Changing Conduct with Changing Demand: Evidence of Coordination among Suppliers of Electricity in California in 2000

Nauman Ilias*, Robert Reynolds

The Brattle Group, Washington DC

Abstract

A number of empirical studies have shown that the market power exercised by the California generators was a significant cause for the high prices of wholesale electricity in California in 2000. Some studies have found that the generators' behavior was consistent with unilateral market power as opposed to collusion, particularly under high-demand conditions. The intuition for this result is that the California market was characterized by price-inelastic demand and short-term capacity constraints. Therefore, firms could profitably withhold capacity unilaterally without the need to engage in collusive behavior. This paper shows that the California generators were indeed withholding significant amounts of capacity, even though that capacity was capable of providing energy at a cost below the prevailing price. However, we also show that the generators behavior varied with demand conditions. In particular, we find that the behavior was more collusive during periods with intermediate levels of demand (*i.e.* when demand was neither too low nor too high), while it was consistent with unilateral pricing in very low and very high demand hours. We compute withholding and conjectural variation parameters by deciles of demand, and find that both measures are higher on average for the middle deciles relative to the low and high deciles.

JEL classification: L1; L4

---

* Corresponding author.
E-mail address: nauman.ilias@brattle.com

1. Introduction

California's experiment with deregulation of the electricity industry suffered a set-back in the summer of 2000 when wholesale electricity prices increased dramatically. The prices were higher nearly five-fold in the summer of 2000 relative to the same period in 1998 and 1999. This elevation in prices triggered a collapse of the organized trading market by the beginning of 2001 as the investor-owned utilities, which were the net purchasers of power in California, were not allowed to pass on the increased wholesale prices to retail consumers. This caused the utilities to lose their creditworthiness. The State of California had to intervene and purchase power on behalf of the utilities to "keep the lights from going out".

A number of empirical studies have attributed the increase in price in 2000 to the exercise of market power by the five generators ("Big-5") in California that were created as part of the deregulation process. Joskow and Kahn (2002) conducted an analysis of price-cost margin and a comparison analysis of supplier behavior in the California market. Using publicly available data, they estimated competitive wholesale market benchmark prices and compared the benchmark prices to the prices that were actually observed. The competitive price benchmark used was the short-run marginal cost of supplying electricity from the most expensive generating unit that was required to clear demand.[1] The authors found a significant price gap in that the actual prices were significantly higher than the competitive benchmark in June through September of 2000. Since the benchmark price accounts for market fundamentals such as costs of generation, total demand, and imports into California,[2] the authors attributed the price gap to the market power of the Big-5 generators. Borenstein, *et al.* (2002) and Hildebrandt (2000) found similar results on price gaps using confidential data and an empirical approach analogous to that of Joskow and Kahn (2002). Borenstein, *et al.* (2002) concluded that 50 percent of the total electricity expenditures could be attributed to market power in the summer of 2000.

In their analysis of supplier behavior, Joskow and Kahn (2000) showed that the Big-5 generators withheld supply from the market when it would have been profitable for a generator

---

[1] This approach to measuring market power in wholesale electricity market was pioneered by Wolfram (1999) in her study of electricity market operating in U.K. The same approach has been applied in other studies of the California market, e.g. Borenstein, Bushnell and Wolak (2002) and Hildebrandt (2000).

[2] The market fundamentals accounted for were: (i) prices for natural gas, which is an input in the production of electricity from steam and combustion turbines; (ii) costs for Nitrogen Monoxide (NOx) emission, which is a by-product of the generation process; (iii) load levels which capture the total demand; and (iv) and import levels which represents out-of-state supply into California.

without market power to supply more. The authors examined the hours in which the market-clearing price was high enough for the California generators to economically supply their capacity. They found that there was a substantial "output gap" between the maximum possible levels of generation and observed levels, consistent with the notion of market power that the generators were withholding supply to elevate prices. This conclusion is also supported by other studies of the California market. For example, Puller (2007) showed that, very often during the summer of 2000, the Big-5 firms observed prices above marginal cost, yet failed to utilize their available capacity.

Although there is evidence of some form of market power, less is know about the type of pricing regime that led to the exercise of market power. The conditions prevailing in the California market were such that the market power exercised by the suppliers could have been due to unilateral withholding on the part of generators and/or tacit collusion among them. Such characteristics as the non-storable nature of electricity, very low short-run demand elasticity, capacity constraints, and a large fraction of demand being satisfied in the spot market in California, incentevized the suppliers to unilaterally withhold output to drive up prices when demand was high. At the same time, attributes such as the homogeneity of the product, knowledge of rivals' production costs, publicly available data on market-clearing prices and demand forecasts, as well as repeated interactions among the suppliers, could have facilitated collusion among the generators. An understanding of the underlying pricing game is important for the design of electricity markets. For example, if market power is likely to be exercised by collusive means, the policy maker can change market rules so as to make it difficult for the suppliers to reach and sustain a collusive agreement.

A couple of studies have attempted to distinguish whether the market power in the California market was unilateral or collusive. Using data on the bids submitted by suppliers, Wolak (2003) computed hourly elasticity of residual demand facing each of the Big-5 firms from 1998 to 2000. He used the average hourly value of the inverse of the firm-level residual demand elasticity over the period from June to September of each year as a summary measure of the extent of unilateral market power of each generator. For each firm, Wolak found that his measure of unilateral market power was significantly higher in 2000 relative to the corresponding values in 1998 and 1999. He concluded that the higher prices in 2000 were, therefore, a result of unilateral market power rather than collusion. However, Wolak did not

consider whether the computed firm-level residual demand elasticity was consistent with the price-cost margin faced by each firm. For the same value of demand elasticity, a higher margin implies a more collusive pricing regime. Therefore, one would have to examine the extent to which margins were higher in the summer of 2000 relative to the two prior years to determine whether the market power was indeed unilateral rather than collusive.

In another study, Puller (2007) conducted empirical analysis to test whether firm-level behavior was more consistent with unilateral market power or tacit collusion. He compared actual prices to simulated prices under three benchmark models of competition – competitive, Cournot, and joint monopoly pricing for peak hour 18 (5-6 PM) from 1998 to 2000. He found that for hour 18, actual prices were very close the price that would result if all Big-5 firms acted as Cournot competitors in all three years. Puller also computed a firm-level conjectural variation parameter for hour 18 by econometrically estimating the supply function for each Big-5 firm. The average value of the conjectural variation parameter, also referred to as the conduct parameter, quantifies the level of collusion in the industry. He found that although individual firm behavior was slightly less competitive in the second half of 2000, there was generally no evidence of tacit collusion among the firms and the hypothesis of Cournot pricing could not be rejected. Puller concluded that the high prices in 2000 were driven more by changes in costs and less elastic demand than by changes in firm conduct.

This study contributes to the existing literature on market power in the California electricity market by demonstrating that the conduct of the generators varied with demand conditions in the summer of 2000. In particular, we show that the Big-5 generators' behavior was more collusive during periods with intermediate levels of demand (*i.e.* when demand was neither too low nor too high), while it was consistent with non-cooperative Cournot pricing in very low and very high demand hours. Taking advantage of very detailed and reliable data that were made available during the litigation that followed the electricity crisis, we measure withholding by the Big-5 generators and also estimate a conjectural variation parameter to quantify the average level of collusion. To examine changes in firm conduct with changing demand, we divide the hourly data from the summer months of 2000 into deciles based on demand level, and compute the average level of withholding and the conduct parameter for each decile. We find that withholding follows an inverted U or hump-shaped pattern in demand. In particular, average withholding increases initially with demand deciles, reaching a maximum

value at the 6th decile, before declining for higher decile levels. Secondly, we find that the conjectural variation parameter also shows an inverted-U shaped pattern in demand. Specifically, the conduct parameter takes values around 1 for lower deciles (1st and 2nd), ranges between 2 and 3 for the middle deciles, and decreases to values around 1 for higher deciles (9th and 10th). A value of 1 for the conduct parameter corresponds to Cournot pricing, while a value of 2.5 corresponds to the case when the five firms are pricing like a two-firm duopoly. Thus, the Big-5 generators behaved as collusively as a two-firm duopoly during the intermediate levels of demand, while their behavior was consistent with Cournot pricing in very low and very high demand hours.

Our results are consistent with those of Joskow and Kahn (2002) and Puller (2007) in that we find significant levels of withholding by the Big-5 generators even under very conservative assumptions used to measure withholding. Our dataset allows us to measure withholding in each hour during the summer of 2000 with a fair degree of precision, as we are able to account for most if not all of the physical and economic constraints that each generating unit faced in each hour. Secondly, although it appears otherwise, our finding of tacit coordination among the Big-5 firms under certain demand conditions is not inconsistent with Puller's (2007) result that the generators' behavior reflected unilateral market power. Puller's results were based on analysis of hour 18 only during which demand is generally at or close to its peak in each day. Our results also indicate that market power was unilateral in nature during high-demand periods.

2. Institutional Background

Joskow (2001) provides a detailed description of California's restructuring program for its electricity industry. We discuss below some aspects of the deregulation process that are relevant to this paper. These aspects include the divestiture of fossil generating capacity, the creation of institutions to facilitate trading of electricity, and the events that led up to the meltdown of the deregulated market in 2000 and 2001.

Prior to the restructuring, California's electricity industry was organized around three vertically-integrated investor owned utilities which operated generation, transmission and distribution networks. The utilities, namely Pacific Gas & Electric Company (PG&E), Southern California Edison Company (SCE), and San Diego Gas & Electric Company (SDG&E), were regulated by the California Public Utilities Commission (CPUC). While the IOUs supplied a

large fraction of their retails customer's needs, they also depended on importing a significant amount of power for utilities in other Western states, Canada, and Mexico.

The electricity market in Californian was deregulated in April, 1998. As part of the deregulation process, PG&E and SCE were required to divest their fossil-fueled generation capacity, which was comprised of gas-fired and steam and combustion turbines, to mitigate horizontal market problems. This capacity of approximately 17,000 MW constituted close to half of the generation capacity within California. More importantly, this capacity was at the top end of the supply stack meaning that it was generally the marginal supply source that balanced supply and demand for most of the time. All of this capacity was divested in 1998 and 1999 to five independent generating companies, which did not have other generation capacity within California. These "Big-5" firms are Williams (in affiliation with AES), Duke, Dynegy, Reliant, and Mirant (also knows as Southern). Each firm ended up with roughly a fifth of the divested capacity.

Two institutions were created to facilitate the transmission and trading of energy. The first was the California Power Exchange (CALPX), which ran the "forward" markets, namely the day-ahead and hour-ahead wholesale market. It was intended that the bulk of the power needed for California would be traded through the CALPX. The second institution was the California Independent System Operator (CAISO) which was responsible for operating the transmission network, balancing last-minute ("real-time") supply and demand requirements, and maintaining the overall short-term reliability of the system by procuring operating reserve services ("ancillary services"). Both CALPX and CAISO were non-profit corporations, and ran electricity markets on a uniform single-price auction basis for each hour of the day. For example, the PX took the hourly day-ahead supply and demand bids and stacked them to form aggregate supply and demand curves for each hour. The hourly market clearing price was then determined by the intersection of these aggregate supply and demand curves. All buyers paid the uniform market clearing price and all sellers were paid this price.

The three utilities were net-purchasers of electricity in the wholesale PX and ISO markets. Furthermore, although the wholesale market prices were deregulated, retail prices were fixed for up to four years in that the utilities could not charge a price higher than $65/MWh to retail customers. It was assumed that the wholesale price could not be higher than the regulated retail price.

The competitive wholesale market, which began functioning in April 1998, faced a number of flaws in the market design in the first couple of years. There were also some episodes of market power during very high demand periods. However, by and large, the wholesale market prices for energy were reasonably close to pre-reform projections. It was not until mid-May 2000 that whole prices began to rise above historical levels. Price increased dramatically in June and stayed high for the rest of summer months. The average CAISO real-time price in the summer months was close $140/MWh which was almost five times the average price between April 1998 and May 2000.

The IOUs were not allowed to pass on the increase in price to the retail consumers. Consequently, the three utilities began to lose significant amounts of money. Wholesale prices fell modestly in October 2000 and then increased significantly in November and December 2000. By mid-December, the utilities were paying almost $400/MWh for power in the wholesale market and reselling it for $65/MWh to retail customers. The utilities' continued requests for permission to increase retail prices were either rejected or deferred for further consideration by the CPUC. By early January 2001, PG&E and SCE were effectively insolvent. Supply shortages and involuntary curtailments of supplies to individual consumers soon followed. The PX ceased operating its forward market on January 31, 2001 in response to utility credit problems. By the end of January, the state of California, through the California Department of Water Resources (CDWR), began to buy power that "kept the lights from going out" in California. CDWR spent about $8 billion through June 2001 doing so. Thus, in just six months, the electricity reform program had collapsed.

Joskow (2001) identified five interdependent factors that contributed to the elevation of wholesale prices in California. These were: (a) increased prices for natural gas, which was the primary input in the production of electricity from steam and combustion turbines; (b) increased costs for Nitrogen Monoxide (NOx) emissions (NOx is a by-product of the generation process); (c) large increases in electricity demand in California; (d) reduced imports from other States; and (e) market power problems. Joskow attributed a third of the wholesale price to market power in the summer month of 2000, after accounting for change in the fundamental demand and supply conditions.

3. Data

The data used in our analysis was furnished during the litigation that followed after the California electricity crisis. At the request of the IOUs and the State of California (collectively known as the "California Parties"), the Federal Energy Regulatory Commission (FERC) opened an investigation into the justness and reasonableness of wholesale prices in the CALPX and the CALISO. The FERC ultimately ordered the CAISO, in the "Refund Proceeding" (henceforth referred to as the FRP), to recalculate the price to reflect what prices would have prevailed in a competitive market, in order to aid FERC in calculating refunds owed to the California consumers. During the FRP, a significant amount of data was made available by the CAISO, the California generators, and other industry experts. We had access to that confidential data as we filed testimony on behalf of the California Parties showing evidence of withholding by the Big-5 generators during the crisis period (Reynolds, 2003a, 2003b). These documents contain details on the description of the data and our methodology to compute withholding.[3]

During the FRP, alternate sources of data were made available for several of the metrics. We present the results in this paper using the most conservative measur available. For example, whenever there is a choice of which input price series to use in the computation of marginal cost for the California generators, we rely on the series which gives a higher estimate of the marginal cost, and hence a lower value for withholding and the conduct parameter.

4. Estimation of Withholding

4.1. Definition

"Withholding" is defined as the failure to produce energy (or to provide ancillary services)[4] from generating capacity that is capable of providing such energy at a marginal cost that is below the prevailing price. When a generator withholds output, it sacrifices the profits that it would have earned on that foregone output. However, if the withholding leads to a

[3] The data furnished during the FRP was made public subsequent to the filing of the testimonies by various parties.

[4] Ancillary services are services other than the provision of energy that are required to maintain system reliability. These services generally require that the supplier set aside capacity that can respond within a specified time period to a directive to increase output.

sufficiently large increase in the prevailing price, that sacrifice in profits can be more than offset by the increase in profits from the price increase realized on the firm's remaining output (*i.e.*, the output that is not withheld). Thus, if a firm is seen to be withholding output, it can be concluded that the firm believed such withholding would lead to a sufficiently large price increase to make the withholding profitable. In other words, if a firm withheld output, it can be inferred that the firm had market power, exercised that power, and caused prices to be elevated.

Withholding can be divided into two categories: "physical withholding" and "economic withholding." Physical withholding refers to a situation in which capacity that was available and economic at the prevailing market price was not bid into the market. Economic withholding refers to a situation in which capacity that was available and economic at the prevailing market price was bid at a price that was higher than its marginal cost and the market price, so that the bid was not accepted. Thus, such capacity was not dispatched. Since each of the categories of withholding are likely to have similar effects on the market, we do not distinguish between the types of withholding.

Table 1 provides some simplified examples to help illustrate our definition of withholding. Basically, withholding occurred when a generating unit was not actually dispatched, but the unit would have been dispatched had it been bid at its marginal cost. In Case 1, while the unit had a marginal cost of $50 per MWh, it was bid at $100 per MWh and the market price was $150 per MWh. In this case, the unit would have been dispatched at the actual bid (as well as at a bid equal to its marginal cost), and there was no withholding. In Case 2, the market price was $75 per MWh and the unit was not dispatched at the actual bid of $100. In this case, there was withholding because the capacity would have been dispatched had it been bid at its marginal cost. In Case 3, the market price was below the marginal cost of the unit. Hence, there was no withholding because the unit would not have been dispatched even if it had been bid at marginal cost. Similarly, there was no withholding in Case 6 even though the unit was not bid. In Cases 4 and 5, there was withholding because the unit was not bid and not dispatched, but would have been dispatched had it been bid at its marginal cost.

It is important to note that bidding at prices above marginal costs, in and of itself, does not constitute withholding under our definition. The reason is that bidding above marginal costs may or may not have affected dispatch of the unit. Referring back to Figure 1, Cases 1 and 3 are examples in which bidding above marginal costs did not constitute withholding, whereas Case 2

is an example in which bidding above marginal costs did constitute withholding. Similarly, failing to bid available capacity, in and of itself, did not constitute withholding under our definition, because again that failure may or may not have affected dispatch. Referring to Figure 1, Cases 4 and 5 are examples in which failing to bid did constitute withholding whereas Case 6 is an example in which it did not.

Our calculation for withholding is for the CAISO real-time energy market, which is the "market of last resort" in that it represents the last opportunity for the generators to sell energy in an hour. As such, the "opportunity" cost of selling power at an alternate time does not exist for the real-time market.[5] Therefore, when measuring withholding from the real-time energy market, one does not need to consider returns from alternative uses of the capacity for that hour in computing the marginal costs used in the analysis.[6]

In what follows, we describe briefly the components of the withholding computation. The details for the calculation can be found in Reynolds (2003a, 2003b).

4.2. Computation

We compute withholding (WH) as the difference between producible economic capacity (PEC) and supplied output (SO). Producible economic capacity (PEC) is the amount of capacity that: (a) can be produced given all of the constraints on production and (b) is economic, meaning that the marginal cost of producing energy from that capacity was below the real-time market price; and supplied output (SO) is the amount of output actually supplied, including energy generated and capacity reserved for ancillary services. In other words, producible economic capacity is the amount of capacity that could have been economically supplied while supplied output is the amount that was actually supplied. Withholding is the gap between these two amounts.

---

[5] When considering options to sell power for a given hour before bids are due in the real-time energy market for that hour, generators face an "opportunity" cost. That is, if the generator decides to sell power "now" for delivery in a future hour, it loses the opportunity to sell that power for delivery in that hour at a later time (*i.e.*, closer to the time of delivery). If there is a possibility that the price for such delivery will be higher later, by selling "now" the generator gives up the opportunity to earn greater profits by selling "later."

[6] This is not to say that focusing on the CAISO real-time energy market eliminates all alternative use issues. In particular, we discuss below certain inter-temporal issues related to environmental regulations.

4.2.1 Producible Economic Capacity

Absent any constraints, producible economic capacity for each unit in each hour would simply equal the effective capacity of the unit. For this analysis, we have used the lowest effective capacity value for each unit reported by the California generators during the FRP. However, there are a number of factors that can limit producible capacity to a value below effective capacity in any given hour. These include outages, reserve shutdowns, ramping limits, environmental regulations, and marginal costs relative to market prices.

Outages are events when a generating unit has to be taken down for servicing or repairs. Outages can be either full (*i.e.*, the unit is not capable of producing any power) or partial (*i.e.*, the unit can produce some power, but not up to its full capability). Reserve shutdowns are events when a unit is taken offline for economic reasons. For example, if an operator expects that there will be a period of low prices such that the operator cannot earn a positive margin from operating the plant and selling the output, the operator may turn the plant off for that period. For purposes of this analysis, we have conservatively assumed that the outages and reserve shutdowns as reported in the data were entirely "legitimate" in the sense that they did not represent withholding.[7]

The available capacity is also constrained by ramping limits. Power plants cannot instantaneously change the level of their output. Rather, units must ramp up (or down) over time to reach the desired operating level. Our analysis of withholding attempts to account for this limitation.[8] Our analysis also considers the fact that certain generating units were subject to environmental regulation that could have affect the opportunity cost and/or dispatch of those units.[9]

---

[7] This is a conservative assumption because testimony filed during the FRP showed that outage rates during the summer of 2000 were significantly higher than their historical average, implying that the generators were physically withholding capacity by declaring "false" outages. See, for example, Hanser (2003).

[8] In particular, if a unit was not operating at its producible economic capacity in hour *t* because of a legitimate reason, we assumed that the maximum available capacity in the following hour *t+1* equaled the average amount that would be produced if the unit ramped up from its operating level in hour *t* until it reached its full producible economic capacity for hour t+1 or it reached the end of the hour t+1.

[9] We conservatively excluded from the withholding analysis all generating units that were subject to: (a) limits on daily NOx emissions; (b) regulation on cooling water discharge during certain periods of the year; (c) annual limit on the number of hours that could be operated. The last restriction was applicable on all the combustion turbine units in the CAISO area. In all, the excluded unit constituted about 10% of the Big-5 capacity.

The final constraint is marginal cost of operating the unit relative to the CAISO real-time energy market price. For a given generating unit in a given hour, the amount of capacity that is economic in the CAISO real-time energy market is that portion of the unit's capacity (above the capacity that is committed outside the CAISO market) that has a marginal cost below the CAISO real-time energy market price.[10] We compute hourly marginal cost curves for each unit using data on fuel costs (natural gas prices in this case), heat rate curves which specify the efficiency of each unit at various production levels, variable operation and maintenance costs (O&M), cost of NOx emissions, and emission curves which specify the quantity of emission at various production levels.[11, 12]

4.2.2 Supplied Output

Supplied output measures the extent to which a unit was actually used in a given hour. Supplied output is the sum metered generation, undispatched ancillary services, and decremental instructions.

Metered generation is the actual amount of energy provided by a unit in a given hour. We do not consider capacity that was awarded in the form of ancillary services to have been withheld. However, awarded ancillary service capacity could have been dispatched in the CAISO real-time energy market. Because dispatch of such ancillary services capacity would register as metered generation, to avoid double-counting such generation we added only *undispatched* ancillary services to metered generation. Finally, units were sometimes asked to decrease (decrement) their output by CAISO in the event that CAISO over-procured energy. Because we do not consider power that was not produced as a result of decremental instructions

---

[10] The capacity committed outside the CAISO was generally through sales to the PX day-ahead and hour-ahead markets.

[11] For the computation of economic capacity for each generating unit and hour, we assume that the market price in the hour is exogenous. In the calculation of the marginal cost, the fuel cost is the product of the incremental heat rate of the units, which is a measure of the efficiency with which each unit turns fuel into electricity, and the price of the natural gas. For the O&M cost, we used the value of $6/MWh, which is the rate adopted in the FRP. The NOx emissions cost is equal to the product of the NOx emission rate for the unit (lbs/MWh) and the relevant NOx emissions cost ($/lb). For each generating unit, the heat rates as well as the emission rates varied with the level of production.

[12] To be conservative in our marginal cost calculations, we use the gas price series based on certain published indices of California gas price, even though this series was higher on average than the series used by FERC it its Refund Proceeding. Similarly, we use the NOx emission costs that were furnished by the experts sponsored by the California generators. These NOx costs were higher on average than those provided by another expert. See details in Reynolds (2003a, 2003b).

by the CAISO to be withholding, we added back the amount of the decrements in the calculation of supplied output.

Our computation of withholding is not only more precise but also more conservative than that of Joskow and Kahn (2002). Joskow and Kahn relied on publicly available data for estimating withholding. In particular, they did not have access to unit level data on outages, reserve shutdowns, ramping constraints, undispatched ancillary capacity, and decremental instructions. Therefore, their analysis did not fully account for peculiarities in unit level behavior in each hour. Moreover, not being able to compute hourly unit-level marginal costs, they focused their analysis only on a set of high-price hours when it would have been economical for all units to supply.

4.3. Withholding Results

We compute hourly withholding during peak hours from 10 AM to 10 PM between June and October of 2000.[13] We ignore hours when there is transmission congestion across zones, since a generating unit may not have been able to supply into a congested zone even if it was willing to do so. Congestion took place in approximately 27% of the peak hours in the relevant period.[14]

Table 2 shows average hourly withholding numbers by month for each of the Big-5 generators. We find that there is significant heterogeneity among the Big-5 firms in the level of withholding. Duke shows the lowest levels of withholding with an average of 25 MW across the five months.[15] Williams shows the second lowest level of withholding, with an average of 56 MW. Dynegy, Mirant, and Reliant show an average withholding level of 132 MW, 188 MW, and 235 MW, respectively. The average aggregate withholding across the Big-5 generators was 636 MW, which was approximately 6% of the aggregate energy generated by the Big-5 during this

---

[13] The CAISO defines "peak" as hours between 7AM and 10PM everyday of the week except Sunday. We excluded the hours from 7AM to 9AM because a number of generating units were found to be ramping up during those early morning hours.

[14] During congestion hours, only suppliers with units located in the congestion zone can typically meet an incremental increase in the demand for energy in that zone. This means that there are fewer suppliers in the zone and hence the potential for unilateral maker power is higher during such hours.

[15] These low values could be attributed to the fact that during this period, Duke had a significant portion of its capacity under long-term contracts. Therefore, it did not have a strong incentive to withhold supply in the real-time market.

period. Furthermore, the combined withholding was at least 1000 MW in 19.2% of the hours in the sample. Although we have not undertaken an analysis of the direct effect of withholding on market prices, it is worthwhile to mention that FERC made Reliant pay $13.8 million for withholding close to 1000 MW from the PX market over a two-day period in June 2000.[16]

We would like to emphasize that our computation of withholding is quite conservative as we consider all outages and reserve-shutdown events as legitimate. To demonstrate the conservativeness of our calculation, we compute withholding without giving credit to the generators for the reserve shutdown hours. In particular, since reserve shutdowns could be a form of physical withholding, we re-estimated withholding under the assumption that the capacity on reserve shutdown was available if it was economic to supply, *i.e.* its marginal cost was below the prevailing market clearing price. The results, shown in last column of Table1, indicate an average withholding of 1,277 MW which twice as large as the base case.

Next we present withholding figures by deciles of real-time CAISO demand in Figure 1. Given that there are approximately 6,000 hours in the relevant period, each decile contains observations on 600 hours. Each dot on the chart represents combined withholding by the five generators in an hour. We find that there is significant variation in average hourly withholding in each decile. For example, for the 1st decile, withholding ranges from -250MW to $1750 MW. The average withholding for this decile is 270 MW which is close to 1% of the average ISO load for that decile. Furthermore, average withholding increases initially with deciles reaching a maximum of 1,010 MW at the 6th decile before declining for higher decile levels. In the middle deciles the average withholding is between 2.5% to 3% of the average ISO load.

The inverted-U or hump-shaped pattern of withholding is consistent with the notion that coordination among the generators was largest when peak demand is neither too low nor too high. At low peak demand levels (1st and 2nd deciles) coordination is presumably not profitable because there is excess fringe supply in the market that can negate any restriction in the output by the Big-5 generators. At high demand levels (8th, 9th, and 10th deciles) coordination is

[16] See "Order Approving Stipulation and Consent Agreement," issued January 31, 2003 in FERC Docket No. PA02-2-001. The episode involved bidding in the CalPX day-ahead market. At issue were the bids submitted on June 20-21 for energy to be delivered on June 21-22. The FERC's investigation of this episode eventually led to a settlement in which Reliant agreed to pay $13.8 million and abide by other conditions related to its conduct. The $13.8 million settlement was based on the FERC staff's assessment of the maximum effect of Reliant's withholding on prices in the CalPX day-ahead market on the two days

presumably not necessary as firms may be able to elevate prices unilaterally under tight demand conditions when the fringe is capacity constrained. The incentives for coordination are plausibly highest in the middle deciles when, even though the Big-5 generator may not be able to unilaterally affect market prices, working together they are able to substantially restrict output to profitably elevate prices.

5. Prices and Margins

Figure 2 presents the range and average real-time ISO prices by deciles. Each dot on the chart represents an hourly market clearing price. The average price increases monotonically in deciles, going up from $42/MWH in the 1st decile to $365/MWH in the 10th decile. The ISO had set price caps during this period to limit the exercise of market power. The cap, which was set at $750/MWh in June, was lowered to $500/MWh in July and lowered further to $250/MWh in August. Therefore, a concentration of dots are seen at $250, $500, and $750 for higher decile levels.

To see if high prices implied high margins, we estimated the marginal cost for each firm in each hour. The marginal cost for a firm in an hour is assumed to be the cost of its most expensive unit not running at full capacity in that hour. A unit was assumed to be running at full capacity if it was operating at 90% or more of the capacity that was available during the hour, *i.e.* capacity that was not subject to forced outages, scheduled outages, or reserve shutdowns. We found that in many instances, a firm had multiple units with different costs that were not running at capacity. This could happen because a high cost unit might have had to run at a low load level to be able to ramp up to full load in subsequent hours. Therefore, we computed the marginal costs in two ways: by using the marginal cost of the most expensive unit (method 1) and by taking the average of the marginal cost for units with spare capacity, weighting each unit by its spare capacity (method 2). The first approach is more conservative as it yields high values of costs.

Figure 3 shows the range and average marginal cost by deciles for the two methods. The marginal costs increase in deciles. This is to be expected, since the firms employ increasingly expensive units to meet increasing demand. However, the increase in average marginal cost in

---

in question. Total sales transacted in the CalPX day-ahead market on those two days were about $103 million. Thus, the estimated impact of Reliant's withholding amounted to about 13% of the sales dollars.

going from the 1st decile to the 10th decile is not as dramatic as the increase in market clearing prices observed in Figure 2. Thus, the marginal cost increases from $67/MWh for the 1st decile to $95/MWh for the 10th decile for the more conservative method. Furthermore, marginal cost estimates are on average higher by 10% for the more conservative method.

Given the market clearing price and firm-level marginal costs, we can compute firm-level margins. Figure 4 shows the range of and average margins by deciles for both measures of marginal costs. The maximum margin is computed using the most expensive unit not operating at capacity, while the mean margin is computed using the average of the marginal cost for units with spare capacity. We find that the margins increase monotonically in deciles. The mean margins are actually negative for the first two deciles before turning positive in 3rd decile. For the more conservative method 2, the average margins increase sharply from 7% in the 3rd decile to 48% in the 7th decile and to 70% in the 10th decile.

High margins are indicative of market power. However, they can be caused by different types of market power. In fact, there are at least three alternate explanations for the high margins observed for the middle and high deciles. Firstly, high margins could reflect scarcity rents. The aggregate supply curve is close to being vertical at high demand levels. We observe that price hit the cap in more than 75% of the hours when the total ISO load was more than 38,000 MW. The high margins during these hours, therefore, may be reflective of scarcity rents.

Secondly, during high demand hours, the generators could be in a position to exercise unilateral market power. Since total demand was nearly inelastic during the period of study, individual firms faced residual demand functions (rival supply subtracted from total demand) that were inelastic. Hence, the firms would have an incentive to unilaterally increase the market price. Such a strategy would be profitable as the demand nears the industry's capacity; if one firm were to withhold capacity in order to drive up the price, other firms would have limited ability to increase output.

Thirdly, high margins may be a result of coordination among the suppliers. Market characteristics that are generally required to facilitate collusion among the generators were present in the California market. For example, the generators could obtain publicly available information about rival's marginal cost as well as output. Similarly, demand side information was also common knowledge as the ISO published its demand forecast. Moreover, repeated

interactions among the Big-5 suppliers were conducive to monitor and enforce collusive arrangements.

6. Empirical Framework to Distinguish Between Unilateral and Collusive Behavior

We examined the behavior of the Big-5 firms using a Cournot model with conjectures as in Puller (2007). As Puller points out, the quantity-setting Cournot model is more appropriate than a price-setting model because capacity constraints prevent any single firm from undercutting and supplying the entire market. However, the Cournot assumption is a major simplification, since the firms were in fact bidding supply function into the PX and ISO markets. The advantage of the Cournot with conjectures is that this specification is fairly tractable and can be used to distinguish between the three sources of market power. We closely follow Puller's exposition in specifying the model and deriving the first order conditions.

Firm *i* chooses the quantity of output in period *t* to maximize profit subject to capacity constraint:

$$\max_{q_{it}} \; P(q_{it} + q_{-it}).q_{it} - C_{it}(q_{it}) \quad s.t.\; q_{it} \le k_{it}$$

The first-order condition for an interior solution at the optimal quantity q_{it}^* is:

$$P(q_{it}^* + q_{-it}) - c_{it}(q_{it}^*) + \theta_{it}.P_t^{'}.q_{it}^* - \lambda_{it}^* = 0 \qquad (1)$$

where $c_{it}(q_{it}^*)$ is marginal cost, λ_{it}^* captures scarcity rent, and $\theta_{it} \equiv \frac{dQ_t^*}{dq_{it}} = 1 + \sum_{j \neq i} \frac{\partial q_{jt}}{\partial q_{it}}$ is the firms' conjecture about the effect of increasing its output on total industry output. The parameter $\theta_{it} = \{0,1,N\}$ corresponds to perfect competition, Cournot, and joint monopoly pricing (under symmetry), respectively. The first-order condition can be transformed into the Lerner index:

$$\frac{P(q_{it}^* + q_{-it}) - c_{it}(q_{it}^*) - \lambda_{it}^*}{P(q_{it}^* + q_{-it})} = \theta_{it} * \frac{q_{it}^*}{Q_t} * \frac{1}{\varepsilon_{DRt}} \qquad (2)$$

The three scenarios of market power discussed above are captured by equation 2. First, market power originating from scarcity is reflected by the condition ($\theta_{it} = 0$ and $\lambda_{it}^* > 0$). Under this condition, firms utilize all capacity that is economic (*i.e.* market cost less than market clearing price), and margins signal the value of additional capacity. Second, unilateral market power is represented by the condition ($\theta_{it} = 1$ and $\lambda_{it}^* \ge 0$). In this case, firms unilaterally withhold a

portion of their capacity to raise price on the infra-marginal units that are not withheld. Finally, joint market power is captured by the condition $(\theta_{it} = N \text{ and } \lambda_{it}^{*} \geq 0)$, whereby firms engage in jointly withholding capacity to raise price, leading to an outcome that would be observed if the industry were a monopoly. Note, the condition $(1 \leq \theta_{it} \leq N \text{ and } \lambda_{it}^{*} \geq 0)$ represents outcomes that are between Cournot and monopoly. In this case, the inferred conduct is imperfect collusions, as though the firms are engaged in joint withholding capacity, the resulting price is below the joint monopoly level.

The conjectural variations model as laid out above has been subject to criticism on theoretical grounds. In particular, the theoretical literature has shown that the behavioral parameter represents a consistent equilibrium only under specific information assumptions and need not represent a Nash equilibrium generally (Lindh, 1992). Despite these criticisms, the empirical literature in industrial organization has continued to estimate and use the conduct parameter on the grounds that it is a static-equivalent of a dynamic model. We take a similar stand in our use of the conjectural variations model. Since the Folk theorem tells us that a range of conduct are Nash equilibria in a dynamic game, the conduct parameter should be viewed as only a measure of the elasticity adjusted Lerner index that summarizes the industry's competitiveness.

The econometric estimation of the conduct parameter is also problematic. In particular, although the conduct parameter is supposedly the "average" margin over varying demand conditions, in practice the parameter is estimated as "marginal" response of markup to demand shocks. This is because the parameter is estimated as the slope of a regression of price or margin on demand conditions. Corts (1999) has shown analytically and through simulation that the average markup may not generally equal the marginal markup, which implies that the econometric estimates of the parameter are biased.[17] This study avoids the problem because we have very good data which enables us to measure the conduct parameter directly. Looking at equation 2, we have hourly data on market prices (P), firm-level marginal costs (c_{it}), and quantities (q_{it}, Q_t). Furthermore, as discussed in the next section, we can also estimate on an hourly basis the elasticity of residual demand (ε_{DRt}) facing the big-5 generators directly using actual bid data in the real-time market.

7. Estimation of Elasticity of Residual Demand

We compute elasticity of residual demand on an hourly basis using the actual bids submitted in the CAISO real-time energy markets. The method was developed by Wolak (2003) and involves first computing the aggregate demand for electricity in the CAISO's real-time energy market and then subtracting from that the total amount supplied at the market clearing price by all fringe participants. To compute the slope at the market-clearing price, we find the closest price above the market-clearing price such that the residual demand is less than the value at the market-clearing price. Denote the market-clearing price as *P*, closest price above as *P*(low), and the associated value of residual demand as *DR*(*P*(low)). Next we find the closest price below the market-clearing price such that the residual demand is greater that the value at the market clearing price. Call this price *P*(high) and the associated demand as *DR*(*P*(high)). The elasticity of residual demand is then equal to the arc elasticity between the two points and is given by:

$$\varepsilon_{DR} = \frac{DR(P(high)) - DR(P(low))}{P(high) - P(low)} x \frac{P(high) + P(low)}{DR(P(high)) + DR(P(low))} \quad (3)$$

Figure 5 shows the range of and average elasticity estimates by demand deciles. The elasticity decreases (in absolute value) with deciles, going from -4.6 in the 1st decile to -3.2 in the 5th decile and to -0.2 in the 10th decile. Note that for a number of hours the computed elasticity was larger than 10 in absolute value. Since higher elasticity implies a larger value for the conduct parameter all else equal, we conservatively assign an elasticity of 10 to all high elasticity hours.

8. Estimation of Conjectural Variation Parameter

With the computation of an hourly elasticity measure, we have all the necessary data to compute the conduct parameter. Assuming that scarcity rent parameter (λ_{it}^*) is zero when the capacity is not binding, we can compute the conduct parameter from equation 2 as follows:

$$\theta_{it} = \frac{P(q_{it}^* + q_{-it}) - c_{it}(q_{it}^*)}{P(q_{it}^* + q_{-it})} * \varepsilon_{DRt} * \frac{1}{s_{it}} \quad (4)$$

---

[17] Puller (2007) makes an attempt to control for this bias in his econometric estimation.

The conduct parameter for a Big-5 firm *i* in any hour is the product of firm *i*'s price-cost margin, elasticity of demand facing the Big-5 firms, and the inverse of firm i's share of output in that hour. As stated earlier, the parameter $\theta_{it} = \{0,1,N\}$ corresponds to perfect competition, Cournot, and joint monopoly pricing (under symmetry), respectively.

We set the conduct parameter value equal to zero for the hours in which a firm faced non-positive margins or was operating at full capacity.[18] There were some hours in which the residual demand curve facing the Big-5 firms was essentially vertical, *i.e.* the elasticity of residual demand was zero. Since each firm is able to unilaterally affect price under such tight demand conditions, we assigned a value of one to the conduct parameter for these hours. Figure 6 shows the range of and average conduct parameter by decile. We observe an inverted-U shaped pattern for the average value of the conduct pattern, similar to the pattern observed for withholding. For the more conservative measure of marginal cost, the parameter has a mean of 0.8 for the 1st decile and 1.4 for the 2nd decile. These figures imply that on average the Big-5 was behaving as Cournot competitors at low demand levels. Similarly, the conduct parameter shows relatively low average values for deciles 9 and 10, which is consistent with the notion that the generators were exercising unilateral market power during these hours when the residual demand was essentially inelastic. For the middle deciles, the conduct parameters average between 2 and 3 for the conservative measure of marginal cost. This implies that the Big-5 firms were behaving more collusively during this period. Since the conduct parameter is 2.5 for a duopoly, the observed values suggest that the 5 firms were acting at least as collusively as a duopoly during the period when demand was neither too low nor too high.

Table 3 displays the mean, standard error, and standard deviation of the values of the conduct parameter for the two measures of marginal costs by decile and by firm. As was the case with the withholding figures, we find significant heterogeneity among the firms in the conduct parameter, with Duke and Williams showing lower values and Reliant and Mirant showing higher values.

9. Conclusion

In accordance with previous studies, this paper has shown that the generators in California were exercising market power in the summer of 2000 by withholding significant

amounts of capacity even though that capacity was capable of providing energy at a cost below the prevailing price. However, unlike other studies that have argued that the exercised market power was unilateral in nature, we have shown that the generators' behavior varied with demand conditions. In particular, the behavior was more collusive during periods with intermediate levels of demand (*i.e.* when demand was neither too low nor too high) while it was consistent with unilateral pricing in very low and very high demand hours.

An understanding of the underling pricing regime is important for the design of an electricity market. For example, if market power arising from collusive pricing is a concern, market designers can alter the structure of ownership by requiring additional divestitures. They can also change the information available to market participants to make it difficult to reach and sustain collusive agreements.

[18] We assume that a firm is running at full capacity when a firm has all of its plants operating at 90% or more of the available capacity. Since we cannot distinguish whether such a firm is running at full capacity or withholding a small proportion of its capacity to elevate prices, we conservatively assumer the former.

References

Borenstein, Severin, Bushell, James, Wolak, Frank, 2002. Measuring market inefficiencies in California's restructured wholesale electricity market. American Economic Review 92, 1376-1405.

Corts, Kenneth S., 1999. Conduct parameters and the measurement of market power. Journal of Economometrics 88, 225-250.

Hanser, Philip, 2003. Prepared Direct Testimony of Philip Hanseron Behalf of the California Parties. FERC Dockets EL00-95-000 *et al.*

Hildebrandt, Eric, 2000. Declaration of Eric Hildebrandt. FERC Dockets EL00-95-000 *et al.*

Joskow, Paul L., 2001. California's Electricity Crisis. Oxford Review of Economic Policy 17, 365-388.

Joskow, Paul, Kahn, Edward, 2000. A quantitative analysis of pricing behavior in California's wholesale electricity market during summer 2000. Energy Journal 23, 1-35.

Lindh, Thomas, 1992. The inconsistency of consistent conjectures: Coming back to Cournot. Journal of Economic Behavior and Organization, 69-90.

Puller, Steve L., 2007. Pricing and firm conduct in California's deregulated electricity market. Review of Economics and Statistics 89, 75-87.

Reynolds, Robert J., 2003 a. Prepared Direct Testimony of Robert J. Reynolds, Ph.D. on Behalf of the California Parties. FERC Dockets EL00-95-000 *et al.*

Reynolds, Robert J., 2003 b. Prepared Rebuttal Testimony of Robert J. Reynolds, Ph.D. on Behalf of the California Parties. FERC Dockets EL00-95-000 *et al.*

Wolak, Frank A., 2003. Measuring unilateral market power in wholesale electricity markets: the California market, 1998-2000. American Economic Review, 425-430

Wolfram, Catherine D., 1999. Measuring duopoly in British electricity spot market. American Economic Review 89, 105-826.

Table 1: Simplied Exampled to Illustrate Definition of Withholding

Case	Marginal Cost	Bid Price	Market Price	Withholding?
1	50	100	150	No
2	50	100	75	Yes - Economic
3	50	100	30	No
4	50	None	150	Yes - Physical
5	50	None	75	Yes - Physical
6	50	None	30	No

Table 2: Summary of Withholding Results for On-Peak Hours

	Average Hourly Withholding (MW)						Withholding	% of Hours Withholding	Total Withholding including
	Duke	Dynegy	Mirant	Reliant	Williams	Total	% of Generation	>1000 MW	Reserve Shutdown
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Jun-00	17	137	10	251	68	483	4.7	13.3	1,431
Jul-00	43	134	234	286	98	795	7.7	29.9	1,740
Aug-00	13	132	215	203	34	597	4.7	20.9	725
Sep-00	30	156	227	231	39	682	6.1	17.8	1,273
Oct-00	20	104	256	202	43	624	6.8	14.5	1,214
Average	25	132	188	235	56	636	6.0	19.2	1,277

Notes:
On-Peak Hours are defined as hours between 10 AM and 10 PM in all days of the week, except Sunday. Congestion hours are excluded.

Table 3: Theta Mean and Theta Max (By Decile and by Company)

	Decile (obs=620)										Company				
Theta Mean	1	2	3	4	5	6	7	8	9	10	Duke	Dynegy	Mirant	Reliant	Williams
Mean	**1.13**	**1.96**	**2.79**	**3.64**	**3.19**	**3.46**	**2.59**	**2.01**	**1.41**	**0.73**	**1.72**	**2.57**	**3.00**	**2.66**	**1.49**
Standard Error	0.103	0.130	0.143	0.159	0.148	0.146	0.125	0.110	0.085	0.028	0.084	0.093	0.105	0.094	0.069
Standard Deviation	2.55	3.22	3.55	3.95	3.67	3.62	3.10	2.73	2.10	0.69	2.95	3.28	3.68	3.29	2.41

	Decile										Company				
Theta Max	1	2	3	4	5	6	7	8	9	10	Duke	Dynegy	Mirant	Reliant	Williams
Mean	**0.83**	**1.38**	**2.05**	**2.90**	**2.63**	**2.96**	**2.30**	**1.89**	**1.35**	**0.73**	**1.65**	**1.58**	**2.93**	**2.18**	**1.17**
Standard Error	0.086	0.112	0.128	0.149	0.138	0.139	0.118	0.107	0.081	0.027	0.082	0.071	0.104	0.086	0.057
Standard Deviation	2.14	2.77	3.18	3.71	3.42	3.44	2.94	2.66	2.01	0.68	2.89	2.50	3.65	3.02	2.01

Figure 1: Withholding by Decile

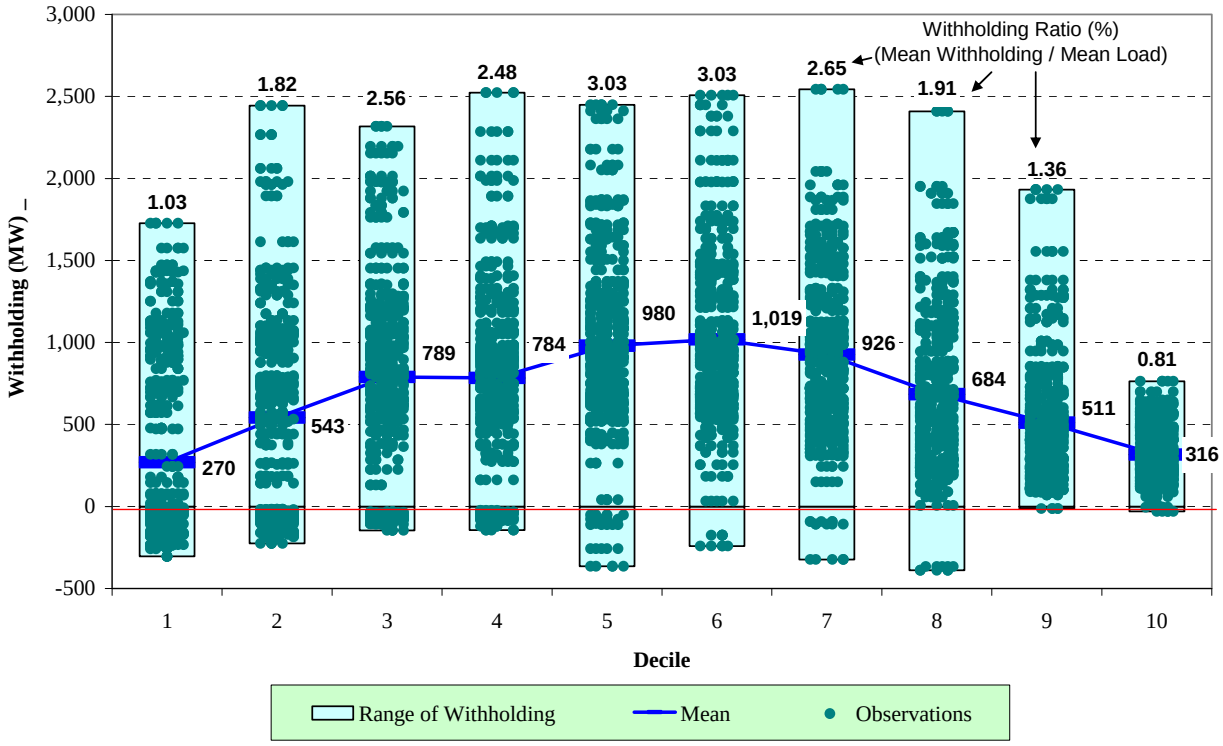

Figure 2: Real Time Market Clearing Price by Decile

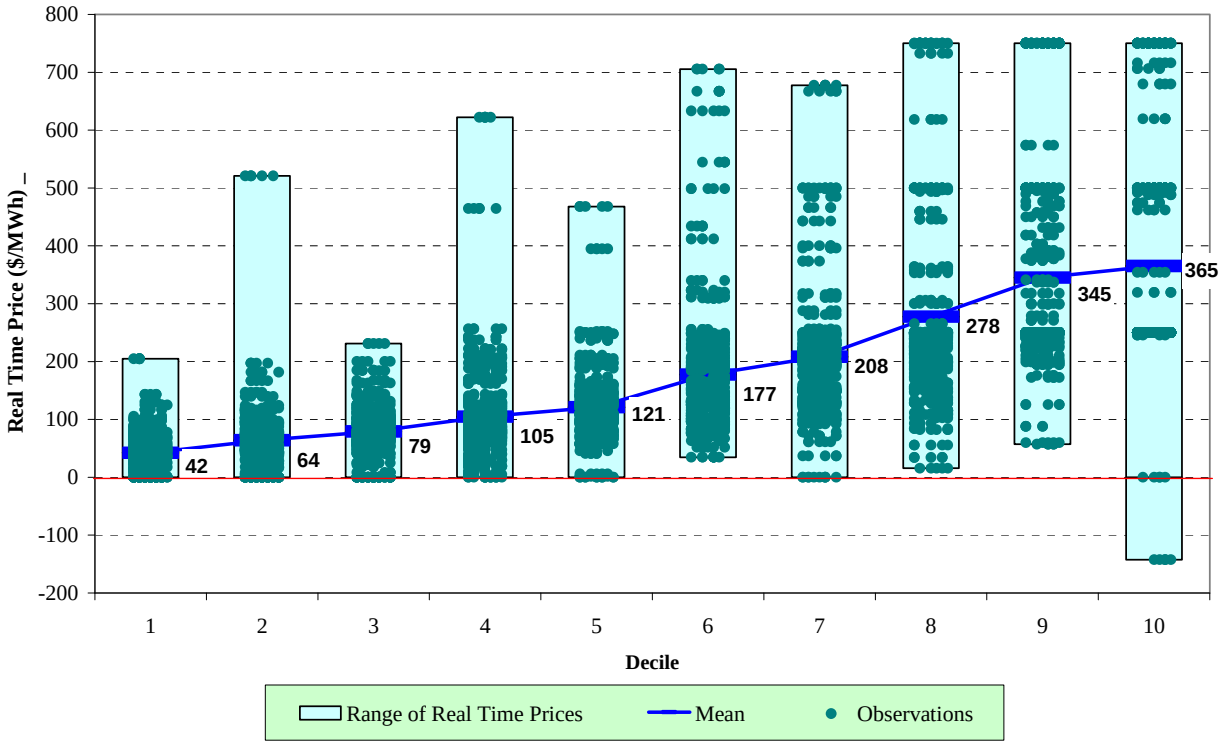

Figure 3: Mean and Maximum Marginal Cost by Decile

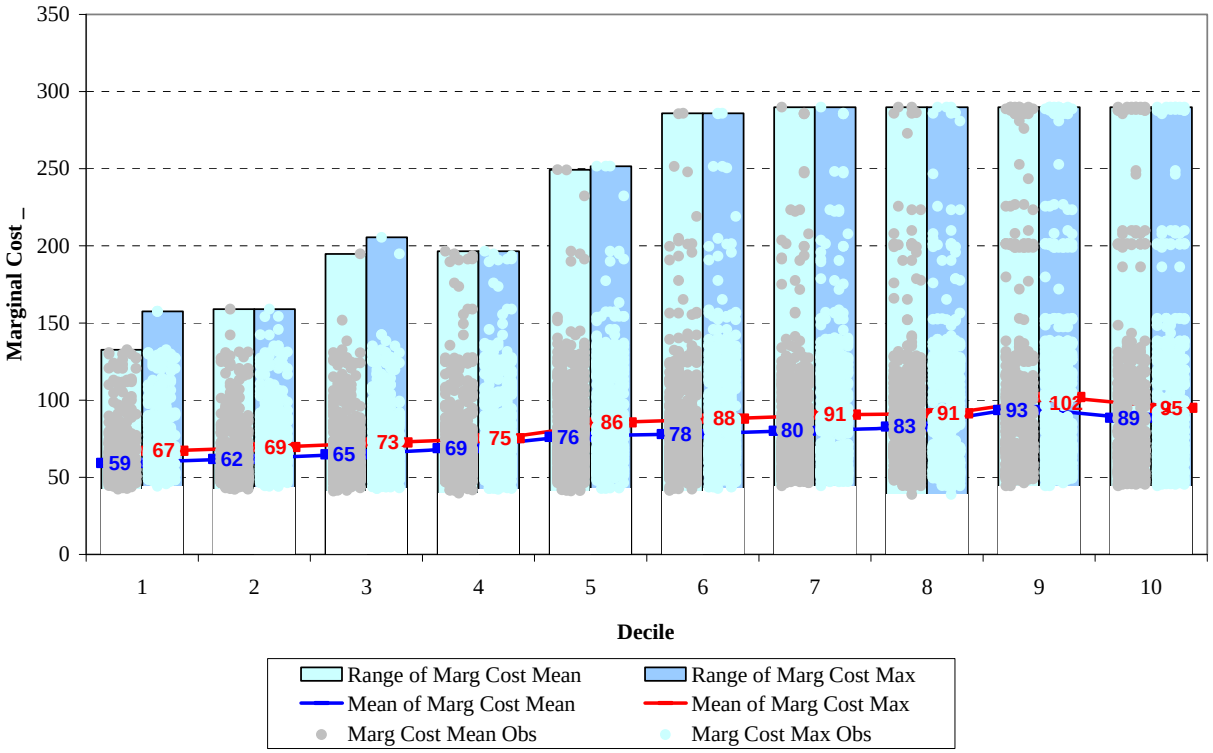

Figure 4: Mean and Maximum Margins by Decile

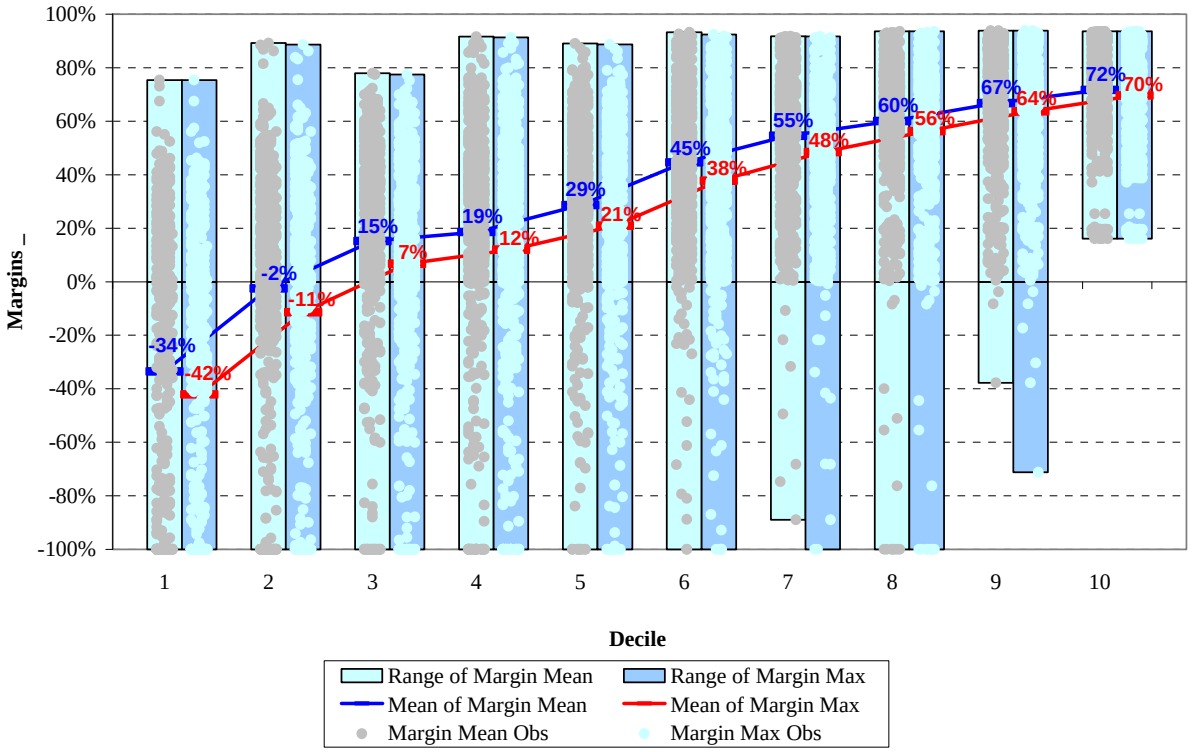

Figure 5: Elasticity of Residual Demand by Decile

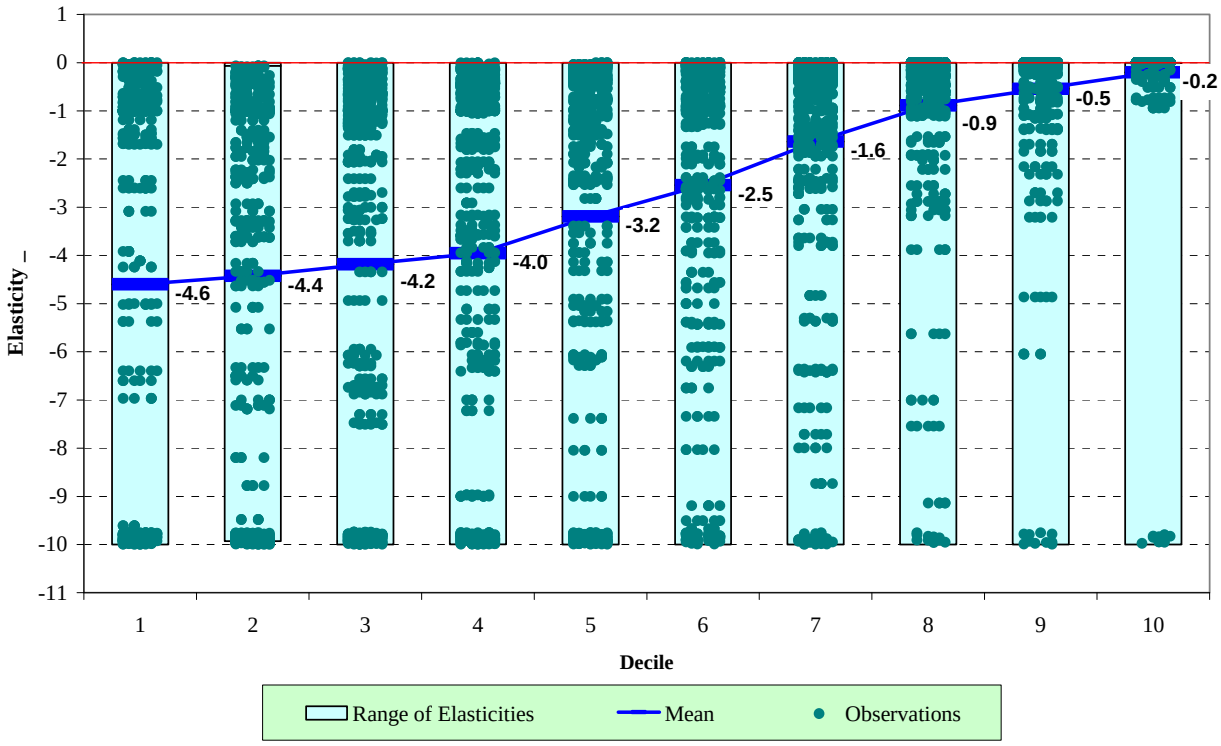

Figure 6: Theta Mean and Theta Max by Decile