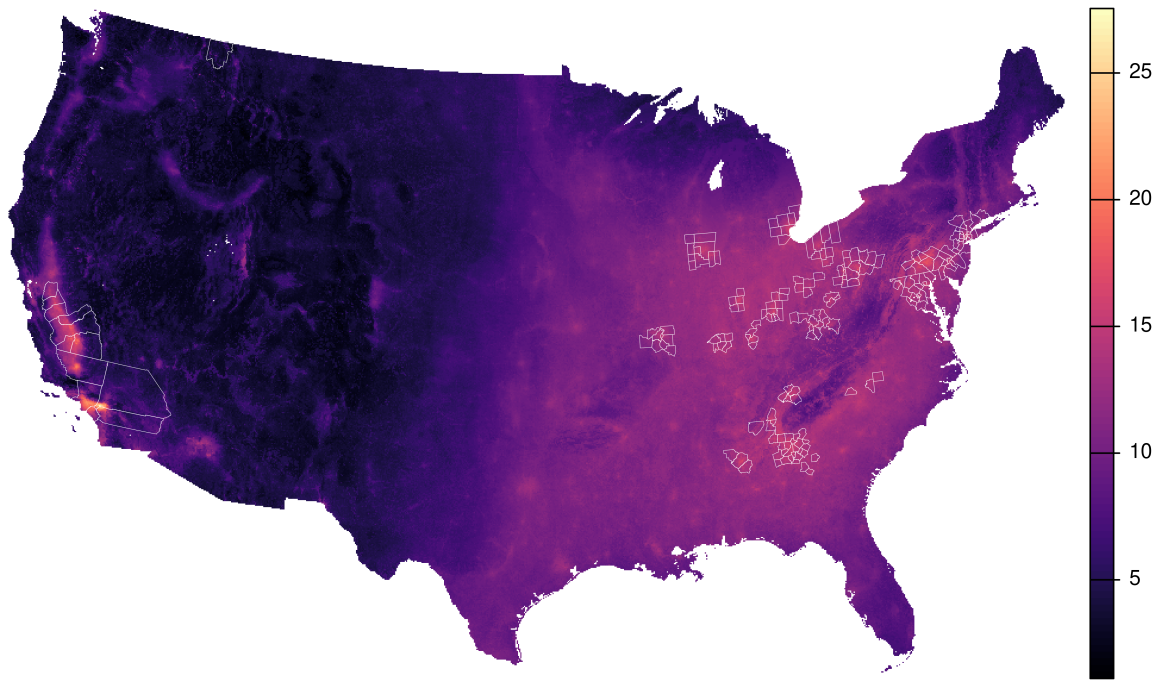
3.4 Conclusion

This chapter revisits previous literature evaluating the impact of the Clean Air Act's nonattainment designations on fine particulate concentrations. Replicating and extending the analysis of Sager and Singer (2024), I find that the results are generally robust to the use of our re-evaluated predicted- $\text{PM}_{2.5}$ data. However, major endogeneity concerns remain, particularly regarding the placement of monitors and the use of predicted- $\text{PM}_{2.5}$ data as a proxy for ground truth. Our estimate contain a large degree of uncertainty that is correlated with the placement of monitors. Furthermore, in my previous chapter, I showed that the model accuracy of the SPCV–SPCV is roughly an R^2 of 0.63, which is much lower than most of the literature suggests.

Our findings emphasize the necessity of rigorous validation and evaluation of predicted- $\text{PM}_{2.5}$ models, particularly in the context of regulatory impact assessments. The use of propensity scores to weight the DiD analysis provides a promising avenue for addressing endogeneity concerns, but further research is needed to refine these methods and improve the accuracy of predicted- $\text{PM}_{2.5}$ data.

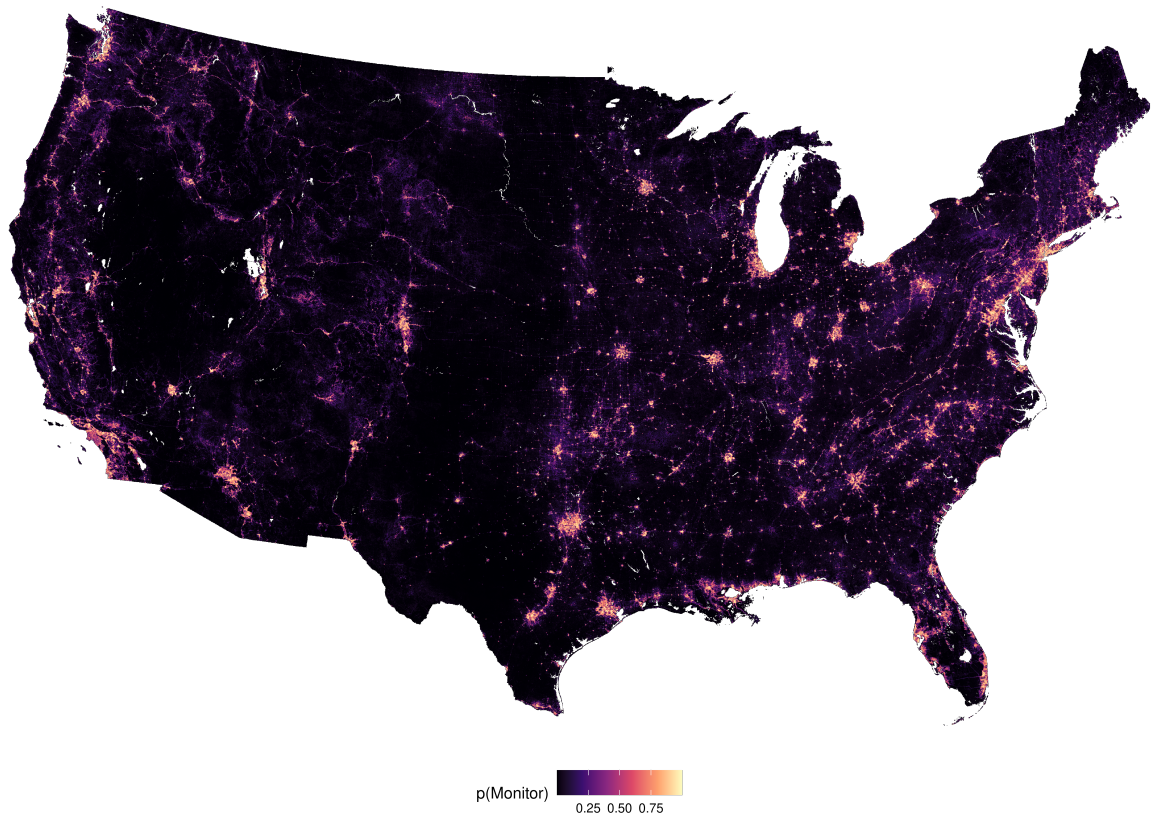
3.5 Figures and Tables

Figure 13. PM_{2.5} Design Values and nonattainment counties



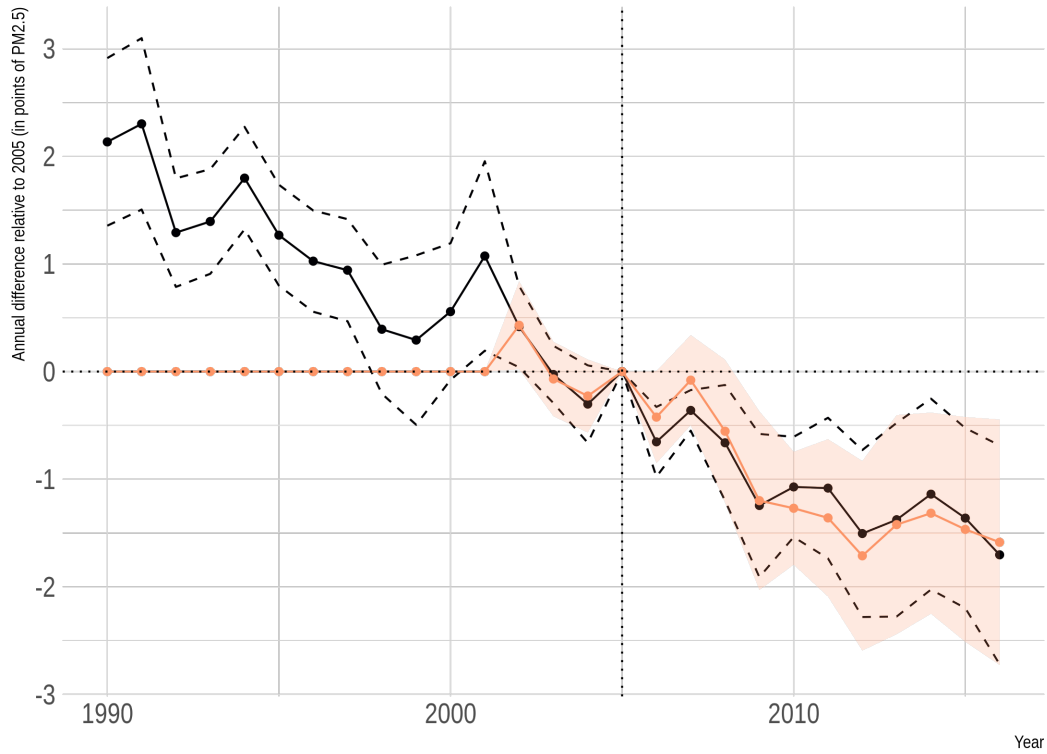
This figure shows baseline (2002-2004) PM_{2.5} Design Values at a 1 km by 1 km resolution using SPCV-SPCV predictions generated in Chapter 4 of this dissertation. Counties drawn in white are designated in nonattainment in 2005.

Figure 14. PM_{2.5}



This figure shows the predicted propensity scores for each 1km grid cell in the United States. The color gradient highlights the varying degree of similarity between monitored and unmonitored areas, closely following major population centers.

Figure 15. Event Study of the CAA on $PM_{2.5}$ Levels



This figure shows coefficient estimates from a regression that includes a treated dummy interacted with years, controlling for year and tract level fixed effects. In yellow, dotted line show point estimates of our $PM_{2.5}$ (SPCV-SPCV) predictions (2002-2016) and the shaded area shows 95 percent confidence intervals. In black, the dotted line shows the point estimates replicate Figure 3 Panel B of Sager and Singer (2024) using the original predicted- $PM_{2.5}$ (Meng et al., 2019) (2000-2016) and the dashed line shows the 95 percent confidence intervals.

Table 3. Nonattainment Status and Changes in PM_{2.5}

<i>Part A: 2002–2004 to 2006–2008</i>						
		Meng		SPCV–SPCV		
		DiD	DiDwb	DiD	DiDwb	DiDp
		(1)	(2)	(3)	(4)	(5)
Tract-level	Nonattainment	-1.47	-0.49	-0.58	-0.09	-0.35
		0.34	0.10	0.07	0.01	0.08
	N	72,043	72,043	63,618	63,618	63,814
Pixel-level	Nonattainment	–	–	-0.07	-0.04	-0.02
		–	–	(0.01)	(0.01)	(0.01)
	N	–	–	5,022,687	5,022,687	5,022,687
<i>Part B: 2002–2004 to 2011–2013</i>						
		Meng		SPCV–SPCV		
		DiD	DiDwb	DiD	DiDwb	DiDp
		(1)	(2)	(3)	(4)	(5)
Tract-level	Nonattainment	-2.35	-0.56	-1.05	-0.02	-0.90
		(0.27)	(0.10)	(0.08)	(0.08)	(0.09)
	N	72,043	72,043	63,618	63,618	63,814
Pixel-level	Nonattainment	–	–	-0.15	-0.09	0.02
		–	–	(0.03)	(0.02)	(0.013)
	N	–	–	5,022,687	5,022,687	5,022,687

Notes: The table shows coefficient estimates for the treatment effect of nonattainment status on change in PM_{2.5} levels between pre- and post-treatment periods. Each panel-column combination reports separate regressions. Column (1) uses simple DiD; column (2) uses DiD with baseline controls; column (3) uses DiD with SPCV–SPCV predictions from Chapter 4 of this dissertation; column (4) uses DiD with baseline controls and SPCV–SPCV predictions; and column (5) uses DiD with SPCV–SPCV predictions and propensity score weights. The first two rows report tract-level estimates, while the last two rows report pixel-level estimates. Standard errors are reported in parentheses and clustered at the county level.

CHAPTER IV

HIGH RESOLUTION GEOSPATIAL DATABASE OF PM_{2.5} CONCENTRATIONS IN THE CONTIGUOUS UNITED STATES (2002–2019)

4.1 Introduction

Fine-scale estimates of ambient air pollutant concentrations are increasingly important tools in environmental and health research. In fields such as public health, geography, and environmental economics, such data enables the study of exposure patterns, health impacts, and policy effectiveness at previously inaccessible spatial resolutions.

A growing class of empirical models, ranging from traditional land-use regression to modern machine learning algorithms, aims to predict pollutant concentrations at locations and times where direct measurements are unavailable. These methods leverage observed pollution data and a set of spatiotemporal predictors to produce fine-scale estimates, often at regional, national, or global scales (Meng et al., 2019; Reid et al., 2021; Van Donkelaar et al., 2021).

Advances in remote sensing and computational methods have expanded both the inputs and methodologies used in pollution prediction models. Modern approaches often incorporate satellite-derived Aerosol Optical Depth (AOD), meteorological reanalysis products, chemical transport model outputs, and high-resolution land use data. Concurrently, statistical models have evolved from simple regression-based approaches to more complex machine learning-based methods capable of capturing nonlinearities and interactions across high-dimensional input spaces.

Despite their increasing popularity and sophistication, predictive models of pollution exposure are often treated as black boxes by applied researchers. Many existing datasets are proprietary or inadequately documented, limiting

transparency and replicability. Even when inputs and outputs are openly available, the modeling choices—including feature selection, imputation strategies, validation procedures, and uncertainty quantification—are seldom disclosed, leaving users to either rely on pre-generated estimates with limited context or to embark on labor-intensive efforts to reconstruct datasets and processing workflows from scratch.

This chapter addresses these challenges by developing a fully transparent and reproducible modeling pipeline for predicting monthly average PM_{2.5} concentrations across the contiguous United States (CONUS) at a 1 km by 1 km resolution, from 2002 to 2019.

To assess model generalizability, we employ a nested, spatial cross-validation strategy. In companion work, we show that the choice of validation strategy can substantially influence model performance metrics. In particular, standard observation-level cross-validation tends to overstate performance due to spatial autocorrelation and data leakage. We advocate for spatially explicit validation strategies that better reflect the intended use case of predicting pollutant concentrations in unmonitored areas.

The modeling pipeline developed here is structured with transparency and reproducibility in mind, clearly documenting data processing steps, methodological decisions, and validation strategies. While designed with the intention of future public dissemination, the pipeline is currently in an development stage. This structured approach ensures methodological clarity and positions the work for eventual sharing and use by the broader research community in environmental health and policy.

4.2 Data and Methods

4.2.1 Monitoring Data. The primary training data consist of regulatory-grade measurements of $\text{PM}_{2.5}$ collected from the U.S. Environmental Protection Agency’s Air Quality System (AQS). These data include 24-hour average $\text{PM}_{2.5}$ concentrations from monitors distributed across the contiguous United States. Observations span the period from 2002 to 2019 and are aggregated from daily to monthly means to serve as the outcome variable in the modeling framework.

These monitors are not uniformly distributed. They tend to be concentrated in urban and industrial areas, with sparse coverage in rural regions. The number and location of active monitors vary across time due to instrument deployment, relocation, or decommissioning. On average, approximately 1,452 monitors report observations in any given month. The irregular and incomplete spatial coverage of these monitors motivates the use of spatially explicit prediction models to estimate pollution levels in unmonitored areas.

4.2.2 Covariates. We assemble a diverse set of spatial and temporal covariates known to influence $\text{PM}_{2.5}$ concentrations. These covariates are drawn from satellite observations, reanalysis data, chemical transport models, and land use sources. All variables are resampled to a common $1 \text{ km} \times 1 \text{ km}$ grid using bilinear interpolation or nearest-neighbor aggregation, depending on data type.

Aerosol Optical Depth (AOD) measurements are obtained from the MODIS MAIAC product at 470 nm and 550 nm wavelengths, derived from the Aqua and Terra satellite platforms. MODIS surface reflectance (Band 3) is included to address retrieval bias. To enhance temporal coverage and address missingness due to cloud cover and other retrieval limitations, we incorporate aerosol-related variables from the MERRA-2 reanalysis product.

Chemical transport model outputs are included to provide physically based estimates of pollutant levels. We use total and component-specific $\text{PM}_{2.5}$ concentrations from the CMAQ model, available at a 12 km resolution and downscaled to the 1 km target grid. These simulations are treated as predictive features rather than targets.

Meteorological variables are drawn from the NCEP North American Regional Reanalysis. We aggregate daily data to monthly means and include variables such as temperature, precipitation, wind speed, boundary layer height, and humidity, among others.

To characterize local emissions and terrain, we incorporate land cover classifications from the National Land Cover Database (NLCD), calculated as the modal class per grid cell. Road density is calculated from TIGER shapefiles. Elevation metrics (mean, standard deviation, minimum, and maximum) are extracted from the GMTED2010 digital elevation model. Vegetation activity is represented using monthly Normalized Difference Vegetation Index (NDVI) values from MODIS.

We additionally compute spatial lags of $\text{PM}_{2.5}$ using inverse distance weighting across a range of buffer distances, following approaches used in prior literature. These lags are designed to capture spatial dependence in pollutant levels while mitigating overfitting by excluding immediate neighbors.

All covariates are aligned to the analysis grid and subjected to preprocessing. For categorical variables, missing values are imputed using the most frequent category within each pixel over time. Continuous variables are imputed using a progressive rolling average scheme: missing values are filled using a 3-month rolling mean, extended to 5, 7, and 9-month windows if needed.

4.2.3 Model Training. We use LightGBM, a gradient boosting machine learning algorithm, to predict monthly average PM_{2.5} concentrations. Model training is conducted using the `mlr3` framework in R. The model objective is to minimize the mean squared error (MSE), defined as:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (4.1)$$

where y_i denotes the observed PM_{2.5} concentration at location i and $\hat{y}_i$ represents the corresponding model prediction. The trained model is then applied across all grid cells to generate a complete spatiotemporal surface of PM_{2.5} concentrations across the CONUS.

4.2.4 Model Validation. To assess model performance and ensure generalizability to unmonitored areas, we evaluate predictive accuracy using multiple cross-validation strategies. Standard k -fold cross-validation, which partitions monitor-month observations randomly, is widely used in the literature but may overstate performance due to spatial and temporal autocorrelation. For example, when data from the same monitor appear in both training and test sets, the model may exploit spatial proximity rather than learn generalizable patterns.

To address this limitation, we implement spatial cross-validation by partitioning the data into geographically distinct clusters. This approach ensures that the model is tested on locations that are spatially separate from those used in training, providing a more realistic evaluation of predictive performance.

We also apply nested resampling, a more rigorous evaluation framework that separates model tuning from model assessment. The outer loop partitions the data for performance evaluation, while the inner loop is used to optimize hyperparameters. We explore both spatial and non-spatial versions of this nested

strategy. This layered resampling scheme prevents overfitting to the validation set and yields a less biased estimate of true model performance.

All model evaluations are reported using standard metrics such as R^2 and root mean squared error (RMSE). The full modeling pipeline, including data sources, code, and output rasters, is made publicly available to facilitate transparency and reproducibility.

4.3 Results

4.3.1 Model Performance. The predictive performance of our LightGBM model is evaluated using spatial and nested cross-validation techniques to provide realistic and unbiased assessments of its accuracy. Using standard metrics, the model demonstrates strong predictive capability across the study region. Spatially explicit validation strategies yield an R^2 of approximately 0.64, indicating robust performance in predicting monthly average PM_{2.5} concentrations even in areas without direct monitoring. These results underscore the model's generalizability and suitability for generating reliable predictions in diverse geographic contexts.

4.3.2 Prediction Uncertainty. We further quantify prediction uncertainty through quantile regression, generating 95% prediction intervals for PM_{2.5} estimates. Across the contiguous United States, uncertainty exhibits spatial variation, with wider intervals observed in regions distant from monitoring networks, particularly rural areas and the western United States. The average width of the prediction interval is 6 $\mu\text{g}/\text{m}^3$, highlighting that predictions are generally precise yet sufficiently broad to reflect realistic uncertainty. This explicit quantification of uncertainty is essential for accurate interpretation of predictions in downstream health and environmental policy analyses.

4.3.3 Dataset and Repository Overview. A central contribution of this project is the fully open-source dataset and modeling pipeline made publicly accessible through our dedicated repository. This repository comprehensively documents each stage of data processing, model training, and prediction generation, allowing users to replicate and adapt our methods seamlessly.

The dataset provided includes clearly organized directories:

- **Raw Input Data (data-raw):** Original datasets sourced from census data, EPA EQUATES, GMTED2010 elevation, MERRA-2 reanalysis, MODIS AOD, NDVI, surface reflectance, air quality monitors, NCEP-NCAR reanalysis, NLCD land cover, and smoke data.
- **Processed Data (data-processed):** Cleaned and standardized datasets, including census, EPA EQUATES, MODIS AOD, NDVI, surface reflectance, monitor summaries, and combined MODIS products aligned at 1 km x 1 km resolution.
- **Additional Data (data-additional):** Supplementary processed data and resources for further analysis.

The repository also includes fully documented R scripts, structured into sequential processing steps:

- **000 – Download raw data:** Scripts for downloading datasets from public sources (EPA EQUATES, MODIS products, census data).
- **001 – Process raw data:** Scripts to preprocess raw data into analysis-ready formats.
- **002 – Join data with monitors:** Integrates processed datasets with air quality monitor observations.

- **003 – Unfold raster into dataframe:** Converts raster data into structured data frames suitable for downstream analysis.
- **004 – Predict:** Prepares predictor variables, conducts spatial modeling, and generates PM_{2.5} predictions.

Figure 16 illustrates a stacked raster of monthly average PM_{2.5} predictions at a 1 km by 1 km resolution, generated using the SPCV–SPCV cross-validation.

Detailed documentation clearly outlines each script’s purpose, inputs, and outputs, facilitating easy execution and modification. The repository ensures transparency, reproducibility, and methodological rigor, substantially lowering barriers to entry for researchers aiming to validate, replicate, or extend our predictive modeling pipeline.

4.4 Conclusion

This study develops a transparent and reproducible PM_{2.5} prediction pipeline, enhancing methodological rigor and clarity in environmental exposure research. By systematically documenting data processing steps, modeling decisions, and validation procedures, the pipeline addresses critical gaps in existing datasets and methodological transparency. While currently an internal resource, this modeling framework is structured for eventual public release, positioning it to support future environmental health research and informed policy-making efforts effectively.

4.5 Figures

Figure 16. Stacked Raster of Monthly Average PM_{2.5} Predictions



This figure shows monthly predicted-PM_{2.5} concentrations for the contiguous United States at a 1 km by 1 km resolution using SPCV-SPCV predictions. Layers are stacked to show the temporal progression of PM_{2.5} concentrations across the first four months of 2017.

APPENDIX A
APPENDIX TO CHAPTER ONE

Table A.4. GitHub events with proportional breakdowns

Name	Description	Proportion
Push	One or more commits are pushed to a repository branch or tag	0.466
Issue comment	Activity related to an issue or pull request comment	0.112
Create	A Git branch or tag is created	0.108
Watch	When someone stars a repository	0.086
Pull request	Activity related to pull requests	0.078
Issues	Activity related to issues	0.051
Pull request review comment	Activity related to pull request review comments	0.034
Delete	A Git branch or tag is deleted	0.027
Fork	A user forks a repository	0.021
Gollum	When a wiki page is created for a repository	0.008
Release	Activity related to a release	0.004
Member	Activity related to repository collaborators	0.003
Public	When a private repository is made public	0.001

Notes: The sample includes observations of all GitHub events in the 50 most-active cities between 2013–2019. The proportional breakdown is of all events three weeks before and after DST (representing 16,394 city-day observations of 174,505 unique users).

Table A.5. 50 most-active GitHub cities (ranked by total events) in the Northern Hemisphere in 2013–2019

Rank	City	Events	Pushes	Early / Late / Never DST
1	San Francisco	26,220,671	11,464,699	Early
2	London	19,511,879	8,840,797	Late
3	New York	18,571,890	9,123,456	Early
4	Seattle	16,706,522	7,232,050	Early
5	Paris	10,261,588	4,729,269	Late
6	Berlin	9,999,729	4,087,816	Late
7	Tokyo	9,889,374	4,539,373	Never
8	Boston	9,405,851	4,396,030	Early
9	Los Angeles	7,365,915	3,474,078	Early
10	Moscow	6,312,342	3,006,135	Never
11	Toronto	6,174,122	2,934,192	Early
12	Chicago	5,960,846	2,946,564	Early
13	Austin	5,834,262	2,738,333	Early
14	Portland	5,793,351	2,712,860	Early
15	Bengaluru	5,359,385	2,542,372	Never
16	Washington	4,941,728	2,398,345	Early
17	Amsterdam	4,295,359	1,986,714	Late
18	Vancouver	4,219,302	1,948,889	Early
19	Montreal	3,983,904	1,876,482	Early
20	Denver	3,782,576	1,766,305	Early
21	Stockholm	3,782,445	1,810,495	Late
22	Munich	3,444,518	1,520,287	Late
23	Seoul	3,434,359	1,774,131	Never
24	Madrid	3,375,069	1,532,040	Late
25	Kyiv	3,373,661	1,665,148	Late
26	Atlanta	3,135,574	1,608,506	Early
27	Philadelphia	2,880,876	1,373,885	Early
28	Zurich	2,877,658	1,282,288	Late
29	San Diego	2,853,542	1,477,020	Early
30	Delhi	2,604,071	1,185,269	Never
31	Hamburg	2,536,831	1,161,703	Late
32	Prague	2,483,811	1,095,295	Late
33	Pittsburgh	2,359,704	1,269,121	Early
34	Raleigh	2,313,654	991,554	Early
35	Oslo	2,295,868	1,092,017	Late
36	Warsaw	2,258,030	1,064,295	Late
37	Saint Petersburg	2,170,095	968,918	Never
38	Minneapolis	2,051,118	971,689	Early
39	Copenhagen	2,029,562	914,542	Late
40	Cambridge	2,012,722	969,881	Late
41	Helsinki	1,995,306	1,000,439	Late
42	Dallas	1,972,359	960,276	Early
43	Phoenix	1,845,819	869,805	Never
44	Barcelona	1,841,706	905,971	Late
45	Dublin	1,821,963	924,427	Late
46	Tulsa	1,798,579	772,251	Early
47	Ottawa	1,724,724	794,807	Early
48	Cologne	1,677,009	752,390	Late
49	Lyon	1,600,373	733,745	Late
50	Budapest	1,579,333	790,950	Late

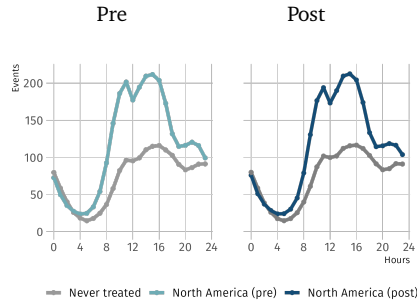
Table A.6. Daylight Saving Time across treated locations, 2013–2019

Year	North America	European Union
2013	March 10	March 31
2014	March 9	March 30
2015	March 8	March 29
2016	March 13	March 27
2017	March 12	March 26
2018	March 11	March 25
2019	March 10	March 31

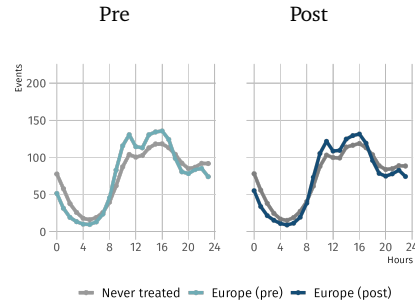
Figure A.17. Constructing hourly difference-in-differences estimates

A: Mean counts of total events by treatment status, DST ± 1 week

(i) North America and the never treated

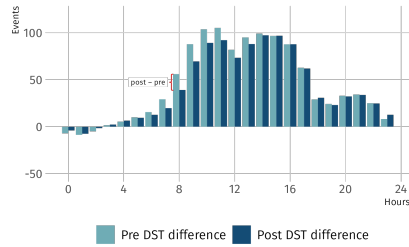


(ii) EU and the never treated

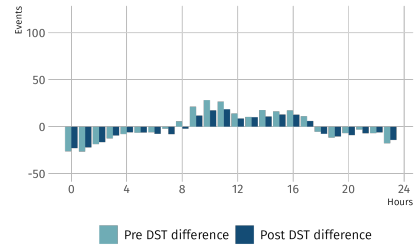


B: Post- and pre-DST differences in total number of events (“treatment – control”)

(i) North America



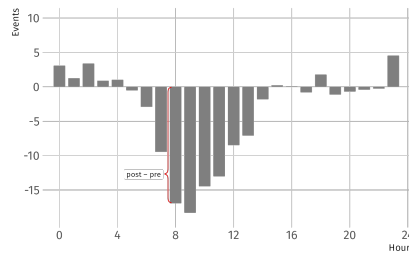
(ii) European Union



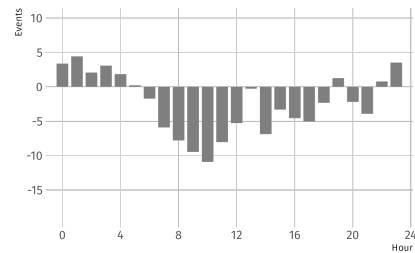
C: Difference in differences

(Post-DST “treatment – control” – Pre-DST “treatment – control”)

(i) North America



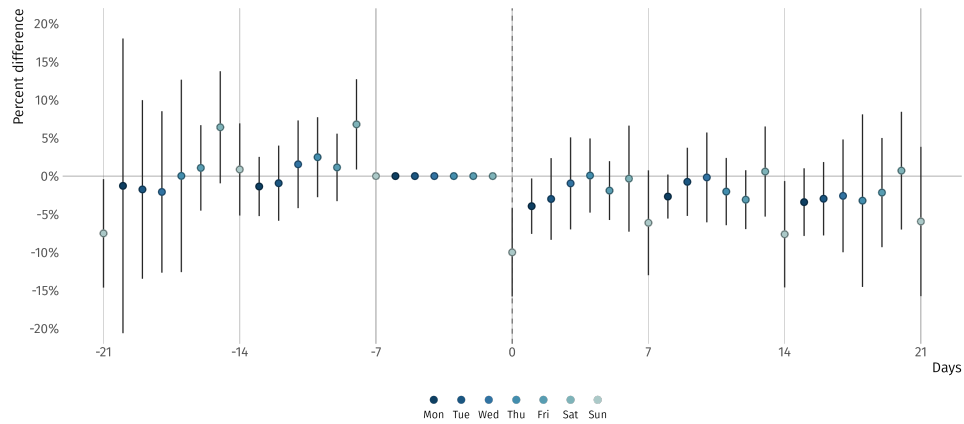
(ii) European Union



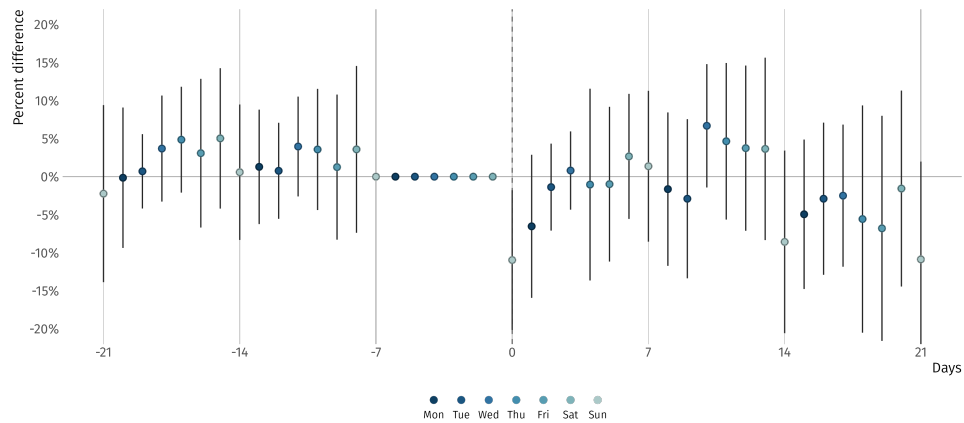
Notes: In Panel A we plot the mean number of hourly events in the one week before and one week after DST, separately for North America and Europe. In Panel B we plot the difference in means, subtracting hourly averages by treatment status separately for pre- and post-treatment period. In Panel C we plot the “second difference” in means, subtracting the post-treatment differences by pre-treatment differences.

Figure A.18. Do we see different responses across North America and Europe?

A: Total events, North America and never treated



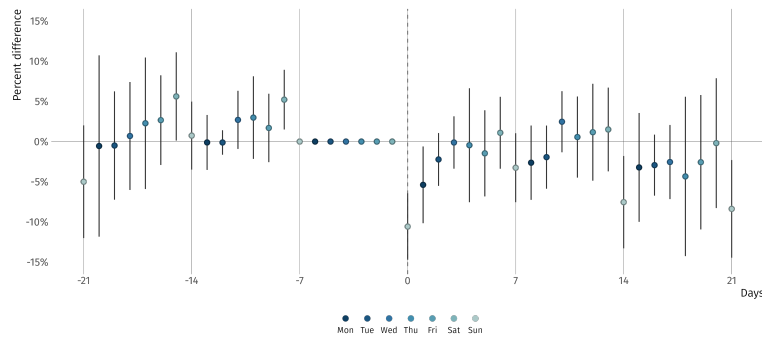
A: Total events, European Union and never treated



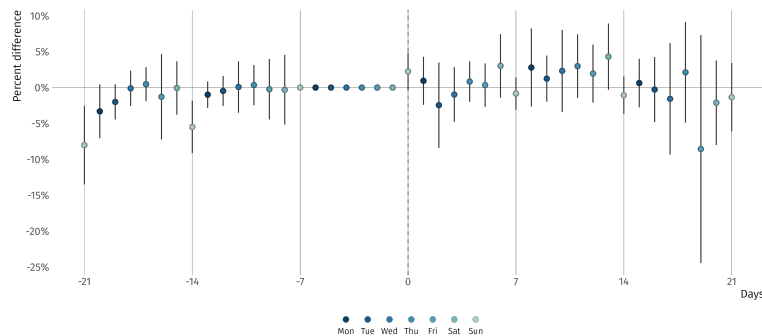
Notes: The sample includes the observations of the total (logged) number of GitHub events three weeks before and after DST, for the 50 most-active cities. In both panels we estimate seven separate models (i.e., one model for each day of the week). In Panel A we model the (logged) total number of GitHub events that originate in North America and never-treated cities (representing 10,374 city-day observations). In Panel B we model the (logged) total number of GitHub events that originate in Europe and never-treated cities (representing 9,772 city-day observations).

Figure A.19. Daily estimates of the effect of DST and the return to standard time on the total number of daily GitHub events?

A: Spring transition to DST, as in 2B.
Separate models (by day of week), with events
normalized to the week prior to treatment



B: Fall return to standard time.
Separate models (by day of week), with events
normalized to the week prior to treatment



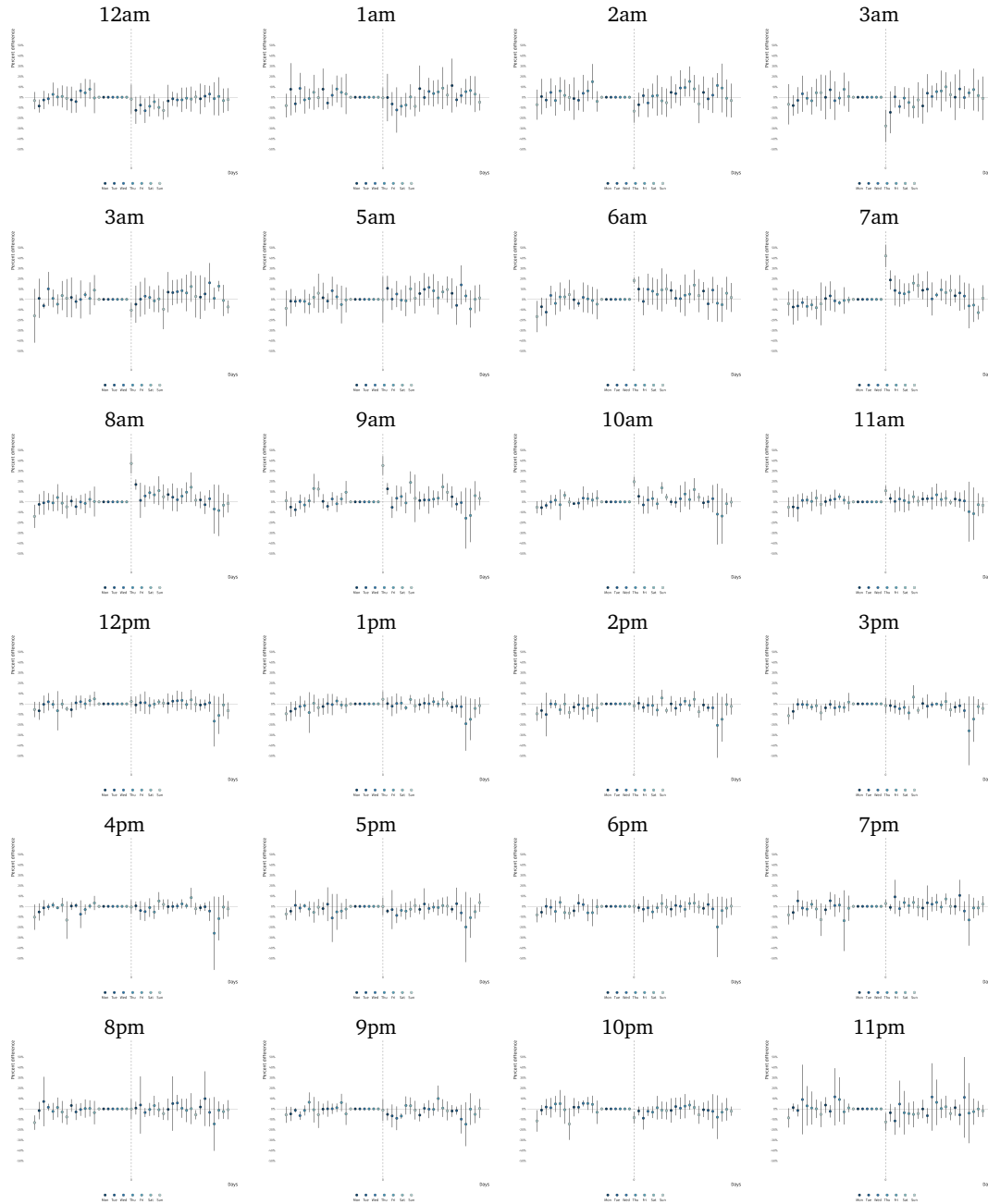
Notes: The sample includes the observations of the total (logged) number of GitHub events three weeks before and after DST, for the 50 most-active cities. Panel A is the same as was reported in ./figures/chapter01/Figure 2B. In Panel B we estimate similar models for the fall return to standard time—seven separate models (i.e., one model for each day of the week) where the city fixed effect then implicitly absorbs any day-of-week differences. In Panel A, observations represent 16,394 city-day observations of 174,505 unique users. In Panel B, observations represent 15,778 city-day observations of 174,505 unique users.

Figure A.20. Hour-specific estimates of the effect of DST on GitHub events



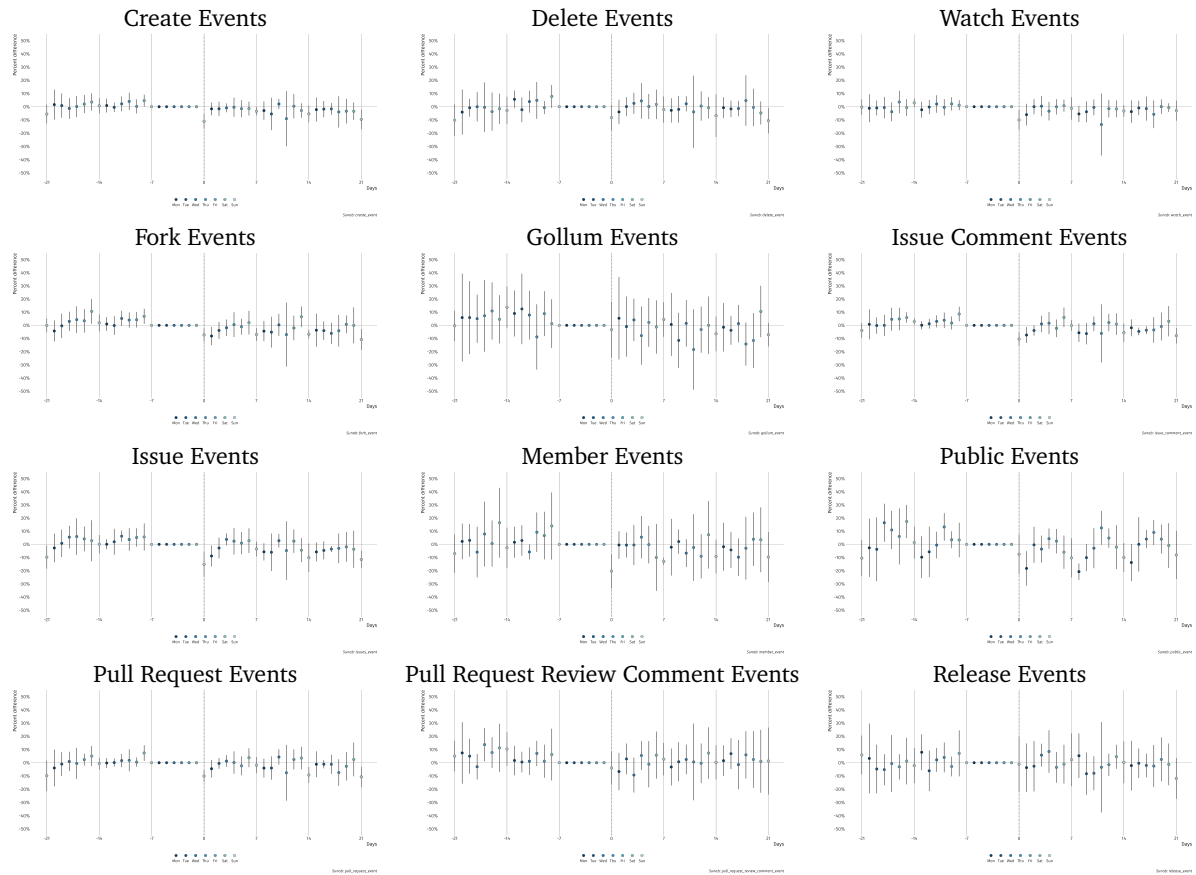
Notes: Estimates are extracted from models of the (logged) total number of GitHub events. The sample includes observations three weeks before and after DST (represent 393,162 city-hour observations of 174,505 unique users).

Figure A.21. Hour-specific estimates of the fall return to standard time



Notes: Estimates are extracted from models of the (logged) total number of GitHub events. The sample includes observations three weeks before and after the return to standard time (representing 378,672 city-hour observations of 174,505 unique users).

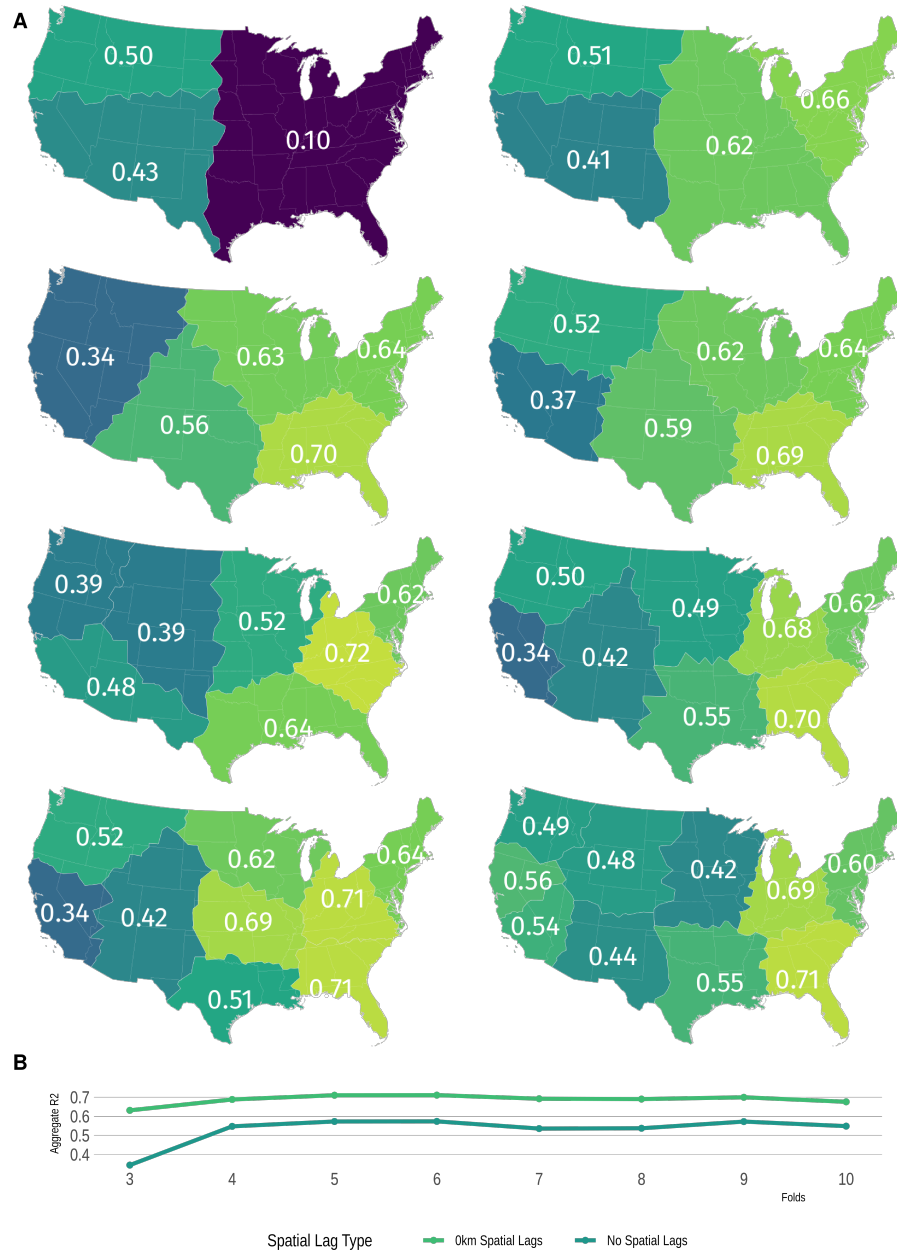
Figure A.22. Additional GitHub events, and the declines in activity induced by DST



Notes: Estimates are extracted from models of the (logged) number of events, separately by event category. (Recall, from Table A1, that “Pushes” account for 46.6 percent of all GitHub events—this figure collectively accounts for all other activities.) The sample includes observations three weeks before and after DST (representing 16,394 city-day observations of 174,505 unique users).

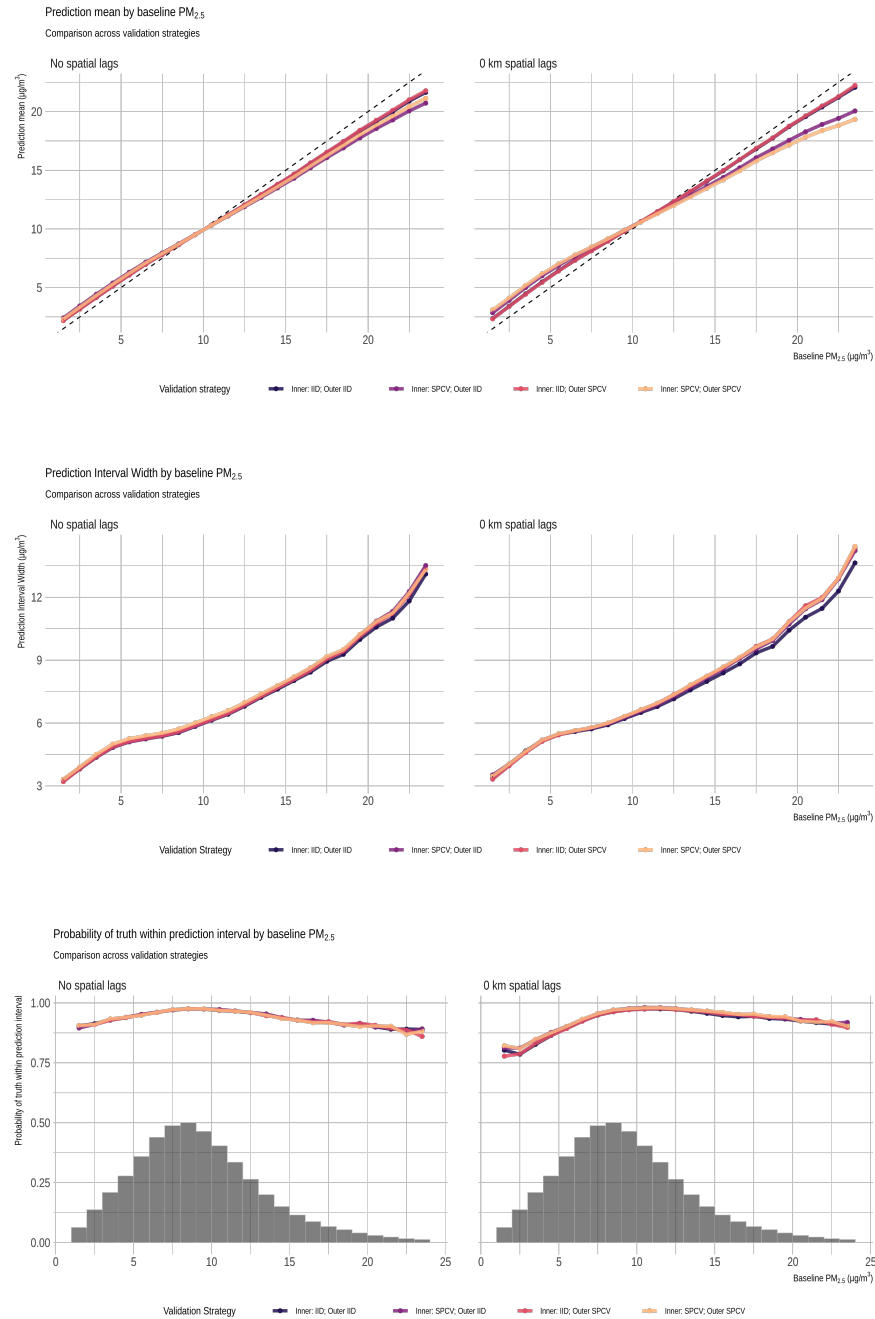
APPENDIX B
APPENDIX TO CHAPTER TWO

Figure B.1. Changes in PM_{2.5} prediction accuracy as the number of spatial cross-validation splits increases.



Each map presents the spatial distribution of the cross-validation splits for the outer loop of the nested resampling strategy. Filled values represent R^2 of the model trained on the split. The number of splits increases from left to right, with the top left map representing three splits and the bottom right map representing ten splits.

Figure B.2. In-sample $PM_{2.5}$ prediction accuracy.



Panel A compares binned predicted $PM_{2.5}$ values to binned true $PM_{2.5}$ values for pixels with monitors. Panel B plots prediction intervals across the same binned true $PM_{2.5}$ values. Panel C shows the probability that the true value falls within the prediction interval for each model.

REFERENCES CITED

- Alix-García, J., & Millimet, D. L. (2023, September). 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. Retrieved from <http://dx.doi.org/10.1086/723723> doi: 10.1086/723723
- Barnes, C. M., & Wagner, D. T. (2009). Changing to Daylight Saving Time cuts into sleep and increases workplace injuries. *Journal of Applied Psychology*, 94(5), 1305.
- Belzer, D., Hadley, S., & Chin, S.-M. (2008). *Impact of extended daylight saving time on national energy consumption report to congress* (Tech. Rep.). DOE/EE (USDOE Office of Energy Efficiency and Renewable Energy (EE)).
- Berk, M., Dodd, S., Hallam, K., Berk, L., Gleeson, J., & Henry, M. (2008). The effect of daylight saving. *Sleep and Biological Rhythms*, 6(1), 22–25.
- Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., . . . Lindauer, M. (2023). Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *WIREs Data Mining and Knowledge Discovery*, 13(2), e1484. Retrieved from <https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1484> doi: <https://doi.org/10.1002/widm.1484>
- Boser, A. (2024, March). Validating spatio-temporal environmental machine learning models: Simpson’s paradox and data splits. *Environmental Research Communications*, 6(3), 031003. Retrieved from <http://dx.doi.org/10.1088/2515-7620/ad2e44> doi: 10.1088/2515-7620/ad2e44
- Bünnings, C., & Schiele, V. (2021). Spring forward, don’t fall back: The effect of Daylight Saving Time on road safety. *Review of Economics and Statistics*, 103(1), 165–176.
- Childs, M. L., Li, J., Wen, J., Heft-Neal, S., Driscoll, A., Wang, S., . . . Burke, M. (2022). Daily local-level estimates of ambient wildfire smoke pm2.5 for the contiguous us. *Environmental Science & Technology*, 56(19), 13607-13621. Retrieved from <https://doi.org/10.1021/acs.est.2c02934> (PMID: 36134580) doi: 10.1021/acs.est.2c02934

- Colmer, J., Hardman, I., Shimshack, J., & Voorheis, J. (2020, July). Disparities in PM2.5 air pollution in the United States. *Science*, 369(6503), 575–578. Retrieved 2024-05-20, from <https://www.science.org/doi/10.1126/science.aaz9353> (Publisher: American Association for the Advancement of Science) doi: 10.1126/science.aaz9353
- Considine, E. M., Hao, J., deSouza, P., Braun, D., Reid, C. E., & Nethery, R. C. (2023). Evaluation of model-based pm2.5 estimates for exposure assessment during wildfire smoke episodes in the western u.s. *Environmental Science & Technology*, 57(5), 2031-2041. Retrieved from <https://doi.org/10.1021/acs.est.2c06288> (PMID: 36693177) doi: 10.1021/acs.est.2c06288
- Costa-Font, J., Fleche, S., & Pagan, R. (2024a). The labour market returns to sleep. *Journal of Health Economics*, 93, 102840.
- Costa-Font, J., Fleche, S., & Pagan, R. (2024b). The welfare effects of time reallocation: Evidence from daylight saving time. *Economica*, 91(362), 547–568.
- Currie, J., Voorheis, J., & Walker, R. (2023, January). What Caused Racial Disparities in Particulate Exposure to Fall? New Evidence from the Clean Air Act and Satellite-Based Measures of Air Quality. *American Economic Review*, 113(1), 71–97. Retrieved 2024-06-06, from <https://pubs.aeaweb.org/doi/10.1257/aer.20191957> doi: 10.1257/aer.20191957
- Czeisler, C. A., Duffy, J. F., Shanahan, T. L., Brown, E. N., Mitchell, J. F., Rimmer, D. W., ... others (1999). Stability, precision, and near-24-hour period of the human circadian pacemaker. *Science*, 284(5423), 2177–2181.
- Danesh Yazdi, M., Wang, Y., Di, Q., Zanobetti, A., & Schwartz, J. (2019). Long-term exposure to pm2.5 and ozone and hospital admissions of medicare participants in the southeast usa. *Environment International*, 130, 104879. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0160412019300364> doi: <https://doi.org/10.1016/j.envint.2019.05.073>
- Di, Q., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., ... Schwartz, J. (2019, September). An ensemble-based model of pm2.5 concentration across the contiguous united states with high spatiotemporal resolution. *Environment International*, 130, 104909. Retrieved from <http://dx.doi.org/10.1016/j.envint.2019.104909> doi: 10.1016/j.envint.2019.104909

- Di, Q., Kloog, I., Koutrakis, P., Lyapustin, A., Wang, Y., & Schwartz, J. (2016, April). Assessing pm2.5 exposures with high spatiotemporal resolution across the continental united states. *Environmental Science & Technology*, 50(9), 4712–4721. Retrieved from <http://dx.doi.org/10.1021/acs.est.5b06121> doi: 10.1021/acs.est.5b06121
- Di, Q., Wang, Y., Zanobetti, A., Wang, Y., Koutrakis, P., Choirat, C., ... Schwartz, J. D. (2017). Air pollution and mortality in the medicare population. *New England Journal of Medicine*, 376(26), 2513-2522. Retrieved from <https://www.nejm.org/doi/full/10.1056/NEJMoal702747> doi: 10.1056/NEJMoal702747
- Doleac, J. L., & Sanders, N. J. (2015). Under the cover of darkness: How ambient light influences criminal activity. *Review of Economics and Statistics*, 97(5), 1093–1103.
- Gaski, J. F., & Sagarin, J. (2011). Detrimental effects of daylight-saving time on SAT scores. *Journal of Neuroscience, Psychology, and Economics*, 4(1), 44.
- Gibson, M., & Shrader, J. (2018, December). Time use and labor productivity: The returns to sleep. *The Review of Economics and Statistics*, 100(5), 783–798. Retrieved 2022-04-04, from <https://direct.mit.edu/rest/article/100/5/783-798/58488> doi: 10.1162/rest_a_00746
- Giuntella, O., & Mazzonna, F. (2019, May). Sunset time and the economic effects of social jetlag: Evidence from US time zone borders. *Journal of Health Economics*, 65, 210–226. Retrieved 2022-04-04, from <https://linkinghub.elsevier.com/retrieve/pii/S0167629618309718> doi: 10.1016/j.jhealeco.2019.03.007
- Gregory-Allen, R. B., Jacobsen, B., & Marquering, W. (2010). The Daylight Saving Time anomaly in stock returns: Fact or fiction? *Journal of Financial Research*, 33(4), 403-427.
- Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., ... Townshend, J. R. G. (2013, November). High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science*, 342(6160), 850–853. Retrieved 2023-06-11, from <https://www.science.org/doi/10.1126/science.1244693> doi: 10.1126/science.1244693

- Hu, X., Belle, J. H., Meng, X., Wildani, A., Waller, L. A., Strickland, M. J., & Liu, Y. (2017, June). Estimating PM_{2.5} Concentrations in the Conterminous United States Using the Random Forest Approach. *Environmental Science & Technology*, 51(12), 6936–6944. Retrieved 2023-01-17, from <https://pubs.acs.org/doi/10.1021/acs.est.7b01210> doi: 10.1021/acs.est.7b01210
- Hu, Z., Huang, L., Zhai, X., Yang, T., Jin, Y., & Xu, Y. (2023, December). Quantifying Individual PM_{2.5} Exposure with Human Mobility Inferred from Mobile Phone Data. *Sustainability*, 16(1), 184. Retrieved 2023-12-29, from <https://www.mdpi.com/2071-1050/16/1/184> doi: 10.3390/su16010184
- Jain, M. (2020, January). The benefits and pitfalls of using satellite data for causal inference. *Review of Environmental Economics and Policy*, 14(1), 157–169. Retrieved from <http://dx.doi.org/10.1093/reep/rez023> doi: 10.1093/reep/rez023
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: with applications in r*. Springer. Retrieved from <https://faculty.marshall.usc.edu/gareth-james/ISL/>
- Jbaily, A., Zhou, X., Liu, J., Lee, T.-H., Kamareddine, L., Verguet, S., & Dominici, F. (2022, January). Air pollution exposure disparities across US population and income groups. *Nature*, 601(7892), 228–233. doi: 10.1038/s41586-021-04190-y
- Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016, August). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301), 790–794. Retrieved 2024-06-05, from <https://www.science.org/doi/10.1126/science.aaf7894> doi: 10.1126/science.aaf7894
- Jin, L., & Ziebarth, N. R. (2020). Sleep, health, and human capital: Evidence from daylight saving time. *Journal of Economic Behavior & Organization*, 170, 174–192.
- Kamstra, M. J., Kramer, L. A., & Levi, M. D. (2000). Losing sleep at the market: The daylight saving anomaly. *American Economic Review*, 90(4), 1005-1011.
- Kantermann, T., Juda, M., Mellow, M., & Roenneberg, T. (2007). The human circadian clock's seasonal adjustment is disrupted by Daylight Saving Time. *Current Biology*, 17(22), 1996–2000.

- Ke, G., Meng, Q., Finely, T., Wang, T., Chen, W., Ma, W., ... Liu, T.-Y. (2017, December). Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems 30 (nips 2017)*. Retrieved from <https://www.microsoft.com/en-us/research/publication/lightgbm-a-highly-efficient-gradient-boosting-decision-tree/>
- Kelly, J. T., Jang, C., Timin, B., Di, Q., Schwartz, J., Liu, Y., ... Bell, M. L. (2021, 05). Examining pm2.5 concentrations and exposure using multiple models. *Environmental Research*, 196, 110432. Retrieved from <https://linkinghub.elsevier.com/retrieve/pii/S0013935120313293> doi: 10.1016/j.envres.2020.110432
- Kotchen, M. J., & Grant, L. E. (2011). Does Daylight Saving Time save energy? Evidence from a natural experiment in Indiana. *Review of Economics and Statistics*, 93(4), 1172–1185.
- Kountouris, Y., & Remoundou, K. (2014). About time: Daylight Saving Time transition and individual well-being. *Economics Letters*, 122(1), 100–103.
- Kuehnle, D., & Wunder, C. (2016). Using the life satisfaction approach to value Daylight Saving Time transitions: Evidence from Britain and Germany. *Journal of Happiness Studies*, 17(6), 2293–2323.
- Lahti, T., Sysi-Aho, J., Haukka, J., & Partonen, T. (2011). Work-related accidents and Daylight Saving Time in Finland. *Occupational Medicine*, 61(1), 26–28.
- Lahti, T. A., Leppämäki, S., Lönnqvist, J., & Partonen, T. (2006). Transition to Daylight Saving Time reduces sleep duration plus sleep efficiency of the deprived sleep. *Neuroscience Letters*, 406(3), 174–177.
- The Lancet Commission on pollution and health. (2018, February). *The Lancet*, 391(10119), 462–512. Retrieved 2024-10-28, from [https://doi.org/10.1016/S0140-6736\(17\)32345-0](https://doi.org/10.1016/S0140-6736(17)32345-0) (Publisher: Elsevier) doi: 10.1016/S0140-6736(17)32345-0
- Lang, M., Binder, M., Richter, J., Schratz, P., Pfisterer, F., Coors, S., ... Bischl, B. (2019, dec). mlr3: A modern object-oriented machine learning framework in R. *Journal of Open Source Software*. Retrieved from <https://joss.theoj.org/papers/10.21105/joss.01903> doi: 10.21105/joss.01903
- McDermott, G. R., & Hansen, B. (2021). Labor reallocation and remote work during COVID-19: Real-time evidence from GitHub. *National Bureau of Economic Research WP29598*.

- Meng, J., Li, C., Martin, R. V., van Donkelaar, A., Hystad, P., & Brauer, M. (2019, May). Estimated Long-Term (1981–2016) Concentrations of Ambient Fine Particulate Matter across North America from Chemical Transport Modeling, Satellite Remote Sensing, and Ground-Based Measurements. *Environmental Science & Technology*, 53(9), 5071–5079. Retrieved 2023-01-17, from <https://pubs.acs.org/doi/10.1021/acs.est.8b06875> doi: 10.1021/acs.est.8b06875
- Meyer, H., Reudenbach, C., Wöllauer, S., & Nauss, T. (2019). Importance of spatial predictor variable selection in machine learning applications – moving from data reproduction to spatial prediction. *Ecological Modelling*, 411, 108815. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0304380019303230> doi: <https://doi.org/10.1016/j.ecolmodel.2019.108815>
- Monk, T. H., & Aplin, L. C. (1980). Spring and autumn Daylight Saving Time changes: Studies of adjustment in sleep timings, mood, and efficiency. *Ergonomics*, 23(2), 167–178.
- Nordhaus, W. D. (1997). Do real-output and real-wage measures capture reality? The history of lighting suggests not. *The Economics of New Goods*, 58, 29–66.
- Park, Y., Kwon, B., Heo, J., Hu, X., Liu, Y., & Moon, T. (2020, January). Estimating pm2.5 concentration of the conterminous united states via interpretable convolutional neural networks. *Environmental Pollution*, 256, 113395. Retrieved from <http://dx.doi.org/10.1016/j.envpol.2019.113395> doi: 10.1016/j.envpol.2019.113395
- Polonik, P., Ricke, K., Reese, S., & Burney, J. (2023, June). Air quality equity in US climate policy. *Proceedings of the National Academy of Sciences*, 120(26), e2217124120. Retrieved 2024-05-20, from <https://www.pnas.org/doi/full/10.1073/pnas.2217124120> (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.2217124120
- Potapov, P., Turubanova, S., Hansen, M. C., Tyukavina, A., Zalles, V., Khan, A., ... Cortez, J. (2021, December). Global maps of cropland extent and change show accelerated cropland expansion in the twenty-first century. *Nature Food*, 3(1), 19–28. Retrieved 2023-06-11, from <https://www.nature.com/articles/s43016-021-00429-z> doi: 10.1038/s43016-021-00429-z

- Power, M. C., Lynch, K. M., Bennett, E. E., Ying, Q., Park, E. S., Xu, X., ... Whitsel, E. A. (2024). A comparison of pm2.5 exposure estimates from different estimation methods and their associations with cognitive testing and brain mri outcomes. *Environmental Research*, 256, 119178. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0013935124010831> doi: <https://doi.org/10.1016/j.envres.2024.119178>
- Proctor, J., Carleton, T., & Sum, S. (2023). *Parameter recovery using remotely sensed variables*. Retrieved from <http://dx.doi.org/10.3386/w30861> doi: 10.3386/w30861
- Ratledge, N., Cadamuro, G., De La Cuesta, B., Stigler, M., & Burke, M. (2022, November). Using machine learning to assess the livelihood impact of electricity access. *Nature*, 611(7936), 491–495. Retrieved 2024-06-05, from <https://www.nature.com/articles/s41586-022-05322-8> doi: 10.1038/s41586-022-05322-8
- Reid, C. E., Considine, E. M., Maestas, M. M., & Li, G. (2021, April). Daily PM2.5 concentration estimates by county, ZIP code, and census tract in 11 western states 2008–2018. *Scientific Data*, 8, 112. Retrieved 2023-08-25, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8055869/> doi: 10.1038/s41597-021-00891-1
- Reid, C. E., Jerrett, M., Petersen, M. L., Pfister, G. G., Morefield, P. E., Tager, I. B., ... Balmes, J. R. (2015, March). Spatiotemporal Prediction of Fine Particulate Matter During the 2008 Northern California Wildfires Using Machine Learning. *Environmental Science & Technology*, 49(6), 3887–3896. Retrieved 2024-05-29, from <https://doi.org/10.1021/es505846r> (Publisher: American Chemical Society) doi: 10.1021/es505846r
- Requia, W. J., Di, Q., Silvern, R., Kelly, J. T., Koutrakis, P., Mickley, L. J., ... Schwartz, J. (2020, September). An ensemble learning approach for estimating high spatiotemporal resolution of ground-level ozone in the contiguous United States. *Environmental science & technology*, 54(18), 11037–11047. Retrieved 2023-10-03, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7498146/> doi: 10.1021/acs.est.0c01791
- Rhee, J., Dominici, F., Zanobetti, A., Schwartz, J., Wang, Y., Di, Q., ... Christiani, D. C. (2019, July). Impact of Long-Term Exposures to Ambient PM2.5 and Ozone on ARDS Risk for Older Adults in the United States. *Chest*, 156(1), 71–79. Retrieved 2024-11-14, from <https://linkinghub.elsevier.com/retrieve/pii/S0012369219307111> doi: 10.1016/j.chest.2019.03.017

- Robinson, C., Hohman, F., & Dilkina, B. (2017). A Deep Learning Approach for Population Estimation from Satellite Imagery. In (p. 47–54). New York, NY, USA: Association for Computing Machinery. Retrieved from <https://doi.org/10.1145/3149858.3149863> doi: 10.1145/3149858.3149863
- Sager, L., & Singer, G. (2024). Clean Identification? The Effects of the Clean Air Act on Air Pollution, Exposure Disparities and House Prices. *American Economic Journal: Economic Policy*.
- Schratz, P., Muenchow, J., Iturrutxa, E., Richter, J., & Brenning, A. (2019). Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling*, 406, 109-120. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0304380019302145> doi: <https://doi.org/10.1016/j.ecolmodel.2019.06.002>
- Schulte, N., Li, X., Ghosh, J. K., Fine, P. M., & Epstein, S. A. (2020, oct). Responsive high-resolution air quality index mapping using model, regulatory monitor, and sensor data in real-time. *Environmental Research Letters*, 15(10), 1040a7. Retrieved from <https://dx.doi.org/10.1088/1748-9326/abb62b> doi: 10.1088/1748-9326/abb62b
- Shi, L., Zhu, Q., Wang, Y., Hao, H., Zhang, H., Schwartz, J., ... Liu, P. (2023, January). Incident dementia and long-term exposure to constituents of fine particle air pollution: A national cohort study in the United States. *Proceedings of the National Academy of Sciences*, 120(1), e2211282119. Retrieved 2024-05-20, from <https://www.pnas.org/doi/full/10.1073/pnas.2211282119> (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.2211282119
- Smith, A. C. (2016). Spring forward at your own risk: Daylight Saving Time and fatal vehicle crashes. *American Economic Journal: Applied Economics*, 8(2), 65–91.
- Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2), 175–199.
- Swenson, J. J., Carter, C. E., Domec, J.-C., & Delgado, C. I. (2011, April). Gold Mining in the Peruvian Amazon: Global Prices, Deforestation, and Mercury Imports. *PLoS ONE*, 6(4), e18875. Retrieved from <https://dx.plos.org/10.1371/journal.pone.0018875> doi: 10.1371/journal.pone.0018875

- Toro, W., Tigre, R., & Sampaio, B. (2015). Daylight Saving Time and incidence of myocardial infarction: Evidence from a regression discontinuity design. *Economics Letters*, 136, 1–4.
- Ubah, J. I., Orakwe, L. C., Ogbu, K. N., Awu, J. I., Ahaneku, I. E., & Chukwuma, E. C. (2021, December). Forecasting water quality parameters using artificial neural network for irrigation purposes. *Scientific Reports*, 11(1), 24438. Retrieved 2023-06-11, from <https://www.nature.com/articles/s41598-021-04062-5> doi: 10.1038/s41598-021-04062-5
- van Donkelaar, A., Martin, R. V., Brauer, M., Hsu, N. C., Kahn, R. A., Levy, R. C., ... Winker, D. M. (2016, March). Global estimates of fine particulate matter using a combined geophysical-statistical method with information from satellites, models, and monitors. *Environmental Science & Technology*, 50(7), 3762–3772. Retrieved from <http://dx.doi.org/10.1021/acs.est.5b05833> doi: 10.1021/acs.est.5b05833
- van Donkelaar, A., Martin, R. V., Li, C., & Burnett, R. T. (2019, January). Regional estimates of chemical composition of fine particulate matter using a combined geoscience-statistical method with information from satellites, models, and monitors. *Environmental Science & Technology*, 53(5), 2595–2611. Retrieved from <http://dx.doi.org/10.1021/acs.est.8b06392> doi: 10.1021/acs.est.8b06392
- Van Donkelaar, A., Hammer, M. S., Bindle, L., Brauer, M., Brook, J. R., Garay, M. J., ... Martin, R. V. (2021, November). Monthly Global Estimates of Fine Particulate Matter and Their Uncertainty. *Environmental Science & Technology*, 55(22), 15287–15300. Retrieved 2023-09-08, from <https://pubs.acs.org/doi/10.1021/acs.est.1c05309> doi: 10.1021/acs.est.1c05309
- Wagner, D. T., Barnes, C. M., Lim, V. K. G., & Ferris, D. L. (2012). Lost sleep and cyberloafing: Evidence from the laboratory and a Daylight Saving Time quasi-experiment. *Journal of Applied psychology*, 97(5), 1068-76.
- Wang, Y., Marshall, J. D., & Apte, J. S. (2024). U.s. ambient air monitoring network has inadequate coverage under new pm2.5 standard. *Environmental Science & Technology Letters*, 11(11), 1220-1226. Retrieved from <https://doi.org/10.1021/acs.estlett.4c00605> doi: 10.1021/acs.estlett.4c00605

- Wei, J., Li, Z., Cribb, M., Huang, W., Xue, W., Sun, L., ... Song, Y. (2020, March). Improved 1 km resolution PM_{2.5} estimates across China using enhanced space–time extremely randomized trees. *Atmospheric Chemistry and Physics*, 20(6), 3273–3289. Retrieved 2024-05-29, from <https://acp.copernicus.org/articles/20/3273/2020/> (Publisher: Copernicus GmbH) doi: 10.5194/acp-20-3273-2020
- Wei, J., Li, Z., Lyapustin, A., Sun, L., Peng, Y., Xue, W., ... Cribb, M. (2021, January). Reconstructing 1-km-resolution high-quality PM_{2.5} data records from 2000 to 2018 in China: spatiotemporal variations and policy implications. *Remote Sensing of Environment*, 252, 112136. Retrieved 2024-05-29, from <https://linkinghub.elsevier.com/retrieve/pii/S0034425720305095> doi: 10.1016/j.rse.2020.112136
- Wu, X., Braun, D., Schwartz, J., Kioumourtzoglou, M. A., & Dominici, F. (2020). Evaluating the impact of long-term exposure to fine particulate matter on mortality among the elderly. *Science Advances*, 6(29), eaba5692. Retrieved from <https://www.science.org/doi/abs/10.1126/sciadv.aba5692> doi: 10.1126/sciadv.aba5692
- Xu, Z., Li, J., Lv, Z., Wang, Y., Fu, L., & Wang, X. (2021, July). A graph spatial-temporal model for predicting population density of key areas. *Computers & Electrical Engineering*, 93, 107235. Retrieved 2023-06-11, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8494484/> doi: 10.1016/j.compeleceng.2021.107235