Imperfect Markets versus Imperfect Regulation in U.S. Electricity Generation

Steve Cicala*
University of Chicago
Click here for the latest version.

January 22, 2017

Abstract

This paper estimates changes in electricity generation costs caused by the introduction of market mechanisms to determine output decisions in service areas that were previously using command-and-control-type operations. I use the staggered transition to markets from 1999-2012 to evaluate the causal impact of liberalization using a nationwide panel of hourly data on electricity demand and unit-level costs, capacities, and output. To address the potentially confounding effects of unrelated fuel price changes, I use machine learning methods to predict the allocation of output to generating units in the absence of markets for counterfactual production patterns. I find that markets reduce production costs by $3B per year by reallocating output among existing power plants: Gains from trade across service areas increase by 20% based on a 10% increase in traded electricity, and costs from using uneconomical units fall 20% from a 10% reduction in their operation.

*I am grateful to Gary Becker, Jim Bushnell, Thom Covert, Tatyana Deryugina, Edward Glaeser, Michael Greenstone, David Hémous, Lawrence Katz, Ryan Kellogg, Erin Mansur, Kevin Murphy, Erica Meyers, Morten Olsen, Jim Sallee, Andrei Shleifer, Chad Syverson, Roberton Williams III, Matthew White, and seminar participants at Harvard, MIT, Cornell, Chicago, Yale, the UC Energy Institute, the EEE Session of the 2015 NBER Summer Institute, UIUC, Wharton, and Brown for helpful comments and suggestions. Sébastien Phan, Julien, Sauvan, Songyuan Ding, Enrique Chazaro-Acosta, Xianying Fan, Mary Vansuch, and Dan Pechi provided excellent research assistance. This paper has been reviewed by the Energy Information Administration to ensure no confidential data has been disclosed. All errors remain my own. e-mail: scicala@uchicago.edu

1 Introduction

When regulation brings its own host of distortions and inefficiencies, the mere existence of a market failure is insufficient to ensure government intervention will improve welfare. Instead, by comparing the distortions under potential regulatory regimes, one can identify superior policies as those with relatively fewer imperfections (Kahn (1979); Joskow (2010)). This paper undertakes such an evaluation in the context of U.S. wholesale electricity markets, which have replaced command-and-control-type operations in some areas.

To do so I construct a virtually complete hourly characterization of supply and demand of the U.S. electrical grid from 1999 - 2012. Data on fuel costs, capacities, heat efficiency, and operations of nearly all generating units at the hourly level allows me to construct power supply curves (known as the "merit order") for each of 98 "Power Control Areas" (PCAs), as well as observe the units that were chosen to operate to meet demand at any moment in time. These curves allow me to calculate two key welfare measures for each PCA-date-hour: "out of merit" losses from dispatching higher marginal cost units relative to installed capacity, and the gains from trading electricity across areas. Market power losses manifest themselves as out of merit production (Borenstein et al. (2002); Mansur (2001)), as do normal grid operations, such as maintenance, refueling, start-up costs, and transmission congestion (Davis and Wolfram (2012); Mansur (2008); Reguant (2014)). In either case, the increased operational costs are observationally equivalent as the distance between the realized cost of operations and cost from utilizing only the lowest-cost installed capacity.

While prior papers have evaluated one of these outcomes during single instances of market transition, I develop a framework and compile the necessary data to examine both outcomes over the history of market transitions since 1999. I use the staggered creation and expansions of wholesale electricity markets over this period to estimate the causal impact of using markets to allocate production on these welfare measures. I employ a differences-in-differences (DD) framework to estimate changes in gains from trade and out of merit losses following the transition to market dispatch against PCAs that have not undergone any regulatory changes. This approach finds gains from trade increase by upwards of 30% after adopting market dispatch due to a 10% increase in electricity traded. There is also a 10% decrease in out of merit operations, reducing these costs by nearly 20%.

The simple DD approach is susceptible to the confounding effects of fuel price fluctuations (over time and across areas) when estimating counterfactual outcomes: Fuels prices shift supply curves, making historical outcomes poor counterfactuals

for what would have happened today under a different set of prevailing fuel prices. This means one might estimate changes in the gains from trade without any actual changes in production patterns because the value of offset production scales with fuel prices. This issue motivates a policy function approach in which I estimate each system operators' rules for dispatching units in a given year, and compare outcomes the following year against those predicted by the policy function. I show how the treatment effect can be estimated by comparing changes in the quality of fit of this rule across areas that switch to market dispatch against areas with no change in regulation.

Estimating dispatch probabilities with out-of-sample validity is a pure prediction problem for which recent developments in the machine learning literature have proven to be particularly effective (Kleinberg et al. (2015)). I use the random forest algorithm of Breiman (2001) to non-parametrically estimate policy functions, then embed the results in a DD framework to estimate causal treatment effects. This part of the paper complements the recent work of Burlig et al. (2016), who also use machine learning methods (Least Absolute Shrinkage and Selection Operator) to predict counterfactual outcomes.

This approach yields estimates smaller in magnitude than the simple DD estimates for gains from trade, suggesting fuel price confounding. I find that production costs are reduced by about three billion dollars per year due to market-based improvements in allocating output to lower cost units, with these savings split between reduced output from uneconomical units and gains from trade by 2:1.

It should be noted at the outset that my estimates measure changes in how output is allocated *given the installed capacity, costs, and patterns of demand*. It would not be unreasonable to suspect that market dispatch has affected investment incentives, which are likely to be an important source of welfare changes. In addition, my estimates measure the average effect of market dispatch, which itself has been heterogeneous both with respect to pre-existing institutions (i.e. power pools, bilateral markets, or smoke-filled rooms), and with respect to the rules of the markets implemented (uniform or locational marginal prices, virtual bidding, market monitors, etc.). However, given the even greater differences between market and traditional dispatch methods these estimates should be informative regarding the performance of the relatively new mechanisms that currently determine how over 60% of generating capacity in the United States is utilized.

The paper is organized as follows: in the next section I describe the structure of electricity generation and transmission in the United States, and the institutional

details that will facilitate estimation. The third section describes how out of merit costs and gains from trade are measured in electricity generation, and the fourth section describes the data. The fifth section presents an estimation strategy motivated by this setting. The sixth section presents causal estimates of the impact of markets on gains from trade and out of merit costs. The final section concludes.

2 Background on Power Control Areas and Dispatch in the United States

The U.S. electricity grid developed over the 20^{th} century based on a mix of IOUs, government-owned utilities (municipal, state, and federal), and non-profit cooperatives. All of these organizations tended to be vertically integrated, so they owned the power plants, the transmission system, and the delivery network within their respective, exclusively operated territories. The entity that determines which power plants operate to meet demand is called a "Balancing Authority." A single Balancing Authority controls the transmission system and dispatches power plants within a "Power Control Area," or PCA. When vertically-integrated, the Balancing Authority and Utility have often been one-in-the-same, as with the service territory and the PCA.[1] These areas operate with relative autonomy over their assets, and transmission lines that connect areas enable flows between them.

The national grid consists of three large Interconnections: East, West, and Texas (with relatively little capacity to transmit power between them). Figure I shows the approximate configurations of the U.S. Electricity Grid in 1999 and 2012.[2] The boundaries between Interconnections are denoted in Panel A by the thick black lines separating Texas and the West (unchanged over the period). The red lines denote regions of the North American Electric Reliability Corporation (NERC) who coordinate their operations in order to preserve the stability of the transmission system (when large plants go down for maintenance, for example). The tangle of power control areas reflects the legacy of local monopolies that have been the principal architects of the U.S. electricity grid.

Although the Public Utility Regulatory Policies Act of 1978 (PURPA) opened the door for independent power generation (by requiring IOUs to buy their output at

[1] Exceptions include the New York and New England Power Pools, which formed in response to The Great Northeast Blackout of 1965, as well as smaller utilities that do not control dispatch directly. Regional reserve margin coordination was also formalized during this time with the establishment of the National Electric Reliability Council.

[2] The exact geographic boundaries of PCAs often defy straightforward demarcation. This map is based on U.S. counties, with the predominant PCA receiving assignment of the entire county–and is therefore approximate for visualization purposes. In addition, a number of small or hydro-only PCAs are merged with the larger neighboring areas that provide the majority of their (fossil-based) energy.

"avoided cost"), the growth of such producers was impeded by discriminatory transmission practices (Joskow (2000)). Because the IOUs owned the transmission system, they could effectively shut independent producers out of wider markets by denying transmission access.[3] This began to change with the Energy Policy Act of 1992, which required the functional separation of transmission system owners and power marketers–they were no longer allowed to use their wires to prevent or extract the surplus from trades across their territory. These changes were codified on April 24, 1996 with FERC orders 888 and 889, which required open-access, non-discriminatory tariffs for wholesale electricity transmission.

Open-access created greater potential for wholesale electricity markets, which were initially conducted through bilateral contracts for power. In this decentralized setting, contracts would typically specify the amount of electricity to be generated by one utility under a set of conditions, transmitted across a particular area, and withdrawn from the system by the purchasing utility. Mansur and White (2012) give examples showing why the nature of congestion in electricity transmission networks renders decentralized markets particularly poorly suited for identifying all of the potential gains from trade. In particular, transmission lines are constrained by net flows of power. When this is the case, there are production externalities that may allow otherwise infeasible bilateral trades to occur by coordinating offsetting transactions to keep net flows below transmission capacity. Identifying such potential trades in this type of decentralized market is a challenge akin to coordinating simultaneous multilateral exchanges (Roth et al. (2004)).

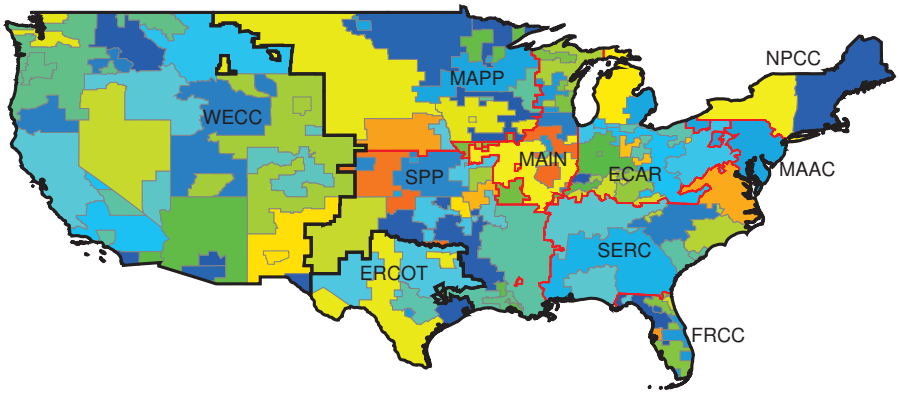
Operationally, balancing authorities have relied on engineering estimates of costs to devise dispatch algorithms to determine which plants within the PCA operate, and separately schedule any other operations requested by utilities (for bilateral trades). Centralized wholesale electricity markets ("market dispatch") integrate dispatch operations in to an auction for electricity. In day-ahead auctions, for example, generators submit bids to produce electricity, and only those below the price needed to meet projected demand are called on to operate. These auctions incorporate feasibility constraints, so calling on higher-priced units to operate due to transmission congestion allows for the direct revelation of the cost of shortcomings in the transmission system.[4] Day-ahead markets establish financial obligations to produce, which are sub-

[3]Examples of IOUs exercising market dominance can be found in Appendix C of FERC Order 888.

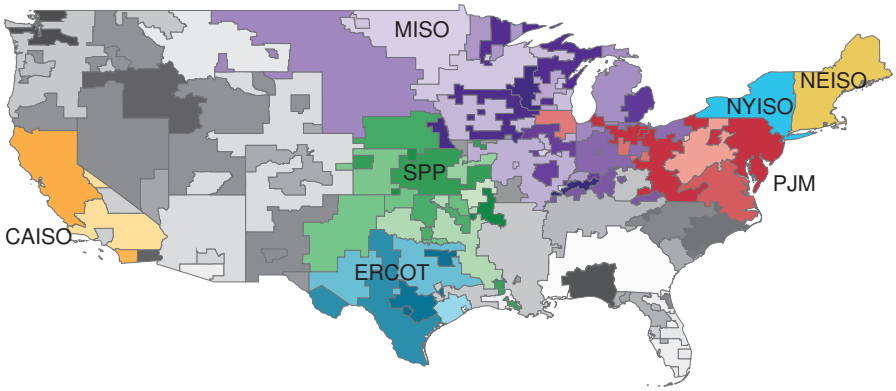
[4]In particular, auctions using the "Standard Market Design" yield "Locational Marginal Prices" (LMPs) which denote the market-clearing price at each of the points of withdrawal from the system. When LMPs are identical everywhere, the system is said to be uncongested.

Figure I: U.S. Electrical Grid as Power Control Areas

(a) Approximate PCA Configuration in 1999

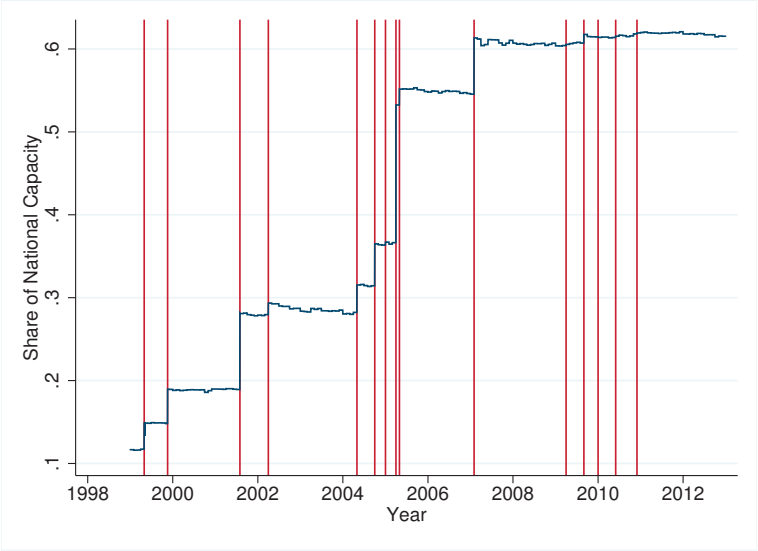


(b) PCAs by Market Dispatch in 2012



Note: Thick black lines in Panel A denote Interconnection boundaries, red lines denote NERC Reliability Regions. Boundaries are approximate.

Figure II: Share of Generating Capacity Dispatched by Markets



Note: Vertical red lines indicate dates of transition to market-based dispatch.

sequently either met with production in the real time market or unwound by buying back one's allocated output at the real time price (Wolak (2000); Hortacsu and Puller (2008); Ito and Reguant (2016); Cramton (2003); Jha and Wolak (2013); Borenstein et al. (2008), among others).

As of 2012, 60 of the 98 PCAs operating in 1999 had adopted market dispatch, either during the initial creation of a new market or as part of the expansion of an existing market. Adopting market dispatch is a discrete change in the decision algorithm that allocates output to generating units: the local PCA cedes control of their transmission system to an Independent System Operator, who conducts the auctions.

All told, there have been 15 distinct events in which PCAs have transitioned to market dispatch overnight. Figure II denotes each of these events with a vertical red line, and shows that over the period of study markets have expanded from covering about 10% of capacity to roughly 60%. The remaining areas have retained their traditional dispatch methods, though a number have continued to explore the possibility of joining existing markets.[5] This variation in market adoption forms the basis of the empirical strategy for causal estimates by allowing the comparison of changes in allocative efficiency following the transition to market dispatch relative to areas that

[5]For example, the East Kentucky Power Cooperative joined PJM on 6/1/2013, there was a major southern expansion of MISO on 12/18/2013, and Pacificorp has formally begun to explore the possibility of joining CAISO.

have not undergone such changes over the same period.

The transition from command-and-control to market dispatch is related to, but distinct from the movement toward restructured electricity markets in the United States (Joskow and Schmalensee (1988)). In particular, the changes to dispatch and transmission described thus far were undertaken by the Federal government.[6] The end of cost-of-service regulation of vertically-integrated IOUs was initiated by states. It is important to distinguish between these developments, for although all states that adopted restructuring legislation eventually adopted market dispatch, many areas began participating in these markets while preserving their traditional regulatory framework.[7] I therefore focus my attention on the cost of generating electricity, rather than the retail price of power delivered to consumers, whose relationship with their local utility may or may not have changed over this period.

Vulnerability to the exercise of market power has been a primary focus of the research on wholesale electricity markets to date. From the UK (Wolfram (1999); Wolak and Patrick (1997)), Spain (Ito and Reguant (2016); Reguant (2014)), New Zealand and Australia (Wolak (2012)) abroad, to California (Borenstein et al. (2002); Bushnell et al. (2008); Joskow and Kahn (2002); Puller (2007); Borenstein (2002)), PJM (Mansur (2001, 2008)), and Texas (Hortacsu and Puller (2008)) in the United States, one could fairly characterize these vulnerabilities as robust. Against these losses, there is sparse evidence of allocative efficiency gains from market dispatch, with the notable exception of Mansur and White (2012) who study one of the 15 market expansion events described above. Instead, liberalization studies have focused on state-led deregulatory events to estimate within-plant changes: reduced maintenance time (Davis and Wolfram (2012),Cropper et al. (2011)), labor and fuel costs (Fabrizio et al. (2007); Cicala (2015)), and capital intensity of pollution abatement equipment (Fowlie (2010); Cicala (2015)). On the other hand, the actual rate at which heat is converted to electricity (heat rate) has proven largely unaffected by the nature of regulatory oversight (Fabrizio et al. (2007); Wolfram (2005); Cropper et al. (2011)).

While market imperfections are certainly cause for concern, evidence of their existence is not proof of their inferiority (Joskow (2010)). The relevant question for policymakers considering what to do about the current regulatory situation is: do

[6]The ERCOT system in Texas is the exception because this Interconnection does not cross state lines, and is therefore not subject to FERC jurisdiction on many matters. However, Texas does participates in the North American Electric Reliability Corporation (NERC), which has been designated by FERC as the electricity reliability organization for the United States.

[7]Examples include Indiana, West Virginia, and parts of Kentucky in the Pennsyvania-Jersey-Maryland (PJM) Interconnection, most of the Midwest ISO (MISO), and all of the Southwest Power Pool (SPP).

markets (including all of their flaws) outperform the alternative methods for deciding which plants should operate in order to satisfy demand for electricity?

3 Measuring Welfare in Electricity Generation

The approach I use to measure welfare combines the within-PCA methods of Borenstein et al. (2002) (BBW), with the Mansur and White (2012) view of gains from trade across PCAs. Each PCA has a narrowly defined "merit order" in which the fixed, installed generating capacity is lined up in order of increasing marginal cost (effectively a supply curve for the area). Each generating unit has a nameplate rating that constrains the maximum amount of electricity it is capable of generating at any moment. Its cost per MWh is based on its heat rate, cost of fuel, and emissions fees, making the supply curve a step function.[8] "Economic dispatch" solves this constrained cost minimization problem to meet a given level of demand without damaging plants by exceeding their nameplate capacity.

To fix ideas, let $C_{pt}(Q_{pt})$ denote the observed cost of producing total quantity of electricity Q_{pt} in PCA p at hour t, which has N_{pt} MW of capacity installed. Further, define $C^*_{pt}(Q_{pt})$ as the cost of generation from the Q_{pt} lowest-cost MW of PCA p in merit order, indexed by i:

$$C^*_{pt}(Q_{pt}) = \sum_{i=0}^{Q_{pt}} c_{pt}(i) \tag{1}$$

$$where \quad Q_{pt} = \sum_{i=0}^{N_{pt}} q_{pt}(i); \quad q_{pt}(i) \in [0,1] \quad \forall \quad i$$

where $c_{pt}(i)$ is the cost of dispatching the i^{th} lowest cost MW in PCA p at time t.[9] Thus the observed cost of generation can be written as $C_{pt}(Q_{pt}) = \sum_{i=0}^{N_{pt}} c_{pt}(i)q_{pt}(i) = \mathbf{c'_{pt}q_{pt}}$, the inner product of costs and production as vectors in the merit order.

Out of Merit Costs

I will refer to a unit as operating "out of the merit order" when it is called on to operate to help meet Q_{pt} MW of demand although it is not one of the Q_{pt} cheapest MW of installed capacity based on its marginal cost. There are a number of reasons to fire

[8] Labor costs are unavailable, but relatively small compared to fuel costs. Commercial vendors of unit production cost data (such as SNL or Platts) often include a 10-20% markup over fuel costs to account for labor, operations, and maintenance costs.

[9] The unit dispatch problem partitions the N_{pt} MW of capacity into distinct units (with common costs), and chooses how much to generate from each unit subject to nameplate rating constraints. While this is an identical problem, indexing MW according to i creates a stable metric of the merit order, while indexing units themselves may shuffle as fuel prices vary.

up units that are out of merit: Plants must occasionally go off-line for maintenance, or are forced to shutdown unannounced, causing more expensive units to fill the gap. Transmission constraints may make it infeasible for the least-cost units to meet local demand. Large units require time and fuel to substantially change their output (ramping and start-up costs) which may exceed the cost of firing up a more nimble out of merit unit (Reguant (2014); Cullen (2011); Mansur (2008)). Large units may also continue operating when out of merit to prevent having to pay larger start-up costs from a cold start (idling). These are all real physical constraints that make out of merit operation the true cost-minimizing allocation of output. The cost of these constraints can be measured by the incrementally higher cost unit that must be used: $C_{pt}(Q_{pt}) - C^*_{pt}(Q_{pt})$.

This can be seen in Panel A of Figure III, which plots a hypothetical (smooth) supply curve against the perfectly inelastic demand that must be met in a particular moment to avoid a blackout. The welfare costs of dispatching units out of merit is simply the additional cost of output from these units relative to dispatching the lowest cost units installed in the area. It is important to emphasize that these are the *gross* costs, which are often incurred to avoid the even larger costs of following the strict merit order.

This out of merit loss is also the loss borne when market power is exerted. A firm may increase the market clearing price by taking an economical unit "down for maintenance," forcing an otherwise out of merit unit to operate (presumably to collect rents on co-owned inframarginal units). Because demand is completely inelastic (in real-time operations), the welfare loss is the incremental operating costs caused by taking economical units offline (Borenstein et al. (2002)).

It should be clear that legitimate maintenance, congestion, etc. is observationally equivalent from a welfare perspective to the exertion of market power–they differ by intent only. Mansur (2008) and Reguant (2014) note that failing to account for start-up and ramping costs will lead one to over-attribute the gap between the merit order and observed dispatch to market power when only accounting for normal maintenance and outages. The same is true when failing to account for transmission constraints (Ryan (2013); Borenstein et al. (2000)). I will side-step these issues completely by abstaining from assigning motives to the observed gap between idealized economic dispatch and what is observed in the data. Firms may well continue to exert market power, but also reduce downtime among low-cost units, as in Davis and Wolfram (2012)–my interest is in how the net of these impacts brings a PCA closer or farther

away from the merit order.[10]

Gains From Trade

When importing electricity from another area, one saves having to fire up a more expensive unit at the cost of the imports. When exporting, one gains any additional revenue beyond that required to generate the power. Panel B of Figure III considers the gains from trade between two areas as in Mansur and White (2012), effectively a fixed-factor Heckscher-Ohlin model. The red line continues to represent demand in the "Local" PCA of Panel A. Superimposed on this is the mirror image supply and demand figure from a "Foreign" PCA. The width of the x-axis is the sum of the demand of the two areas. If the two areas were to operate in autarky, the cost of meeting this demand would be the area under the upper envelope of the supply curves, meeting at the solid demand line. These two areas would reduce their joint production costs if they instead produced at the vertical dotted line, the lower envelope of their supply curves, as in any standard trade example.

The challenge in measuring these surpluses in this setting is that I do not observe with whom a PCA is trading–these simple bilateral examples do not exist in an interconnected electricity grid with indistinguishable electrons. Instead, I lean heavily on the following argument: PCAs pay (and are compensated) at the margin of their merit order. When I observe an area importing electricity (as in the Local PCA of Panel B), I infer that if they were paying more (less) than their marginal cost of generating, they would reduce (increase) their imports until these costs were in balance. Similarly, an exporting area must at least be covering its production costs–and if they are more than doing so, they would increase their exports until the analogous balance were reached. This may seem strong in the presence of transmission constraints until one considers these costs as part of the exchange: The inability to equate marginal generation costs reflects the shadow price of insufficient transmission. Thus I assume an importing area equates its marginal generation cost to the transmission-inclusive price of electricity generated elsewhere (and similarly for exporting areas). With this assumption I can measure PCA p's gains from trade in hour t with load L_{pt}, as S_{pt} by looking at each PCA individually without needing to know the source and destination

[10]The critique of Mansur (2008) remains if unit operators fail to account for the wear of ramping once facing prices in a wholesale market. In this case short-run operations will move closer to the merit order, but damage to the units will go unaccounted for. Such activities have been mentioned during personal interviews with market participants–but those costs did not remain hidden for long.

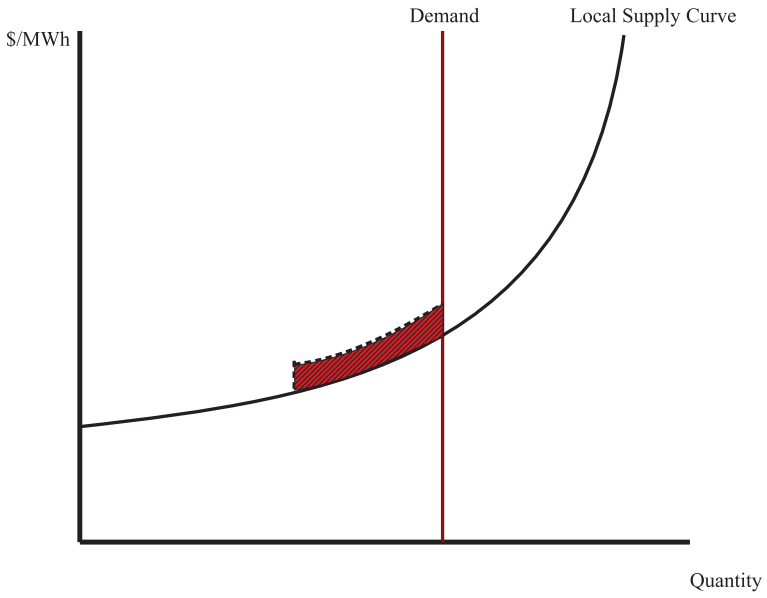
of each MWh traded:

$$S_{pt} = C^*_{pt}(L_{pt}) - C^*_{pt}(Q_{pt}) + c_{pt}(i = Q_{pt}) * [Q_{pt} - L_{pt}] \tag{2}$$

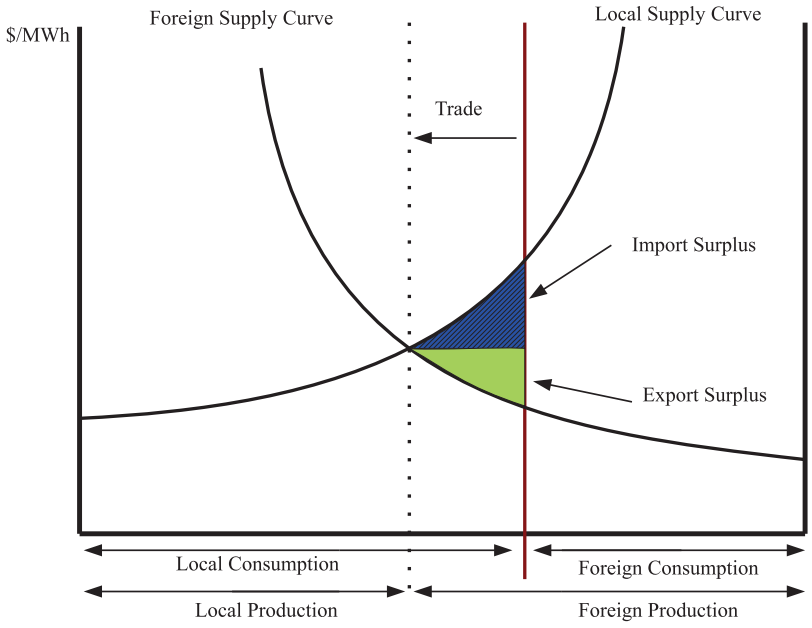
The difference of the first two terms is the cost of meeting load according to the merit order, as if in autarky, net of the merit order cost of the actually observed production. The rectangle between supply and demand is formed at the marginal merit order cost of production, on net yielding the triangle below the supply curve between supply and demand in an area that is importing, and the triangle above the supply curve between demand and supply for an exporting area.

Figure III: Welfare Measurement in Electricity Markets

(a) Out of Merit Losses



(b) Gains From Trade



4 Data

This study draws from a disparate and incongruous set of data sources to synthesize an essentially complete characterization of U.S. electricity production at the hourly, generating unit level from 1999-2012 (over 530 million unit-hour observations). This section presents an overview of the data, while the details of data construction can be found in the Data Appendix.

Hourly Load Data

The demand side consists of a balanced panel of hourly load (consumption) from the 98 major U.S. power control areas (PCAs) that dispatched power plants in 1999 to meet demand. This data has been reported annually to the Federal Energy Regulatory Commission on Form 714, "Annual Electric Balancing Authority and Planning Area Report." Record-keeping challenges at FERC requires this data to be supplemented with equivalent data from regional authorities and markets (Western Interconnection, ERCOT, PJM, NYISO, NEISO, and NERC). In instances that original administrative data is unavailable (or reporting policies/boundaries change), I employ LASSO to estimate missing demand based on weather, population, and employment. Combined with cross-validation to maximize out-of-sample accuracy, this procedure delivers predictions within 4% of the realized values on average (see the Data Appendix). Small municipal authorities that do not actually conduct dispatch of fossil- or nuclear-powered plants are added to the load of their principal suppliers or customers, yielding 98 total PCAs.

Figure IV summarizes the electricity load data. The US consumes a bit less than 4,000 TWh (billions of kilowatt-hours) annually. Panel A shows that electricity consumption increased from 1999 until the Great Recession, and was relatively flat through 2012. Panel A also highlights the seasonal nature of electricity usage: summer cooling and winter heating can increase usage by over a third of temperate seasonal usage on a month-to-month basis, with much larger swings during peak usage. Panel B plots hourly usage over the course of the week, averaged over the 14-year study period. Here too there are large swings in usage both over the course of the day and the week. The key fact to remember when interpreting these figures is that production must move exactly in sync with these demand swings, and that utilities must have enough generation capacity to meet demand at the moment of peak usage. Thus every downward swing also represents vast quantities of generating capacity becoming idle.

As a demonstration of real-time patterns of demand, I have animated one year's worth of hourly load here. This animation shows the East-to-West flow of electricity

demand as usage follows local clocks. It also reflects the daily and seasonal patterns shown in Figure IV, while highlighting the substantial variation around these averages: peak demand can be as much as 2.5 times average annual usage, can be quite persistent during summer months in the South and Southern Plains, and generally varies less in temperate areas of the Pacific Northwest.

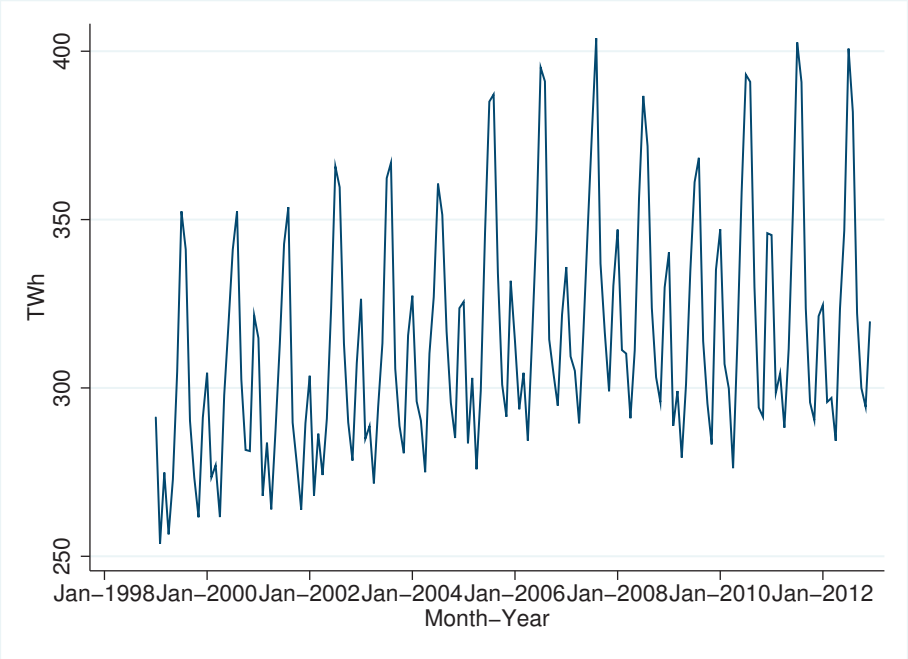
Hourly Generation Data

The supply side is based on data from the Energy Information Administration (EIA), merged with hourly gross generation reported to the Environmental Protection Agency (EPA) with Continuous Emissions Monitoring Systems (CEMS), as well as daily production at nuclear-powered units from the Nuclear Regulatory Commission (NRC). Boilers from the EPA are matched to generators' monthly net generation and heat rates via Forms EIA-767 and EIA-923, "Annual Steam-Electric Plant Operation and Design Data / Power Plant Operations Report." Hourly production in the data is the gross generation from CEMS scaled by the ratio of monthly gross-to-net generation from EIA at the unit level. I then merge this data on heat rates and hourly production with coal and oil fuel costs under a non-disclosure agreement with the EIA (from Forms EIA-423, "Monthly Report of Cost and Quality of Fuels for Electric Plants," EIA-923 and Form FERC-423, "Monthly Report of Cost and Quality of Fuels for Electric Plants"). This is shipment-level data, reported monthly by generating facilities with a combined capacity greater than 50MW. I use spot-market coal prices to measure the opportunity cost of coal burned rather than contract prices. Natural gas prices are from 65 trading hubs around the country reported by Platts, Bloomberg, and NGI (not EIA), and are quoted daily. Plants are linked to their nearest trading hub along the pipeline network. Areas with emissions markets for Sulfur and Nitrogen Oxides include the cost of pollution based on measured emissions and monthly market prices from BGC Environmental Brokerage Services.

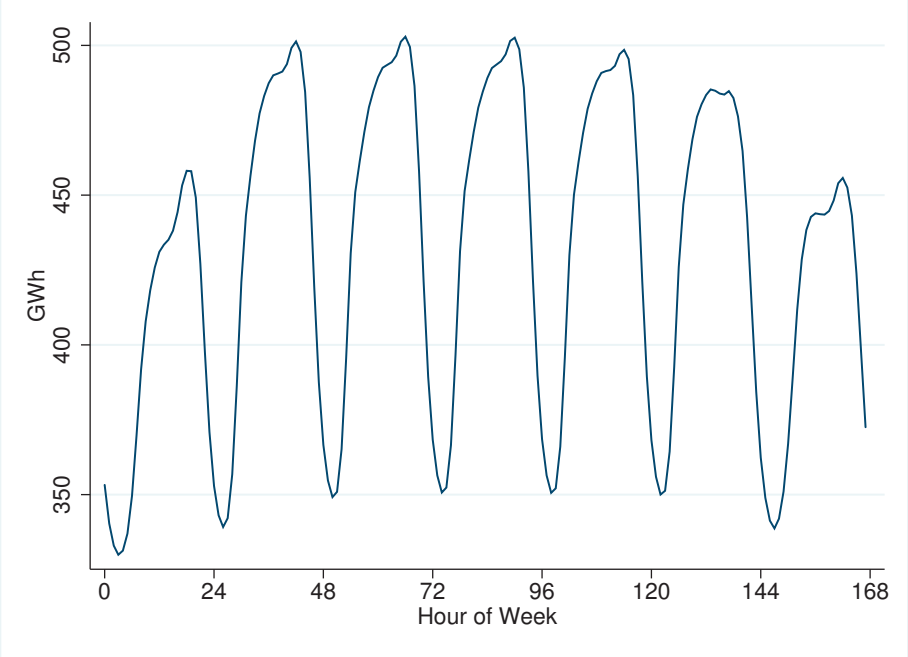
Generation from hydro-powered units either comes directly from from the source (i.e. Tennessee Valley Authority, U.S. Bureau of Reclamation, etc.), or is based on the streamflow of the nearest downstream gage from the U.S. Geological Survey's Streamgage Network (linked through analysis of the National Hydrography Database). Because the cost of reservoir-based hydropower is the opportunity cost of the water, I price hydropower based on the marginal cost of fossil generation in the merit order that is being supplanted. Run-of-river hydro is priced at zero. Hydropower units >10MW were classified as reservoir or run-of-river based on internet searches and/or satellite images.

Figure IV: Electricity Load over Time

(a) Total Monthly Load



(b) Average Load by Hour of Week



Hourly generation is unavailable from a number of smaller fossil-fired units (whose net generation rarely exceeds 3% by NERC region-year). Power from these units is distributed across the hours of the month in an intuitive manner: having produced nW MWh in a month, where W is the unit's nameplate capacity, I assume that the unit produced at maximum capacity during the n hours of highest demand observed over the course of the month. This replicates the behavior of a dispatcher who employs a threshold rule of when to generate from a unit (assuming no start-up costs or ramping constraints), while allowing observed behavior to dictate what threshold was employed each month.

Figure V presents the aggregate annual statistics for electricity generation in the US. Coal supplied the energy for roughly half of the electricity generated from 1999 - 2012, but has been in decline since 2007. From that time, natural gas has grown from rough parity with nuclear (20%) to 30%, almost entirely at the expense of coal-fired generation. Following a nearly three-fold increase from 1999 - 2008, Panel B shows that fossil fuel expenditures fell by approximately 50% from the peak in 2008 from the combined effects of reduced demand overall and the massive reallocation of output to units burning cheaper gas thanks to the advent of hydraulic fracturing (Hausman and Kellogg (2015); Linn et al. (2014); Knittel et al. (2014)). Fossil fuel expenditures averaged about \$72B/year over these 14 years, thus the complete dataset tracks the burning of \$1T of fuel at the plant-generation unit-hour level.

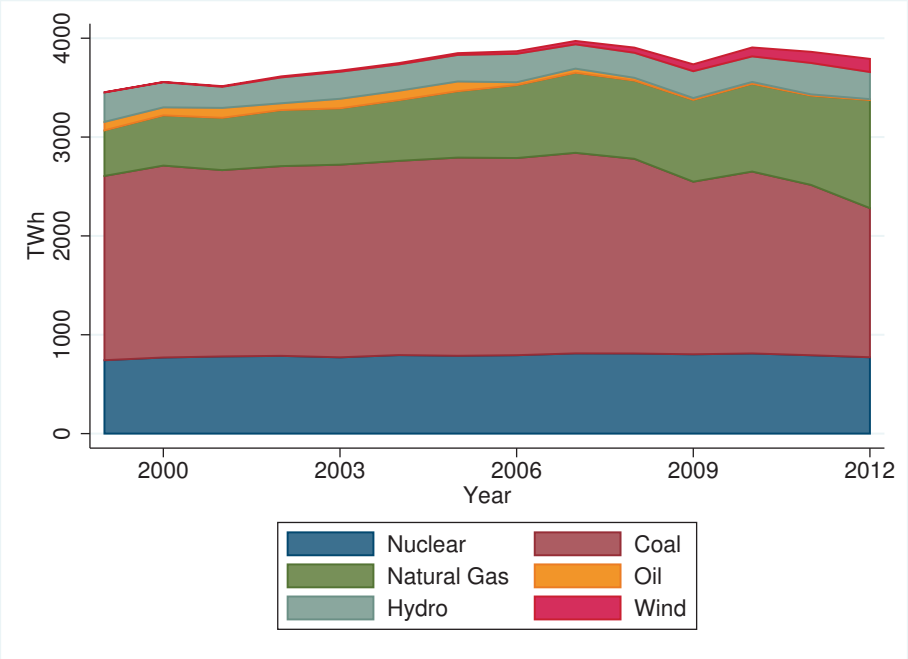
Matching Supply and Demand

Because the supply data is built up from microdata independently from the demand side, it is important to ensure congruence between the data sources–there is nothing institutional about their reporting to ensure they agree. Beginning with the 1999 configuration of the electrical grid, I match plants to their initial PCA from the EPA eGRID database. New capacity since that time is matched to PCA either directly or based on historical utility service territory in the case that the PCA territory has changed. These associations are then checked based on power plant names reported by PCAs in FERC 714. I then compare the implied monthly totals from the supply side of the data against those reported by the PCAs to FERC. In total, about 99% of reported generation from FERC 714 can be accounted for in the supply-side data. About 3% of net generation does not fit neatly in to a single power control area because multiple PCAs report a share of output from large plants as their own. In these cases, the plant is assigned to the PCA with greatest dispatch authority.

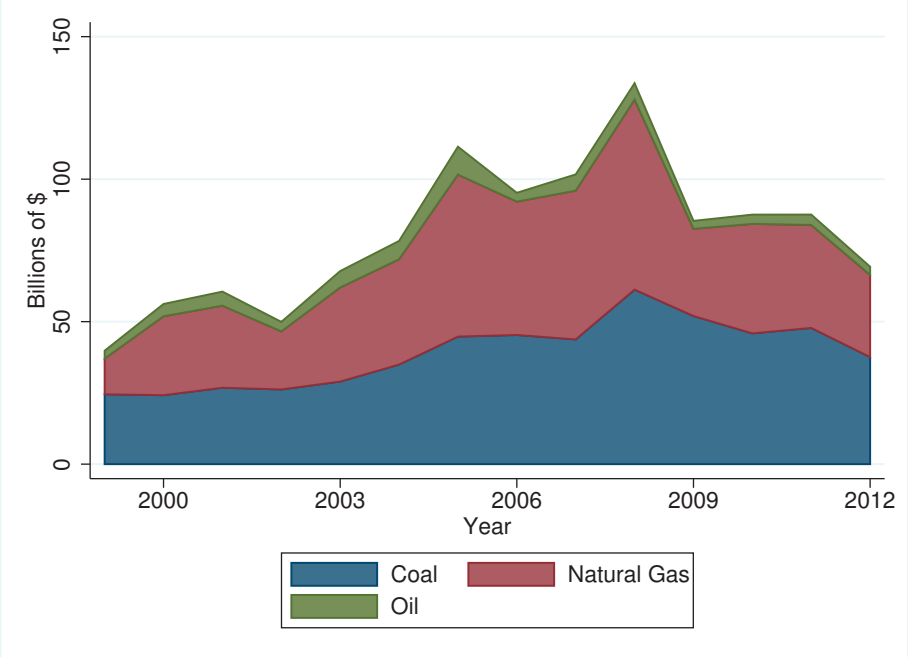
Figure VI breaks down generation by data source, and shows the quality of the

Figure V: Annual Net Generation and Fuel Cost by Source

(a) Net Generation



(b) Fuel Cost



match between supply and demand. The top black line in Panel A is identical to the total monthly load shown in Panel A of Figure IV. After totaling the generation observed (or calculated) based on high-frequency data, the remaining numbers reported at the monthly level result in totals that almost exactly match the demand side of the data. Panel B gives a closer view of what is missing by calculating the gap (as imports or exports) every hour across PCAs, then adding them separately up to the monthly level, measuring the volume of trade across areas. The first striking statistic is that roughly 90% of generation is effectively consumed in its local PCA–while PCAs are interconnected, they continue to largely produce energy for their own consumption. To my knowledge, these statistics are new: regulatory bodies typically report the *net* flow of electricity between areas, which fails to reflect the real-time interdependence among PCAs (or lack thereof).

The remaining gap between imports and exports as I observe them is due to imports from outside of the US (which have grown over this period to about 1% of supply (Energy Information Administration (2012) Table 2.13). Based on the framework presented in Section 3, not observing this generation effectively values it as an import from outside of each PCA, which is valued as displaced local generation. The production costs and exporter surpluses (mostly from Canada) are outside of the data.

5 Estimating Counterfactual Operations

I use the staggered timing of market creation and expansion to arrive at an estimate of the causal impact of the transition to market-based electricity dispatch. These events are defined as the PCAs formally ceding control of their transmission system to an Independent System Operator, who conducts auctions to allocate output to generating units. As demonstrated in Figure II, these are discrete events–typically demarcated prominently in the history of each market. These events suggest a DD approach, using areas without regulatory change to estimate counterfactual outcomes after one has adjusted for common shocks and time-invariant differences:

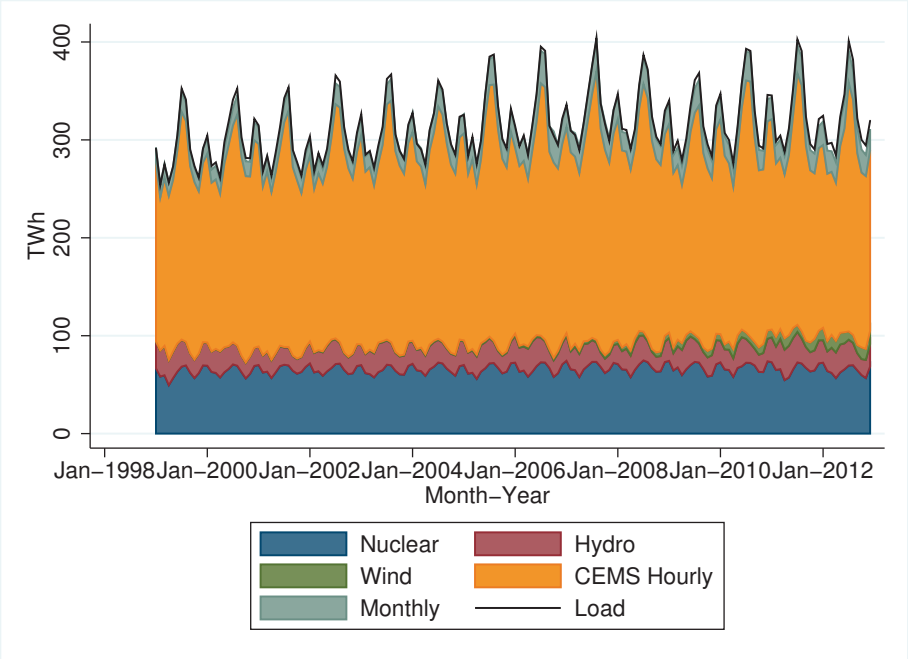
$$y_{pt} = \tau D_{pt} + \gamma_p + \delta_{tr} + \varepsilon_{pt} \tag{3}$$

where y_{pt} is the logged value of the outcome variable, and D_{pt} is an indicator of market dispatch. This approach may also allow PCA-specific coefficients on flexible measures of demand (taken to be exogenous since consumers do not face the real-time cost of electricity). [11] The time fixed effects δ_{tr} are included at the date-hour-region

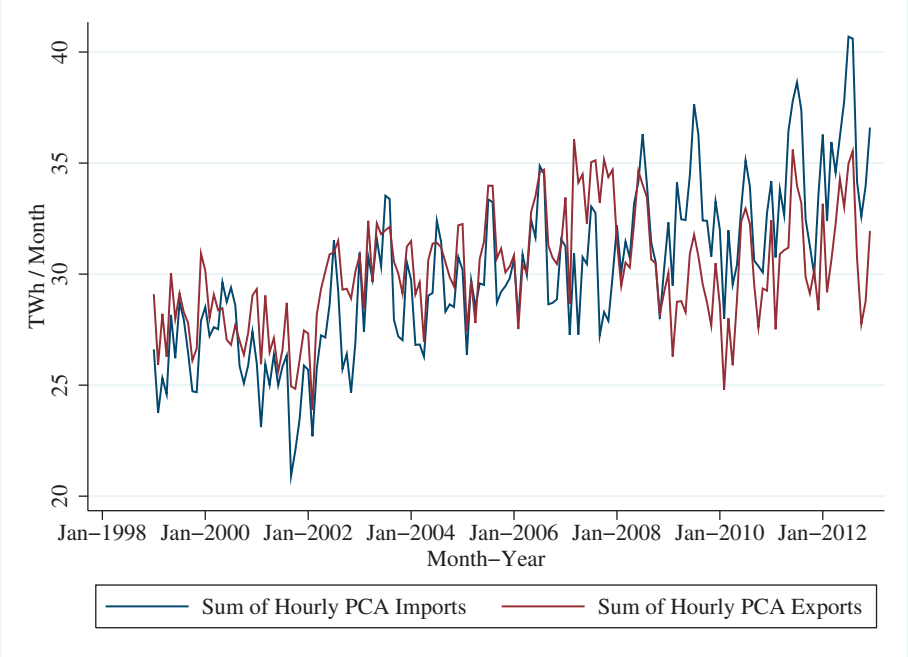
[11] This would account for heterogenous time-invariant relationships between the outcome vari-

Figure VI: Annual Net Generation and Fuel Cost by Source

(a) Monthly Net Generation by Data Source



(b) Monthly Trade Across PCAs



level to account for spatial and time-varying unobservables, particularly with respect to fuel prices (Cicala (2015)). τ measures the average effect of market dispatch, and should be interpret as an Average Treatment on the Treated (ATT)–it measures the effect in the areas that have adopted market dispatch. Interpreting this as an Average Treatment Effect (ATE) requires the stronger assumption that PCAs in the South and West have the same potential benefits from market integration–rather than the continued business-as-usual assumption required for the validity of the ATT. One should keep in mind that markets themselves are heterogenous, and their rules change over time. Thus a single "treatment effect" of markets as conceived here takes the average of these various institutional changes, compared to the various institutions that preceded the transition to market dispatch.

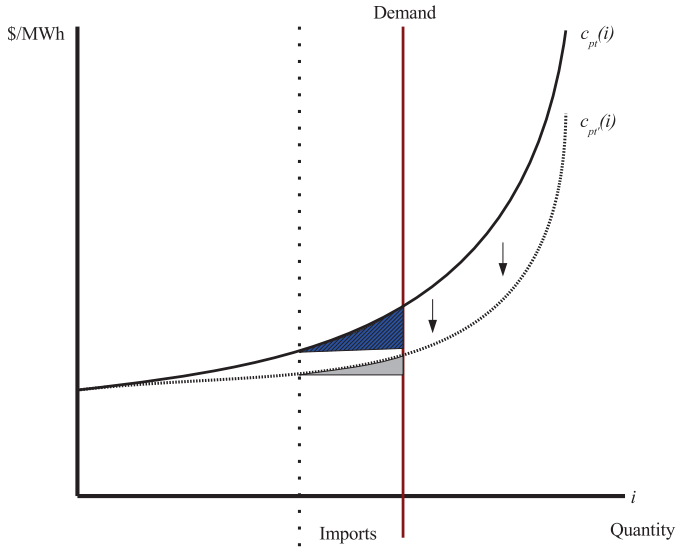
A Policy Function Approach to Counterfactuals

The causal effect of markets on gains from trade or out of merit dispatch is the difference between an observed outcome in a market area and what that outcome *would have* been but for the market–holding production capacity, fuel costs, and demand fixed. Although DD forms a natural starting point for the analysis, it is insufficient to simply estimate the change in outcomes following market introduction, even relative to areas without any regulatory change: Within a PCA, outcomes (holding demand fixed) are confounded by varying fuel prices over time, which change the cost of operating a given unit out of merit, or the value of offset production through trade. Contemporaneous differences across areas are confounded by the fact that PCAs differ in their installed capacity, and are therefore differentially affected by common time-varying shocks.

To illustrate this problem, Figure VII presents the import gains of the "local" supply curve from Figure III. The lighter grey addition represents the import gains realized in this same area, but under a different set of fuel prices, represented by the dashed curve. Though there may be no differences in production between these two curves, the difference in prices yields different gains from trade. If one curve were realized during regulation, and the other after the introdutction of markets, a simple difference would indicate that gains from trade had changed, although there had been no change in behavior. Because each PCA differs in the composition of units, common fuel price shocks affect areas differently, comparing changes in neighboring areas will

able and demand–the fact that some areas are more prone to congestion in times of high demand, for example. In addition, controlling for demand also accounts for the possibility that differences in outcomes across areas might be driven by regional trends (such as population) instead of how dispatchers allocate production taking demand as given.

Figure VII: Cost Changes Unrelated to Deregulation Confound Counterfactual Estimates



Note: This figure shows how measured gains from trade change with the price of fuel, holding demand and traded quantities constant. Using gains from trade in period t' as the counterfactual for what would have happened in t in the absence of treatment would yield a predicted change in outcome in spite of no behavioral change.

fail to correct for this confounding.

I propose a policy function approach that builds upon the "generation regressions" of Davis and Hausman (2016) to overcome this issue: I use historical patterns of unit-level production given load, unit capacity, and position in the merit order to estimate predicted allocations of production.[12] I apply these predicted quantities to observed unit costs to estimate what production costs would have been if not for treatment.

Let a policy function for PCA p in year y be the probability that the PCA orders generation from the i^{th} MW of capacity of the merit order in hour t, conditional upon covariates X_{ipt} (such as load, month of year, hour of day, and nameplate capacity of the unit producing the i^{th} MW) and treatment, D_{pt}

$$\psi_{py}\left(i, X_{ipt}, D_{pt}\right) = Pr\left[q_{pt}(i) = 1 | X_{ipt}, D_{pt}\right] \tag{4}$$

[12]Here, policy refers to a rule that maps states in to actions, without any reference to the optimality of that rule, as is typically implied in the use of this term in the control theory literature.

With this notation, a PCA can be expected to produce the total amount

$$\mathbb{E}(Q_{pt}|X_{ipt}, D_{pt}) = \sum_{i=0}^{N_{pt}} \psi_{py}\left(i, X_{ipt}, D_{pt}\right)$$

Expected costs of production are based on the inner product of costs and the policy function in vector form, $\mathbf{c_{pt}}\prime\psi_{\mathbf{pt}}\left(\mathbf{X_{pt}}, D_{pt}\right)$.

To operationalize these policy functions for causal inference, some assumptions are required. To economize on notation I adopt the 'Potential Outcomes' framework popularized by Rubin (1974), in which a generic outcome of can be thought of taking on value Y_{pt}^0 in the absence of treatment, and Y_{pt}^1 if treated. Thus estimating the causal impact when Y_{pt}^1 is observed requires estimating Y_{pt}^0, which is not. Here the outcomes being evaluated are functions of production allocations and costs, capacities, and demand: $Y^D = F\left(\mathbf{q_{pt}^D}, \mathbf{X_{pt}^D}\right)$, such as gains from trade as denoted in equation (2).

Assumption 1. *Demand, unit production costs and capacities are invariant to treatment:*

$$X_{ipt}^0 \quad = X_{ipt}^1 \quad = X_{ipt}$$

This assumption narrows the set of potential outcomes to focus on the question, how does market dispatch affect the allocative efficiency of meeting demand? Real-time pricing for retail customers is nearly nonexistent during the sample period, so that consumers' behavior is invariant to hourly production costs. Although I have shown elsewhere (Cicala (2015)) that prices paid for coal (but not gas) depend on plant-level regulations, this study focuses on allocative efficiency changes–how production moves across power plants holding costs fixed. The brief time horizon evaluated after the introduction of market dispatch is intended to hold the capital stock fixed so that the observed supply function for each area is invariant to treatment.

Assumption 2. *Parallel trends in unobservables and evolution of the policy function:*

$$\begin{aligned} Y_{pt}^{0} &= F\left(\mathbf{q_{pt}^{0}}, \mathbf{X_{pt}}\right) \\ &= F\left(\psi_{\mathbf{p,y-1}}^{\mathbf{0}}, \mathbf{X_{pt}}\right) + \underbrace{F\left(\mathbf{q_{pt}^{0}}, \mathbf{X_{pt}}\right) - F\left(\psi_{\mathbf{p,y}}^{\mathbf{0}}, \mathbf{X_{pt}}\right)}_{\textit{Contemporaneous Error}} + ... \\ &\quad + \underbrace{F\left(\psi_{\mathbf{p,y}}^{\mathbf{0}}, \mathbf{X_{pt}}\right) - F\left(\psi_{\mathbf{p,y-1}}^{\mathbf{0}}, \mathbf{X_{pt}}\right)}_{\textit{Change in Policy Function}} \\ &= F\left(\psi_{\mathbf{p,y-1}}^{\mathbf{0}}, \mathbf{X_{pt}}\right) + \delta_{tr} + \gamma_{p} + \upsilon_{pt} \end{aligned}$$

where $E\left(\upsilon_{pt}\right) = 0$

This assumption forms the basis of the estimation strategy, using the allocation of production based on operations in year $y-1$ to predict operations in year y. There are two forms of error with this approach: the difference between the true outcome and the value based on the contemporaneous policy function, and the difference induced by the evolution of policy functions from year-to-year. Assumption 2 decomposes these errors in to a PCA-specific, time-invariant component, a regional contemporaneous shock, and noise. This allows, for example, for out-of-sample predictions based on last year's operations to persistently be off by an amount that varies by PCA, while also accounting for contemporaneous regional shocks to fuel prices.

Assumption 3. *Conditional Independence of Treatment for Control Outcomes and Policy Function Measurement Error*

$$Y_{pt}^{0}\,,\ \left(\psi_{\mathbf{p,y-1}}^{\mathbf{0}} - \hat{\psi}_{\mathbf{p,y-1}}^{\mathbf{0}}\right) \perp\!\!\!\perp D_{pt} | X_{pt}$$

That treatment is conditionally independent of control outcomes allows the identification of an average treatment on the treated (ATT). The second part of this assumption ensures that using estimated values of counterfactual outcomes will not bias estimates of the treatment effect. Rather than including these estimates as a generated regressor, this assumption allows a modified DD-type estimating equation in which the dependent variable is the departure from the outcome predicted by the estimated policy function:

$$Y_{pt} - F(\psi_{\mathbf{p,y-1}}^{\mathbf{0}}, \mathbf{X_{pt}}) = \tau D_{pt} + \delta_{tr} + \gamma_{p} + \upsilon_{pt} \tag{5}$$

There are a number of potential threats to the validity of this research design. First and foremost, the stable unit treatment value assumption (SUTVA) requires

that the treatment status of markets that become PCAs does not affect the outcomes of other areas. This will be violated, for example, if the expansion of PJM facilitates the delivery of electricity from the Tennessee Valley Authority (TVA), which is not dispatched by markets. Using TVA as a control PCA will understate the true effect of market dispatch when their exports change due to the policy change. This estimation framework also assumes that outcomes change immediately with the change in treatment status. However, sudden massive changes tend not to be conducive to keeping the lights on. The pre-period may be contaminated if PCAs began to change their dispatch policies in preparation for the transition to markets. On the other hand, the treatment effect may take time to fully manifest itself as PCAs learn how to use the market to improve their operations (or exert market power).

As is standard in DD research designs, unrelated, differential trends between treatment and control also threaten the validity of estimates. Aside from including PCA-specific trends, the policy function approach mitigates this issue by transforming the dependent variable in to the residual of behavior predicted by the prior year's policy. This kicks the threat of differential trends up one level (for quantities, not prices), requiring an unrelated trend in how well last year's policy matches that of this year. Such problems become evident with event study-style estimates leading up to the time of treatment.

On interpretation, the supply curves I construct are based on fuel and emissions prices, while variable labor, operations, and maintenance costs are ignored. Although these other costs are small relative to total variable cost, they create distance between my measured merit order and the true marginal cost of power. The treatment effect on the costs I observe may be well-measured, but it will be a biased estimate of the overall change in allocative efficiency if something about the transition to market dispatch changes these errors–such as reduced labor costs in markets as in Fabrizio et al. (2007). The small share of non-fuel costs multiplied by the modest impact of restructuring renders the potential magnitude of this bias quite small. There is likely to be greater measurement error concerning exact fuel prices and unit capacities. I reduce these errors to the extent possible by using daily gas prices at geographically disperse hubs (to account for pipeline congestion), and by using the implied capacity based on observed operations from CEMS (maximum hourly net generation by season) rather than the round figures reported to EIA. Again, these errors bias my causal estimates only to the extent that they are non-stationary and correlated with market dispatch.

Regarding inference, estimates using this approach are presented with standard

errors calculated by block-bootstrapping PCA-months, with regular DD estimates clustered at the PCA-month. This reflects the thought experiment that the observed data (a complete census of operations) is drawn from a super-population of operations to allow for the inference of potential outcomes–and that each months' fluctuations in demand allow for an independent observation for each PCA. If one believes that there are truly really only 98 (PCA) independent observations, the reported standard errors roughly double. Conversely, the standard errors become infinitesimal if one follows the existing literature, having studied one area at a time with independence assumed across fine time units.

Machine Learning Estimation of the Policy Functions

The policy function approach removes the role of fuel price variation in the estimation of counterfactual outcomes for a given allocation of output: Instead it is the quantities themselves that are predicted, then applied to the observed prices to calculate counterfactual behavior.

Estimating the policy functions requires balancing flexibility and risk of overfitting. On one hand, the probability of running a unit is a complex, unknown function of the variables system operators use to make decisions–simple approximations are unlikely to deliver high-quality predictions of behavior. On the other hand, overly-flexible specifications may provide the illusion of superior fit, but perform poorly out-of-sample. Since the estimated treatment effect comes from changes in the quality of fit between predicted and observed behavior, it is particularly important to prevent overfitting from showing up as illusory treatment effects.

This is a pure prediction problem, for which recent tools from machine learning are well-suited. I use the "random forest" algorithm of Breiman (2001) as implemented by Wright and Ziegler (2016). This nonparametric estimation algorithm draws bootstrap samples of the data and calculates means of the outcome variable for random partitions of the explanatory variables. It then aggregates these weak predictions across the bootstrap samples to form robust estimates without functional form assumptions.

More formally, for PCA p and year $y-1$ with sample size $N_{p,y-1}$, random forest draws $N_{p,y-1}$ pairs (q_{ipt}, X_{ipt}) with replacement from that PCA-year. It then "grows" a regression tree as follows: starting from a single node, it randomly selects a set of variables from $\mathbf{X_{p,y-1}} \subseteq \mathbb{R}^p$ where p is the dimension of $\mathbf{X_{p,y-1}}$. It then splits the data along these dimensions at cut-points that make the subsequent nodes as homogenous as possible with respect to the outcome (Breiman et al. (1984)), forming two nodes. Each of these nodes are subsequently split using the same method until a pre-specified

(and here, cross-validated) number of observations remain at each final node, referred to as leaves (or perfect uniformity is achieved). Using θ_m to denote the random vector used to draw the bootstrap sample and determine which explanatory variables are used to split at each node of tree m, the tree produces a set of leaves $l = 1, ..., L$ that partition the space of explanatory variables ($\mathbb{R}^p$) in to rectangular subspaces, $\mathbb{R}_l$. The prediction of the tree given a particular x is obtained by averaging over the outcomes of the observations in the leaf to which x belongs, $l(x, \theta_m)$. Following Meinshausen (2006), the prediction for a vector of covariates x can be thought of as a weighted mean of the entire sample of the original data, depending upon each observation's inclusion in the bootstrapped sample and terminal leaf position

$$\hat{\psi}_m(x) = \sum_{i=1}^{N_{p,y-1}} w_i(x, \theta_m) q_i$$

$where$

$$w_i(x, \theta_m) = \frac{1\left\{X_i \in \mathbb{R}_{l(x,\theta_m)}\right\}}{\sum_{j=1}^{N_{p,y-1}} 1\left\{X_j \in \mathbb{R}_{l(x,\theta_m)}\right\}}$$

with $1\{\cdot\}$ denoting an indicator function that is one when the statement in the braces is true, zero otherwise. The prediction from a single tree provides a poor prediction–it does not use all of the underlying data and over-fits the data it does use–Breiman (2001) shows that as the number of trees grown in this way increases, the quality of out-of-sample predictions stabilizes.[13] Continuing with the weighted-average interpretation, one draws a number of iid θ_m vectors to grow M trees, then calculates the final weights each observation receives in the final prediction as

$$w_i(x) = \frac{1}{M} \sum_{m=1}^{M} w_i(x, \theta_m)$$

Predictions for policy functions are made out-of-sample for year y for data with explanatory variables $\mathbf{X_{p,y}}$ by calculating

$$\hat{\psi}_{\mathbf{p,y-1}}(\mathbf{X_{p,y}}) = \frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{N_{p,y-1}} w_i(\mathbf{X_{p,y}}, \theta_m) q_i$$

[13] Scornet et al. (2015) establish the consistency of random forests grown in this way as estimators of the conditional expectation function in the presence of an additive error. Wager and Athey (2016) establish consistency and asymptotic normality results more broadly in the context of causal inference using the unconfoundedness assumption for estimating treatment effects conditional upon terminal leaf partitions, and review related results.

where each i indexes observations of the production data of PCA p in year $y-1$.

The core motivation for methods such as random forest from the machine learning literature has been its performance in out-of-sample prediction. This remains true in this setting as well, as demonstrated in Figure VIII. The metric of performance here is the out-of-sample residual sum of squares, divided by that of a simple OLS regression of unit operations on the covariates used in the random forest estimation (separately by PCA, including month and hour of day as dummies). The x-axis separates units by their position in the merit order, as a percentile of costs of installed capacity for each PCA-hour to create a common scale. The figure is constructed using data from areas without market dispatch to show quality of fit in the control group. The dashed line shows the performance of a more flexible OLS specification: a second-order polynomial of all terms, estimated separately by PCA for each month and hour of day. While this specification fits the data better than the simpler one, random forest far outperforms throughout the merit order. It delivers a superior fit to observed operations uniformly across the merit order, and is particularly good at predicting baseload and peaking operations.

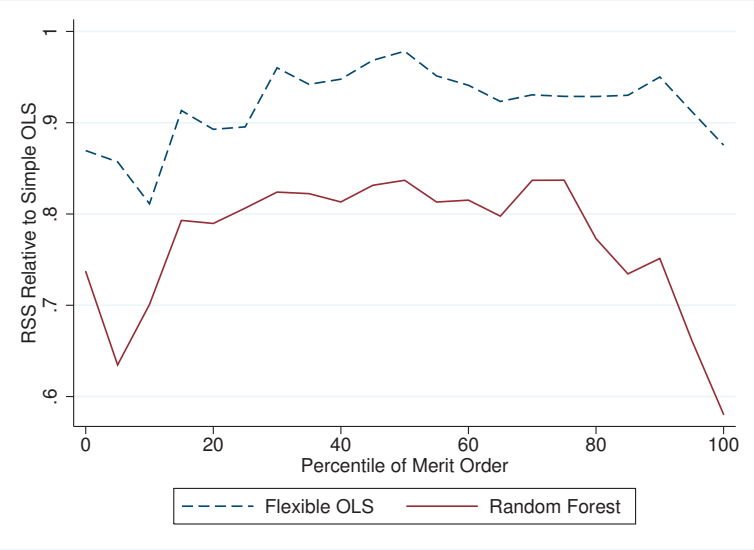
It is important to note that this estimation framework has been designed so that *all* predicted values of the policy function come from out-of-sample estimates. The treatment effect is based on *changes* in how well last year's operations predict this year's operations. Using the change in how well observed behavior in year $y-1$ fits predictions estimated during year $y-1$ against using year $y-1$'s predictions for year y risks baking-in an in-sample/out-of-sample change in fit.[14] It also requires iterative estimation of placebo treatment dates among areas that never receive treatment, an exceptionally high computational burden in this setting (or the assumption of time invariance of overfitting issues).

6 Results

Tables I through IV present the main results as Average Treatment on the Treated estimates to measure the impact of market dispatch on allocative efficiency in electricity production. The first columns are based on straight DD estimates that includes date-hour-region and PCA fixed effects. The second column flexibly controls for the

[14] Burlig et al. (2016) deal with this issue by randomly selecting a placebo date to separate in-sample/out-of-sample data in the control group, then including an indicator of out-of-sample prediction. This makes the estimated treatment effect the relative deterioration of fit going from in-sample to out-of-sample (non-contemporaneously). The approach taken here avoids making assumptions on the in-sample/out-of-sample transitions, instead evaluating the quality of out-of-sample predictions made from contemporaneous training periods. The cost of this approach is that a year of baseline outcomes lacks out-of-sample predictions.

Figure VIII: Relative Prediction Quality of Random Forest in Control Group



Notes: This figure plots the relative residual sum of squares (RSS) based on out-of-sample predictions of random forest and a flexible OLS specification where the numeraire is the RSS of an OLS specification with linear terms and indicators for month and hour of day. Explanatory variables include: position in the merit order, nameplate capacity, and load. All predictions are estimated separately by PCA.

effect of load on the outcome variables (allowing a different slope for each quartile of each PCA's load distribution). This permits each area to have persistent idiosyncratic relationships between demand and how it goes about meeting that demand with out of merit generation and trade. The third column adds PCA-specific time trends. The final column transforms the outcome variable to be the difference between the observed outcome and that predicted by the policy function, as described in Section 5.

All specifications also include separate dummies for greater than 24 months prior, and greater than 24 months after the transition to markets. This serves two functions: For the first three specifications, this prevents long-term responses to market dispatch (and potential confounders) from loading on to the short-term DD estimates. For the policy function estimates, "treatment" only occurs when predicting behavior for a period with a different status of market dispatch. I predict from the year before dispatch out two years afterwards (and year-on-year otherwise). Subsequent predictions are based on behavior after markets have begun, making treatment effectively an impulse during this initial window. A post-24-month indicator allows this new period to have a different mean than pre-treatment. Changes in observation counts between these tables indicates the extent to which PCAs operate exactly according to my measure of the merit order: zeroes are dropped in the logarithmic specifications when the merit order is followed so that no generation is out of merit. The drop in observations between the DD and policy function specifications in Tables I through IV is because the baseline period is held out to ensure all observations for the policy function estimates are from out-of-sample calculations.[15]

Beginning with quantities, Table I indicates a roughly 10% increase in traded volumes following the adoption of market dispatch. Because these numbers look PCA-by-PCA, an increase in exports in one area will be complemented with increases in imports in other areas, but not show up additively in the coefficients–and thus do not double-count trade volumes. These estimates are relatively stable across specifications, and do not change in a statistically significant manner when using policy functions to predict counterfactual operations. This is also true for out of merit generation, with the exception of a drop when including PCA-specific trends. However, estimates return to their original levels in the final specification, suggesting the coarser linear trend projects over changes that are more subtly accounted for with

[15]To avoid losing the first year completely (which includes the New York and New England transitions), the held-out data is from every-other day for the first year. The impacts in these markets were relatively large, but dropping them does not change the overall estimates substantially.

policy functions.

One striking measure of the work being done here by the policy functions is to compare the R^2 of the models across specifications. Removing the outcomes predicted by the machine learning algorithm leaves substantially less variation in the dependent variable, and the control variables have far less power in explaining the variation that remains.

To ensure these results are not the artifacts of pre-existing time trends, Figure IX estimates the model of column (2), including separate dummies for each month measuring the time until (or since) market dispatch adoption. Note that this specification only measures the effect for the initial transition to market dispatch: performance changes among incumbents (with whom the area is trading) following market expansion are not included. While not as clean as one might like for event study-style figures, they make clear that the overall estimates are not due to long-term trends. There is an overall level shift in Panel (a) corresponding with the onset of treatment, while it appears the slide to a new, lower level of out of merit generation occurs over a few months.

Tables III and IV estimate the welfare impacts of this reallocation of output based on changes in production costs. These results indicate substantial impacts of market dispatch: over 30 log points for gains from trade, and 20% reductions in out of merit costs. One should note the substantial reduction in observations between the two tables. This is because the specifications in Table III condition upon positive gains from trade: in roughly 25% of PCA-hours, there is sufficiently little trade that both supply and demand land on the same generator, which yields zero surplus (as described in Section 3).

That the value of this trade exceeds the change in volume implies that there is a substantial gap between the cost of electricity whose production increases versus that being displaced–equating the marginal cost of power across areas would yield zero net benefits of an additional MWh traded. Similarly for out of merit costs, these results imply that it is the relatively expensive out of merit units whose production is reduced by market dispatch.

Figure X presents the main results on cost reductions relative to the onset of treatment. The additional volatility of coefficients in these figures relative to the quantity estimates highlights the dependence of the welfare estimates on fuel prices mentioned above: production costs scale with input prices in this Leontief setting, so volatility in fuel prices is directly translated in to volatility of the welfare impacts of a given change in behavior.

While there is a nice jump in gains from trade in Panel (a), longer-term trends play a more prominent role in the overall shape of the plots than in Figure IX. That differential trends in fuel prices might confound estimates as presented in Figure X motivates the policy function estimates of Figure XI. This figure plots the analogous event-time coefficients, but with the dependent variable transformed in to the residual between observed outcomes and those predicted by the policy function. Although this approach adds volatility relative to the straight DD estimates, it makes clear that the estimated treatment effects are not due to differential trends: there are unambiguous breaks in the series for both outcomes and an absence of pre-trends. The timing of these breaks also correspond with the transition to markets, though reductions in out of merit costs begin the month prior to market dispatch.

There are two possible contributing factors to why the results are lower for the policy function approach for gains from trade. First, estimating a slightly smaller shift in output naturally yields smaller cost reductions (less output is offset). Second, as highlighted in Figure VII, the simple DD estimates are potentially confounded by fuel price changes. Higher fuel prices steepen the supply curve, yielding greater gains from trade compared to periods with lower prices. Straight DD estimates make such comparisons, while the basis of the policy function approach is to compare a given supply curve to itself.

With over \$4.2B in trade surplus in market PCAs annually, applying the gains from trade treatment uniformly over the treated territories is worth nearly \$1B/year. For out of merit costs, these estimates applied to the \$10B accrued in market PCAs raises the overall impact of this institutional change on cost reductions to about \$3B per year.

Table I: Market Dispatch on *Log*(Trade Volume)

	(1)	(2)	(3)	(4)
Market Dispatch	0.125***	0.118***	0.095***	0.098***
	(0.021)	(0.020)	(0.021)	(0.029)
Log(Load)		Yes	Yes	Yes
PCA Trend			Yes	Yes
Policy Function				Yes
Clusters	16464	16464	16464	15910
PCAs	98	98	98	98
R^2	0.557	0.595	0.613	0.099
Obs.	12001882	12001882	12001882	11352820

Note: All specifications include PCA and Region-Date-Hour Fixed Effects. Demand controls are PCA-specific. Standard errors clustered by PCA-Month in parentheses. * p<0.1, ** p<0.05, *** p<0.01

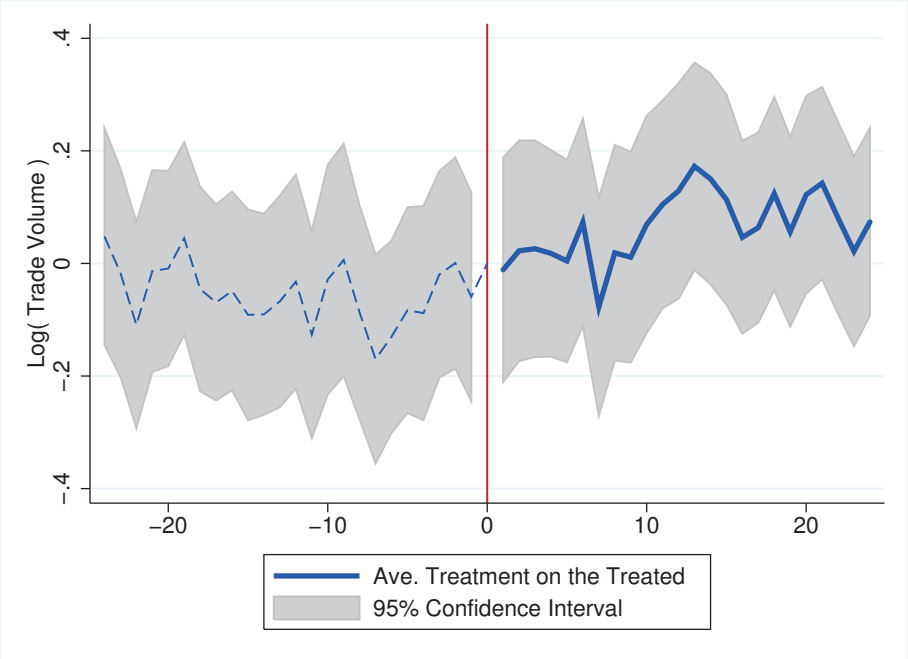
Table II: Market Dispatch on *Log*(MWh Out of Merit)

	(1)	(2)	(3)	(4)
Market Dispatch	–0.111***	–0.113***	–0.061***	–0.117***
	(0.018)	(0.017)	(0.017)	(0.019)
Log(Load)		Yes	Yes	Yes
PCA Trend			Yes	Yes
Policy Function				Yes
Clusters	16448	16448	16448	16430
PCAs	98	98	98	98
R^2	0.841	0.852	0.863	0.157
Obs.	11648909	11648909	11648909	11428353

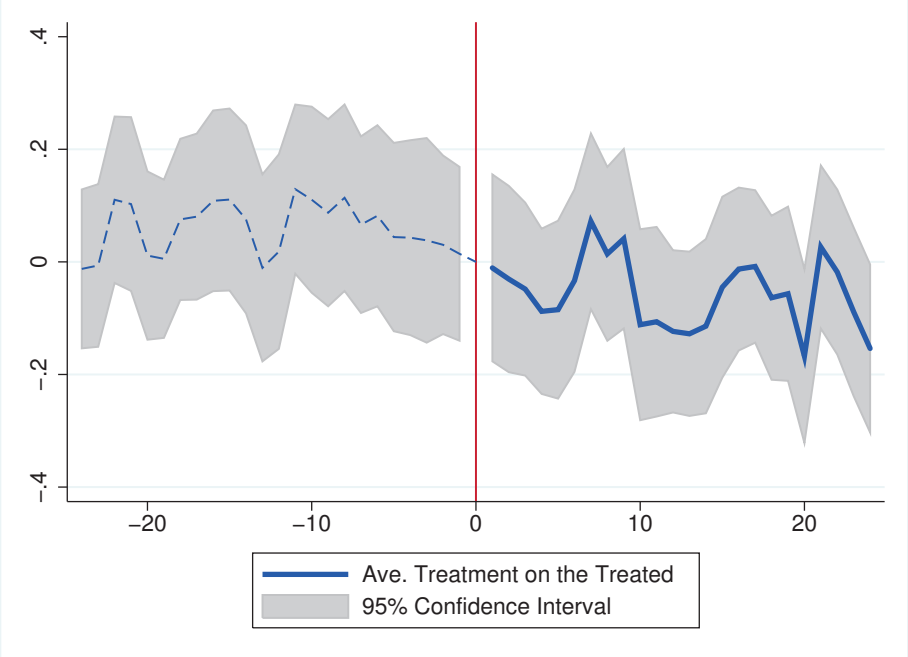
Note: All specifications include PCA and Region-Date-Hour Fixed Effects. Demand controls are PCA-specific. Standard errors clustered by PCA-Month in parentheses. * p<0.1, ** p<0.05, *** p<0.01

Figure IX: Treatment Effects by Months to Market: Quantities

(a) *Log*(Trade Volume)



(b) *Log*(MWh Out of Merit)



Note: These figures are based on regressing logged outcomes on a set of indicator variables for each month until (after) the transition to market dispatch, PCA-specific controls for load, date-hour-region and PCA fixed effects. The month prior to treatment is normalized to zero. Confidence intervals are based on clustering at the PCA-month level.

Table III: Market Dispatch on *Log*(Gains from Trade)

	(1)	(2)	(3)	(4)
Market Dispatch	0.312***	0.350***	0.258***	0.191***
	(0.045)	(0.042)	(0.041)	(0.048)
Log(Load)		Yes	Yes	Yes
PCA Trend			Yes	Yes
Policy Function				Yes
Clusters	16424	16424	16424	15814
PCAs	98	98	98	98
R^2	0.494	0.579	0.617	0.125
Obs.	8671235	8671235	8671235	8098935

Note: All specifications include PCA and Region-Date-Hour Fixed Effects. Demand controls are PCA-specific. Standard errors clustered by PCA-Month in parentheses. * p<0.1, ** p<0.05, *** p<0.01

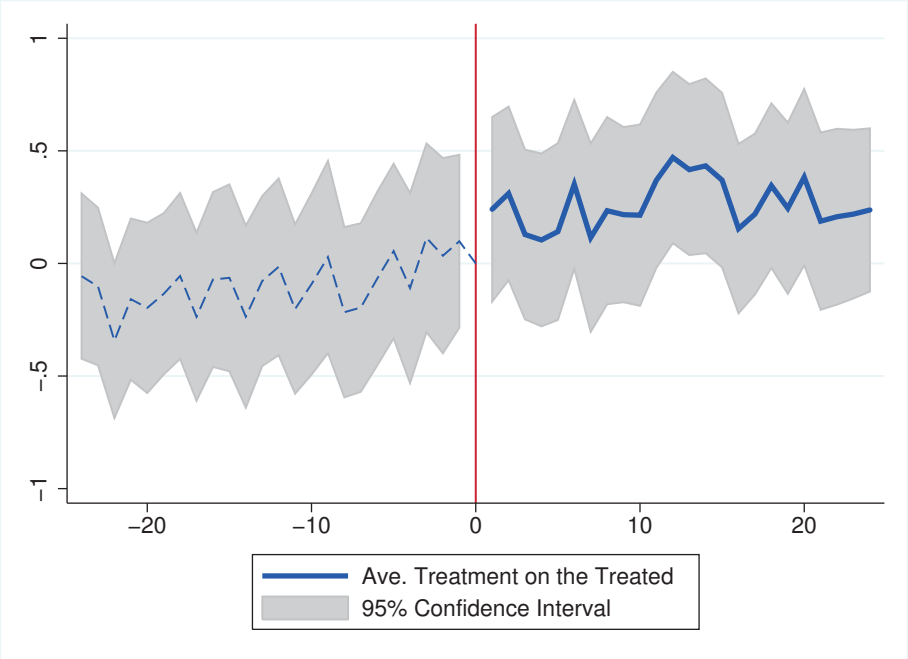
Table IV: Market Dispatch on *Log*(Out of Merit Costs)

	(1)	(2)	(3)	(4)
Market Dispatch	–0.186***	–0.164***	–0.009	–0.179***
	(0.034)	(0.033)	(0.033)	(0.033)
Log(Load)		Yes	Yes	Yes
PCA Trend			Yes	Yes
Policy Function				Yes
Clusters	16450	16450	16450	16444
PCAs	98	98	98	98
R^2	0.775	0.793	0.812	0.302
Obs.	11648731	11648731	11648731	11427266

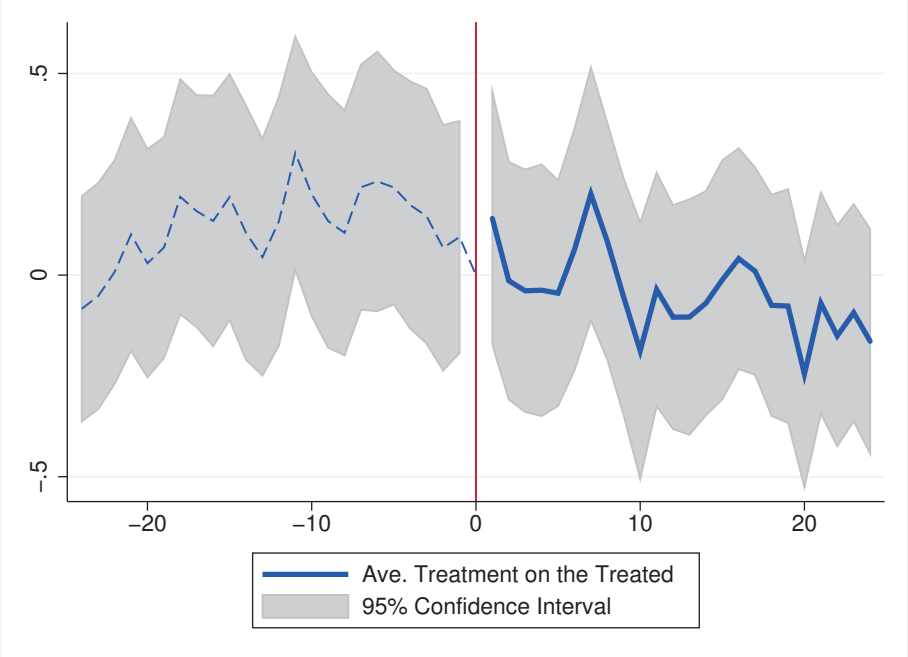
Note: All specifications include PCA and Region-Date-Hour Fixed Effects. Demand controls are PCA-specific. Standard errors clustered by PCA-Month in parentheses. * p<0.1, ** p<0.05, *** p<0.01

Figure X: Treatment Effects by Months to Market: Welfare

(a) *Log*(Gains from Trade)



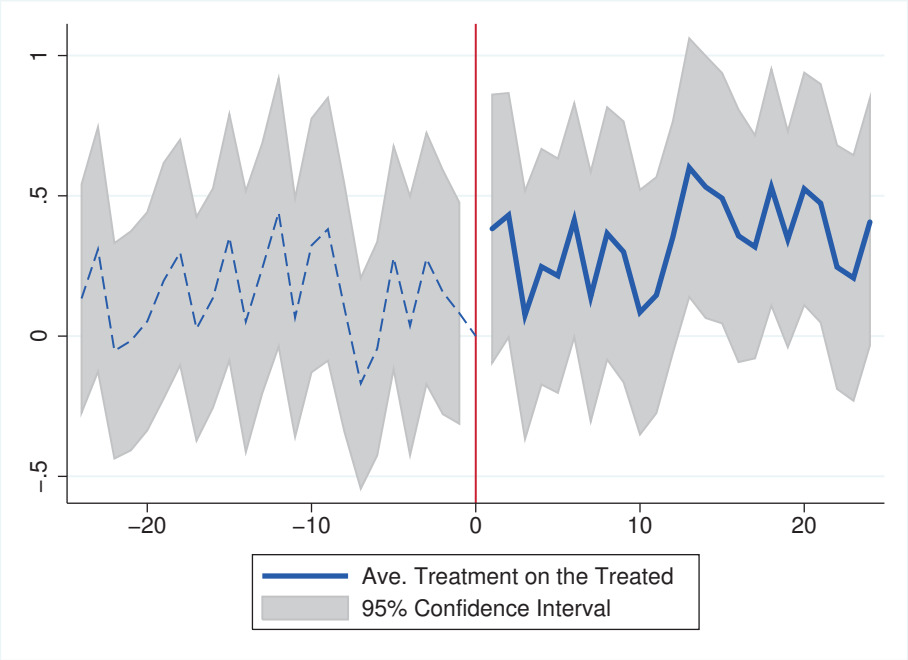
(b) *Log*(Out of Merit Costs)



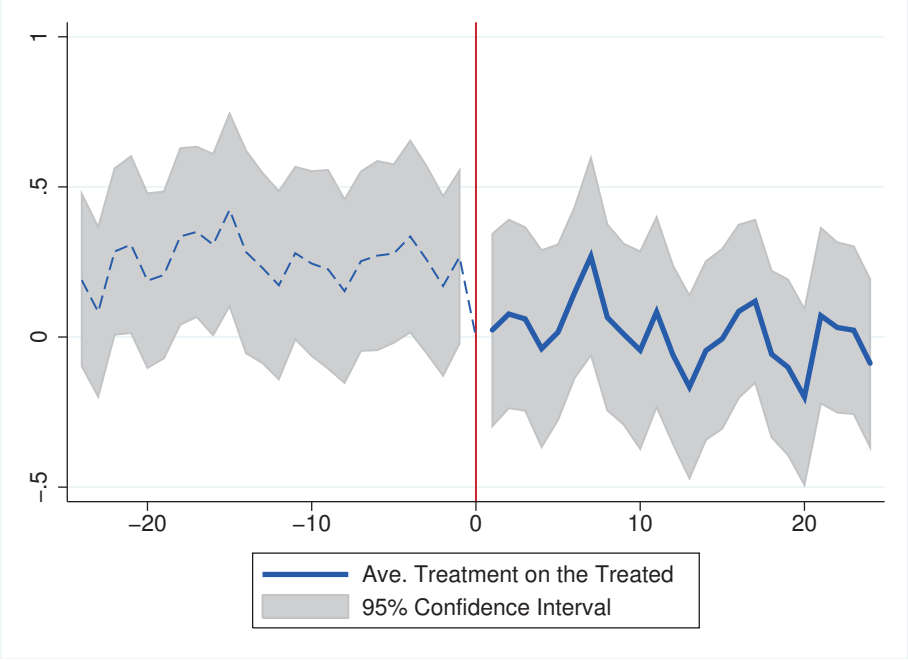
Note: These figures are based on regressing logged outcomes on a set of indicator variables for each month until (after) the transition to market dispatch, PCA-specific controls for load, date-hour-region and PCA fixed effects. The month prior to treatment is normalized to zero. Confidence intervals are based on clustering at the PCA-month level.

Figure XI: Policy Function-Based Treatment Effects by Months to Market

(a) *Log*(Gains from Trade)



(b) *Log*(Out of Merit Costs)



Note: These figures are based on regressing the difference between logged outcomes and those predicted by the policy function on a set of indicator variables for each month until (after) the transition to market dispatch, PCA-specific controls for load, date-hour-region and PCA fixed effects. The month prior to treatment is normalized to zero. Confidence intervals are based on clustering at the PCA-month level.

Heterogeneity over the Year

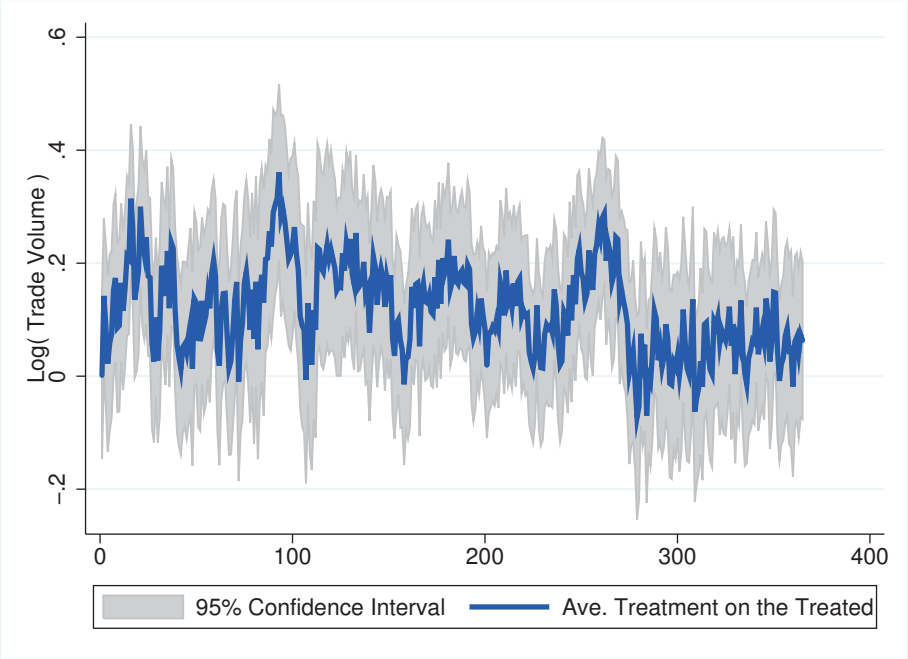
The richness of the data allows for the examination of heterogeneous treatment effects in order to better understand the forces driving the overall point estimate. In that spirit, Figures XII and XIII interact the treatment variable with the day of the year, and plot the corresponding coefficients. For quantities, Figure XII shows the strong complementarity between the two measures: The biggest reductions in out of merit generation occur during the low demand periods utilities use to perform maintenance on their large units (and refuel nuclear-powered units). How do they manage to reduce these outrage costs? Panel (a) shows that trade volumes increase during these periods in market areas. This indicates that markets keep utilities from favoring their own higher-cost units during maintenance, and instead coordinate supply of lower-cost power across PCAs. These results complement the prior findings of Davis and Wolfram (2012) that merchant nuclear units reduce their down-time overall by showing that markets facilitate reducing production costs for the down-time that remains.

For gains from trade, Figure XII shows how fuel prices are not simply confounders, but also drivers of treatment effect heterogeneity. There are substantial increases in gains from trade during peak summer months even with smaller trade volumes because more expensive units' production is supplanted with traded generation. The overall treatment effects measured earlier are in fact weighted averages of quite large treatment effects during shoulder seasons and summer, but much smaller effects during the winter months.

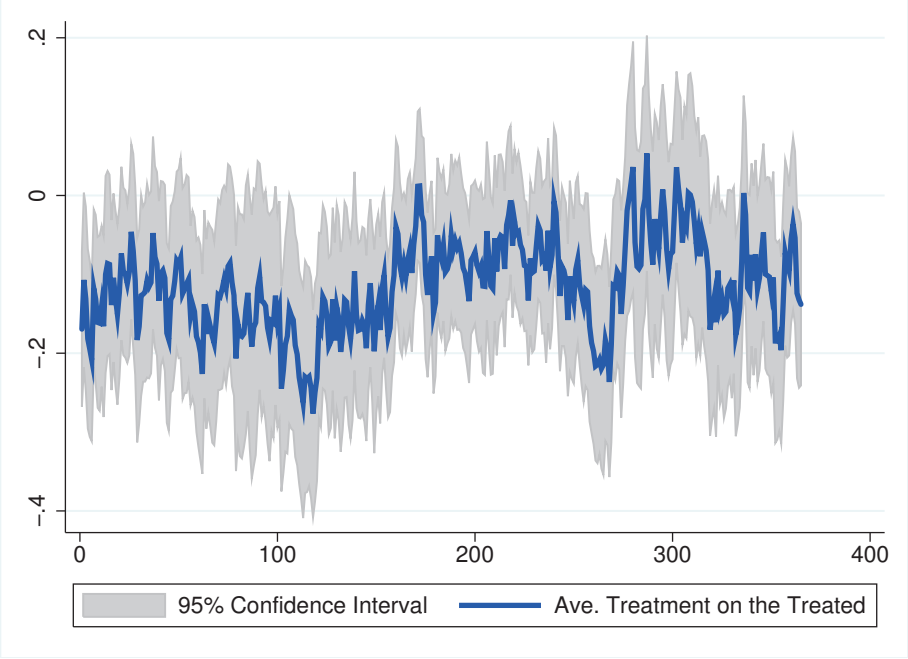
Panel (b) highlights the key opposing forces at play when switching to a regulated area: On one hand, generators have an increased incentive to ensure their low-cost generators are available for production, on the other hand peak periods of demand create the potential to profitably exercise market power by taking an economical unit offline. Thus even though there is a reduction in the quantity of out of merit generation in XII.b, the reduction in out of merit costs is low relative to the value of offset generation (high during these peak periods). On net, diligent market monitoring has made these strategies of withholding production more difficult, and the effect overall is reduced out of merit costs throughout the year.

Figure XII: Treatment Effects by Day of Year: Quantities

(a) *Log*(Trade Volume)



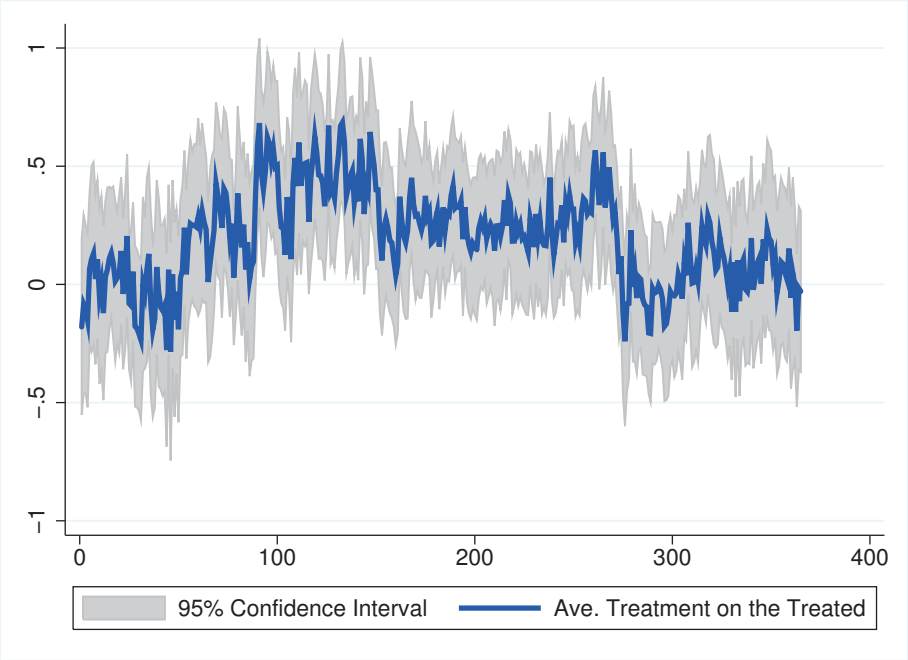
(b) *Log*(MWh Out of Merit)



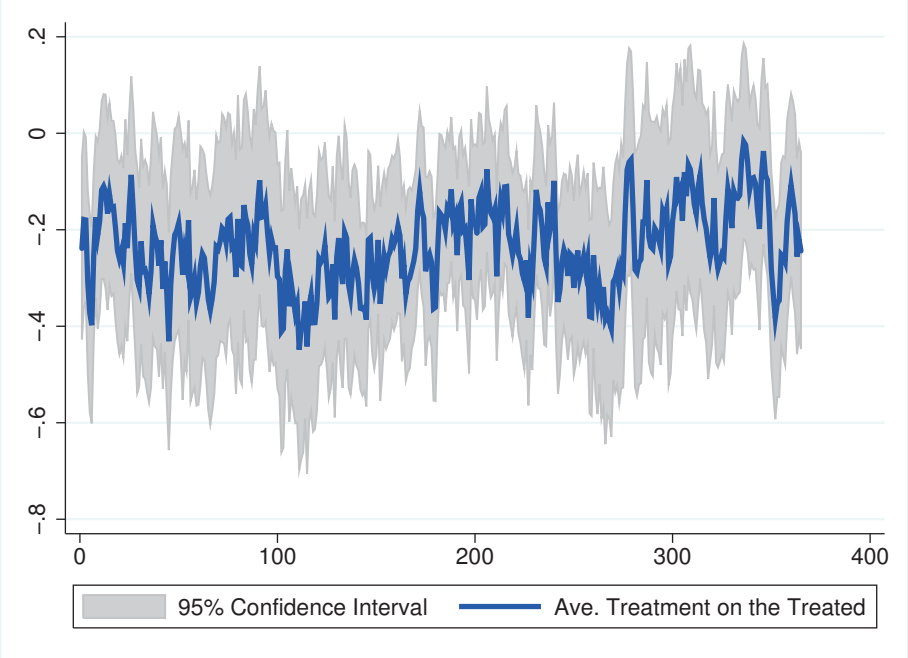
Note: These figures are based on regressing logged outcomes on a set of indicator variables for each day of the year interacted with market dispatch, along with date-hour-region and PCA fixed effects. Confidence intervals are based on clustering at the PCA-month level.

Figure XIII: Treatment Effects by Day of Year: Policy Function Estimates

(a) *Log*(Gains From Trade)



(b) *Log*(Out of Merit Costs)



Note: These figures are based on regressing the difference between logged outcomes and those predicted by the policy function on a set of indicator variables for each month until (after) the transition to market dispatch, PCA-specific controls for load, date-hour-region and PCA fixed effects. The month prior to treatment is normalized to zero. Confidence intervals are based on clustering at the PCA-month level.

7 Conclusion

In this paper I use the recent introduction of wholesale electricity markets in some areas as a natural experiment to evaluate the performance of markets relative to the policy-relevant counterfactual: centralized dispatch by a regulated private or government local monopolist. In constructing a fourteen year panel of hourly operations, I am able to infer gains from trade at any moment of time based on the amount of electricity being produced and consumed in an area, and the installed generating capacity that might have been used to equate local supply and demand. Observing production hourly at the generating unit level allows me to calculate the difference between actual production costs, and those that would have been realized if only the most economical (based on marginal fuel cost) units were utilized. I estimate how the introduction of wholesale markets affected these two measures of welfare, interpreted as the net impact of market power problems and improved coordination on production costs. I find that market-based dispatch has caused a roughly 20% increase in the gains from trade due to reallocated production across power control areas, while also reducing out of merit costs by 20%–a reduction in production costs of about $3B per year.

While the estimated allocative efficiency improvements caused by market dispatch are substantial, they are likely part of a much bigger story. These short-run estimates are based on responses to institutional changes imposed on a grid that was built for reliability rather than massive trans-regional exchange. This inherently imposes an upper bound on the potential gains that might be observed with this estimation strategy, but is a constraint that may be relaxed over time as locational marginal prices reveal profitable transmission investments.

About 40% of electricity in the United States continues to be generated by plants called upon to operate based on the decision-making of the local balancing authority. Policymakers are therefore faced with the question of whether markets should be expanded or scaled-back. The answer depends on the balance between market failures and regulatory shortcomings. While market power is certainly a concern for market monitors (Wolak (2012) shows their work is critical), my results suggest the benefits realized by more efficient allocation of output though market-based dispatch have far outweighed such losses.