MIT CEEPR

MIT Center for Energy and Environmental Policy Research

Working Paper Series

Learning from Supply Shocks in the Energy Market: Evidence from Local and Global Impacts of the Shale Revolution

BORA OZALTUN

JANUARY 2021

CEEPR WP 2021-002

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Learning from supply shocks in the energy market: Evidence from local and global impacts of the shale revolution

Bora Ozaltun

January 3, 2021

Abstract

In this thesis, we carry out three studies of the local and global impacts of supply shocks in energy markets, and also analyze certain properties of these markets. First, the relationship between US power plants and local air pollution is assessed from 2003 to 2016, by exploiting the information provided by the large deviations that occurred during that period due to the shale revolution. Next, fossil fuel trade is analyzed from a networks perspective, quantifying its properties. Finally, a general equilibrium model of fossil fuel trade is constructed to simulate the impact of a supply shock to a given country and in order to understand the impact of the shale revolution.[1]

[1]This paper was written as a thesis under the supervision of Christopher R. Knittel and John N. Tsitsiklis

Chapter 1

Understanding Energy Markets, Power Plants and $PM_{2.5}$ Pollution: Evidence from Shale

This chapter studies how hydraulic fracturing, through changing the relative prices of coal and natural gas, has impacted Particulate Matter concentrations across the US. Satellite estimates of pollution, fossil fuel prices, and power plant specific data is aggregated for each month between 2003 and 2016. The data is aggregated in two geographic formats: (1) for each county in the Contiguous United States (excluding Alaska), (2) for an equal size grid layout. For the grid layout, power plant fuel prices is predicted using supervised learning to create a more granular dataset. Results suggest that an increase in coal prices is associated with reduced pollution, increases in natural gas prices is associated with an increase in pollution and this impact varies based on the capacity available in the geographic unit.

1.1 Introduction

As discussed in the Introduction, hydraulic fracturing has transformed US energy prices. The dramatic changes in natural gas wellhead prices in the US can be seen

in Figure 1-1. Looking at the price paid by US electricity generators: In June of 2008, the average price of natural gas was over $12 per million BTU. During this same month, the average price of US generator purchased coal was just over $2 per million BTU. By April of 2012, the price of natural gas had fallen to below $3 per million BTU, while coal prices were largely unchanged. The change in relative prices have led to large shifts away from coal-fired generation to natural gas generation. Hydraulic fracturing (fracking) has transformed US energy prices[1]. In June of 2008, coal generation accounted for over 56 percent of US utility generation, while natural gas was 13.7 percent. In September 2019, coal was only 33.1 percent of generation, while natural gas was 38.8 percent.

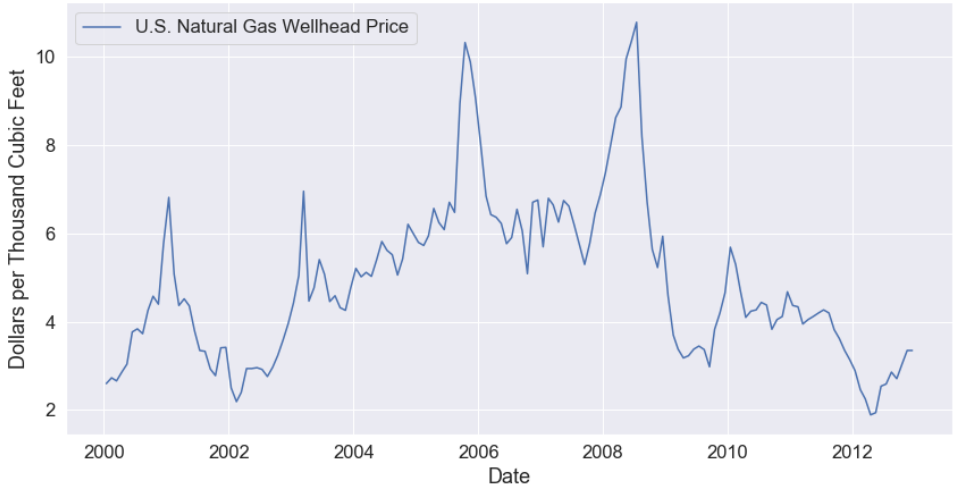

Figure 1-1: US Wellhead Natural Gas Prices between 2000 and 2018. Data can be found on EIA's website.

The reduction in pollution resulting from fracking likely had positive health benefits as the negative health impacts of ambient air pollution are well known at this point. In the US, the Environmental Protection Agency (EPA) has set National Ambient Air Quality Standards (NAAQS) for six categories of pollutants through

[1]For a more detailed discussion please see Hausman and Kellogg (2015).

the Clean Air Act. The category that this chapter will focus on is Particulate Matter (PM), specifically the sub category of PM that has a diameter less than 2.5 micrometers: $PM_{2.5}$. EPA's website documents that $PM_{2.5}$ can cause many health impacts, such as: premature cardiovascular related death, aggravated asthma, increased respiratory symptoms, as well as other problems[2].

Historically, there have been many cases where densely populated urban areas have had to deal with high pollution levels. Over the last decade, this has been most prevalent in India, China, and other countries that have been rapidly industrializing, with $PM_{2.5}$ levels in some regions regularly exceeding 100 micrograms per meter cube. Relative to levels of pollution that have been seen in the past, ambient air pollution in the US has been low in the 21st century. This success can be partly attributed to federal government policies such as the Clean Air Act and the Acid Rain program as well as other state level policies.

PM concentrations are a result of primary (direct) and secondary (indirect) impacts. If particulates are emitted from various sources such as combustion engine vehicles, forest fires and some industrial processes, this is called primary PM. If particulates are formed through various chemical reactions such as sulfur dioxide (SO_2), nitrogen oxides (NO_X) and organic compounds (OC), then this is often referred to as secondary particles. A major source of these secondary particles is from power plants [3]. SO_2 and NO_X emissions for US power plants can be found using the Air Markets Program Data (AMPD)[4]. Figures 1-2 and 1-3 visualize the data from AMPD for SO_2 and NO_X. Figure 1-2 shows the average (and 95% CI) emissions of a facility in the AMPD dataset, while Figure 1-3 visualizes the total emissions from power plants in the AMPD dataset (for a discussion on the decreasing emissions see). In both Figures, it can be seen that even though there has been a decreasing impact of coal power plants on ambient air quality, relative to natural gas power plants, we would expect them to play a larger role in secondary PM concentration.

The NAAQS for $PM_{2.5}$ concentrations in the US regulates on two different time

[2]https://www.epa.gov/pm-pollution/health-and-environmental-effects-particulate-matter-pm
[3]EPA Regulatory Data, Chapter 3: Emissions and Air Quality Impacts
[4]More information can be found at: https://ampd.epa.gov/ampd/

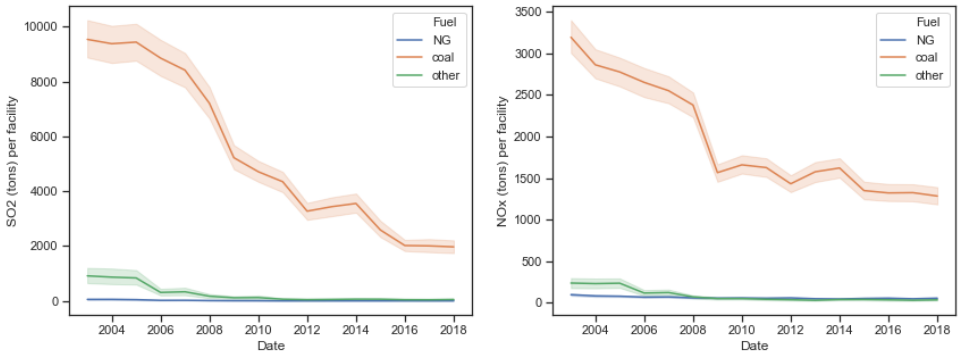

Figure 1-2: AMPD data on facility level pollution

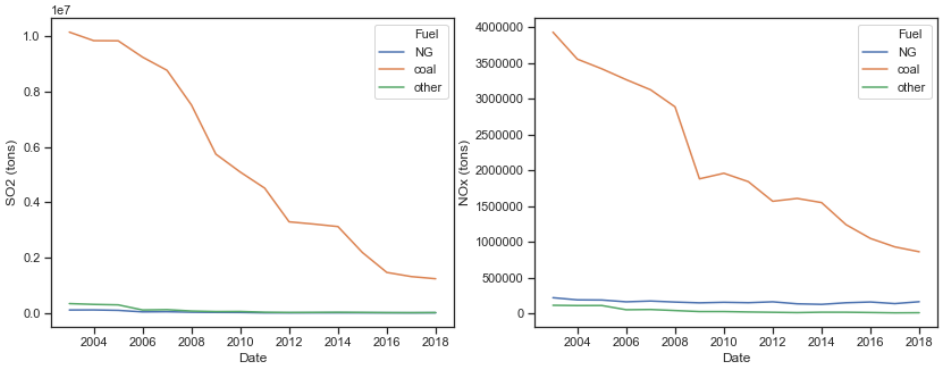

Figure 1-3: AMPD data on total pollution

scales: (1) 24-hour average $PM_{2.5}$ pollution levels in the US can't exceed 35 micrograms per cubic meter and (2) annual average of $PM_{2.5}$ can't exceed 12 micrograms per cubic meters. These levels are monitored through the Air Quality System (AQS). Across the US, there are thousands of these monitors, but since these monitors only collect data at a given point and they are not distributed in an equidistant manner within the US, they are usually seen as not being a useful dataset when trying to understand the impacts of pollution on a national scale. To deal with this issue, researchers employ different modelling techniques to extrapolate and estimate $PM_{2.5}$ in a spatio-temporal manner. The historical average $PM_{2.5}$ concentrations using the AQS data can be seen in Figure 1-4. This data can be found on the EPA's website on Particulate Matter Trends [5].

Figure 1-4 shows that the average US $PM_{2.5}$ concentrations have mostly been within the regulated standard. The major benefit policy makers are trying to achieve through emission standards on $PM_{2.5}$ is one of public health. Even in the absence of high levels of pollution, there is still evidence that long-term health effects of pollution are negative. A large study conducted on a cohort of 312,944 individuals in Europe found that even low levels of $PM_{2.5}$ concentrations (which fall within standards in the US and Europe) have negative health impacts. This study by Raaschou-Nielsen et al. (2013) found that there are still significant health risks associated with low levels of long term exposure to PM_{10} and $PM_{2.5}$.

Methods of evaluating the health impacts of pollution change from discipline to discipline. Epidemiology relies on atmospheric chemistry models and health cohort data to determine impacts. Economics approaches the problem using statistical tools to infer causal relationships between pollution and health indicators (note that some epidemiology papers employ statistical methods just like some economics papers employ insight from atmospheric chemistry).

There has been a lot of work linking the health effects of particulate matter to short and long-term health consequences in epidemiology. One review of such literature can be found in Du et al. (2016), which discusses a long list of studies and their results. Some of the epidemiology literature tries to develop a functional represen-

[5] `https://www.epa.gov/air-trends/particulate-matter-pm25-trends`

tation for what the risk associated with PM2.5 exposure is. Burnett et. al. (2014) develop a functional form for the relative risk associated to long-term exposure to PM2.5. They use a sigmoid function structure with parameters to learn their relative risk function. Their data relies on many different health studies, ones that even include the health impact of smoking cigarettes. This method becomes especially useful in prediction problems, where one's primary goal is to predict a counterfactual scenario, but on the other hand adds a certain structure to the relationship between pollution and health impacts.

From an economics standpoint, one paper that analyzes the health impacts of pollution is Chen et al. (2013). This paper looks at what they call the "Huai River policy", which is a difference in policy between the government north of the river and the government south of the river. The authors take advantage of this difference to determine the health effects of the total suspended particulates (TSPs) impact on health. There are also studies linking air pollution to infant health. Currie, Greenstone and Meckel (2017) looks at over 1 million births in Pennsylvania and their distance from fracking to evaluate the impacts it has on health.

It is apparent from research within epidemiology and economics that there are many negative health impacts associated with exposure to pollution. This raises the question: what are the causes of this pollution? As discussed before, there are many different primary and secondary impacts that make up the $PM_{2.5}$ concentration of a given place at a given time. This chapter will focus on the impacts of the energy markets and power plants. In particular, it will try to understand how county level power plant information (such as capacity, net generation and fuel consumption) and associated fuel prices play a role in impacting $PM_{2.5}$ concentrations.

As discussed previously, power plants play a major role in $PM_{2.5}$ concentrations in the US and understanding the relationship between power plant properties and $PM_{2.5}$ can help address policy measures towards regulating $PM_{2.5}$ due to its negative health implications. Fossil fuel consumption by power plants in the US has drastically changed since the mid-2000s. This change is mainly due to the supply shock created by the shale revolution, which has led to a drastic change in the fuel mix of electricity net generation. Mainly, natural gas plants have gone from providing 17.1% to 35.1%

of electricity between 2001 to 2016. This has resulted in coal power plants becoming less favorable. This exogenous shock to the energy market creates the foundation of our study, which utilizes this shock to understand the relationship between natural gas markets, coal markets, power plants and $PM_{2.5}$ pollution in the Contiguous United States. One study that is similar in nature to this chapter is Johnsen et al. (2018), which isolates the impact of shale on natural gas prices using an IV design as well as finding a negative impact on pollution due to coal generation. They use monitor level $PM_{2.5}$ values to estimate the impact of decreased coal generation. In their analysis, they use 537 of the monitors spanning 363 counties from 2007 to 2012.

The rest of the chapter is organized in the following manner: Section 1.2 will discuss the data sources and data aggregation process, Section 1.3 will discuss the final panel data and the model setup that is used to analyze this dataset and Section 1.4 will discuss the results. Section 1.5 will conclude the chapter.

1.2 Data

The study relies on unifying three different categories of data. The first dataset is spatio-temporal $PM_{2.5}$ concentration. The second category is different power plant information for each county and month specifically concerning natural gas and coal power plants. The last category is fuel price variables concerning coal and natural gas. The time span that we will cover is from 2003 to 2016 (inclusive). The analysis will setup two separate data frames. The first setup aggregates information on the county, month, year identifier. The second setup aggregates information for a smaller, equal sized (in latitude and longitude) grid, month and year.

For the county level analysis, variables in a given month, year and county include net generation, consumption and capacity information related to the coal and natural gas power plants in that given county. It also has a weighted $PM_{2.5}$ value which is further explained in the next subsection. Finally, it contains coal and natural gas fuel prices. For the grid level analysis, a similar setup is constructed, where power plants are assigned to the grid which they are located in. For this setup, we predict power plant level fuel prices in order to produce granular values. This results in a panel

data where the unique identifier is month, year and grid (which can be separated into row and column).

1.2.1 $PM_{2.5}$

As discussed in the introduction, $PM_{2.5}$ is known to have negative health impacts. The EPA has been regulating these pollutants since 1997. In 2006, the standard was updated to its current format which is that the 24-hour average concentration can't exceed 35 $\mu g/m^3$ and the annual standard can't exceed 15 $\mu g/m^3$. The EPA monitors and regulates $PM_{2.5}$ concentrations using the AQS monitoring system. In the case that one of the monitors records data that violates the given standard, the state that the monitor is in is fined. The data is collected real-time and can be accessed on the EPA's website.

One of the main issues with these monitors is that they are sparsely located, which limits the geography in which an analysis can be conducted. Less than 20% of all counties in the US have an Air Quality System (AQS) monitor. This prohibits being able to conduct analysis for the majority of US counties. Furthermore, since these estimates are point estimates, extrapolating these variables to a county or state level without using any atmospheric chemistry would lead to results that ignore spatial variations in $PM_{2.5}$ concentrations. Finally, these point estimates also lack the ability to capture the impact that pollution is having on the population in a given county, since the population in a given community is not spread out evenly. In order to create a measure of $PM_{2.5}$ pollution that a given community is exposed to, one needs to incorporate the variability of pollutants spatially as well as the variability of population concentration spatially.

In order to capture the spatial variability of pollution, atmospheric chemistry usually employs modelling. One method of getting modelled $PM_{2.5}$ concentrations is using chemical transport models. One common software tool used is GEOS-Chem. Another modelling technique used that has become more popular recently is to exploit statistical methods on datasets to create estimates of pollution. These two methods differ in their methodology in that the chemical transport models are struc-

tural models that detail the intricate chemical reactions that take place in the atmosphere calibrating the results using data. Statistical modelling techniques use atmospheric chemistry as a guide in deciding what parameters are important in capturing the true functional form of particulates in a given location at a given time.

In this chapter, we use pre-computed $PM_{2.5}$ concentrations that are modelled using statistical methods. In most of the models that are statistical, the main variable used is satellite estimates of Aerosol Optical Depth (AOD). The reason is that data for AOD is reported on a spatio-temporal matter globally from earth observing satellites. This data is publicly available from institutions such as NASA. This data is combined with other spatio-temporal weather data as the main parameters used to predict $PM_{2.5}$. This model is trained and fitted using ground truth data on $PM_{2.5}$. The dataset used in this chapter is satellite-based $PM_{2.5}$ prediction computed in Van Donkelaar et al. (2019)[6]. Resolution of the grids is 0.01 by 0.01 latitude by longitude respectively for every month between 2000 and 2016 in North America. This dataset provides us with estimates of the spatial variability of $PM_{2.5}$. In order to capture the spatial distribution that is associated with population, we weight the data by spatio-temporal population values. In order to incorporate population, we use a spatio-temporal population density values from NASA's Socioeconomic Data and Applications Center (SEDAC).

For the purposes of this chapter, the $PM_{2.5}$ data is aggregated in two different methods. All final datasets use the data produced by Van Donkelaar et al. (2019). For the county level analysis, the $PM_{2.5}$ values and the population data from SEDAC is reformatted in order to get population weighted $PM_{2.5}$ values for each county, year and month. This reformatting of the data from a gridded, spatio-temporal dataset into a panel data format takes place in multiple stages. The process follows the following procedure: (1) First, we reformat both datasets to be in the same format, (2) Next, we create a new dataset that is the product of the population and $PM_{2.5}$ raster datasets at each grid point at each time interval, (3) Then, a county level shape file is used to get the sum of the new dataset and the population dataset within the

[6]I am grateful to the authors for sharing the data. These estimates are openly available on the labs website: `http://fizz.phys.dal.ca/~atmos/martin/?page_id=140`

county boundaries and (4) Finally, these two values are divided to get the final value. For each county i and time interval t, this can be written in the following matter

$$PM_{2.5\,i,t} = \frac{\sum_{\forall k \in \mathcal{Z}(i,t)} PM^{raster}_{2.5,k,t} * population_k}{\sum_{\forall k \in \mathcal{Z}(i,t)} population_k} \tag{1.1}$$

where $PM^{raster}_{2.5,k,t}$ is the raster dataset at point k in time period t, $\mathcal{Z}(i,t)$ represents the set of points in the raster dataset that are in the county i at time t. Note that in our setting $\mathcal{Z}(i,t) = \mathcal{Z}(i)$ since the population distribution file and county shape file used for this calculation is from the 2000 census. Meaning that, these results do not capture the change in population distribution that has occurred since 2000 (or the change in county shape).

For the grid level analysis, the $PM^{raster}_{2.5,k,t}$ dataset is used in its grid format. Based on the granularity parameter that is determined, the dataset is then downsampled. For the results in this chapter, the grid is downsampled by a factor of 10. So, compared to the original raster dataset of 4550 by 9300, which is $\sim$ 43 *million* float values, the new file is 455 by 930 for each month and year. Additionally, a neighborhood is defined for each grid point to incorporate surrounding power plants. The downsampling is done for two reasons: (1) 0.01 by 0.01 latitude longitude grid will very seldom have a power plant located in the grid, let alone a coal and a natural gas power plant, which would require a large neighborhood to be defined, (2) the original grid makes the final regressions and the whole data processing exercise computationally expensive [7].

1.2.2 Power Plants

Datasets related to the consumption, net generation, price and capacity of coal and, natural gas in the US is used for this project. These datasets are aggregated from the Energy Information Agency (EIA) and the Environmental Protection Agency (EPA). Data related to net generation and consumption is gathered from the EIA-

[7] Note that robustness checks will be conducted in the future to see the sensitivity of the results to the specific grid size.

906, EIA-920 and EIA-923 forms. Additionally, nameplate capacity is found using the EIA-860 forms. Using the plant IDs, we can derive a dataset for each month, year and power plant that has consumption and net generation data.

For the county level analysis, this data for consumption and net generation is then mapped to a dataset that contains month, year and county. The allocation of consumption and net generation data into counties is done according to the county shape files. Capacity data is mapped in a similar manner. The only difference with capacity data is that the EIA-860 is an annual report, which meaning that this dataset is unique in terms of power plant and year. Initially, we produce county level variables that sum all the capacity, consumption and net generation of coal and natural gas within its borders for each time period. For capacity, the annual number is assumed for each month. We also calculate additional variables based on distance from the centroid[8]. The aggregation of the data is done similarly for the grid based analysis, but instead of county shape, the binning is conducted into the rectangular grids that each $PM_{2.5}$ value represents.

1.2.3 Fuel Price Data

For the county level analysis, we use natural gas and coal prices reported by the EIA. More specifically, we use state level price data for both natural gas and coal prices. From the EIA, we are able to find two datasets for coal prices. The first is state level coal prices averaged over all sectors on an annual basis. The second is nationwide cost of coal at electric generating plants on a monthly basis. For most of the analysis, we use the annual coal prices for each state. The reason for this is: (1) the spatial variability of coal prices is more important then the higher frequency temporal variability and (2) the limiting time frequency with power plant capacities is also annual which makes it a better fit overall. For natural gas, we are able to find monthly state level price for natural gas delivered to the residential customer. These three variables are then mapped to the county level for each month and year.

For the grid level analysis, we want to incorporate more granular price data than

[8]These results are not shown in this chapter, but the general trend in the results are similar

state level data. The ideal data for this scenario would be price of natural gas and coal at the power plant level. The EIA-923 and FERC-423 provide partial sets of data that contain the price information for a relatively small subset of power plants. Using this partial power plant level coal and natural gas price dataset, we can setup a supervised learning problem, where we use features of a power plant at a given time to predict the price of coal or natural gas at that power plant. The goal here being that once we have power plant level monthly price data, all power plant and fuel type data will be unique to a power plant in a given month and year.

One thing to note about the grid level analysis is that even though this analysis is more granular, it doesn't incorporate every grid, since the price data that is being used is at a power plant level. For example, say we want to incorporate some grid j that doesn't have any power plants in it or in its n^{th} neighbor. In this scenario, if we want to understand the impact of price on $PM_{2.5}$ in this grid point, it isn't clear how to allocate the price of coal or natural gas. Since the realized price of coal or natural gas at the power plant that has contributed to the pollution level at that given grid point is not obvious. A price of zero would indicate to the regression that a zero price has some associated pollution level (which would likely be low since this grid doesn't have any power plants in a n grid radius), when at a price of zero we would expect all power plants to be burning coal or natural gas, which would definitely increase pollution.

1.2.4 Features used to predict fuel price data

For the prediction problem, we need to develop a set of predictor variables that we have for each power plant, month, and year. Additionally, we need to have associated price data for some subset of this power plant, month, and year. The set of variables used for coal price prediction and gas price prediction are very similar (we consider a larger set of gas price predictors than coal price predictors, the reason being predicting gas prices is a harder problem). One important aspect of the partial data used to learn a prediction function is that it be a representative sample of the power plant prices that we don't have data on. Figure 1-5 visualizes the spatiotemporal

distribution of this dataset. The data doesn't seem to be concentrated in one given location or time.

Initially, we use the partial data on power plant fuel prices that we have values on to construct a state level monthly average price variable. Additionally, since this is a time series data, we incorporate the average price in a state for the previous three months and the two future months. Due to the spatial and temporal variation of the data, averaging this data on a state level for each month does not incorporate overfitting concerns. The goal of this variable is to incorporate some information of the price at power plants near the one we are trying to predict. Two robustness checks are run to make sure this is the case: (1) by leaving a subset of this dataset out of this variable as a test set and calculating the mean squared error (MSE) after training the model show similar MSE to that of the Cross Validation MSE when including the variable in the analysis, (2) when leaving the current time period out, the cross validation error does not change much. These robustness checks signal that incorporating this variable isn't causing bias in the model. This helps incorporate past, current, and future trends. Note that adding previous and future time trends lead to a loss of the first three months and the last two months. Next, we incorporate the price data that was used in the county level analysis (annual state level coal prices and monthly state level natural gas prices). The latitude an longitude of the plant, the year, month, and state variables. For natural gas, due to the high volatility, we include LNG imports and LNG import prices for each month year for the US (data can be found on EIA's website). Finally, power plant level net generation, consumption, and capacity variables are incorporated.

1.2.5 Final datasets

We can briefly outline the county level and grid level datasets. Both datasets are identified based on a geographic area, month, and year. For a given geographic area, month, and year, the datasets contain associated $PM_{2.5}$ values, net generation, consumption, capacity, and price values for coal and natural gas power plants. The data spans from 2003 to 2016. For the county level dataset, the $PM_{2.5}$ value is

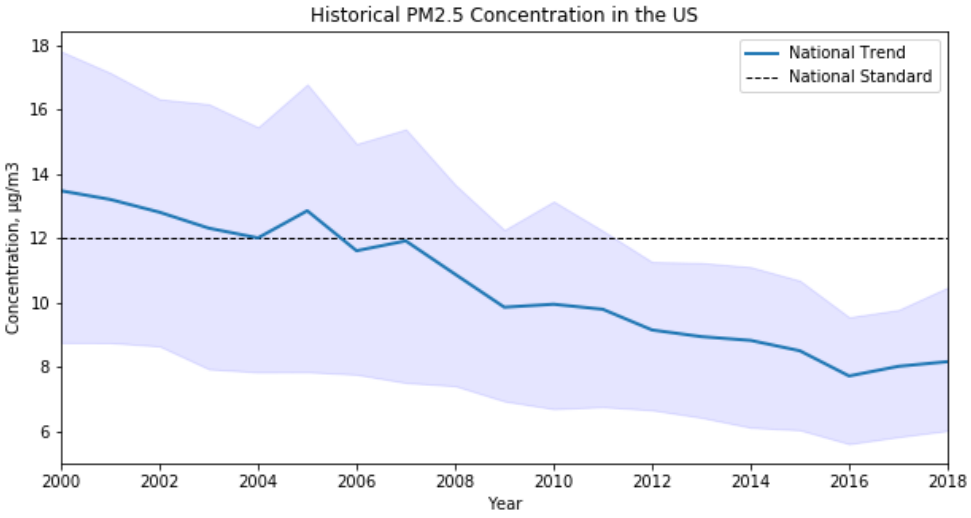

Figure 1-4: EPA's historical $PM_{2.5}$ air quality estimates based on 412 sites

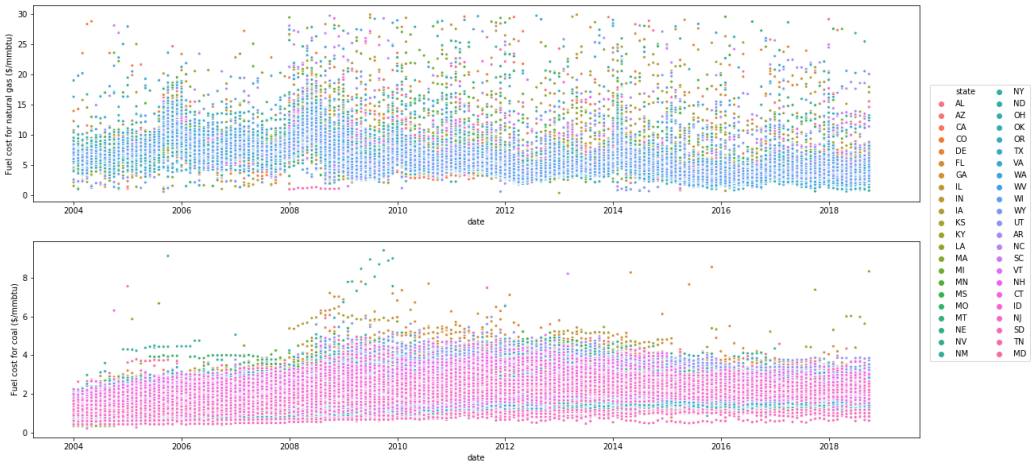

Figure 1-5: Spatiotemporal distribution of the power plant level price data.

weighted by population. For the grid analysis, each grid is equal sized in terms of latitude and longitude. For the grid analysis, the price data is not incorporated for any grid that doesn't have a coal or natural gas plant located within the grid or its defined neighborhood.

1.3 Methods

In this section, we will discuss the different analyses that will be run on the datasets. *County level analysis* will detail the county, year, monthly regressions that are run, *predicting power plant prices* will discuss the machine learning (ML) framework that is used to predict power plant level fuel prices, and *grid level analysis* will detail the grid level regressions that will be run.

1.3.1 County level analysis

The final county level dataset is in a panel data format where we have values for each month, year, and county in the lower 48 states of the US. Our main econometric framework leverages changes in natural gas and coal prices across time and space. We progressively add controls for county and time effects.
To start off, we can start with a basic framework:

$$PM_{2.5,i,t} = \beta X_{i,t} + \alpha_i + \nu_t + \omega_{i,t} + \epsilon_{i,t} \tag{1.2}$$

where $PM_{2.5,i,t}$ is the $PM_{2.5}$ values at location i at time t, $X_{i,t}$ is some subset of the variables described in the previous section (natural gas and coal prices, consumption, capacity, or net generation) and $\alpha_i, \nu_t, \omega_{i,t}$ are the fixed effects. We begin with a set of county, year, and month-of-year fixed effects. We then add county-specific trends to this specification and end with county-by-year fixed effects. We also use different transformations of the variables based on how we would expect the interactions to vary. In the next framework we can look at the case where we want to incorporate potential transformations:

$$\Phi_{PM}(PM_{2.5,i,t}) = \beta\Phi_X(X_{i,t}) + \alpha_i + \nu_t + \omega_{i,t} + \epsilon_{i,t} \tag{1.3}$$

where $\Phi_\cdot$ represents a dictionary of transformations that are deemed appropriate. Finally, we might look at potential interactions between variables:

$$\Phi_{PM}(PM_{2.5,i,t}) = \beta\Phi_X(X_{i,t}) + \theta(\Phi_{X_1}(X_{1,i,t}) \times \Phi_{X_2}(X_{2,i,t})) + \alpha_i + \nu_t + \omega_{i,t} \tag{1.4}$$

where, $X_{1,i,t}$ and $X_{2,i,t}$ are two subsets of $X_{i,t}$ and $\times$ represents the interaction between the two terms. The standard errors used for all analyses are robust standard errors using county level clustering.

1.3.2 Predicting power plant fuel prices

As discussed in the previous section, the ideal dataset we would like is fossil fuel prices at a power plant level, which is partially available, but the dataset has a lot of missing values. In order to get past this issue, we train a supervised learning algorithm to predict fuel prices at a power plant for a given month and year.

Before doing so in the results section, we can setup the basic framework: our goal is to attain an approximation $\hat{f}(x)$ for the function $f(x)$ that in expectation outputs the price for a given power plant in a given month and year given a set of features x which is associated with that power plant in the given month and year. In other words:

$$y = f(x) + \epsilon \tag{1.5}$$

where ϵ is some zero mean noise. This will give us the relationship, $f(x) = E[y|X = x]$. In order to achieve a good function approximation, we can use ML algorithms that try to approximate this relationship given a set of data $D = \{y_i, x_i\}_{i=1}^n$, where y_i is the outcome of interest and x_i is a vector of features.

In the case of predicting power plant fuel prices for a given month, year: i is an index for the specific $(month, year, power\ plant, fuel)$ combination, $y_i \in R^+$ is the

power plant fuel price for index i, and x_i is the feature vector for the power plant fuel price.

For a given model evaluation, mean-squared-error (MSE) is used as the evaluation and 5-fold cross-validation is used to avoid overfitting and hyper parameter tuning. Pre-processing of the training data includes feature engineering methods. Models tested are all regressions (since y is real valued): k-Nearest-Neighbors (kNN), Random Forest, XG-Boost, Ridge, Lasso, Elastic Nets.

1.3.3 Grid Level Analysis

The overall method used to analyze the grid-level data is similar to that done for the county-level data. The goal of incorporating the grid data is to be able to get higher resolution results without the abstract shapes of counties. For the grid-level data, we have a certain rectangular grid resolution ($M \times P$), where each grid covers the same latitude and longitude. This choice is based off of the shape that the predicted $PM_{2.5}$ data is in.

In addition to the given data in a grid, we can also aggregate information about surrounding grids. Lets call these surrounding grids neighbors, and the n^{th} neighborhood is the set of rectangles that are n grids away. Figure 1-6 visualizes this relationship for a grid i and its 1^{st} and 2^{nd} neighborhood. We can incorporate the n^{th} neighbor information, as we might expect that the $PM_{2.5}$ levels in one grid are a result of power plants in the surrounding grids. Similar to the county level analysis, the setup of the regression will be that given in Equations 1.2, 1.3, and 1.4. Instead of county, there will be grid identifiers.

1.4 Results

This section discusses the results of using the data and model framework outlined in the previous sections. The goal of this section is to understand the relationship between natural gas, coal and ambient air pollution in the US between 2003 and 2016. The results exploit the major shift in natural gas and coal prices as well as the

shift in power plant fuel type that was described in the introduction. First, we will discuss the results of the county level dataset, next we will discuss the price prediction problem, and lastly we will discuss the grid level results using the predicted prices.

1.4.1 County level analysis

Table 1.1 shows us different regressions run on the impact of prices on emissions in a county. All three are fixed effect regressions with robust county clustered standard errors. The first column follows the framework of Equation 1.2 and the second and third column follows the framework of Equation 1.3. The transformation in the second column is $log(\cdot)$ and $arcsinh(\cdot)$ in the third column (denoted $hsin(\cdot)$). The X variable in this Table is natural gas and coal prices. The Fixed Effects (FEs) for this table are county fixed effects as well as month-of-year fixed effects. We can see significant and consistent results that would suggest that an increase in natural gas prices in a given state increases the pollution in a county and an increase in coal prices decreases the emissions. This is what we would expect intuitively since increased natural gas price would lead us to use a substitute fuel, which in this case pollutes more on average. An increase in coal prices would lead us to decrease the amount of coal used, which means that the substitute fuel for coal pollutes less. One way of stating the results in Column (1) is that an increase of \$1 in coal prices decreases county level $PM_{2.5}$ pollution by -0.51 $\mu g/m^3$ and an increase of \$1 in natural gas prices increases county level $PM_{2.5}$ pollution by 0.14 $\mu g/m^3$. This is quite a large impact if we think of average $PM_{2.5}$ levels in the US that can be seen in Figure 1-4.

Table 1.2 and Table 1.3 have the same setup as Table 1.1, but the X variable used is total county net generation and consumption of natural gas and coal (respectively). The results are similar in intuition, but provide more heterogeneity in the data. Price data only had state level variation, whereas net generation and consumption data has county level variation. The results are a bit harder to interpret in a familiar manner since MWh or MMBtu are not as clear as dollars. One interesting point from these Tables can be seen in the NG coefficients in columns 2 and 3 in both Tables. The

pollution impact of changes in NG is no longer significant, which is not immediately clear when comparing it to the price setting. We can justify that it isn't significant since increasing net generation or consumption on a plant level increases pollution independent of fuel type (natural gas or coal), while contradicting with the fact that it also might potentially replace future coal net generation or consumption. This potentially explains the difference seen to the price variables that are a big driver of net generation and consumption as well as power plant transitions.

Turning back to price data, we might be interested in how the parameters are impacted for varying FE settings. The results of this can be seen in Figure 1.4. The table shows three alternative settings of FEs, including, county, year, month-of-year, and county trends. We choose a log-log formulation in this table. The results are significant to the same level in all three setups which shows the robustness of the regressions.

Table 1.5 is trying to capture the relationship between fuel prices, county level capacity values as well as interactions between prices and county level capacity values. The framework for this table can be seen in Equation 1.4, where we have transformations to variables as well as interactions. For this table, all X values are log transformed. We perform four sets of regressions, where regression (1) only regresses on fuel prices, regression (2) incorporates capacity in a given county into the mix, regression (3) adds interaction variables and regression (4) transforms the y variable, weighted $PM_{2.5}$, using $log(\cdot)$.

In regression (2) of Table 1.5, we can see that the coefficients of the price variables are relatively constant. The capacity coefficients indicate that as we see an increase in coal and natural gas capacity in a given county, we will also see an increase in pollution. But, there will be more of an increase in pollution with a percentage increase in coal capacity in a county than if there is a percentage increase in natural gas capacity. We can also see that the coefficient of the natural gas capacity is not significant. Meaning that the impacts of a percentage change in natural gas capacity can have ambiguous results.

The next two columns incorporate interactive variables between capacity parameters and price. The main difference between them is that the y variable is the

weighted PM for regression (3) and the logarithm of weighted PM for regression (4). These results also reveal interesting dynamics.

The interactions between coal capacity and coal prices suggest that the impact is negative. Both columns show that this result is significant to the 99^{th} percentile. This would indicate that an increase in coal prices when there is coal capacity around you decreases the weighted $PM_{2.5}$. Or in other words, if you have coal capacity around you, you are better off when coal prices are high. The coefficient of the interactions between coal prices and gas capacity suggest that its impact on prices is positive. The coefficient is significant to the 99^{th} percentile for column (3), yet isn't significant for column (4). These results suggest given that there is coal capacity around you, an increase in gas prices increases pollution. Or in other words, if you have coal capacity around you, you are worse off when gas prices are high.

The interactions between natural gas capacity and coal prices suggest that the impact is positive. The coefficient isn't significant in both columns, but has the same sign. This would indicate that an increase in coal prices when there is gas capacity around you increases the weighted $PM_{2.5}$. We can understand why this is not significant, since we generally see less pollution when coal prices increase. Since prices are usually at a state level or even regional, it is not unreasonable that this variable's impact is not clear. The interactions between natural gas capacity and natural gas prices suggest that the impact is negative. The coefficient is significant to the 99^{th} percentile in both columns. This would indicate that an increase in natural gas prices when there is natural gas capacity around you decreases the weighted $PM_{2.5}$ (independent of the coal capacity). Or in other words, if you have natural gas capacity around you, you are better off when natural gas prices are high.

Table 1.5's results are novel because of the strong relationship it shows between the energy market, power plants and pollution. The coefficients are intuitive in a certain way as we are able to directly show significant impacts of prices and capacity as well as their interactions on county level weighted pollution. When the interactions aren't clear, we see that the significance also decreases, which can be seen when interacting coal capacity and natural gas prices as well as natural gas capacity and coal prices. Table 1.6 shows similar regressions, but in a log-log transformation setup.

The results are consistent in sign and significance (some difference in point estimates is obviously expected).

These regressions that determine relationships between the energy market and ambient air pollution allows us to produce counterfactual scenarios. How would pollution be different if for a given year, natural gas prices were to increase from $8 to $16? For the year of 2013, this is done in Figure 1-7. The figure illustrates the increase in pollution levels for the year of 2013 if the nationwide natural gas price were to increase from $8 per thousand cubic feet to $16 per thousand cubic feet.

1.4.2 Predicting power plant fuel prices

The prediction problem detailed in the previous sections is to predict fuel prices at the power plant level for coal and natural gas power plants. Given a power plant (that uses coal or natural gas) in a given month and year, we want to predict the price of fuel. The problem is constructed in a supervised learning fashion, with a set of predictive variables and associated coal and natural gas prices. The quantity being predicted y_i is the coal or natural gas price at a given power plant, month, year. Two separate learning processes are run on both coal and natural gas prices. The variables for both processes are discussed in the data section. We can separate the set of predictive variables into two: categorical variables and continuous variables. Month, year, and state of the power plant are considered categorically, whereas net generation, consumption, and other price variables are continuous.

Given the data and methods discussed in the previous section, the data is initially transformed in the following way: (1) The categorical variables are encoded into indicator columns (one-hot vectors in ML literature, FEs in economics literature), (2) The continuous variables that have a skewed distribution are transformed using power transformations (specifically the box-cox transformation). The price variable that is being predicted is transformed using a logarithm.

After the predictive set of variables is constructed and transformed, the learning process is started. A set of regressions are tested using kNN, XG-Boost, Random Forest, Ridge, LASSO, and Elastic Net for both coal and natural gas prediction. Af-

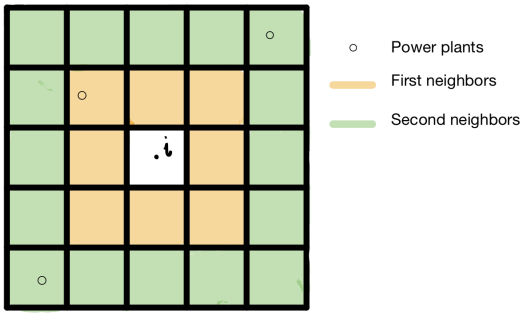

Figure 1-6: Visualization of a grid and its 1^{st} and 2^{nd} neighborhood..

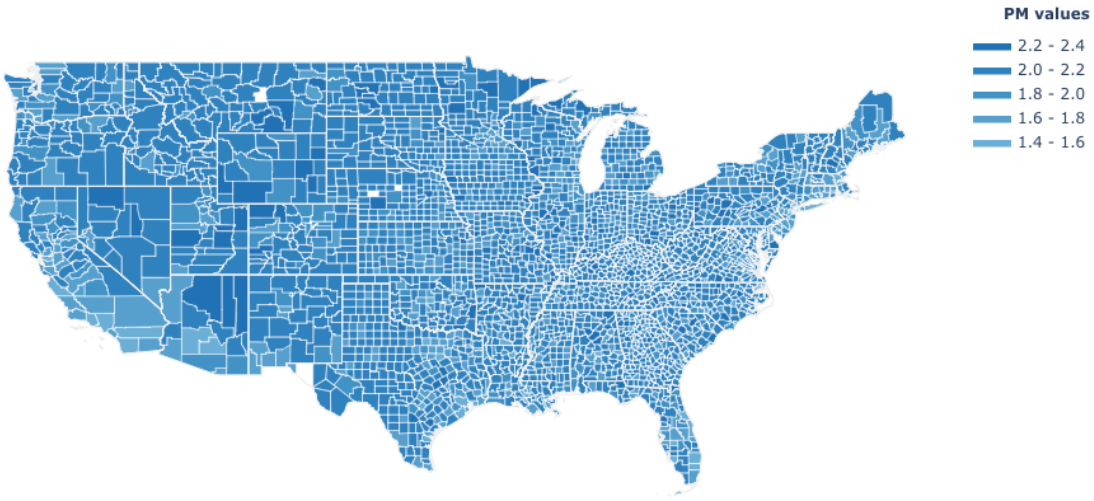

Figure 1-7: Change in Counterfactual $PM_{2.5}$ concentrations when the price of natural gas is increased from \$8 to \$16 per Thousand Cubic Feet NG for the year of 2013

ter comparing the cross validation score, Random Forest regression is chosen for coal price prediction and XG-Boost is chosen for natural gas price prediction. For each regression method, a set of hyperparameters is also tested to optimize performance. Each result is evaluated using a 5-fold cross validation structure.

The average cross validation score for the logarithm of coal price prediction is 0.006, which is equivalent to an average 0.24 cross validation MSE of coal price. The MSE on the whole dataset using the final configuration is 0.002 for the logarithm of coal price prediction. The optimal model is Random Forest model with the following hyperparameters:$\{n_{estimators} : 300, max\ features : 50, min\ samples\ split : 6\}$. The resulting fit on the whole dataset can be seen in Figure 1-8.

The average cross validation score for the logarithm of natural gas price prediction is 0.05, which is equivalent to an average 2.6 cross validation MSE of natural gas price. The MSE on the whole dataset using the final configuration is 0.03 for the logarithm of natural gas price prediction, showing that there isn't much overfitting. The optimal model is the XG-Boost model with the following hyperparameters:$\{n_{estimators} : 600, max\ depth : 6\}$. The resulting fit on the whole dataset can be seen in Figure 1-9.

As can be seen from the MSE values of the two predictors, the algorithm does a better job predicting coal prices than natural gas prices. A lot of this can be attributed to the difficulty in capturing the spikes in natural gas prices. The reason for using the predicted values is that they are better than the data we have. Compared to the current state level data we have, the predicted data is able to capture the heterogeneity much better. The MSE of using the state level monthly data for natural gas is 24.6. The natural gas prediction and coal prediction increases our accuracy when trying to obtain plant level data.

1.4.3 Grid level analysis

Finally, we are interested in getting similar tables to those produced in the county level analysis section. In this section, we will produce tables similar to 1.6 and 1.4 and discuss the implications. Initially, we can produce these two tables for the year

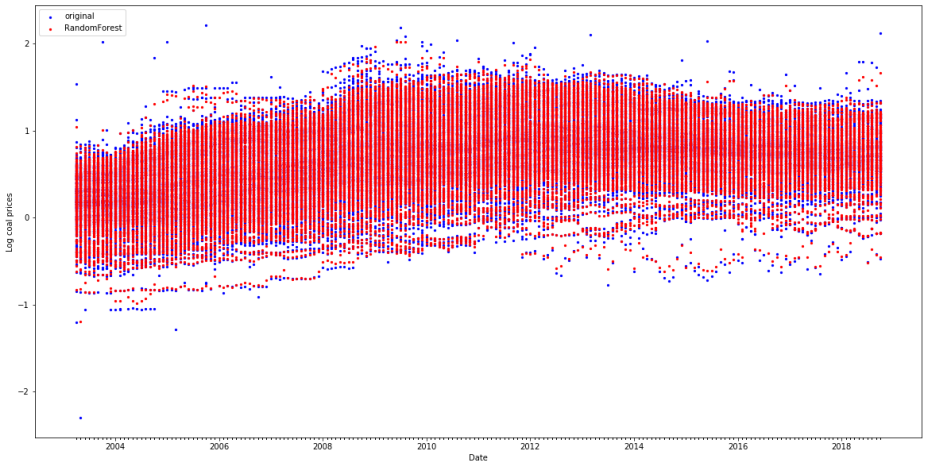

Figure 1-8: Predicted coal prices compared to original coal prices.

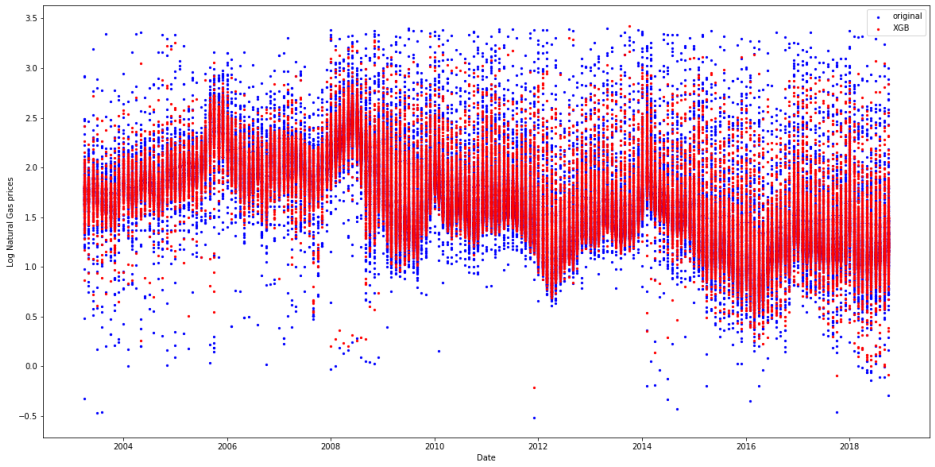

Figure 1-9: Predicted natural gas prices compared to original natural gas prices.

range 2003 to 2016. The grid size for this particular analysis is 0.1 latitude by 0.1 longitude, with 4 levels of neighborhoods considered. Latitude and longitude vary based on location, but as a ballpark estimate one can think of these values being equivalent to $\sim 10\ km$. Which gives a rough radius of (with the neighborhood) of $\sim 90\ km$. Once this grid size and neighborhood is chosen, power plants and their associated characteristics (net generation, consumption, predicted price, capacity) are allocated to the bins for a specific month and year. Finally, this is transformed into a panel data format for analysis.

Tables 1.7 and 1.8 produce the grid equivalent of the county level analysis figures. Before discussing the results, there is some merit in discussing the differences in the variables used in the analysis besides the geographic unit being a grid instead of a county. The capacity variable in this regression is weighted with respect to its level of neighborhood. The idea being here that the impact of capacity should decay relative to its distance from the central grid point. Additionally, the price variable is the average of the mean price of each neighborhood, if a neighborhood doesn't have a power plant, then it isn't incorporated into the average[9]. Additionally, population isn't considered in the $PM_{2.5}$ values. Considering only grids that have both a coal and natural gas power plant in their neighborhood over the time period between 2003 and 2018 results in over ~ 3 million data points.

Looking at Tables 1.7 and 1.8, we can see that the price impacts are similar to those found in the county level analysis and robust to different configurations of FEs. Additionally, the the implications of capacity also show similar behavior and are significant. Finally, the interactions show interesting results as well, where the coal capacity and coal price as well as the gas capacity and gas price impacts are similar to that found in the county level analysis, but the cross interactions are showing negative and significant impacts. It is hard to conclude much on the mixed

[9]A quick example of this can be explained: say we are considering the mean price for grid i with 2 neighborhoods. Assume that within grid i there is one coal power plant $coal_1$ and another coal power plant in the 2^{nd} neighborhood labelled $coal_2$. Assume that there is only one natural gas power plant in the 1^{st} neighborhood: ng_1. Then the mean coal price considered for grid point i is $\frac{coal_{1,price}+coal_{2,price}}{2}$ and natural gas price $ng_{1,price}$ and the coal capacity considered will be $coal_{1,capacity} + coal_{2,capacity}/3$ and natural gas capacity considered will be $ng_{1,capacity}/2$

interactions since they are not as easy to interpret. A final analysis is run on net generation using all grid points ($\sim$ 38 million points), where the log of the weighted net generation of a grid (calculated similar to the weighted capacity) is considered in Table 1.9, the results are similar to that found in the county level analysis.

1.5 Conclusion

In conclusion, in this chapter we are able to use the fluctuations caused by hydraulic fracturing in US energy prices as well as US power plant fuel usage to quantify the relationship between energy prices, power plant fuel usage, and ambient air pollution. Initially, we do this on a county level, that shows the impact of state level price data and county level capacity on county level pollution. Next, in order to run a more granular analysis, we predict plant level fuel prices and us this information to analyze impacts on a grid level. The results are similar to the county level results and show the robustness of the price impact between the years 2003-2016.

VARIABLES	(1) Weighted PM	(2) log(Weighted PM)	(3) hsin(Weighted PM)
Coal Price	-0.514***		
	(0.0456)		
NG Price	0.143***		
	(0.00530)		
log(Coal Price)		-0.0505***	
		(0.0116)	
log(NG Price)		0.150***	
		(0.00944)	
arcsinh(Coal Price)			-0.0334***
			(0.00900)
arcsinh(NG Price)			0.148***
			(0.0101)
Constant	7.605***	1.837***	2.309***
	(0.111)	(0.0278)	(0.0354)
Observations	522,144	522,144	522,144
R-squared	0.632	0.670	0.669

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 1.1: Understanding the impact of prices on pollution. We can see that the results are robust to different transformations

VARIABLES	(1) Weighted PM	(2) log(Weighted PM)	(3) hsin(Weighted PM)
Coal Net Gen	1.31e-06***		
	(1.41e-07)		
NG Net Gen	-3.34e-07**		
	(1.47e-07)		
log(Coal Net Gen)		0.00309***	
		(0.000541)	
log(NG Net Gen)		-0.000114	
		(0.000492)	
hsin(Coal Net Gen)			0.00267***
			(0.000473)
hsin(NG Net Gen)			-0.000222
			(0.000398)
Constant	8.160***	2.141***	2.704***
	(0.00779)	(0.00138)	(0.00119)
Observations	596,736	588,945	596,736
R-squared	0.603	0.652	0.649

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 1.2: Understanding the impact of net generation on pollution. We can see that the results are robust to different transformations

VARIABLES	(1) Weighted PM	(2) log(Weighted PM)	(3) hsin(Weighted PM)
Coal Cons	1.33e-07***		
	(1.45e-08)		
NG Cons	-4.88e-08***		
	(1.82e-08)		
log(Coal Cons)		0.00250***	
		(0.000408)	
log(NG Cons)		-0.000249	
		(0.000371)	
hsin(Coal Cons)			0.00880***
			(0.00180)
hsin(NG Cons)			-0.00146
			(0.00139)
Constant	8.158***	2.142***	2.705***
	(0.00855)	(0.00135)	(0.00138)
Observations	596,736	596,736	596,736
R-squared	0.603	0.651	0.649

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 1.3: Understanding the impact of consumption on pollution. We can see that the results are robust to different transformations

VARIABLES	(1) log(Weighted PM)	(2) log(Weighted PM)	(3) log(Weighted PM)
log(Coal Price)	-0.0453***	-0.0547***	-0.0903***
	(0.0132)	(0.0104)	(0.00839)
log(NG Price)	0.287***	0.0374***	0.0264***
	(0.00506)	(0.00709)	(0.00716)
Observations	522,144	522,144	522,144
R-squared	0.580	0.627	0.638
County	YES	YES	YES
Year	YES	YES	YES
Month-of-Year	NO	YES	YES
County Trends	NO	NO	YES

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 1.4: Understanding the impact of different FEs of the results of Table 1.1. We can see that the results are robust to different FE combinations.

VARIABLES	(1) Weighted PM	(2) Weighted PM	(3) Weighted PM	(4) log(Weighted PM)
Log(Coal Price)	-0.498***	-0.499***	-0.462***	-0.0513***
	(0.0851)	(0.0856)	(0.0832)	(0.0118)
Log(NG Price)	1.627***	1.620***	1.674***	0.159***
	(0.0819)	(0.0819)	(0.0857)	(0.00962)
Log(Coal Capacity)		0.0878***	0.186***	0.0192***
		(0.0135)	(0.0524)	(0.00542)
Log(NG Capacity)		0.00576	0.153***	0.0164***
		(0.0142)	(0.0482)	(0.00487)
Log(Coal Cap * Coal Price)			-0.174***	-0.0117***
			(0.0250)	(0.00184)
Log(Coal Cap * NG Price)			0.0398*	7.01e-06
			(0.0222)	(0.00205)
Log(NG Cap * Coal Price)			0.0407*	0.00637***
			(0.0225)	(0.00183)
Log(NG Cap * NG Price)			-0.0730***	-0.00875***
			(0.0190)	(0.00178)
Observations	522,144	522,144	522,144	522,144
R-squared	0.628	0.629	0.629	0.671
County	Yes	Yes	Yes	Yes
Year	Yes	Yes	Yes	Yes
Month-of-Year	Yes	Yes	Yes	Yes
Month-Year	Yes	Yes	Yes	Yes

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 1.5: Analyzing the impact of Coal and Natural Gas Prices and Capacity on county level, weighted $PM_{2.5}$ concentrations. The capacity values are total county values. All standard errors are robust, county level clustered standard errors. The regression is a yearly fixed effect estimator.

VARIABLES	(1) log(Weighted PM)	(2) log(Weighted PM)	(3) log(Weighted PM)
Log(Coal Price)	-0.0505***	-0.0505***	-0.0513***
	(0.0116)	(0.0116)	(0.0118)
Log(NG Price)	0.150***	0.149***	0.159***
	(0.00944)	(0.00944)	(0.00962)
Log(Coal Capacity)		0.00519***	0.0192***
		(0.00104)	(0.00542)
Log(NG Capacity)		0.000570	0.0164***
		(0.00109)	(0.00487)
Log(Coal Cap * Coal Price)			-0.0117***
			(0.00184)
Log(Coal Cap * NG Price)			7.01e-06
			(0.00205)
Log(NG Cap * Coal Price)			0.00637***
			(0.00183)
Log(NG Cap * NG Price)			-0.00875***
			(0.00178)
Observations	522,144	522,144	522,144
R-squared	0.670	0.670	0.671
County	Yes	Yes	Yes
Year	Yes	Yes	Yes
Month-of-Year	Yes	Yes	Yes
Month-Year	Yes	Yes	Yes

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 1.6: Analyzing the impact of Coal and Natural Gas Prices and Capacity on county level, $PM_{2.5}$ concentrations. The capacity values are total county values. All standard errors are robust, county level clustered standard errors. The regression is a yearly fixed effect estimator.

VARIABLES	(1) log(PM)	(2) log(PM)	(3) log(PM)	(4) log(PM)
log(NG Price)	-0.0372***	-0.0325***	-0.0395***	-0.0431***
	(0.00155)	(0.00146)	(0.00150)	(0.00124)
log(Coal Price)	0.0405***	0.0796***	0.0446***	0.0712***
	(0.00121)	(0.00101)	(0.00117)	(0.00105)
Observations	3,178,529	3,178,529	3,178,529	3,178,529
R-squared	0.552	0.626	0.669	0.635
County	YES	YES	YES	YES
Year	YES	YES	YES	YES
Month-of-Year	NO	YES	YES	YES
Month-Year	YES	YES	YES	NO
County Trends	NO	NO	NO	YES

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 1.7: Understanding the impact of different FEs using the grid level data. We can see that the results are robust to different FE combinations.

VARIABLES	(1) log(PM)	(2) log(PM)	(3) log(PM)
log(NG Price)	-0.291***	-0.0400***	-0.0219***
	(0.0156)	(0.00150)	(0.00270)
log(Coal Price)	0.542***	0.0440***	0.0560***
	(0.0130)	(0.00117)	(0.00219)
Log(Coal Capacity)		0.00951***	0.0138***
		(0.000445)	(0.000833)
Log(NG Capacity)		-0.000194	0.00410***
		(0.000496)	(0.000793)
Log(Coal Cap * Coal Price)			-0.00174***
			(0.000486)
Log(Coal Cap * NG Price)			-0.00152***
			(0.000355)
Log(NG Cap * Coal Price)			-0.00243***
			(0.000434)
Log(NG Cap * NG Price)			-0.00131***
			(0.000290)
Observations	3,178,926	3,178,529	3,178,529
R-squared	0.633	0.669	0.669
County	Yes	Yes	Yes
Year	Yes	Yes	Yes
Month-of-Year	Yes	Yes	Yes
Month-Year	Yes	Yes	Yes

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 1.8: Analyzing the impact of Coal and Natural Gas Prices and Capacity on grid level, $PM_{2.5}$ concentrations. The capacity values are weighted by neighborhood. All standard errors are robust, grid-level clustered standard errors.

VARIABLES	(1) log(PM)
log(Coal Net Generation)	0.0141***
	(0.000476)
log(Natural Gas Net Generation)	-0.000503*
	(0.000293)
Observations	37,587,012
R-squared	0.604

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 1.9: Analyzing the impact of Coal and Natural Gas net generation on grid level, $PM_{2.5}$ concentrations.

Chapter 2

Network Properties of fossil fuel trade

Global energy trade -mostly conducted through fossil fuels such as coal, natural gas, and oil- has a large effect on geopolitics. Understanding the dynamics of the energy trade network, as well as some of its properties, can help inform energy experts and policy makers. In this chapter, global fossil fuel trade data is used to create a network. This network is then analyzed using different network and statistical methods to understand its structure. Results show that contrary to previous claims in the literature, the fossil fuel trade does not follow scale free distribution. Furthermore, a weighted Hyperlink-Induced Topic Search (HITS) algorithm was found to best predict global rankings of major exporters and importers. The InfoMAP statistical tool was then used for community detection in the trade networks, and showed the change (over time) in clusters apparent in natural gas networks. Finally, the energy trade network was analyzed to find how the introduction of hydraulic fracking has affected it, primarily the US.

2.1 Introduction

Networks are ubiquitous in many fields and using tools developed in the Network Science community can help us better understand a given problem's structure. In this section, we will look at trade networks, primarily fossil fuel trade. The network we construct will be weighted and directed, where countries represent nodes and edges represent trade flows. The quantity that is flowing is analyzed in four different settings: coal, natural gas, oil, and the aggregation of all fuel types. Visualizations of the model as well as basic parameters such as average degrees and total number of nodes will be displayed for each year. Further analysis was conducted to see if the networks follow a scale free distribution. Also, potential ranking methods were tested and community detection methods were analyzed.

The United Nations (UN) Comtrade data was used for this project. It is considered to be the most comprehensive and widely used global trade data set. The UN Comtrade data reports on trades between ISO 3 territories. There is a total of 248 territories defined by the ISO 3 standard. We will be calling these territories countries for the sake of clarity. It is also important to note some of the limitations associated to this data before the analysis.

The data is collected from countries based on willingness to participate and report. This means that a country only discloses trades that it is willing to share. This can create a problem overall because of the possibility of inaccurate data. Especially in the highly geopolitical context of energy trade, there is almost surely some discrepancies associated with the data. We are assuming that they are negligible for the sake of this work.

The subset of the data considered in this project comprises of annual fossil fuel trade between the years 2000 and 2016. This data is further processed to determine trades specific to the three main fossil fuels traded: coal, natural gas, and oil. The annual data can be used to create a weighted-directed-graph, where the weights can be specified in two different metrics: US Dollars and Weight in kg. The analysis for this data is conducted in Python. The network was built using NetworkX. Other packages and frameworks were used for the analysis that can be found on the GitHub

repository.

Previous work combining Network dynamics with Fossil Fuel trade is limited to a few papers. All papers use the UN Comtrade data. There is a wider body of literature associated to trade in general, which we will discuss in the next chapter. The few papers that relate fossil fuel trade to networks have some questionable analyses that we will discuss later in this paper. The three main papers on this topic are:

1. Zhong et al. (2015) analyze the fossil fuel trade network looking at network properties. They initially calculate a new parameter for the weights on their networks which they call "emergy". This is to better understand the amount of energy that a certain country is importing and exporting. Concentrating on all three energy types, they analyze different network properties as well as trade relations.

2. Geng et al. (2014) analyze the natural gas trade network and how it has evolved over time. The analysis is for the periods between 2000-2011. They use a minimum spanning tree model to analyze natural gas integration. They conclude that improvements of market integration will promote globalized trade.

3. Lastly, Kaya and Eren (2016) discuss fossil fuel trade using a complex networks approach. They analyze metrics such as connectivity, clustering, assortativity, centrality, degree distribution, and ranking methods.

The rest of this chapter is as follows: Section 2.2 discusses the methods that will be used in the analysis, Section 2.3 discusses the results from using these methods, and Section 2.4 concludes the chapter by summarizing the results.

2.2 Network Methods

This section discusses the properties and analyses that will be used on the data. Initially, basic network properties are studied to better understand the simple structure of the data at hand. Afterwords, more complex network analyses are conducted:

choosing a ranking algorithm, assessing if the data follows scale free properties, and choosing a community detection algorithm. We will briefly discuss these methods before moving on to the results.

Initially we can define simple measures for the network such as in-degree and out-degree measures for a given node. Before doing so, we can define some terminology. For our setup, let the network for year t be G_t with V_t representing the set of nodes and A_t representing the weighted adjacency matrix. For the sake of notation assume that the edge weight going from nodes u to v in G_t is $A_t[u, v]$, which is always nonnegative, and that this value is 0 if there is no edge. The weight of an edge can either be US Dollars or kilograms. For a given node $v \in V_t$, the in-degree is:

$$d_{in}(v) = \sum_{\forall j} A_t[j, v] \tag{2.1}$$

and out-degree:

$$d_{out}(v) = \sum_{\forall j} A_t[v, j] \tag{2.2}$$

Next we can discuss scale-free distributions. A network is considered to have scale-free degree distribution if the fraction of nodes that have degree k scales approximately as $k^{-\gamma}$, where γ is typically in the range $2 < \gamma < 3$. Scale-free distributions have been thought to occur in many different settings and are usually sought after for desirable properties. Testing if a given distribution is scale-free has become more nuanced over time, with increasing rigor being required in making such claims about a network. The approach used in this chapter is that found in Clauset, Shalizi, and Newman (2009) and will be discussed in the results section.

We will next look at ranking algorithms, which initially became common for internet search results. HITS was one of the first algorithms used to do this. The algorithm was developed by Jon Kleinberg and identifies two types of nodes: hubs versus other authorities. Initially designed for the internet, it does not accommodate for weights in its original setup. The key equations are:

$$c = \alpha A^T h$$
$$h = \beta A^T c \tag{2.3}$$

where c and h are scoring vectors and α and β are constants. PageRank is another famous ranking algorithm that was introduced in Google's initial search engine. The associated equation is:

$$w = P^T w \tag{2.4}$$

where w is a scoring vector and $P_{ij} = A_{ij}/d_{out}(i)$.

The final method we use is the InfoMAP community detection algorithm. This is used because it can accommodate weighted-directed networks. The *map equation* that is used in InfoMAP will not be discussed in this section[1].

2.3 Results

This section presents properties and analyses conducted on the data from a network science perspective. We use the methods discussed in the previous section to look at the fossil fuel trade network, and discuss the results for each method.

2.3.1 Network Properties

Initially, basic features of the networks are analyzed for the three fuel types as well as to analyze how it has changed over time. This includes properties such as number of nodes, average degree, and simple graph visualizations. Three different fuel types are analyzed over the 17 years of data to generate average node and degree figures.

Figure 2-1 (a) shows how the total number of nodes has changed over time in the natural gas, coal, and oil networks. The blue line close to 250 is the maximum number of total nodes there can be (based on the number of countries/territories). We can see that a large portion of all territories participate in some manner of trade.

[1]For more information: https://www.mapequation.org/apps/MapDemo.html

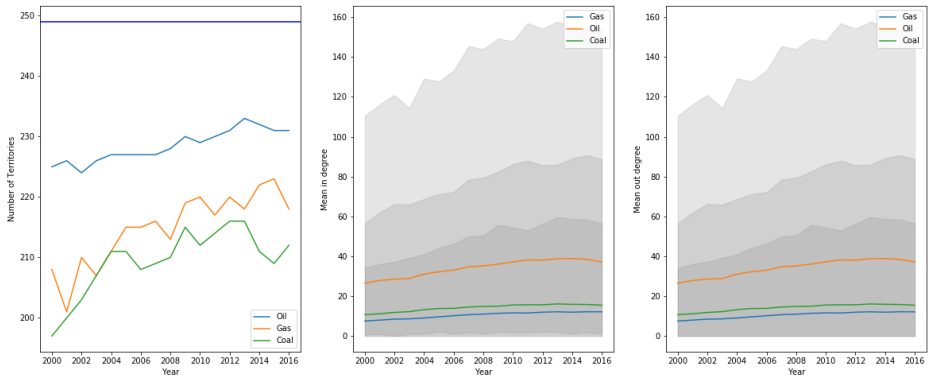

Figure 2-1: This figure is a simple visualization of the number of territories connected in each year (figure to the left), average in degrees for a given node in each year (figure in the middle), and average out degrees for a given node in each year (figure on the right). Note the Shaded areas in the two figures on the right are the 5 to 95 percentiles.

As expected, oil trade is the most wide spread, with natural gas and coal being traded between fewer countries.

Figure 2-1 (b) and (c) show the average degree distribution of networks. This is very informative about fossil fuel trade. We observe the high connectivity of oil trade as well as coal. This high density has a lot to do with the competitiveness of the products. Easily transportable (as well as cost effectively), geography does not play such a constraint as it does with natural gas. Additionally, we can see that in and out degrees have high variation, as can be seen in the 5% and 95% percentiles shaded in the Figures. One final observation on Figure 1 is that even though coal (in the later years) has fewer countries trading it, it seems to involve more trade happening between those countries (compared to natural gas).

These figures help give a good intuition about fossil fuel networks:

1. Oil is the most commonly traded fuel among nations.

2. Gas trade has seen a steady increase in trade over time.

3. Coal has a higher density overall, but isn't traded as commonly.

2.3.2 Degree Distribution

The degree distribution of a network is of high interest from a network science perspective. Depending on the application, such properties can hold favorable or unfavorable properties. Yet the analysis for determining if a network follows a power law distribution is often not conducted properly. When carrying out a literature review on network science properties of fossil fuel trade the results showed exactly this. The three main papers all concluded that fossil fuel networks follow power law distributions.

Zhong et al. (2015) conducted a linear regression on log-log axes to conclude that the degree distribution followed a power law. Geng et al. (2014) conducted the same analysis to conclude that natural gas trade networks followed power law distributions. Kaya and Eren (2016) conducted a more rigorous analysis using KS Statistics. They concluded that fossil fuel networks followed a power law distribution because the hypothesis could not be rejected for p-values above 0.05.

For this section, the approach detailed by Clauset, Shalizi, and Newman (2009) is used. They suggest a three-step process that starts with estimating the parameters x_{min} and α, followed by a goodness-of-fit test between the data and the power law, using KS Statistics. The last step, which is what is detailed in this section is to compare the hypotheses to an alternative distribution using a likelihood ratio test. The alternative distribution that is chosen is the exponential distribution (the power law distribution is the numerator and the exponential distribution is the denominator for the loglikelihood ratio test). The reason for choosing exponential is that to be considered as a power law distribution, being comparatively better than exponential is a bare minimum. This is because of the fact that exponential distribution is a prototypical distribution that is very different from a power-law distribution.

This analysis was conducted using the package detailed in Alstott et al. (2014). Figure 2-2 details the comparison results. The figure details the log likelihood ratio and the associated probability ratio of the two distributions. Each row in Figure

2-2 represents the analysis run on a different network (natural gas, oil, coal, and the combined fossil fuel trade). Finally, the x-axis in each plot represents the year and the y-axis represents the log likelihood ratio/probability ratio (p-value) calculated.

In the first column of Figure 2-2, the null hypothesis is that the dataset is an observation of a power law distribution and the alternative hypothesis is that it is part of an exponential distribution. The first column shows the loglikelihood ratio results, where a value that is positive would suggest that the data is more likely to be from a power law distribution and a value that is negative would suggest that the data is more likely to be from an exponential distribution. In order to quantitatively measure our certainty towards one distribution (we want to be more confident in believing a distribution is exponential if we calculate a loglikelihood of -10 versus -1 and more confident in believing it is power law if the loglikelihood is 10 versus 1) the associated p-value is calculated in the second column. Using the sign of the loglikelihood ratio results, the p-value in column 2 indicates our certainty that the results are more like one distribution than the other. Note that this means the p-value for two years can have a different null hypothesis if the loglikelihood values have different signs for those two years.

Looking at the results in Figure 2-2, it can be seen that for oil and combined fossil fuel trade networks, an exponential distribution is a better hypothesis than a power law distribution. The first column in rows 2 and 4 show consistently negative numbers that indicate that an exponential distribution is a better fit and these p-values show that we can be fairly confident in the results. Next, we can look at the results of the coal networks that are shown in row 3. The loglikelihood can be seen to be mostly negative, but the magnitude of these values are not as high as in the oil and combined settings. This is then reflected in the p-values that show more uncertainty in the loglikelihood results. Finally, we can look at the natural gas results, which seem to not indicate one distribution or another, with the p-values never giving much confidence. The results show significant ambiguity in the natural gas networks, while giving strong evidence to reject the hypothesis that oil and the combined fossil fuel networks follow scale free distribution. Even in the natural gas networks, there is no evidence suggesting that the network follows power law properties.

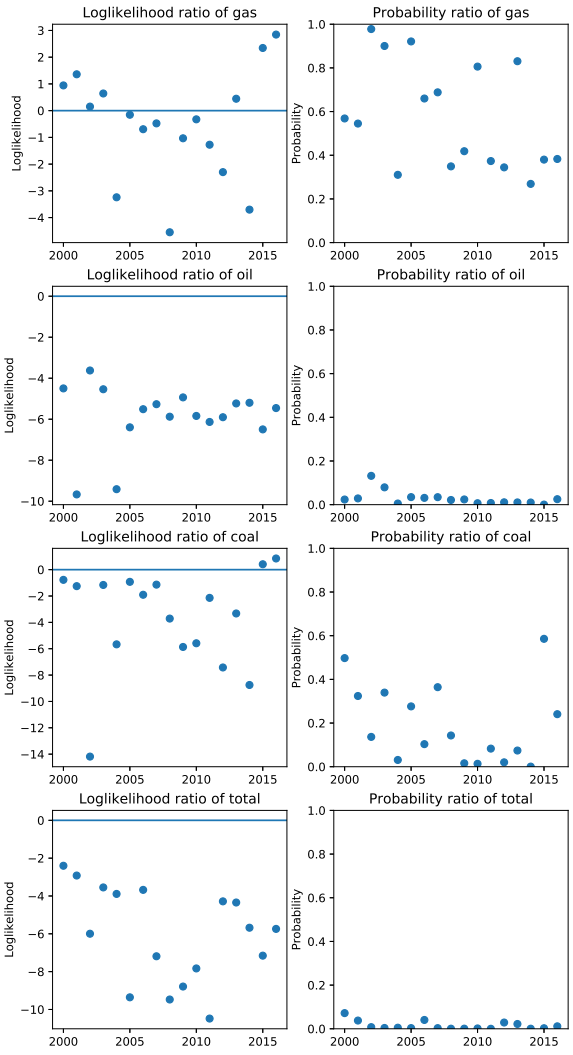

Figure 2-2: Plots of loglikelihood ratio between power law and exponential distributions.

2.3.3 Ranking Methods

Ranking methods in networks is an interesting and widely researched topic. It could also be valuable for trade networks, yielding a hierarchical ranking of countries based on their importance. The measure could help determine a benchmark to understand the role a certain country plays in world fossil fuel trades as well as how their ranking has changed over time. When picking algorithms, it was important that it would be a good fit for our network. This is, the ranking algorithm needed to account for the weights and the directedness of the network. From the different types of ranking algorithms, this project tested three: pageRank, springRank, and HITS.

PageRank is famously used by Google's initial search algorithm. A weighted version was used to create our ranking algorithm. SpringRank is a newer algorithm detailed in Bacco et al. (2017). This algorithm is designed for directed, weighted networks and assumes that interactions are more likely to occur between nodes with similar ranks.

The HITS algorithm has been modified to incorporate weights. The modifications made in order to make the algorithms incorporate weights are given in Deguchi et al. (2014). For the weighted HITS, we replace the A matrix (adjacency matrix) given in equation (3) with the weighted adjacency matrix (W).

After running pageRank (with and without edge weights), HITS (with and without edge weights), and the springRank algorithm, the results showed that the weighted HITS algorithm is best suited for fossil fuel trade networks. When comparing pageRank and HITS, Deguchi et al. (2014) found similar results for other trade networks.

A trivial analysis of the rankings made it apparent that weighted HITS was the most accurate in ranking countries in global energy trade. The rankings from the weighted HITS are given in the appendix. SpringRank's assumption that interactions are more likely to occur between nodes with similar ranks does not match well with fossil fuel trade, indicating that high ranking countries like Saudi Arabia and Russia are likely to be trading with each other. SpringRank's algorithm ranked the territory of Bouvet Island as the highest ranked node. The results did not follow any logical ordering and were left out of the consideration. PageRank does a bit better, but

Oil		Coal	
Hub	Authority	Hub	Authority
Canada	USA	Australia	Japan
Saudi Arabia	China	Indonesia	China
Russia	Japan	Russia	India
Iraq	India	Canada	South Korea

Figure 2-3: Top 4 Hub and Authorities in the Oil and Coal networks.

has discrepancies such as ranking of countries such as Singapore and Netherlands very high. This is interesting because these two countries are considered to be high intensity/frequency ports. But it does not give us any distinction between port countries, high exporting countries, and importing countries. The HITS ranking seems able to separate producers (hubs) and consumers (authorities) in its ranking. Figure 2-3 shows the top 4 hubs and authorities for oil and coal. The oil table on the left shows the ranking for before the shale revolution occurred. This result is very intuitive. The coal table results are less intuitive because the coal trade is less well known. The table reveals an interesting structure of coal networks without having domain expertise.

2.3.4 Community Detection

The initial motivation for running community detection algorithms on fossil fuel trade networks was that it could be a method of determining regional/global partnerships between countries. Separating them into clusters would be a good method of understanding geopolitical partnerships linked through fossil fuel trades. Yet, the results showed that there were very few communities to be detected and also that there are very few algorithms available for detecting communities in directed, weighted networks. Still, some interesting results persisted.

Reading the review by Lanchichinetti and Forunato (2010), it was decided that the only viable option for directed networks is InfoMAP. InfoMAP's algorithm opti-

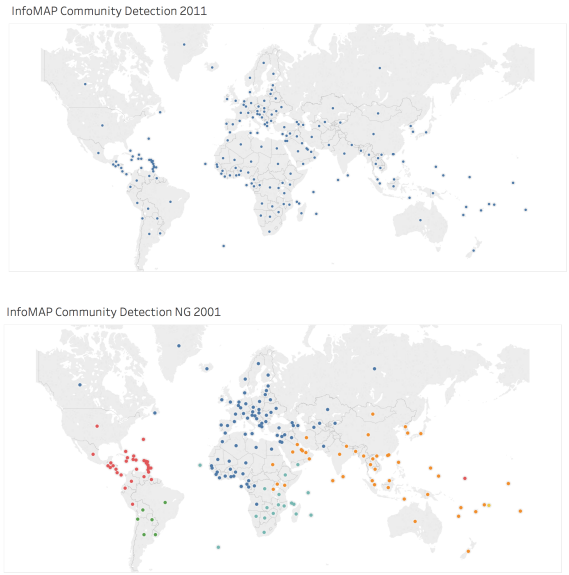

Figure 2-4: InfoMAP Community detection for 2011 (on top) and 2001 (on bottom)

mizes what it calls the map equation, which is detailed in Rosvall et al. (2009).

The results showed that over the past 16 years there has been minimal or no community structure in oil or coal trade networks. This can mostly be attributed to the ease of transporting (oil) and abundance of extracting (coal) the two commodities, leading to highly dense graphs.

This makes it hard for the formation of any communities. This was a good affirmation that the algorithm was working properly. For natural gas trading, the results were more interesting. The community structure of natural gas trading seems to greatly reflect how the industry has evolved. It is becoming less dependent on geographies and more global (with increasing LNG terminals). The difference in community detection between 2001 and 2011 can be seen in Figure 2-4.

2.3.5 Shale Gas & Network Properties

Before going into the impacts of hydraulic fracturing, lets briefly discuss why it is important. Hydraulic fracturing, also known as shale, is a method of excavating gas

and oil, as discussed in the first chapter. It gained immense popularity over the past decade. This has led to massive amounts of natural gas production in the US.

This explosion of fracking is also interesting in another way: it has been a US-centric phenomenon, mainly due to three reasons:

1. There have been environmental concerns due to the process of hydraulic fracturing. This has led to negative public policy in some regions.

2. Property rights in many shale rich states permit horizontal drilling which decreases the market price of leasing land to drill.

3. It is a water intensive process. If the opportunity cost of using water is too high, it isn't viable.

Fracking has affected the composition of the electricity grid as well. Natural gas makes up 30% of the total electricity production compared to 20% just 10 years ago. Such structural changes are unprecedented in the energy sector in such a short time period. This effect has propagated to international fossil fuel trade as well, making the US a major producer of oil and natural gas.

Using the network properties as well as the ranking methods discussed previously, we will discuss some of the effects of shale.

Exports & Imports

Figure 2-5 visualizes US exports and imports over time. We can make two speculative interpretations of this figure:

1. We can see that coal exports are starting to increase around 2008. We see a peak around 2011 and then decreases afterwards. We also see a not so dramatic, but still gradual decrease of coal imports over time. One interpretation of the exports can be the following: with shale starting to play a prominent role in 2008, US coal producers found the US market prices to be below their marginal cost. This led to an increase in US exports, up until 2011 when the exports started to decline, quite dramatically after the year 2012. After 2012, total

Figure 2-5: Historical Natural Gas and Coal Export/Imports in the US

production decreased substantially in the US, potentially causing coal to hit a critical point of not meeting marginal cost for exporting.

2. Natural Gas: This graph is very much in line with what we would expect. Around 2008, US imports of gas decrease. We see less of a dramatic change in exports because of the geographic constraints associated with natural gas. We should expect to see increases in export over time as well as dramatic changes if transportation technology is drastically changed.

Trade Partners

When it comes to trading partners that the US has had for coal and natural gas, we can look at Figure . This figure is mostly as expected. Certain discrepancies that can be discussed are:

1. In 2004, we see a jump in US coal export partners. This is due part to the Asian market opening up to the US. The details of this are given in the 2004 Coal report by the EIA.

2. There is a significant jump in trade partners of coal between 2000-2001. We have not found a good explanation for this phenomenon.

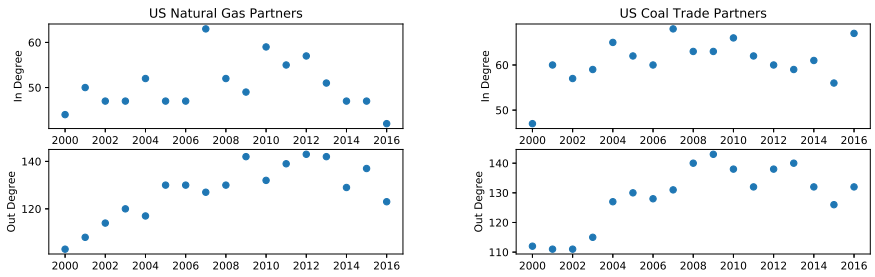

Figure 2-6: Historical Natural Gas and Coal US Trading Partners (by number)

Rankings

Lastly, the algorithm associated to the US were analyzed. Using the rankings from the weighted HITS, Figure 2-7 was created. It shows how the US rank has changed over time and visualizes these changes that have been ensuing since the late 2000s. Figure 2-7 (a) shows the authority ranking of the US. This represents the importance of a node in a network as an authority. It can be seen that the rankings of gas and coal are decreasing over time. This would imply less of a central role in the network for these two commodities. This makes sense as the US has been moving towards consuming its own gas and consuming less coal. Figure 2-7 (b) shows the consistent increase in hub ranking the US has had over time. What is interesting to see is its ranking not change for coal. If the ranking algorithm is accurate, this would imply that increased US exports were not enough to move it up the rankings significantly. One hypothesis for this could be: the US replaced coal exported by countries with lower rankings and that countries ranking above the US did the same. This would imply the market potentially losing density and increasing in centrality around a couple nodes.

2.4 Conclusion

In this chapter, fossil fuel trade is analyzed using network science tools. Simple data aggregation was used to visualize basic network properties such as network size and

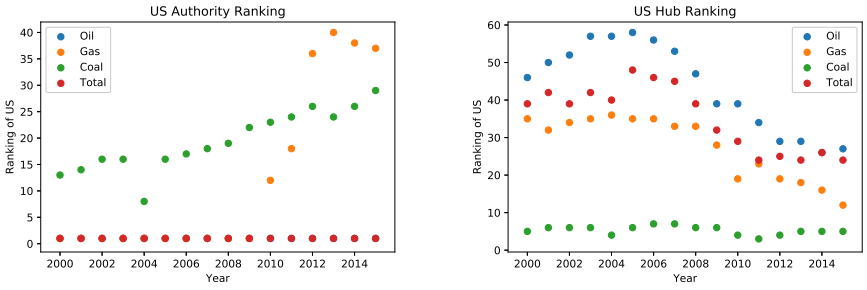

Figure 2-7: Historical US Hub and Authority Rankings

degree properties. The hypothesis that fossil fuel networks follow scale free distributions was shown to fail stronger tests. Several algorithms were tested to create different features of the networks, including ranking algorithms and community detection algorithms. Finally, the network properties and features were used to analyze the impacts of hydraulic fracturing.

Chapter 3

Modelling the Shale Shock: A general equilibrium framework

3.1 Introduction

The previous chapter analyzed the network properties of fossil fuel trade. These analyses were able to demonstrate interesting properties but do not give us insight into causal relationships between changes in fossil fuel trade. These results are useful in determining more formal ways of answering questions such as: (1) Which countries are becoming more influential in global fossil fuel trade? (2) How is clustering of countries changing over time? and (3) What distribution does the trade networks follow/ don't follow? But, they are not able to answer questions related to understanding the impacts of a given event. We now discuss why this is the case.

One might formulate a hypothesis on the impacts of a given shock to certain economic activities such as supply, demand, or trade barriers. If we want to test such a hypothesis, we would need to understand the causal structure of the system. To do so, we must control for the many endogenous components of the system that are interlinked: local demand and production, different factors of production, global trade, the different labor, technology, and natural resources of countries, and many other variables. Additionally, and connected to the previous issue is the high dimen-

sional aspect of trade networks. The time-series information we have (which is in the order of decades) is relatively small relative to the dimensionality of the data for a given year. This makes statistical testing difficult. Compounding this effect is the fact that time-series analysis requires additional understanding of the temporal dimension, which is hard to extract partly due to the two issues of endogeneity and high dimensionality.

All of the reasons mentioned above make inference on trade networks a difficult statistical problem. For this reason, we will stray away from statistical methods in determining the trade impacts of fossil fuel trade. We will move to the literature that covers general equilibrium frameworks, more specifically the international trade literature in economics. In these types of models, an "expert" creates a set of constitutive equations and conservation laws that dictate the model world. The "expert" justifies his/her decisions of these equations and why they would represent a relatively accurate depiction of the simulation that one wants to conduct. Once the "expert" creates and justifies the general equilibrium framework, certain scenarios can be analyzed and interpreted. If the setting of the model is assumed to be representative, then the results can be interpreted as causal.

General equilibrium models tend to work quite well when modelling the physical world, where the equations are known to a high level of certainty. But due to the fact that equations related to the social and economic world are not known with the same certainty, they tend to be scrutinized (with reason) to a higher degree. The assumption of an all-knowing expert that can perfectly model global trade dynamics is of course an extremely strong one, but because of the challenges detailed earlier, which limit us from using statistical methods, it is one that will be made. This of course means that the results cannot always be taken at face value. To heed some of these cautions, we will select a type of model that takes into account the complex behavior of global trade while still having interpretable results.

Traditionally, the international trade literature has focused on Computable General Equilibrium Models (Costinot and Rodríguez-Clare (2014)). These models, mainly aimed to capture the complexity of global markets tend to have two main drawbacks: (1) Their complexity hinders interpretability and (2) many assumptions

are needed for parameters in the equations.

Starting with Eaton and Kortum (2002), there has been a strand of literature that develops simpler model frameworks based on micro-economic foundations and use "gravity equations" as the main structure implemented on bilateral trade. This research is a quantification of Ricardian theory developed by David Ricardo in 1817. This theory describes how two countries are often better off by specializing in the production of a certain good and trading it rather then producing and consuming goods in isolation. The Ricardian theory of comparative advantage is quite old, but has only recently become a common tool for quantitative research in international trade (Eaton and Kortum (2012)). These types of models are often referred to as Structural Gravity Models.

As discussed in Costinot and Rodríguez-Clare (2014), gravity equations, which refer to the structure that is assumed of bilateral trade, is also a relatively old idea dating back to Tinbergen (1962), but this technique was not accepted for a long time since it wasn't founded on economic principles. These gravity equations get their name from physics, since the equation describing to gravity was a point of inspiration in determining bilateral trade structure. The vague idea behind it is that trade relationships between any two countries for a certain good are proportional to economic properties of the country and inversely proportional to the distance between any two countries. Costinot and Rodríguez-Clare (2014) review variations and assumptions of contemporary models that are based on gravity equations. Costinot and Vogel (2015) furthermore review related Ricardo-Roy models that have microeconomic foundations based on multiple factors of production of a certain good. More recently, Farrokhi (2016) uses the frameworks developed by this literature to understand global oil trade and discusses the gains from trade resulting from oil.

As the literature grew, different adaptations of the model framework developed. Most frameworks used counterfactual setups to understand gains from trade by varying trade costs. The method used to set up counterfactual equations is termed "exact hat algebra", and uses data from a baseline year to create a counterfactual equilibrium setup relative to the baseline year. This method helps simplify the equations, by cancelling out many terms that are hard to find data on. This method was first

developed in Dekle et al. (2008). In trade literature, “iceberg trade costs" are usually used as the main source of variation in these counterfactual setups, but these same techniques can be used to analyse the impact from other shocks, such as supply shocks.

When trying to answer questions related to the shale shock in the US, it is important that the model be able to incorporate the unique role that commodities play in the global economy. Fally and Sayre (2018) tries to do exactly this, by creating a two-stage production model with downstream and upstream goods, where the upstream goods represent commodities. The paper uses a counterfactual setup with a shock to “iceberg trade barriers” to evaluate the gains from trade. We will use this model as our main framework and develop similar counterfactual equations, but instead of shocking trade barriers, we will shock commodity supplies in the US and analyze the results. In the next two sections we will set up the model equilibrium and the counterfactual equilibrium, followed by discussing the data the numerical methods used and the results.

3.2 Model

The model discussed in this section is developed in Fally and Sayre (2018). The author's develop a model that at its core incorporates commodities and labor as the two main sources of income and uses final goods as an additional factor of production and international trade with a gravity type structure. The relationship between commodities and final goods creates input-output linkages. This model can incorporate multiple sectors of commodities, final goods, and regions. These will be denoted with subscripts of n or i for regions, g for commodities, and k for final goods, similar to the original paper.

The model sets up demand, cost of production, price indices, bilateral trade equations, and sources of income and looks for an equilibrium. Although this model is based on structural gravity model literature, the authors have tailored the model to incorporate characteristics unique to commodity trade. We refer the reader to their paper for more discussion on the assumptions in setting up the model.

Demand for final good The model has the following quantity of demand in region n for a final good k:

$$D_{n,k} = (P_{n,k}/P_n)^{1-\sigma} a_{nk} E_n \tag{3.1}$$

where $P_{n,k}$ represents price index of final good k in region n, P_n represents the price index of region n, σ is the elasticity of substitution, $a_{n,k}$ is the utility shifter for a given region n and final good k, and E_n is the total income in region n (GDP).

Production Cost of final goods Production of final good k in region i is dependent, by the structure of this model, on wages as well as a set of commodities $G(k)$. Producing final good k in region i will cost:

$$C_{i,k} = A_{i,k} \left[\beta_{i,k,L} w_i^{1-\eta_k} + \sum_{g \in G(k)} \beta_{i,k,g} (P_{i,g})^{1-\eta_k} \right]^{\frac{1}{1-\eta_k}} \tag{3.2}$$

where $A_{i,k}$ is the region final good specific productivity term, w_i is the representative wage in region i, η_k is the elasticity of demand, $P_{i,g}$ is the price index for commodity g in region i.

Price Index for final goods The price index for final goods in a given region is the accumulation of the cost and trade barriers from producing the product in all other regions:

$$P_{n,k} = \left[\sum_i (C_{i,k} \tau_{n,i,k})^{-\theta_k} \right]^{\frac{1}{-\theta_k}} \tag{3.3}$$

where $\tau_{n,i,k}$ is the iceberg trade barrier of importing final good k from region i to region n. We let $\tau = 1$ when $i = n$.

Demand for commodities The production of final goods is the main source of demand for commodity g in region i and industry k, which we will define to be:

$$D_{i,g,k} = \beta_{i,k,g} (P_{i,g}/C_{i,k})^{1-\eta_k} Y_{ik} \tag{3.4}$$

where $\beta_{i,k,g}$ is the factor requirements for commodity g (reflects differences in tech-

nology).

Production Cost of commodities Producing commodities has again constant elasticity of substitution for the factor prices. Production for a commodity is dependent on labor and natural resources. We can define the cost of producing commodity g in region i to be similar to that of producing final goods:

$$C_{i,g} = A_{i,g}\left[\beta_{i,g} r_{i,g}^{1-\rho_g} + (1-\beta_{i,g}) w_i^{1-\rho_g}\right]^{\frac{1}{1-\rho_g}} \tag{3.5}$$

where $\beta_{i,g}$ is a commodity, region specific technology constant and ρ_g is the elasticity of substitution between labor and a commodity. This parameter reflects elasticity of supply.

Price Index for commodity goods The price index for commodity goods in a given region is the same that it is for final goods:

$$P_{n,g} = \left[\sum_i (C_{i,g}\tau_{n,i,g})^{-\theta_g}\right]^{\frac{1}{-\theta_g}} \tag{3.6}$$

Price index of a region The price index of a region is defined as the accumulation of the price index of final goods:

$$P_n = \left[\sum_k a_{n,k}(P_{n,k})^{1-\sigma}\right]^{\frac{1}{1-\sigma}} \tag{3.7}$$

Trade in final goods and commodities The model follows the same convention of trade for final goods and commodities. The trade for good k from region i to region n is given by:

$$X_{n,i,k} = \lambda_{n,i,k} D_{n,k} = \frac{(C_{i,k}\tau_{n,i,k})^{-\theta_k}}{\sum_j (C_{j,k}\tau_{n,j,k})^{-\theta_k}} D_{n,k} \tag{3.8}$$

and similarly trade for commodity g from region i to region n is given by:

$$X_{n,i,g} = \lambda_{n,i,g} D_{n,g} = \frac{(C_{i,g}\tau_{n,i,g})^{-\theta_g}}{\sum_j (C_{j,g}\tau_{n,j,g})^{-\theta_g}} D_{n,g} \tag{3.9}$$

where θ_g and θ_k are the trade elasticities, $\lambda_{n,i,g}$ and $\lambda_{n,i,k}$ are the share of exports which are useful to define for future analysis.

Sources of Income In equilibrium, income for a representative agent from labor is determined by:

$$L_i w_i = \sum_{g \in Commodities} (1 - \beta_{i,g})(w_i / C_{i,g})^{1-\rho_g} Y_{i,g} + \sum_{k \in \text{Final goods}} (\beta_{i,k,L})(w_i / C_{i,k})^{1-\eta_k} Y_{i,k} \tag{3.10}$$

As well as total income from commodities:

$$R_{i,g} r_{i,g} = \beta_{i,g} (r_{i,g} / C_{i,g})^{1-\rho_g} Y_{i,g} \tag{3.11}$$

Which leads to total income (GDP):

$$E_n = L_n w_n + \sum_{g \in Commodities} R_{n,g} r_{n,g} \tag{3.12}$$

Equilibrium conditions are defined by equations (1) - (12). Note that we can additionally define region production of final goods and commodities to be: $Y_{i,k} = \sum_n X_{n,i,k}$ and $Y_{i,g} = \sum_n X_{n,i,g}$. In this setup, the set of given parameters are $\{\sigma, a_{nk}, \beta_{i,k,L}, \eta_k, \beta_{i,k,g}, A_{i,k}, \tau_{n,i,k}, \tau_{n,i,g}, \theta_k, \rho_g, \beta_{i,g}, A_{i,g}, \theta_g, \sigma, R_{i,g}, L_n\}$ and the set of free variables are $\{D_{n,k}, D_{n,g}, C_{i,k}, C_{i,g}, P_{n,k}, P_{n,g}, P_n, X_{n,i,k}, X_{n,i,g}, w_i, r_{i,g}, E_n\}$.

3.3 Counterfactual Setup

In order to study a counterfactual scenario, we are using the technique first introduced in Dekle et al. (2008) which has been termed the “exact hat algebra" method. The observation is that solving the simulation exercise relative to observations you already have can cancel out some of the constants in the equations. To do so, we create a new set of equations that solve for $\widehat{Z} = Z'/Z$, where Z is the value we observe and Z' is the value of variable Z that we are looking to simulate.

The counterfactual equilibrium is given in the following set of equations:

$$\widehat{P_n} = [\sum_k \alpha_{nk} (\widehat{P_{nk}})^{1-\sigma}]^{\frac{1}{1-\sigma}} \tag{3.13}$$

$$\widehat{P_{nk}} = [\sum_i \lambda_{nik} (\widehat{C_{ik}})^{-\theta_k}]^{-\frac{1}{\theta_k}} \tag{3.14}$$

$$\widehat{P_{ng}} = [\sum_i \lambda_{nig} (\widehat{C_{ig}})^{-\theta_g}]^{-1/\theta_g} \tag{3.15}$$

$$\widehat{D_{ng}} = \sum_k d_{nkg} (\widehat{P_{ng}}/\widehat{C_{nk}})^{1-\eta_k} \widehat{Y_{nk}} \tag{3.16}$$

$$\widehat{D_{nk}} = (\widehat{P_{nk}}/\widehat{P_n})^{1-\sigma} \widehat{E_n} \tag{3.17}$$

$$\widehat{Y_{ik}} = \sum_n (X_{nik}/Y_{ik}) (\widehat{C_{ik}})^{-\theta_k} (\widehat{P_{nk}}^{\theta_k}) \widehat{D_{nk}} \tag{3.18}$$

$$\widehat{Y_{ig}} = \sum_n (X_{nig}/Y_{ig}) (\widehat{C_{ig}})^{-\theta_g} (\widehat{P_{ng}}^{\theta_g}) \widehat{D_{ng}} \tag{3.19}$$

$$\widehat{C_{nk}} = [\phi_{ik,L} \widehat{w}_i^{1-\eta_k} + \sum_g \phi_{ik,g} (\widehat{P_{ig}})^{1-\eta_k}]^{\frac{1}{1-\eta_k}} \tag{3.20}$$

$$\widehat{C_{ng}} = [\phi_{ig,R} \widehat{r_{ig}}^{1-\rho_g} + \phi_{ig,L} \widehat{w}_i^{1-\rho_g}]^{\frac{1}{1-\rho_g}} \tag{3.21}$$

$$\widehat{E_n} = e_{iL} \widehat{w}_i + \sum_g e_{ig,R} \widehat{r_{ng}} \widehat{\boldsymbol{R_{ng}}} \tag{3.22}$$

$$\widehat{r_{ng}} = (\widehat{r_{ig}}/\widehat{C_{ig}})^{1-\rho_g} \widehat{Y_{ig}}/\widehat{\boldsymbol{R_{ng}}} \tag{3.23}$$

$$\widehat{w_n} = \sum_g (\phi_{ig,L} Y_{ig}/(w_i L_i))(\widehat{w_i}/\widehat{C_{ig}})^{1-\rho_g}\widehat{Y_{ig}} + \sum_k (\phi_{ik,L} Y_{ik}/(w_i L_i))(\widehat{w_i}/\widehat{C_{ik}})^{1-\eta_k}\widehat{Y_{ik}} \tag{3.24}$$

The derivations for these counterfactual setups can be found in the Appendix. In bold, we can see the exogenous variable that will deliver the supply shock in our counterfactual system: $\widehat{R_{n,g}}$, the region good specific resource allocation. For our counterfactual scenario we plan to increase the resource allocation of natural gas and oil in the US.

Additionally, $Y_{i,k}$ and $Y_{i,g}$ is introduced as region, good specific production. We do so by replacing the X_{nik} and X_{nig} variables with production in order to decrease the number of equations we have to solve for. We can also do this since there is data on region good specific production. Additionally, as in Fally and Sayre (2018), we simplify the equations to variables that we have data on. $e_{n,L}, e_{n,g}, d_{i,k,g}, \phi_{i,k,L}, \phi_{i,k,g}$ are all further discussed in the data section.

The equations from (13)-(24) define our equilibrium state, but notice that these equations can be compressed into two core variables: $\widehat{w_i}$ and $\widehat{r_{n,g}}$, the wages of labor and the cost of a given commodity in a region. This makes sense since these variables are what drives the equations. We can derive a reduced form with just these two equations. For more detail on why this is the case, the reader may refer to the Appendix. In this setup, the set of given parameters are $\{X_{nig}, X_{nik}, Y_{ng}, Y_{nk}, D_{ng}, D_{nk}, E_n, e, \phi, d_{ikg}\}$ and the set of free variables are $\{\widehat{w_i}, \widehat{r_{n,g}}\}$[1].

In addition to the setup above, if the equations from (13)-(24) are satisfied for the case where $\widehat{Z} = Z'/Z = 1$, then we need to add a constraint. This is because when $\widehat{Z} = 1$, (13)-(24) will have infinitely many solutions, occurring each time the elements of $\widehat{w_i}$ and $\widehat{r_{n,g}}$ are equal. The reason for this is that when we are solving the equations relative to a given year, we also have to constrain the wages or resource prices relative to a given region. Choosing some region n_{base}, we need the constraint that $\widehat{w_i} = \frac{\widehat{w_i}}{\widehat{w_{n_{base}}}}$. This allows the solution to be relative to a given region. Note that this makes the base region's wage equal to 1, which means that depending on the

[1]The Data section goes into more detail about how the given parameters are obtained.

region we choose, GDP (E_n) will be different. This can be interpreted as deciding on which region's currency will be used as the base currency.

Given this counterfactual setup, we are interested in modelling the supply shock of the shale revolution. In order to do so we need data for a base year as well as a numerical setup that will solve for the counterfactual results. The next two sections will discuss the data and numerical methods for the counterfactual setup.

3.3.1 Data

The main source of data that is used for the model calibration is the Global Trade Analysis Project (GTAP) version 8. Description of this dataset can be found in Badri et al. (2012). This version has data on global trade and is freely available. For this project, we use the data from 2007, which we choose since it is considered to be right before the large change in US production of shale gas and oil. The GTAP dataset contains information on 129 regions and 57 sectors. In the dataset, most regions represent major countries[2].

As in Fally and Sayre (2018), we divide the 57 sectors into upstream and downstream industries. Unlike their paper, which tries to disaggregate the commodities (such as minerals), we aggregate these sectors into 6 commodity categories: coal, oil, natural gas, metals, food, and minerals. There are three reasons for aggregating: (1) numerical feasibility, (2) it contains the three commodities we are most interested in, and (3) ease of interpretability. Similarly, for downstream industries, we group the variables into 6 general sectors: manufacturing, clothes manufacturing, metal manufacturing, utilities, business, and public administration/defence/health/education. A summary of this aggregation can be seen in Table A.1.

Bringing the data to a flexible and usable format requires some preprocessing. The dataset on the website is found in an encrypted *.har* format. Once a free account is created, a license is provided by GTAP that allows you to open the dataset with their own software. They provide some methods for extracting this data into *GAMS*,

[2]for a more detailed list of these regions and sectors, we direct the reader to the GTAP's website: https://www.gtap.agecon.purdue.edu/default.asp

a optimization software that MIT does not have a license for, so this dataset has to initially be converted to CSV format using a *har2csv.exe* script provided. Once this is done, the *.csv* files can be analyzed in any general programming language. For this project, Python 3.6 was used. The GTAP dataset is a set of datasets that describe global trade as well as input output tables between industries. Variable and dataset definitions can be found on their website[3]. Using the variables in this dataset, we create variables for $Z = \{X_{nig}, X_{nik}, Y_{ng}, Y_{nk}, D_{ng}, D_{nk}, E_n\}$ as well as parameters e, ϕ, d_{ikg}. The discussion on how these variables were defined from the dataset can be found in the Appendix.

3.3.2 Numerical Methods & Calibration

The counterfactual setting that we intend to study can be set up in different ways. Either as solving a system of equations or as an optimization given a set of constraints. To see this, let us first define what we are solving for: $(\widehat{w_n}, \widehat{r_{ng}})$. We can define a vector x of dimension $n + n * g$ that concatenates w_i and r_{ig}. Next we can define $F(\cdot)$ to be the function whose root we are trying to solve for:

$$0 = F(x) = \begin{bmatrix} \widehat{r_{ng}} - f_{\widehat{r_{ng}}}(\widehat{w_n}, \widehat{r_{ng}}) \\ \widehat{w_n} - f_{\widehat{w_n}}(\widehat{w_n}, \widehat{r_{ng}}) \end{bmatrix} \tag{3.25}$$

where $f_{\widehat{w_n}}(\widehat{w_n}, \widehat{r_{ng}})$ and $f_{\widehat{r_{ng}}}(\widehat{w_n}, \widehat{r_{ng}})$ are the right hand sides of equations (23) and (24); respectively.

$$\begin{bmatrix} f_{\widehat{w_n}}(\widehat{w_n}, \widehat{r_{ng}}) \\ f_{\widehat{r_{ng}}}(\widehat{w_n}, \widehat{r_{ng}}) \end{bmatrix} = \begin{bmatrix} ((\widehat{r_{ng}}/\widehat{C_{ng}})^{1-\rho_g}\widehat{Y_{ng}}/\widehat{R_{ng}} \\ \sum_g(\phi_{ng,L}Y_{ig}/(w_nL_n))(\widehat{w_n}/\widehat{C_{ng}})^{1-\rho_g}\widehat{Y_{ng}} + \sum_k(\phi_{nk,L}Y_{nk}/(w_iL_n))(\widehat{w_n}/\widehat{C_{nk}})^{1-\eta_k}\widehat{Y_{nk}} \end{bmatrix} \tag{3.26}$$

Note that, as we discussed in the counterfactual setup, the variables $\{\widehat{C_{ng}}, \widehat{Y_{ng}}, \widehat{C_{nk}}, \widehat{Y_{nk}}\}$ are functions of $(\widehat{w_n}, \widehat{r_{ng}})$, which means that we do not need to solve for them as well. When considering to solve this problem, in addition to $F(x)$, we need to consider the

[3]https://www.gtap.agecon.purdue.edu/models/setsVariables.asp

constraint of a base region that is necessary (discussed in the counterfactual setup section): $\widehat{w}_i = \frac{\widehat{w_i}}{\widehat{w_{n_{base}}}}$. There are two equivalent setups that we can use to solve this problem: (1) A system of equations and (2) Constrained minimization. The first setup would be:

$$\text{Solve x} \in \mathbb{R}^+ \text{ s.t. } \begin{bmatrix} F(x) \\ \widehat{x_1} - \frac{\widehat{x_1}}{\widehat{x_{n_{base}}}} \\ \widehat{x_2} - \frac{\widehat{x_2}}{\widehat{x_{n_{base}}}} \\ . \\ . \\ . \\ \widehat{x_n} - \frac{\widehat{x_n}}{\widehat{x_{n_{base}}}} \end{bmatrix} = 0 \tag{3.27}$$

where $x_1, ..., x_n$ is the $\widehat{w}_i$ vector. The second setup would be the optimization setup:

$$\begin{aligned} x^* = & arg \min_{x \in \mathbb{R}^+} F(x) \\ & \text{s.t. } \widehat{w}_i = \frac{\widehat{w}_i}{\widehat{w_{n_{base}}}} \forall i \in 1, 2, ..., n \end{aligned} \tag{3.28}$$

The equivalence of both settings to solving the problem is achieved under the condition that there exists a unique solution for $x \in \mathbb{R}^+$. There is some warrant in believing that a unique solution exits. In Allen et al. (2020), the authors prove existence and uniqueness for a class of gravity type general equilibrium models. The 6 conditions laid out in the paper are met by the model in this chapter, but the framework that Allen et al. (2020) lays out is for a single aggregate good, compared to the setting in this chapter, which has multiple goods and two factors of production. Although this proof doesn't encompass the whole framework in this paper, it gives some reason to believe that a unique result exists. Python's SciPy package as well as the IpOpt framework are used to solve this numerically.

3.4 Results

As discussed in the data section, data for most variables is taken from the GTAP 8 dataset. The only parameters that are not in this dataset are the elasticity measures θ, ρ, η, and σ. There is a large literature on values to use for these elasticity measures as they can have a large impact on the output. Initially, we will start with standard values that are used in the literature, as well as the Fally and Sayre (2018) paper. Later, we will try to use better informed values, especially in relation to fossil fuels and see how the results change.

As in Simonovska and Waugh (2014) and Fally and Sayre (2018), we set the trade elasticities (θ) to be 5, supply and demand elasticity of 0.6, and σ value of 1 to satisfy the constant elasticity of substitution assumption. We set the supply shock of a doubling of crude oil resources and a 50% increase in natural gas production, similar to the change seen between 2007 and 2017. If we look at what has been seen in the markets, the Henry Hub Natural Gas Spot Price have gone from and average of \$6.97 in 2007 to \$2.99 in 2017, while natural gas production has gone from 20 Million Cubic Feet to 30 Million Cubic Feet in the US (50% increase)[4][5]. Similarly, crude oil prices in the West Texas Intermediate have gone from \$72 in 2007 to \$51, while US crude oil production has gone from 5 million barrels a day to 9.5 million barrels a day ($\approx 2x$ increase). The price data is highly volatile and hard to interpret as both natural gas and oil prices are highly sensitive to many geopolitical issues that the model isn't able to capture. These geopolitical issues can be collusion to fix prices (OPEC) or impacts of domestic market power of natural gas suppliers (shown in Davis and Muehlegger (2010)). We should not expect our model to mimic these prices closely, as it is trying to evaluate the impact of a shale shock alone.

Using a tolerance of 10^{-4} and the setup given in Equation (28), the model converges to give values for the $\widehat{w_i}$ and $\widehat{r_{ng}}$. As an initial figure, we can look at the change in price values for commodities:$\widehat{r_{ng}}$. These can be seen in the heat-map in Figure 3-1.

[4]Production data https://www.eia.gov/dnav/ng/hist/n9050us2a.htm
[5]Price data https://www.eia.gov/dnav/ng/hist/rngwhhdA.htm

Figure 3-1 shows the $\widehat{r_{ng}} = r'_{ng}/r_{ng}$, the relative change in commodity prices in each region. In the figure, there are two general trends: (1) there has been an overall decrease in oil prices, with the largest decrease in the US, and (2) there has been an average decrease in natural gas prices, but the decrease is more irregular (the largest decrease occurring in the US). This should be expected since trade barriers to crude oil are minimal, due to the ease of transporting a barrel of oil. The largest decrease in price is in the US, with $\widehat{r_{USA,oil}} = 0.64$. The decrease in natural gas prices also follows a trend we would expect, which is that US prices decrease the most significant, followed by bordering countries: Mexico and Canada. This suggests an interesting property of the model, which is that it is capturing the property that sharing a border makes natural gas trade easier (the data does not explicitly contain distance between countries). The $\widehat{r_{USA,NG}} = 0.71$.

Next, we would like to see the sensitivity of these results to the elasticity assumptions as well as looking at different elasticity values that might better represent the commodity properties. Assuming that the commodities of interest (coal, oil, and natural gas) have constant elasticities of trade is a bit unrealistic. Estimating trade elasticity has a long history and is still a large area of research. Instead of trying to estimate these values, we can use estimates published in other papers. Primarily, the Broda and Weinstein (2006) estimates fit well, since they provide trade elasticity parameters for the three commodities of interest. The authors find the elasticity of trade to be 22.1 for crude oil, 2.01 for coal and 2.12 for natural gas. Using these parameters, we can evaluate the model again. The results show a relatively different picture. Where natural gas prices are 0.65 times their base year value ($\widehat{r_{USA,NG}} = 0.65$) and oil prices are 0.77 times their base year value ($\widehat{r_{USA,oil}} = 0.77$). These results are not dramatically different from the initial “naive" results, but still show a a relative shift for both price changes. This shows the importance of the elasticity parameter when it comes to the point estimate of the change in natural resource prices. Figure 3-2 shows the $\widehat{r_{ng}} = r'_{ng}/r_{ng}$, the relative change in commodity prices in each region using the Broda and Weinstein (2006) values. Even though the point estimates defer, the distributional price impact is quite similar.

3.5 Conclusion

In this chapter, we set up a general equilibrium framework for trade that incorporates upstream industries (commodities) as well as downstream industries to simulate the impacts of a supply shock to US natural gas and oil supply. Using a counterfactual setup coined “exact hat algebra", the supply shock was simulated relative to the base year of 2007. Numerically the model converged (using a tolerance of 10^{-4}) and the resulting values were analyzed. The results predict that the supply shock to natural gas and oil caused a 25% to 35% decrease in price for both commodities, as well as a cascading effect on trade, which led to a significant drop in oil prices across most countries and a localized (geographically) impact on natural gas prices.

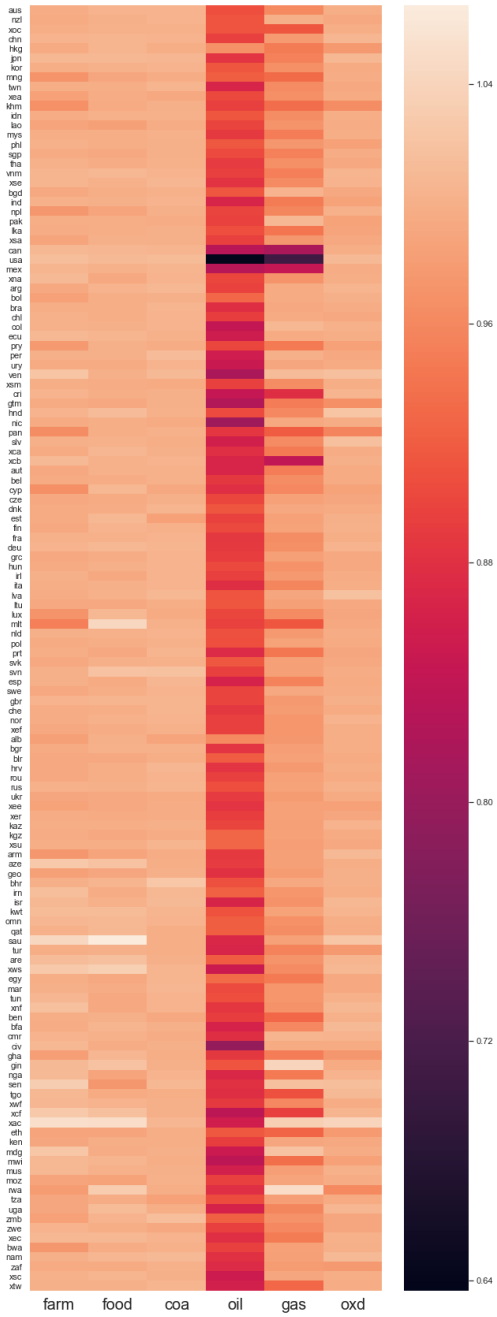

Figure 3-1: Heatmap of $\widehat{r_{ng}}$ values for the initial solution to the General Equilibrium Model.

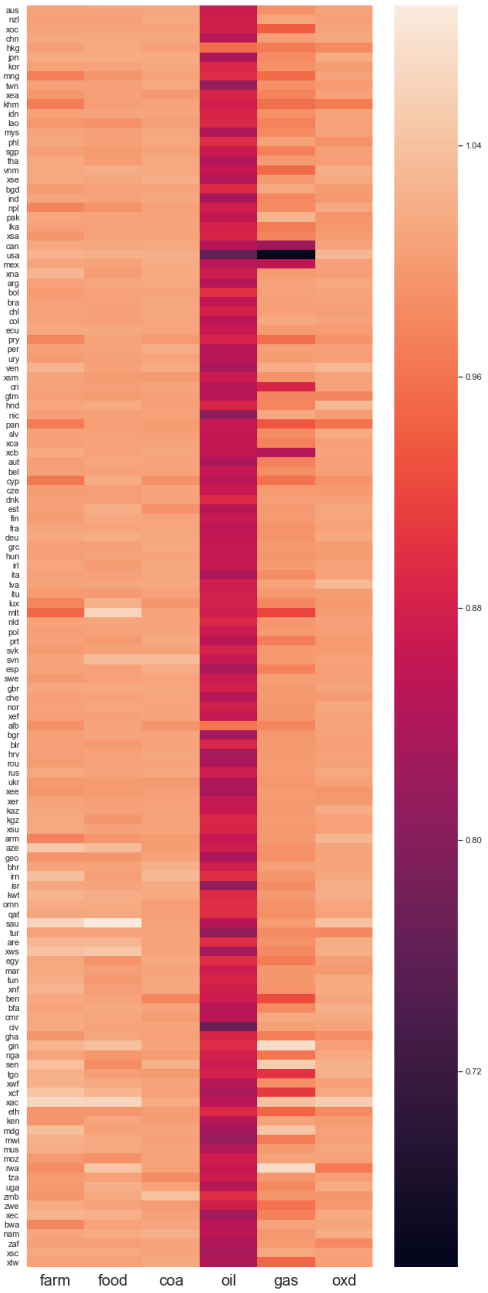

Figure 3-2: Heatmap of $\widehat{r_{ng}}$ values for the solution to the General Equilibrium Model using the Broda and Weinstein (2006) values.

Chapter 4

Conclusion

Using an array of different methods, Chapters 2-4 try to answer questions related to local and global impacts of the shale revolution. There are many areas that this paper does not look at which require future work (some of which has been done by other researchers). One area that this paper does not look at is assessing the benefits and drawbacks of fracking in relation to Climate Change. Recent studies show the methane leakage to be larger than previously thought, quickly bringing the CO_2 impact of using natural gas close to that of coal. One potential area of future work could be a qualitative review of the benefits and drawbacks of fracking. Turning back to the results of Chapters 2-4, we can separate the results into local (Chapter 2) and global (Chapters 3 and 4) impacts.

The methods used in Chapter 2 try to asses the impact that the changing energy composition of US power plants has had on local ambient air pollution. Satellite based, modelled $PM_{2.5}$ data is combined with local power plant properties and price information to show interesting results. The results show that between 2003 and 2018, an increase in coal prices decreased $PM_{2.5}$ concentrations and an increase in natural gas prices increased $PM_{2.5}$ concentration. Furthermore, we are able to show interactions between these prices and power plant capacity information for a given county to understand the varying impacts of prices on counties given the power plant capacity that is located within the county. The results are consistent and significant to robustness checks using additional FEs as well as different transfor-

mations. Counterfactual formulation was used on the results of the regressions to predict the different $PM_{2.5}$ concentrations that might have occurred given higher natural gas prices. Finally, price data is predicted on a power plant level to run a similar analysis on a more granular grid level analysis. The results show similar relationships, adding to the robustness of the results.

The global impacts of shale are analyzed in Chapters 3 and 4. Initially, network properties of fossil fuel networks are analyzed, highlighting the changes that have occurred in international trade of fossil fuel between 2000 and 2016. In Chapter 4, a general equilibrium trade model is used to isolate the impacts that shale has had on US and global prices as well as the cascading effect that has resulted from the shock. The results predict that the supply shock to natural gas and oil caused a 25% to 35% decrease in price for both commodities, as well as a cascading effect on trade, which led to a significant drop in oil prices across most countries and a localized (geographically) impact on natural gas prices.

Appendix A

Tables

Table A.1: Table describing the mapping from the 57-sector GTAP dataset to the 11-sector setup used in this model

#	**11-sector**	**57-sector**
1	food	Paddy rice
2	food	Wheat
3	food	Cereal grains nec
4	food	Vegetables, fruit, nuts
5	food	Oil seeds
6	food	Sugar cane, sugar beet
7	farm	Plant-based fibers
8	farm	Crops nec
9	farm	Cattle,sheep,goats,horses
10	farm	Animal products nec
11	food	Raw milk
12	farm	Wool, silk-worm cocoons
13	farm	Forestry
14	food	Fishing
15	coal	Coal

16	oil	Oil
17	gas	Gas
18	minerals	Minerals nec
19	food	Meat: cattle,sheep,goats,horse
20	food	Meat products nec
21	food	Vegetable oils and fats
22	food	Dairy products
23	food	Processed rice
24	food	Sugar
25	food	Food products nec
26	manufacturing	Beverages and tobacco products
27	clothe manufacturing	Textiles
28	clothe manufacturing	Wearing apparel
29	clothe manufacturing	Leather products
30	manufacturing	Wood products
31	manufacturing	Paper products, publishing
32	manufacturing	Petroleum, coal products
33	manufacturing	Chemical,rubber,plastic prods
34	manufacturing	Mineral products nec
35	metal manufacturing	Ferrous metals
36	metal manufacturing	Metals nec
37	metal manufacturing	Metal products
38	manufacturing	Motor vehicles and parts
39	manufacturing	Transport equipment nec
40	manufacturing	Electronic equipment
41	manufacturing	Machinery and equipment nec
42	manufacturing	Manufactures nec
43	utilities	Electricity
44	utilities	Gas manufacture, distribution

45	utilities	Water
46	manufacturing	Construction
47	business	Trade
48	transportation	Transport nec
49	transportation	Sea transport
50	transportation	Air transport
51	business	Communication
52	business	Financial services nec
53	business	Insurance
54	business	Business services nec
55	business	Recreation and other services
56	PubAdmin/Defence/Health/Educat	PubAdmin/Defence/Health/Educat

Appendix B

Counterfactual derivation

In this section of the appendix, we will derive the counterfactual setup proposed in Chapter 4. Chapter 4 uses the technique first introduced in Dekle et al. (2008), that derives a set of equations of the form: $\widehat{Z} = Z'/Z$. This derivation is skipped in the chapter and will be derived here. The algebra for each equation has a repeating flow: (1) take an equation from the Model section of Chapter 4, (2) set up the counterfactual $Z' = \widehat{Z}Z$, and (3) replace the variables that are not fixed on the right hand side with their counterfactuals.[1].

Counterfactual Demand for final good Since the right hand side is all multiplicative, the result is simple:

$$\begin{aligned} \widehat{D_{n,k}} D_{n,k} &= (\widehat{P_{n,k}} P_{n,k} / (\widehat{P_n} P_n))^{1-\sigma} a_{nk} \widehat{E_n} E_n \\ \widehat{D_{n,k}} &= \frac{(\widehat{P_{n,k}} P_{n,k} / (\widehat{P_n} P_n))^{1-\sigma} a_{nk} \widehat{E_n} E_n}{(P_{n,k}/(P_n))^{1-\sigma} a_{nk} E_n} \\ \widehat{D_{n,k}} &= (\widehat{P_{n,k}} / \widehat{P_n})^{1-\sigma} \widehat{E_n} \end{aligned} \tag{B.1}$$

[1]Note that similar algebra can also be found in Fally and Sayre (2019)'s appendix. Difference is that they use $\hat{\tau}$ and I use $\hat{R}$.

Counterfactual Production Cost of final goods:

$$\begin{aligned}
\widehat{C_{i,k}}C_{i,k} &= A_{i,k}\left[\beta_{i,k,L}(\widehat{w}_i w_i)^{1-\eta_k} + \sum_{g\in G(k)} \beta_{i,k,g}(\widehat{P_{i,g}}P_{i,g})^{1-\eta_k}\right]^{\frac{1}{1-\eta_k}} \\
\widehat{C_{i,k}} &= \left[A_{i,k}^{1-\eta_k}\beta_{i,k,L}(\widehat{w}_i w_i)^{1-\eta_k}/C_{i,k}^{1-\eta_k} + \sum_{g\in G(k)} \beta_{i,k,g}(\widehat{P_{i,g}}P_{i,g})^{1-\eta_k}/C_{i,k}^{1-\eta_k}\right]^{\frac{1}{1-\eta_k}} \\
\widehat{C_{i,k}} &= \left[\phi_{i,k,L}\widehat{w}_i{}^{1-\eta_k}(\widehat{w}_i w_i)^{1-\eta_k}/C_{i,k}^{1-\eta_k} + \sum_{g\in G(k)} \widehat{P_{i,g}}^{1-\eta_k}\phi_{i,k,g}\right]^{\frac{1}{1-\eta_k}}
\end{aligned} \tag{B.2}$$

where $\phi_{i,k,L}$ and $\phi_{i,k,g}$ are the labor and commodity cost shares. These values can be calculated using the GTAP dataset.

Counterfactual Price Index for final goods:

$$\begin{aligned}
\widehat{P_{n,k}}P_{n,k} &= \left[\sum_i (\widehat{C_{i,k}}C_{i,k}\tau_{n,i,k})^{-\theta_k}\right]^{\frac{1}{-\theta_k}} \\
\widehat{P_{n,k}} &= \left[\sum_i (\widehat{C_{i,k}}C_{i,k}\tau_{n,i,k})^{-\theta_k}/P_{n,k}^{-\theta_k}\right]^{\frac{1}{-\theta_k}} \\
\widehat{P_{n,k}} &= \left[\sum_i (\widehat{C_{i,k}})^{-\theta_k}(C_{i,k}\tau_{n,i,k}/P_{n,k})^{-\theta_k}\right]^{\frac{1}{-\theta_k}} \\
\widehat{P_{n,k}} &= \left[\sum_i (\widehat{C_{i,k}})^{-\theta_k}\lambda_{n,i,k}\right]^{\frac{1}{-\theta_k}}
\end{aligned} \tag{B.3}$$

where $\lambda_{n,i,k}$ is the share of final good k being traded from region i to region n. This tensor can be calculated using the GTAP dataset.

Counterfactual Demand for commodities:

$$\widehat{D_{i,g}}\widehat{D_{i,g}} = \sum_{\forall\, k} \beta_{i,k,g}(\widehat{P_{i,g}}P_{i,g}/(\widehat{C_{i,k}}C_{i,k}))^{1-\eta_k} Y_{i,k}\widehat{Y_{i,k}}/D_{i,g,k}$$
$$\widehat{D_{i,g}} = \sum_{\forall\, k} \beta_{i,k,g}(P_{i,g}/C_{i,k})^{1-\eta_k} Y_{i,k}/D_{i,g}(\widehat{P_{i,g}}/\widehat{C_{i,k}})^{1-\eta_k}\widehat{Y_{i,k}} \tag{B.4}$$
$$\widehat{D_{i,g}} = \sum_{\forall\, k} d_{i,k,g}(\widehat{P_{i,g}}/\widehat{C_{i,k}})^{1-\eta_k}\widehat{Y_{i,k}}$$

where $d_{i,k,g}$ is the ratio of input into final good k of commodity g in country i for the baseline year, which we can aggregate data on by using GTAP's input output tables. Note that the model being used assumes that commodities are inputs to final goods and there are no final goods inputted into commodities, which results in a cutting off of part of the input-output tables and re-normalizing.

Counterfactual Production Cost of commodities can be derived:

$$\widehat{C_{i,g}}C_{i,g} = A_{i,g}\left[\beta_{i,g}(\widehat{r_{i,g}}r_{i,g})^{1-\rho_g} + (1-\beta_{i,g})(\widehat{w}_i w_i)^{1-\rho_g}\right]^{\frac{1}{1-\rho_g}}$$
$$\widehat{C_{i,g}} = \left[A_{i,g}^{1-\rho_g}\beta_{i,g}(\widehat{r_{i,g}}r_{i,g})^{1-\rho_g}/C_{i,g}^{1-\rho_g} + A_{i,g}^{1-\rho_g}(1-\beta_{i,g})(\widehat{w}_i w_i)^{1-\rho_g}/C_{i,g}^{1-\rho_g}\right]^{\frac{1}{1-\rho_g}} \tag{B.5}$$
$$\widehat{C_{i,g}} = \left[\phi_{i,g,r}(\widehat{r_{i,g}})^{1-\rho_g} + \phi_{i,g,l}(\widehat{w}_i)^{1-\rho_g}\right]^{\frac{1}{1-\rho_g}}$$

where $\phi_{i,g,r}$ and $\phi_{i,g,l}$ represent the ratio of resource cost and labor cost to total cost of producing commodity g in country i. These can also be derived using the GTAP data.

Counterfactual Price Index for commodity goods: Similar to the price index of final goods:

$$\widehat{P_{n,g}}P_{n,g} = \left[\sum_i (\widehat{C_{i,g}}C_{i,g}\tau_{n,i,g})^{-\theta_g}\right]^{\frac{1}{-\theta_g}}$$
$$\widehat{P_{n,g}} = \left[\sum_i \lambda_{n,i,g}(\widehat{C_{i,g}})^{-\theta_g}\right]^{\frac{1}{-\theta_g}} \tag{B.6}$$

Price index of a region

$$\begin{aligned} \widehat{P_n}P_n &= \left[\sum_k a_{n,k}(\widehat{P_{n,k}}P_{n,k})^{1-\sigma}\right]^{\frac{1}{1-\sigma}} \\ \widehat{P_n} &= \left[\sum_k \alpha_{n,k}(\widehat{P_{n,k}})^{1-\sigma}\right]^{\frac{1}{1-\sigma}} \end{aligned} \tag{B.7}$$

where $\alpha_{n,k}$ is the share of expenditure on good k in region n, which is calculated using GTAP data.

Trade in final goods and commodities: see that $P_{n,g}^{-\theta_g} = \sum_i (C_{i,g}\tau_{n,i,g})^{-\theta_g}$, then:

$$\begin{aligned} \widehat{X_{n,i,k}}X_{n,i,k} &= (\widehat{C_{i,k}}C_{i,k}\tau_{n,i,k})^{-\theta_k}(\widehat{P_{n,k}}P_{n,k})^{\theta_k}\widehat{D_{n,k}}D_{n,k} \\ \widehat{X_{n,i,k}} &= (\widehat{C_{i,k}})^{-\theta_k}(\widehat{P_{n,k}})^{\theta_k}\widehat{D_{n,k}} \end{aligned} \tag{B.8}$$

Now since $X_{n,i,k}$ represents a large set of equations, we will instead summarize this counterfactual equation in production $Y_{i,k} = \sum_{\forall\, n} X_{n,i,k}$, which has the counterfactual:

$$\begin{aligned} \widehat{Y_{i,k}}Y_{i,k} &= \sum_{\forall\, n} \widehat{X_{n,i,k}}X_{n,i,k} \\ \widehat{Y_{i,k}} &= \sum_{\forall\, n}(X_{n,i,k}/Y_{i,k})(\widehat{C_{i,k}})^{-\theta_k}(\widehat{P_{n,k}})^{\theta_k}\widehat{D_{n,k}} \end{aligned} \tag{B.9}$$

Similarly for commodities:

$$\widehat{Y_{i,g}} = \sum_{\forall\, n}(X_{n,i,g}/Y_{i,g})(\widehat{C_{i,g}})^{-\theta_g}(\widehat{P_{n,g}})^{\theta_g}\widehat{D_{n,g}} \tag{B.10}$$

where θ_g and θ_k are the trade elasticities, $\lambda_{n,i,g}$ and $\lambda_{n,i,k}$ are the share of exports which are useful to define for future analysis. This data can be found from a variety of sources, but for consistency, GTAP is used.

Counterfactual Sources of Income

$$
\begin{aligned}
L_i\widehat{w_i}w_i = &\sum_{g\in Commodities}(1-\beta_{i,g})(\widehat{w_i}w_i/(\widehat{C_{i,g}}C_{i,g}))^{1-\rho_g}\widehat{Y_{i,g}}Y_{i,g} \\
&+\sum_{k\in\text{Final goods}}(\beta_{i,k,L})(\widehat{w_i}w_i/(\widehat{C_{i,k}}C_{i,k}))^{1-\eta_k}\widehat{Y_{i,k}}Y_{i,k} \\
\widehat{w_n} = &\sum_g(\phi_{ig,L}Y_{ig}/(w_iL_i))(\widehat{w_i}/\widehat{C_{ig}})^{1-\rho_g}\widehat{Y_{ig}} + \sum_k(\phi_{ik,L}Y_{ik}/(w_iL_i))(\widehat{w_i}/\widehat{C_{ik}})^{1-\eta_k}\widehat{Y_{ik}}
\end{aligned}
\tag{B.11}
$$

As well as total income from commodities:

$$
\begin{aligned}
\widehat{R_{i,g}}\widehat{r_{i,g}}R_{i,g}r_{i,g} &= \beta_{i,g}(\widehat{r_{i,g}}r_{i,g}/(\widehat{C_{i,g}}C_{i,g}))^{1-\rho_g}\widehat{Y_{i,g}}Y_{i,g} \\
\widehat{R_{i,g}}\widehat{r_{i,g}} &= (\widehat{r_{i,g}}/(\widehat{C_{i,g}}))^{1-\rho_g}\widehat{Y_{i,g}}
\end{aligned}
\tag{B.12}
$$

Finally the counterfactual total income:

$$
\begin{aligned}
\widehat{E_n}E_n &= L_n\widehat{w_n}w_n + \sum_{g\in Commodities}\widehat{R_{n,g}}R_{n,g}\widehat{r_{n,g}}r_{n,g} \\
\widehat{E_n} &= e_{i,L}\widehat{w_n} + \sum_{g\in Commodities}\widehat{R_{n,g}}\widehat{r_{n,g}}e_{i,g,R}
\end{aligned}
\tag{B.13}
$$

where $_{i,L}$ and $e_{i,g,R}$ represent the ratio of labor and resources to total income. These can also be derived using the GTAP data.

Bibliography

Treb Allen, Costas Arkolakis, and Yuta Takahashi. Universal gravity. *Journal of Political Economy*, 128(2):000–000, 2020.

Jeff Alstott, Ed Bullmore, and Dietmar Plenz. powerlaw: a python package for analysis of heavy-tailed distributions. *arXiv*, 1305.0215v3, 2014.

Caterina De Bacco, Daniel B. Larremore, and Cristopher Moore. A physical model for efficient ranking in networks. *arXiv*, 1709.09002v3, 2017.

Angel Aguiar Badri Narayanan and Editors (2012). Robert McDougall. Global trade, assistance, and production: The gtap 8 data base, center for global trade analysis. *Purdue University*, 55(3):511–540, 2012.

Christian Broda and David E Weinstein. Globalization and the gains from variety. *The Quarterly journal of economics*, 121(2):541–585, 2006.

Richard T Burnett, C Arden Pope III, Majid Ezzati, Casey Olives, Stephen S Lim, Sumi Mehta, Hwashin H Shin, Gitanjali Singh, Bryan Hubbell, Michael Brauer, et al. An integrated risk function for estimating the global burden of disease attributable to ambient fine particulate matter exposure. *Environmental health perspectives*, 122(4):397–403, 2014.

Curtis Carlson, Dallas Burtraw, Maureen Cropper, and Karen L Palmer. Sulfur dioxide control by electric utilities: What are the gains from trade? *Journal of political Economy*, 108(6):1292–1326, 2000.

Yuyu Chen, Avraham Ebenstein, Michael Greenstone, and Hongbin Li. Evidence on the impact of sustained exposure to air pollution on life expectancy from china's huai river policy. *Proceedings of the National Academy of Sciences*, 110(32):12936–12941, 2013.

Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *Society for Industrial and Applied Mathematics*, 51(4):661–703, 2009.

Arnaud Costinot and Andrés Rodríguez-Clare. Trade theory with numbers: Quantifying the consequences of globalization. 4:197–261, 2014.

Arnaud Costinot and Jonathan Vogel. Beyond ricardo: Assignment models in international trade. *economics*, 7(1):31–62, 2015.

Janet Currie, Michael Greenstone, and Katherine Meckel. Hydraulic fracturing and infant health: New evidence from pennsylvania. *Science advances*, 3(12):e1603021, 2017.

Lucas W Davis and Erich Muehlegger. Do americans consume too little natural gas? an empirical test of marginal cost pricing. *The RAND Journal of Economics*, 41(4):791–810, 2010.

Tsuyoshi Deguchi, Katsihide Takahashi, Hideki Takayasu, and Misako Takayasu. Hubs and authorities in the world trade network using a weighted hits algorithm. *Plos One*, 1709.09002v3(9), 2014.

Robert Dekle, Jonathan Eaton, and Samuel Kortum. Global rebalancing with gravity: Measuring the burden of adjustment. *IMF Staff Papers*, 55(3):511–540, 2008.

Yixing Du, Xiaohan Xu, Ming Chu, Yan Guo, and Junhong Wang. Air particulate matter and cardiovascular disease: the epidemiological, biomedical and clinical evidence. *Journal of thoracic disease*, 8(1):E8, 2016.

Jonathan Eaton and Samuel Kortum. Technology, geography, and trade. *Econometrica*, 70(5):1741–1779, 2002.

Jonathan Eaton and Samuel Kortum. Putting ricardo to work. *Journal of Economic Perspectives*, 26(2):65–90, 2012.

D. Edler and M. Rosvall. The mapequation software package, available online at.

Thibault Fally and James Sayre. Commodity trade matters. Technical report, National Bureau of Economic Research, 2018.

Farid Farrokhi. Global sourcing in oil markets. *PhD diss., Pennsylvania State University*, 2016.

Center for International Earth Science Information Network CIESIN Columbia University. Gridded population of the world, version 4 (gpwv4): Population density, revision 11. 20200501 2018.

Jiang-Bo Geng, Qiang Ji, and Ying Fan. A dynamic analysis on global natural gas trade network. *Applied Energy*, 2014.

Catherine Hausman and Ryan Kellogg. Welfare and distributional implications of shale gas. Technical report, National Bureau of Economic Research, 2015.

Reid Johnsen, Jacob LaRiviere, and Hendrik Wolff. Fracking, coal, and air quality. *Journal of the Association of Environmental and Resource Economists*, 6(5):1001–1037, 2019.

Semanur Soyyigit Kaya and Ercan Eren. Complex network approach to international trade of fossil fuel. *World Academy of Science, Engineering, and Technology International Journal of Economics and Management Engineering*, 10(1), 2016.

Andrea Lancichinetti and Santo Fortunato. Community detection algorithms: a comparative analysis. *arXiv*, 0908.1062v2, 2010.

United Nations. Commodities trading.

Ole Raaschou-Nielsen, Zorana J Andersen, Rob Beelen, Evangelia Samoli, Massimo Stafoggia, Gudrun Weinmayr, Barbara Hoffmann, Paul Fischer, Mark J Nieuwenhuijsen, Bert Brunekreef, et al. Air pollution and lung cancer incidence in 17 european cohorts: prospective analyses from the european study of cohorts for air pollution effects (escape). *The lancet oncology*, 14(9):813–822, 2013.

M. Rosvall, D. Axelsson, and C.T. Bergstrom. The map equation. *The European Physical Journal*, 178(13):23, 2009.

Richard Schmalensee and Robert N Stavins. The so 2 allowance trading system: the ironic history of a grand policy experiment. *Journal of Economic Perspectives*, 27(1):103–22, 2013.

Ina Simonovska and Michael E Waugh. The elasticity of trade: Estimates and evidence. *Journal of international Economics*, 92(1):34–50, 2014.

Aaron Van Donkelaar, Randall V Martin, Chi Li, and Richard T Burnett. Regional estimates of chemical composition of fine particulate matter using a combined geoscience-statistical method with information from satellites, models, and monitors. *Environmental science & technology*, 53(5):2595–2611, 2019.

Weiqiong Zhong, Haizhong An, Wei Fang, Ziangyun Gao, and Di Dong. Features and evolution of international fossil fuel trade network based on value of emergy. *Applied Energy*, 2016.

MIT CEEPR

MIT Center for Energy and Environmental Policy Research

Since 1977, the Center for Energy and Environmental Policy Research (CEEPR) has been a focal point for research on energy and environmental policy at MIT. CEEPR promotes rigorous, objective research for improved decision making in government and the private sector, and secures the relevance of its work through close cooperation with industry partners from around the globe. Drawing on the unparalleled resources available at MIT, affiliated faculty and research staff as well as international research associates contribute to the empirical study of a wide range of policy issues related to energy supply, energy demand, and the environment.

An important dissemination channel for these research efforts is the MIT CEEPR Working Paper series. CEEPR releases Working Papers written by researchers from MIT and other academic institutions in order to enable timely consideration and reaction to energy and environmental policy research, but does not conduct a selection process or peer review prior to posting. CEEPR's posting of a Working Paper, therefore, does not constitute an endorsement of the accuracy or merit of the Working Paper. If you have questions about a particular Working Paper, please contact the authors or their home institutions.

MIT Center for Energy and Environmental Policy Research
77 Massachusetts Avenue, E19-411
Cambridge, MA 02139
USA

Website: ceepr.mit.edu

MIT CEEPR Working Paper Series is published by the MIT Center for Energy and Environmental Policy Research from submissions by affiliated researchers.

For inquiries and/or for permission to reproduce material in this working paper, please contact:

Email ceepr@mit.edu
Phone (617) 253-3551
Fax (617) 253-9845