Assessing the Effectiveness of Scrubber Installation on Air Pollution Emissions Reductions Among Coal-Fired Power Plants: Application of Statistical Methods for Causal Inference

Assessing the Effectiveness of Scrubber Installation on Air Pollution Emissions Reductions among Coal-Fired Power Plants: Application of Statistical Methods for Causal Inference

John Barrett Hansen

Presented to the Department of Applied Mathematics in partial fulfillment of the requirements for a Bachelor of Arts degree with Honors

Harvard College
Cambridge, Massachusetts
April 1, 2015

Acknowledgements

I would first like to thank Professoressa Francesca Dominici, whose mentorship and guidance have proved invaluable in all phases of this project. I would also like to thank Professor Cory Zigler for his methodological expertise that greatly improved the analyses' design, and Dr. Christine Choirat for her unbelievable assistance in data collection and analysis. I would also like to acknowledge the rest of the ARP working group team—Chanmin Kim, Georgia Papadogeorgou, and Evan Peet—for their continuous support and thoughtful feedback over the past year, and Sarah Moon for her peer review.

Table of Contents

1. Introduction 8

1.1. Background on the Acid Rain Program 8

Cap and Trade System 8

1.2. Literature Review and Gaps in Knowledge 9

Environmental Analyses 9

Cap and Trade Analyses 10

Health Impact Analyses 10

Cost Analyses 11

1.3. Scientific Questions 12

2. Data 13

2.1. Overview of Data Sources 13

Air Markets Program Data (AMPD) from EPA Acid Rain Program (ARP) 13

US Energy Information Administration (EIA) 14

2.2. Exploratory Analysis 15

2.3. Definition of Treatment 17

2.4. Definition of Study Populations 18

2.5. Definition of Outcome 20

2.6. Data Challenges 21

Missing Data 21

Outlying Data 21

Competing Forms of Treatment 22

Time-Varying Treatment 23

Nature of an Observational Study 24

2.7. Measured Confounders 24

LogHeat 25

Operating Time 26

Sulfur Content 26

Log Emissions 26

SO_2, NO_x, and $PM_{2.5}$ Scrubbers 26

SO_2 and NO_x Phase 26

Secondary Fuel 27
Regulation Level 27
Geographic Regions 28
Has Start Date 29
2.8. Experimental Datasets 29
Questions 1 & 2 29
Questions 3 & 4 30
3. Methods 31
3.1. Potential Outcomes Framework 31
3.2. Propensity Score Model 32
Theoretical Background 32
Confounders Used in Analysis 33
Subclassify & Assess Balance 35
3.3. Variable Ratio Matching by Covariates 37
Theoretical Background 37
Covariates Used in Analysis 38
3.4. Analysis of Outcome 39
Average Treatment Effect on the Treated (ATT) 39
Two Months Before and After 39
Difference in Means & Regression Adjustment 40
4. Results 41
4.1. Questions 1 & 2 42
4.2. Questions 3 & 4 43
4.3. Sensitivity Analyses 45
Variation of the Propensity Score Model 45
Variation of the Variable Ratio Matching Model 45
5. Discussion 46
Explanation of Results 46
Conclusion 47
6. Appendix 49
Appendix A: Acid Rain Program 49
Appendix B: Unit-level Information 50
Appendix C: Steps in Methodology 51
7. References 58

Abstract

The 1990 amendment to the Clean Air Act implemented a cap-and-trade system that required electricity-generating power plants to dramatically reduce Sulfur Dioxide (SO_2) and Nitrogen Oxide (NO_x) emissions. Plants impacted by this legislation had a variety of compliance options, including decreasing factory operation, purchasing carbon credits, installing scrubbers, and changing fuel inputs. Using data from 1997-2012 of 995 coal-burning power plants, we examine the effectiveness of scrubber installation in reducing SO_2 and NO_x emissions. Specifically, we employ two methods—a propensity score algorithm and a matching algorithm—to estimate: 1) the causal effect of scrubber installation prior 1997 on the emissions during 1997; and 2) the causal effect of scrubber installation at any time during the period 1997-2012 on emissions two months following scrubber installation. Using a propensity score method, we found that pre-1997 SO_2 scrubbers reduced 1997 SO_2 emissions by 68% (95% CI 58% to 76%), and pre-1997 NO_x scrubbers reduced 1997 NO_x emissions by 28% (16%, 38%). Additionally, installing SO_2 and NO_x scrubbers at any time during the period 1997-2012 reduces SO_2 and NO_x emissions by 89% (88%, 90%) and 21% (19%, 24%) two months following installation, respectively. These final two results are corroborated by a matching algorithm, which finds scrubbers cause SO_2 and NO_x emissions decline by 88% (87%, 89%) and by 20%. (17%, 22%) two months following installation, respectively.

1. Introduction

1.1. Background on the Acid Rain Program

In 1990, then President George H.W. Bush signed into law amendments to the Clean Air Act, which included a cap-and-trade system intended to curb the threat of acid rain. Acid rain is formed when harmful chemicals like sulfur dioxide (SO_2) and nitrous oxide (NO_x) mix with moisture in the atmosphere to form sulfuric and nitric acids, and return to the earth's surface through precipitation. Emissions from coal-fired power plants were, and continue to be, the primary source of SO_2 emissions and a leading source of NO_x emissions in the United States.

Title IV of the Clean Air Act established the Acid Rain Program (ARP), a requirement for major emission reductions of both SO_2 and NO_x for American power plants. The goal of this program was to reduce total SO_2 emissions by ten million tons relative to 1980 levels (29.5 million tons per year). This drop was to be achieved mostly through cutting emissions from electricity-generating units (EGUs), a process enacted in two stages. Phase I (1995-1999) required 263 extremely polluting units to significantly reduce their emissions. Phase II, which began in 2000, placed a target SO_2 emissions cap at 8.95 billion tons per year on about 3,200 EGUs, which cut power-sector emissions nearly in half from 1980 levels.

Cap and Trade System

Rather than mandating specific targets for each factory, the ARP created nation-wide emissions caps under which the EGUs collectively had to remain. The government accomplished this by distributing emissions allowances, which together summed to the emissions cap. These allowances could be traded, so that factories generating more tons of emissions than they had in allowances could buy additional ones on the open market (and factories with extra allowances could sell them). The price of the allowances was subject to the market. Extra allowances could also be saved for the following year. Factories with more emissions than allowances at the end of the year were subject to severe penalties.

The fundamental idea behind this program is that the generation of electricity creates an externality; SO_2 and NO_x, byproducts of burning coal, are a societal cost that the factories were not offsetting. Therefore the ARP was a form of Pigouvian taxation, forcing the EGUs to internalize the externality they had created.

This implementation of a national cap was a marked change from previous environmental regulation, which typically mandated individual cuts to emissions (Schmalensee & Stavins, 2012). Intuitively, it is likely to be more efficient because it allows factories to choose the best option for them. In the case of SO_2 and NO_x reduction, a variety of options are available; factories can install a scrubber, change the composition of its inputs (purchase higher quality coal, switch to natural gas, or add secondary fuels), reduce operating time, or purchase allowances. All of these options are designed to be significantly less costly than the penalty for exceeding one's allowances. We expect factories to choose the least costly path; EGUs that can easily reduce emissions will likely do so, and those that cannot will purchase allowances, all while the total amount of harmful chemicals will be reduced.

1.2. Literature Review and Gaps in Knowledge

The ARP has been covered extensively in the literature. It is generally lauded as a success story, as the marked national decreases in SO_2 and NO_x over the past two decades came relatively inexpensively. Though aggregate figures describing emissions declines are well known, a detailed analysis of scrubber effectiveness has never been conducted. Here we outline the literature on the ARP broadly partitioned into four categories—environmental, cap and trade, health, and cost analyses—and identify the gaps in knowledge that this paper fills.

Environmental Analyses

The ARP was implemented primarily with the goal of stopping the acidification of aquatic ecosystems. It was thought that to achieve this it was necessary to reduce SO_2 and NO_x emissions, and by this measure the ARP was a remarkable success. Despite a 25% increase in electricity production over the first 14 years of the program, SO2 emissions fell by 36% (U.S. Environmental Protection Agency 2011b). The program met its long-term goal of reducing EGU annual SO_2 emissions to 8.95 tons by 2007, and in 2010 annual emissions declined even further

to 5.1 tons (Schmalensee, and Stavins, 2012). Appendices A.1 and A.2 provide plots of the decline in total annual emissions.

Whether or not these emissions reductions translated to improved ecosystem health in a cost-effective way remains up for debate. Banzhaf et al (2006) estimate the ecosystem benefits outstripped program costs, but Schmalensee & Stavins estimate the ecosystem benefits to be as low as 25% the costs of the program. Geographic-specific analyses of the program's impact remain unclear, mostly due to the dramatic decline in overall emissions (Burtraw et al., 1998). None of these papers assess the viability of any particular intervention.

De Gouw et al (2014) compare emissions of coal-fired to natural gas powered EGUs in the United States and estimate that the increased use of natural gas has led to a 44% reduction in SO_2 emissions and a 40% reduction NO_x emissions over the past several decades. Our approach differs from this paper's in that it measures the impact of scrubber installations rather than a change in fuel inputs. In addition, it estimates the causal effect from within-unit changes in emissions from an intervention rather than comparing population means.

Cap and Trade Analyses

Chan, Stavins, Stowe and Sweeney (2012) assess the pros and cons of the cap and trade system as a tool to curb emissions. They focus mostly on the trends in national emissions, concluding that the program as a whole was a success, and that the cap and trade system worked toward the goal of reducing national emission levels. It also points out that the market-based approach incentivized technological innovation, since EGUs with a cost-effective way of reducing emissions could bank and sell allowances for a profit. Burtraw and Palmer (2004) also assessed the ARP's market-based approach, and concluded that the policy lends itself well to reproduction in future instances of externality correction. Neither paper delves into the specific tools with which emissions are reduced, focusing on the efficacy of the policy in its entirety.

Health Impact Analyses

Though the ARP was enacted as an environmental measure, its legacy will be its dramatic health benefits. Estimates of the annualized human health benefits range from $50B to $100B (Burtraw et al., 1998; Burtraw, 1999; Chestnut and Mills, 2005; Banzhaf et al., 2006, Schlamansee &

Stavins, 2012). There is of course considerable uncertainty when estimating health benefits, specifically to do with valuing mortality and morbidity, but it is clear the health benefits are on the order of 100 times environmental benefits. As with the environmental and cap and trade analyses, these papers utilize national-level emissions data in making estimates, and do not discuss any particular method of emissions reduction.

Buonocore et al. estimated health impacts of air quality interventions using Community Multiscale Air Quality (CMAQ) simulations, in which various EGUs in the Midwest and Great Lakes regions had emissions set to zero. Similarly, Levy et al. (2009) build a source-receptor model that outlines how pollution is dispersed across space to estimate the health impacts of secondary $PM_{2.5}$, a harmful airborne pollutant formed by SO_2 and NO_x. Neither of these papers addresses the expected decline in emissions from air quality interventions, which would result in more accurate estimates of the health benefits.

Cost Analyses

Most of the literature on costs related to the ARP focus on the cost of compliance for the EGUs. Burtraw (1998) finds that the annualized costs of the program were on the order of $0.5B; Schmalensee & Stavins (2012) estimates the annual cost to be between $0.5B and $2B. These costs are dramatically lower than the program's benefits, which Schmalensee & Stavins (2012) estimate to be between $59-$116B (drawing from Chestnut & Mills (2005), Burtraw et al. (1998) and Burtraw (1999)), and far below $6.1B cost estimated by the EPA in 1990.

The effectiveness of the program is most effectively judged in comparison to alternative regulations like command and control. Chan, Stavins, Stowe & Sweeney (2012) aggregate a series of studies, and find that the cap and trade program was plainly less expensive than alternatives; however, their reported range of 15-90% less expensive leaves the exact degree to which it was cheaper ambiguous (via Carlson, et al., 2000; Ellerman, et al., 2000, 253–295; Keohane, 2003).

Costs of particular elements of the program are also briefly discussed. Schmalensee, and Stavins (2012) point to the $2,000/ton fine for exceeding allowances as a primary reason that compliance was nearly 100%. They argue unknown marginal abatement costs, particularly at the beginning, may have increased total compliance costs as EGUs faced uncertainty regarding compliance

options. Finally, they discussed the impact of the Railroad Revitalization and Regulatory Reform Act of 1976 and the Staggers Rail Act of 1980, which opened up the railroad industry; they argue falling freight costs dramatically lowered the cost of compliance by permitting low sulfur coal from the Powder River Basin in Wyoming and Montana to be transported inexpensively to the many factories east of the Mississippi River.

1.3. Scientific Questions

We constructed a new data set of an unprecedented level of accuracy that provides detailed information on strategies implemented on EGUs to cut emissions. We then conducted statistical analyses for causal inference to ultimately provide new evidence regarding the effectiveness of scrubber installation on emissions.

First, for the units that installed a scrubber at any time before 1997, we will address the following two questions:

1. Does SO_2 scrubber installation at any time before January 1 1997 impact SO_2 emissions during the period 1997-2012?
2. Does NO_x scrubber installation at any time before January 1 1997 impact NO_x emissions during the period 1997-2012?

Second, for the units that installed a scrubber at any time during the period 1997 - 2012, we address the following two questions:

1) Are SO_2 emissions two months following SO_2 scrubber installation lower than SO_2 emissions two months before scrubber installation, compared to units that never install a scrubber?
2) Are NO_x emissions two months following NO_x scrubber installation lower than NO_x emissions two months before scrubber installation, compared to units never install a scrubber?

2. Data

2.1. Overview of Data Sources

In this section we will provide an overview of several sources of data and of the final data set used for the analyses.

Air Markets Program Data (AMPD) from EPA Acid Rain Program (ARP)

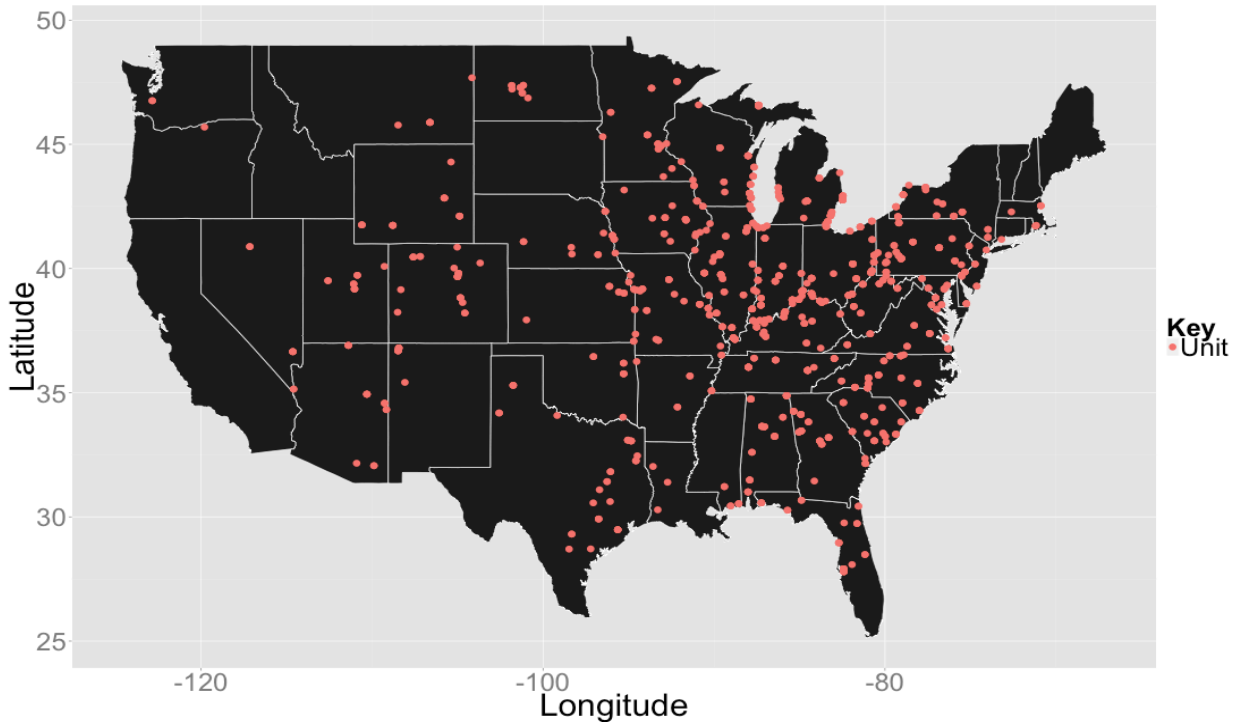
For the period 1995-2012, open-access daily data (found at http://ampd.epa.gov/ampd/) are available at the power plant and unit levels, where a power plant is defined as a facility with one or more generating units. We have data on 4,164 units belonging to 1,248 power plants, the totality of facilities that participated in the ARP. These facilities represent approximately 20% of all US power plants, but account for 75% of fuel combustion emissions and 65% of overall emissions.

For this analysis, we specifically collected: information about each unit (e.g., state, county, latitude, longitude); ARP phase (I, II, opt-in, substitution, compensating); SO_2, NO_x, and CO_2 emissions; heat input, gross load, steam load, operating time, and status; primary and secondary fuel types (e.g., coal, diesel oil, natural gas); and scrubber technologies (whether a scrubber is installed and, if so, the technology it uses) for SO_2, NO_x and particulate matter (PM).

We aggregate daily levels of emissions to monthly values. Among the 4,164 units, 1,174 use coal as their primary input and they will be the focus of our analysis. Figure 1 shows a map of the units used in this paper.

To illustrate the high level of granularity of our data, in Figure 2 we plotted two AMDP data points by their latitude and longitude on Google Maps: a coal-burning facility in Florida with four units (left) and a natural gas facility in New York with four units (right). The smokestacks are clearly visible in both instances.

Figure 1: Map of Unit Locations



The location of all units considered in this analysis. Notice the much higher density of units in the old industrial center of the nation, colloquially known as the "Rustbelt" area.

Figure 2

Two ARP-monitored power plants: coal (left, Apollo Beach, FL) and natural gas (right, Long Island City, NY).

US Energy Information Administration (EIA)

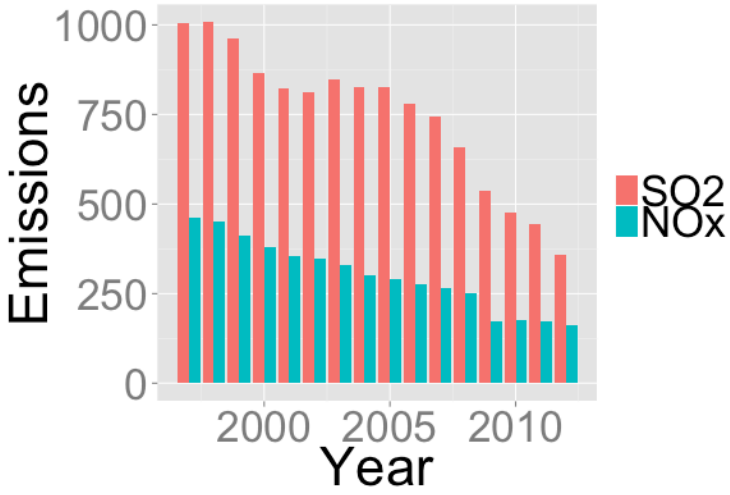
We use monthly and yearly data for the period 1997-2012 at the unit level. This analysis uses only a few attributes from this data source, including the sulfur content of the coal, and the stringency of government-imposed regulations for each plant. This data is also available online, at http://www.eia.gov/.

The final dataset used in this study was created through merging both the EPA and EIA datasets. The bulk of the information used in this analysis comes from the EPA dataset, including the intervention (scrubber) and outcome (emissions) data.

2.2. Exploratory Analysis

Figure 3 shows annual average SO_2 and NO_x emissions from 995 coal-fired power plants during the period 1997 to 2012. Please note that both SO_2 and NO_x have been declining over time, likely due to increasingly stringent air quality regulations.

Figure 3: Mean Yearly Emissions for Coal-Fired Units

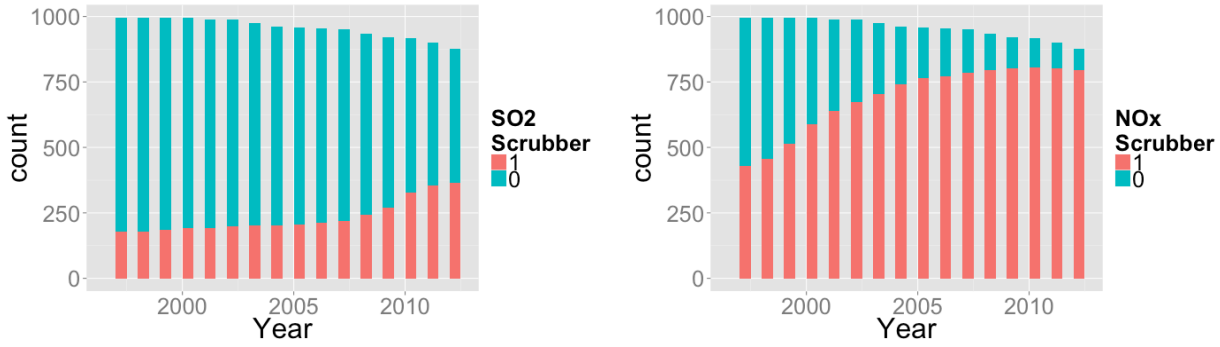


Mean yearly emissions for coal-fired units. Both pollutants undergo steady declines over the time period.

As described in the introduction, this overall decline is closely examined in the literature. The aim of this paper is to contribute to this body of knowledge by isolating the effect of scrubbers, of both the SO_2 and NO_x variety, on this decline. To start, we can examine the prevalence of these technologies in units, as shown in Figure 4.

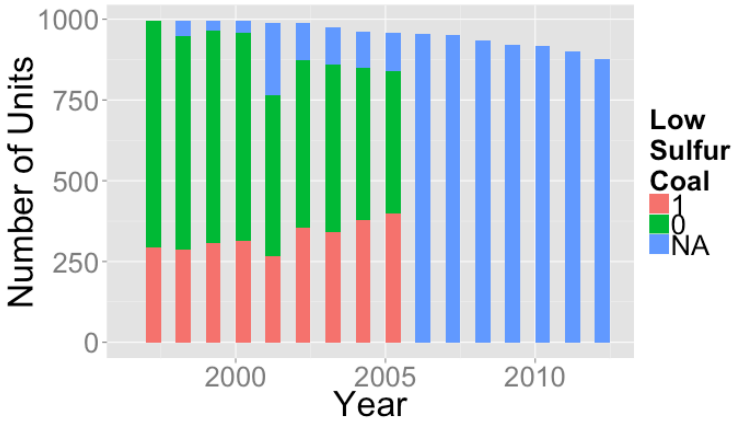
Both SO_2 and NO_x scrubbers see an increase in use over time, but NO_x scrubbers are more prevalent throughout the period. Our dataset also allows us also to observe some of the other methods by which units can lower emissions, such as purchasing lower sulfur coal and reducing operating time. While a formal analysis of these alternative emission-reduction techniques will not be developed here, we can still get a sense for them with figures 5 and 6.

Figure 4: Number of Scrubbers Over Time



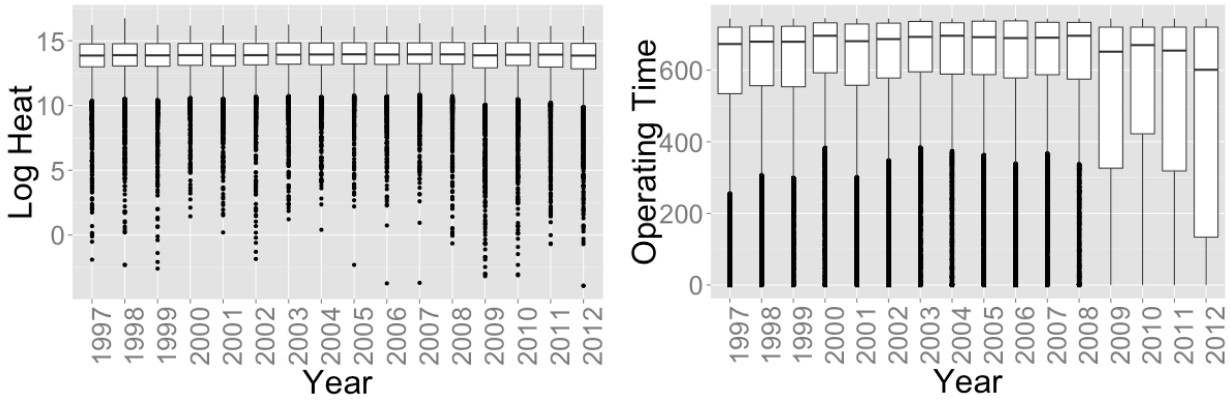
Number of SO_2 scrubbers (left) and NO_x scrubbers (right) as a function of time.

Figure 5: Prevalence of Low Sulfur Coal over Time



Number of units burning low sulfur coal (defined as having below 0.6 lbs/MMBtu) as a function of time. Blue indicates a missing value; notice how the dataset only has entries from 2005 and earlier.

Figure 6: Boxplots of Log Heat and Operating Time by Unit over Time



The average log of heat input (left) and the average operating time in hours (right) per unit as a function of time.

In Figure 5 we see that more units adopt low sulfur coal over time, although our ability to measure this increase abruptly stops in 2006, when the EIA stops recording this attribute. The threshold for low-sulfur coal is 0.6 pounds of sulfur per million British Thermal Units (BTU) (EIA, 2015). This increase in the use of low sulfur coal was largely driven by falling transportation costs caused by the *Railroad Revitalization* and *Staggers Rail* Acts as was outlined in the Literature Review.

Figure 6 shows the distributions of the log of heat input and of Operating Time for units over the 1997-2012 period. There are two attributes of these figures worth noting. First, there are many outliers on the lower end of the distribution. This is a recurring theme in the dataset, and will be addressed more fully in Section 2.5. Second, both quantities are constant over the period, indicating units generally do not change either the operating time or heat input in the face of more stringent regulations.

2.3. Definition of Treatment

For Question 1, the indicator of treatment Z is defined as:

- $Z_i = 1$ if unit *i* has an SO_2 scrubber installed at any time before January 1, 1997
- $Z_i = 0$ if unit *i* does not have a scrubber installed by January 1, 1997

Similarly, for Question 2, it is:

- $Z_i = 1$ if unit *i* has a NO_x scrubber installed at any time before January 1, 1997
- $Z_i = 0$ if unit *i* does not have a scrubber installed by January 1, 1997

Technically, since our data spans only 1997-2012 (with 1995-1996 data for a select few units), we only know with certainty whether or not a unit had a scrubber in January 1997. We would not know, for instance, if a unit installed a scrubber in 1996 and removed if before January 1997. However, given the extraordinarily low rates of scrubber un-installation post-1997, we can infer reasonably safely that all units without scrubbers in January 1997 did not ever have one previously.

For Question 3, the indicator of treatment Z is defined as:

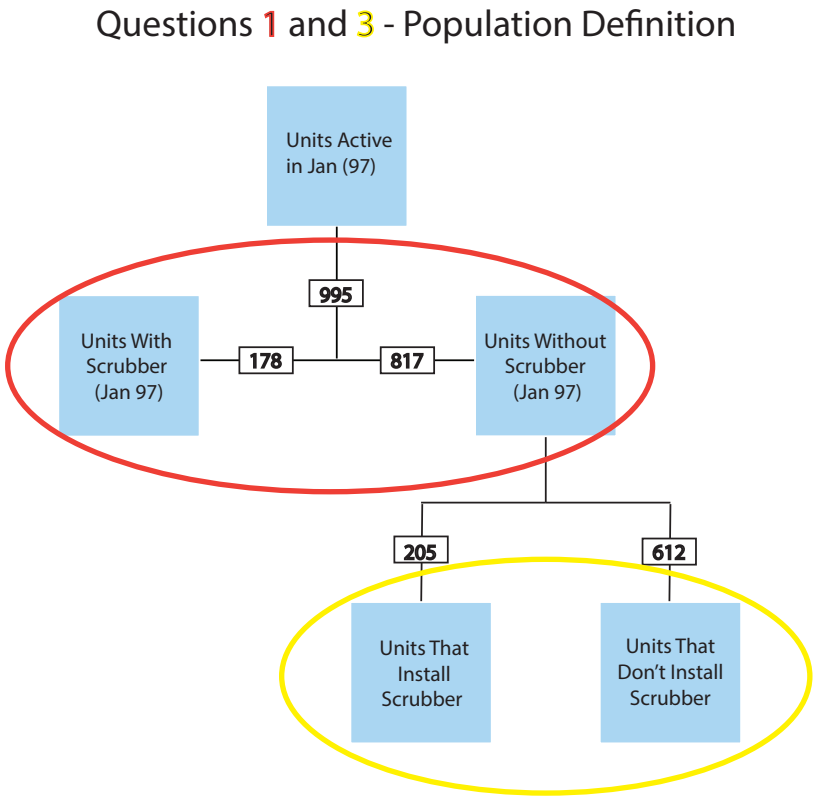
- $Z_{it} = 1$ if unit *i* installed an SO_2 scrubber in month *t* at some point between 1997 and 2012
- $Z_{it} = 0$ if unit *i* does not have an SO_2 scrubber at any point between 1997 and 2012

Finally, for Question 4, we define treatment as:

- $Z_{it} = 1$ if unit *i* installed a NO_x scrubber in month *t* at some point between 1997 and 2012
- $Z_{it} = 0$ if unit *i* does not have a NO_x scrubber at any point between 1997 and 2012

2.4. Definition of Study Populations

Diagram 1: Definition of Study Populations for Questions 1 & 3



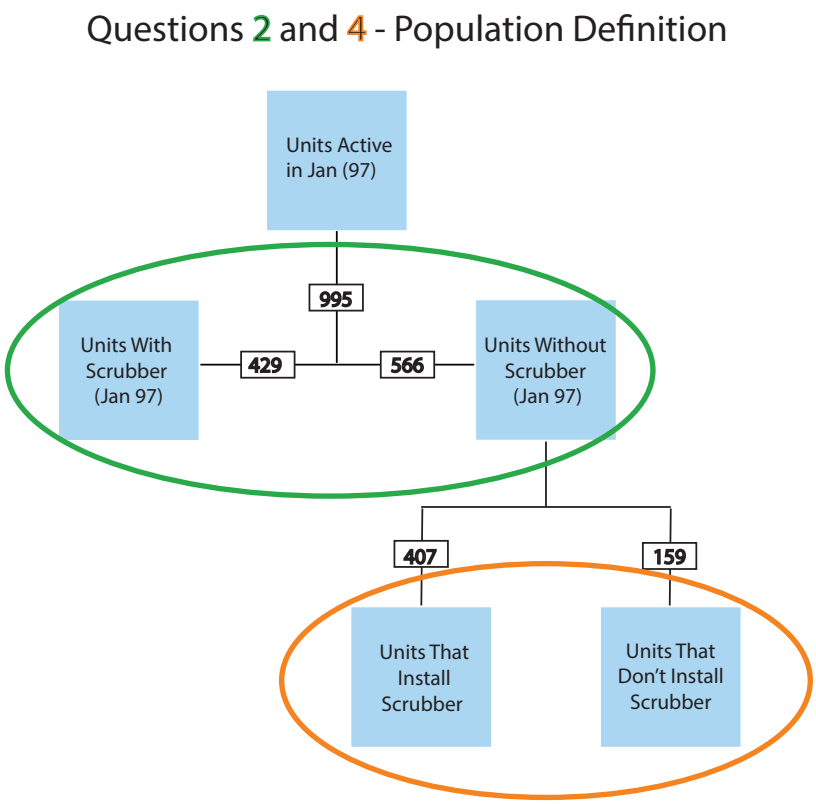
Description of the study populations for Questions 1 & 3. The red circle surrounds the study population for Question 1, in which the treated units are those entering the dataset with a scrubber. Question 3's study population—the group inside the yellow circle—is the same as Question 1's control units. 205 of these units install a scrubber at some point during the period 1997 to 2012 (Q3 treatment), and 612 do not (Q3 control).

There is overlap between the study populations used in each of the analyses. A detailed description of the four study populations is described in Diagrams 1 & 2. Once again, a *unit* is a smokestack within a power plant; each plant may have multiple units.

Units in Questions 1 and 2 are considered *treated* if the unit has an SO_2 or NO_x scrubber in January 1997, respectively, and *control* if no scrubber is present. These two groups are then

compared on their emissions over the course of 1997. We note that there are 12 units that install a NO_x scrubber, and 1 that installs an SO_2 scrubber in 1997; these units are excluded from the control group upon installation of a scrubber so as not to introduce bias.

Diagram 2: Definition of Study Populations for Questions 2 & 4



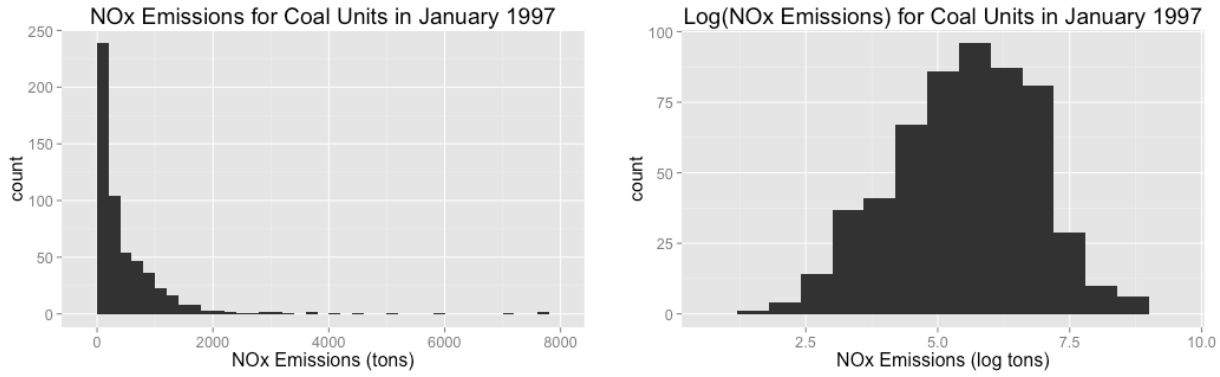
Description of the study populations for Questions 2 & 4. The green circle surrounds the study population for Question 2, in which the treated units are those entering the dataset with a scrubber. Question 4's study population—the group inside the orange circle—is equal to Question 2's control units. 407 of these units install a scrubber at some point during the period 1997 to 2012 (Q4 treatment), and 159 do not (Q4 control).

The control units in Questions 1 then form the study population for Question 3 (and the control units from Question 2 become the study population for Question 4). In each case, the units that eventually install a scrubber, at any point between 1997 and 2012, are considered *treated* for these questions, and *control* if no scrubber is ever installed.

2.5. Definition of Outcome

In determining the effect of these four interventions, we estimate the causal effect of installing a scrubber on log emissions. We use the logarithm because both types of emissions follow a lognormal distribution in the data. Figure 7 shows the distributions for NO_x emissions and its logarithm in one month, January 1997. SO2 emissions exhibit similar behavior.

Figure 7: Lognormal Distribution of NO_x Emissions



Left: Distribution of NO_x Emissions for Coal Units in January 1997. Right: The distribution of the log transformation of NO_x emissions in January 1997.

Emissions are measured in tons, while heat input is in MMBtu. An MMBtu—one million British Thermal Units (BTU)—is approximately 1.055 billion joules. The log transformations mean that we measure emissions in log(tons) and heat in log(MMBtu).

For Questions 1 & 2, we compare the average log emissions during the period 1997 between the treatment and control groups. Each unit's outcome therefore is the average of their log monthly emissions over the 12 month period in 1997. More formally:

$$y_i = \frac{1}{12}\sum_t Y_{i,t}$$

where $Y_{i,t}$ is the monthly emissions for unit i in month t, and t can only take on the 12 monthly values of the year 1997.

For questions 3 & 4, the outcome of interest is the change in log emissions from 2 months before scrubber installation to 2 months after scrubber installation. Please note that the date of scrubber installation may vary across units. The motivation to measure two months before to two months after installation will be explained more thoroughly in the methods section. Formally speaking, the outcome is:

$$y_i = Y_{i,t+2} - Y_{i,t-2}$$

where $Y_{i,t}$ again measures the monthly emissions for unit *i* in month *t*, but this time the year can be any in the range 1997 to 2012.

2.6. Data Challenges

The statistical analyses will need to overcome several challenges in the data, to the extent possible. In this section we outline each in detail.

Missing Data

As alluded to previously, our final dataset, though quite comprehensive, does have a considerable number of missing (omitted) values. For example, sulfur content data is only available through 2006, and approximately 5-10% of each SO_2 emissions, NO_x emissions and heat input are missing in the data. Fortunately, missing values among the heat input and two emission attributes were highly correlated (especially NO_x and SO_2 emissions), reducing the number of units needing to be removed. These omissions appear random with respect to the remaining covariates.

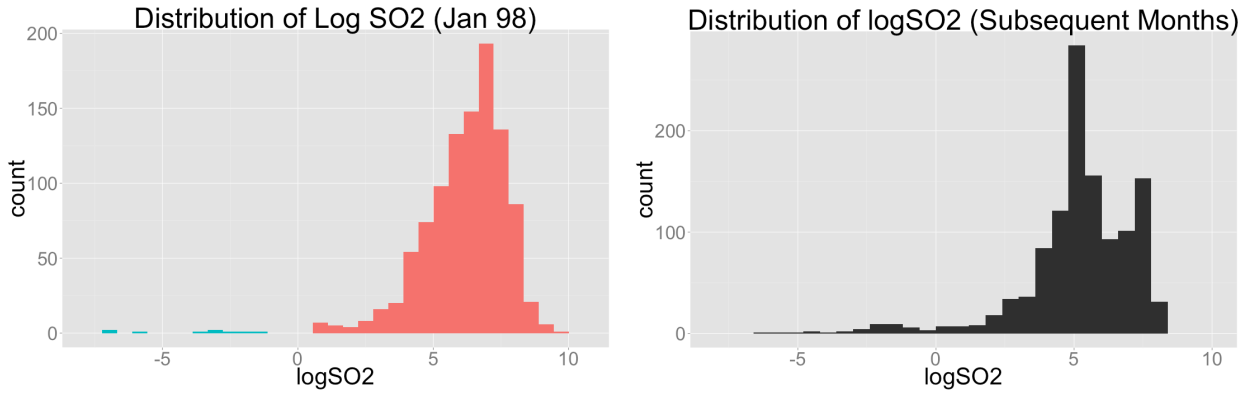
The initial year of operation, denoting the first year a unit was opened, was reported more often in the EIA data for larger, dirtier factories and thus it is not missing at random. That meant we had to exclude this particular covariate.

Outlying Data

Several key covariates had significant left skew in our data. For example, Figure 8 plots the distribution of log SO2 for coal-fired units in January 1998. Though the mean log SO2 value lies just above 6.1, there exist several values below zero. Converting back to tons of emissions, that

means that while the average unit emitted around 450 tons of SO_2 per month, some were recorded with as little as six one-thousands of a ton, or a mere 13 pounds.

Figure 8: Outlying Values



Left: The distribution of logSO2 for coal-fired power plants in January 1998. Right: Distribution of the left panel's teal (below zero) units in subsequent months (1999-2012).

These extraordinarily low values were likely abnormalities. This quite clear by taking the nine units whose January 1998 $logSO_2$ value lies below zero and plotting $logSO_2$ emissions for subsequent months, as done in right panel of Figure 8. While there exists a slightly higher density of points below zero than in the first graph, the shapes of the distributions are largely the same. This means that while units reporting extremely low values in one month are only slightly more likely than average to do so in again in the future. A likely explanation for this phenomenon is faulty monitors. We only worry about outlying values for logHeat, $logCO_2$, $logNO_x$ and $logSO_2$, as they are the only continuous covariates with long tails.

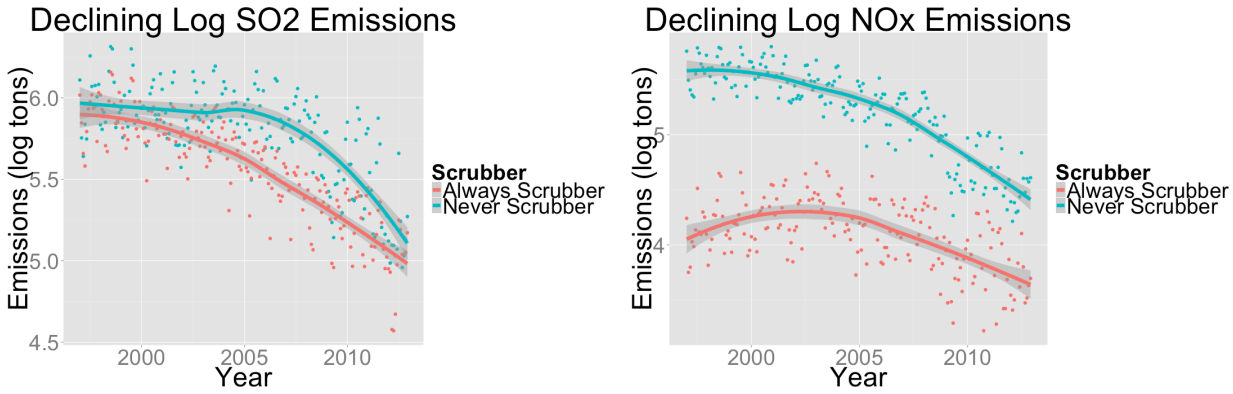
Given the randomness of these outlying values, we opted to simply remove them from our analysis. We used the inner-quartile range (IQR) method to determine outliers, where the IQR is defined as the difference between the 75^{th} percentile and 25^{th} percentile. Any value less than the 25^{th} percentile minus 1.5 times the IQR or greater than the 75^{th} percentile plus 1.5 times the IQR was deemed an outlier and removed from the analysis.

Competing Forms of Treatment

As outlined in the introduction, units needing to cut emissions could choose between several options in addition to installing scrubbers. These alternative options are: 1) purchasing lower

sulfur coal; 2) reducing operating time; and 3) purchasing carbon credits. In theory, units can employ any combination of these strategies. Figure 5 shows the log of emissions plotted as a function of time from 1997 to 2012 for the units that had a scrubber for each month in the data ("Always Scrubbers") and those that never did ("Never Scrubbers"). This division is not one we will use at any future point in the analysis; however, it is of interest because any change in emissions from these units will necessarily have nothing to do with scrubber installation, and therefore will give us a sense of how the other forms of treatment impact emissions.

Figure 9: Declining Emissions over Time for both Pollutants



Left: Average LogSO_2 emissions for Always Scrubber and Never Scrubber units from 1997-2012. Right: logNOx emissions for Always Scrubber and Never Scrubber units over the same time period.

LOWESS (locally weighted scatter-plot smoothing) lines fitted to the points all show a decline between 1997 and 2012. Further complicating the problem, the declines are dissimilar for units that have a scrubber for the entire period and those without. "Always SO_2 Scrubber" units have similar log emissions to "Never SO_2 Scrubber" units at both the start and end of our time period, but took very different paths in between. In contrast, the "Always NO_x Scrubber" units experienced half a decade of rising emissions before slowly declining, while the "Never NO_x Scrubber" units experience a steady decline for the whole period. The general downward trend, especially among the units that never installed a scrubber, clearly indicates that these units employed other emission-reducing techniques during this period beyond scrubber installation.

Time-Varying Treatment

The most obvious challenge to the classical experimental setup is that units do not all receive treatment at the same time. This occurs for two primary reasons. First, the ARP did not mandate

the same timetable for all factories; as described in the introduction, there were two phases in which units could be placed. Even within those phases, there is no strict requirement any particular technology is used, nor a specific timeline on which it should be implemented. Second, the staggered nature of the emissions requirements meant that units with relatively low emissions did not need to invest in emissions reduction equipment right away. Finally, for questions 1 & 2, we face the additional challenge that we do not know the precise time of installation: we just know that it occurred prior 1997.

Nature of an Observational Study

As with any observational study, this analysis challenges us to isolate the treatment effect from a series of confounders. We would have liked to have gathered the owners of each factory and tossed coins to determine who installs a scrubber in order to ensure a randomized experiment. For a multitude of reasons, this was clearly impossible. Each owner made a choice of whether or not to install a scrubber, the reasoning behind which is not clearly labeled in the data. That said, our dataset comes with a series of other attributes that we use to prune and otherwise arrange the data so as to isolate the treatment effect from whatever confounders might also drive the decision to install a scrubber.

2.7. Measured Confounders

We used a number of covariates in our dataset. We will first address the continuous variables utilized, described more thoroughly in Table 1.

Table 1: Measured Continuous Confounders

Confounder	Description	Units
logHeat	The logarithm of the amount of heat generated by the plant.	Log(MMBtu)
logSO2	The logarithm of a unit's SO2 emissions.	Log(Tons)
logNOx	The logarithm of a unit's NOx emissions.	Log(Tons)
logCO2	The logarithm of a unit's CO2 emissions.	Log(Tons)
Sulfur Content	The pounds of sulfur in the coal per million BTU of potential energy	Lbs/MMbtu
Operating.Time	The amount of time the plant is operating.	Hours

We used also used a number of categorical variables in this analysis. To incorporate this information into our models, we split them into a series of dummies. Table 2 provides a summary of the categorical variables, along with a description of their meaning and a list of the values the variables can take.

Table 2: Measured Discrete Confounders

Confounder	Description	Levels
Scrubbers	What types of scrubbers the unit has in operation	SO_2 Scrubber, NO_x Scrubber, $PM_{2.5}$ Scrubber
SO_2 Phase	The SO_2 emissions requirements as laid out in Title IV of the Clean Air Act.	Phase 1, Phase 2, Substitution
NO_x Phase	The NO_x emissions requirements as laid out in Title IV of the Clean Air Act.	Early Election, Phase 1 Group 1, Unknown
Secondary Fuel	The second input to electricity production (if any).	Diesel, Natural Gas, Other, None
Regulation	The government level at which the most stringent regulations are imposed.	Federal, State, Local, Unknown
Region	The geographic region in which the unit is located.	West, Midwest, South, Rustbelt, Atlantic, Northeast
Has IYMO	Whether or not the value for initial year & month of operation (iymo) is missing.	1, 0

Here we provide a more comprehensive description of the covariates in the tables.

LogHeat

The dataset has three covariates—Heat Input, Gross Load, and Steam Load—that all capture the unit's energy volume. These three covariates are highly correlated. Because Heat Input has the least amount of missing data points, we use this covariate as a comprehensive measure for unit electricity production.

This covariate is one of the most important in the data because it is our closest proxy to factory size; emissions can change dramatically with scrubbers or other interventions, but the factory will continue creating electricity. As such, we are very careful to ensure balance for this covariate is particularly strong.

Operating Time

Operating time is less valuable than log heat because there is less variation; the majority of plants operate as long as they can, with a few units each month that for whatever reason (repairs, local restrictions) operated less frequently. The distributions of both this confounder and LogHeat are visible in Figure 6 in the Exploratory Analysis Section. Please note that a quadratic term for Operating Time is also present in several of the analyses.

Sulfur Content

The data also reports the coal's sulfur content. This is an important covariate because the composition of the coal has a large impact on emissions. Changing the coal input is another form of treatment entirely, and needed to be controlled for. Because of a long right tail, sometimes the logarithm of this quantity is utilized.

Log Emissions

The logs of SO_2 and NO_x serve time both as a dependent variable and as a covariate, depending on whether an SO_2 or NO_x intervention is being tested. We include both—and logCO_2—because each provides more information on factory size, the regulatory environment and the factory's response.

SO_2, NO_x, and $PM_{2.5}$ Scrubbers

Each of Questions 1-4 identifies the installation of either an SO_2 or NO_x scrubber as the intervention of interest. However, information regarding whether the unit has installed any of the other types of scrubbers available intuitively is valuable in determining its propensity for treatment.

SO_2 and NO_x Phase

SO_2 and NO_x phases refer to the way in which the unit enters the ARP. The options for SO_2 Phase are Phase I, Phase II, Substitution and Compensating. As mentioned in the introduction, Phase I units were the larger emitters entered into the program in 1995. Phase II began in 1999, and implemented a total emissions cap that affected most power plants in the nation. For sample size reasons we do not include compensating plants in the analysis. We have complete data for

the SO_2 phases, but unfortunately must categorize some NO_x phase values as *unknown*. The distributions for both are available in Appendix B.1.

Secondary Fuel

Units can use multiple fuels in electricity generation, and in the data units have a primary and a secondary fuel listed. This measure is very imprecise however, because proportions are not provided; this means a unit using 51% coal and 49% diesel will appear identical to one using 99% coal and 1% diesel. Units may use no secondary fuel, and can also have multiple. There are a variety of names for these fuels that we broadly group into diesel, natural gas, other, and none. While some units in the data used a different form of coal as a secondary fuel, none of the units in our experimental datasets did.

We create indicators for each of the four categories. The simplest to interpret is *none*: if the secondary fuel field is empty, the indicator for none is one. Since units can have multiple fuels, the indicator for natural gas is one if one of the secondary fuels is a type of natural gas types. The same process is followed for diesel. This means that these two indicators are not mutually exclusive, since both natural gas and diesel could be present. Finally, this means that we should interpret the "other" value—which encompasses residual oil, wood, and unknown inputs, among others—as meaning the unit has some secondary input that is neither diesel nor natural gas.

Regulation Level

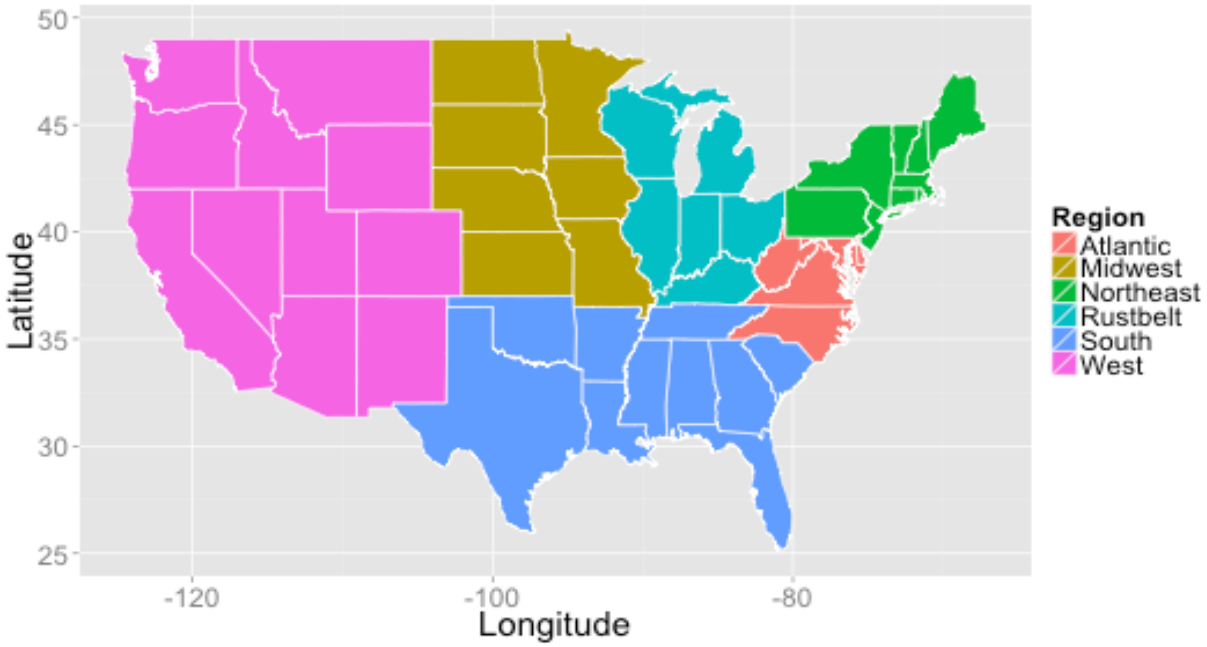
There are three types of regulations in the data: SO_2, NO_x and $PM_{2.5}$ Regulations. Each of these attributes indicates the authority from which the most stringent governmental regulations were imposed on the pollutant in question. Some regulation levels, particularly for nitrogen, are reported as missing in the data. We can interpret a missing value as meaning that either the regulation is unknown or that there is none, which we hypothesize might occur in exceptional circumstances for plants that are "grandfathered" in (meaning they are exempt from any new regulations). This means that for each of the three regulation types, we have four possible options: Federal, State, Local, or NA, where "Local" refers to the county or municipal level. The distribution is shown in Appendix B.3.

We examine all three regulation types because correlation between the three is not particularly high; only 346 units (35%) have the same value across all three outputs (i.e. State-level SO_2 regulation, State-level NO_x regulation and State-level $PM_{2.5}$ regulation). This means each type of regulation provides new information that could confound the analysis, and are therefore all considered.

Geographic Regions

There are coal-fired units all across the United States, but the Rustbelt and Eastern parts of the country had much higher densities than the western or southern regions. Figure 1 in Section 2.1.1 shows this distribution. Given that some states had very few observations, we chose to group states into larger regions. Figure 10 illustrates how we grouped the states.

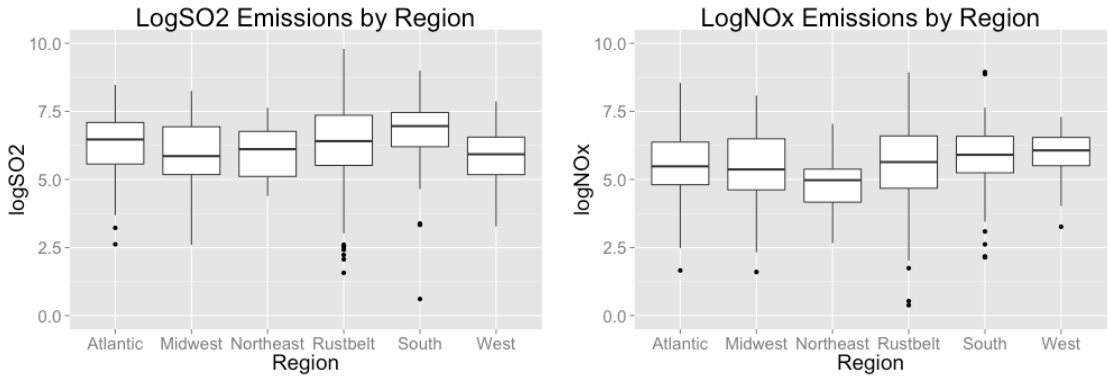
Figure 10: Map of States by Assigned Geographical Region



States colored by assigned geographical region. Historical trends in factory construction and purpose were taken into consideration.

Figure 11 provides some exploratory analysis on regional differences. We can see that the South has some of the dirtiest factories, at the top of the pack for both SO_2 and NO_x. The Rustbelt and Western regions also rank up in dirtiness, while the Midwest and Northeast start off with relatively clean factories.

Figure 11: Regional Differences in Emissions



Left: LogSO2 emissions in January 1997 by region. Right: LogNOx emissions in January 1997 by region. The South and Rustbelt have dirty factories in terms of both SO_2 and NO_x while the Northeast is relatively clean.

Has Start Date

We also included an indicator for whether or not the unit had a value listed for their start date, or initial year and month of operation (as called in the data). A high percentage of observations are missing this attribute (over 25% of the experimental dataset), which is not random; units with a value listed are older and dirtier on average. Given the non-randomness, we include an indicator for whether the unit has this value or not.

2.8. Experimental Datasets

Questions 1 & 2

To address questions one and two, we restricted the dataset to the 995 units that:

- Use coal as their primary fuel,
- Are active in January 1997,
- Have non-outlying data on Emissions, Heat Input and Sulfur Content in 1997,
- Are intended for electricity production, and
- Did not select a compensation compliance plan.

As mentioned in Section 2.4, we took care to exclude units that install a scrubber at some point during 1997 once they undergo treatment. Given that our analysis focuses on the impact of scrubbers on emissions from coal power plants, we only considered those that primarily use coal

as an input. Being active in January 1997 is required because we fit the propensity score model on this month's observations.

As described in the Data Challenges section, we define an outlying value as one that lies more than 1.5 inter-quartile ranges below the first quartile or above the third quartile of that quantity's empirical distribution. We require heat input and log emissions to fall within this range to keep the normal shape of their distributions.

The casualties of the cuts are almost exclusively on the left of the distribution, a phenomenon we suspect stems from the aggregation of daily to monthly data; units operating for only a handful of days or with a faulty monitor for any part of a month would record lower emissions or heat levels than expected. It is important to note that observations corresponding to incomplete (i.e. NA) values or zero also qualify as outlying, and are consequently also omitted.

Table 3: Distribution of Treated and Control Units for Q1 & Q2

Intervention	Total Units	Number Treated	Number Control
Question 1	995	178	817
Question 2	995	429	566

We included the electricity production and compensating compliance restrictions because of sample size concerns. Given that the other three criteria are met, there are only six units not meant for electricity production, and only one on a compensation plan. To avoid possible bias from such small samples, we chose to simply eliminate those units.

Questions 3 & 4

To address questions 3 & 4 we employed two methods: a propensity score algorithm and a variable ratio matching algorithm. For the propensity score model, we restricted the dataset used for Questions 1 & 2 further by including only the units without scrubbers in January 1997 (the control groups). More specifically, the control group used in Question 1, those who had not installed an SO_2 scrubber in January 1997, comprised the experimental dataset in Question 3; similarly, the control units in Question 2 were used as the experimental units in Question 4. The relationships between the study populations were outlined graphically in Section 2.4. There were 817 total units considered for Question 3, and 566 total units for Question 4.

For our matching algorithm, we used a slightly different dataset. The basic requirements for this group are that they:

- Use coal as their primary input, and
- Are active at some point between 1997 and 2012.

We further required that the treated units:

- Are active two months before and two months after scrubber installation, and
- Have non-outlying values for log heat input and log emissions for those months.

We placed no additional requirements on the control units, knowing that any units we might have excluded would likely not be matched to any treated units. For this second model, there were 235 units remaining for Question 3, and 478 for Question 4.

3. Methods

3.1. Potential Outcomes Framework

This analysis' approach is rooted in potential-outcomes methods for causal inference (Rubin 1978, Holland 1986, Rubin 2008), which interpret causal effects as consequences of specific, well-defined actions within an experimental paradigm (Hernan 2005, Hernan et al. 2008, Zigler and Dominici 2014). The central idea is to consider the regulatory intervention as a (possibly hypothetical) experiment in which there is an "observed condition" (e.g., SO_2 scrubber installation, Z=1) and a "world avoided condition" (e.g., no SO_2 scrubber installation, Z=0). In this hypothetical experiment, if populations (EGUs) could be randomly assigned to these conditions, differences in observed outcomes (log emissions) would be interpreted as causal effects of the intervention. For a specified outcome *Y*, we have two *potential outcomes*:

- $Y_i(Z=1)$ denotes the potential outcome that would be observed if the unit had installed a scrubber.
- $Y_i(Z=0)$ denote the potential outcome that would be observed if the *same* unit had not installed a scrubber.

In the context of Question 3, Y_i would denote the difference between log SO_2 emissions 2 months after scrubber installation and log SO_2 emissions 2 months before scrubber installation. If unit *i* indeed has a SO_2 scrubber, then $Y_i(Z=1)$ denotes the observed emissions change for that unit, whereas $Y_i(Z=0)$ denotes the *counterfactual* emissions change for unit *i*; that is, the *unobserved* change in SO_2 emissions that would have *potentially* occurred at *the same* unit *i* if the scrubber had not been installed. Therefore, the causal effect of scrubber installation (Z) on emissions (Y) at unit *i* is defined:

$$Y_i(Z=1) - Y_i(Z=0)$$

Since, at most, one potential outcome can actually be observed for each unit, average causal effects are estimated comparing average values of $Y_i(Z=0)$, observed on Z=0 units, to average values of $Y_i(Z=1)$ observed on Z=1 units.

3.2. Propensity Score Model

Theoretical Background

One salient challenge in estimating the causal effects of an intervention Z on an outcome Y is that the power plant units are not randomized to *Z=1* (scrubber installation) versus *Z=0* (no scrubber installation). Without randomization, it is necessary to confound for adjustment. Possible confounders for the causal effects of scrubber installation include differences in size, location, and other factors that determine whether an emissions control strategy is adopted.

We employ a particularly valuable tool for confounding adjustment in this context, the propensity score (Rosenbaum and Rubin 1983). The goal of propensity scores is to address confounding by forming groups of *Z=0* and *Z=1* units that have similar propensity scores, as units with similar propensity scores can be regarded as similar on the basis of observed confounding factors (Rubin 2008, Stuart 2010). A unit's propensity score is the probability it will receive treatment given a set of covariates. In statistical notation, a propensity score S can be defined as:

$$S = \Pr(Z = 1 \mid X = x)$$

We can estimate the propensity score vector S by running a logistic regression of the form Logit(S) = βX, where the logit transformation of vector S is the product of vector β and covariate matrix X. The logit transformation, also known as the log-odds transformation, of a probability *p* is defined as *Log(p/(1-p))*. Any model relating a binary outcome to a set of confounders is acceptable, and logistic regressions are the most widely used method for propensity score estimation (Stuart, 2010).

Using the propensity score, it is possible to estimate the counterfactual emissions that would have occurred at a treated unit had no scrubber been installed, provided the assumption of *ignorability* (to be addressed further in the next section) is satisfied. This can be done by using observed data on emissions for units that did not install a scrubber, but that were comparable on the basis of confounding factors to the unit that installed the scrubber. Using observed data on *Z=0* units to inform counterfactual emissions scenarios acknowledges that emissions changed during this time period for reasons other than scrubber installation.

This data-informed approach represents a refinement over, for example, an approach that (hypothetically) assumes that emissions would have remained constant if scrubbers had not been installed. The ability to confirm that observed confounders are in fact balanced between *Z=0* and *Z=1* units is a key feature of the approach that is in contrast to, for instance, a regression model that permits no such confirmation and may implicitly extrapolate to derive effect estimates from units that are not comparable (King and Zeng 2006, Rubin 2008, Stuart 2010). We conduct sensitivity analyses to gauge the extent to which analyses are susceptible to the essential yet un-testable assumption that the propensity score includes all relevant confounding information (Rosenbaum and Rubin 1983, Vanderweele and Arah 2011).

Confounders Used in Analysis

Propensity score methods rely on the concept of *ignorability*, which assumes that there are no unobserved differences between treatment groups, conditioning on the observed covariates (Stuart, 2010). To meet this assumption, it is imperative to include all covariates highly correlated with both the treatment assignment and outcome in the propensity score model (Glazerman et al., 2003). In general, there is little harm in having too many covariates in the propensity score model; irrelevant attributes will have trivial impacts on the propensity score,

whereas omitting important ones can add significant bias (Stuart, 2010). This means it is normally beneficial to include more rather than fewer confounders when specifying the propensity score model.

Table 4: Covariates Used in Propensity Score Model

Covariate	Question 1	Question 2	Question 3	Question 4
logHeat	✓	✓	✓	✓
logNO_x	✓	×	✓	✓
logSO_2	×	✓	✓	✓
logCO_2	✓	✓	✓	✓
Operating Time	✓	✓	✓	✓
Has SO_2 Scrubber	×	✓	×	✓
Has NO_x Scrubber	✓	×	✓	×
Has $PM_{2.5}$ Scrubber	✓	✓	✓	✓
Sulfur Content	✓	✓	✓	✓
SO_2 Phase Indicators	✓	✓	✓	✓
NO_x Phase Indicators	×	✓	×	✓
Secondary Fuel Indicators	✓	✓	✓	✓
SO_2 Regulation Indicators	✓	✓	✓	✓
NO_x Regulation Indicators	×	✓	×	✓
$PM_{2.5}$ Regulation Indicators	✓	✓	✓	✓
Geographic Indicators	✓	✓	✓	✓
Has Start Date	✓	✓	✓	✓

There are three notable exceptions to this general rule. First, we must not include any covariates that may have been affected by treatment, as these would be consequences—not predictors—of the treatment (Rosenbaum, 1984; Frangakis and Rubin, 2002; Greenland, 2003). Second, covariates highly correlated with the treatment, but not the outcome, should not be included as these covariates are instruments, and it would be virtually impossible to extricate these attributes' impact on the outcome from the treatment's (Stuart, 2010). Finally, we must be careful not to include so many covariates as to overfit the propensity score model. By including too many covariates, we run the risk of separating the treatment groups' propensity score distributions

(Schafer & Kang, 2008). We aim for at least 10 treated observations (or 20 total observations) per each covariate, as any fewer has been shown to generate high mean squared errors of the estimate (Lowe et al., 2013).

All four propensity score models are estimated using January 1997 values. Table 4 lists the covariates used for each of the interventions. The only difference between the pre-1997 and post-1997 analyses is the omission of SO_2 or NO_x emissions. This is done because these attributes are consequences of the treatment. Though other emissions observations also occur after treatment for Questions 1 and 2, they are included in the propensity score model because we verified that scrubber installations have no impact on other types of emissions, and we hypothesize that these values might still be indicative of power plant features that could confound the analysis.

For analyses examining the effect of SO_2 scrubber installation post-1997, we omit the indicator for having an SO_2 scrubber because the units in the experimental dataset necessarily have no scrubber (the same is true for the NO_x analyses). All other covariates are included or excluded from the models based on their correlation with the treatment and the outcome.

Subclassify & Assess Balance

After calculating propensity scores for each unit in the analyses, we compare the range of scores for the treatment and control units to find the *common support*, or the range of overlapping scores between the two treatment groups. All observations outside the region of common support are discarded to ensure comparable units exist across the two groups (as done in Heckman *et al.*, 1997; Dehejia and Wahba, 1999). Appendix C.1 and C.2 have the distributions of propensity scores for all interventions.

Once the extreme values are eliminated, we stratify the units into subclasses. Subclassification is a propensity score matching methodology in which experimental units are grouped with similar units, as measured by the propensity score (Rosenbaum & Rubin, 1985). Stratifying into five subclasses—the number used for all analyses in this paper—has been shown to remove at least 90% of the bias in the estimated treatment effect due to all of the covariates that went into the propensity score (Rosenbaum & Rubin, 1985). To confirm that improved balance has been attained, we can compare the covariate standardized differences in means between treated and untreated before and after subclassification.

Appendix C.3 assesses the new balance by plotting the standardized differences of the original datasets with our new subclassified ones. For each covariate in the original dataset, the standard difference is the difference in means between the treatment and control groups divided by the pooled standard deviation between the two treatment groups. For each covariate in the new dataset, the difference in means is calculated by taking a weighted average of the individual subclass standardized difference in means, with the *number of treated units* as the weights. This is done to estimate the Average Treatment Effect (ATT) and will be explained further in the Analysis of Outcome section (Stuart, 2010). We use the pooled standard deviation of the original dataset as the denominator for both sets of vectors (as recommended in Stuart, 2010). This choice allows us to fairly judge improvements in covariate balance.

Table 5: Distribution of Units by Subclass After Limits to Common Support

Intervention	Treatment	Subclass 1	Subclass 2	Subclass 3	Subclass 4	Subclass 5	Total
Question 1	Treatment	14	18	30	29	66	157
	Control	466	126	114	67	30	803
Question 2	Treatment	77	53	56	73	141	400
	Control	303	89	67	41	49	549
Question 3	Treatment	31	28	31	25	81	196
	Control	341	83	66	27	31	548
Question 4	Treatment	11	28	144	39	130	352
	Control	38	30	49	9	6	132

Given the time-varying nature of treatment, it is necessary to take additional precautions to ensure that balance is retained throughout the time period. The balance reported in Appendix C.3 measures only that in January 1997; this does not guarantee that balance will be sustained until 2012. Consequently, we also calculate the balance in following years. Across interventions 1-3, balance remains strong, as illustrated in Appendix C.4. Question 4's balance becomes suspect after 2002, indicating that the matching method may be a more reliable estimate for this intervention.

3.3. Variable Ratio Matching by Covariates

In addition to stratifying by propensity score, we run a second analysis in which we match treated units to comparable control units. Though we check and show that the propensity score stratification creates balance that is maintained over time, it definitely deteriorates by the final years. A second, independent analysis that produces similar results would lend credibility to the propensity score methodology's estimates.

Theoretical Background

Matching algorithms, like propensity score methods, attempt to solve the problem of non-randomization by grouping treatment (*Z=1*) units with similar control (*Z=0*) units. However, rather than computing an additional value (the propensity score) and matching on that, pure matching algorithms pair treatment units with the *closest* control unit (or units), as measured by a pre-determined distance measure.

In this paper, we match units with the R package *Coarsened Exact Matching* (CEM), an algorithm that coarsens continuous variables into bins and matches exactly on these new discrete values (King et al., 2014). It matches with a variable ratio matching algorithm, meaning treatment units are paired with any and all control matches (Ming and Rosenbaum, 2001). Variable ratio matching is distinct from K:1 nearest neighbor matching in that it allows each treated unit to have a different number of matches. This flexibility gives it an advantage over K:1 matching because it does not suffer from the bias-variance tradeoff—in which the decreased variance from additional matches results in increased bias from inclusion of worse matches—that results from fixing the number of matches per treated unit (Stuart, 2010).

The CEM approach also sidesteps the challenge of multidimensional matching presented by Chapin (1947) by allowing for more matches per unit without requiring the introduction of a propensity score, which has been challenged in recent literature for possibly adding bias and model dependence (Chapin, 1947; King et al., 2015). Additionally, CEM has been shown to outperform more traditional matching methods—such as those based on Mahalanobis distance, nearest neighbors, or propensity scores—on both creating balance and reducing the mean squared error of the treatment effect (King et al., 2011).

Once the algorithm generates matches for the treated units, we can perform analysis on this reduced dataset as if it were a randomized control trial (Ho et al., 2007). This method is in contrast to taking the treatment effect within each matched set and averaging. It is also more flexible and allows for more sophisticated analysis of outcome tools, such as regression adjustment, which we use in this paper.

Covariates Used in Analysis

We use a similar set of covariates as those used in the propensity score model, but due to the challenges associated with matching on a high number of attributes we restricted the covariates to those very highly correlated with treatment and outcome. Table 6 shows the remaining covariates for each intervention. Note that elements of the form ***I(X)*** denote an indicator that the statement *X* is true.

Table 6: Covariates Used in CEM Algorithm

Intervention	Covariates
SO_2	Year, Month, logSO_2, logNO_x, logHeat, I(Phase 1), I(Sulfur Regulation is NA), I(Particulate Regulation is NA)
NO_x	Year, Month, logSO_2, logNO_x, logHeat, I(Sulfur Regulation is NA), I(Particulate Regulation is NA)

Units are matched exactly on month so as to eliminate any seasonal differences in emission changes. They are matched within three years, but not exactly on year. The year is only important because emissions decline over time; if two units separated by a few years are equivalent in terms of emissions, then they would likely respond equivalently to treatment. Units are also matched on all emissions values, and in the SO_2 intervention are matched on the indicator of whether the unit was in Phase 1, which was meant for the highest polluters. LogCO_2, though highly correlated with the outcome, is omitted due to its 0.999 correlation with logHeat.

Somewhat surprisingly, the dummies for having missing values for SO_2 and NO_x were highly associated with both outcomes. This result is especially unexpected given that missing SO_2 regulatory values have more impact on changes in NO_x emissions than any NO_x regulations. This occurrence lends credence to the hypothesis that at least some missing values indicate no restrictions even on the federal level, reflecting a grandfathering process.

3.4. Analysis of Outcome

Average Treatment Effect on the Treated (ATT)

In assessing the outcome using a propensity score or matching algorithm, we have two choices of estimand: the Average Treatment Effect (ATE), or the Average Treatment Effect on the Treated (ATT). The ATT should be used when we are interested in the treatment effect on a narrow portion of the population (Imbens, 2004). Though an argument can be made for using the ATE in these interventions (to estimate how any unit's emissions would change by installing a scrubber) we choose to estimate the ATT because, due to unit size, fuel inputs (i.e. low sulfur coal in the vicinity), or some other factor, some units may never install a scrubber. Estimating the ATT focuses us on the causal effect of a scrubber only on the units that might have feasibly installed one, which avoids possibly introducing unmeasured confounder bias from extrapolating to dissimilar populations (Austin, 2011). Regardless, estimates for the ATT and ATE are often very similar in practice (Imbens, 2004).

Applied to propensity score methods, the ATE and ATT differ only in the weights used on the subclasses. The ATE calls for the subclasses to be weighted by the number of units in each, while the ATT requires the subclasses to be weighted just by the number of treated units in each (Imbens, 2004). These weights are used both in the assessing of balance phase (as shown in Section 3.3) and in pooling the variances to estimate the standard error for a difference in means estimation (as done in Lunceford & Davidian, 2004). K:1 matching measures the ATT by definition, as the control units are supplied expressly to provide counterfactuals for the treated units.

Two Months Before and After

The choice to measure a change in emission from two months before to two months after installation was made to maximize the number of observations available for analysis. First, the month of installation must be ruled out from analysis; since our dataset reports monthly figures, we can at best know the month a unit installs a scrubber. The installation could have occurred on any day in the month, meaning we must compare full months on either side of the installation month.

The months immediately preceding and following scrubber installation were ruled out due to unusually high levels of both missingness and outlying heat and emissions values. We attribute this phenomenon to installation-related procedures that caused either full or partial unit shutdowns. Measuring the change in emissions from two months before and after—as opposed to one month on either side—adds no additional validity threats to the established methodology.

Difference in Means & Regression Adjustment

The outcome for Questions 1 & 2 is simply the weighted difference in means of log emissions between treated and non-treated units by subclass. The outcome of interest is emissions in 1997, but since we have data through 2012 we can also estimate the impact of pre-1997 scrubber installations on all years between 1997 and 2012.

In the propensity score and matching algorithms for Questions 3 & 4, we calculate the treatment effect after adjusting for changes in key covariates via a linear regression model. Regression and matching should not be thought of as competing forms of analysis, and in fact have been shown to complement each other very well (Rubin, 1973b; Carpenter, 1977; Rubin, 1979; Robins and Rotnitzky, 1995). In tandem, matching balances the experimental dataset, removing systematic bias, and regression adjustment cleans up any remaining residual imbalances (Stuart, 2010). We use the following two models of regression adjustment, each of which is presented in the Results section.

$$Y = \beta_0 + \beta_1 * Z$$

$$Y = \beta_0 + \beta_1 * Z + \beta_2 * \Delta(\log Heat)$$

In these model, Z is the indicator of treatment and the outcome Y is the change in log emissions from time *t-2* to time *t+2*, as defined in Section 2.5. Model 1 is simply a regression of the outcome on treatment, and so is equivalent to estimating the difference in means. Model 2 includes the change in log heat over the same time period *t-2* to *t+2*. This adjustment is designed to eliminate noise in the effect estimation resulting from changing unit behavior that was not captured by the matching algorithms. In both cases, β_1 is interpreted as the treatment effect.

Regression adjustment has been shown to actually increase bias when the relationship between the covariate and the outcome is even slightly non-linear (Cochran and Rubin, 1973). Plots

demonstrating the linear relationship log Heat has with both log SO_2 and log NO_x emissions are visible in Appendix C.5 Appendix C.6 demonstrates the non-linear relationship between Operating Time and emissions, explaining why it was not used in this regression adjustment.

In the propensity score model, the above regression adjustment model is ideally calculated within each subclass and subsequently averaged across strata (Lunceford & Davidian, 2004). However, given the relatively small samples in some of the subclasses (refer to Table 5), we choose to regress the outcome on a series of interactions between the treatment *Z* and indicators for each of the five subclasses, and the change in logHeat. In practice, this means the *Z* above is substituted for a row vector of five interactions between the treatment indicator and each of the five subclasses.

$$Z = \sum_i \delta_i * Z_i$$

Each Z_i is understood to be 1 if the unit receives treatment and is in subclass *i*, and 0 otherwise. This regression structure assumes that logHeat affects the outcome uniformly across all five subclasses, which is a more stringent assumption than the alternative but is not unreasonable given the relationship between logHeat and emissions presented in Appendix C.5. The final point estimate is a weighted average of the five $\delta i * \beta_1$ estimates, with the number of treated units used as the weights. Since Var(a*X) = a^2 (Var(X)), the weights are squared in the weighted sum to estimate the standard error. Please note that for the reasons presented in this section, we favor the regression adjustment model takes over the difference in means as an estimator of the treatment effect.

4. Results

In this section, we present the results of the methodology outlined thus far. As outlined in Section 1.3, there are four interventions being assessed in this study:

1) The impact of SO_2 scrubbers installed before 1997 on 1997 log SO_2 emissions
2) The impact of NO_x scrubbers installed before 1997 on 1997 log NO_x emissions

3) The impact of SO_2 scrubbers installed after 1997 on the change in log SO_2 emissions two months before installation to two months after
4) The impact of NO_x scrubbers installed after 1997 on the change in log NO_x emissions two months before installation to two months after

Within these interventions, we measure several different quantities. For Questions 1 & 2, because we have the benefit of 15 additional years of outcome data (1998-2012), we show the causal effect applied to these future years as well.

Table 7: Interpreting Models in Results Section

Matching Method	Weighted Mean	Regression Adjustment
Propensity Score Subclassification	Model 1	Model 2
Variable Ratio Matching	Model 3	Model 4

For each of Questions 3 & 4, we measure the outcome with a propensity score method and a matching algorithm. We provide two estimates for each method and question: a weighted difference in means (Models 1 & 3), and a regression-adjusted point estimate that accounts for changes in the log of heat input over the same time period (Models 2 & 4). The names for the models are visible in Table 7. We further note that in this section, parentheses following point estimates are understood to denote the 95% confidence interval.

4.1. Questions 1 & 2

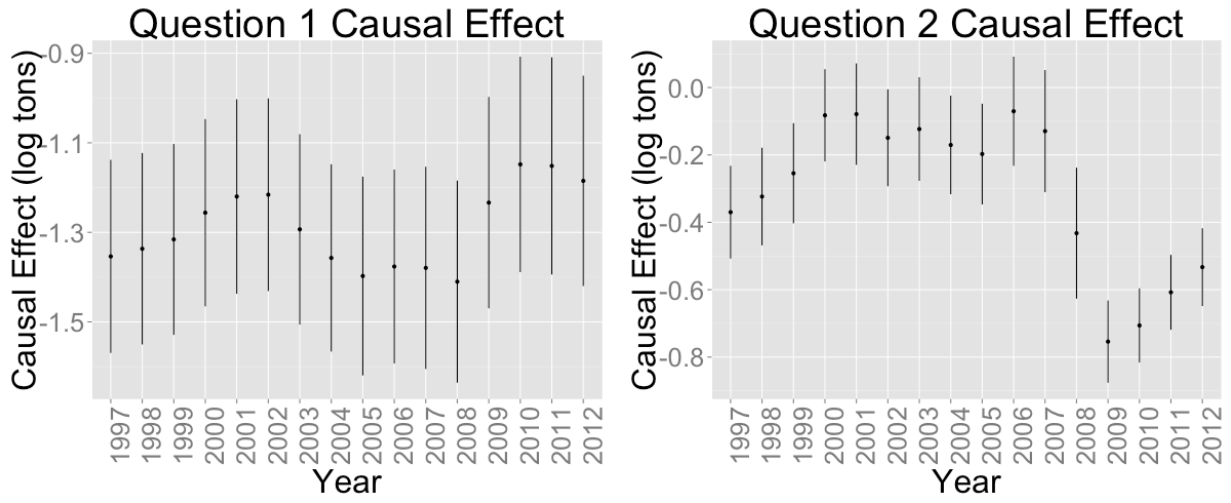
Table 8 shows the causal effect of installing both SO_2 and NO_x scrubbers on 1997 emissions. Both results are significant to well below the 0.1% level. SO_2 scrubber installations reduce log SO_2 emissions by 1.36 (1.15, 1.58) log tons in 1997 on average, which translates a 74% decline in tons of SO_2. NO_x scrubbers reduce log NO_x emissions by 0.33 (0.18, 0.48) log tons in 1997 on average, which equates to a 28% decline in tons of NO_x.

The causal effects of installing SO_2 and NO_x scrubbers by year are summarized in Figure 12. The difference in SO_2 emissions between the treated and untreated units remains significantly below zero for the duration of the period. In contrast, the causal effect of installing a NO_x scrubber quickly disappears before recurring in 2008.

Table 8: Summary of Key Results for Questions 1 & 2

Intervention	ATT (95% CI)	Percent Change (95% CI)	Standard Error	p
Question 1	-1.36	-74%	0.11	<0.001
	(-1.58, -1.15)	(-79%, -68%)		
Question 2	-0.33	-28%	0.08	<0.001
	(-0.48, -0.18)	(-38%, -16%)		

Figure 12: Causal Effect of Scrubber Installation Over Time



Causal effect of installing an SO_2 (left) or NO_x (right) scrubber over time, with 95% confidence intervals of the estimate.

4.2. Questions 3 & 4

Table 9 presents the propensity score models' results to Questions 3 & 4. Both estimates for Question 3 indicate SO_2 scrubbers are incredibly effective, removing 80-90% of SO_2 after installation. Model 3.2 estimates the scrubbers as 3.8% more effective at 88.8%, just outside Model 3.1's 95% confidence interval. Please note the much smaller standard error in Model 3.2, likely resulting from the additional term cleaning up a lot of the residual noise.

The models for Question 4 seriously disagree about the causal effect of installing a scrubber. Model 4.1 estimates a 0.33 *increase* in NO_x emissions following NO_x scrubber installation, while Model 4.2 estimates a 0.24 *decrease* in emissions. Both estimates are significant to the 0.1% level. Given that 71% of units have a positive change in log heat over the time period (average

increase of 0.59 log MMBtu, which is an 81% increase in heat input), it appears likely that Model 4.1 is picking up a lot of noise that Model 4.2 adjusts for. We also remind the reader that Question 4 had the least reliable balance of the four propensity score models.

Table 9: Propensity Score Results

Intervention	Model No.	ATT (95% CI)	% Change (95% CI)	Standard Error	p
Question 3	3.1	-1.88 (-2.13, -1.63)	-85% (-88%, -80%)	0.13	<0.001
	3.2	-2.19 (-2.26, -2.12)	-88.8% (-89.5%, -88%)	0.03	<0.001
Question 4	4.1	0.33 (0.18, 0.48)	39% (19%, 62%)	0.08	<0.001
	4.2	-0.24 (-0.27, -0.21)	-21% (-24%, -19%)	0.016	<0.001

Table 10 presents our final four results, the outcomes of the variable ratio matching algorithm. The results are remarkably similar to the ones in Table 9; the point estimates for the percent change in SO_2 emissions are nearly identical to those from the propensity score method, and the two models for Question 4 produce similarly contradictory results. The results are all significant to the 5% level, with three of the four significant to below the 0.1% level. Both interventions have high match rates, as 72% of treated units in Question 3 and 64% of treated units in Question 4 receive matches.

Table 10: Matching Algorithm Results

Intervention	No. Treated Matched	Model No.	ATT (95% CI)	% Change (95% CI)	Standard Error	p
Question 3	113/157 (72%)	3	-1.87 (-2.07, -1.66)	-85% (-87%, -81%)	0.09 -	<0.001 -
	- -	4	-2.13 (-2.25, -2.01)	-88% (-89%, -87%)	0.06 -	<0.001 -
Question 4	167/260 (64%)	3	0.16 (0.01, 0.31)	17% (1%, 36%)	0.08 -	0.031 -
	- -	4	-0.22 (-0.25, -0.18)	-20% (-22%, -17%)	0.02 -	<0.001 -

4.3. Sensitivity Analyses

Variation of the Propensity Score Model

The inputs to the propensity score models were selected based on their high correlation with the treatment and outcome, with three exceptions; the attributes must be pre-treatment, not be instruments, and not be too many in number so as to overfit the model. In this section we'll take the "kitchen sink" approach and add any and all covariates that we hypothesize to have an impact on the outcome. Table 11 shows the results derived from this more liberal approach.

Table 11: Estimations under Variation in the Propensity Score Model

Intervention	ATT (95% CI)	Percent Change (95% CI)	Standard Error	p
Question 1	-1.21 (-1.51, -0.91)	-70% (-78%, -60%)	0.15	<0.001
Question 2	-0.37 (-0.51, -0.23)	-31% (-40%, -21%)	0.07	<0.001
Question 3	-2.17 (-2.24, -2.10)	-88.5% (-89.3%, -87.7%)	0.036	<0.001
Question 4	-0.24 (-0.27, -0.21)	-22% (-24%, -19%)	0.016	<0.001

The results for Questions 3 & 4 in this table are the estimates from the regression-adjusted model (Models 3.2 & 3.2). The additional covariates led to slightly more separation in the multi-dimensional histograms (e.g. Question 3 had 10% fewer observations resulting from additional cuts in the "limit to common support" step), but the estimates are very close to those presented in the previous section. The estimates for Questions 3 and 4 are particularly close, within one percentage point each time. Estimates for Questions 1 and 2 are within three to four percentage points of the pervious estimates, and well within the 95% confidence intervals.

Variation of the Variable Ratio Matching Model

The above analysis shows that changing marginal inputs to the propensity score model has little effect on the final estimates. However, with the CEM method we can also adjust the proximity of

values required for a match. Here we present two alternatives for each of Questions 3 & 4, one with more stringent matching requirements and one with more relaxed ones. Both are estimated using the regression-adjustment method (Models 3.4 & 4.4). As a reminder, the models presented above matched 72% of SO_2 and 64% of NO_x treated units.

Table 12: Estimations under Variations in the Coarsening Requirements

Intervention	Stringency (% Treated Matched)	ATT (95% CI)	Percent Change (95% CI)	Standard Error	p
Question 3	Stringent (39%)	-1.68 (-1.92, -1.46)	-81% (-85%, -77%)	0.11	<0.001
Question 3	Relaxed (87%)	-2.08 (-2.14, -2.01)	-87.5% (-88.3%, -86.7%)	0.032	<0.001
Question 4	Stringent (35%)	-0.21 (-0.26, -0.15)	-19% (-23%, -14%)	0.029	<0.001
Question 4	Relaxed (77%)	-0.25 (-0.28, -0.22)	-22% (-24%, -20%)	0.014	<0.001

These estimates are also in the same ballpark as the original estimates. The two estimates for Question 4 and the second for Question 3 are again nearly identical to the previous estimates. The stringent alternative for Question 3 is the most different estimate, but it also has the widest confidence intervals. -1.68 falls within the 95% confidence interval of the original matching algorithm's estimate.

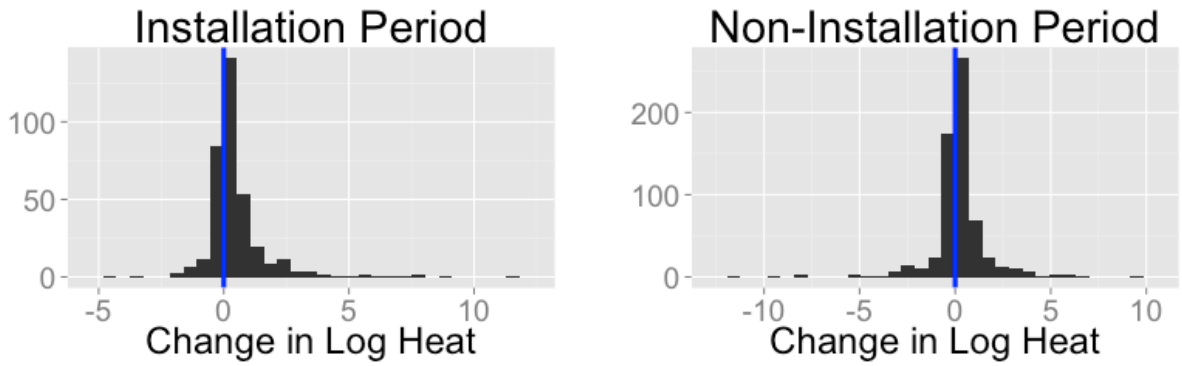
5. Discussion

Explanation of Results

The highly statistically significant and negative results virtually across the board indicate that scrubbers are an effective air quality control measure. SO_2 scrubbers are significantly more effective, with efficacy rates approaching 90 percent, while NO_x scrubbers can expect to remove around 20-30 percent of NO_x emissions. The consistent figures across the three primary analyses (pre-1997, post-1997 propensity score, post-1997 matching) and the series of sensitivity analyses gives confidence in these estimates as representative of the technologies' true impact.

The extremely conflicting estimates of NO_x scrubber effectiveness are striking, particularly because the same pattern emerges in both the propensity score and matching methods. The fact that heat input rises on average over the five-month span centered on scrubber installation is itself unsurprising; units installing a scrubber are likely to have been close to a NO_x emissions cap, and so can ramp up production upon scrubber installation. However, the 81% increase in heat input is quite dramatic. Figure 13 contrasts the change in log heat over the five-month period centered on NO_x scrubber installation for the treated units with a random five-month span, showing how anomalous the occurrence is.

Figure 13: Change in Log Heat over Installation and Non-Installation Periods



Change in log Heat for treated units two months before to two months after NO_x scrubber installation (left) and over a random five-month period (right).

It is worth noting that these treated units' gross load, a highly correlated attribute that measures electricity generation, rises by 52% over this period. This is also a sizable jump, but does not completely account for the rise in heat input. More investigation into how scrubbers operate is needed to fully understand this phenomenon.

Conclusion

In this paper, we linked several data sources and applied two different matching methods to estimate the causal effect of installing SO_2 and NO_x scrubbers on emissions in American coal-fired power plants. We found that installing an SO_2 scrubber before 1997 results in 74% (68%, 79%) lower SO_2 emissions in 1997, and installing a NO_x scrubber before 1997 results in 28% (16%, 38%) lower NO_x emissions in 1997. Furthermore, SO_2 scrubbers installed during the period 1997-2012 result in 88.8% (88%, 89.5%) lower SO_2 emissions, and NO_x scrubbers

installed between 1997-2012 result in 21% (19%, 24%) lower NO_x emissions, relative to units that do not install a scrubber. These results are corroborated by a matching algorithm, which finds SO_2 scrubbers cause SO_2 emissions to fall 88% (87%, 89%) and NO_x scrubbers to cause NO_x emissions to fall 20% (17%, 22%).

One strength of this analysis is our ability to link information on our intervention (scrubber installation) with the outcome variable: emissions. We also have access to alternative interventions and potential confounders—such as the sulfur content of coal, secondary fuels, operating time, heat input and relevant government regulations—that serve as controls in our statistical models. We successfully applied two different methods, subclassification by propensity score and variance ratio matching, to the interventions. In addition, we examined questions in which we considered the treatment as time invariant (Questions 1 & 2), and as a pre-post analysis (Questions 3 & 4).

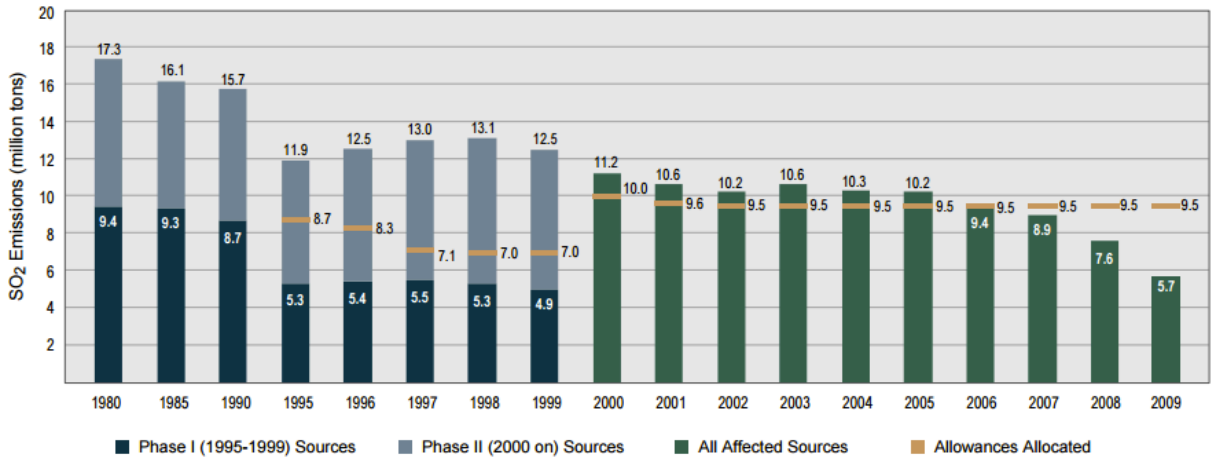
The biggest limitation to this analysis is its nature as an observational study. Because treatment is not randomized, it is possible our estimates suffer from unmeasured confounder bias, in which a latent variable biases the outcome. The dataset also suffers from many missing and outlying values. The initial year of operation and sulfur content attributes were partially excluded from the analysis due to such high quantities of missing values. The omission of these two attributes, and any others not measured in our dataset, may have biased our estimates.

As far as we know, this paper is the first comprehensive study to quantitatively evaluate the efficacy of a particular air quality intervention on emissions. The assembled data and proposed methodology has applications for future air quality analysis. Furthermore, the results of this analysis can hopefully be used in conjunction with future studies on the effectiveness of changing to low-sulfur coal or other fuel inputs to inform researchers, policy-makers and the energy industry on the efficacy of available options for air quality regulations.

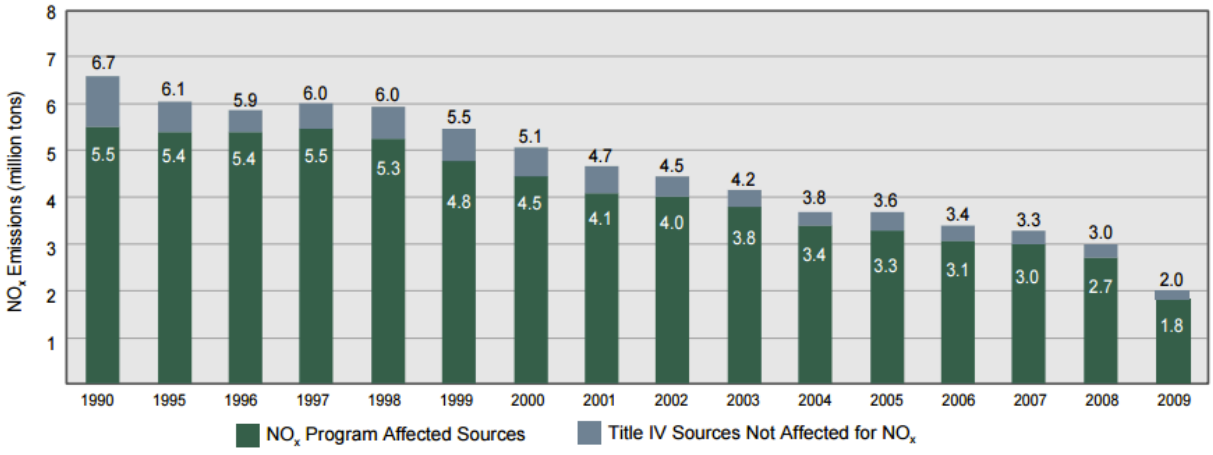
6. Appendix

Appendix A: Acid Rain Program

APPENDIX A.1: ANNUAL SO_2 EMISSIONS FROM EGUS IN THE UNITED STATES (1980-2009)



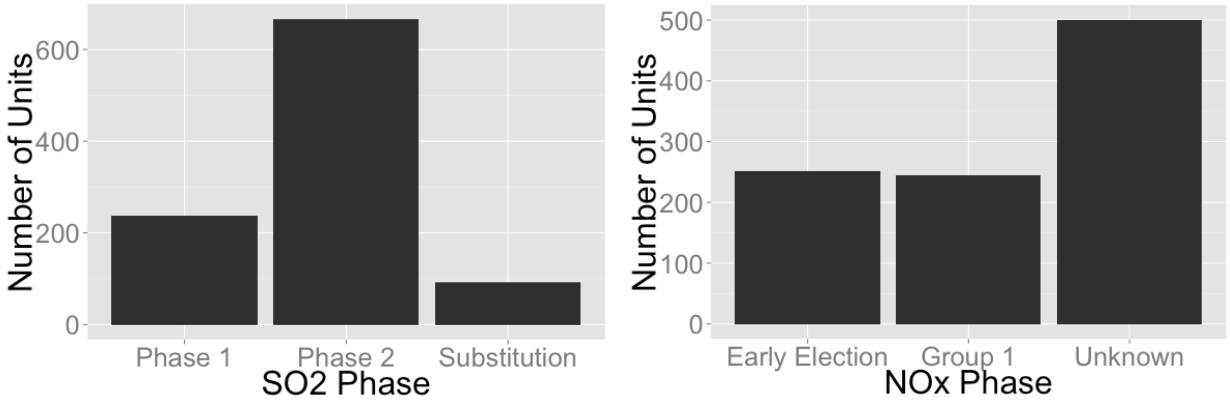
APPENDIX A.2: ANNUAL NO_X EMISSIONS FROM EGUS IN THE UNITED STATES (1990-2009)



Source: EPA, 2010

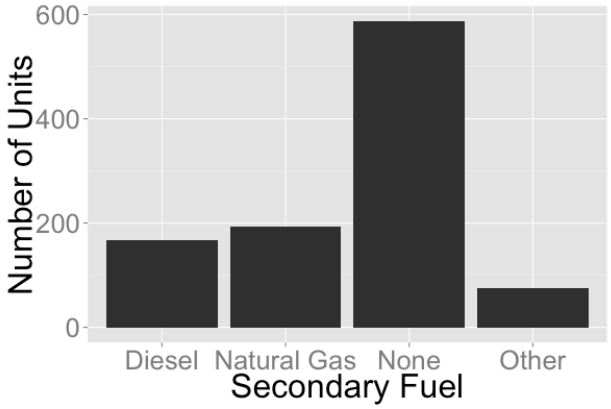
Appendix B: Unit-level Information

APPENDIX B.1: DISTRIBUTION OF SO2 AND NOX PHASES



Distributions of SO_2 (left) and NO_x (right) Phases. Notice the high volume of unknown data points on the NO_x Phase histogram.

APPENDIX B.2: DISTRIBUTION OF SECONDARY FUEL INPUTS



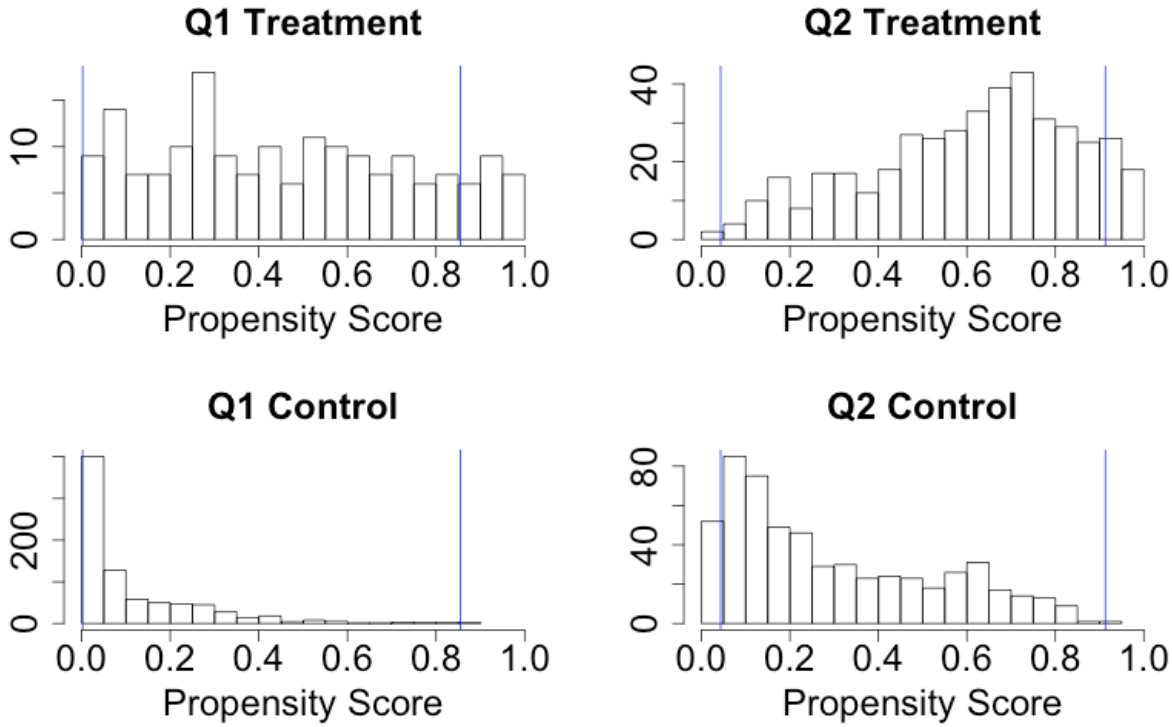
The frequency of secondary fuel inputs in January 1997 units. Note that the diesel and natural gas categories are not mutually exclusive (there are 30 units that employ both fuels).

APPENDIX B.3: DISTRIBUTION OF REGULATION STRINGENCIES

	Federal	**State**	**Local**	**NA**
Sulfur	152	763	31	51
Nitrogen	316	263	1	417
Particulate	106	814	26	51

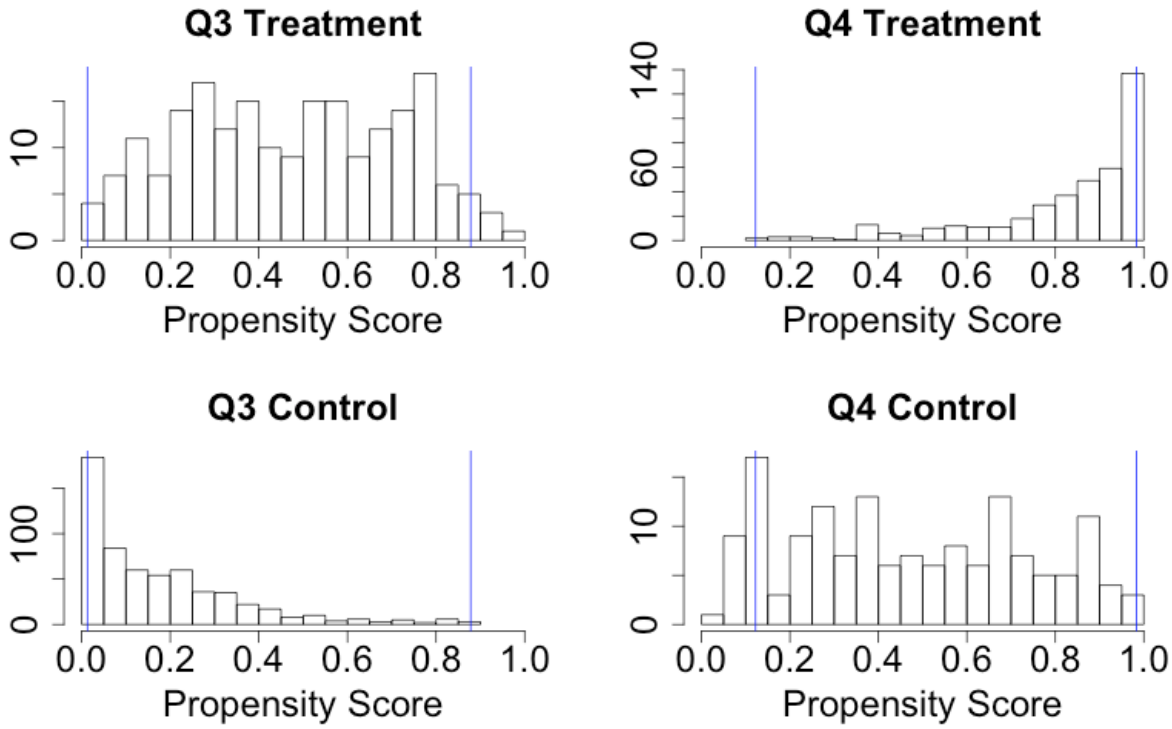
Appendix C: Steps in Methodology

APPENDIX C.1: DISTRIBUTION OF PROPENSITY SCORES (QUESTIONS 1 & 2)

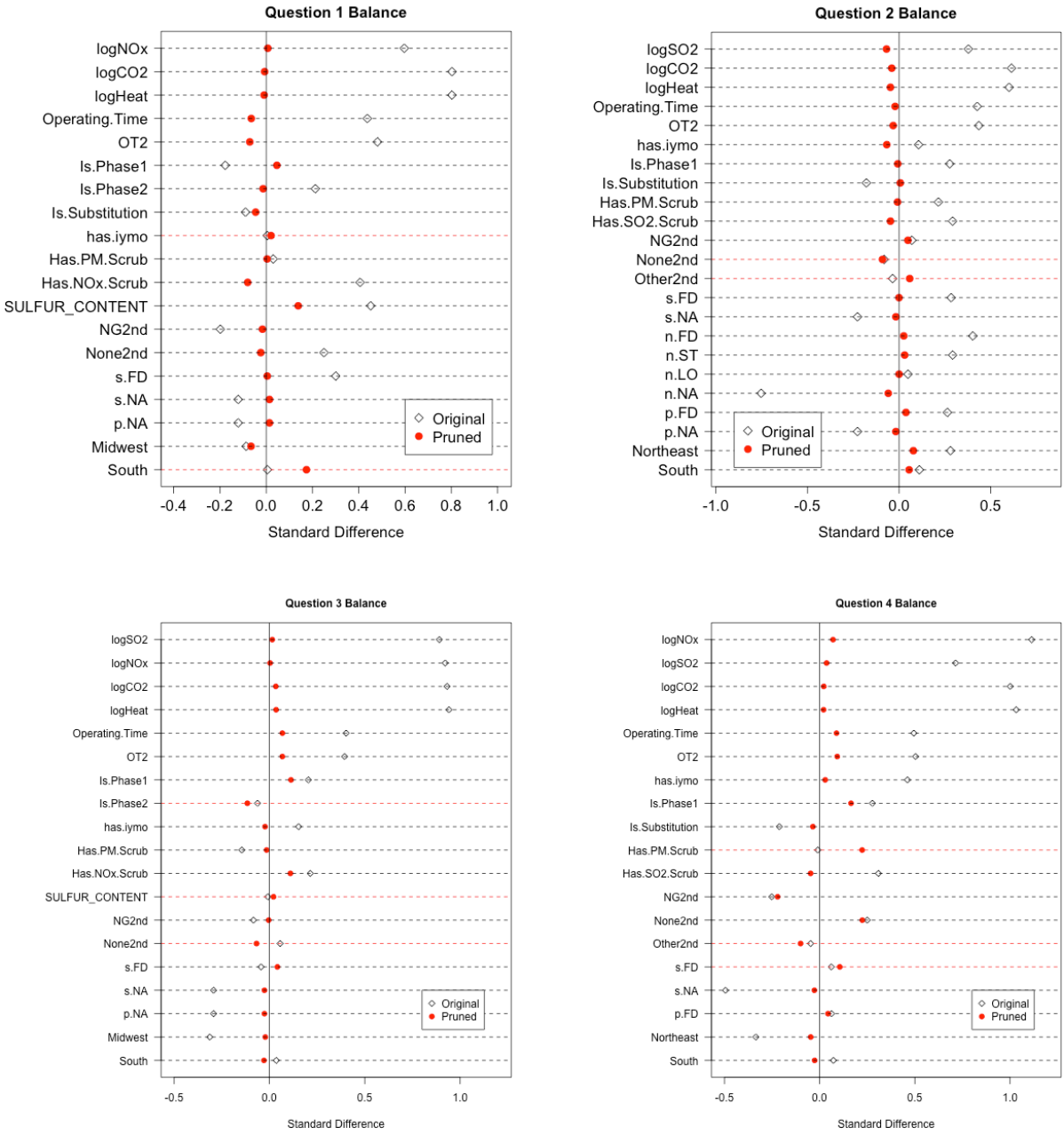


Distribution of propensity scores for Questions 1 (left) & 2 (right), by both treatment and control. The blue lines denote the limit of common support region.

APPENDIX C.2: DISTRIBUTION OF PROPENSITY SCORES (QUESTIONS 3 & 4)

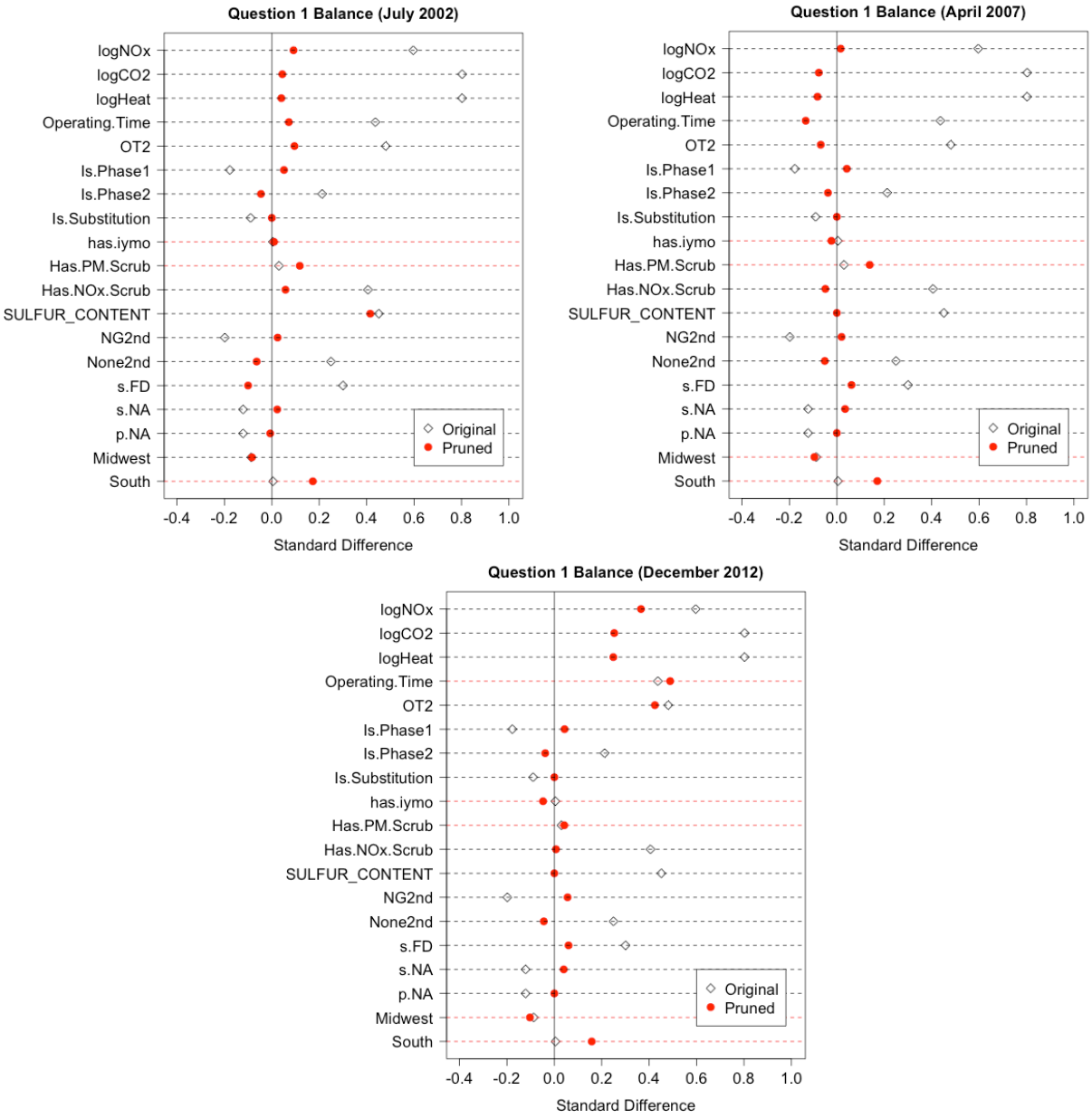


APPENDIX C.3: BALANCE ASSESSMENTS



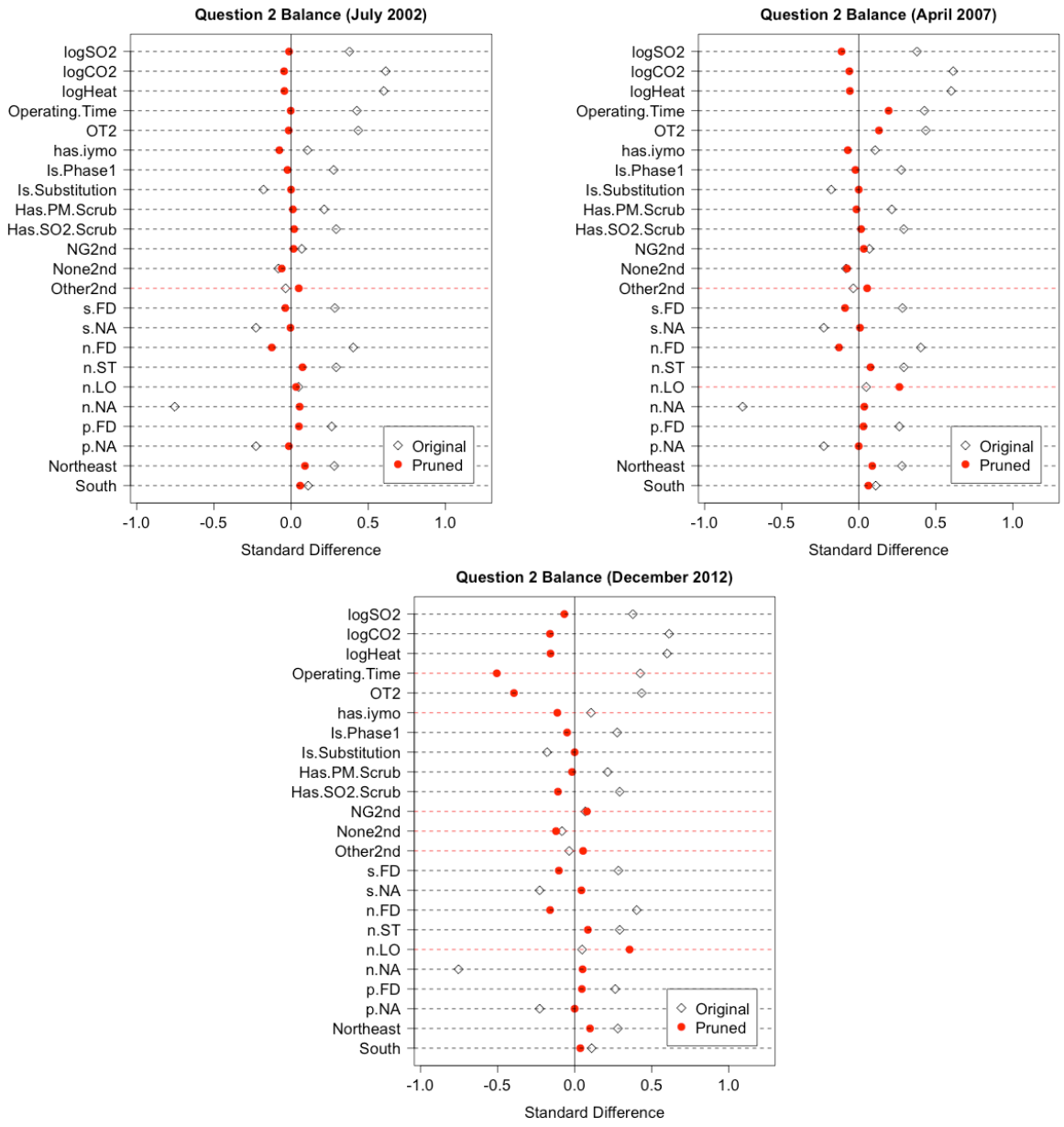
Balance Assessments: Original (white) vs Subclassified (red) standardized difference in means

APPENDIX C.4: FUTURE BALANCE ASSESSMENTS



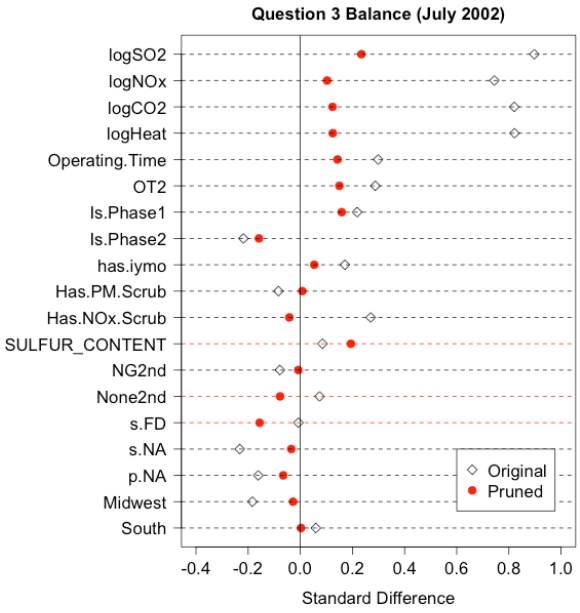
Balance as measured by standardized difference in means between treatment groups for Question 1 in July 2002 (top left), April 2007 (top right), and December 2012 (bottom). Balance clearly deteriorates by 2012, but is very strong through April 2007 on all covariates other than sulfur content.

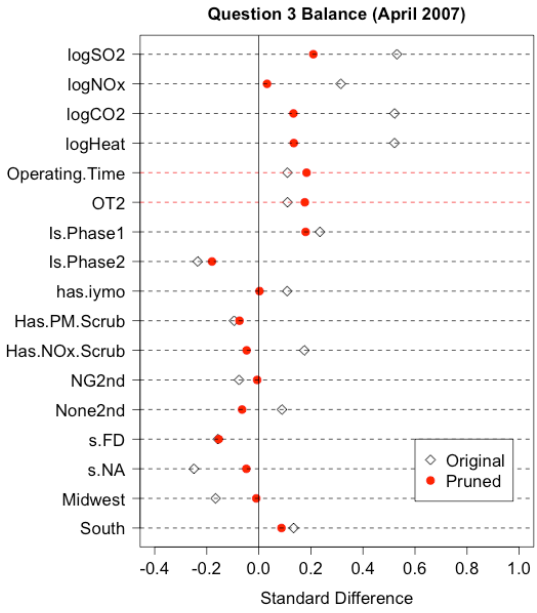
APPENDIX C.4: FUTURE BALANCE ASSESSMENTS



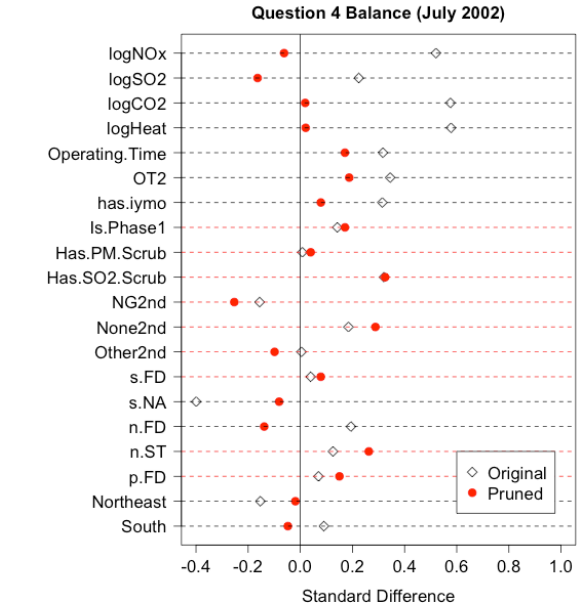
Balance as measured by standardized difference in means between treatment groups for Question 2 in July 2002 (top left), April 2007 (top right), and December 2012 (bottom). Balance also deteriorates by 2012, though less than that of Question 1, and is very strong through April on all covariates.

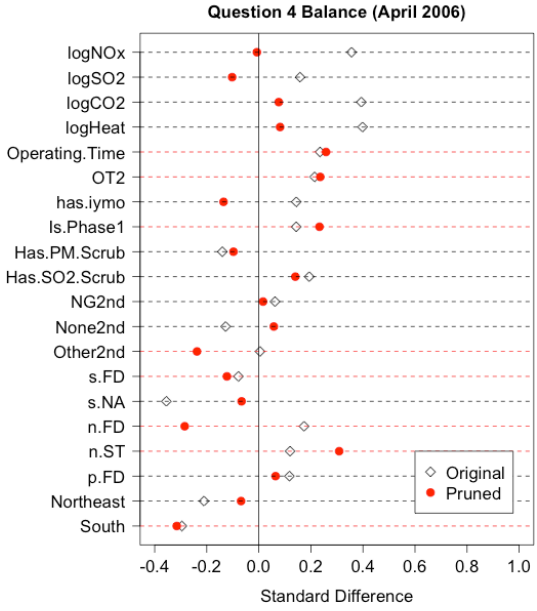
APPENDIX C.4: FUTURE BALANCE ASSESSMENTS



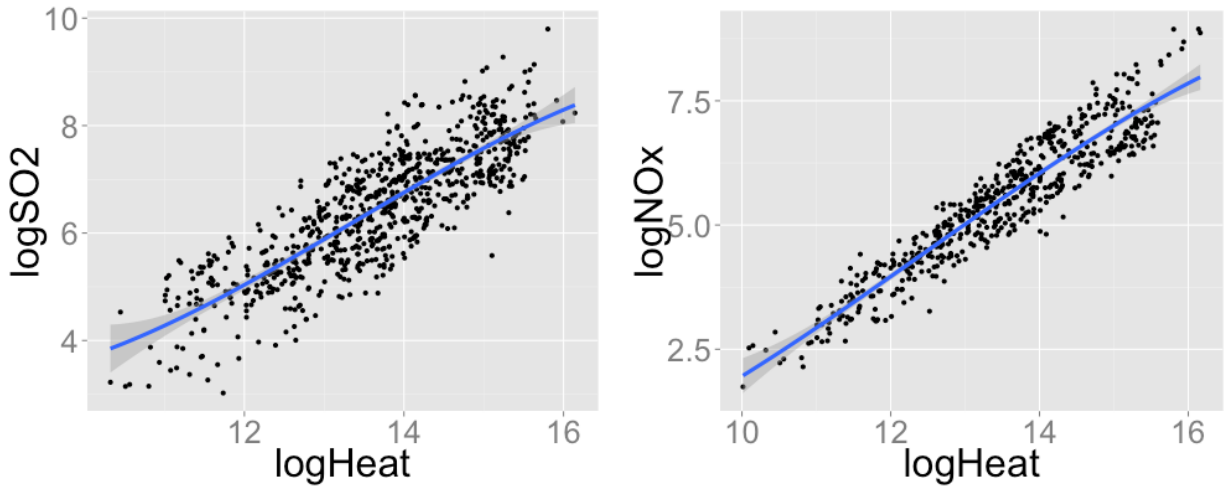


Balance for Question 3 remains strong through 2007. No more observations are taken past this point as the number of treated units yet to receive treatment becomes very low.



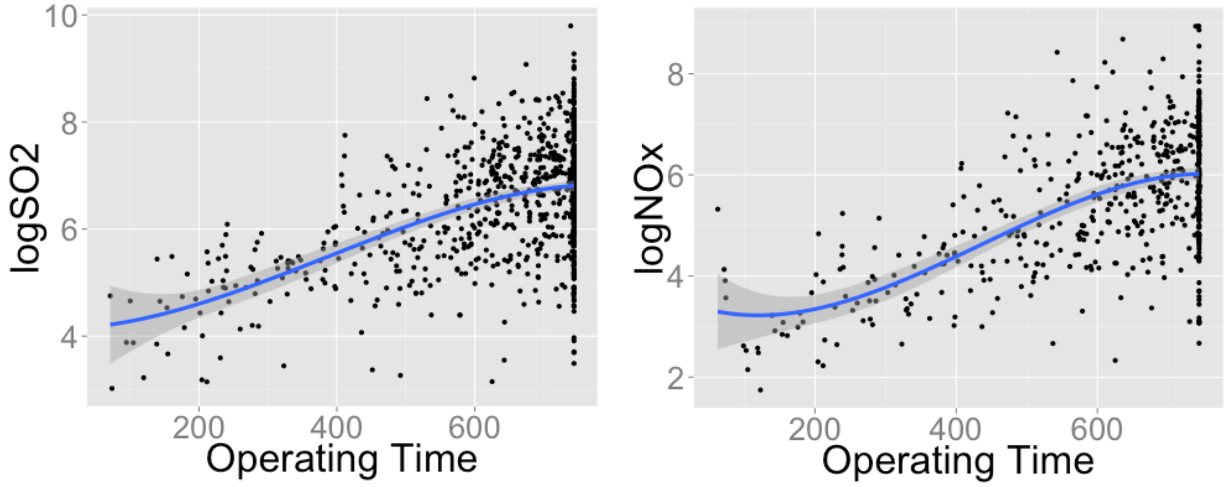


Appendix C.5: Linearity of log Heat with Emissions



LogHeat plotted against logSO2 (left) and logNOx (right) for the month January, 1997. Both are fitted to third degree polynomials and have a clearly linear relationship.

Appendix C.6: Non-linearity of Operating Time with Emissions



Operating Time plotted against logSO2 (left) and logNOx (right) for the month January 1997. The curvature shows that this relationship is clearly non-linear. Notice the high density of points at 744 hours, which is the number of hours in a 31-day month.

7. References

Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behavioral Research. 2011; 46:399–424.

Banzhaf, H. Spencer, Dallas Burtraw, David Evans, and Alan Krupnick. 2006. "Valuation of Natural Resource Improvements in the Adirondacks." Land Economics 82(3): 445 – 64

Burtraw, Dallas, Alan Krupnick, Erin Mansur, David Austin, and Deirdre Farrell. 1998. Cost and Benefi ts of Reducing Air Pollutants Related to Acid Rain. Contemporary Economic Policy 16(4):379 – 400.

Burtraw, Dallas. Cost Savings, Market Performance and Economic Benefits of the U.S. Acid Rain Program. Discussion Paper 98-28-REV. Washington, D.C.: Resources for the Future, March.

Burtraw D, Palmer K. 2004. The SO2 Cap-and-Trade Program in the United States: a "Living legend" of Market Effectiveness, In Choosing Environmental Policy: Comparing Instruments and Outcomes in the United States and Europe. Winston Harrington, Richard

D. Morgenstern and Thomas Sterner, Eds. Resources for the Future: Washington D.C.

Carpenter R. Matching when covariables are normally distributed. Biometrika 1977;64(2):299–307.

CEM, http://gking.harvard.edu/cem/ Iacus, S. M., King, G. and Porro, G. cem: Coarsened exact matching software. 2009

Chan, G, Stavins, R, Stowe, R, Sweeney, R. The SO_2 Allowance Trading System and the Clean Air Act Ammendments of 1990: Reflections on Twenty Years of Policy Innovation. Harvard Environmental Economics Program, Harvard Kennedy School. 2012.

Chapin, F. Experimental designs in sociological research. Harper; New York. 1947.

Cochran WG, Rubin DB. Controlling bias in observational studies: A review. Sankhya: The Indian Journal of Statistics, Series A 1973;35:417–446.

Frangakis, Constantine E., and Donald B. Rubin. "Principal stratification in causal inference." Biometrics 58.1 (2002): 21-29.

Glazerman S, Levy DM, Myers D. Nonexperimental versus experimental estimates of earnings impacts. The Annals of the American Academy of Political and Social Science 2003;589:63–93.

Heckman JJ, Hidehiko H, Todd P. Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. Review of Economic Studies. 1997;64:605–654.

Imbens G.W. Nonparametric estimation of average treatment effects under exogeneity: A review. The Review of Economics and Statistics. 2004;86:4–29.

Keohane, Nathaniel. What did the market buy? Cost savings under the US tradable permit program for sulfur dioxide. Yale Center for Environmental Law and Policy Working Paper. 2003.

King, G. and L. Zeng. The dangers of extreme counterfactuals. Political Analysis. 2006;14(2): 131-159.

King, Gary, and Richard Nielsen. Why Propensity Scores Should Not Be Used For Matching. Copy at http://j. mp/1FQhySn Export BibTex Tagged XML Download Paper 452. 2015.

Iacus, Stefano M, Gary King, and Giuseppe Porro. 2011. Multivariate Matching Methods That Are Monotonic Imbalance Bounding. Journal of the American Statistical Association 106 (493): 345-361. Copy at http://j.mp/lxfJch

Lowe, Wilfrid Kouokam. Over-fitting of Propensity Score Models-does it matter? Methodology. 2013.

Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in Medicine. 2004; 23: 2937–2960.

Ming K, Rosenbaum PR. A note on optimal matching with variable controls using the assignment algorithm. Journal of Computational and Graphical Statistics 2001;10:455–463.

Robins J, Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. Journal of the American Statistical Association 1995;90:122–129.

Rosenbaum P, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983; 70:41-55.

Rosenbaum P.R., Rubin D.B. Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association. 1984;79:516–524.

Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. The American Statistician. 1985; 39:33–38.

Rubin DB. The use of matched sampling and regression adjustment to remove bias in observational studies. Biometrics 1973b;29:185–203.

Rubin, D.B. Bayesian inference for causal effects: the role of randomization. Annals of Statistics. 1978; 6(1): 34-58.

Rubin, D. B. For objective causal inference, design trumps analysis. Annals of Applied Statistics. 2008; 2(3): 808-840.

Schafer, J. L. and Kang, J. Average causal effects from nonrandomized studies: A practical guide and simulated example. Psychological Methods. 2008; 13, 279.

Schmalensee, Richard, and Robert Stavins. The SO2 allowance trading system: The ironic history of a grand policy experiment. No. w18306. National Bureau of Economic Research, 2012.

Stuart, E. A. (2010). "Matching methods for causal inference: A review and a look forward." Statistical Science 25(1): 1-21.

U. S. Energy Information Agency. Online Glossary. 2015. *http://www.eia.gov/tools/glossary*

U. S. Environmental Protection Agency. Acid Rain Program: Instructions for Phase 2 NOx Compliance. 1990; 40 CFR 76.9.

U. S. Environmental Protection Agency. NO_x Reductions under the Acid Rain Program. 2012. *http://www.epa.gov/airmarkets/progsregs-old/arp/nox.html*

U.S. Environmental Protection Agency 2011b. National Emissions Inventory (NEI) Air Pollutant Emissions Trends Data: 1970–2011. www.epa.gov/ttn/chief/trends/index.html.

U. S. Government Accountability Office. Freight Railroads: Updated Information on Rates and Other Industry Trends. 2007; Publication No. GAO-07-1245T.

Vanderweele, T. J. and O. A. Arah. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. Epidemiology. 2006; 22(1): 42-52.

Winston, Clifford. The Success of the Staggers Rail Act of 1980. AEI-Brookings Joint Center for Regulatory Studies, Related Publication. 2006; 05-24.