Methods for Estimating the Health Effects of Exposure to Point Sources of Emissions Using Large-Scale and Diverse Data Sources

Methods for Estimating the Health Effects of Exposure to Point Sources of Emissions Using Large-Scale and Diverse Data Sources

a dissertation presented
by
Kevin Cummiskey
to
The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts
April 2018

Dissertation advisor: Corwin Zigler Kevin Cummiskey

Methods for Estimating the Health Effects of Exposure to Point Sources of Emissions Using Large-Scale and Diverse Data Sources

Abstract

There is a well-documented association between exposure to fine particulate matter ($PM_{2.5}$) and numerous health outcomes, with some evidence suggesting $PM_{2.5}$ originating from coal combustion may have different health impacts. These studies typically estimate exposure to coal-derived $PM_{2.5}$ based on the presence of certain chemical tracers measured in the air near exposed populations. Interpreting such user-defined source profiles requires a certain degree of subjectivity and approximation, and such approaches do not consider the contributions of individual coal power plants. This limits their relevance for informing air quality management interventions that must, ultimately, be implemented at individual sources (e.g., through scrubber installation, closing inefficient plants, etc.). Existing literature that does focus on specific point sources uses computationally expensive models for pollution transport, thus limiting their applicability to only a few power plants or groups of power plants.

In chapter one, we employ a recently-developed, reduced-complexity air quality model to provide the first national study of the association between long-term exposure to emissions from individual coal power

plants and Ischemic Heart Disease hospitalization. The study provides a novel combination of observed data, statistical methods, and tools from environmental engineering. Rooting the approach to causal inference methods to isolate the coal emissions/health relationship represents an important step towards establishing the causal links between emissions and health necessary to drive policy changes.

In chapter two, we provide the first investigation of whether a purely statistical, data-driven approach to source-receptor mapping can reproduce knowledge typically produced by complex chemical transport models. The ability to do so would provide a more computationally nimble approach to estimate S-R relationships in a wider variety of settings. Specifically, we consider daily sulfur dioxide (SO_2) emissions from 385 coal-fired power plants operating in the U.S. in 2005, and estimate a source-receptor mapping to 732 EPA Air Quality System (AQS) monitor locations measuring daily fine particulate matter ($PM_{2.5}$). Results were framed as an "emissions network" – power plants and monitors are nodes and significant associations between their daily time-series define edges in the network – representing an annual pattern in coal emissions transport for 2005. The results of the proposed approach were shown to hold some promise in capturing general patterns of pollution transport and source-specific exposures, but was limited in its ability to recover individual source-receptor links relative to a recently proposed

reduced complexity CTM. Our investigation uncovered several statistical challenges for which we provide initial progress towards addressing, with future refinements holding promise for improving the fidelity of the purely statistical approach.

In chapter three, we explore the value of the statistical, data-driven approach to source-receptor mapping to evaluate how source-receptor relationships vary over time. Specifically, we use the statistical methods to explore seasonal variability (winter, spring, summer, and fall) in coal emissions transport using daily SO_2 emissions from coal-fired power plants operating in the United States from 2005-2010 and daily $PM_{2.5}$ concentrations at air quality monitors. We fit four emissions networks per year (one each season) from 2005-2010 and compared them across seasons and years at various levels of granularity. Our results point to important short-term variability in source-receptor mappings that may not be captured in annual models.

Contents

Listing of figures

List of Tables

With unmeasurable gratitude, I dedicate this work to my family. To my parents, thank you for your unwavering love and commitment. To my children, Andrew and Ashley, you bring such joy to my life, and your contributions to my work transcend the words written on these pages. Most of all, to my wife, Bokyeon, your love and support made this endeavor possible. I love you so much.

Acknowledgments

I am extremely grateful to my dissertation advisor, Cory Zigler, for his mentorship and steadfast commitment to my development. He brings a new and important perspective to air pollution health effects studies, and being his student was a thrilling and challenging experience from which, as a researcher and scientist, I have grown tremendously. I would also like to thank the other members of my dissertation committee, Brent Coull and Chanmin Kim, for the many insightful recommendations that shaped this work. In addition, Chanmin's assistance in getting the early stages of this research going was indispensable. I am very grateful to Christine Choirat for the many data analysis tools supporting my work (and the work of many others). Thank you to Lucas Henneman for providing extensive expertise in atmospheric chemistry and environmental engineering throughout this research; your contributions have greatly improved this work. Thank you to Joel Schwartz and Petros Koutrakis for their critical reviews of this work (Chapters 1 and 2, respectively). I would also like to thank Jelena Follweiler and Paige Williams for all their support, and my PhD cohort for their camaraderie.

Lastly, to the Department of Mathematical Sciences at the United States Military Academy, I am tremendously grateful for the opportunity to continue to serve our nation by educating, training, and inspiring its future military leaders.

1

A Source-Oriented Approach to Coal Power Plant Emissions Health Effects

1.1 Introduction

Long-term exposure to ambient fine particulate matter of less than 2.5 micrometers in diameter ($PM_{2.5}$) has been associated with increased mortality and morbidity related to various cardiovascular and respiratory conditions [9, 21, 29–31]. While the association between overall $PM_{2.5}$ exposure and health outcomes is well-documented, $PM_{2.5}$ varies in chemical composition and there is increasing evidence particles originating from different sources impact

health differently [19, 24, 27, 36]. The National Academy of Sciences identified this area of research as a top priority because it "could result in targeted control strategies that would specifically address these sources having the most significant effects on public health"[26].

Most existing research on the comparative health effects of different sources of $PM_{2.5}$ relies on source apportionment methods to apportion $PM_{2.5}$ mass measured at population locations into broad source categories, based in part on the availability of certain chemical tracers associated with specific types of sources. Despite the value of using source apportionment in health effects studies, such methods are limited by their reliance on broad source categories identified by chemical tracers. Rather, it may be valuable to estimate the health impact associated with exposure to a specific set of point sources, for example, all coal-fired power plants operating in the United States. The ability to evaluate health impacts associated with specific point sources is an important step towards providing policy-relevant information that can inform specific, targeted interventions. Moreover, observational epidemiological studies are particularly challenged by the threat of confounding; areas exhibiting high exposure may share important differences from areas with low exposure. The ubiquity of potential confounding in studies of air pollution warrants the use of statistical methods anchored to causal inference methodology designed to more explicitly address the threat of confounding.

The goal of this paper is to refine and complement the existing evidence of the health burden of coal combustion with a novel analysis approach that uses a reduced complexity air quality model to characterize exposure and combines this with statistical methods for causal inference and confounding adjustment to isolate the association between public health and emissions from a specific set of coal-fired power plants. Specifically, we estimate the association between a United States ZIP code having high exposure to coal combustion

emissions and ischemic heart disease (IHD), mortality from which has been previously linked to long-term exposure to $PM_{2.5}$ [16, 25, 36]. Our study of IHD events offers additional granularity to the study of long-term $PM_{2.5}$ exposure and IHD which has, to date, focused on mortality.

We characterize coal power plant $PM_{2.5}$ exposure with a recently-developed, reduced-complexity air quality model called the Intervention Model for Air Pollution (InMAP, available at http://spatialmodel.com/inmap/).[35] Specifically, we use InMAP to quantify the influence of coal emissions from 783 coal-fired power generating units at over 21,000 ZIP code locations in the Northeast, Industrial Midwest, and Southeast United States. While InMAP is used to characterize coal emissions exposure, ZIP code level measures of total ambient $PM_{2.5}$ mass concentrations come from state-of-the-art predictions from Di et al (2016), derived from a neural-network combining information from monitoring data, land-use regression, remote-sensing satellite data, and GEOS-Chem simulations [8]. We link both measures with health outcomes available among 19,726,981 Medicare beneficiaries living in these regions to isolate the association between exposure to coal power plants and IHD events.

A key component of our analysis approach is the use of propensity score matching, a method for confounding adjustment with advantages over traditional regression models. Grounding the statistical analysis to an explicit causal inference method such as propensity score matching is essential for informing policy interventions, particularly in observational epidemiological studies of air pollution [10]. The matching procedure is specifically designed to identify and mitigate the threat of confounding by framing the investigation as a hypothetical controlled experiment where each ZIP code is regarded as either high-exposed to coal power plant emissions or a control location.

To our knowledge, the combination of a reduced-complexity air quality model with modern statistical methods for causal inference and confounding adjustment deployed in an analysis of Medicare health outcomes and state-of-the-art data fused estimates of $PM_{2.5}$ represents the largest scale study to date of the health impacts of coal power plant pollution emissions.

1.2 Methods

1.2.1 Data Sources and Study Population

We compiled basic background and demographic information from the Center for Medicare and Medicaid Services on 19,726,981 Medicare beneficiaries residing in 21,351 ZIP codes in the Northeast, Industrial Midwest, and Southeast regions of the United States in 2005. These three regions account for most coal power generation in the United States and have been subjected to national regulations intended to limit interstate transfer of air pollution emissions. An IHD event was defined as a hospital admission with a primary discharge diagnosis of ICD-9 410-414, or 429. Population demographic data was augmented with information from the US Census Bureau (year 2000) and the CDC Behavioral Risk Factor Surveillance System. Total annual emissions data for 783 coal-fired generating units was obtained from continuous emissions monitors provided in the EPA's Air Markets Program Data (AMPD). Power plant stack features (height, diameter) were obtained from the EPA's National Emissions Inventory (NEI 2014).

For a secondary analysis, we obtained $PM_{2.5}$ exposure predictions for 2005, from Di et al (2016), derived from a neural-network combining information from monitoring data, land-use regression, remote-sensing satellite data, and GEOS-Chem simulations [8].

Ultimately, the analysis data set contained data on 21,351 U.S. ZIP codes, each having a measure of IHD hospitalization rate in 2005, measures of population demographics, weather, and total $PM_{2.5}$ concentration, and a measure of coal power plant emissions exposure.

1.2.2 Classifying Coal Power Plant Exposure Using InMAP

To create the primary exposure metric, we classified each ZIP code as either a high-exposed location to coal power plant emissions, or a control location, by combining data on power plant emissions with the results from InMAP. InMAP uses output from a widely-used chemical transport model, WRF-Chem, to estimate changes in annual average $PM_{2.5}$ concentrations on a variable spatial grid attributable to annual changes in precursor emissions at user-prescribed locations.

Total annual SO_2 emissions during 2005, geographic coordinates, and stack features for each coal-fired generating unit were input into InMAP. For each location on the variable spatial grid, InMAP estimated the total annual change in $PM_{2.5}$ attributable to a 100% emissions reduction for all coal generating units. These grid estimates were aggregated to obtain ZIP code estimates. While the InMAP output can be interpreted as the annual change in $\mu g/m^3$ of $PM_{2.5}$ concentration, we refer to these estimates as the influence of coal emissions and use these values simply to classify ZIP codes (recall that total $PM_{2.5}$ mass predictions come from [8]). Based upon the distribution of these influences, an appropriate cutoff to classify locations as either high-exposed or control was selected, with this cutoff varied in sensitivity analyses.

1.2.3 Confounding Adjustment with Propensity Score Matching

Isolating the association between high exposure to coal power plant emissions and IHD hospitalizations by comparing high-exposed to

control locations requires adjusting for various population and climatological factors, also referred to as confounders, that differ between locations in both exposure groups.

One common tool for confounding adjustment in such settings is propensity score matching, which, in this context, attempts to match high-exposed locations to control locations that are comparable on the basis of possible confounding factors. Towards this goal, we estimated propensity scores for each ZIP code as the predicted probability of being high-exposed from a logistic regression model that included a broad set of covariates including socioeconomic and demographic variables, smoking rates, weather attributes, and characteristics of the Medicare population. See Appendix A.1 for specific covariates included in the propensity score model. A secondary analysis augments propensity scores to include overall $PM_{2.5}$ mass.

After propensity score estimation, each high-exposed location was matched to a control in the same region and with similar estimated propensity score (see Appendix A.3 for details). After matching, the threat of confounding can be empirically assessed by checking whether covariate distributions are comparable in matched locations, before conducting any analysis of IHD hospitalizations.

1.2.4 Model for IHD Hospitalizations

The outcome of interest is the number of IHD hospitalizations in the Medicare population in 2005 in each ZIP code. Poisson regression was used to estimate the incident rate ratio (IRR) for IHD hospitalization comparing high-exposed to control locations in the propensity score matched data set. All models included the covariates from the propensity score model in the IHD model to adjust for residual confounding remaining after the matching process. To reflect the regional nature of air pollution, a separate Poisson regression model

was fit to the matches for each region.

1.2.5 Sensitivity Analyses

We present three sensitivity analyses. To assess the sensitivity to the exposure cutoff defining high-exposed ZIP codes, we performed the analysis under a range of different cutoffs delineating high coal emissions exposed ZIP codes from controls. To assess sensitivity to the specific propensity score method for confounding adjustment, we analyzed the data by stratifying (instead of matching) locations on propensity score quintile and included an indicator of propensity score quintile in the model for IHD. This method typically results in many more locations in the IHD analysis and helps assess sensitivity of results to the specific locations chosen in the matching process.

Finally, to more directly address the prospect of unmeasured spatial confounding that could arise if there are unmeasured differences between ZIP codes that are located far from one another, an alternative method, Distance Adjusted Propensity Score Matching (DAPSm), was used to construct a matched data set based on estimated propensity scores and geographic proximity of matches (see Appendix A.4) [28].

1.2.6 Secondary Analysis: Adjusting for $PM_{2.5}$

Importantly, results of the primary analysis do not adjust for total $PM_{2.5}$ mass in the propensity score or outcome models. As a consequence, estimates of the association between high coal emissions exposure and IHD include both the impact of elevated coal emissions themselves and any associated increase in overall ambient $PM_{2.5}$ mass.

As a secondary analysis, we augment the propensity score model to include the annual average ambient $PM_{2.5}$ concentration at each ZIP code. This analysis estimates the increased association between coal

emissions exposure and IHD among ZIP codes that have been matched to have similar annual average total $PM_{2.5}$ mass. Without restrictive assumptions, this analysis is difficult to interpret, as it amounts to adjusting for a variable "on the causal pathway" between coal emissions influence and IHD hospitalization; some areas may have total $PM_{2.5}$ mass that is largely a consequence of the coal influence exposure. Under the assumption that annual average total $PM_{2.5}$ mass is not measurably affected by the high coal exposure metric (after adjusting for other variables in the propensity score), estimates in this secondary analysis address whether elevated coal power plant emissions are associated with IHD among areas with the same annual average total $PM_{2.5}$ mass. Such a result would indicate differential health impact of coal-derived $PM_{2.5}$ relative to other sources that might make up total $PM_{2.5}$ mass; even for ZIP codes with the same annual ambient $PM_{2.5}$ concentration, those with high coal emissions exposure exhibit different health impacts. Appropriateness of this restrictive assumption relates to whether annual average total $PM_{2.5}$ mass, which is generally a consequence of many sources (including, but not limited to power plants), is not measurably affected by the high coal exposure metric, after adjusting for other variables in the propensity score. Such may be the case, for example, in areas where $PM_{2.5}$ derived from coal power plant emissions represents a small proportion of overall ambient average $PM_{2.5}$ mass relative to other local pollution sources, or where coal emissions exposure determines the chemical composition of the ambient average but not the total mass.

1.3 Results

The study population experienced 537,369 IHD events in over 18 million person-years, a rate of 2856 events per 100,000 person-years.

These rates were 2657 per 100,000 in the Northeast, 3025 per 100,000 in the Industrial Midwest, and 2922 per 100,000 in the Southeast. Table 1.3.1 contains descriptive statistics of all covariates.

1.3.1 Exposure Classification with InMAP

Figure 1.3.1 (top panel) depicts the InMAP estimates of the influence of coal emissions on each ZIP code. We selected a cutoff value of $4.0\,\mu g/m^3$ in the coal emissions influence to classify locations as either high-exposed or controls because this value is between the two modes in the distribution (Figure 1.3.1: top). Using this cut-off, there were 6,625 high-exposed and 14,726 control locations. In terms of coal emissions influence, the high-exposed locations were in the 80th percentile of all U.S. ZIP codes and in the 69th percentile of ZIP codes included in the study area.

1.3.2 Propensity Score Matching Results

High-exposed locations were matched to controls in the same region with similar propensity scores. Figure 1.3.1 (bottom) depicts the locations of ZIP codes in the propensity score matched data. The matched data set consisted of 3,720 high-exposed and 3,720 controls, which is 35% of the original 21,351 locations. There were 190,339 IHD events in over 6 million person-years in the matched data set. IHD rates per 100,000 person-years in the high-exposed (controls) were 3170 (3162) in the Industrial Midwest, 2843 (2589) in the Northeast, and 2853 (2720) in the Southeast. Standard diagnostics for propensity score matching were performed (see Appendix A.3) to ensure the comparability of the high-exposed and control locations, thus mitigating the threat of confounding from these variables.

	Industrial Midwest		Northeast		Southeast	
	Controls	High-exposed	Controls	High-exposed	Controls	High-exposed
Number of ZIPs	5576	1347	3852	3861	5298	1417
IHD events	25.44 (39.34)	22.03 (34.63)	19.62 (28.53)	27.88 (44.21)	27.16 (38.59)	27.35 (31.85)
Person-years	845.66 (1258.85)	707.94 (1086.57)	795.01 (1102.47)	992.81 (1481.91)	912.10 (1267.69)	999.36 (1132.52)
$PM_{2.5}$	13.96 (2.06)	15.10 (1.34)	11.34 (1.79)	13.98 (1.45)	12.97 (2.20)	14.53 (1.57)
logPop	8.07 (1.66)	8.04 (1.62)	8.27 (1.49)	8.44 (1.73)	8.63 (1.53)	9.00 (1.40)
PctUrban	0.36 (0.42)	0.33 (0.40)	0.44 (0.44)	0.55 (0.44)	0.43 (0.43)	0.47 (0.42)
PctBlack	0.05 (0.13)	0.04 (0.11)	0.04 (0.10)	0.12 (0.19)	0.22 (0.24)	0.22 (0.23)
PctHisp	0.02 (0.06)	0.01 (0.02)	0.04 (0.08)	0.04 (0.08)	0.05 (0.10)	0.03 (0.05)
PctHighSchool	0.39 (0.10)	0.41 (0.10)	0.33 (0.10)	0.35 (0.13)	0.33 (0.08)	0.30 (0.09)
MedianHHInc	40.19 (14.85)	36.42 (11.51)	46.83 (19.34)	47.92 (20.72)	33.64 (11.64)	40.15 (13.16)
PctPoor	0.12 (0.10)	0.13 (0.09)	0.10 (0.08)	0.10 (0.08)	0.17 (0.10)	0.13 (0.08)
PctFemale	0.50 (0.04)	0.51 (0.04)	0.51 (0.03)	0.51 (0.04)	0.51 (0.04)	0.51 (0.04)
PctOccupied	0.88 (0.13)	0.89 (0.10)	0.84 (0.17)	0.90 (0.11)	0.87 (0.10)	0.89 (0.09)
PctMovedIn5	0.40 (0.11)	0.41 (0.11)	0.39 (0.10)	0.39 (0.13)	0.46 (0.11)	0.47 (0.12)
MedianHValue	92.34 (55.44)	81.91 (33.19)	137.63 (103.43)	138.67 (109.63)	85.60 (54.02)	106.03 (55.61)
smokerate2000	0.28 (0.03)	0.30 (0.03)	0.25 (0.03)	0.25 (0.04)	0.28 (0.03)	0.26 (0.03)
avtmpf	283.58 (2.04)	285.40 (1.08)	281.61 (1.72)	284.80 (1.69)	292.46 (3.45)	289.34 (1.68)
avrelh	0.01 (0.00)	0.01 (0.00)	0.01 (0.00)	0.01 (0.00)	0.01 (0.00)	0.01 (0.00)
mean_age	74.91 (1.31)	74.54 (1.33)	75.15 (1.27)	75.21 (1.48)	74.72 (1.29)	74.59 (1.13)
Female_rate	0.56 (0.05)	0.55 (0.06)	0.55 (0.06)	0.56 (0.06)	0.56 (0.05)	0.58 (0.05)
White_rate	0.95 (0.14)	0.96 (0.11)	0.94 (0.13)	0.86 (0.20)	0.80 (0.21)	0.80 (0.21)

Table 1.3.1: Mean (standard deviation) of ZIP code-level variables in the raw dataset. See Table A.1.1 in the Appendix for variable definitions.

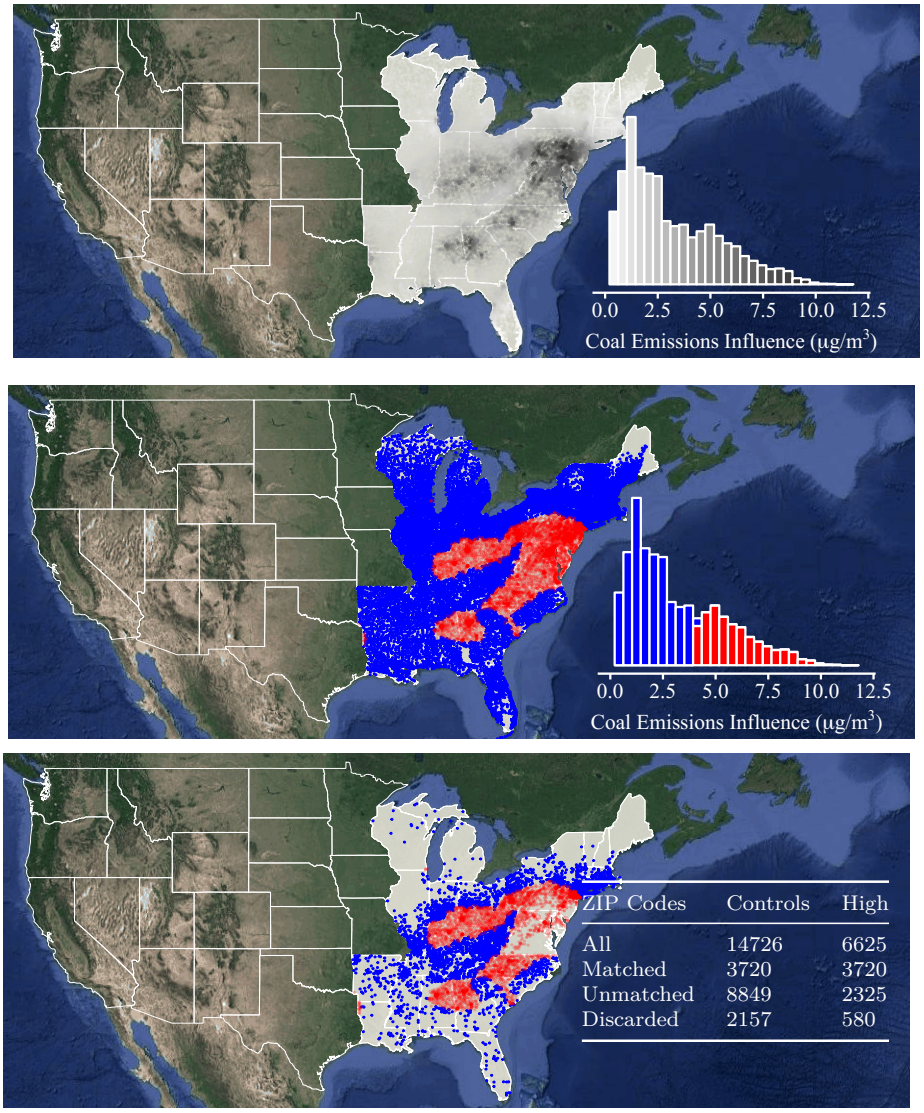


Figure 1.3.1: Top - InMAP estimates of the influence of emissions from 783 coal-fired generating units (2005) on ZIP codes in the eastern U.S.. Middle - high-exposed (red) and control (blue) locations in the full data set. Bottom - high-exposed (red) and control (blue) locations in the propensity score matched data set.

1.3.3 Analysis of IHD in the Matched Data

Using Poisson regression fit to the matched data in each region, the estimated IRRs for annual IHD hospitalizations comparing high-exposed locations to controls were 1.016 (95% CI: 0.998, 1.035) in the Industrial Midwest, 1.077 (95% CI: 1.060, 1.094) in the Northeast, and 1.058 (95% CI: 1.042, 1.075) in the Southeast. This indicates a significant increase in the rate of IHD hospitalizations in high-exposed locations in the Northeast and Southeast after adjusting for covariates. There was no significant increase in the Industrial Midwest.

1.3.4 Sensitivity to Propensity Score Procedure

Two sensitivity analyses, one stratifying on estimated propensity scores instead of matching, and another using DAPS matching, resulted in similar IRR estimates as the primary analysis. The stratified analysis included 18,614 ZIP code locations, 87% of the full data set, in the IHD model. For further details and results on DAPS matching, see Appendix A.4.

1.3.5 Sensitivity to the Definition of High vs. Low Exposure

Figure 1.3.2 (top) shows estimated IRRs for the exposure/IHD relationship for various cutoffs ranging from 3.0 to 5.0 $\mu g/m^3$ for delineating high coal emissions exposed ZIP codes from controls. This range was selected because these cutoff values are between the two modes of the coal emissions influence distribution. In the Northeast and Southeast, positive associations were observed at every cutoff. In the Industrial Midwest, lower cutoffs resulted in negative associations. Larger cutoffs in that region resulted in either small or no effects with the point estimates becoming more positive with higher cutoffs.

Figure 1.3.2 (bottom) shows the mean coal emissions influence in the high-exposed and controls, by region, for the range of cutoffs. It depicts the contrasts, in terms of the continuous coal emissions influence, between the two exposure groups. The high-exposed locations in the Northeast have the highest coal emissions influence and the contrast between the high-exposed and controls is the largest among the regions. In the Industrial Midwest, the high-exposed areas have a comparatively low coal emissions influence and the contrasts are the smallest. Note the IRR estimates increase with larger contrasts in the Industrial Midwest.

1.3.6 Secondary Analysis Results: Adjusting for $PM_{2.5}$

Analysis	Industrial Midwest	Northeast	Southeast
Primary	1.02 (1.00, 1.04)	1.08 (1.06, 1.09)	1.06 (1.04, 1.08)
Secondary	1.01 (0.99, 1.03)	1.02 (1.01, 1.04)	1.05 (1.03, 1.07)

Table 1.3.2: Estimated IRRs associated with IHD hospitalizations in the primary ($PM_{2.5}$ unadjusted) and secondary ($PM_{2.5}$ adjusted) analyses.

Table 1.3.2 contains estimated IRRs for IHD hospitalizations comparing high-exposed to control locations in both the primary ($PM_{2.5}$ unadjusted) and secondary ($PM_{2.5}$ adjusted) analyses, showing that the secondary analysis yields attenuated IRRs in each region. In interpreting these results, it is important to consider the relationship between $PM_{2.5}$ and the coal influence exposure in each region.

In the Southeast, total $PM_{2.5}$ mass concentrations in the propensity score matched data were only slightly higher (see Appendix A.5) in the high-exposed locations than the controls suggesting that

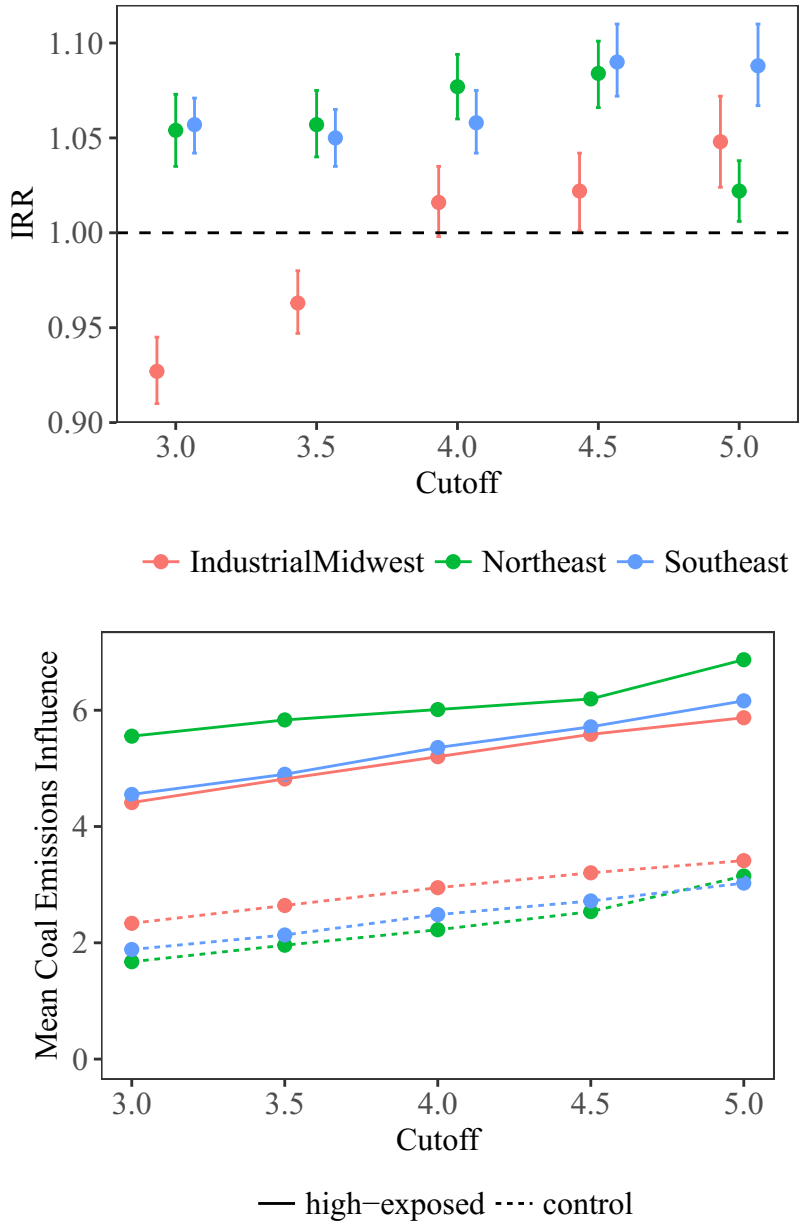


Figure 1.3.2: Top - estimated IRRs associated with IHD hospitalizations comparing high-exposed locations to controls for various cutoffs. Bottom - mean coal emissions influence by exposure group at the cutoffs.

high-exposed and control areas are differentiated by the amount of coal-derived $PM_{2.5}$, even though total annual average $PM_{2.5}$ mass is similar. This provides some support for interpreting results from the secondary analysis as effects of evelaved coal-derived emissions in areas with similar total $PM_{2.5}$ mass. In other words, in the Southeast, elevated coal emissions influence may be associated with increased IHD above and beyond what might be due to associations with overall $PM_{2.5}$ mass, possibly due to other characteristics of the $PM_{2.5}$. The lack of an association in the Industrial Midwest precludes a similar interpretation of the secondary analysis in that region.

In the Northeast, however, there was a more pronounced association between high-exposure and total $PM_{2.5}$ concentrations (see Appendix A.5), reflecting the prominent influence of coal emissions in determining overall $PM_{2.5}$ mass in this region. As a result, the IRR effect estimate in the secondary analysis was particularly attenuated (primary IRR: 1.08, secondary IRR: 1.02). Given the lack of independence between exposure and total $PM_{2.5}$ concentrations, we cannot interpret the secondary analysis as being informative about the relative toxicity of particles from coal emissions.

1.4 Discussion

We have deployed new computational and statistical tools to investigate the health impact of exposure to emissions originating from coal power plants. For this analysis, we considered populations in 21,351 ZIP code locations in the Northeast, Industrial Midwest, and Southeast regions of the United States and their exposure to the emissions from 783 coal-fired generating units in 2005. Our results showed an increased rate in IHD hospitalizations in the Northeast (IRR: 1.08, 95% CI: 1.06, 1.09) and the Southeast (IRR: 1.06, 95% CI: 1.04, 1.08) for high-exposed locations. Importantly, this is an

impact on annual hospitalization rates, whereas most previous studies of IHD admissions and air pollution have only addressed short-term exposure and daily rates. No significant association was found in the Industrial Midwest (IRR: 1.02, 95% CI: 1.00, 1.04). While there are no directly comparable studies, our results are broadly consistent with existing work. The source-apportinment analysis in Thurston et al (2016) estimated a hazard ratio for IHD mortality of 1.05 (95% CI: 1.02, 1.08) per $10\,\mu g/m^3$ increase in coal combustion $PM_{2.5}$.[36]

The lack of an association in the Industrial Midwest may be the result of the relative spatial homogeneity in coal power plant exposure in that region, reducing the exposure gradient used for comparison and the ability to detect health effects. The Industrial Midwest had the smallest difference in mean coal emissions influence between the high-exposed and control locations. In addition, the highest and least exposed ZIP codes were underrepresented there. In terms of coal emissions influence, only 1.4% (5.0%) of Industrial Midwest ZIP codes were among the highest (lowest) 10% of ZIP code exposure levels in the propensity score matched data (for comparison, the analogous values in the NE were 17% (16.5%)).

At lower cutoffs, negative associations estimated in the Industrial Midwest can be attributed to controls in the matched data being densely located in two areas with high IHD rates derived from other causes. Specifically, coal mining in southeast Kentucky and southern West Virginia and steel production near Lake Erie result in locally high IHD rates. [14, 20] At higher cutoffs, these areas, which are consistently characterized as control locations, are not included in the matched data due to more suitable controls being located elsewhere.

Losses in predictive accuracy are a potential issue for reduced-complexity models like InMAP. It is important to reiterate that our strategy does not use outputs from InMAP directly in the health-outcomes analysis, it merely uses the output to characterize

ZIP codes as "high-exposed" or controls. Thus, our design provides robustness to losses in predictive accuracy from the use of reduced-complexity models as we require only accurate estimates of the relative ranking of coal emissions influence among ZIP codes.

While the propensity score analysis employed here is designed to provide a more rigorous account of the threat of confounding, it comes with important limitations. The method relies on a binary exposure classification of an inherently continuous exposure. Dichotomizing the continuous exposure permits more targeted adjustment for confounding, including empirical assessment of the extent to which covariates are "balanced" between high and low exposed groups. Nonetheless, dichotomization of a continuous exposure results in a loss of information and the inability to characterize a complete exposure-response relationship. It can also be viewed as a type of classical measurement error. This limitation was evident in the Industrial Midwest, where exposures between the two groups were closer on a continuous scale than the other regions.

Another inherent feature of the investigation is the presence of ambient $PM_{2.5}$ mass as an intermediate variable "on the causal pathway" between coal emissions influence and IHD hospitalizations. Our primary analysis did not provide any adjustment for ambient $PM_{2.5}$ mass, permitting IRR estimates to include any effect of high coal emissions influence that is due to the resulting increase in total $PM_{2.5}$ mass. In a secondary analysis that adjusts for $PM_{2.5}$ in the propensity score and outcome models, we evaluated the possibility of interpreting results as differential effects of coal emissions influence among areas with the same overall $PM_{2.5}$ mass. We provided some evidence that the assumptions required for such interpretation are reasonable in the Industrial Midwest and Southeast, but more targeted analysis of this point, possibly with methods emanating from the literature on mediation analysis or principal stratification, are

warranted. [13, 32]

2

A Data-Driven Approach to Source-Receptor Mapping of Power Plant Emissions to Exposed Populations.

2.1 Introduction

Source-receptor mapping is the process of determining which populations are impacted by the air pollution derived from the emissions from a set of pollution sources and is a critical component of high-stakes, public-policy interventions to improve air quality. The central challenge is long-range pollution transport – emissions from,

for example, power plant smokestacks, chemically react in the atmosphere and move hundreds of kilometers before reaching exposed populations. While rich-observed data resources on emissions and ambient air quality are increasingly available, deterministic chemical transport models (CTMs) remain the primary tool for source-receptor mapping, providing mathematical representations of pollution transport processes. However, their computational burden limits the total number of times the procedure can be practicably run in any given investigation. Typically, implementation considers only a few sources or groups of sources, or a limited number of scenarios representing specific time frames or specific emissions scenarios can be accommodated [5, 22, 23]. As deterministic physical/chemical models, they also confront limitations in terms of uncertainty quantification and validation [2, 4, 17, 37, 40].

A more flexible and computationally nimble approach to source-receptor mapping could permit consideration of a wider variety of scenarios relevant to policy evaluation, and so-called "reduced-complexity" CTMs are generating interest for such a purpose [35]. In a similar spirit, this paper considers methods to leverage increasingly rich-observed data on sources and receptors towards exploring an entirely data-driven approach: using only observed emissions and monitored air quality measurements, we employ a rigorous statistical modeling approach to quantify associations between emissions from specific point sources and air quality at a set of population locations. Specifically, we consider daily sulfur dioxide (SO_2) emissions from 385 coal-fired power plants operating in the U.S. in 2005, and estimate a source-receptor mapping to 732 EPA Air Quality System (AQS) monitor locations measuring daily fine particulate matter ($PM_{2.5}$). The purely statistical approach, anchored only to observed data, offers a promise of computational scalability and flexibility relative to CTMs. In addition, the simplicity

of the approach has the potential to aid in transparency and communicability in public policy assessments relative to highly specialized and technical CTMs. However, the statistical challenges associated with such an approach and the reliability for capturing true source-receptor relationships have never been seriously addressed. We aim to document both as an important first step towards statistical and observation-based approaches to problems that have historically been only addressed with deterministic physical-chemical models.

Towards the goal of estimating and then validating the statistical approach, we frame a source-receptor matrix as an "emissions network", a directed, unweighted network with two types of nodes (power plants and monitors). To specify the presence or absence of an edge between any monitor/power plant pair in the network, we fit a generalized additive model predicting the daily time-series of $PM_{2.5}$ for the monitoring location from the daily SO_2 emissions from the power plant, adjusting for seasonality, day of the week, and weather. An edge in the network is defined as a power plant/monitor pair with a significant association between daily air quality at the monitor and recent emissions from the power plant.

One statistical challenge in detecting edges in this manner is the isolation of regional pollution patterns emanating from power plants from those related more to local sources of pollution such as traffic, residential heating, and local industry. To address this challenge, we pre-process the daily pollution time series data with a spatial wavelet decomposition following [1]. The decomposition parses the daily $PM_{2.5}$ time series into components that vary at high spatial frequency (presumably corresponding to variations in local sources such as traffic) and those that vary at a low spatial frequency (presumably corresponding to regional pollution sources). We confine estimation of the source-receptor network to the low spatial frequency component, as the secondary formation of $PM_{2.5}$ originating from power plants is

expected to vary at a more regional scale. In this sense, the spatial wavelet decomposition can be viewed as filtering the high-frequency variation from local sources, which is particularly important in light of the positioning of many monitors in the vicinity of local pollution sources that account for most of the variability in $PM_{2.5}$.

A second statistical challenge is the occurrence of extreme $PM_{2.5}$ events. These events affect large geographic areas, typically for several days, with $PM_{2.5}$ concentrations reaching levels several times higher than normal. The air pollution comes from a variety of manmade (traffic, power generation, etc) and natural sources (forest fires, wind-blown soil, etc), and meteorological conditions during these events facilitate its suspension in the air as fine particulate matter. Given the relatively small day-to-day variability in $PM_{2.5}$ on normal days, extreme $PM_{2.5}$ events exert undue influence on models relating daily emissions and $PM_{2.5}$ time-series. Specifically, not accounting for these days in the model would result in many extraneous edges between affected areas and any power plants simultaneously increasing their emissions, regardless of proximity to the affected areas. The spatial wavelet decomposition is not effective in filtering these events because of the wide geographic area affected. Thus, we adjust for temperature and relative humidity, two important meteorological factors in extreme $PM_{2.5}$ events, in our models.

To assess the reliability of the approach, we evaluate our emissions network across several domains. First, we assess its adherence to established regional patterns in long-range pollution transport. Second, we use the emissions network to rank monitor locations total exposure to coal emissions and compare them to output from a recently developed, reduced-complexity air quality model called Intervention Model for Air Pollution (InMAP) [35]. Lastly, we investigate correspondence between individual edges in the network and output from InMAP.

2.2 Methods

2.2.1 Data

We obtained daily SO_2 emissions data measured by continuous monitoring systems for 385 coal-fired power plants operating in the United States from January 1, 2005 to December 31, 2005 [38]. SO_2 was selected because it is a byproduct of coal combustion and a precursor in the formation of $PM_{2.5}$.

We obtained $PM_{2.5}$ concentrations from 651 AQS monitors operating in the eastern United States during the same time period. Of these monitors, 82 measured the concentration daily and 569 measured the concentration every third day. For those monitors measuring every third day, we imputed the missing values using linear interpolation. We restricted our analysis to monitors in the eastern United States because the majority of coal combustion for power generation occurs in this region. The eastern United States is divided into three regions: Industrial Midwest (IMW), Northeast (NE), and Southeast (SE). Figure 2.2.1 depicts the geographic regions, coal power plants, and monitors.

2.2.2 Spatial Decomposition of $PM_{2.5}$

We used the 2-dimensional wavelet decomposition of Antonelli et al (2017) to decompose the $PM_{2.5}$ spatial surface into high and low spatial frequency components. Seven levels of wavelet functions were estimated as in [1] and the cutoff between high (levels 6-7) and low (levels 1-5) frequency levels determined by visual inspection. Each day was decomposed independently of the other days. For the remainder of this paper, $PM_{2.5}$ refers to the low frequency component of the total $PM_{2.5}$ obtained from the decomposition, meant to

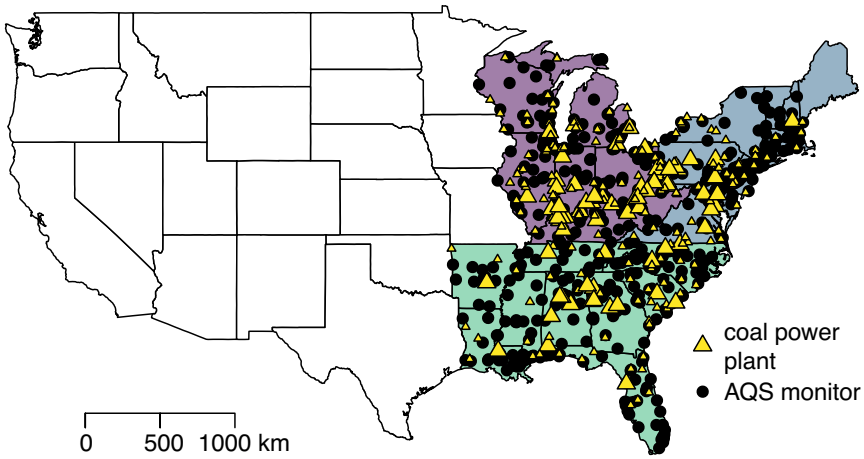


Figure 2.2.1: Map of geographic regions, coal-fired power plants, and eastern AQS monitors in 2005. Large triangles: highest 20% of power plants in terms of average daily SO_2 emissions.

represent the regional pollution more sensitive to secondary $PM_{2.5}$ from regional sources.

2.2.3 Statistical Models for Emissions and Air Quality

Let PM_t be the low frequency component of the total $PM_{2.5}$ on day t at a monitor. Let $SO2_t$ be the sulfur dioxide emissions reported from a power plant on day t (note: for simplicity, subscripts for each power plant and monitor are excluded from $SO2_t$ and PM_t). For each power plant/monitor pair within 1000km, we fit the following generalized additive model:

$$\log(PM_t) = a_0 + \sum_{l=0}^{3} \beta_l(SO2_{t-l}) + f(t) + \mathbf{W}_t + \mathbf{g}(temp_t, humid_t) + \varepsilon_t \quad t = 1, \ldots, 365 \tag{2.1}$$

where $l = 0, \ldots, 3$ is the lag in days between emissions and $PM_{2.5}$, $f(t)$ is a smooth function of time (cubic regression splines with eight knots),

and W_t a vector of indicators for the day of the week. $\mathbf{g}(\cdot)$ are smooth functions (cubic regression splines with 5 degrees of freedom) of the temperature, humidity, and an interaction between the two on day t. The strategy for estimating the source-receptor network fits 168,084 models of this type, one for each power plant/monitor pair. As each such model is estimated independently, computation distributed in parallel across 150 cores required only 20 minutes of computation time.

The smooth functions of time in the model isolates the short-term association between emissions and $PM_{2.5}$. This term is necessary because seasonal trends in emissions and air quality typically result in correlations between most power plants and monitors, regardless of their physical proximity. During warm weather periods, power plants increase output in response to increased energy demands, resulting in higher emissions. At the same time, warmer temperatures enhance the conversion of primary pollutants to $PM_{2.5}$ in the atmosphere, resulting in higher concentrations of $PM_{2.5}$ even for the same level of primary pollutants. Using a smooth function of time to adjust for seasonal trends is a common approach in studies of short-term health effects of $PM_{2.5}$ [7, 11, 18].

The smooth functions of temperature and humidity are included to adjust for extreme $PM_{2.5}$ events, which occur when air pollution is present in certain meteorological conditions, of which temperature and humidity are an important component. These events are typically short-duration, lasting as short as a few days in many cases. Therefore, the smooth function of time is not effective in removing the influence of these extreme events on the models.

In Equation 3.1, we estimate the association between each of several lag days in emissions and a particular day's $PM_{2.5}$. This reflects the complicated nature of long-range pollution transport, as changing weather patterns and other factors typically make emissions from

several days contribute to air quality on a given day. While regression estimates of the coefficients of individual lags will be highly unstable due to multicollinearity, we focus on the overall association of several days by summing the β coefficients. This method is also commonly used in short-term health effects studies of air pollution [33].

For each power plant/monitor pair, the model is used to compute a p-value testing the null hypothesis of no association between emissions from that power plant and $PM_{2.5}$ at that monitor. We adjust these p-values for multiple comparisons using the Benjamini-Hochberg procedure. Pairs with positive overall effects and adjusted p-values less than 0.05 are regarded as edges in the estimated emissions network.

2.2.4 Validation against long-range pollution transport patterns

Descriptive statistics of the network are provided and edge patterns are compared to known long-range pollution transport patterns in the United States. This level of validation is largely heuristic, designed primarily to give a rough sense of whether the procedure corresponds to general expectations.

2.2.5 Validation of relative monitor exposures to coal emissions

For a more refined validation, we investigate the concordance of the emissions network to InMAP in terms of the relative exposures to coal emissions at monitor locations. In addition, we perform a similar comparison with annual average sulfate, a byproduct of coal emissions.

Emissions network exposure

For each edge in the network connecting a power plant to a monitor, we define a measure of the power plant's influence on the monitor as a function of the geographic distance between the plant/monitor pair and average daily emissions at the plant. Specifically, the influence of

power plant j on monitor i is:

$$x_{ij} = \frac{1}{distance_{ij}} \times avgemissions_j \times I(edge_{ij} = 1) \quad j \in \mathbb{P}, i \in \mathbb{M} \tag{2.2}$$

where $\mathbb{P}$ and $\mathbb{M}$ are the sets of power plants and monitors, respectively, distance_{ij} is the geographic distance between j and i, $avgemissions_j$ is the average daily SO_2 emissions from j, and $I(edge_{ij} = 1)$ is 1(0) if there is (not) an edge in the emissions network between j and i. After calculating x_{ij} for each (i, j) pair representing an edge in the network, we obtain a single measure for each monitor by summing over all power plants. That is, to obtain a monitor-specific quantity of coal emissions exposure, we calculate $x_i = \sum_{j \in \mathbb{P}} x_{ij}$, where $x_{ij} = 0$ for (i, j) pairs that do not have an edge in the network. This quantity is not intended to be interpretable in an absolute sense, but to provide a relative quantity for validating general patterns of how a given location is impacted by pollution transport from power plants, as derived from significant edges in the network.

InMAP exposure

We compare the quantity x_i to analogous measures of coal-emissions exposures obtained from the reduced complexity air quality model InMAP. InMAP uses preprocessed physical and chemical information from a chemical transport model to estimate the annual change in $PM_{2.5}$ ($\mu g/m^3$) at tens of thousands of U.S. locations on a variable spatial grid attributable to user-specified changes in emissions. It is computationally fast, allowing this task to be performed individually for a large number of sources. Using InMAP, we estimated the annual change in $PM_{2.5}$ at each location on the grid associated with a 100% reduction in the emissions from each power plant. This task was performed individually for each of the power plants. The resulting grid-level outputs were aggregated to obtain ZIP code level values by

performing a spatial overlay of over ZIP code shapefiles. Monitors were assigned the value of their ZIP code. The result is an InMAP-derived analog to x_{ij}, denoted with x_{ij}^{InMAP}, designed to quantify the amount of $PM_{2.5}$ at a given monitor i attributable to SO_2 emissions from power plant j. The sum over all power plants is calculated for each monitor location (x_i^{InMAP}), and compared against the analogously-calculated x_i.

Sulfate exposure

Sulfate is a byproduct of coal combustion and is measured by a small subset of AQS monitors measuring $PM_{2.5}$. For these monitors, we compare x_i to the annual average sulfate, $x_i^{sulfate}$.

2.2.6 Validation of edge selection

Finally, we investigate the correspondence of estimated edges between individual sources and receptors in the network and InMAP output. Note that InMAP does not produce "edges" directly, rather, it provides a continuously-scaled measure of the relationship between sources and receptors. To investigate correspondence with the estimated edges in the network, we use InMAP output to first classify power plant/monitor pairs as "High InMAP pairs" or "Low InMAP pairs" based on a percentile cutoff in the distribution of InMAP output. Then, we compare the percent of estimated network edges among power plant/monitor pairs classified as "high InMAP pairs" to the percent of estimated network edges among power plant/monitor pairs classified as "low InMAP pairs." Correspondence between the estimated network and InMAP would be indicated by a comparatively high percent of estimated edges among High InMAP pairs. A variety of cutoffs for defining high/low InMAP pairs were considered to evaluate sensitivity.

2.3 Results

2.3.1 Descriptive statistics

Figure 2.3.1 shows the total daily SO_2 emissions from coal-fired power plants in 2005 and the average daily $PM_{2.5}$ by region.

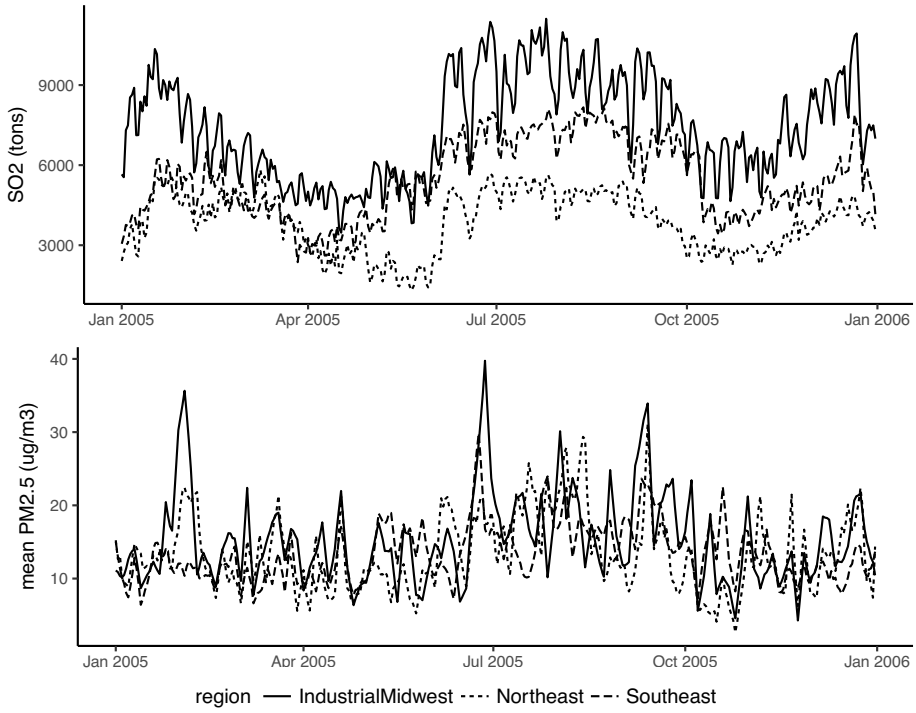


Figure 2.3.1: Total sulfur dioxide (SO_2) emissions in tons from coal power plants and average fine particulate matter ($PM_{2.5}$) in $\mu g/m^3$ by region in 2005.

Table 2.3.1 is a summary of power plant and monitor characteristics by region. We considered 168,084 power plant/monitor pairs and corresponding models of the form (3.1). Based on estimates from the models, 25,221 pairs (22.3%) resulted in edges in the emissions network.

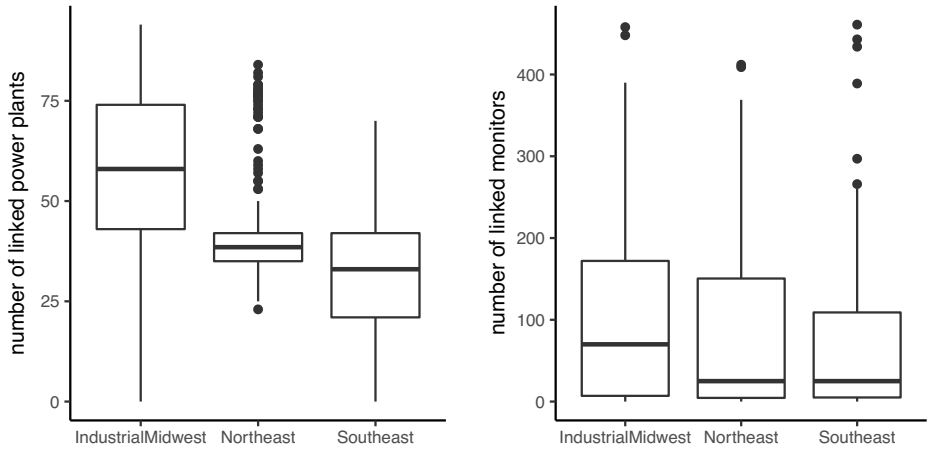
	Industrial Midwest	Northeast	Southeast
power plants	125	63	73
avg daily SO2 emissions (tons)	59.4	61.3	74.7
avg stdev (tons)	43.3	37.3	46.5
median degree	70	25	25
unlinked power plants	8	5	7
monitors	223	180	248
avg daily $PM_{2.5}$ ($\mu g/m^3$)	14.4	13.4	13.7
avg stdev in daily $PM_{2.5}$ ($\mu g/m^3$)	5.2	4.9	4.0
median degree	58	39	33
unlinked monitors	1	0	6

Table 2.3.1: Summary of power plant and monitor characteristics by region.

2.3.2 Power plant and monitor connectivity

Figure 2.3.2 depicts node degree (number of connected edges) distributions for monitors (a) and power plants (b). Five monitors did not link to any power plants because of insufficient measurements. These monitors made $PM_{2.5}$ measurements on less than 30 days each during the year and were removed from the remainder of the analyses.

The power plant degree distribution is highly skewed, most power plants link to a few or no monitors. Many factors influence whether power plants form edges with monitors. For example, power plants operating continuously throughout the year were more likely to be high degree than those that operate intermittently. Appendix B.1 investigates other factors associated with power plant connectivity in the network, indicating the importance of factors such as operating frequency, average emissions, and emissions variability.



(a) Monitor degree distribution (b) Power plant degree distribution

Figure 2.3.2: Histogram of degree distributions for monitors and power plants.

2.3.3 Validation against long-range pollution transport patterns

Figure 2.3.3 depicts the distance and direction from power plants to monitors by power plant region for various aspects of the estimated emissions network. Row 1 depicts the number of power plant/monitor pairs available in each region, regardless of whether they are connected with an edge, and simply describes where the monitors are located in relation to the power plants. Row 2 depicts the distance and direction of estimated edges and describes where the linked monitors are located in relation to the power plants. Row 3 depicts the percent of monitors within 250km of power plants resulting in edges. Rows 4-6 depict similar percentages for 250-500, 500-750, and 750-1000km, respectively.

For Industrial Midwest power plants, most monitors are located to the east, southeast, and south (Figure 2.3.3, row 1). Monitors to the east are most likely to be linked as distance increases. For Northeast power plants, monitors are typically located to the west and

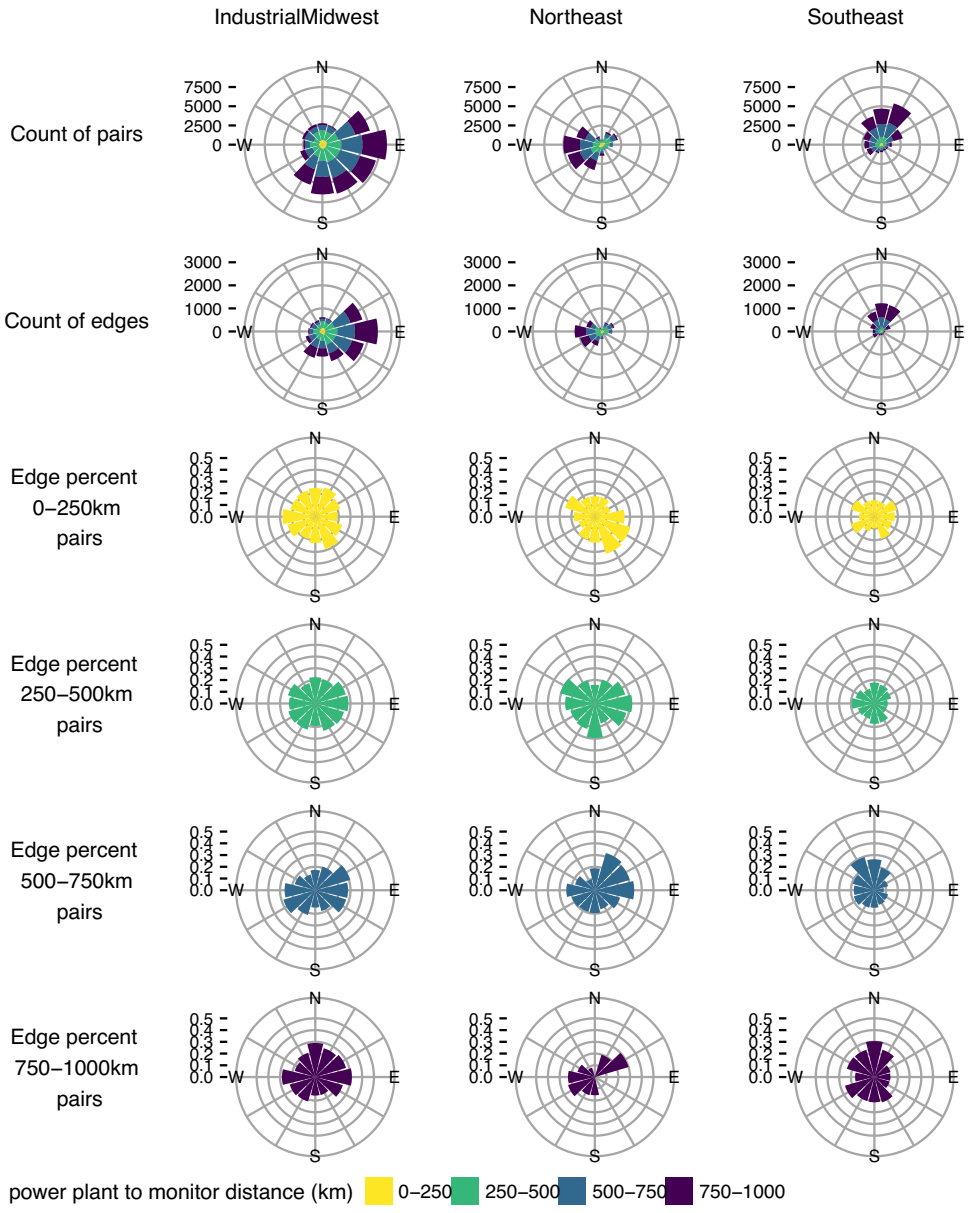


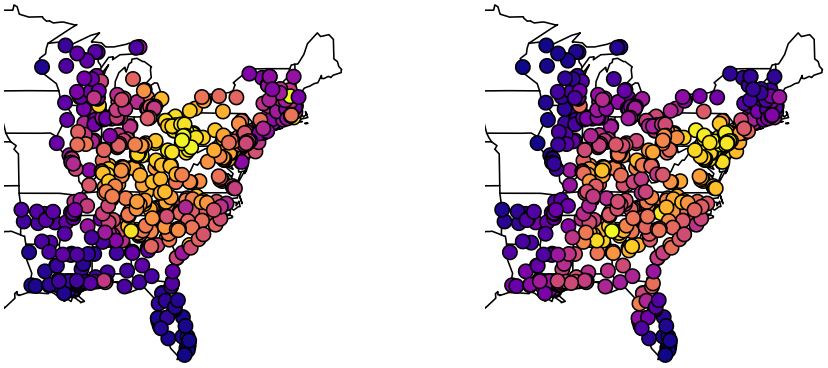
Figure 2.3.3: Compass direction from power plants to monitors by power plant region. The y-axis variable is the same across rows (row 1 - count of monitor/power plant pairs, row 2 - count of edges, rows 3 to 6 - percent of pairs that are edges). Colors represent four distance categories.

southwest. However, these power plants are much more likely to be linked to monitors to the east and northeast, consistent with weather patterns in the region. This is particularly evident in the highest distance category, where power plants link almost exclusively to monitors to the east and to very few of the many monitors located to the west and southwest. For Southeast power plants, there is not a dominant transport pattern.

2.3.4 Validation of relative monitor exposures to coal emissions

InMAP

Figure 2.3.4 depicts the relative exposures to coal emissions at monitors using the emissions network and an analogous measure from InMAP. Spearman's rank-order correlation coefficient comparing the methods is 0.80, 0.40, and 0.63 in the Industrial Midwest, Northeast, and Southeast, respectively. See Appendix B.2 for a more detailed comparison of the two methods.

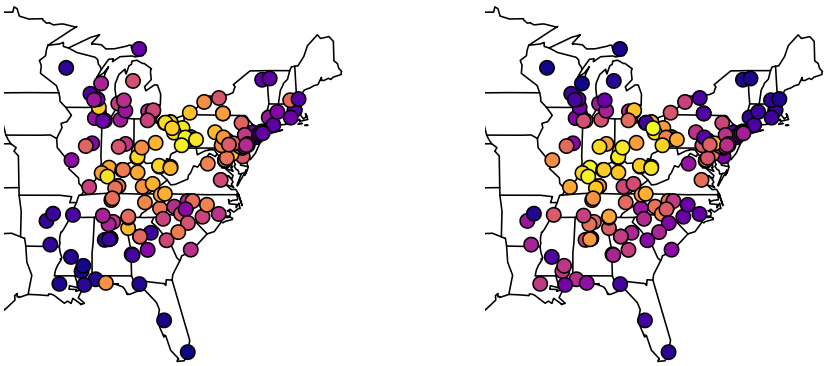


(a) Emissions network (b) InMAP

Figure 2.3.4: Relative exposures to coal emissions at monitors using the emissions network and InMAP. Lighter monitors are higher exposed.

Sulfate

Figure 2.3.5 depicts the network-derived exposure and annual average sulfate at monitors for 163 monitors measuring both $PM_{2.5}$ and sulfate. Spearman's rank-order correlation coefficient are 0.67 in the Industrial Midwest, 0.70 in the Northeast, and 0.33 in the Southeast.



(a) Emissions network (b) Sulfate

Figure 2.3.5: Exposures to coal emissions estimated from the network and annual average sulfate. Lighter monitors are higher exposed.

2.3.5 Validation of individual edge selection

Table 2.3.2 compares edge percents in high InMAP to low InMAP power plant pairs for two different cutoff values. Edges were more likely to form between high InMAP pairs in the Northeast for both cutoffs. Edges were less likely in high InMAP pairs in the Southeast and there was no difference in the Industrial Midwest.

2.3.6 Negative Associations

There were 12,446 (11.0%) power plant monitor pairs for which the time series models indicated significant negative associations between

Power Plant Region	InMAP cutoff (perc.)	InMAP cutoff (μg/m³)	Edges High (%)	Edges Low (%)
IndustrialMidwest	90	0.040	0.25	0.24
Northeast	90	0.020	0.26	0.17
Southeast	90	0.039	0.12	0.19
IndustrialMidwest	80	0.015	0.24	0.24
Northeast	80	0.003	0.22	0.17
Southeast	80	0.013	0.13	0.20

Table 2.3.2: Comparison of edge percents in high InMAP vs low InMAP pairs for two cutoffs.

daily emissions and daily ambient $PM_{2.5}$. The negative associations are discussed in the next section.

2.4 Discussion

We have explored an entirely statistical and data-driven approach to source-receptor mapping, designed to answer questions in air pollution research typically addressed with computationally intensive, deterministic models. The conceptual simplicity and computational scalability of the approach can potentially extend source-receptor modeling over a wider variety of time frames, subsets of sources and receptors, and regulatory scenarios than can be currently supported by state-of-the-science chemical transport models. Anchoring the approach exclusively to observed measurements and reasonably simple statistical models has the potential to add transparency and communicability to public policy assessments.

After confronting the most salient statistical challenges, the statistically-estimated source-receptor mapping was shown to have several strengths and several weaknesses. Our approach results in a

network that has some of the characteristics of general pollution transport patterns in the eastern United States. Edges from power plants in the Industrial Midwest and Northeast were more likely to be with montitors to the east. When compared with the output from a reduced-complexity air quality model (InMAP), the statistical approach characterizes the relative exposure levels to coal power plant emissions comparably, with agreement in exposure metrics varying by region. Some noteworthy differences between the statistical procedure and InMAP output are discussed here. In the Industrial Midwest, the statistical approach places higher exposures in northern Ohio, near the Cleveland area, relative to other areas in that region. The two methods are less correlated in the Northeast, potentially the result of an inability to distinguish between small absolute differences in exposure in the densely located monitors in the I-95 corridor. There is also disagreement between relative exposures in western Pennsylvania (Pittsburgh area) and the mid-Atlantic states, with our method placing higher exposures at the former. The two methods are highly correlated in the Southeast, with the biggest disagreement occurring in locations on the Gulf of Mexico.

Statistically estimating individual power plant/monitor edges in the network proved more challenging in terms of correspondence with the output from the reduced-complexity CTM. There is not convincing evidence the statistical approach can reliably identify the critical power plant/monitor relationships determined by InMAP. The statistical approach estimated edges at a higher rate in the high InMAP pairs in only one region (Northeast), perhaps because these power plants tended to operate more frequently throughout the year and were located in a region with more consistent weather patterns. In the Southeast, edges were actually more likely in the low InMAP pairs. The high prevalence of large, seasonally-operating power plants in that region may explain this contradictory result.

Importantly, the reliability of this comparison should also be viewed in light of the fact that InMAP contains its own limitations, and cannot necessarily serve as a "gold standard" comparison of source-receptor matrices. For example, InMAP's reliance on annual input data leads it to generally assign the highest exposures to locations in the direction of the annual average wind vector from an emissions source, which may in and of itself explain some lack of correspondence with the statistical approach, in particular when assessing edges between sources and receptors that are not connected by consistent weather patterns. In addition, InMAP uses total annual emissions from each power plant. Therefore, a large power plant, operating for only a few months of the year, would have the same impact on locations as a smaller power plant, operating continously throughout the year, with the same annual emissions. However, the ability of the statistical approach to detect edges from these two power plants will be different, given its reliance on variability in emissions and $PM_{2.5}$ to identify edges.

The statistical approach also relies on the low frequency component of the wavelet decomposition of the entire ambient $PM_{2.5}$ surface as a proxy for coal-derived $PM_{2.5}$, but such decomposition is complicated by many factors. First is the proxy nature of this strategy; while SO_2 emissions from coal power plants are likely to play an important role in regional pollution variability, the low frequency $PM_{2.5}$ component is likely driven by additional sources or emissions as well. Furthermore, areas with dense monitor coverage, typically urban, have more information for detecting and filtering local sources of pollution, thus providing an advantage over areas, typically rural, with sparse monitor coverage. Despite this advantage, urban areas present their own challenge. The large concentration of local sources (traffic, residential heating, etc) combine to form another type of low spatial frequency pollution, known as urban background pollution, making it

difficult to isolate regional pollution such as power plants in the low frequency component. Refinement of parameters in the wavelet decomposition to account for regional differences in air pollution sources, transport, and monitor density could result in improvements.

When statistical associations are detected between emissions and air quality, it does not necessarily mean the emissions caused the air quality changes. In particular, we found extreme $PM_{2.5}$ events, such as regional haze, resulted in many "extraneous" edges between affected monitors and any power plants with changing emissions during the event, regardless of their physical proximity. These events may also explain many of the negative associations detected between emissions and air quality. Adjusting for weather in the time-series models was the primary strategy for reducing the influence of these events, and seemed to be effective when compared with unadjusted, preliminary analyses. This strategy may result in some decreased ability to detect edges, given the highly correlated nature of weather, emissions, and air quality. A potential improvement to the current method would be to filter these events by temporally decomposing the $PM_{2.5}$ time-series.

In addition to extreme $PM_{2.5}$ events, similarity of power plant operating schedules likely leads to extraneous edges. If one power plant impacts a location's air quality, any plant with a similar operating schedule is also likely to form an edge with that same location, regardless of true transport patterns. These types of extraneous edges may be common in the network, as many power plants operate seasonally. For example, there was a 60% increase from May 1st to July 1st in the number of power plants operating in 2005. A greater understanding of operating dependencies could result in improved filtering of extraneous edges.

The statistical models used for estimating the network in this work were intentionally simple and designed to establish a baseline of performance using little more than the daily time series data

themselves. Future directions to refine the data driven approach include adding spatial components, or some other structure, to the models in order to share information among similar power plants/monitor. Such models could provide more spatial consistency to the source-receptor mappings and allow for edges to form between pairs with characteristics that currently inhibit edges. In addition, this feature could help filter extraneous edges and improve conditions for comparisons with deterministic models. Lastly, our approach was completely agnostic with regard to knowledge of pollution transport. A refined approach could combine some knowledge from chemical transport models into statistical models.

2.5 Conclusion

We have provided the first investigation of whether a purely statistical data-driven approach to source receptor mapping can reproduce knowledge typically produced by complex chemical transport models. The ability to do so would provide a more computationally nimble approach to estimate S-R relationships in a wider variety of settings. The results of the proposed approach were shown to hold some promise in capturing general patterns of pollution transport and source-specific exposures, but was limited in its ability to recover individual source-receptor links relative to a recently proposed reduced complexity CTM. Our investigation uncovered several statistical challenges for which we provide initial progress towards addressing, with future refinements holding promise for improving the fidelity of the purely statistical approach.

3

A Longitudinal Analysis of Source-Receptor Mappings from Coal Power Plant Emissions Networks.

3.1 Introduction

Recent work has explored the possibility of a purely data-driven and statistical approach to source-receptor mapping, the process of determining which air pollution sources affect which locations' air quality[6]. Such decisions are usually performed by deterministic, chemical transport models with important limitations in flexibility

and scalability. The proposed data-driven approach, initially applied to mapping sulfur dioxide (SO_2) emissions from coal power plants to fine particulate matter ($PM_{2.5}$) concentrations measured at USEPA Air Quality System monitors in the United States in 2005, uses statistical models to detect associations between source-specific daily emissions time-series and location-specific daily air quality time-series. Results were framed as an "emissions network" – power plants and monitors are nodes and significant associations between their daily time-series define edges in the network – representing an annual pattern in coal emissions transport for 2005. Several aspects of the network were consistent with known air pollution transport patterns and output from a reduced-complexity air quality model.

One of the key advantages of the observation-based, statistical approach over deterministic models is its flexibility; it can readily be applied under a variety of different scenarios and time frames without the complex setup and computational resources required for CTMs, potentially even more flexibly than existing reduced-complexity CTMs that rely on annual summaries [35]. In this paper, we explore the value of this computational flexibility to evaluate how source-receptor relationships vary over time. Specifically, we use the statistical methods to explore seasonal variability (winter, spring, summer, and fall) in coal emissions transport using daily SO_2 emissions from coal-fired power plants operating in the United States from 2005-2010 and daily $PM_{2.5}$ concentrations at air quality monitors. Evidence of seasonal variability in coal emissions exposures attributable to weather/transport patterns could be leveraged to generate exposure contrasts of interest in health outcomes studies. To isolate seasonal variability from the overall downward trend in coal power generation and increased air quality monitoring during the period, we restricted our analysis to 168 coal power plants and 345 air quality monitors that were continuously operating over the six year period.

The initial work proposing the statistical approach documented two critical statistical challenges [6]. First, it is difficult to detect associations with coal emissions at monitors located near local sources of pollution, such as traffic, that cause most of the day-to-day variability at those monitors. For this reason, we perform a spatial decomposition, each day, of $PM_{2.5}$ concentrations across monitors using the wavelet decomposition of [1]. Second, extreme $PM_{2.5}$ events result in many extraneous edges between affected monitors and power plants varying their emissions during the event. Therefore, we adjust for key determinants of extreme $PM_{2.5}$ events in the time-series models, namely, daily temperature and humidity. See [6] for more discussion on each of these challenges.

To evaluate seasonal variability in source-receptor patterns, we fit four emissions networks per year (one each season) from 2005-2010 and compared them across seasons and years at various levels of granularity. First, we explored whether there is seasonal variability in the geographic patterns of long-range air pollution transport. For each network, we summarized the network density, defined as the percent of possible power plant/monitor connections resulting in edges, by compass direction from power plant to monitor. Second, we investigated seasonal variability in coal emissions exposure at monitor locations. In each network, the primary measure of a monitor's exposure to coal emissions is its degree, or number of connected power plants. Summaries of monitor degree and year-to-year correlations are reported. A secondary exposure measure that weights each edge by both the distance from power plot to monitor and the emissions from the power plant was also considered. These measures quantify monitors' exposures to coal emissions in general, but are agnostic to which specific power plants are contributing to the exposures. Therefore, lastly, we considered the consistency of specific power plant/monitor relationships over time by measuring the concordance

of edge/non-edge assignments between networks in consecutive years.

3.2 Methods

3.2.1 Data

Daily SO_2 emissions were obtained from continuous emissions monitoring systems for coal-fired power plants operating in the United States from December 1, 2004 to November 30, 2010 [38]. We restricted our analysis to the 168 continuously operating power plants, defined as operational for at least one day per season for the entire period. The seasons were winter, spring, summer, and fall, beginning on the first day of December, March, June, and September and ending on the last day of February, May, August, and November, respectively.

We obtained $PM_{2.5}$ concentrations measured at AQS monitors operating in the eastern United States during the same time period. We restricted analysis to a continuously operating subset of these monitors, defined as measuring $PM_{2.5}$ on at least 100 days per year for the entire period. 345 monitors were included in the analysis. Of them, 18% measured $PM_{2.5}$ concentration daily, with the remainder measuring every third day. For these monitors, daily measurements were estimated by linearly interpolating between consecutive measurements.

3.2.2 Spatial decomposition and network formation

To filter local variability in $PM_{2.5}$ that is unlikely attributable to regional sources of air pollution, we performed a spatial decomposition of the daily $PM_{2.5}$ surface using the two-dimensional wavelets method of [1]. This spatial decomposition isolates a low spatial frequency component of the daily $PM_{2.5}$ surface believed to correspond with regional pollution sources and excludes a high spatial

frequency component believed to correspond to local sources of pollution such as nearby traffic. The decomposition is performed on each day independently. See [6] for a more detailed descriptun of the decomposition.

Next, we fit four emissions networks per year (one per season) from 2005-2010. An emissions network consists of nodes – power plants and monitors – and edges between them. Edges correspond to positive, significant associations between power plant daily SO_2 emissions and monitor $PM_{2.5}$ time-series, after adjusting for day of the week, temperature, and humidity using a generalized additive model. Specifically, for each power plant/monitor pair (for simplicity of notation, subscripts for each are omitted in Equation 3.1) within 1000km, the model is:

$$\log(PM_t) = \alpha_0 + \sum_{l=0}^{3} \beta_l(SO2_{t-l}) + f(t) + \mathbf{W}_t + \mathbf{g}(temp_t, humid_t) + \varepsilon_t \quad t = 1, \ldots, 92 \tag{3.1}$$

where PM_t and $SO2_t$ are the $PM_{2.5}$ concentration at the monitor and SO_2 emissions from the power plant on day t, and $l = 0, \ldots, 3$ is the lag in days between them. $f(t)$ is a smooth function of time (cubic regression splines with five knots), W_t a vector of indicators for the day of the week, and $\mathbf{g}(\cdot)$ are smooth functions (cubic regression splines with 5 degrees of freedom) of the temperature, humidity, and an interaction between the two on day t. In Equation 3.1, the quantity of interest is the sum of the β coefficients, as coefficient estimates for individual lags will be highly unstable due to multicollinearity. This method is commonly used in short-term health effects studies of air pollution [33]. For further discussion of this model, see [6]. The strategy for estimating the source-receptor network fits 1,391,040 models of this type, one for each power plant/monitor pair per season per year. As each such model is

estimated independently, computation distributed in parallel across 120 cores required 45 minutes of computation time.

3.2.3 Network Metrics for Quantifying Longitudinal S-R Variability

Seasonal variability in long-range coal emissions transport

In this section, we investigate seasonal and year-to-year variability in geographic patterns of long-range pollution transport. Towards this end, we summarize the network density, or percent of possible connections resulting in edges, by geographic direction from power plant to monitor, for each geographic region.

Seasonal variability in coal emissions exposures

In this section, we investigate seasonal and year-to-year variability in coal emissions exposures using two measures. The first is node degree, defined as the number of connected edges. For a monitor, it is both an indicator of overall exposure and the scope of the intervention required to improve air quality at the location. For a power plant, it is an indicator of the geographic scope of the area affected by its emissions and the value of potential emissions reductions there. For this measure, we report both the median and year-to-year correlations, by season.

As a secondary measure of exposure for each monitor, we weight each edge in the network by geographic distance and average daily SO_2 emissions during the period. Specifically, the weights are:

$$w_{ij} = \frac{1}{distance_{ij}} \times avgemissions_j \quad j \in \mathbb{P}, i \in \mathbb{M} \tag{3.2}$$

where $\mathbb{P}$ and $\mathbb{M}$ denote the power plant and monitor sets, respectively, distance_{ij} is the geographic distance between j and i, and $avgemissions_j$ is the average daily SO_2 emissions from j. Then, we obtain a single

exposure measure, x_i, for each monitor by summing over its connected edges.

$$x_i = \sum_{j \in \mathbb{P}} w_{ij} I(edge_{ij} = 1) \quad i \in \mathbb{M} \tag{3.3}$$

$I(edge_{ij} = 1)$ is 1 (0) if there is (not) an edge in the emissions network between i and j. We use this quantity to compare relative exposures to coal emissions in different time periods.

Seasonal variability in individual power plant/monitor relationships

In this section, we investigate the consistency of individual power plant/monitor relationships over time, a more granular assessment than the previous section which considered overall exposures at monitor locations. The previous section did not consider whether the exposures were attributable to the same or different power plants. Here, we compare the concordance of edge/non-edge assignments between consecutive years, by season, using three different measures. Edge (non-edge) concordance is the percent of edges (non-edges) in the previous year's network that were again edges (non-edges) in the next year's network. Overall concordance, the percent of all agreeing pairs, both edges and non-edges, is also reported.

3.3 Results

3.3.1 Daily emissions and air quality

Figure 3.3.1a depicts average daily and annual $PM_{2.5}$ concentrations measured at AQS monitors in the eastern United States. $PM_{2.5}$ concentrations were highest in the summer and lowest in the spring and fall. The average annual $PM_{2.5}$ concentration decreased 30% from 14.0 µg/m³ in 2005 to 10.7 µg/m³ in 2010.

Figure 3.3.1b depicts the daily total and annual daily average SO_2

emissions in thousands of tons from 2005-2010 for coal power plant facilities operating in the United States. Coal emissions were highest in the summer and winter. There was a 53% decrease in the average daily emissions from 2005 to 2010 (21.0 to 9.8 thousand tons per day).

Monitor characteristics

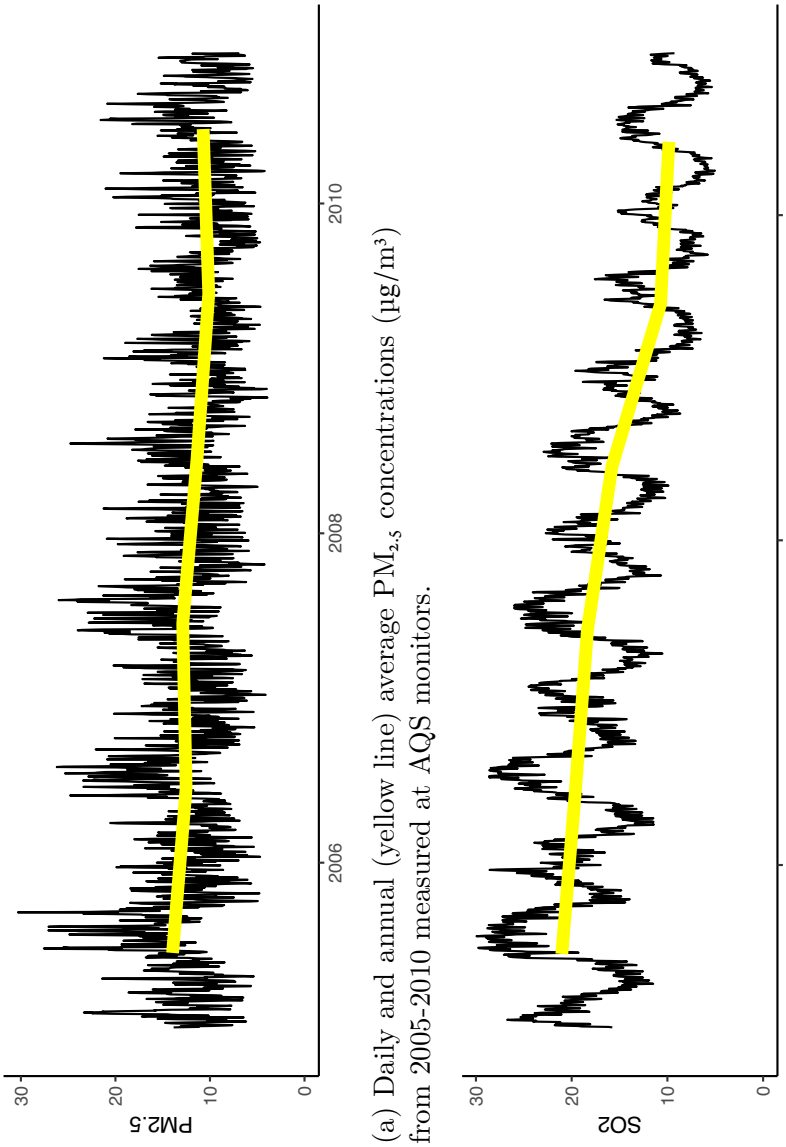
region	monitors	median distance to power plants (km)	2005 avg $PM_{2.5}$ ($\mu g/m^3$)	2010 avg $PM_{2.5}$ ($\mu g/m^3$)
IndustrialMidwest	105	521.4	14.7	10.9
Northeast	119	607.3	13.6	10.1
Southeast	121	671.3	13.9	10.6

Table 3.3.1: Monitor characteristics by region.

Power plant characteristics

region	power plants	median distance to monitors (km)	2005 daily avg SO_2 emissions (1,000 tons)	2010 daily avg SO_2 emissions (1,000 tons)
IndustrialMidwest	83	618.6	5.24	2.46
Northeast	39	507.7	3.19	1.19
Southeast	46	659.0	3.64	1.46

Table 3.3.2: Power plant characteristics by region.



(a) Daily and annual (yellow line) average $PM_{2.5}$ concentrations ($\mu g/m^3$) from 2005-2010 measured at AQS monitors.

(b) Daily and annual daily (yellow line) averageSO_2 emissions from coal power plants in thousands of tons.

Figure 3.3.1: Daily $PM_{2.5}$ and SO_2 emissions from 2005-2010.

3.3.2 Network Metrics for Quantifying Longitudinal S-R Variability

Seasonal variability in long-range coal emissions transport

Winter networks were the most dense; summer networks were the least dense. Aggregated over the six-year period, network densities were 0.25, 0.18, 0.11, and 0.17 in the winter, spring, summer, and fall, respectively. Spring and fall were typically between summer and winter, in terms of network density, with notable exceptions in the spring of 2006 and the fall of 2010 in the Northeast and Industrial Midwest. Figures 3.3.2-3.3.4 depict network density by geographic orientation from power plant to monitor for each season in the time period, aggregated over power plants in each geographic region. In many networks, power plants were more likely to make connections with monitors in specific directions. However, it does not appear that such directional trends extend across seasons in the same year nor across years in the same season.

Seasonal variability in coal emissions exposures

Figure 3.3.5 depicts the median monitor degree by geographic region and season and is a measure of exposure to coal emissions during each time period. Median monitor degrees were typically highest in the winter and lowest in the summer. From year-to-year, they were most variable in the winter and spring. The distribution of monitor degrees within each season/year (these plots are not shown) were typically symmetric about the medians depicted in Figure 3.3.5.

Figure 3.3.6 depicts the median power plant degree by geographic region and season. It is a measure of the geographic scope of the influence of emissions from a typical power plant. Power plants were

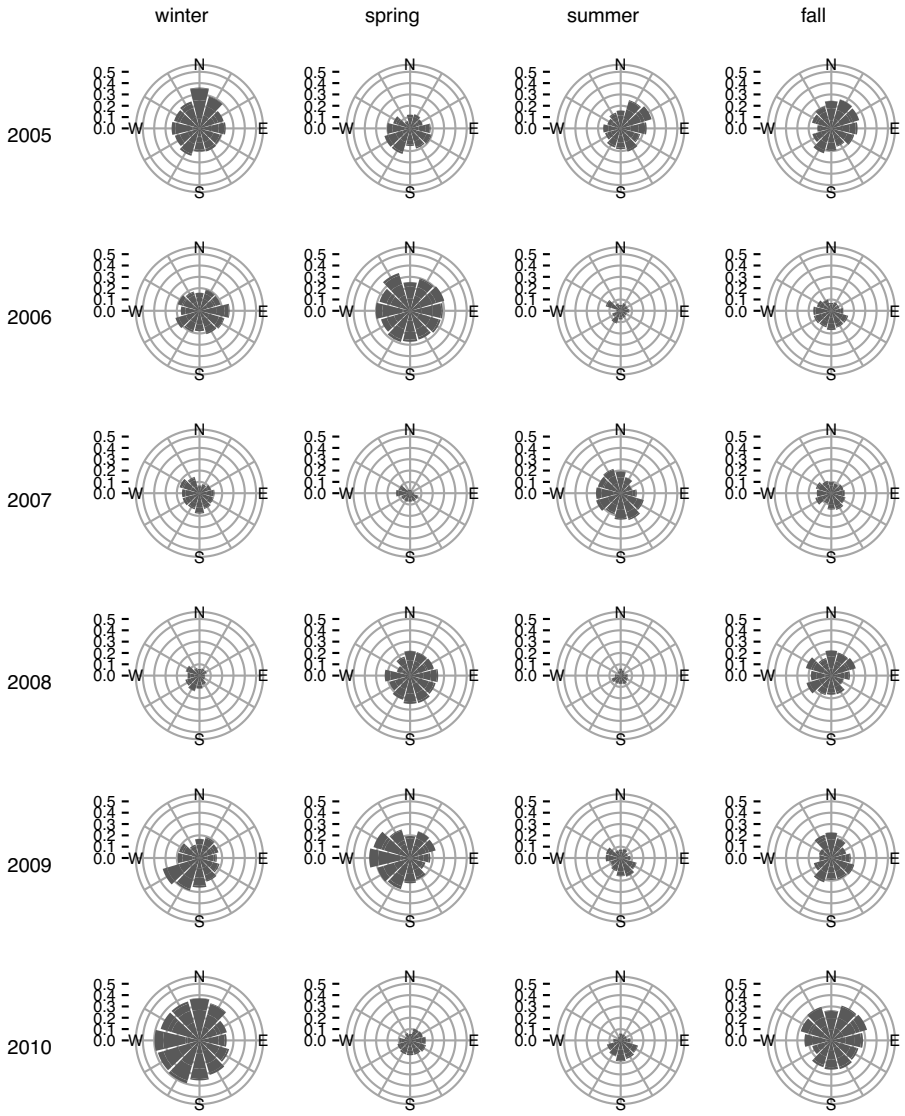


Figure 3.3.2: Industrial Midwest: network density by direction from power plants to monitors within 1000km, indicating geographic patterns in long-range transport of coal emissions.

connected to the most monitors in the winter and the least in the summer. There is more variability in median power plant degree within the same season than in median monitor degree. In addition,

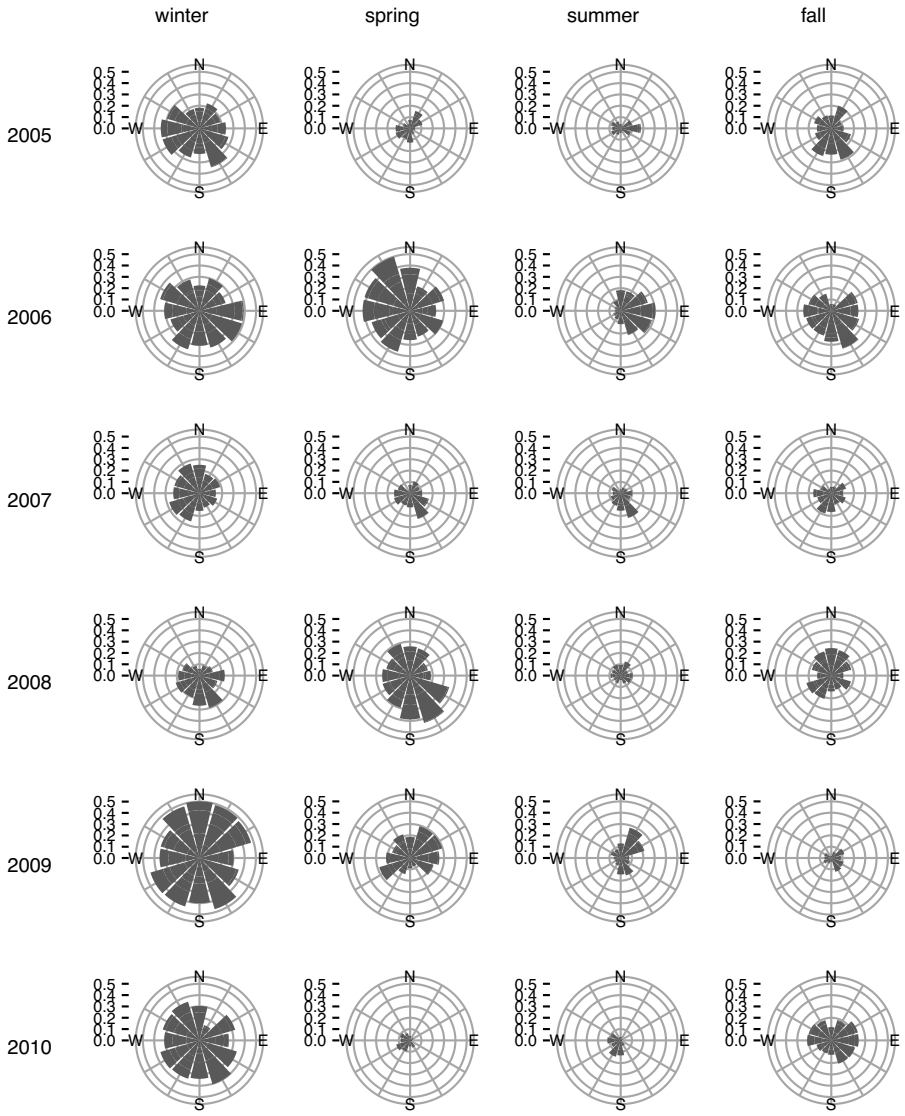


Figure 3.3.3: Northeast: network density by direction from power plants to monitors within 1000km, indicating geographic patterns in long-range transport of coal emissions.

the distribution of power plant degrees (not shown in this paper) was typically skewed, with most power plants linked to a small number of

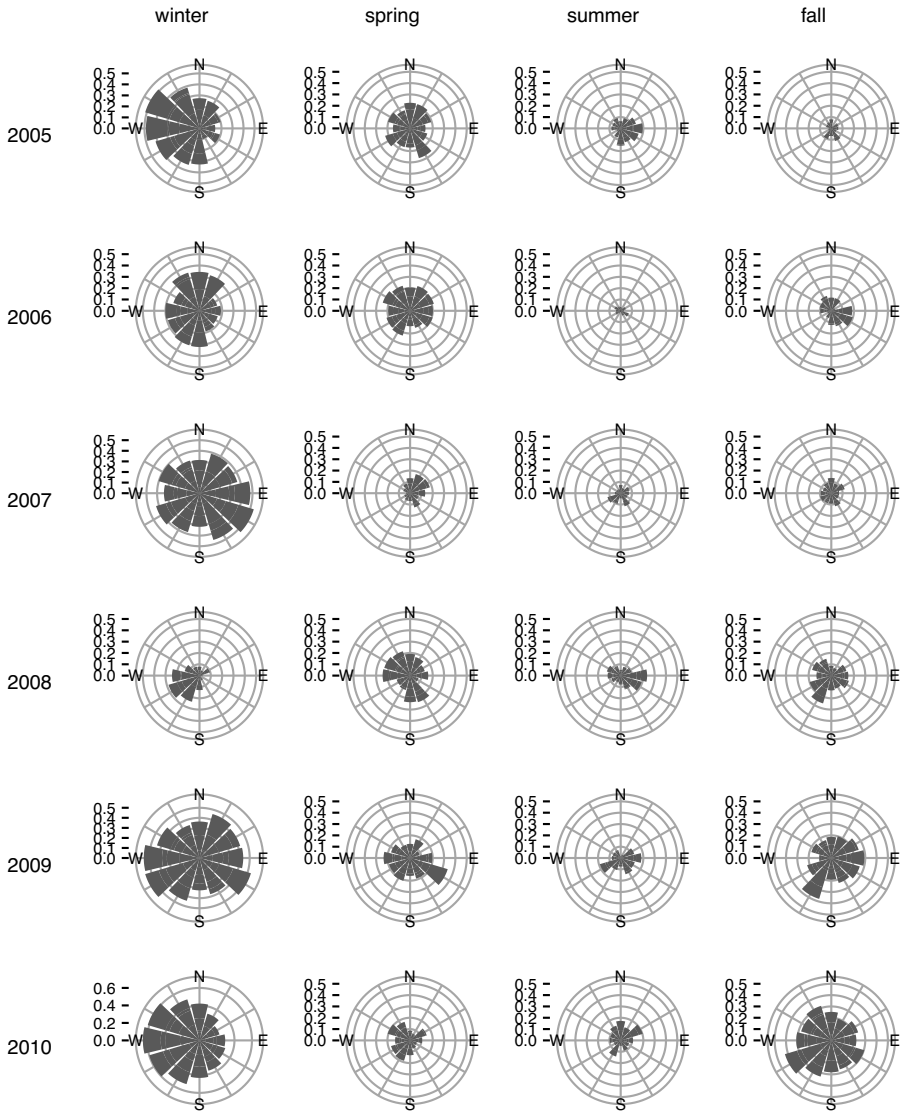


Figure 3.3.4: Southeast: network density by direction from power plants to monitors within 1000km, indicating geographic patterns in long-range transport of coal emissions.

monitors and fewer power plants linked to many monitors.

Figure 3.3.7 depicts the correlation in monitor degrees comparing the current year to the same season in the previous year. Large,

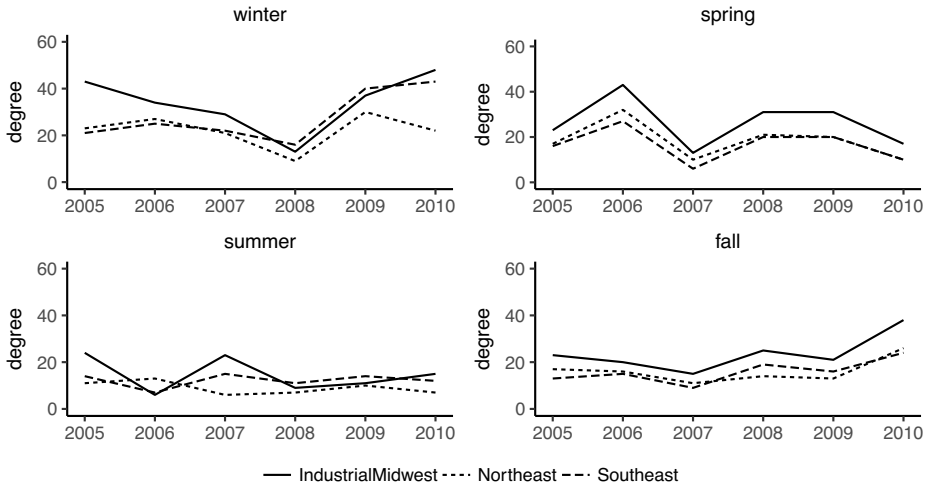


Figure 3.3.5: Median monitor degree by season, describing how strongly monitoring locations are subject to power plant pollution.

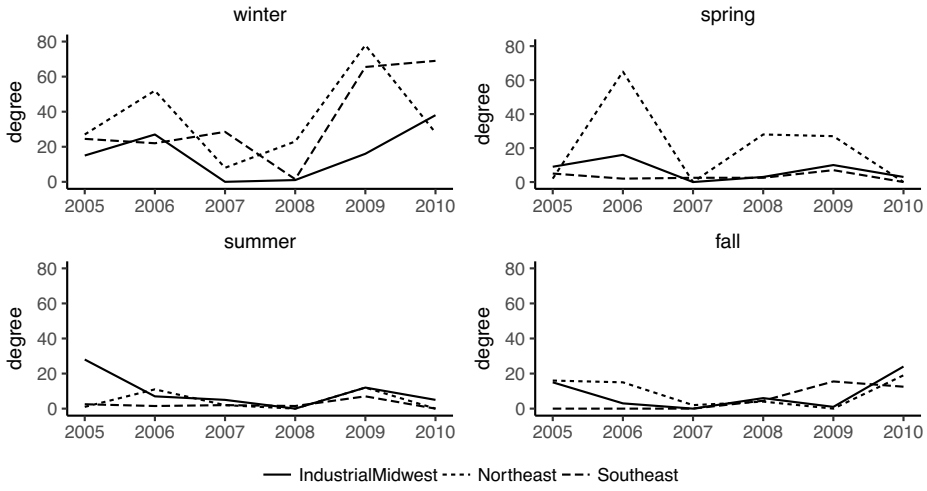


Figure 3.3.6: Median power plant degree by season, describing the scope of the area affected by emissions from individual power plants.

positive correlations indicate monitors linked to many (few) power plants in the current year were also linked to many (few) power plants in the previous year and is a measure of the consistency of exposures from year-to-year.

In the winter, spring, and fall, correlation are especially positive (with fall 2008 in the Industrial Midwest an exception), potentially indicating relative exposures are consistent across the years in that season. In the summer, correlations were mostly negative in the Industrial Midwest and Northeast. In 2006-2007, 2008-2009, and 2009-2010, locations that were high (low) exposed in the first year were more likely to be low (high) exposed in the second year.

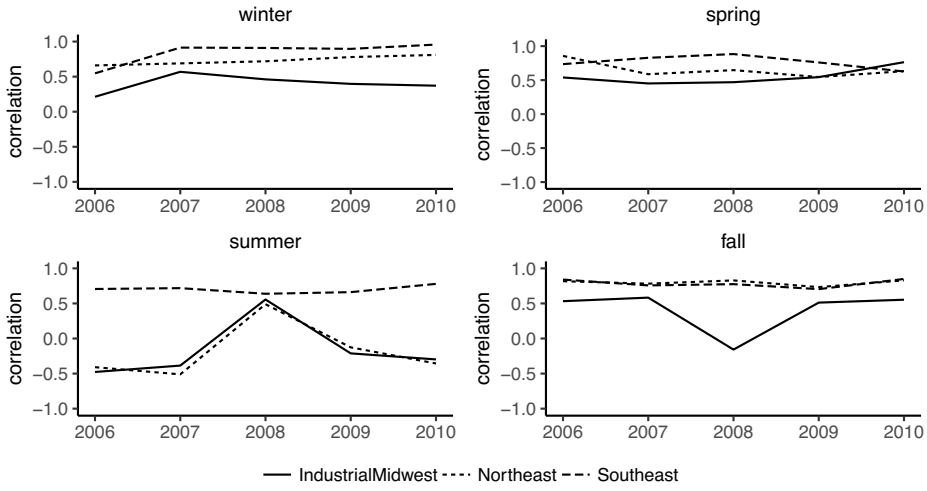


Figure 3.3.7: Spearman's rank-order correlation coefficient comparing monitor degrees from year to year by season, describing the consistency of relative exposures amongst the monitors from year to year in the same season.

Figure 3.3.8 depicts the correlation in monitor degree between consecutive time periods. For example, the large, positive correlations in spring 2005 indicate similar monitors were high (low) exposed in spring 2005 as in winter 2005. The figure shows that exposures in the Southeast were positively correlated for the entire six-year period, thus indicating the relative exposures amongst monitor locations in this region were consistent over the entire time period. In Figure 3.3.8, the negative correlation comparing Summer 2006 to Spring 2006

was the result of a one-month shift in the geographic location of high-exposed monitors. Typically, the high-exposed monitors are located across the Ohio River Valley and Pennsylvania. In Summer 2006, the highest exposures were located along the I-95 corridor in the Mid-Atlantic states and southern New England (see Appendix C.1 for detailed exposure maps). This shift was primarily the result of Northeast power plants connecting with monitors to their west in Spring 2006 and to their east in Summer 2006 (Figure 3.3.3). A similar shift in high exposures away from the Industrial Midwest occurred in Spring and Summer of 2009, but in this case, emissions from Industrial Midwest power plants were primarily responsible (Figure 3.3.2). Lastly, in Figure 3.3.8, the Industrial Midwest and Northeast display very similar season-to-season relationships as each other, with the Southeast differentiated from the other regions.

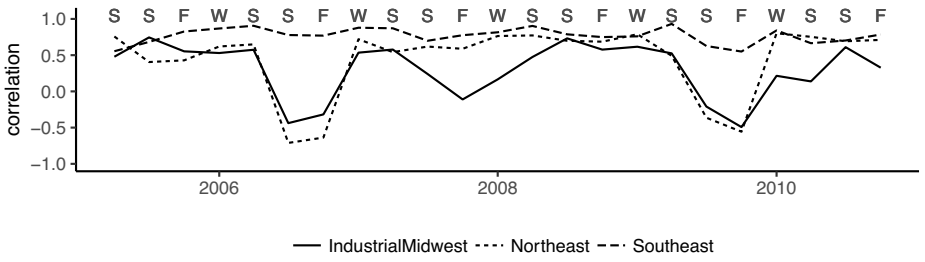
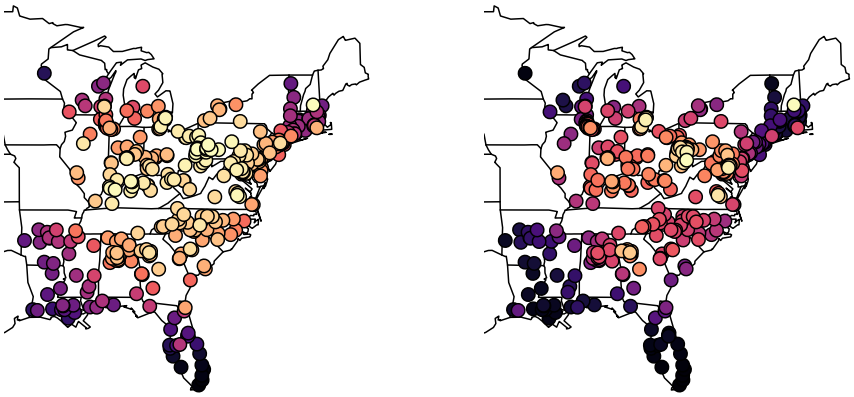


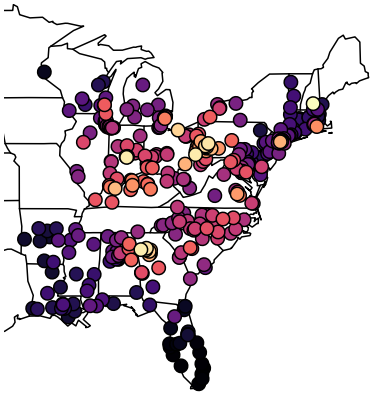
Figure 3.3.8: Spearman's rank-order correlation coefficient comparing monitor degrees in consecutive time periods, describing the consistency of relative exposures amongst the monitors from one season to the next.

Figure 3.3.9 depicts seasonal maps of the distance and emissions weighted monitor exposures (Equation 3.3). Across seasons, the highest exposed areas were in the Ohio River Valley and in an area from western North Carolina to northern Alabama. Exposures were the highest in the winter. See Appendix C.1 for individual exposure

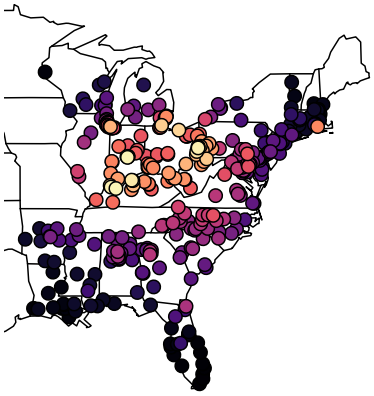


(a) winter

(b) spring



(c) summer



(d) fall

Figure 3.3.9: Seasonal monitor exposures, averaged over 2005-2010, to coal emissions using distance/emissions weighted edges in the networks. Lighter areas are higher exposed.

maps for each network.

Seasonal variability in individual power plant/monitor relationships

Figure 3.3.10 depicts the edge, non-edge, and overall concordance of power plant/monitor pairs in the network in consecutive years. Edge concordances were typically very low, indicating that individual power

plant/monitor relationships in the network change each year. Overall concordance and non-edge concordance were relatively high because network density in each time period (see Section 3.3.2) is relatively small.

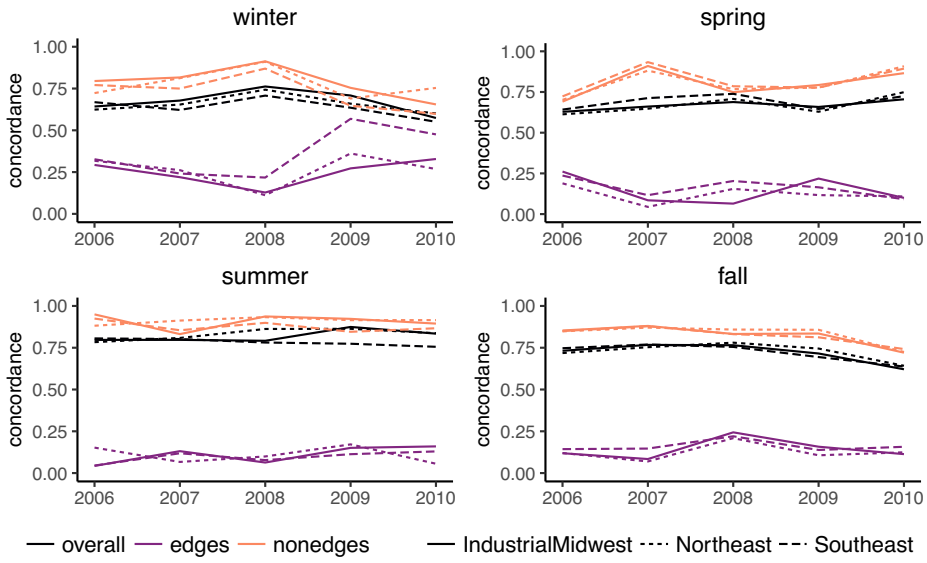


Figure 3.3.10: Concordance of each year's power plant/monitor pairs with those of the same season in the previous year.

3.4 Discussion

In this paper, we investigated the seasonal variability in several aspects of coal emissions transport and exposures from 2005-2010 using a newly-developed, observation-based, statistical approach to source-receptor mapping. Using this approach, four networks were fit per year (one for each season) and assessed for seasonal variability in coal emissions exposures, geographic patterns in long-range pollution transport, and the stability of individual power plant/monitor relationships over time.

There are striking differences between winter and summer networks, despite similar overall coal power plant emissions in these two seasons. Daily average SO_2 emissions (thousands of tons) were highest in the summer (10.9) and winter (10.2), and lowest in the fall (7.4) and spring (6.9). Winter networks had the highest densities, median node degrees, and weighted exposures of the four seasons. On the other hand, summer was typically the lowest in these measures. Together, these results point to more clearly-defined and detectable relationships between population exposures from individual power plants in the winter. A likely explanation is the different plume dispersion characteristics that dominate the two seasons. In summer, emissions plumes disperse the most both horizontally and vertically in the atmosphere, mixing with emissions from other pollution sources before affecting the air quality at monitors. Therefore, even though emissions and $PM_{2.5}$ are high in the summer, the affects of individual power plants are more diluted than in the winter, when plumes tend to stay narrower.

Many individual networks exhibited clear geographic patterns in coal emissions transport within the time period they represented, as network densities were higher in some orientations between power plant and monitor than others. (Figures 3.3.2-3.3.4). However, there were no apparent seasonal or annual trends in these patterns across different networks, which suggests significant seasonal variability in long-range pollution transport patterns. An important implication of high seasonal variability is annual exposures derived from annual input-based, deterministic models may not accurately represent shorter term exposures.

There were substantial shifts in the highest-exposed geographic locations from spring to summer in two different years (2006,2009). In both cases, these shifts affected the Industrial Midwest and Northeast, but not the Southeast, in terms of relative exposures of locations

(Figure 3.3.8. In addition, in both cases, the highest exposed locations shifted from the Industrial Midwest, which is typical (Figure 3.3.9), to farther east in the Mid-Atlantic and southern New England states (Figure C.1.1. There were no clear meteorological or other explanations for these events. Further investigation of these periods may provide valuable information on the ability of the models to detect relationship under varying conditions.

In assessing the efficacy of our statistical approach, it is noteworthy that year-to-year correlations in monitor degree (Figure 3.3.7) were mostly high, but edge concordances (Figure 3.3.10) were mostly low. Taken together, this suggests the relative exposures amongst a group of locations were stable over time, but the specific power plants contributing to their exposures were not. In other words, monitors located in geographic areas with many power plants were higher exposed, but the specific power plants contributing to the exposure are either highly variable or cannot be reliably detected. Owing to the complexities of atmospheric processes, the statistical approach, as described in this paper, may only be able to capture broad trends in air pollution transport and exposure, but not be able to reliably determine individual power plant/monitor relationships.

Our results suggest high seasonal variability in coal emissions transport patterns. The relatively simple methods described in this paper could be improved in a several ways described in [6], including adding spatial information in determining power plant/monitor relationships and incorporating information from the output of chemical transport models. In addition, this particular study focused on SO_2 emissions. SO_2 gas, which is converted to sulfate (an important $PM_{2.5}$ component) primarily in the warmer months, is not the only bi-product of coal burning that contributes to elevated $PM_{2.5}$ concentrations [34]. NO_x gas is converted to nitrate (another important $PM_{2.5}$ component) primarily in the winter months, and its

emissions are typically highly correlated with SO_2. In the mid-2000s, however, regulatory policies [39] required summertime NO_x emissions controls on power plants to reduce O_3 concentrations; during these periods, NO_x and SO_2 emissions were less correlated. Both the differences in chemical pathways that convert NO_x and SO_2 to $PM_{2.5}$ and varying implementation time lines of air quality controls have the potential to confound the results here, and suggest an opportunity for future related work that investigates different air pollutant such as NO_x.

Appendices

A

A Source-Oriented Approach to Coal Power Plant Emissions Health Effects

A.1 Data Sources

Abbreviation	Description	Source
$PM_{2.5}$	Average annual concentration ($\mu g/m^3$) of $PM_{2.5}$ in 2005 (secondary analysis only)	Di et al[8]
PctOccupied	Percent of housing units occupied	Census 2000
PctUrban	Percent residing in an urban area	Census 2000
logPop	log(total population)	Census 2000
MedianHHInc	Median household income (thousands of $)	Census 2000
PctHighSchool	Percent with a high school degree	Census 2000
PctFemale	Percent female	Census 2000
PctBlack	Percent African-American	Census 2000
PctPoor	Percent living below poverty threshold	Census 2000
PctMovedIn5	Percent moved in last 5 years	Census 2000
MedianHValue	Median house values (thousands of $)	Census 2000
mean_age	Mean age of the Medicare population	Medicare 2005
Female_rate	Percent female (Medicare pop.)	Medicare 2005
White_rate	Percent Caucasian (Medicare pop.)	Medicare 2005
avrelh	Average relative humidity (2005)	Di et al[8]
avtmpf	Average temperature (2005)	Di et al[8]
smokerate2000	County smoking rate (2000)	Dwyer-Lindgren et al [12]

Table A.1.1: Covariates included in the propensity score model.

A.2 Geographic Regions

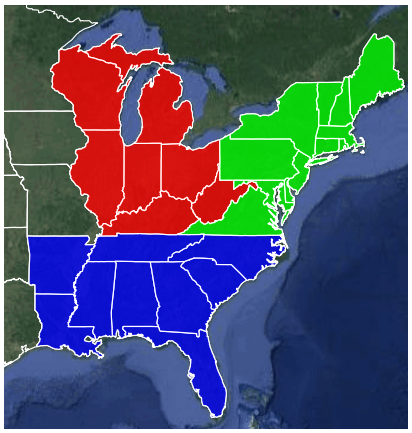


Figure A.2.1: United States regions (Industrial Midwest, Northeast, Southeast) included in this study.

A.3 Propensity Score Matching

We estimated the propensity score, which is the predicted probability of being high-exposed conditional on covariates, of each ZIP code using logistic regression. Figure A.3.1 shows the distribution of the estimated propensity scores for high-exposed and control locations. The propensity score distributions are very different, indicating stark differences in the characteristics of high-exposed and control locations and a strong threat of confounding in unadjusted, health-outcome comparisons. The purpose of the propensity score matching algorithm is to match high-exposed locations to controls with similar propensity scores.

We used a 1:1 nearest neighbor algorithm with caliper implemented in the R MatchIt package [15]. The caliper is the maximum allowable difference in propensity scores between matched locations. We used calipers equal to 20% of the pooled standard deviation of the logit of the propensity score, as suggested in Austin (2011) [3]. These calipers were 0.38, 0.61, and 0.31 for the Industrial Midwest, Northeast, and the Southeast, respectively.

In addition, locations with propensity scores outside the mutual support of the two groups' propensity scores were discarded from the analysis to prevent extrapolation beyond the observed range of covariate profiles common to both exposure groups.

The matching process resulted in 3,720 of the 6,625 (56%) high-exposed locations receiving matches with similar propensity scores. Table A.3.1 provides descriptive statistics of the matched data set. After matching, we reviewed several diagnostics to ensure the matching process successfully balanced covariates, which would adjust for confounding. Figure A.3.2 depicts one common diagnostic, the standardized mean difference, for each covariate in the raw and propensity score matched data. The standardized mean difference is

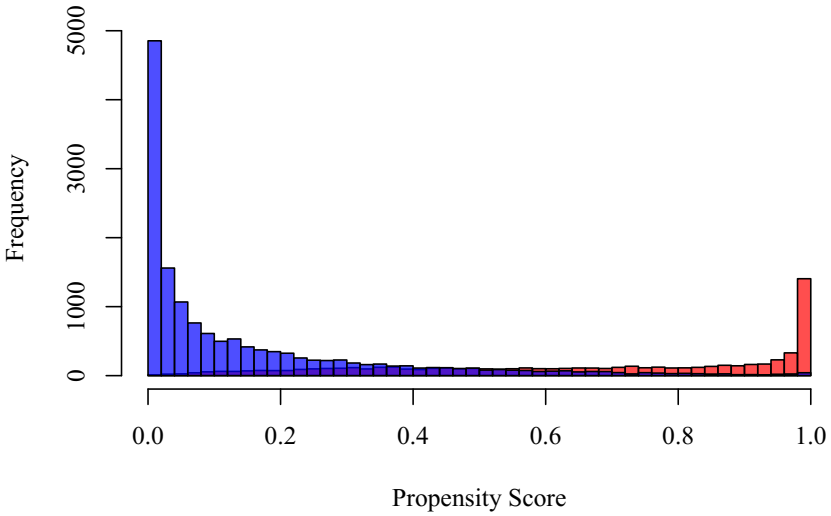


Figure A.3.1: Distribution of estimated propensity scores for high-exposed (red) and control (blue) locations before matching.

the difference in means between the high-exposed and controls, divided by the pooled standard deviation of the two groups [3]. Differences are close to zero in the matched data for each covariate in each region, except for average temperature and humidity, indicating that covariates in the matched data are "balanced" (on average) between high-exposed and matched control locations. The ability to confirm such balance is a key benefit of using propensity scores. Average temperature and humidity were adjusted for in the outcome model. Thus, the threat of confounding due to these factors is minimized.

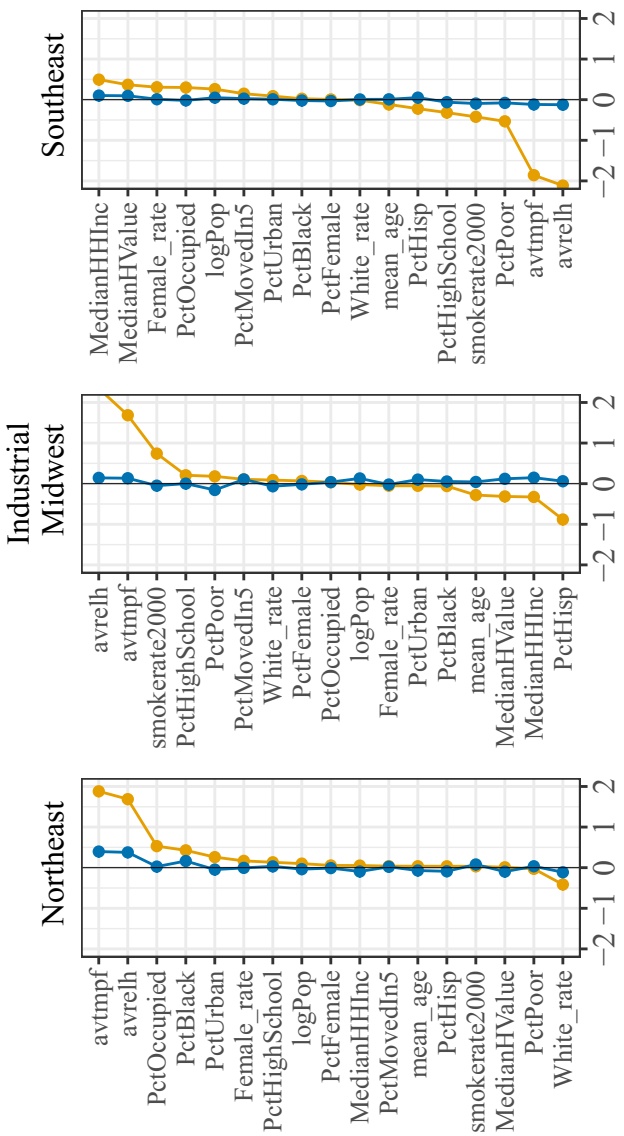


Figure A.3.2: Standardized mean difference (SMD) between the high-exposed and control groups for each covariate in the raw (orange) and propensity score matched (blue) data. Variable abbreviations are in Appendix A.1.

	Industrial Midwest		Northeast		Southeast	
	Controls	High-exposed	Controls	High-exposed	Controls	High-exposed
Number of ZIP codes	1234	1234	1266	1266	1220	1220
IHD events	19.11 (30.55)	21.85 (34.96)	28.95 (35.81)	29.93 (45.30)	26.63 (31.09)	26.86 (32.13)
Person-years	604.20 (995.62)	689.26 (1080.49)	1118.45 (1343.46)	1052.92 (1521.75)	979.04 (1143.19)	941.33 (1100.74)
$PM_{2.5}$	14.71 (1.71)	15.06 (1.36)	12.38 (1.51)	13.89 (1.62)	14.13 (1.49)	14.53 (1.58)
logPop	7.77 (1.62)	7.98 (1.62)	8.63 (1.53)	8.56 (1.76)	8.81 (1.46)	8.88 (1.39)
PctUrban	0.29 (0.38)	0.33 (0.40)	0.61 (0.43)	0.59 (0.43)	0.42 (0.41)	0.42 (0.41)
PctBlack	0.03 (0.09)	0.04 (0.11)	0.06 (0.13)	0.09 (0.18)	0.23 (0.25)	0.22 (0.23)
PctHisp	0.01 (0.01)	0.01 (0.02)	0.06 (0.11)	0.05 (0.11)	0.03 (0.06)	0.03 (0.04)
PctHighSchool	0.41 (0.10)	0.41 (0.10)	0.34 (0.11)	0.35 (0.13)	0.32 (0.09)	0.32 (0.08)
MedianHHInc	34.16 (14.06)	35.86 (11.34)	50.16 (22.76)	48.22 (23.25)	37.04 (13.62)	38.38 (11.68)
PctPoor	0.15 (0.10)	0.14 (0.09)	0.10 (0.08)	0.10 (0.09)	0.14 (0.08)	0.13 (0.08)
PctFemale	0.51 (0.04)	0.51 (0.04)	0.51 (0.03)	0.51 (0.04)	0.51 (0.04)	0.51 (0.04)
PctOccupied	0.88 (0.10)	0.89 (0.10)	0.89 (0.13)	0.89 (0.13)	0.89 (0.08)	0.89 (0.09)
PctMovedIn5	0.40 (0.11)	0.41 (0.11)	0.38 (0.10)	0.38 (0.13)	0.46 (0.10)	0.46 (0.12)
MedianHValue	76.54 (53.65)	80.53 (32.75)	165.89 (12656)	155.30 (15.15)	92.13 (53.69)	97.46 (43.70)
smokerate2000	0.30 (0.03)	0.30 (0.03)	0.25 (0.03)	0.25 (0.03)	0.27 (0.03)	0.27 (0.03)
avtmpf	285.20 (1.64)	285.35 (1.08)	283.17 (1.46)	283.84 (1.43)	289.68 (2.39)	289.48 (1.62)
avrelh	0.01 (0.00)	0.01 (0.00)	0.01 (0.00)	0.01 (0.00)	0.01 (0.00)	0.01 (0.00)
mean_age	74.48 (1.28)	74.54 (1.35)	75.36 (1.34)	75.25 (1.47)	74.57 (1.16)	74.58 (1.13)
Female_rate	0.55 (0.05)	0.55 (0.06)	0.57 (0.06)	0.57 (0.06)	0.58 (0.05)	0.58 (0.05)
White_rate	0.96 (0.09)	0.96 (0.11)	0.91 (0.18)	0.88 (0.20)	0.80 (0.22)	0.80 (0.21)

Table A.3.1: Mean (standard deviation) of ZIP code-level variables in the propensity score matched dataset.

A.4 Distance Adjusted Propensity Score Matching

DAPSm allows the investigator to modify the relative importance of propensity score similarity and geographic distance in selecting matches by specifying a weight between zero (geographic distance matching) and one (propensity score matching). Using DAPSm, instead of propensity score matching, typically results in matched data sets with geographically closer matches, but with some additional covariate imbalance. In our analysis, we created DAPS matched data sets for a range of weights using the DAPSm package in R [28] and selected the largest weight for which the standardized mean difference (SMD) of all covariates was less than 0.15.

Figure A.4.2 shows the SMD of each covariate in the DAPSm data set for a range of weights. For this sensitivity analysis, we used data sets obtained for weights 0.9975, 0.985, and 0.9975 in the Industrial Midwest, Northeast, and Southeast, respectively. Figure A.4.1 depicts the locations of the DAPSm data, which are geographically closer than the propensity score matched data in Figure 1.3.1. Table A.4.1 compares the IRRs for IHD estimated using the two matching methods.

Analysis	Industrial Midwest	Northeast	Southeast
Propensity Score Matched	1.02 (1.00, 1.04)	1.08 (1.06, 1.09)	1.06 (1.04, 1.08)
DAPS Matched	1.00 (0.98, 1.02)	1.06 (1.04, 1.08)	1.05 (1.03, 1.05)

Table A.4.1: Comparison of the propensity score and DAPS matched estimates of IRRs for IHD hospitalizations associated with high-exposure to coal power plant emissions.

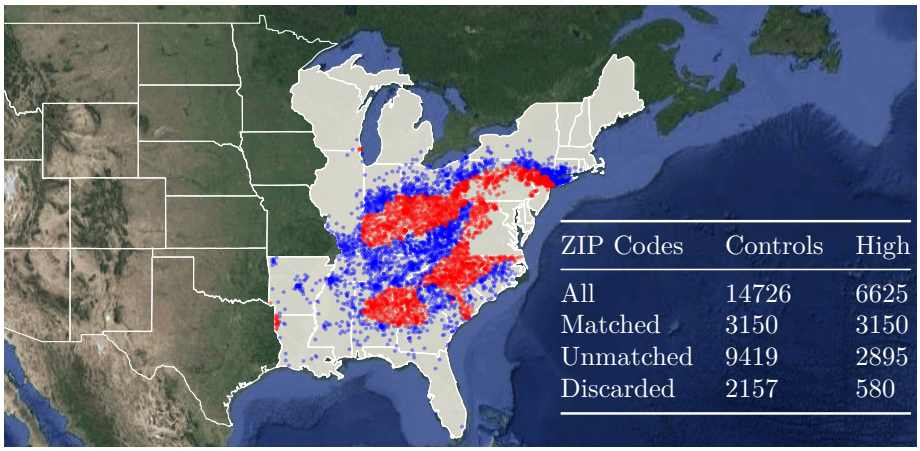


Figure A.4.1: High-exposed (red) and control (blue) locations in the Distance Adjusted Propensity Score matched data.

A.5 Secondary Analysis

In interpreting the secondary analysis, it is important to consider the relationship between the exposure and total $PM_{2.5}$ mass concentration. When total $PM_{2.5}$ mass concentrations are similar in the high-exposed and controls, the secondary analysis can be interpreted as the effects of coal power plant influence among areas with similar total $PM_{2.5}$ mass, indicating characteristics of the coal-derived $PM_{2.5}$ itself, other than just total mass, may be responsible for increased IHD.

To assess this relationship, a common measure is the standardized mean difference, which is the difference in means between the high-exposed and controls, divided by the pooled standard deviation of the two groups [3]. The standardized mean differences comparing total $PM_{2.5}$ mass concentration in the high-exposed and controls were 0.23, 0.96, and 0.26 in the Industrial Midwest, Northeast, and Southeast, respectively. Further interpretations of these results are provided in the main text.

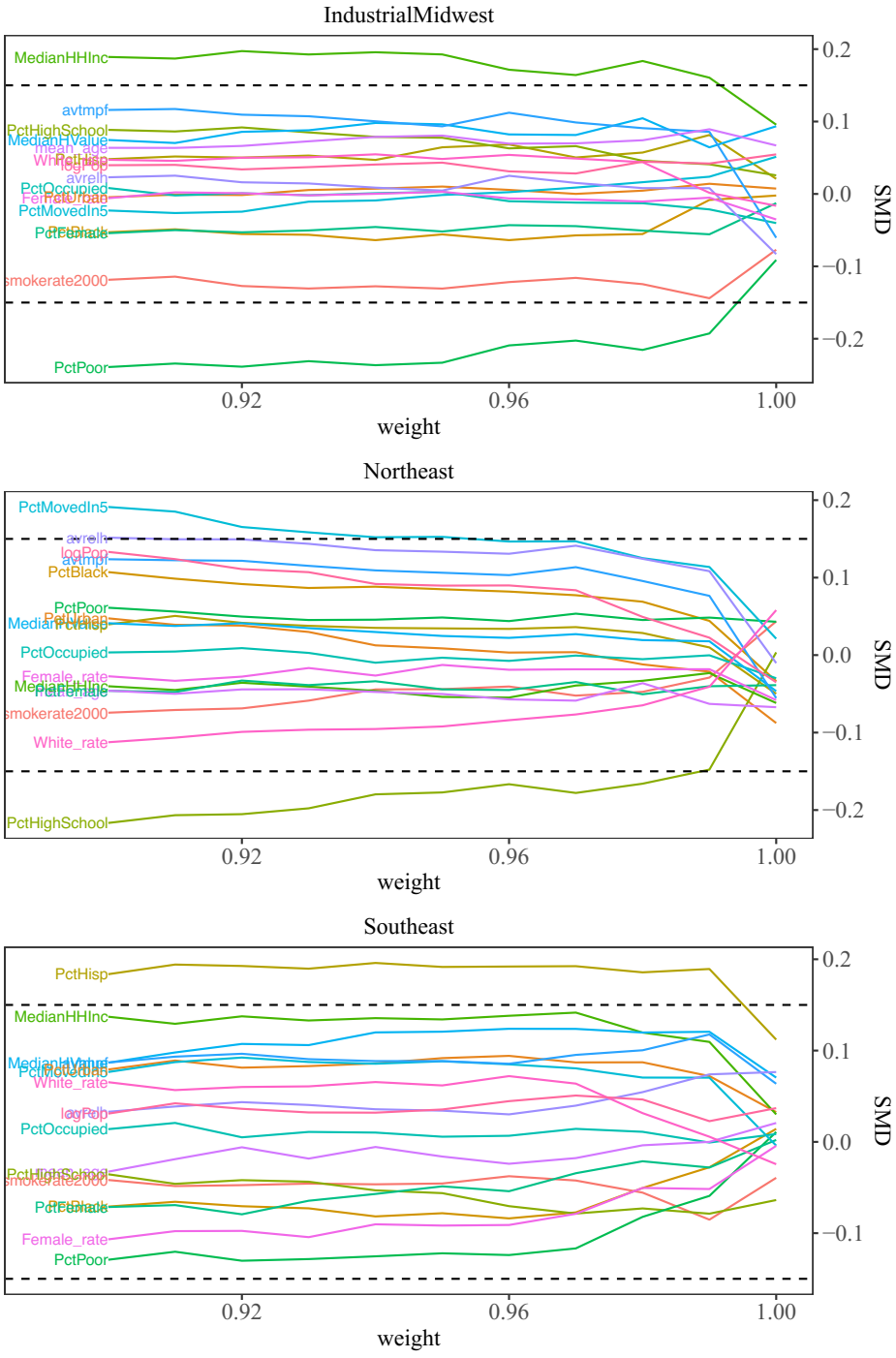


Figure A.4.2: Covariate standardized mean differences in DAPSm data sets for various weights.

B

A Data-Driven Approach to Source-Receptor Mapping of Plant Emissions to Exposed Populations

B.1 Power plant characteristics by degree

Figure B.1.1 contrasts high and low degree power plants in various characteristics. High degree power plants were those with degree higher than the median power plant.

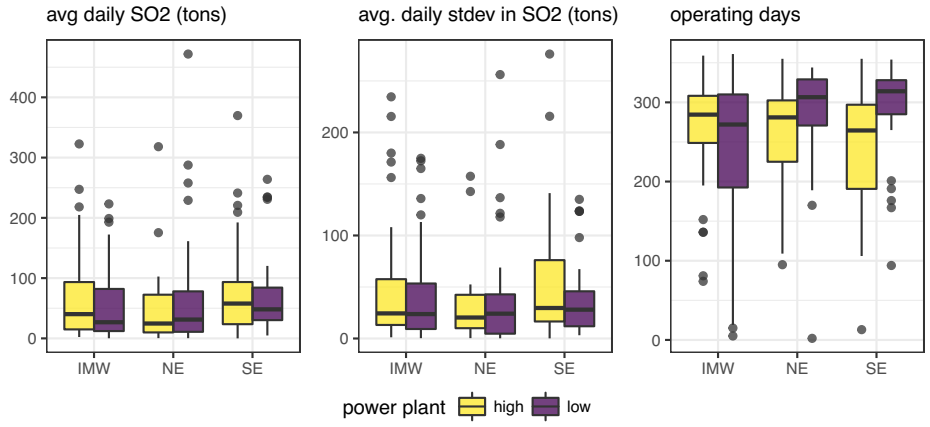


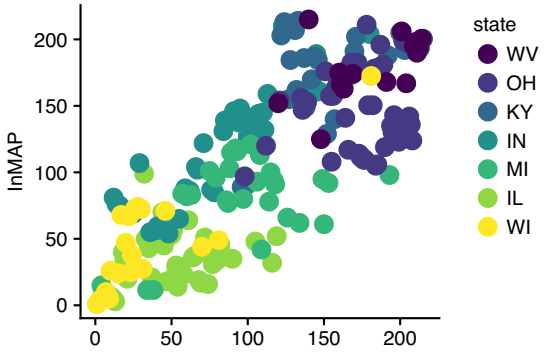
Figure B.1.1: The average daily emissions (left), average standard deviation in daily emissions (center), and number of operating days for power plant by high and low degree. The median degree is out of 365 (right) by high linked and unlinked (degree zero) power plants.

B.2 Monitor Exposures

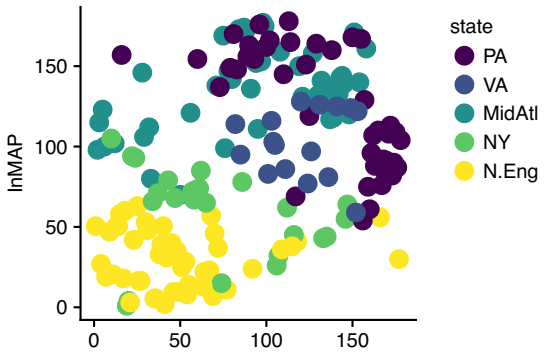
Figure B.2.1a depicts the relative exposure of each monitor in the Industrial Midwest to coal emissions according to the two methods. Both methods tend to agree on the relative exposures, listed here in increasing order, of monitors in Wisconsin, Illinois, Indiana, and West Virginia. The emissions network ranks Ohio and Michigan higher exposed, and Kentucky less exposed, than InMAP.

In the Northeast (Figure B.2.1b), the methods tend to agree on New England, Virginia, and eastern Pennsylvania. The emissions network ranks western Pennsylvania and New York higher, and parts of the mid-Atlantic less exposed than InMAP.

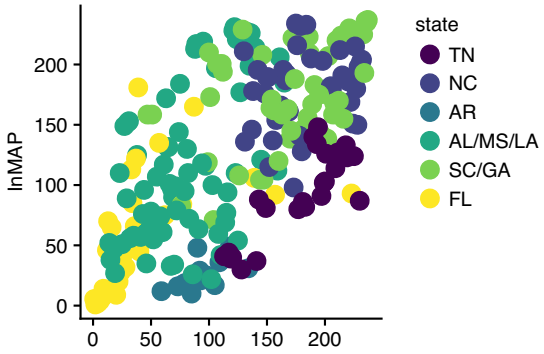
In the Southeast (Figure B.2.1c), the methods mostly agree across the region, with the most disagreement coming from the emissions network underestimating exposures along the coast of the Gulf of Mexico.



(a) Industrial Midwest (ρ = 0.80).



(b) Northeast (ρ = 0.40).



(c) Southeast (ρ = 0.63).

Figure B.2.1: Ranked InMAP and emissions network exposures at monitor locations by geographic region. ρ is the rank-order correlation coefficient.

C

A Longitudinal Analysis of Source-Receptor Mappings from Coal Power Plant Emissions Networks.

C.1 Weighted monitor exposures for each network: 2005-2010, by season

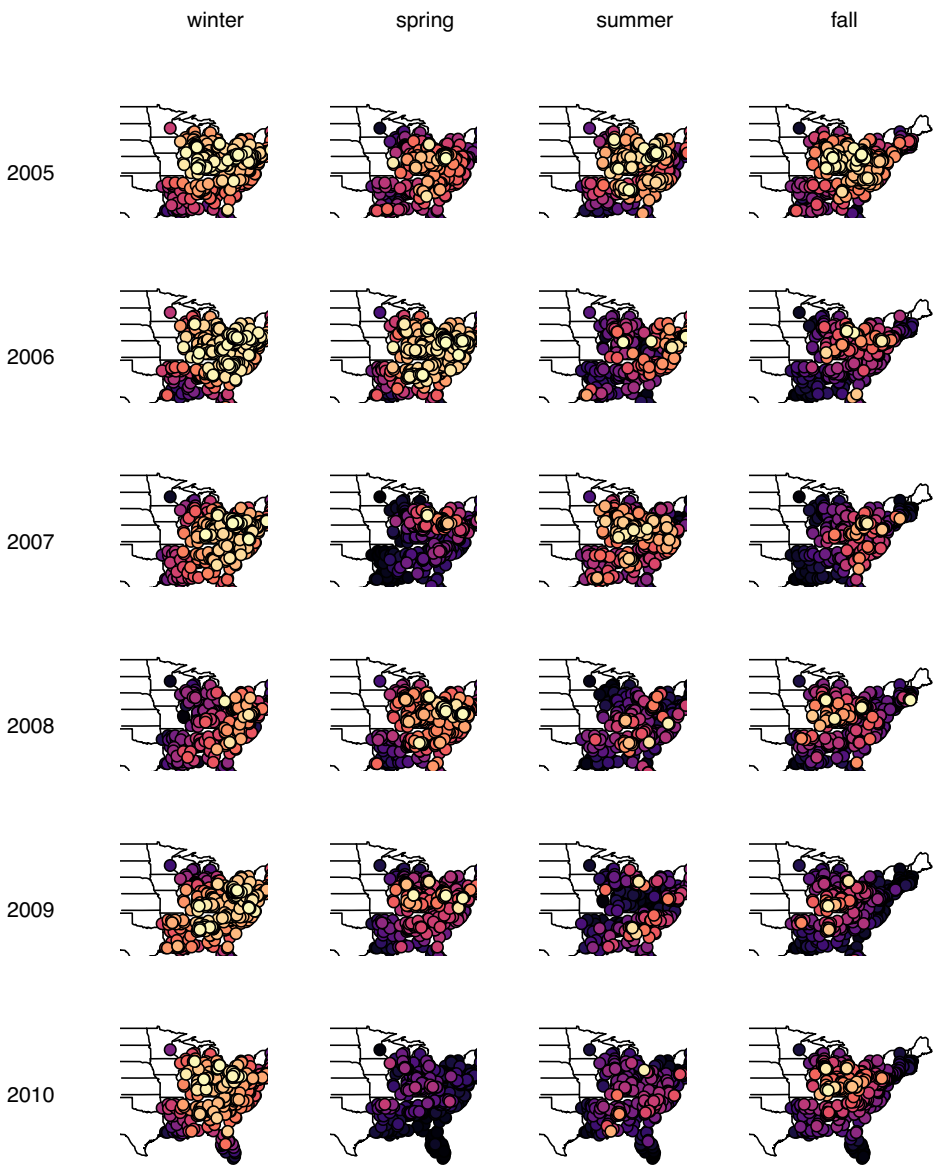


Figure C.1.1: Weighted coal emissions exposures by year and season from 2005 to 2010.

References

[1] Joseph Antonelli, Joel Schwartz, Itai Kloog, Brent A Coull, et al. Spatial multiresolution analysis of the effect of PM2.5 on birth weights. The Annals of Applied Statistics, 11(2):792–807, 2017.

[2] KW Appel, KM Foley, JO Bash, RW Pinder, RL Dennis, DJ Allen, and K Pickering. A multi-resolution assessment of the community multiscale air quality (cmaq) model v4. 7 wet deposition estimates for 2002-2006. Geoscientific Model Development, 4(2):357, 2011.

[3] Peter C Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behavioral Research, 46(3):399–424, 2011.

[4] Sivaraman Balachandran, Jorge E Pachon, Yongtao Hu, Dongho Lee, James A Mulholland, and Armistead G Russell. Ensemble-trained source apportionment of fine particulate matter and method uncertainty analysis. Atmospheric environment, 61: 387–394, 2012.

[5] Jonathan J Buonocore, Xinyi Dong, John D Spengler, Joshua S Fu, and Jonathan I Levy. Using the Community Multiscale Air Quality (CMAQ) model to estimate public health impacts of PM2.5 from individual power plants. Environment International, 68:200–208, 2014.

[6] Kevin Cummiskey, Lucas Henneman, Chanmin Kim, Brent Coull, Christine Choirat, Petros Koutrakis, and Corwin Zigler. A data-driven approach to source-receptor mapping of power plant emissions to exposed populations. Unsubmitted Draft.

[7] Frank C Curriero, Karlyn S Heiner, Jonathan M Samet, Scott L Zeger, Lisa Strug, and Jonathan A Patz. Temperature and mortality in 11 cities of the eastern united states. American journal of epidemiology, 155(1):80–87, 2002.

[8] Qian Di, Itai Kloog, Petros Koutrakis, Alexei Lyapustin, Yujie Wang, and Joel Schwartz. Assessing PM2.5 exposures with high spatiotemporal resolution across the continental United States. Environmental Science & Technology, 50(9):4712–4721, 2016.

[9] Douglas W Dockery, C. Arden Pope, Xiping Xu, John D Spengler, James H Ware, Martha E Fay, Benjamin G Ferris Jr, and Frank E Speizer. An association between air pollution and mortality in six US cities. New England Journal of Medicine, 329 (24):1753–1759, 1993.

[10] Francesca Dominici and Corwin Zigler. Best practices for gauging evidence of causality in air pollution epidemiology. American Journal of Epidemiology, page kwx307, 2017.

[11] Francesca Dominici, Roger D Peng, Michelle L Bell, Luu Pham, Aidan McDermott, Scott L Zeger, and Jonathan M Samet. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. JAMA, 295(10):1127–1134, 2006.

[12] Laura Dwyer-Lindgren, Ali H Mokdad, Tanja Srebotnjak, Abraham D Flaxman, Gillian M Hansen, and Christopher JL Murray. Cigarette smoking prevalence in US counties: 1996-2012. Population Health Metrics, 12(1):1, 2014.

[13] Constantine E Frangakis and Donald B Rubin. Principal stratification in causal inference. Biometrics, 58(1):21–29, 2002.

[14] Michael Hendryx. Mortality from heart, respiratory, and kidney disease in coal mining areas of Appalachia. International archives of occupational and environmental health, 82(2):243–249, 2009.

[15] Daniel E Ho, Kosuke Imai, Gary King, Elizabeth A Stuart, et al. Matchit: nonparametric preprocessing for parametric causal inference. Journal of Statistical Software, 42(8):1–28, 2011.

[16] Gerard Hoek, Ranjini M Krishnan, Rob Beelen, Annette Peters, Bart Ostro, Bert Brunekreef, and Joel D Kaufman. Long-term air pollution exposure and cardio-respiratory mortality: a review. Environmental Health, 12(1):1, 2013.

[17] Michael Jerrett, Michelle C Turner, Bernardo S Beckerman, C Arden Pope III, Aaron van Donkelaar, Randall V Martin, Marc Serre, Dan Crouse, Susan M Gapstur, Daniel Krewski, et al. Comparing the health effects of ambient particulate matter estimated using ground-based versus remote sensing exposure estimates. Environmental health perspectives, 125(4):552, 2017.

[18] JE Kelsall, Jonathan M Samet, SL Zeger, and J Xu. Air pollution and mortality in philadelphia, 1974–1988. American journal of epidemiology, 146(9):750–762, 1997.

[19] Francine Laden, Lucas M Neas, Douglas W Dockery, and Joel Schwartz. Association of fine particulate matter from different sources with daily mortality in six US cities. Environmental Health Perspectives, 108(10):941, 2000.

[20] Deborah D Landen, James T Wassell, Linda McWilliams, and Ami Patel. Coal dust exposure and mortality from ischemic heart disease among a cohort of US coal miners. American journal of industrial medicine, 54(10):727–733, 2011.

[21] Johanna Lepeule, Francine Laden, Douglas Dockery, and Joel Schwartz. Chronic exposure to fine particles and mortality: an extended follow-up of the Harvard Six Cities study from 1974 to 2009. Environmental Health Perspectives, 120(7):965, 2012.

[22] Jonathan I Levy and John D Spengler. Modeling the benefits of power plant emission controls in massachusetts. Journal of the Air & Waste Management Association, 52(1):5–18, 2002.

[23] Jonathan I Levy, John D Spengler, Dennis Hlinka, David Sullivan, and Dennis Moon. Using calpuff to evaluate the impacts of power plant emissions in illinois: model sensitivity and implications. Atmospheric Environment, 36(6):1063–1075, 2002.

[24] Morton Lippmann. Toxicological and epidemiological studies of cardiovascular effects of ambient air fine particulate matter (pm2. 5) and its chemical components: coherence and public health implications. Critical Reviews in Toxicology, 44(4):299–347, 2014.

[25] Kristin A Miller, David S Siscovick, Lianne Sheppard, Kristen Shepherd, Jeffrey H Sullivan, Garnet L Anderson, and Joel D Kaufman. Long-term exposure to air pollution and incidence of cardiovascular events in women. New England Journal of Medicine, 356(5):447–458, 2007.

[26] National Research Council and others. Research priorities for airborne particulate matter: IV. Continuing research progress, volume 4. National Academies Press, 2004.

[27] Bart Ostro, Michael Lipsett, Peggy Reynolds, Debbie Goldberg, Andrew Hertz, Cynthia Garcia, Katherine D Henderson, and Leslie Bernstein. Long-term exposure to constituents of fine particulate air pollution and mortality: results from the california teachers study. Environmental Health Perspectives, 118(3):363, 2010.

[28] Georgia Papadogeorgou, Christine Choirat, and Corwin M Zigler. Adjusting for unmeasured spatial confounding with distance adjusted propensity score matching. Biostatistics, page kxx074, 2018.

[29] C. Arden Pope, Richard T Burnett, George D Thurston, Michael J Thun, Eugenia E Calle, Daniel Krewski, and John J Godleski. Cardiovascular mortality and long-term exposure to particulate air pollution epidemiological evidence of general pathophysiological pathways of disease. Circulation, 109(1): 71–77, 2004.

[30] C. Arden Pope III, Michael J Thun, Mohan M Namboodiri, Douglas W Dockery, John S Evans, Frank E Speizer, and Clark W Heath Jr. Particulate air pollution as a predictor of mortality in a prospective study of US adults. American Journal of Respiratory and Critical Care Medicine, 151(3_pt_1): 669–674, 1995.

[31] C. Arden Pope III, Richard T Burnett, Michael J Thun, Eugenia E Calle, Daniel Krewski, Kazuhiko Ito, and George D Thurston. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. JAMA, 287 (9):1132–1141, 2002.

[32] James M Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. Epidemiology, pages 143–155, 1992.

[33] Joel Schwartz. The distributed lag between air pollution and daily deaths. Epidemiology, 11(3):320–326, 2000.

[34] John H Seinfeld and Spyros N Pandis. Atmospheric chemistry and physics: from air pollution to climate change. John Wiley & Sons, 2016.

[35] Christopher W Tessum, Jason D Hill, and Julian D Marshall. Inmap: A model for air pollution interventions. PloS one, 12(4): e0176131, 2017.

[36] George D Thurston, Richard T Burnett, Michelle C Turner, Yuanli Shi, Daniel Krewski, Ramona Lall, Kazuhiko Ito, Michael Jerrett, Susan M Gapstur, W Ryan Diver, et al. Ischemic heart disease mortality and long-term exposure to source-related components of US fine particle air pollution. Environmental Health Perspectives (Online), 124(6):785, 2016.

[37] Daniel Q Tong and Denise L Mauzerall. Spatial variability of summertime tropospheric ozone over the continental united states: Implications of an evaluation of the cmaq model. Atmospheric Environment, 40(17):3041–3056, 2006.

[38] United States Environmental Protection Agency. Air Markets Program Data. https://ampd.epa.gov. Accessed: 2016-06-15.

[39] United States Environmental Protection Agency. Regulatory Impact Analysis for the NOx Budget Trading Program. EPA-452/R-98-003. www3.epa.gov/ttn/ecas/docs/ria/transport_ria_final-nox-sip-call-vol2_1998-12.pdf, 1998. Last accessed on 11 April 2018.

[40] Haofei Yu, Armistead Russell, James Mullholland, Talat Odman, Yongtao Hu, Howard H Chang, and Naresh Kumar.
Cross-comparison and evaluation of air pollution field estimation methods. Atmospheric Environment, 2018.