5-2022

Essays on Nonparametric Benchmarking of Energy Firms and Natural Gas Market Integration

Matthew Cronin
mlcroni@g.clemson.edu

Essays on Nonparametric Benchmarking of Energy Firms and Natural Gas Market Integration

---

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Economics

---

by
Matthew Lee Cronin
May 2022

Accepted by:
Professor Paul W. Wilson, Committee Chair
Professor Matthew S. Lewis
Professor Christy Zhou
Professor F. Andrew Hanssen

Abstract

My dissertation focuses on issues related to the technical efficiency of energy firms as well as the integration of U.S. natural gas markets during the era of shale gas expansion of the early 2000s. In the first chapter, I use nonparametric methods to estimate changes in efficiency and productivity of natural gas pipelines in the U.S. during the period of shale gas expansion during 2007–2018. This period, known as the Shale Revolution, saw an increase in shale gas production from less that 5 billion cubic feet per day (BCF/d) in 2007 to over 60 BCF/d in 2018. This increase in production coincides with increases in capacity and infrastructure improvements to the transportation network of natural gas pipelines. Previous research in energy economics has not examined changes in efficiency and productivity of natural gas pipelines during the Shale Revolution using modern nonparametric techniques. To guide my estimation procedure I use new nonparametric tests to make inference on the properties of the production set of pipelines. These nonparametric tests indicate that the production set for pipelines changes with production year, is convex, and does not exhibit constant returns to scale. Moreover, I find strong evidence of increases in the technical efficiency and productivity of pipelines during the period of shale gas expansion from 2007 through 2018.

In the second chapter, I use nonparametric methods to estimate changes in productivity and efficiency of major electric utilities in the U.S. during 2001–2019. Starting in 2007, an increased supply of natural gas resulted in large-scale adoption of natural gas fueled energy production as well as an overall increase in installed generating capacity. Data on major electric utilities, collected by the Federal Energy Regulatory Commission (FERC), indicate that the overall installed generating capacity of utilities increased by 13 percent during 2001–2013, with an increasing share of capacity coming from plants fueled by natural gas. However, this increase in capacity coincides with a decrease in the growth of utility-scale electricity consumption and a 12 percent decrease in utility

energy generation. The evidence suggests a decline in the productivity of energy generation from electric utilities during the sample period. In addition, using modern nonparametric techniques reveals a decrease in mean technical efficiency of utilities during the period of decline in energy generation. Finally, other modern nonparametric tests show that the production frontier of electric utilities changes depending on the production year and share of natural gas fueled generating capacity a utility has out of their total capacity. This indicates that nonparametric efficiency estimates should be conditional on time and the share of natural gas fueled generating capacity.

In the third chapter, I use unit root tests to measure the degree to which geographically dispersed natural gas markets in the U.S. became integrated into the same market from 1996 through 2019. Previous papers examining natural gas market integration in earlier periods suggest that some regional natural gas markets in the U.S. became integrated into the same market during the early 1990s. This was likely the result of regulatory reform instituted by FERC in the late 1980s and early 1990s. Nonetheless, previous studies also indicate that some regions remained distinct markets, particularly the eastern and western U.S. Unit root tests of the price-gaps between geographically dispersed price hubs fail to reject that the eastern and western gas markets were distinct markets prior to the early 2000s. After 2001 a higher proportion of the east and west price-gaps became stationary, suggesting that more price hubs in each region responded to similar market shocks. This provides some evidence that the eastern and western U.S. natural gas markets may have been integrated into the same market starting in the early 2000s. Moreover, the qualitative results of this paper hold under four different unit root tests.

Acknowledgments

I first would like to thank Professor Paul W. Wilson, my advisor and chairman of my dissertation committee. He has been a great teacher and has always offered excellent feedback and advice. He has kept me focused and I credit his guidance for helping me become a better researcher, writer, economist, and person.

I am also grateful for the advice from my other committee members, Professors Matthew Lewis, Andrew Hanssen, and Christy Zhou. Their helpful feedback and comments on my papers have been indispensable in finalizing my dissertation.

Nora, my parents, and my brother, Eric, have also been a source of great help during the past few years. I am incredibly grateful for their love and encouragement.

Finally, to all my office mates and friends at Clemson (a.k.a. The Squad), thank you. I could not have done this without all your support and shared laughs.

Contents

List of Tables

List of Figures

Chapter 1

Nonparametric Benchmarking of Natural Gas Pipelines: Changes in Productivity and Efficiency of Natural Gas Pipelines During The Shale Revolution

1.1 Introduction: Natural Gas Transportation

Most natural gas production occurs far away from major population centers or market regions. Transporting natural gas from the wellhead to a market region requires a vast network of pipelines and processing facilities. Pipelines that cross state boundaries, or interstate pipelines, account for approximately 63 percent of natural gas pipelines in the U.S., while the remaining 37 percent are smaller intrastate pipelines or local delivery lines. The Federal Energy Regulatory Commission (FERC) oversees interstate pipelines' construction approvals, shipping rates, and contracts.

FERC considers two events in the past 30 years to be highly important to the natural gas pipeline industry. The first event was the unbundling of the sale (marketing) and transportation of wholesale natural gas. Prior to 1992, pipelines were vertically integrated firms that transported and marketed natural gas at a "bundled rate" to customers, such as local distribution companies. These vertically integrated pipeline firms could reduce competition in the sale of gas by preventing competing marketers from using their pipelines. In 1992, FERC instituted regulation under FERC Order 636 which required pipeline firms to unbundle the marketing and transportation of natural gas. FERC's goal was to allow marketers to compete in a deregulated competitive market. However, FERC Order 636 still regulates pipelines and prohibits them from marketing gas. Thus, pipelines can only operate as transporters of natural gas, and they cannot discriminate in the choice of marketers they transport gas for.

The second major event was the rapid increase of U.S. shale gas production, which started around 2007 and resulted from technological improvements in gas and oil production. This period, known as the Shale Revolution, started with developments in fracking and directional drilling which made shale gas economically feasible to extract. Figure 1.1 shows average daily production levels of shale gas by shale bed from January 2000 through December 2018. During this time, U.S. shale gas production increased from under 3.6 billion cubic feet of gas per day (BCF/d) to 65.7 BCF/d, an increase by a factor of 18.25 in daily production levels. Because of this, marketers needed an expanded pipeline network to transport larger volumes of produced gas away from wellheads to markets. FERC Order 636 prohibited marketers from developing their own pipelines to transport gas, thus pipeline companies responded to the increased demand for gas transportation and developed more marketer financed pipeline projects to connect shale beds to markets. Marketers that contract with pipelines to help finance the development of a new pipeline, in exchange for a share of capacity on the pipeline to transport gas, are known as anchor shippers. Figure 1.2 shows new pipeline capacity attributed to anchor shippers during 1999–2017, and depicts an upward trend in pipeline capacity attributed to anchor shippers since 2007. In addition, anchor shipper capacity attributed to natural gas producers increased in the years after the start of the Shale Revolution. This suggests there was an increased need for pipelines to transport gas away from production regions to markets.

To make inference on changes in the technical efficiency and productivity of U.S. natural

gas pipelines during the Shale Revolution, I use free-disposal hull (FDH) and data envelopment analysis (DEA) estimators. These estimators involve enveloping a sample of observed inputs and outputs of firms to estimate the production set and frontier (i.e. the set of efficient input and output combinations). My sample contains annual data on the inputs and outputs of U.S. natural gas pipelines during the years 1996–2018, a period that encompasses the Shale Revolution which started around 2007. While there are many papers that estimate the efficiency and productivity of natural gas pipelines in the U.S., to my knowledge, none examine efficiency or productivity trends of pipelines during the Shale Revolution while using modern nonparametric techniques.

Given the increase in natural gas production during the Shale Revolution, coupled with Order 636, regulators at FERC began to recognize a greater reliance on natural gas pipelines to serve electrical load in 2012, as they began to hold series of conferences on gas and electric system coordination (Annibali et al., 2012). The increase in natural gas supply resulted in the drop of the price of U.S. natural gas, and this changed the relative price of natural gas to other fuel sources, such as coal. This led to the adoption of more gas-fired energy generation and a substitution away from coal-burning power plants. Moreover, natural gas is also a fuel used by pipelines to power compressor stations. Thus, pipelines faced a changing cost set during the Shale Revolution. Furthermore, Kahn (1971) and O'Neill (2005) explain that natural gas pipelines are oligopolies where several pipelines may serve the same market. Increased pipeline build out during the Shale Revolution may have enhanced inter-pipeline competition for shipping contracts to transport shale gas to similar markets. This may have resulted in the adoption of cost saving technology by some pipelines. Dreskin and Boss (2010) comment on technological changes in pipelines during the early periods of the Shale Revolution. They argue that many pipelines began to serve similar markets and discuss how this created an incentive to reduce costs. Technological improvements such as more powerful compressor stations and increased pipeline operating pressures to transport gas increased the throughput volume of pipelines at a lower operating cost. Consequently, pipelines faced increased utilization and cost improvements during this period. This paper adds to the energy economics literature by shedding light on how the technical efficiency and productivity of pipelines changed during the Shale Revolution.

This paper also adds to the literature on pipeline efficiency by using recently developed nonparametric methods to make inference. Although studies use nonparametric methods to examine

pipeline technical efficiency (e.g. Jamasb et al. (2008) and Nieswand et al. (2010) both use DEA to estimate efficiency of pipelines just prior to the start of the Shale Revolution), to my knowledge previous studies of natural gas pipelines only report point estimates of pipeline efficiency and do not make inference on the properties of the production set, changes in efficiency over time, nor anything else. I use modern techniques to make inference on changes in technical efficiency over time and test for properties of the production set to guide selection of an appropriate estimator. These tests include tests of (i) changes in the production frontier due to environmental factors, (ii) convexity of the production set and (iii) returns to scale of the production set.

Prior research on interstate pipelines is also limited to small sample sizes. Nonparametric efficiency estimators are similar to other nonparametric estimators as they suffer from the curse of dimensionality. It is well-known that the convergence rates of nonparametric efficiency estimators decrease as the dimensionality of the problem (i.e. the specified number of inputs and outputs) increases, thereby increasing the order of estimation error. Jamasb et al. (2008) use the variable returns to scale (VRS) and constant returns to scale (CRS) DEA estimators with 39 observations and four dimensions. This may overstate the efficiency of pipelines as more pipelines will populate the frontier without dimension reduction. See Kneip et al. (2015) and Wilson (2018) for technical details. While Nieswand et al. (2010) use dimension reduction techniques on pipeline data used for DEA estimation, they apply eigensystem decomposition of the correlation matrix which may not always be reliable for dimension reduction. For this reason and others, Wilson (2018) proposes using eigensystem decomposition of the moment matrix. Given the small number of pipelines observed in U.S. pipeline data, this paper utilizes dimension reduction techniques so five dimensions can be specified with a nonparametric frontier estimator that achieves the parametric rate of convergence.

The remainder of this paper is organized as follows. Section 1.2 discusses the statistical model, estimators of efficiency, and tests for properties of the frontier. Section 1.3 discusses the sample used to estimate efficiency. Section 1.4 reports the results and findings of the estimates. Finally, a summary and overview of the conclusions are reserved for Section 1.5.

1.2 Statistical Model

1.2.1 Modeling Conditional Efficiency

To model pipeline efficiency, I use conditional efficiency measures which were first described by Cazals et al. (2002), and later discussed by Daraio and Simar (2005, 2007a,b) and Daraio et al. (2018). Consider a process that generates random vectors (X, Y, T). Inputs are denoted $X \in \mathbb{R}^p_+$, while $Y \in \mathbb{R}^q_+$ are outputs, and $T \in \mathbb{R}^1$ indicates the production year. To establish notation let the lower case letters (x, y, t) indicate particular realizations of the above random vectors. The production set, ignoring T,

$$\Psi = \{(x, y) \in \mathbb{R}^{p+q}_+ \mid x \text{ can produce } y\}, \tag{1.1}$$

is all the pairs of input and output quantities that are feasible given the production technology. However, as described in Section 1.2.3, the production year may affect the boundary of the production set. Conditioning on T leads to the conditional production set,

$$\Psi^t = \{(x, y) \in \mathbb{R}^{p+q}_+ \mid x \text{ can produce } y \text{ when } T = t\}, \tag{1.2}$$

which consists of all input and output quantities that are feasible given the production technology in year t. The efficient frontier of (1.2) is defined as the extreme points of Ψ^t, or

$$\Psi^{t\partial} = \{(x, y) \in \Psi^t \mid (\gamma^{-1}x, \gamma y) \notin \Psi^t \, \forall \, \gamma > 1, \, T = t\}. \tag{1.3}$$

Several assumptions about Ψ^t are made. These assumptions are similar to standard assumptions in production theory, and follow that of Shephard (1970). Moreover, these assumptions are required to use results established in Kneip et al. (2015), described in more detail below. I assume Ψ^t is closed, $(x, y) \notin \Psi^t$ if $x = 0, y \geq 0, y \neq 0$ (all production requires the use of some inputs), and $\forall (x, y) \in \Psi^t, (i)\, \tilde{x} \geq x \implies (\tilde{x}, y) \in \Psi^t$, and $(ii)\, \tilde{y} \leq y \implies (x, \tilde{y}) \in \Psi^t$ (the production set allows for strong disposability). Assuming that Ψ^t is closed implies that the set of efficient points, or the frontier, is contained in the production set (i.e. $\Psi^{t\partial} \in \Psi^t$). The second assumption ensures there is "no free lunch", or the production of a nonzero output vector requires the expenditure of a nonzero input vector. Finally, strong disposability imposes weak monotonicity of the frontier.

In a given year t, I observe data on inputs and outputs at the level of the pipeline. A pipeline's technical efficiency is measured by their distance from their observed input and output quantities to $\Psi^{t\partial}$. The Farrell (1957) input efficiency measure,

$$\theta(x, y \mid t) := \inf\{\theta \mid (\theta x, y) \in \Psi^t\}, \tag{1.4}$$

indicates by how much firms must proportionally scale back inputs while producing the same level of output to operate on $\Psi^{t\partial}$. Likewise, the Farrell (1957) output efficiency measure,

$$\lambda(x, y \mid t) := \sup\{\lambda \mid (x, \lambda y) \in \Psi^t\}, \tag{1.5}$$

indicates by how much firms must proportionally expand output given the same level of inputs to operate on $\Psi^{t\partial}$. As shown in Wilson (2011, Figure 6.1) a firm operating off of the efficient frontier can have differing input and output measures of efficiency. Thus, the hyperbolic graph measure of efficiency,

$$\gamma(x, y \mid t) := \inf\{\gamma > 0 \mid (\gamma x, \gamma^{-1} y) \in \Psi^t\}, \tag{1.6}$$

proposed by Färe et al. (1985), indicates the amount a firm must simultaneously scale down inputs and increase output by the same factor, γ, to operate on $\Psi^{t\partial}$. The benefit of (1.6) is that a firm's distance to the frontier is measured along a hyperbolic path and this avoids the issue of a firm having differing measures under (1.4) and (1.5).

1.2.2 Estimation Methods

Of course Ψ^t is not observed and must be estimated with a random sample of inputs and outputs by production year, $S_n = \{(X_i, Y_i, T_i)\}_{i=1}^n$, where i indexes the pipeline. Different nonparametric efficiency estimators require different assumptions about the properties of the production set. To begin, Deprins et al. (1984) propose estimating the FDH of the observed input and output vectors in S_n, or

$$\widehat{\Psi}_{FDH,n} := \bigcup_{(X_i,Y_i)\in S_n} \{(x, y) \in \mathbb{R}^{p+q} \mid y \leq Y_i, x \geq X_i\}, \tag{1.7}$$

to estimate the production set. Daraio and Simar (2005) discuss conditioning FDH estimates with respect to variables that may impact the boundary of the production set. In my case, if the produc-

tion year impacts the frontier of the production set, the FDH estimator conditional on T,

$$\widehat{\Psi}^t_{FDH,n} := \bigcup_{(X_i,Y_i)\in S_n} \{(x,y) \in \mathbb{R}^{p+q} \mid y \leq Y_i, x \geq X_i, T = t\}, \tag{1.8}$$

allows for the frontier to vary depending on T.

While the FDH estimators do not impose convexity on the production set, DEA estimators of the production set can be used if the productions set is convex. Farrell (1957) proposes DEA estimators of the production set in the unconditional case. The unconditional VRS-DEA estimator is given by the convex hull of (1.7), or

$$\widehat{\Psi}_{VRS,n} := \{(x,y) \in \mathbb{R}^{p+q} \mid y \leq \mathbf{Y}\boldsymbol{\omega}, x \leq \mathbf{X}\boldsymbol{\omega}, \mathbf{i}'_\mathbf{n}\boldsymbol{\omega} = 1, \boldsymbol{\omega} \in \mathbb{R}^n_+\}, \tag{1.9}$$

where $\mathbf{X} = (X_1, ..., X_n)$ and $\mathbf{Y} = (Y_1, ..., Y_n)$ are $(p \times n)$ and $(q \times n)$ input and output matrices, in year t, $\mathbf{i}'_\mathbf{n}$ is an $(n \times 1)$ vector of ones, and $\boldsymbol{\omega}$ is a $(n \times 1)$ vector of weights. Daraio and Simar (2007b) extend the VRS-DEA estimator to the conditional case, where in my application, conditioning (1.9) on T obtains the VRS-DEA estimator conditional on T, or

$$\widehat{\Psi}^t_{VRS,n} := \{(x,y) \in \mathbb{R}^{p+q} \mid y \leq \mathbf{Y}\boldsymbol{\omega}, x \leq \mathbf{X}\boldsymbol{\omega}, \mathbf{i}'_\mathbf{n}\boldsymbol{\omega} = 1, \boldsymbol{\omega} \in \mathbb{R}^n_+, T = t\}. \tag{1.10}$$

If CRS is assumed, the unconditional CRS-DEA estimator estimates the production set as the conical hull of (1.9), denoted $\widehat{\Psi}_{CRS,n}$, which can be obtained by removing the constraint, $\mathbf{i}'_\mathbf{n}\boldsymbol{\omega} = 1$, in (1.9). Similarly, the conditional CRS-DEA estimator, $\widehat{\Psi}^t_{CRS,n,h}$, is obtained by removing the same constraint from (1.10).

The conditional CRS-DEA, VRS-DEA and FDH estimators of efficiency are obtained by replacing Ψ^t with $\widehat{\Psi}^t_{CRS,n}$, $\widehat{\Psi}^t_{VRS,n}$ and $\widehat{\Psi}^t_{FDH,n}$, respectively, into equations (1.4)–(1.6). Input and output-oriented DEA estimates can be computed using linear programming methods, while hyperbolic-oriented estimates are non-linear programs that can be solved using numerical methods proposed by Wilson (2011). In addition, the FDH estimates can be solved using numerical methods as well. See Wilson (2011) for technical details. Furthermore, all estimates and tests described in this paper are done using the *FEAR* software developed by Wilson (2008).

1.2.3 Testing Hypothesis

To guide my estimation procedure I conduct several tests. First I test whether the production year affects the frontier. In this case, the production year can act as an environmental factor since it may influence the production process. The issues of environmental factors with nonparametric efficiency estimators have been discussed by Simar and Wilson (2011a,b) and Daraio et al. (2018). In my case, the production year, T, can impact the distribution of efficiency estimates, the efficient frontier, or both. If T only impacts the distribution of efficiency estimates, then all pipelines in all production years face the same attainable frontier and FDH and DEA efficiency estimates do not need to be conditioned on production year. In this case, the separability condition, as described in Simar and Wilson (2011a,b) and Daraio et al. (2018), is said to hold with respect to T. However, if T affects the frontier, unconditional FDH and DEA efficiency estimates are meaningless. In this case, unconditional efficiency estimates will benchmark some firms in some production year t relative to a frontier that is not attainable in t. After I test for separability, I then test for convexity versus non-convexity of the production set, then CRS versus VRS. These tests reduce the chance of selecting an estimator that relies on assumptions that are inconsistent with the true structure of the production set.

I test for separability along the lines of Daraio et al. (2018). In my application, the test involves testing the null hypothesis of

$$H_0 : \Psi^t = \Psi \, \forall t \in T \tag{1.11}$$

versus

$$H_1 : \Psi^t \neq \Psi \text{ for some } t \in T. \tag{1.12}$$

The test for separability requires randomly shuffling the observations in the sample and splitting the sample into two sub-samples of equal size, or differing by one if the cardinality of the sample is odd. Using the FDH estimator, the procedure includes estimating unconditional efficiency (restricting the frontier to be the same for all $t \in T$) with the first sub-sample, and conditional efficiency (allowing the frontier to change depending on $t \in T$) with the second sub-sample. I use the test statistics proposed by Daraio et al. (2018) to estimate the difference between the mean unconditional and mean conditional efficiency estimates. Since this test relies on a random sample-split, the uncertainty

from a single samples-split is reduced by using the information from multiple sample-splits through a bootstrap procedure proposed by Simar and Wilson (2020). Using this method, I randomly split the sample 100 times and calculate the sample average of the test statistics over the 100 splits. I then use 1,000 bootstrap replications to estimate the p-value of the average of the test statistics. If the average test statistic is statistically significant, separability is rejected. Moreover, the distribution of p-values from the test statistics generated over the 100 sample splits are compared to the uniform distribution on $[0, 1]$ using the Kolmogorv-Smirnov (KS) test.[1] Since the 100 p-values used to construct the KS statistic are not independent, the bootstrap procedure discussed by Simar and Wilson (2020) is used to estimate the distribution of the KS statistic and make inference on the uniformity of the 100 p-values. If uniformity of the p-values is rejected, then separability can also be rejected. For technical details see Daraio et al. (2018) and Simar and Wilson (2020).

In order to test for convexity versus non-convexity and CRS versus VRS, I use results from Kneip et al. (2015, 2016) and Simar and Wilson (2020). Kneip et al. (2015) provide central limit theorems for FDH and DEA efficiency estimates and Kneip et al. (2016) use these results to develop specific tests for convexity, returns to scale, and differences in mean efficiency across groups of producers. As with the test for separability, the tests for convexity and returns to scale rely on randomly splitting the sample, but the uncertainty of doing so is reduced by using multiple sample-splits through the bootstrap procedure proposed by Simar and Wilson (2020). Furthermore, I follow the testing strategy of Apon et al. (2015, Figure 1) and test convexity before testing for returns to scale.[2] To test for convexity, I first randomly split the sample and obtain VRS-DEA estimates of efficiency with one sub-sample and FDH estimates with the other.[3] The test uses information from 1,000 bootstrap replications and 100 sample-splits to determine if the difference between the mean VRS-DEA and FDH estimates are statistically significant. If so, convexity should be rejected. As with the test for separability, I also examine the distribution of the 100 p-values generated from the sample-splits and use the KS test as described above. In this case, if the uniformity of the p-values is rejected the hypothesis of a convex production set should be rejected. If the test fails to reject convexity, returns to scale are tested. The test for returns to scale is similar to the test for convexity,

[1] As discussed by Simar and Wilson (2020), the uniform distrbution on $[0, 1]$ is the distribution of p-values under the null hypothesis due to the probability integral transform.

[2] Apon et al. (2015) only test returns to scale if convexity is rejected. However, Kneip et al. (2021) examine non-convex production sets with CRS frontiers.

[3] In both the test for convexity and returns to scale, the estimators are conditioned on T if separability with respect to T is rejected.

but the FDH estimator is replaced with the CRS-DEA estimator. For technical details see Apon et al. (2015), Kneip et al. (2015), Kneip et al. (2016), and Simar and Wilson (2020).

1.3 Data

1.3.1 Data on Natural Gas Pipelines and Variable Specification

The sample is an unbalanced panel of inputs and outputs by pipeline and production year. The data collected contain information on U.S. natural gas pipelines during the years 1996–2018 and were collected from the Form 2 Financial Data on Major Natural Gas Pipelines (Form 2) produced by FERC. The Form 2 are financial and engineering data natural gas pipelines report to FERC and are made publicly available. These are the most commonly used data for measuring productivity and efficiency of natural gas pipelines in the U.S.

I specify pipelines as having $p = 1$ input, operating expenditure (OPEX), and $q = 4$ unique outputs including total delivery volume, total horsepower, length, and peak delivery volume.[4] Jamasb et al. (2008) note that having a single monetary cost input avoids the issue of non-reporting of some physical quantities of traditional inputs in the Form 2, such as tons of steel. In addition, I assume that all pipeline firms face the same input prices in a given year t. Jamasb et al. (2006) and Jamasb et al. (2008) note that pipeline costs are largely comprised of globally traded commodities, such as steel, which is used to replace pipes and compressor station turbines, and natural gas, which is used to fuel compressor stations. These input prices are unlikely to vary across pipeline firms. Thus, with pipelines facing the same input prices in the same year, the input orientation measure of technical efficiency, as described above, is the proportion by which the OPEX of producing output quantities, y, can be reduced. This is a measure of cost efficiency as described in Simar and Wilson (2019b).

Although input prices are assumed to not vary across firms, prices for natural gas declined after the Shale Revolution. Thus pipelines face a different cost set over time due to technological advances in natural gas production. To control for this Jamasb et al. (2008) suggest removing fuel cost from the cost input measure. However, I choose to include fuel costs. Thus, tests for changes in

[4]Delivery volume is the quantity of gas delivered by a pipeline in a production year, measured in dekatherms. Horsepower is the sum of the horsepower ratings across all compressor stations in a pipeline. Total length is the total mileage of pipe in a pipeline. Finally, peak delivery volume is a proxy for a pipeline's delivery capacity and measures the largest daily delivery volume reported by a pipeline in a production year.

efficiency over time will also reflect cost reduction due to technological advancement in the upstream production of natural gas. Estimation results that do not include total fuel costs are reserved for Appendix B.

The $q = 4$ outputs were chosen based of a review of previous studies evaluating the efficiency of pipelines. Previous papers on natural gas pipelines use various specifications for inputs and outputs. Table 1.1 contains a review of previous research examining the efficiency of pipelines. Traditionally horsepower, length, tons of steel, or other measures of pipeline capacity have been used as inputs, while delivery volume has been a common output. However, the ability to specify these inputs and outputs was possible due to the availability of physical quantities of some additional inputs such as labor. For example, Aivazian et al. (1987), Sickles and Streitwieser (1992), and Granderson (2000), include labor as an input. To my knowledge, FERC Form 2 data productions have not included data on pipeline employees since at least 1991. Thus, more recent studies, which include Hess (2000), Jamasb et al. (2008), and Nieswand et al. (2010), do not include labor or other physical inputs specified in older studies. In addition, studies that have used horsepower or length as an input relied on parametric methods, with the exception of Sickles and Streitwieser (1992). Therefore, I choose a specification of inputs and outputs that has been previously used for nonparametric estimation of pipeline efficiency and relies on data currently available in the Form 2.

Two papers rely on nonparametric DEA estimation of pipeline efficiency using available Form 2 data. Jamasb et al. (2008) examine 39 interstate pipelines from 1996–2004 and specifies two models by stipulating delivery volume, total length, and compressor station horsepower as outputs, while using total expenditure (TOTEX) as an input. This is known as TOTEX benchmarking, where TOTEX is considered an input or expenditure on a pipeline, while outputs are seen as cost drivers. The outputs of length and compressor station volume are included as capacity measures or capital outputs. While the length of the pipeline may change seldom from year to year, horsepower can be added fairly frequently. In addition, delivery volume captures a pipeline's ability to compete with other pipelines for delivery contracts. Jamasb et al. (2008) note that delivery volume accounts for a pipelines ability to use better management, trading techniques, or other unobserved approaches to increase throughput. Thus, under TOTEX benchmarking, efficiency measurements estimate how well pipelines maximize output quantities (while holding TOTEX fixed), minimize TOTEX (while

holding output fixed), or do both. The authors note the difficulties associated with measuring TOTEX accurately, which includes measuring capital costs. They offer an alternative model where revenue is utilized as an input variable.

Nieswand et al. (2010) offer a similar approach to Jamasb et al. (2008), however OPEX, is used as an input. This is known as OPEX benchmarking. While OPEX includes operating and maintenance expenditure, it excludes capital costs. While Jamasb et al. (2006) and Nieswand et al. (2010) both note that TOTEX and OPEX are correlated measures, OPEX benchmarking avoids the complication of the measurement error associated with recording capital costs. Like Jamasb et al. (2008), Nieswand et al. (2010) include delivery volume, system length, and horsepower in their set of outputs, but also include peak delivery volume as another proxy for pipeline capacity. As with TOTEX benchmarking, OPEX benchmarking estimates how well pipelines maximize output quantities (while holding OPEX fixed), minimize OPEX (while holding output fixed), or do both. Nieswand et al. (2010) also demonstrate the efficacy of dimension reduction with OPEX benchmarking and their specified outputs.[5] Because of this, I follow the approach of Nieswand et al. (2010) and have the same specified inputs and outputs. Table 1.2 provides summary statistics for the sample of specified inputs and outputs. In addition, Table 1.3 shows the number of pipelines observed in each year of the sample, where the number of pipelines observed each year ranges from 43, in 1996, to 72, in 2018.

1.3.2 Dimension Reduction Method

The curse of dimensionality is a well known issue in nonparametric benchmarking, where the convergence rates of DEA and FDH estimators decrease as the number of inputs and outputs, $p+q$, increases. The rate of convergence for the VRS-DEA and FDH estimators have been established by Kneip et al. (1998), Park et al. (2000), and Daouia et al. (2017). With $p + q = 5$ dimensions the VRS-DEA estimator converges at rate $n^{\frac{1}{3}}$, while the FDH estimator converges at a slower rate $n^{\frac{1}{5}}$.

To mitigate the slow rate of convergence, I use eigensystem methods along the lines of Wilson (2018) to reduce the dimensionality of the problem. More specifically, I find the $(n \times 1)$ first

[5] Nieswand et al. (2010) reduce the dimensions of their sample through eigensystem decomposition of the correlation matrix. As noted in Section 1.1, this may not always be reliable for dimension reduction. Because of this, and other reasons noted in Wilson (2018), I use dimension reduction methods described in Section 1.3.2.

principal component, Y^*, of the moment matrix, $\mathbf{Y}'\mathbf{Y}$, of my observed $(n \times q)$ output matrix, $\mathbf{Y}$. Y^* is the matrix product of the observed output matrix and the eigenvector corresponding to the largest eigenvalue of $\mathbf{Y}'\mathbf{Y}$. The ratio of the largest eigenvalue of the moment matrix of $\mathbf{Y}$ to the sum of all of its eigenvalues measures the amount of independent linear information Y^* contains from the q columns of $\mathbf{Y}$. From the sample described above this ratio ranges from 0.975 to 0.994 when limiting the sample to specific years, and equals 0.979 when pooling the entire sample.[6] Results from Wilson (2018) suggest that reducing the dimensions by replacing $\mathbf{Y}$ with Y^* is reasonable since little information is lost. In my application, I estimate Y^* using the pooled sample of data from 1996 through 2018. This is done to maximize the amount of data used to estimate the principal components. See Wilson (2018) for technical details.

With dimension reduction the convergence rate for the VRS-DEA estimator improves from $n^{\frac{1}{3}}$ to $n^{\frac{2}{3}}$, and the convergence rate for the FDH estimator improves from $n^{\frac{1}{5}}$ to $n^{\frac{1}{2}}$. In this case, the VRS-DEA estimator achieves a convergence rate faster than the standard parametric rate, while the FDH estimator converges at the parametric rate. Because of this, I conduct all estimates using dimension-reduced data.

1.4 Estimation and Results

I first test for separability with respect to T to determine if efficiency estimates should allow for potentially a different frontier each production year. Table 1.4 reports several iterations of tests for separability of Ψ^t with respect to T using the FDH estimator.[7] The top portion of Table 1.4 reports results for separability tests for consecutive year pairs. In this case, T is treated as a binary variable. The results depict some evidence of the frontier for pipelines changing from one year to the next, but in most cases Ψ^t does not appear to change. It is not unreasonable to assume that the technology for pipelines is not rapidly changing from year to year, but it should be noted that under almost all orientations changes in Ψ^t are apparent in the years 2008 though 2015, a period of time just after the start of the Shale Revolution.

The second portion of Table 1.4 reports results for a similar set of separability tests, to

[6] In this case when pooling the entire sample, the first principal component of the output matrix contains 97.9 percent of the independent linear information contained in $\mathbf{Y}$.

[7] The FDH estimator is used as it does not rely on convexity of production set and is the safest estimator to use for testing separability without testing for convexity prior.

those described above, but the sample is pooled in the intervals of 1996–2000, 2001–2006, 2007–2012, and 2013-2018. Here I test for separability of consecutive time interval pairs. In this case, even when the data are pooled over longer time periods, separability is not always rejected. This again, only suggests that the frontier may remain the same for consecutive periods of time.

The third and fourth portions of Table 1.4 report tests for separability that no longer treat T as a binary outcome. The third portion of Table 1.4 reports test results that use the same pooling of years as in the second portion of Table 1.4, but I test separability across all four time intervals, and not just consecutive interval pairs. likewise, the fourth portion of Table 1.4 reports results of separability tests across all 23 years in the sample. The tests strongly reject separability of the frontier in both cases. Moreover, Table 1.5 reports the results of the KS tests corresponding to the p-values generated from the sample-splitting procedure used to develop the estimates in Table 1.4. In this case, the rejection rate of the uniformity of the p-values is similar to the rate of rejections observed in Table 1.4. This provides additional evidence against the null hypothesis, separability with respect to time. Thus, I assume the alternative hypothesis, (1.12), otherwise known as non-separability, as described in Daraio et al. (2018). This allows the production set to vary by production year. Assuming separability under non-separability of the production set renders efficiency estimates meaningless, since $\Psi^{t\partial}$ may not be attainable by some firms in certain years. Thus, I condition my estimates on production year and allow the possibility of a different frontier every production year.

As stated above, I follow the testing strategy noted in Apon et al. (2015) and test for convexity then returns to scale, if necessary. Table 1.6 presents the results of the tests for convexity. Out of the 69 tests, convexity is rejected only 7 times, or at a rate of $7/69 \approx 0.101$. Moreover, in sample years where tests reject convexity, rejections only occur under one orientation. Thus, I find little evidence to reject convexity under this set of tests. However, I also examine the distribution of the p-values from the sample-splitting procedure used to calculate the test statistics reported in Table 1.6. Table 1.7 reports the KS test results comparing the distributions of the p-values to the uniform distribution on $[0, 1]$. I find strong evidence against the uniformity of the p-values, with a $37/69 \approx 0.536$ rejection rate. This offers stronger evidence against the convexity of the production set. Given this evidence against convexity, I assume that the production set is non-convex.

Table 1.8 presents the results of the tests of CRS versus VRS. The results indicate CRS is

always rejected at the 1 percent level under the output orientation, and is strongly rejected under the hyperbolic orientation, with only one test failing to reject CRS. Table 1.9 reports the results of the corresponding KS tests and shows similar results in rejecting CRS. Under the input orientation, under both sets of tests, the results are more ambiguous where the tests reject the null only in some years. Despite only rejecting CRS in a few instances under the input orientation, I relax the assumption of CRS and allow for VRS.

These results suggest that the FDH estimator, conditional on production year, will yield consistent estimates of technical efficiency with my sample. While Jamasb et al. (2008) rely on both the CRS and VRS-DEA estimators and Nieswand et al. (2010) rely on the VRS-DEA estimator, I find evidence against separability with respect to time, convexity, and CRS. This suggests that the FDH estimator conditional on production year is likely the only consistent estimator of the technical efficiency of natural gas pipelines. Moreover, while the VRS-DEA estimator has a faster convergence rate than the FDH estimator, simulations by Wilson (2018) show that the FDH estimator with dimension reduction yields less estimation error than the VRS-DEA estimator with dimension reduction. Given these results, all further estimates are done using the FDH estimator.

Table 1.10 presents results on changes in mean efficiency over the sample period. In this test, I use the estimator for differences in mean efficiency across groups of producers, allowing the frontier to vary by group, as described by Kneip et al. (2016). If mean pipeline efficiency is greater in year t than in period $t-1$, the test statistics are large and positive. Conversely, drops in technical efficiency result in a large negative test statistic.[8] The results in the top portion of Table 1.10 compare the mean technical efficiencies of pipelines in consecutive year pairs. In this case, year-to-year changes in technical efficiency varies over the sample period. I find that all instances of drops in technical efficiency, where the null is rejected at the 1 percent level, are contained in the set of years 2001 through 2008. This finding is consistent with Jamasb et al. (2008) and Moss (2008) who both suggest that the in the early 2000s, prior to the Shale Revolution, a series of mergers in the natural gas pipeline industry may have hindered technical efficiency change. However, there are instances of

[8]Note that the tests of differences in mean efficiency, developed by Kneip et al. (2016), require independence between the groups of producers. When testing for changes in mean efficiency over time, pipelines are observed at multiple points in time resulting in a covariance issue. However, Wilson and O'Loughlin (2021) perform similar tests for U.S. municipalities, and they argue that any covariance is likely positive due to inertia. In my case, inertia likely plays a role as well where a pipeline that performs poorly (or well) in one year is likely to continue performing poorly (or well) in a subsequent year. Consequently, ignoring positive covariance makes my tests for changes in mean efficiency conservative as I bias towards failing to reject the null.

significant increases to technical efficiency in this same time interval. Notably, technical efficiency improves during 2003–2004 in the hyperbolic orientation, with significance at the 5 percent level, and during 2005–2006, with significance at the 5 percent level in the input orientation and at the 1 percent level in the output and hyperbolic orientations.

From the period during 2008–2013, the estimates indicate that technical efficiency is non-decreasing with instances of improvement in all orientations. Specifically, technical efficiency improves during 2009–2010 and 2012–2013 under the input orientation, during 2008–2009 and 2011–2012 under the output orientation, and during 2008–2010 and 2012–2013 under the hyperbolic orientation. These improvements to technical efficiency reject the null at at least the 5 percent level with the majority of rejections occurring at the 1 percent level. This provides strong evidence of technical efficiency improvements to the natural gas pipeline industry in the years following the start of the Shale Revolution.

The second portion of Table 1.10 depicts net changes in mean technical efficiency. I compare mean efficiencies in the years 2007, 2008, 2009, 2010, and 2011 to the last year in the sample, 2018. There is strong evidence of technical efficiency improvements from 2007, 2008, and 2009 through 2018, with rejections of the null at the 1 percent level under all orientations. In addition, I find evidence for technical efficiency improvement in the output orientation from 2010 and 2011 through 2018, with rejections of the null at the 1 percent level. This supports the hypothesis that during the years following the start of the Shale Revolution the technical efficiency of pipelines improved.

Because I reduce the dimensionality of pipeline outputs such that $p = q = 1$, I can directly measure the mean productivity of pipelines over time. The mean productivity of pipelines in year t is defined as

$$\widehat{Productivity}_t = n_t^{-1} \sum_{i}^{n_t} \frac{Y_{i,t}^*}{OPEX_{i,t}^*}, \tag{1.13}$$

and the standard Lindeberg-Fellar Central Limit Theorem can be used to make inference on productivity changes over time.[9] Table 1.11 reports the test statistics of differences in the productivity estimates. The results suggest that from year to year the productivity of pipelines does not change as all the estimates measuring year-over-year productivity change are not statistically different from

[9]Similar covariance issues as noted in footnote 8 arise when examining productivity over time. However, similar reasoning applies here where any covariance is likely positive making the tests conservative.

zero. However, examining changes from 2007, 2010, and 2011 to 2018 shows that the net change of natural gas pipeline productivity was positive and statistically significant at at least the 10 percent level. This suggests that during the period of the Shale Revolution pipelines were able to increase output quantities for a fixed quantity of OPEX. For example, mean productivity in 2007 is estimated at 15.2 and increases to 28.6 by 2018, implying an 88.2 percent increase in productivity over this period. This is consistent with Dreskin and Boss (2010) who discuss how engineering improvements to pipelines allowed increased throughout at a lower cost. However, it is unclear if productivity improvements are largely driven by these technical improvements or if productivity improvements are driven by increased utilization of pipelines.[10] Moreover, I observe pipeline entry and exit in my sample, which both can positively impact estimated technical efficiency and productivity. To address this, I examine "older" pipelines in my sample that are observed every production year. These pipelines may be constrained in adopting newer technology which removes the effect of technological improvements. A discussion of this analysis and the results are reported in Appendix A.

I next test for change in technology over the sample period to determine the direction in which the frontier is changing. In particular let $Z_i^t = (X_i^t, Y_i^t)$ denote pipeline i's input-output pair at time t where $t \in \{1, 2\}$. Change in technology relative to pipeline i's position in time period 1 and 2 is given by

$$\mathcal{T}_i = \left[\frac{\gamma(Z_i^2 \mid \Psi^1)}{\gamma(Z_i^2 \mid \Psi^2)} \times \frac{\gamma(Z_i^1 \mid \Psi^1)}{\gamma(Z_i^1 \mid \Psi^2)} \right]^{1/2}. \tag{1.14}$$

This expression is the hyperbolic-oriented analog of the measure of change in technology index that appears in the Malmquist decompositions of Ray and Desli (1997), Gilbert and Wilson (1998), Simar and Wilson (1998), and Simar and Wilson (1999), and consists of a geometric mean of two ratios. The first ratio in $\mathcal{T}_i$ measures any shift in Ψ^∂ relative to i's position in period 2, while the second ratio measures any shift in Ψ^∂ relative to i's position in period 1. Values of $\mathcal{T}_i$ greater (less) than 1 indicate an upward (downward) shift in technology. Values of $\mathcal{T}_i$ equal to 1 indicate the technology is not shifting.

Estimates of $\mathcal{T}_i$ are made by substituting the hyperbolic FDH Estimator in for Ψ^t, where $t \in T$. Inference on the estimates are made using the central limit theorem results from Simar and Wilson (2019a). The top portion of Table 1.12 presents the estimates of change in technology in

[10] In this case, the Shale Revolution decreased the price of natural gas and increased demand for natural gas fueled energy production as well as pipelines to transport fuel.

consecutive year pairs. The results indicate that the technology of the pipelines changes abruptly year-over-year, but neither consistently increases or regresses.[11] At the start of the Shale Revolution technology increased to a large degree during 2007–2008, but regressed from 2008 through 2010. This regress could be attributed to the decrease in energy consumption during the Great Recession. Technology would then increase during 2010–2012, but then again regress during 2012–2013.

The bottom portion of Table 1.12 presents tests for net changes in technology from 2007, 2008, 2009, 2010, and 2011 through 2018. In this case, there is strong evidence of net change in technology in the upward direction when comparing 2007, 2009, 2010 and 2011 to the technology in 2018 with three rejections of the null at less than the 1 percent level and one rejection at the 5 percent level. When comparing 2008 to 2018 the technology indicates a strong net technical regress. However, the year-over-year change in technology from 2007 through 2008 was the largest increase in the technology in year-over-year terms. This indicates that the technology increased sharply in 2008, then regressed the following year, but would then increase in net terms by 2018. Thus, there is strong evidence of the technology increasing in net terms from the beginning of the Shale Revolution to the end of the sample. In addition, Figure 1.3 and Figures C.1–C.22, in Appendix C, show plots of the FDH over time, and do not provide strong evidence of technical regress during the sample period.

Finally, to confirm that the increases in technical efficiency are not driven by a decreasing number of observed firms over time, I examine the number of pipelines observed each production year and the number of pipelines operating on $\Psi^{t\partial}$. Table 1.3 shows the number of observed pipelines in the sample by year. In this case, the number of observed pipelines increases in the sample period with only three instances where the pipeline counts drop by at most two pipelines.[12] This suggests the increasing number of pipelines that define $\Psi^{t\partial}$ of the FDH is not driven by a reduction in the number of observed pipelines each year, and because the Form 2 data are representative of all major interstate pipelines, the number of competing pipelines increased over the sample period. This is consistent with Dreskin and Boss (2010) who explain that increased pipeline entry during the period of the Shale Revolution created an incentive to adopt cost saving technology. Although technical

[11] This test is more sensitive to changes in the frontier than the tests for separability, as it is comparing the geometric mean of ratios of efficiency estimates. This provides further evidence that conditioning estimates on T is necessary for consistent efficiency estimates.

[12] From 2009 through 2010 the pipeline counts drop from 64 to 62, from 2014 though 2015 the pipeline counts drop from 69 to 67, and from 2015 through 2016 the pipeline counts drop from 67 to 66.

efficiency did increase and increased pipeline competition arguably played a role in this, Annibali et al. (2012) discuss how increased utilization of pipelines and the decline in gas costs affected pipeline operations as well. In this case, both increased capacity utilization and cost saving technology could have improved the technical efficiency and productivity of natural gas pipelines.

1.5 Summary and Conclusion

In the analysis outlined above, I use nonparametric DEA and FDH estimation techniques to measure the technical efficiency of pipelines during the period of expansion in the supply of shale gas after 2007. Using tests for convexity, returns to scale and separability I find that the production set for pipelines exhibits convexity, VRS, and non-separability with respect to production year. These results have not been considered in previous work using nonparametric techniques to examine pipeline efficiency. The results on efficiency estimations show that technical efficiency prior to the Shale Revolution declined, however there is strong evidence for improvements to technical efficiency and productivity during the Shale Revolution. These findings suggest that the Shale Revolution provided an opportunity for pipelines to capture efficiency gains from declining fuel costs and increasing utilization (due to fracking) or from improved transmission technology induced by pipeline competition to deliver shale gas. Future research should parse the impacts of these two effects on pipeline technical efficiency.

Table 1.1: Summary of Literature

Author(s)	Inputs	Outputs	Observations	Techniques	Data
Callen (1978)	(1) Horsepower (2) Tons of Steel	(1) Delivery Volume	28 U.S. Interstate Pipelines (1965)	Econometric Production Function	FPC Annual Statistics
Aivazian et al. (1987)	(1) Horsepower (2) Tons of Steel (3) Compressor Fuel Volume (4) Labor	(1) Delivery Volume × Length	14 U.S. Interstate Pipelines (1953 – 1979)	Econometric Production Function	FPC Annual Statistics
Sickles and Streitwieser (1992)	(1) Horsepower (2) Tons of Steel (3) Compressor Fuel Volume (4) Labor	(1) Delivery Volume × Length	14 U.S. Interstate Pipelines (1977 – 1985)	SFA (Translog), DEA	FERC: Form–2
Ellig and Giberson (1993)	(1) Sales Volume (2) 3rd Party Delivery Volume (3) Delivery Volume (4) Length (5) Gas Price	(1) O&M Expense	50 Texan Pipelines (1989 – 1990)	Econometric Cost Function	Railroad Commission of Texas Annual Report
Granderson (2000)	Price and Cost Share of: (1) Horsepower (2) Tons of Steel (3) Compressor Fuel Volume (4) Labor	(1) Compressor Fuel Volume	14 U.S. Interstate Pipelines (1977 – 1989)	SFA: Translog Cost Function	Department of Energy Annual Statistics of Natrual Gas Pipeline Companies
Hess (2000)	(1) Total OPEX (2) Transmission Assets (in $)	(1) Revenue	47 U.S. Interstate Pipelines (1996 – 2005)	SFA: Cobb–Douglass	FERC: Form–2
Jamasb et al. (2008)	(1) TOTEX (2) Revenue	(1) Delivery Volume (2) Length (3) Horsepower	39 U.S. Interstate Pipelines (1996 – 2004)	DEA: CRS and VRS	FERC: Form–2
Nieswand et al. (2010)	(1) OPEX	(1) Delivery Volume (2) Length (3) Horsepower (4) Peak Delivery Volume	37 U.S. Onshore Interstate Pipelines (2007)	DEA: VRS & PCA	FERC: Form–2

Table 1.2: Summary Statistics: Form 2 Data 1996–2018

	Min	1st Q	Median	Mean	3rd Q	Max
OPEX [MM $]	0.181	9.442	29.630	67.759	96.547	536.942
Delivery Volume [MM Dth.]	0.813	174.156	412.695	626.802	832.521	5,525.098
Peak Delivery Volume [MM Dth.]	0.049	0.738	1.680	2.356	2.931	14.755
HP [K HP]	0.556	59.011	142.090	327.159	472.714	2,274.891
Length [K Mi.]	0.050	0.392	1.482	3.319	5.263	23.061
Y^*	7.923	183.801	416.654	706.915	935.213	5,938.682

This table shows summary statistics for the sample collected form Form 2 data. The principal component used in all analysis is denoted as Y^*.

Table 1.3: Count of Observed Pipelines by Year

Year	Pipeline Count	Pipeline Count on $\Psi^{t\partial}$	Percent of Pipelines on $\Psi^{t\partial}$
1996	43	10	0.23
1997	44	9	0.20
1998	44	10	0.23
1999	47	8	0.17
2000	47	9	0.19
2001	47	7	0.15
2002	50	7	0.14
2003	51	7	0.14
2004	51	5	0.10
2005	53	4	0.08
2006	55	7	0.13
2007	55	6	0.11
2008	62	6	0.10
2009	64	7	0.11
2010	62	9	0.15
2011	66	10	0.15
2012	70	12	0.17
2013	69	14	0.20
2014	69	10	0.14
2015	67	15	0.22
2016	66	13	0.20
2017	68	15	0.22
2018	72	13	0.18

Table 1.4: Separability Test With Respect to Time (FDH Estimator)

	Input–Orientation		Output–Orientation		Hyperbolic	
Period	Statistic	p-value	Statistic	p-value	Statistic	p-value
1996–1997	1.235	0.170	2.981	0.172	4.111	0.474
1997–1998	0.255	0.880	3.511	0.046**	5.340	0.018**
1998–1999	1.366	0.135	3.741	0.011**	3.789	0.386
1999–2000	0.774	0.489	3.672	0.129	4.034	0.253
2000–2001	0.568	0.585	3.074	0.271	4.843	0.248
2001–2002	0.140	0.871	4.099	0.047**	4.262	0.388
2002–2003	0.940	0.388	3.325	0.296	4.951	0.285
2003–2004	1.200	0.267	2.809	0.912	5.459	0.554
2004–2005	1.191	0.152	4.082	0.146	6.472	0.136
2005–2006	1.650	0.155	3.542	0.259	7.376	0.028**
2006–2007	1.232	0.509	5.174	0.079*	6.500	0.160
2007–2008	2.434	0.004***	5.350	0.002***	9.518	0.000***
2008–2009	2.330	0.026**	3.942	0.089*	8.124	0.003***
2009–2010	2.613	0.006***	3.164	0.368	7.481	0.000***
2010–2011	1.773	0.012**	3.165	0.255	4.980	0.120
2011–2012	1.868	0.000***	4.112	0.004***	7.241	0.000***
2012–2013	2.331	0.000***	2.476	0.321	4.984	0.000***
2013–2014	1.701	0.000***	4.633	0.008***	4.932	0.005***
2014–2015	0.954	0.014**	2.767	0.069*	4.200	0.012**
2015–2016	0.299	0.386	2.116	0.103	3.088	0.381
2016–2017	−0.283	0.970	2.247	0.212	2.909	0.620
2017–2018	0.232	0.228	2.344	0.049**	3.571	0.046**
[96, 00]–[01–06]	0.092	0.259	5.713	0.002***	9.990	0.001***
[01, 06]–[07–12]	1.015	0.059*	4.610	0.570	14.381	0.000***
[07, 12]–[13–18]	4.910	0.000***	5.022	0.032**	14.092	0.000***
[96, 00], [01, 06], [07, 12], [13, 18]	5.973	0.000***	8.418	0.000***	6.710	0.000***
96, 97,...,18	14.600	0.000***	16.440	0.000***	19.000	0.000***

The test statistic, described in Daraio et al. (2018), is computed using the FDH estimator with dimension-reduced data, such that $p = q = 1$. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 1996–2018. The p-values are developed using 100 sample-splits and 1,000 bootstrap replications using techniques described in Simar and Wilson (2020). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table 1.5: Separability Test (KS Test) With Respect to Time (FDH Estimator)

	Input–Orientation		Output–Orientation		Hyperbolic	
Period	Statistic	p-value	Statistic	p-value	Statistic	p-value
1996–1997	0.412	0.218	0.567	0.407	0.745	0.512
1997–1998	0.247	0.713	0.655	0.109	0.815	0.093*
1998–1999	0.432	0.169	0.626	0.075*	0.570	0.977
1999–2000	0.385	0.279	0.600	0.421	0.695	0.522
2000–2001	0.330	0.381	0.570	0.588	0.742	0.540
2001–2002	0.192	0.832	0.700	0.095*	0.727	0.553
2002–2003	0.385	0.296	0.558	0.724	0.761	0.544
2003–2004	0.475	0.140	0.585	0.892	0.744	0.887
2004–2005	0.399	0.197	0.556	0.738	0.615	0.978
2005–2006	0.455	0.257	0.497	0.901	0.609	0.967
2006–2007	0.439	0.422	0.692	0.472	0.702	0.945
2007–2008	0.546	0.038**	0.590	0.362	0.811	0.194
2008–2009	0.570	0.045**	0.488	0.909	0.693	0.756
2009–2010	0.500	0.098*	0.572	0.617	0.720	0.569
2010–2011	0.522	0.027**	0.508	0.719	0.714	0.682
2011–2012	0.466	0.000***	0.598	0.213	0.644	0.277
2012–2013	0.528	0.000***	0.491	0.712	0.574	0.548
2013–2014	0.521	0.001***	0.748	0.025**	0.746	0.091*
2014–2015	0.374	0.016**	0.624	0.052*	0.667	0.234
2015–2016	0.271	0.130	0.516	0.167	0.579	0.740
2016–2017	0.231	0.402	0.542	0.236	0.607	0.720
2017–2018	0.275	0.057*	0.565	0.043**	0.644	0.213
[96, 00]–[01–06]	0.215	0.688	0.470	0.991	0.835	0.648
[01, 06]–[07–12]	0.478	0.047**	0.545	0.998	0.831	0.621
[07, 12]–[13–18]	0.652	0.000***	0.433	1.000	0.702	0.956
[96, 00], [01, 06], [07, 12], [13, 18]	0.843	0.001***	0.766	0.012**	0.710	0.008***
96, 97,...,18	0.997	0.000***	0.999	0.000***	0.998	0.000***

The test statistic is from the KS test as described in Simar and Wilson (2020). In this case, I use the KS test to test for uniformity of the p-values estimated from the multiple sample-splits used in the tests presented in Table 1.4. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 1996–2018. The p-values are developed using 100 sample-splits and 1,000 bootstrap replications using techniques described in Simar and Wilson (2020). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table 1.6: Convexity Test

	Input–Orientation		Output–Orientation		Hyperbolic	
Year	Statistic	p-value	Statistic	p-value	Statistic	p-value
1996	−1.567	0.999	−1.055	0.992	−1.564	1.000
1997	−1.549	0.993	−0.904	0.938	−1.314	0.983
1998	−2.035	0.999	−0.553	0.706	−1.289	0.973
1999	−1.932	0.998	−0.706	0.750	0.229	0.126
2000	−1.214	0.953	0.116	0.198	1.102	0.006***
2001	−0.639	0.722	−0.292	0.454	0.796	0.008***
2002	−0.960	0.954	−0.802	0.899	−0.763	0.898
2003	−2.230	0.998	−1.743	0.984	−1.644	0.971
2004	−1.662	0.973	−1.301	0.915	−0.791	0.694
2005	−4.589	1.000	−2.965	1.000	−4.059	1.000
2006	−2.598	0.999	−0.348	0.445	−1.126	0.848
2007	−1.516	0.876	−1.660	0.911	−2.366	0.986
2008	−2.882	0.978	−2.285	0.931	−2.057	0.890
2009	−4.157	1.000	−0.900	0.631	−2.259	0.993
2010	−2.056	0.972	−1.084	0.637	−0.910	0.541
2011	−1.694	0.997	−0.652	0.788	−0.811	0.856
2012	−3.813	1.000	0.614	0.016**	−3.021	1.000
2013	−0.039	0.354	0.221	0.208	−0.043	0.362
2014	−2.848	1.000	0.218	0.092*	−1.052	0.883
2015	−2.148	1.000	0.544	0.039**	−1.382	0.985
2016	−0.411	0.764	0.766	0.024**	0.344	0.121
2017	−1.153	0.999	0.813	0.011**	−0.296	0.759
2018	−2.669	1.000	0.138	0.241	−1.655	0.999

The test statistic, described in Kneip et al. (2016), is computed using dimension-reduced data, such that $p = q = 1$. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 1996–2018. The p-values are developed using 100 sample-splits and 1,000 bootstrap replications using techniques described in Simar and Wilson (2020). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table 1.7: Convexity Test (KS Test)

	Input–Orientation		Output–Orientation		Hyperbolic	
Year	Statistic	p-value	Statistic	p-value	Statistic	p-value
1996	0.439	0.002***	0.392	0.011**	0.388	0.006***
1997	0.427	0.017**	0.317	0.112	0.377	0.051*
1998	0.504	0.001***	0.276	0.276	0.408	0.046**
1999	0.485	0.009***	0.351	0.157	0.292	0.330
2000	0.438	0.037**	0.238	0.482	0.394	0.076*
2001	0.312	0.221	0.294	0.290	0.295	0.288
2002	0.325	0.127	0.320	0.147	0.340	0.102
2003	0.443	0.049**	0.504	0.012**	0.437	0.071*
2004	0.434	0.093*	0.484	0.025**	0.353	0.238
2005	0.697	0.001***	0.493	0.032**	0.628	0.000***
2006	0.590	0.002***	0.359	0.214	0.389	0.158
2007	0.415	0.219	0.487	0.074*	0.512	0.048**
2008	0.611	0.041**	0.535	0.141	0.519	0.154
2009	0.720	0.000***	0.314	0.454	0.516	0.052*
2010	0.535	0.037**	0.348	0.508	0.398	0.345
2011	0.439	0.018**	0.377	0.072*	0.303	0.224
2012	0.700	0.001***	0.260	0.248	0.719	0.000***
2013	0.305	0.197	0.371	0.072*	0.294	0.228
2014	0.656	0.000***	0.270	0.423	0.410	0.102
2015	0.571	0.003***	0.248	0.368	0.477	0.014**
2016	0.238	0.246	0.316	0.102	0.190	0.470
2017	0.378	0.013**	0.269	0.099*	0.224	0.225
2018	0.658	0.000***	0.202	0.346	0.570	0.000***

The test statistic is from the KS test as described in Simar and Wilson (2020). In this case, I use the KS test to test for uniformity of the p-values estimated from the multiple sample-splits used in the tests presented in Table 1.6. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 1996–2018. The p-values are developed using 100 sample-splits and 1,000 bootstrap replications using techniques described in Simar and Wilson (2020). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table 1.8: Returns to Scale Test

	Input–Orientation		Output–Orientation		Hyperbolic	
Year	Statistic	p-value	Statistic	p-value	Statistic	p-value
1996	3.653	0.041**	6.494	0.000***	5.391	0.003***
1997	3.389	0.087*	5.403	0.000***	5.233	0.009***
1998	2.986	0.101	5.170	0.000***	5.134	0.007***
1999	1.942	0.232	4.801	0.000***	4.362	0.007***
2000	1.579	0.405	4.174	0.000***	3.977	0.006***
2001	0.532	0.540	2.678	0.001***	2.316	0.023**
2002	0.927	0.514	2.906	0.004***	2.333	0.044**
2003	1.600	0.470	3.836	0.000***	3.292	0.056*
2004	1.219	0.492	3.206	0.007***	2.820	0.087*
2005	1.422	0.627	3.606	0.005***	2.560	0.271
2006	2.271	0.243	5.203	0.000***	5.342	0.001***
2007	1.562	0.615	3.679	0.003***	3.694	0.063*
2008	0.784	0.753	4.147	0.000***	4.008	0.019**
2009	2.314	0.369	6.134	0.000***	6.099	0.007***
2010	1.786	0.479	4.430	0.000***	4.422	0.017**
2011	3.491	0.047**	5.614	0.000***	5.092	0.005***
2012	3.236	0.036**	6.678	0.000***	7.661	0.000***
2013	2.322	0.062*	3.712	0.000***	3.818	0.005***
2014	2.548	0.114	5.728	0.000***	6.014	0.000***
2015	2.886	0.040**	5.537	0.000***	5.567	0.000***
2016	3.352	0.007***	4.894	0.000***	4.757	0.001***
2017	3.276	0.015**	5.012	0.000***	4.549	0.003***
2018	4.180	0.009***	6.661	0.000***	6.756	0.000***

The test statistic, described in Kneip et al. (2016), is computed using dimension-reduced data, such that $p = q = 1$. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 1996–2018. The p-values are developed using 100 sample-splits and 1,000 bootstrap replications using techniques described in Simar and Wilson (2020). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table 1.9: Returns to Scale Test (KS Test)

	Input–Orientation		Output–Orientation		Hyperbolic	
Year	Statistic	p-value	Statistic	p-value	Statistic	p-value
1996	0.785	0.057*	0.904	0.001***	0.847	0.012**
1997	0.832	0.023**	0.951	0.000***	0.945	0.000***
1998	0.797	0.044**	0.918	0.000***	0.916	0.004***
1999	0.442	0.592	0.861	0.001***	0.742	0.040**
2000	0.567	0.315	0.919	0.000***	0.885	0.002***
2001	0.228	0.548	0.739	0.001***	0.663	0.032**
2002	0.345	0.541	0.827	0.001***	0.697	0.032**
2003	0.520	0.467	0.857	0.001***	0.785	0.024**
2004	0.380	0.601	0.773	0.004***	0.706	0.103
2005	0.408	0.812	0.697	0.037**	0.525	0.496
2006	0.607	0.271	0.837	0.002***	0.802	0.018**
2007	0.552	0.434	0.825	0.002***	0.833	0.014**
2008	0.339	0.773	0.843	0.002***	0.695	0.041**
2009	0.600	0.403	0.979	0.000***	0.865	0.014**
2010	0.518	0.567	0.852	0.001***	0.857	0.020**
2011	0.835	0.024**	0.939	0.000***	0.836	0.017**
2012	0.499	0.479	0.944	0.000***	0.814	0.013**
2013	0.621	0.117	0.695	0.010***	0.794	0.017**
2014	0.769	0.032**	0.930	0.000***	0.907	0.001***
2015	0.577	0.210	0.880	0.000***	0.824	0.013**
2016	0.887	0.001***	0.949	0.000***	0.984	0.000***
2017	0.832	0.010***	0.978	0.000***	0.947	0.001***
2018	0.775	0.045**	0.972	0.000***	0.942	0.001***

The test statistic is from the KS test as described in Simar and Wilson (2020). In this case, I use the KS test to test for uniformity of the p-values estimated from the multiple sample-splits used in the tests presented in Table 1.8. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 1996–2018. The p-values are developed using 100 sample-splits and 1,000 bootstrap replications using techniques described in Simar and Wilson (2020). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table 1.10: Test For Equivalency of Mean Efficiency: FDH Estimator

	Input–Orientation		Output–Orientation		Hyperbolic	
Period	Statistic	p-value	Statistic	p-value	Statistic	p-value
1996–1997	−1.779	0.075*	−1.316	0.188	−0.583	0.560
1997–1998	0.527	0.598	−0.533	0.594	0.221	0.825
1998–1999	−0.908	0.364	0.687	0.492	1.829	0.067*
1999–2000	0.699	0.485	−0.816	0.414	−0.877	0.381
2000–2001	0.584	0.559	0.518	0.604	−0.05	0.960
2001–2002	−1.699	0.089*	−1.25	0.211	−2.689	0.007***
2002–2003	−1.922	0.055*	−2.601	0.009***	−3.129	0.002***
2003–2004	1.106	0.269	0.256	0.798	2.251	0.024**
2004–2005	−2.562	0.010***	−4.664	0.000***	−4.887	0.000***
2005–2006	2.067	0.039**	6.869	0.000***	5.377	0.000***
2006–2007	0.851	0.395	−0.941	0.346	−1.347	0.178
2007–2008	−2.359	0.018**	−3.074	0.002***	−3.043	0.002***
2008–2009	0.374	0.709	4.483	0.000***	2.150	0.032**
2009–2010	4.084	0.000***	−0.88	0.379	5.354	0.000***
2010–2011	−0.216	0.829	1.127	0.260	−0.691	0.490
2011–2012	−0.508	0.611	2.581	0.010***	−1.517	0.129
2012–2013	3.485	0.000***	−1.183	0.237	2.329	0.020**
2013–2014	−2.490	0.014**	0.002	0.999	−2.182	0.029**
2014–2015	1.56	0.119	1.204	0.229	1.23	0.219
2015–2016	2.405	0.016**	1.266	0.206	2.220	0.026**
2016–2017	−1.422	0.155	0.337	0.736	−0.746	0.456
2017–2018	−1.435	0.151	−0.879	0.379	−1.899	0.058*
2007 & 2018	3.833	0.000***	5.460	0.000***	4.199	0.000***
2008 & 2018	5.629	0.000***	8.209	0.000***	6.651	0.000***
2009 & 2018	5.713	0.000***	3.979	0.000***	4.460	0.000***
2010 & 2018	1.062	0.288	4.326	0.000***	−0.517	0.605
2011 & 2018	1.415	0.157	3.471	0.001***	−0.063	0.950

The test statistic, described in Kneip et al. (2016), is computed using the FDH estimator with dimension-reduced data, such that $p = q = 1$. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 1996–2018. Since separability tests suggest that separability with respect to T does not hold, the estimator allows for a different frontier every production year. One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table 1.11: Productivity Estimates

Period	Statistic	p-value
1996–1997	0.785	0.432
1997–1998	1.452	0.146
1998–1999	0.803	0.422
1999–2000	−1.047	0.295
2000–2001	1.559	0.119
2001–2002	−0.014	0.989
2002–2003	−0.378	0.705
2003–2004	−0.610	0.542
2004–2005	−0.357	0.721
2005–2006	0.325	0.745
2006–2007	−1.313	0.189
2007–2008	1.495	0.135
2008–2009	−0.885	0.376
2009–2010	−1.200	0.230
2010–2011	−0.011	0.992
2011–2012	1.427	0.154
2012–2013	−1.149	0.251
2013–2014	1.490	0.136
2014–2015	−0.834	0.404
2015–2016	−0.707	0.480
2016–2017	−0.292	0.771
2017–2018	0.555	0.579
2007–2018	2.546	0.010***
2008–2018	−0.585	0.558
2009–2018	0.743	0.457
2010–2018	1.895	0.058*
2011–2018	2.189	0.028**

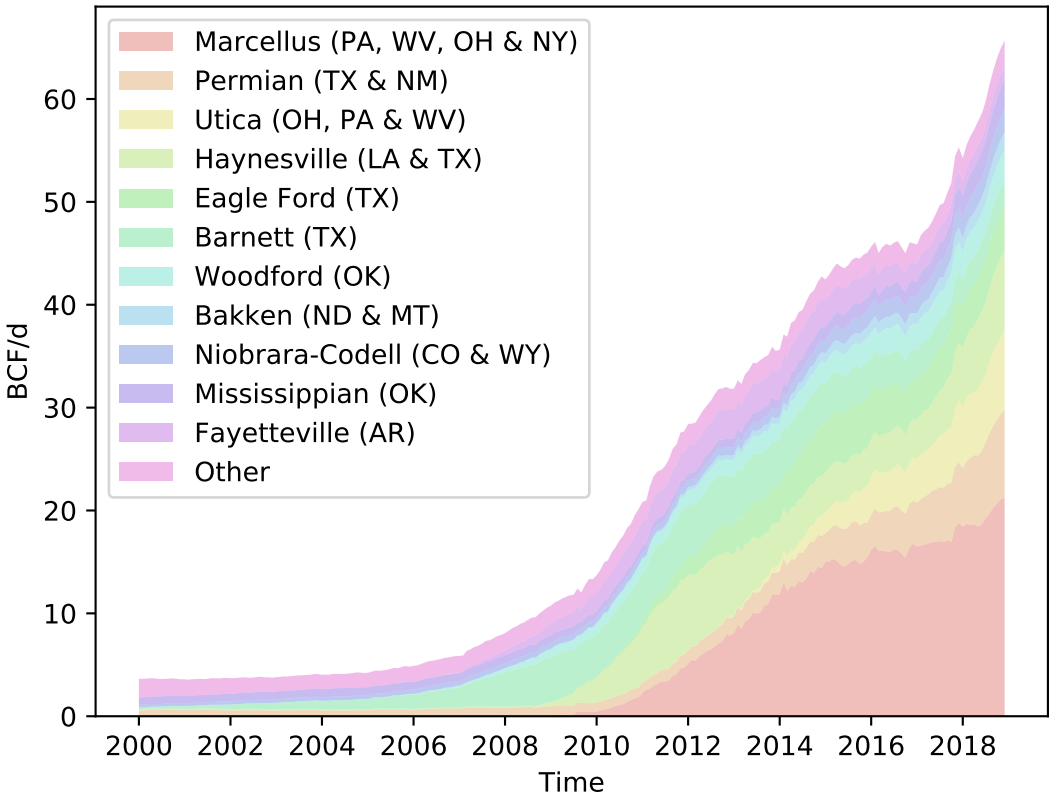
Productivity tests are conducted with dimension-reduced data ($p = q = 1$). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table 1.12: Tests for Change in Technology

Period	$\mathcal{T}$	p-value
1996–1997	1.054	0.001***
1997–1998	1.022	0.000***
1998–1999	0.953	0.000***
1999–2000	1.033	0.000***
2000–2001	1.034	0.126
2001–2002	1.079	0.000***
2002–2003	1.073	0.000***
2003–2004	0.854	0.000***
2004–2005	1.200	0.000***
2005–2006	0.765	0.000***
2006–2007	1.055	0.005***
2007–2008	1.377	0.000***
2008–2009	0.822	0.000***
2009–2010	0.812	0.000***
2010–2011	1.005	0.842
2011–2012	1.348	0.000***
2012–2013	0.785	0.000***
2013–2014	1.214	0.000***
2014–2015	0.932	0.000***
2015–2016	0.920	0.000***
2016–2017	1.004	0.303
2017–2018	1.124	0.000***
2007–2018	1.146	0.039**
2008–2018	0.838	0.000***
2009–2018	1.016	0.000***
2010–2018	1.235	0.001***
2011–2018	1.263	0.000***

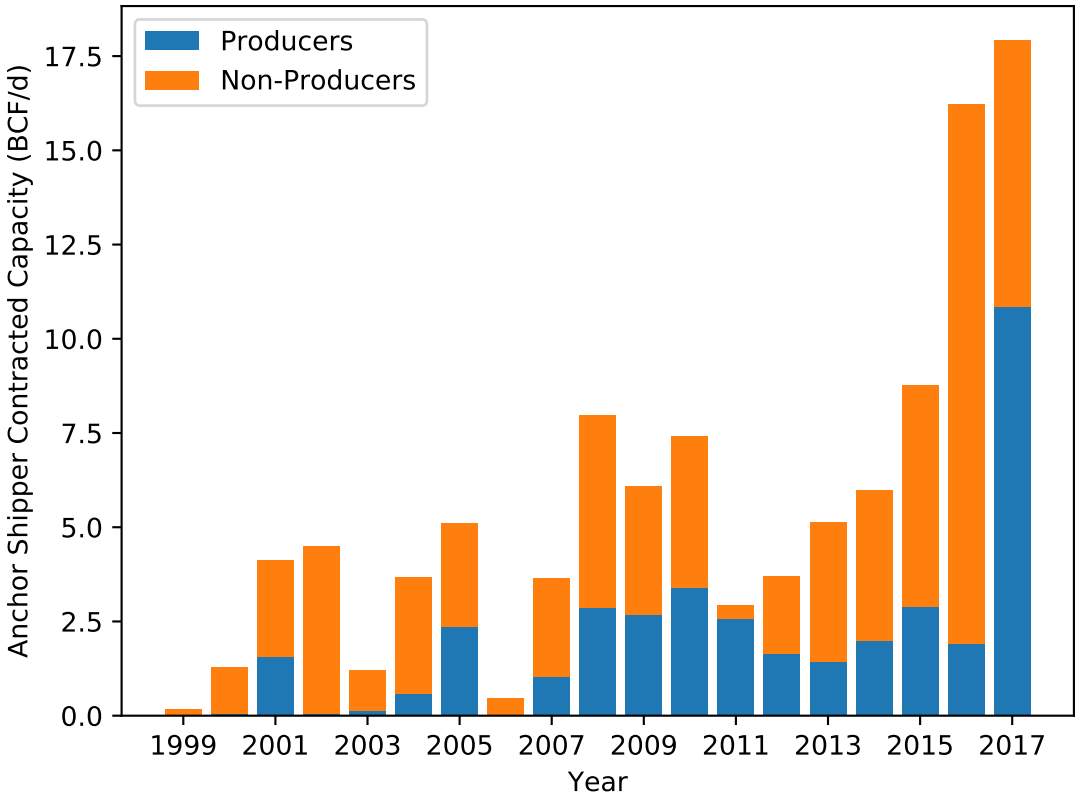
Tests for change in technology are conducted with dimension-reduced data ($p = q = 1$). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Figure 1.1: Daily U.S. Shale Gas Production By Month: January 2000 – December 2018



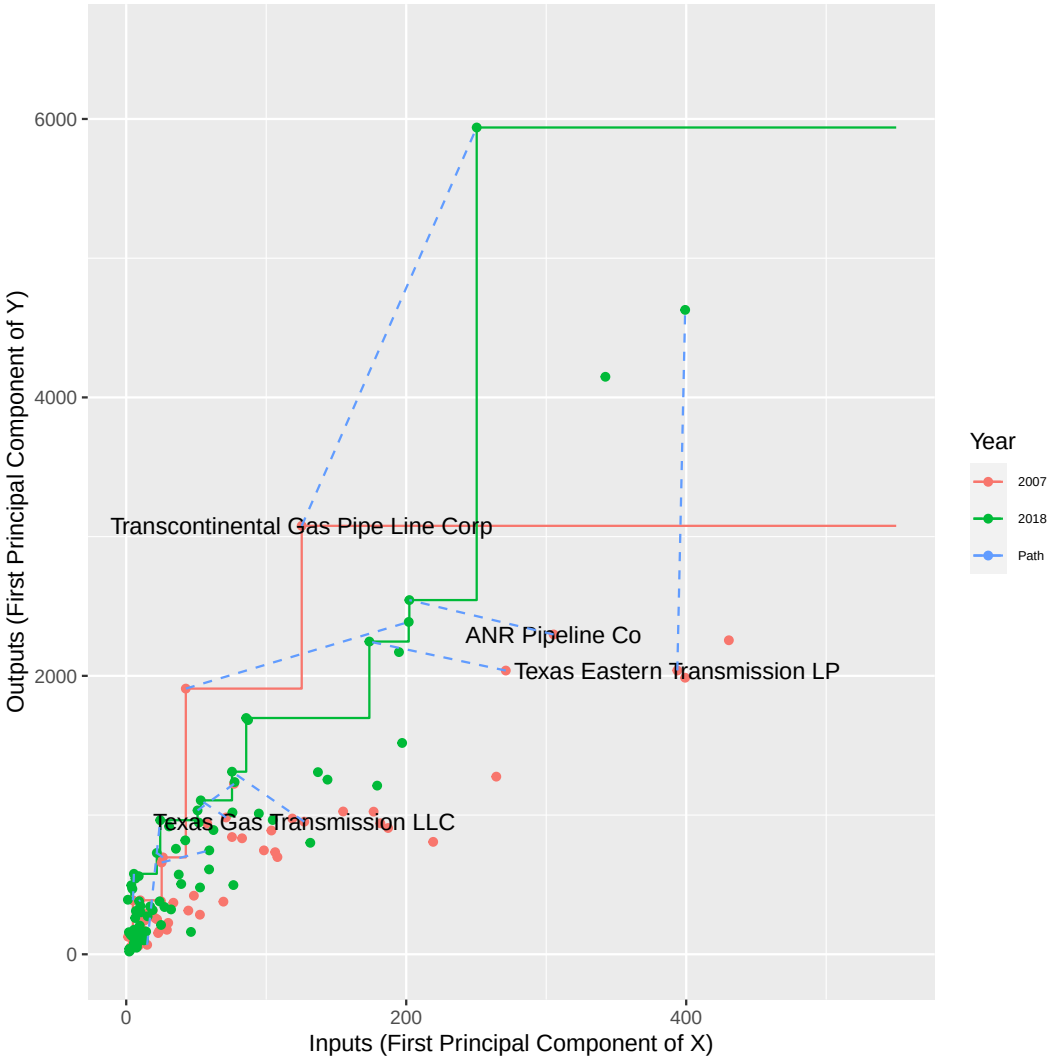
This figure presents data from the U.S. Energy Information Administration on daily U.S. shale gas production by shale bed.

Figure 1.2: Certified Anchor Marketer Capacity: 1999 – 2017



This figure presents data on pipeline capacity, from newly certified pipeline projects, that have been contracted to an anchor shipper. These data were collected from pipeline certificate documents from FERC.

Figure 1.3: FDH: 2007 Versus 2018



Chapter 2

Nonparametric Benchmarking of U.S. Electric Utilities: Changes in Productivity and Efficiency of Electric Utilities From 2001–2019

2.1 Introduction

During 2007–2017, energy consumption in the U.S. declined. Prior to 2007, energy consumption in the U.S. consistently increased every year, corresponding with increases in gross domestic product over the same period. The US Energy Information Administration (March 2021, 2009) indicate that from 1995 to 2007 energy consumption in the U.S. increased from 90,931 trillion British thermal units (BTUs) to 100,893 trillion BTUs, an 11 percent increase. However, during 2007–2009 energy consumption declined by 7 percent to 93,942 trillion BTUs. This was the first multi-year decline in energy consumption in the US. This event itself is unsurprising as it coincides

with the financial crisis during 2007–2008. However, since this decline, the demand for energy has remained flat.

The reasons for the decrease in growth of energy consumption have been widely discussed in previous research. Several studies have suggested that increased efficiency in energy consumption has reduced the demand for electricity. For example, Nadel and Herndon (2014) and Nadel and Young (2014), suggest that energy efficiency programs, warmer winter weather, and distributed rooftop solar generation have been strong contributors to the decline in the production and consumption of energy in the US. In addition, Davis (2017) explains increased usage of more efficient lightning in households and the rapid emergence of light emitting diode or LED bulbs has created more end-user efficiency, reducing household energy consumption. Other explanations for end-user efficiency include more efficient building construction and the increased adoption of more efficient appliances, as Nadel et al. (2015) suggests.

A major event concurrent to the decline in energy consumption was the expansion of natural gas supply due to technology shocks from developments in fracking and directional drilling. In 2009, the average natural gas wellhead price declined by about 50 percent and has remained within \$3–\$5 per million BTUs (MMBtus). This changed the relative prices of coal and natural gas, and thus electricity producers began to substitute away from coal to natural gas as a major fuel source. Coincidentally, the proliferation of renewable portfolio standards, implemented by many state governments and regulators, increased investment in renewable energy generating plants. Both events have coincided with an increase in installed electric generating capacity in the US. According to the US Energy Information Administration (2009, 2020), total utility-scale generating capacity in the U.S. increased from 1,087,791 megawatts (MW) in 2007 to 1,197,917 MW in 2019, a 10 percent increase.

Given the decline in energy consumption during the early 2000s and the expansion of generating capacity, one might expect that the productivity and efficiency of energy producers declined during this period. However, environmental factors affecting the production process of energy producers should be considered, as the feasible production set of producers may differ over time or due to investment in different types of generating capacity. I use nonparametric estimators to measure the technical efficiency and productivity of U.S. electric utilities during 2001–2019. Moreover, I use modern techniques to make inference on changes in technical efficiency over time and

test for properties of the production set. These tests include tests of (i) changes in the production frontier due to environmental factors, (ii) convexity of the production set and (iii) returns to scale of the production set. Previous studies using nonparametric methods to examine electric utility technical efficiency provide only point estimates. Examples include Färe et al. (1989), who examine the effects of environmental regulation on utility efficiency, Sarkis and Cordeiro (2012), who examine the ecological efficiency of utilities, as well as von Geymueller (2009) and Omrani et al. (2015), who discuss the efficiency of transmission operations within utilities. No studies measure the efficiency and productivity of U.S. electric utilities during the simultaneous decrease in energy consumption and increase in generating capacity that occurred after 2007 while using modern nonparametric techniques to make inference.

To my knowledge, the Federal Energy Regulatory Commission (FERC) provide the only publicly available data on U.S. electric utility operations. Moreover, these are the most commonly used data for estimating the efficiency and productivity of electric utilities and power generators. However, these data only include major utilities as defined by FERC resulting in a small sample of utilities being observed each year.[1] A small sample size creates issues when using nonparametric efficiency estimators as they are subject to the curse of dimensionality. It is well-known that the convergence rates of nonparametric efficiency estimators decrease as the dimensionality of the problem (i.e the specified number of inputs and outputs, increases, thereby increasing the order of estimation error).

While Omrani et al. (2015) use dimension reduction techniques to study Iranian electric utilities, they only reduce their dimensions from 16 to 8 total inputs and outputs to estimate the efficiency of 37 utilities in a single year. Thus, their estimates are likely overstating how efficient Iranian electric utilities are. They apply eigensystem decomposition of the correlation matrix which may not always be reliable for dimension reduction. For reasons given in Wilson (2018), it is better to apply eigensystem decomposition of the moment matrix. Given the small number of electric utilities observed in U.S. utility data, this paper utilizes dimension reduction techniques along the lines of Wilson (2018) so four dimensions can be reduced to two dimensions with only a small loss of information.

[1]FERC defines a major utility has having one million megawatt hours of production or more, 100 megawatt hours of annual sales for resale, 500 megawatt hours of annual power exchange delivered, or 500 megawatt hours of annual wheeling for others (deliveries plus losses).

The remainder of this paper is organized as follows. Section 2.2 discusses the statistical model, estimators of efficiency, and tests for properties of the frontier. Section 2.3 discusses the sample used to estimate efficiency of electric utilities. Section 2.4 reports the results and findings of the estimates. Finally, Section 2.5 provides a summary and overview of the conclusions.

2.2 Statistical Model

2.2.1 Modeling Conditional Efficiency

To model electric utility efficiency, I use conditional efficiency measures which were first described by Cazals et al. (2002), and later discussed by Daraio and Simar (2005, 2007a,b) and Daraio et al. (2018). Consider a process that generates random vectors (X, Y, T, Z). Inputs are denoted $X \in \mathbb{R}_+^p$, while $Y \in \mathbb{R}_+^q$ are outputs, $T \in \mathbb{R}^1$ denotes the production year, and $Z \in \mathbb{R}^r$ indicates an r-dimensional vector of continuous environmental factors where, in my application, $r = 1$.[2] To establish notation let the lower case letters (x, y, t, z) indicate particular realizations of the above random vectors. The production set, ignoring (T, Z),

$$\Psi = \{(x, y) \in \mathbb{R}_+^{p+q} \mid x \text{ can produce } y\}, \tag{2.1}$$

is all the pairs of input and output quantities that are feasible given the production technology. However, as described in section 2.2.3, environmental factors may affect the boundary of the production set. Thus, conditioning the production set on (T, Z) might be required, where

$$\Psi^{t,z} = \{(x, y) \in \mathbb{R}_+^{p+q} \mid x \text{ can produce } y \text{ when } (T, Z) = (t, z)\} \tag{2.2}$$

is the conditional production set, or all the pairs of input and output quantities that are feasible given the production technology subject to (t, z). The efficient frontier of (2.2) is defined as

$$\Psi^{t,z\partial} = \{(x, y) \in \Psi^{t,z} \mid (\gamma^{-1}x, \gamma y) \notin \Psi^t \, \forall \, \gamma > 1\}. \tag{2.3}$$

[2]Environmental factors are neither inputs nor outputs and are not a choice variable under the control of the electric utility, but still can influence the production process. Likewise the vector of production years, T, may influence the production process and is treated as a discrete environmental factor.

Several assumptions about $\Psi^{t,z}$ are made. These assumptions follow those of Shephard (1970) and are similar to standard assumptions made in production theory. Moreover, these assumptions are required to use results established in Kneip et al. (2015), described in more detail below. I assume $\Psi^{t,z}$ is closed, $(x, y) \notin \Psi^{t,z}$ if $x = 0, y \geq 0, y \neq 0$ (all production requires the use of some inputs), and $\forall (x, y) \in \Psi^{t,z}, (i)\, \tilde{x} \geq x \implies (\tilde{x}, y) \in \Psi^{t,z}$, and $(ii)\, \tilde{y} \leq y \implies (x, \tilde{y}) \in \Psi^{t,z}$ (the production set allows for strong disposability). Assuming that $\Psi^{t,z}$ is closed implies that the set of efficient points, or the frontier, is contained in the production set (i.e. $\Psi^{t,z\partial} \in \Psi^{t,z}$). The second assumption ensures there is "no free lunch", or the production of a nonzero output vector requires the expenditure of a nonzero input vector. Finally, strong disposability imposes weak monotonicity of the frontier.

In a given year t, I observe data on inputs, outputs, and other variables that could potentially be environmental factors at the level of the utility. An electric utility's technical efficiency is measured by their distance from their observed input and output quantities to $\Psi^{t,z\partial}$. The input efficiency measure conditional on (T, Z), along the lines of Daraio and Simar (2007b),

$$\theta(x, y \mid t, z) := \inf\{\theta \mid (\theta x, y) \in \Psi^{t,z}\}, \tag{2.4}$$

indicates by how much firms must proportionally scale back inputs while producing the same level of output to operate on $\Psi^{t,z\partial}$. Likewise, extending (2.4) to output efficiency gives,

$$\lambda(x, y \mid t, z) := \sup\{\lambda \mid (x, \lambda y) \in \Psi^{t,z}\}, \tag{2.5}$$

which indicates by how much firms must proportionally expand output given the same level of inputs to operate on $\Psi^{t,z\partial}$. As shown in Wilson (2011, Figure 6.1) a firm operating off of the efficient frontier can have differing input and output measures of efficiency. An alternative measure is the hyperbolic graph measure of efficiency conditional on (T, Z), defined as

$$\gamma(x, y \mid t, z) := \inf\{\gamma > 0 \mid (\gamma x, \gamma^{-1} y) \in \Psi^{t,z}\}. \tag{2.6}$$

The hyperbolic graph measure indicates the amount a firm must simultaneously scale down inputs and increase output by the same factor, γ, to operate on $\Psi^{t,z\partial}$. The benefit of (2.6) is that a firm's distance to the frontier is measured along a hyperbolic path and this avoids the issue of a firm having

differing measures under (2.4) and (2.5).

2.2.2 Estimation Methods

Of course $\Psi^{t,z}$ is not observed and must be estimated with a random sample of observed quantities of inputs, outputs, environmental factors and production years, $S_n = \{(X_i, Y_i, T_i, Z_i)\}_{i=1}^n$, where $X_i \in \mathbb{R}_+^p$, $Y_i \in \mathbb{R}_+^q$, $T_i \in \mathbb{R}^1$, $Z_i \in \mathbb{R}^r$, and i indexes the utility. Different nonparametric efficiency estimators require different assumptions about the properties of the production set. To begin, Deprins et al. (1984) propose estimating unconditional efficiency with the free disposal hull (FDH) of the observed input and output vectors in S_n, or

$$\widehat{\Psi}_{FDH,n} := \bigcup_{(X_i,Y_i)\in S_n} \{(x,y) \in \mathbb{R}^{p+q} \mid y \leq Y_i, x \geq X_i\}. \tag{2.7}$$

If environmental factors should be considered, Daraio and Simar (2005) propose conditioning FDH estimates with respect to continuous environmental factors using a bandwidth, h. In my case the FDH estimator conditional on (T, Z),

$$\widehat{\Psi}^{t,z}_{FDH,n,h} := \bigcup_{(X_i,Y_i)\in S_n} \{(x,y) \in \mathbb{R}^{p+q} \mid y \leq Y_i, x \geq X_i, T = t, Z \in [z-h, z+h]\}, \tag{2.8}$$

allows the frontier to to vary depending on T or Z.

While the FDH estimators do not impose convexity on the production set, data envelopment analysis (DEA) estimators of efficiency can be used if the production set is convex. Farrell (1957) propose DEA estimators in the unconditional case. The unconditional variable returns to scale (VRS) DEA estimator is given by the convex hull of (2.7),

$$\widehat{\Psi}_{VRS,n} := \{(x,y) \in \mathbb{R}^{p+q} \mid y \leq \mathbf{Y}\boldsymbol{\omega}, x \leq \mathbf{X}\boldsymbol{\omega}, \mathbf{i}_\mathbf{n}'\boldsymbol{\omega} = 1, \boldsymbol{\omega} \in \mathbb{R}_+^n\}, \tag{2.9}$$

where $\mathbf{X} = (X_1, ..., X_n)$ and $\mathbf{Y} = (Y_1, ..., Y_n)$ are $(p \times n)$ and $(q \times n)$ input and output matrices, in year t, $\mathbf{i}_\mathbf{n}'$ is an $(n \times 1)$ vector of ones, and $\boldsymbol{\omega}$ is a $(n \times 1)$ vector of weights. Daraio and Simar (2007b) extend the VRS-DEA estimator to the conditional case by conditioning (2.9) on (T, Z) to

obtain the conditional VRS-DEA estimator,

$$\widehat{\Psi}^{t,z}_{VRS,n,h} := \{(x,y) \in \mathbb{R}^{p+q} \mid y \leq \mathbf{Y}\boldsymbol{\omega}, x \geq \mathbf{X}\boldsymbol{\omega}, \mathbf{i}'_{\mathbf{n}}\boldsymbol{\omega} = 1, \boldsymbol{\omega} \in \mathbb{R}^n_+, T = t, Z \in [z-h, z+h]\}, \quad (2.10)$$

which is locally convex, where the localization is controlled by h. Both the conditional and unconditional VRS-DEA estimator do not make any assumptions about return to scale of the production set. If constant returns to scale (CRS) is assumed, the unconditional CRS-DEA estimator estimates the conical hull of (2.9), denoted $\widehat{\Psi}_{CRS,n}$, which can be obtained by removing the constraint, $\mathbf{i}'_{\mathbf{n}}\boldsymbol{\omega} = 1$, in (2.9). Similarly, the conditional CRS-DEA estimator, $\widehat{\Psi}^{t,z}_{CRS,n,h}$, is obtained by removing the same constraint from (2.10).

Input and output-oriented DEA estimates can be computed using linear programming methods, while hyperbolic-oriented estimates are non-linear programs that can be solved using numerical methods proposed by Wilson (2011). In addition, the FDH estimates can be solved using numerical methods as well. See Wilson (2011) for technical details. Furthermore, all estimates and tests described in this paper are performed using the *FEAR* software package developed by Wilson (2008).

2.2.3 Testing Hypothesis

Prior to estimating technical efficiency, I conduct several tests to guide my estimation procedure. First I test whether environmental factors, (T, Z), affect the frontier. The issues of environmental factors with nonparametric efficiency estimators have been discussed by Simar and Wilson (2011a,b) and Daraio et al. (2018). To begin, environmental factors, (T, Z), can impact the distribution of efficiency estimates, the efficient frontier, or both. If (T, Z) only impacts the distribution of efficiency estimates, then all utilities face the same attainable frontier and FDH and DEA efficiency estimates do not need to be conditioned on environmental factors. In this case, the separability condition, as described in Simar and Wilson (2011a,b) and Daraio et al. (2018), is said to hold with respect to the environmental factors. However, if (T, Z) affects the frontier, unconditional FDH and DEA efficiency estimates are meaningless. In this case, unconditional efficiency estimates will benchmark some firms relative to a frontier that is not attainable due to the environmental factors those firms face. After I test for separability, I then test for convexity versus non-convexity of the production set, then CRS versus VRS. These tests reduce the chance of selecting an estimator

that relies on assumptions that are inconsistent with the true structure of the production set.

To test for separability with respect to (T, Z), I use methods from Daraio et al. (2018) to test

$$H_0 : \Psi^{t,z} = \Psi \, \forall (t, z) \in (T, Z) \tag{2.11}$$

versus

$$H_1 : \Psi^{t,z} \neq \Psi \text{ for some } (t, z) \in (T, Z). \tag{2.12}$$

Moreover, limiting the sample to single year $t' \in T$ and testing

$$H_0 : \Psi^{t',z} = \Psi \, \forall z \in Z \tag{2.13}$$

versus

$$H_1 : \Psi^{t',z} \neq \Psi \text{ for some } z \in Z, \tag{2.14}$$

allows me to disentangle the effects of T and Z on the frontier, and assess if Z alone impacts the frontier. The test for separability involves randomly shuffling observations in the sample, of size n, and splitting the sample into two groups. Using the FDH estimator, the procedure initially estimates unconditional efficiency, ignoring (T, Z) (restricting the frontier to be the same for all (t, z)), with the first group S_{n_u} (of size n_u), and then estimates efficiency conditional on (T, Z), with the second group S_{n_c} (of size n_c), where $n_u + n_c = n$ and $S_{n_u} \cap S_{n_c} = \emptyset$. The conditional efficiency estimates allow the frontier to change with different $t \in T$ and across different h-neighborhoods of z, where optimal bandwidths are found using the least squares cross validation (LSCV) procedure in accordance with Li et al. (2013). Since this test requires randomly splitting the sample into two sub-samples, the uncertainty from a single sample-split is reduced by gathering information from multiple sample-splits through a bootstrap procedure proposed by Simar and Wilson (2020). This test uses information from 1,000 bootstrap replications and 100 sample-splits to calculate test statistic, proposed by Daraio et al. (2018), which estimates the difference between the mean conditional and unconditional efficiency estimates.[3] If these test statistics are statistically significant, separability is rejected. In addition, the p-values from the 100 sample splits are collected and their

[3]The number of utilities in my data observed each year T ranges from 85 to 113. Simulation results from Simar and Wilson (2020) suggest 100 sample splits to estimate the test statistic and 1,000 bootstrap replications to estimate the distribution of test statistics performs well with sample sizes as small as 100 observations.

distribution is compared to the uniform distribution on $[0,1]$ using the Kolmogorv-Smirnov (KS) test.[4] The bootstrap procedure discussed by Simar and Wilson (2020) is used to estimate the distribution of the KS statistic since the 100 p-values used to construct the KS statistic are not independent. If uniformity of the p-values is rejected, then separability can also be rejected. For technical details see Daraio et al. (2018) and Simar and Wilson (2020).

The convergence rate of the conditional FDH estimator is slower than that of the unconditional FDH estimator, due to localizing the estimates of the frontier. Therefore, the sample-split sizes are selected such that $n_u^\kappa = n_c^{\kappa/(r\kappa+1)}$ subject to $n_u + n_c = n$, where $\kappa = \frac{1}{p+q}$ is the convergence rate of the unconditional FDH estimator, as shown by Park et al. (2000), and $1/(r\kappa+1)$ is the convergence rate of the conditional FDH estimator, as shown by Daraio et al. (2018). Moreover, as noted in Section 2.2.1, $r = 1$ in my application. This method results in an uneven sample-split and is efficient as more information is provided to the conditional estimator. This sample-splitting technique is proposed by Kneip et al. (2016).

Tests of convexity versus non-convexity and CRS versus VRS are also done to infer whether the appropriate estimated production set should be estimated with the FDH, VRS-DEA, or CRS-DEA estimators (with conditioning if needed). I use results from Kneip et al. (2015, 2016), and Simar and Wilson (2020) in order to do this. Kneip et al. (2015) provide central limit theorems for FDH and DEA efficiency estimates, and Kneip et al. (2016) use these results to develop specific tests for convexity, returns to scale, and differences in mean efficiency across groups of producers. As with the tests for separability, these tests rely on a random sample-split where the split sizes are selected with an uneven sample-split along the lines of Kneip et al. (2016). In addition, multiple sample-splits are used along with the bootstrap method of Simar and Wilson (2020) to reduce uncertainty. Furthermore, I follow the testing strategy of Apon et al. (2015, Figure 1) and test convexity before testing for returns to scale.[5] To summarize, I first conduct the test for convexity by randomly splitting the sample and obtaining conditional VRS-DEA estimates of efficiency with one sub-sample and conditional FDH estimates with the other, if separability is rejected. If separability holds, the unconditional estimators can be used. The test uses information from 1,000 bootstrap replications and 100 sample-splits to determine if the difference between the mean conditional VRS-

[4] As discussed by Simar and Wilson (2020), the uniform distrbution on $[0,1]$ is the distribution of p-values under the null hypothesis due to the probability integral transform.

[5] Apon et al. (2015) only test returns to scale if convexity is rejected, however Kneip et al. (2021) does show that CRS technology can exhibit a non-convex production set.

DEA and conditional FDH estimates are statistically significant. If so, then convexity is rejected. If the test fails to reject convexity, returns to scale are tested. The test for returns to scale is similar to the test for convexity, but the conditional FDH estimator is replaced with the conditional CRS-DEA estimator. As with the test for separability, the KS tests are conducted for both the test for convexity and returns to scale. For technical details see Apon et al. (2015, Figure 1), Kneip et al. (2015), Kneip et al. (2016), and Simar and Wilson (2020).

2.3 Data

2.3.1 Data on Electric Utilities and Variable Specification

The sample is an unbalanced panel of inputs and outputs by electric utility and production year collected from the Form 1 Financial Data on Major Electric Utilities (Form 1) produced by FERC. The Form 1 data consist of financial and engineering data that electric utilities report to FERC, and are publicly available. These are the most commonly used data for measuring productivity and efficiency of utilities and power plants in the US, and, to my knowledge, are the only publicly available, comprehensive data on U.S. electric utility operations. The data collected contain information on electric utilities during the years 2001–2019.

Table 2.1 contains a review of previous research examining the technical efficiency of electric utilities or power plants. While many specifications of inputs and outputs have been used to describe electric utility operations, several papers include generating capacity, fuel, and labor as inputs, while net generation of electricity is treated as an output. For example, Färe et al. (1989) specify inputs as the number of workers, heat content of fuel, and installed generating capacity, while a single output is measured in megawatt hours of net generation. Likewise, Sarkis and Cordeiro (2012) and Bernstein (2020) use similar specifications.[6] Other papers examine other facets of electric utility operations (e.g von Geymueller (2009), Growitsch and Hess (2009), and Omrani et al. (2015)) such as transmission operations and cost efficiency, and thus have specified inputs and outputs differently.

[6]It should be noted that Färe et al. (1989) included precipitator costs as an input to examine the effect of environmental regulation on efficiency scores. Likewise Sarkis and Cordeiro (2012) included emissions as outputs to incorporate "bad" outputs into the estimation of the production frontier. Finally, Bernstein (2020) examined efficiency at the level of the power plant and included additional plant level variables, such as hours connected to load, as inputs.

I specify inputs and outputs of utilities along lines similar to these previous studies. Specifically, I include $p = 3$ inputs including heat content of fuel measured in MMBtus, generating capacity measured in MW, and number of employees. In addition, I specify $q = 1$ output consisting of net generation of energy measured in gigawatt hours (GWh). It should be noted that net generation includes electricity generated by the power plants operated by the utilities and is exclusive of energy purchased from other producers. The chosen inputs and outputs allows for the nonparametric estimation of a production function with capital, labor, and fuel as inputs, and energy as the single output.

Figures 2.1 through 2.4 show sample means of the included inputs and outputs over time. Examining the output, net generation, shows that while electric utility output increased over 2001–2007, there was a downward trend in output for the remainder of the sample. More specifically, average net generation declined from 17,843 GWh in 2007 to 15,694 GWh in 2019, a 12 percent decline. These data are consistent with findings from other studies, such as Nadel and Herndon (2014), Nadel and Young (2014), and Davis (2017) who discuss the decline in energy demand due to energy efficiency, as well reports from the US Energy Information Administration (March 2021). With regards to inputs, generating capacity trends upward from 2001 through 2013, with a 13 percent increase in capacity to 4.3 GW in 2013 from 3.8 GW in 2007. From 2013 to the end of the sample, the growth in capacity slows down where total installed capacity remains near 4.3 GW at the end of the sample period. Finally, the average number of employees decreases and there is no distinct trend in the average heat content of fuel used by utilities during the sample period.

A benefit of using the Form 1 data is the ability to examine plant-level data within each utility. Power plant details such as type and the fuel type are available for each plant, allowing examination of the resource mix by each utility. Figure 2.5 shows the average share of generating capacity attributed to natural gas fueled plants over time, while Figure 2.6 shows the average share of generating capacity attributed to coal fueled plants over time. These data confirm that during the period of shale gas expansion in the early 2000s, natural gas fueled energy production increased, while coal generation decreased. While the choice to change resource mix can be made by the utility, this is almost always subject to regulation. For example, regulatory requirements such as renewable portfolio standards may inhibit some utilities from expanding natural gas fueled generation, in favor of renewable generation. In addition, local utility regulation may dictate whether power plant

investment gets approved or not. Thus, I use the share of natural gas generation used by each utility, denoted as Z, as a continuous environmental variable.

2.3.2 Dimension Reduction Method

As noted in Section 2.1, the curse of dimensionality is a well-known issue in nonparametric benchmarking, since‘ the convergence rates of FDH and DEA estimators decrease as the number of inputs and outputs, $p+q$, increases. The rate of convergence for the VRS-DEA and FDH estimators are established by Kneip et al. (1998), Park et al. (2000), and Daouia et al. (2017). With $p+q=4$ dimensions the VRS-DEA estimator converges at rate $n^{\frac{2}{5}}$, while the FDH estimator converges at a slower rate $n^{\frac{1}{4}}$.

To mitigate the slow rate of convergence, I use eigensystem methods along the lines of Wilson (2018) to reduce the dimensionality of the problem. More specifically, I find the $(n \times 1)$ first principal component, X^*, of the moment matrix, $\mathbf{X'X}$, of my observed $(n \times p)$ output matrix, $\mathbf{X}$. X^* is the matrix product of the observed output matrix and the eigenvector corresponding to the largest eigenvalue of $\mathbf{X'X}$. The ratio of the largest eigenvalue of $\mathbf{X'X}$ to the sum of all of its eigenvalues measures the amount of independent linear information X^* contains from the p columns of $\mathbf{X}$. With the sample of inputs from all production years, this ratio is 0.971. Results from Wilson (2018) suggest that reducing the dimensions by replacing $\mathbf{X}$ with X^* is reasonable since little information is lost. See Wilson (2018) for technical details.[7]

With dimension reduction the convergence rate for the VRS-DEA estimator improves from $n^{\frac{2}{5}}$ to $n^{\frac{2}{3}}$, and the convergence rate for the FDH estimator improves from $n^{\frac{1}{4}}$ to $n^{\frac{1}{2}}$. In this case, the VRS-DEA estimator achieves a convergence rate faster than the standard parametric rate, while the FDH estimator converges at the parametric rate. Because of this, I compute all estimates using dimension-reduced data. Table 2.2 provides summary statistics for the sample of specified inputs and outputs as well as the first principal components of the input matrix.[8]

[7] The estimated first principal component of the input matrix is derived using the pooled sample of data from 2001–2019. This is done to maximize the amount of data used to estimate the principal components. If different principal components were estimated for each year in the sample, the ratios described above would range from 0.964 to 0.991, suggesting dimension reduction is still feasible with a sample limited to each year.

[8] Trivially, the first principal component of the output vector is the vector of net generation.

2.4 Estimation and Results

I first test for separability of the production set with respect to (T, Z). Table 2.3 reports results of the tests for separability with respect to (T, Z) using the FDH estimator with 1,000 bootstrap replications and 100 sample-splits.[9] For all tests, I estimate efficiency using the input, output and hyperbolic orientations as robustness checks. The first row, denoted Test 1, of Table 2.3 reports the test statistics corresponding to the differences between the unconditional and conditional FDH estimates. The second row, denoted Test 2, shows the corresponding KS test. Of the six tests, two reject the null at the 1 percent level, one rejects at the 5 percent level, and one rejects at the 10 percent level. This is evidence that separability of the frontier does not always hold with respect to (T, Z). As discussed above, limiting the sample to individual years in T and testing for separability with respect to Z allows me to separate the effects of T and Z on the frontier. Table 2.4 presents the test statistics corresponding to the differences between the FDH estimates conditional on Z and unconditional estimates, for each year in the sample. Table 2.5 reports the KS test results corresponding to the p-values from the sample-splitting procedure used to create the test statistics in Table 2.4. Both tables show numerous rejections of rejections of the null. This is evidence that Z impacts the frontier faced by electric utilities. Previous research does not consider the possibility of gas share of generation impacting the frontier faced by utilities. These results suggest that the attainable frontier of utilities varies with $z \in Z$. Thus, all efficiency estimates will be conditioned on Z as well as T.

Table 2.6 reports the results of the tests of convexity versus non-convexity using the test statistics from Kneip et al. (2016) and Table 2.7 reports the results of the corresponding KS tests. Both sets of tests show similar rates of rejections to the tests for separability, where the rates are high enough to make one cautious in assuming the null hypothesis. In addition, Table 2.8 and Table 2.9 report the results for the returns to scale tests using the test statistics from from Kneip et al. (2016) and the corresponding KS tests. Like the tests for convexity, the returns to scale tests have rejection rates sufficient enough to reject the null. In this case, I find enough evidence to reject the convexity and CRS of the production set. Thus, for all remaining tests I use the conditional FDH estimator.

[9] The FDH estimator is used as it does not rely on convexity of $\Psi^{t,z}$, and is the safest estimator to use for testing separability without testing the convexity of $\Psi^{t,z}$ prior.

Figure 2.7 plots the point estimates of mean efficiency using the FDH estimator, conditional on T and Z, over time. As with the tests for separability, convexity, and returns to scale, efficiency is estimated in the input, output, and hyperbolic orientation as robustness checks.[10] Under all orientations, the point estimates show technical efficiency trending upward from 2001 through 2008. This largely coincides with the increase in net generation from 2001 through 2007 as shown in Figure 2.1. However, the trend in efficiency stops increasing after 2008, with a temporary drop in estimated efficiency from 2011 through 2015 and a decline in efficiency from 2017 to the end of the sample period.

Table 2.10 presents the results for tests in changes in conditional mean efficiency using the FDH estimator. In this test, I use the estimator for differences in mean efficiency across groups of producers, allowing the frontier to vary by group, as described by Kneip et al. (2016). Each row corresponds to an interval of time for which changes in mean efficiency are tested, and the utilities in different years are treated as different groups of producers. The test statistic is asymptotically distributed standard normal, and is large and positive (negative) when there are increases (decreases) in mean efficiency.[11] In year-over-year terms, mean efficiency changes occasionally throughout the sample. There is an increase in mean efficiency during 2001–2002 under all orientations, with a subsequent decrease under the hyperbolic orientation over 2002–2003. During 2006–2007, technical efficiency increases again under all orientations, but would remain flat from 2008 through 2010, then decrease in the input and hyperbolic orientations from 2010 through 2011. Efficiency would then increase over 2015–2016 in the input orientation and over 2016–2017 in the output orientation, but then decline from 2017 through the end of the sample period under all orientations.

The bottom portion of Table 2.10 contains results for tests of changes in mean efficiency in net terms for some year t to the end of the sample period, 2019. The results indicate that over 2001–2019, 2003–2019, and 2006–2019 there were statistically significant increases in technical efficiency, under the input and hyperbolic orientations. As shown in Figure 2.1, these increases

[10] While (2.4) and (2.6) by construction are weakly less than one, (2.5) is weakly greater than one. For ease of comparison, Table 2.10 and Figure 2.7 present the output-oriented measures as the Farrell metric or reciprocal of (2.5). Thus, efficiency measures are in $(0, 1]$ with efficiency measures of one indicating presence on the frontier.

[11] Note that the tests of differences in mean efficiency, developed by Kneip et al. (2016), require independence between the groups of producers. When testing for changes in mean efficiency over time, utilities are often observed at multiple points in time resulting in a covariance issue. However, Wilson and O'Loughlin (2021) perform similar tests for U.S. municipalities, and they argue that any covariance is likely positive due to inertia. In my case, inertia likely plays a role as well where a utility that performs poorly (or well) in one year is likely to continue performing poorly (or well) in a subsequent year. Consequently, ignoring positive covariance makes my tests for changes in mean efficiency conservative as I bias towards failing to reject the null.

coincide with the increase in utility net generation observed early in the sample period. However, over 2008–2019, 2009–2019, and 2010–2019 there are estimated declines in technical efficiency under various orientations. These coincide with the period of decline in net generation, that occurs after 2007. To summarize, the results indicate that technical efficiency did increase, in net terms, from the beginning of the sample period to the end, where these increases correspond with increases in utility output that occurred prior to 2007. However, the subsequent decline in utility output coincides with a decrease in technical efficiency of utilities observed after 2007.

Finally, because I have a single output, net generation, and use dimension reduction to obtain a single input, X^*, I can directly measure the mean productivity of utilities in a given production year t as

$$\widehat{Productivity}_t = n_t^{-1} \sum_{i}^{n_t} \frac{Net\,Generation_{i,t}}{X^*_{i,t}}, \tag{2.15}$$

where i indexes the utilities and n_t is the number of utilities observed in production year t. Moreover, the standard Lindeberg-Feller Central Limit Theorem can be used to make inference on changes in productivity over time.[12] Table 2.11 reports the test statistics of differences in the productivity estimates. The results indicate declines in productivity in year-over-year terms from 2007–2014. This corresponds with the period of decline in net generation as shown in Figure 2.1. Although there is an increase in productivity from 2017–2018, coinciding with a large increase in net generation over the same time period, there is strong evidence for declines in productivity, in net terms. The results at the bottom portion of Table 2.11 indicate that under all tests for net change in productivity the test statistics are negative and large in magnitude, indicating productivity declined over the sample period.

2.5 Summary and Conclusion

In the analysis outlined above I use nonparametric methods and dimension reduction techniques to measure the technical efficiency and productivity of electric utilities from 2001–2019. Moreover, I apply modern nonparametric tests to determine if a utility's reliance on natural gas generation changed the production set the utility faced. This is done to determine if efficiency estimates should be conditional on the amount of gas generation a utility has. The results indicate that

[12] Similar covariance issues as noted in footnote 11 arise when examining productivity over time. However, similar reasoning applies here where any covariance is likely positive making the tests conservative.

separability with respect to production year and reliance on natural gas generation does not hold, thus the production set for utilities changes with respect to these variables. This is something previous work examining utility efficiency using nonparametric methods has not considered. Moreover, I find little evidence against convexity of the production set.

While tests for changes in mean efficiency suggest that the technical efficiency of utilities did increase when comparing early production years in the sample to the end, the increases in technical efficiency cease after the 2006. After 2008, during the initial decline of net generation and electricity consumption, declines in the efficiency of utilities are observed. In addition, the productivity of utilities declined each year from 2007-2014. The results provide evidence that utilities became less efficient and less productive after energy consumption declined after 2007.

Table 2.1: Summary of Literature

Author(s)	Inputs	Outputs	Observations	Techniques	Data
Färe et al. (1989)	(1) Labor (2) Fuel (MMBtu) (3) Capital (MW) (3) Precipitator Cost	(1) Net Generation (MWh)	23 U.S. Electric Utilities (1969 & 1975)	DEA	FERC Form 1
Growitsch and Hess (2009)	(1) TOTEX	(1) Energy (2) Customers	109 U.S. Electric Utilities (1994 – 2005)	SFA	FERC Form 1
von Geymueller (2009)	(1) Transmission Materials and Supplies ($) (2) Transmission Salaries and Wages ($) (3) Transmission Line Length (Miles) (4) Total Transformer Capacity (MW)	(1) Transmission of Electricity (MWh)	50 U.S. Electric Utilities (2000 – 2006)	DEA & Dynamic-DEA	FERC Form 1
Sarkis and Cordeiro (2012)	(1) Total Heat Input (MMBtu) ($) (2) Boiler Capacity (MMBtu) ($) (3) Generator Capacity (MW)	(1) Net Generation (MWh) (2) NO_x (3) SO_2 (4) CO2 Emissions	437 U.S. Electric Plants (1998)	VRS-DEA	E.P.A. Emissions and Generation Resource Database
Omrani et al. (2015)	(1) Transformer Capacity (MVA) (2) Number of Transformers (3) Terrestrial Network Length (Km) (4) Ariel Network Length (Km) (5) Number of Employees (6) Area (km^2)	(1) Energy Delivery (MWh) (2) Energy Consumption by Others (3) Industrial Energy Consumption (4) Household Energy Consumption (5) Industrial Customers (6) Household Customers (7) Other Customers (8) Number of Street Lights	37 Iranian Utilities (2010)	CRS-DEA & PCA (Correlation Matrix) 8 Dimensions After PCA	Iranian Power Generation Publications
Bernstein (2020)	(1)Installed Capacity (MWh) (2) Number of Employees (3) Fuel (Barrels of Oil & Cubic Feet of Gas) (4) Net Peak Demand (5) Plant Hours Connected to Load	(1) Net Generation (KWh)	3,553 Power Plants (1994-2016)	SFA	FERC Form 1

Table 2.2: Summary Statistics: Form 1 Data 2001–2019

	Min	1st Q	Median	Mean	3rd Q	Max
Net Generation [GWh]	0.720	3,029.230	8,784.700	16,420.660	20,571.300	122,266.350
Total Capacity [100 MWs]	0.264	9.454	24.548	39.147	50.476	281.714
Labor [100 Employees]	0.010	1.650	3.770	6.994	8.572	48.970
Fuel [K MMBtu]	0.000	0.076	0.313	6.075	2.357	333.710
X^*	0.304	9.633	24.862	40.219	52.593	280.231
Z	0.000	0.102	0.301	0.337	0.504	1.000

This table shows summary statistics for the sample collected form Form 1 data. The principal component used in all analysis is denoted as X^*.

Table 2.3: Separability Test With Respect to (T, Z) (FDH Estimator)

	Input–Orientation		Output–Orientation		Hyperbolic	
	Statistic	p-value	Statistic	p-value	Statistic	p-value
Test 1	2.410	0.000***	0.407	0.636	2.679	0.002***
Test 2	0.718	0.073*	0.348	0.517	0.725	0.023**

The test statistic for test 1, described in Daraio et al. (2018), is computed using the FDH estimator with dimension-reduced data, such that $p = q = 1$. The test 2 statistic is from the KS test as described in Simar and Wilson (2020). The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 2001–2019. The p-values are developed using 100 sample-splits and 1,000 bootstrap replications using techniques described in Simar and Wilson (2020). Sample-splits are done unevenly to apply more data to the estimator with the slower rate of convergence, as described by Kneip et al. (2016). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table 2.4: Separability Test With Respect to Z (FDH Estimator)

	Input–Orientation		Output–Orientation		Hyperbolic	
Period	Statistic	p-value	Statistic	p-value	Statistic	p-value
2001	−0.584	0.995	−0.273	0.552	0.242	0.737
2002	0.265	0.871	−0.892	0.931	0.047	0.843
2003	0.396	0.788	−0.779	0.879	−0.536	0.941
2004	0.837	0.741	−0.852	0.940	0.973	0.467
2005	1.987	0.379	−0.739	0.884	1.245	0.464
2006	1.054	0.850	−0.787	0.921	0.303	0.849
2007	1.876	0.658	−1.011	0.956	2.248	0.487
2008	3.702	0.034**	−0.416	0.859	3.374	0.108
2009	3.721	0.061*	0.071	0.801	3.727	0.048**
2010	4.636	0.024**	−0.294	0.877	3.882	0.397
2011	1.573	0.672	−0.589	0.906	0.626	0.687
2012	1.180	0.873	−1.237	0.992	0.545	0.847
2013	3.687	0.009***	3.075	0.087*	3.181	0.260
2014	3.830	0.016**	2.843	0.107	3.231	0.145
2015	2.851	0.466	−0.233	0.816	2.005	0.606
2016	3.578	0.019**	2.637	0.126	3.181	0.040**
2017	3.329	0.091*	2.908	0.095*	3.035	0.067*
2018	3.331	0.046**	2.544	0.345	3.106	0.095*
2019	2.952	0.242	1.587	0.698	2.629	0.376

The test statistic, described in Daraio et al. (2018), is computed using the FDH estimator with dimension-reduced data, such that $p = q = 1$. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 2001–2019. The p-values are developed using 100 sample-splits and 1,000 bootstrap replications using techniques described in Simar and Wilson (2020). Sample-splits are done unevenly to apply more data to the estimator with the slower rate of convergence, as described by Kneip et al. (2016). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table 2.5: Separability Test (KS Test) With Respect to Z (FDH Estimator)

	Input–Orientation		Output–Orientation		Hyperbolic	
Period	Statistic	p-value	Statistic	p-value	Statistic	p-value
2001	0.367	0.695	0.590	0.296	0.291	0.693
2002	0.310	0.936	0.631	0.230	0.364	0.631
2003	0.443	0.541	0.619	0.234	0.467	0.275
2004	0.387	0.758	0.583	0.379	0.415	0.524
2005	0.626	0.402	0.607	0.215	0.424	0.551
2006	0.269	0.982	0.583	0.417	0.278	0.856
2007	0.512	0.736	0.786	0.073*	0.665	0.534
2008	0.922	0.050**	0.776	0.158	0.904	0.046**
2009	0.935	0.057*	0.786	0.224	0.928	0.058*
2010	0.965	0.070*	0.780	0.383	0.896	0.343
2011	0.454	0.849	0.667	0.476	0.513	0.552
2012	0.413	0.873	0.752	0.209	0.467	0.789
2013	0.902	0.046**	0.798	0.241	0.741	0.636
2014	0.925	0.038**	0.841	0.094*	0.867	0.201
2015	0.769	0.561	0.533	0.849	0.601	0.636
2016	0.916	0.042**	0.730	0.258	0.856	0.049**
2017	0.904	0.067*	0.766	0.249	0.786	0.150
2018	0.858	0.098*	0.690	0.536	0.804	0.124
2019	0.842	0.171	0.499	0.932	0.744	0.346

The test statistic is from the KS test as described in Simar and Wilson (2020). In this case, I use the KS test to test for uniformity of the p-values estimated from the multiple sample-splits used in the tests presented in Table 2.4. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 2001–2019. The p-values are developed using 100 sample-splits and 1,000 bootstrap replications using techniques described in Simar and Wilson (2020). Sample-splits are done unevenly to apply more data to the estimator with the slower rate of convergence, as described by Kneip et al. (2016). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table 2.6: Convexity Test Conditional on Z

	Input–Orientation		Output–Orientation		Hyperbolic	
Year	$\widehat{T}_{KSW}$	p-Value	$\widehat{T}_{KSW}$	p-Value	$\widehat{T}_{KSW}$	p-Value
2001	0.182	0.495	0.227	0.696	−0.169	0.873
2002	0.210	0.042**	−0.094	0.923	−0.052	0.348
2003	−0.293	0.200	−0.259	0.959	0.109	0.183
2004	−0.195	0.878	0.998	0.085*	−0.257	0.948
2005	0.189	0.247	0.161	0.817	−0.135	0.725
2006	−0.043	0.280	0.489	0.261	−0.310	0.856
2007	−1.798	0.596	0.792	0.010***	−1.816	0.831
2008	−3.257	1.000	0.346	0.387	−3.077	0.998
2009	−1.427	0.545	0.313	0.349	−0.682	0.591
2010	−1.166	0.177	0.370	0.287	−0.252	0.228
2011	0.269	0.265	0.173	0.545	0.026	0.474
2012	−2.007	0.719	0.780	0.008***	−1.164	0.636
2013	−3.401	0.973	−0.951	0.611	−1.580	0.332
2014	−0.408	0.343	−0.488	0.719	−0.926	0.641
2015	−1.988	0.775	0.135	0.602	−1.110	0.603
2016	−1.393	0.367	0.960	0.009***	−0.242	0.063*
2017	−0.487	0.018**	0.397	0.244	−0.890	0.582
2018	−0.564	0.883	0.209	0.560	−0.198	0.805
2019	−1.422	0.555	0.275	0.397	−0.810	0.526

The test statistic, described in Kneip et al. (2016), is computed using dimension-reduced data, such that $p = q = 1$. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 2001–2019. The p-values are developed using 100 sample-splits and 1,000 bootstrap replications using techniques described in Simar and Wilson (2020). Sample-splits are done unevenly to apply more data to the estimator with the slower rate of convergence, as described by Kneip et al. (2016). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table 2.7: Convexity Test (KS Test) Conditional on Z

	Input–Orientation		Output–Orientation		Hyperbolic	
Year	$\widehat{T}_{KSW}$	p-Value	$\widehat{T}_{KSW}$	p-Value	$\widehat{T}_{KSW}$	p-Value
2001	0.322	0.810	0.622	0.343	0.449	0.122
2002	0.417	0.095*	0.591	0.175	0.432	0.020**
2003	0.540	0.024**	0.575	0.142	0.590	0.001***
2004	0.486	0.605	0.640	0.491	0.479	0.581
2005	0.483	0.253	0.642	0.275	0.481	0.196
2006	0.477	0.120	0.626	0.227	0.530	0.031**
2007	0.489	0.462	0.715	0.044**	0.597	0.077*
2008	0.794	0.002***	0.680	0.146	0.679	0.004***
2009	0.452	0.460	0.748	0.024**	0.575	0.188
2010	0.686	0.014**	0.698	0.179	0.500	0.049**
2011	0.580	0.217	0.517	0.467	0.533	0.199
2012	0.529	0.350	0.760	0.017**	0.512	0.226
2013	0.823	0.012**	0.322	0.827	0.428	0.831
2014	0.410	0.653	0.509	0.134	0.477	0.152
2015	0.544	0.262	0.609	0.380	0.383	0.494
2016	0.424	0.779	0.621	0.344	0.441	0.172
2017	0.456	0.382	0.643	0.373	0.352	0.535
2018	0.556	0.081*	0.703	0.317	0.560	0.121
2019	0.386	0.794	0.727	0.113	0.301	0.790

The test statistic is from the KS test as described in Simar and Wilson (2020). In this case, I use the KS test to test for uniformity of the p-values estimated from the multiple sample-splits used in the tests presented in Table 2.6. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 2001–2019. The p-values are developed using 100 sample-splits and 1,000 bootstrap replications using techniques described in Simar and Wilson (2020). Sample-splits are done unevenly to apply more data to the estimator with the slower rate of convergence, as described by Kneip et al. (2016). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table 2.8: Returns to Scale Test Conditional on Z

	Input–Orientation		Output–Orientation		Hyperbolic	
Year	$\widehat{T}_{KSW}$	p-Value	$\widehat{T}_{KSW}$	p-Value	$\widehat{T}_{KSW}$	p-Value
2001	1.891	0.010***	1.749	0.228	1.547	0.003***
2002	0.970	0.267	0.980	0.813	0.742	0.046**
2003	1.424	0.087*	1.388	0.842	0.967	0.015**
2004	1.740	0.008***	1.613	0.109	1.620	0.000***
2005	1.877	0.042**	1.154	0.347	1.478	0.011**
2006	1.494	0.079*	0.867	0.460	1.302	0.008***
2007	1.149	0.161	0.800	0.254	0.683	0.153
2008	1.474	0.113	0.983	0.363	1.143	0.107
2009	0.890	0.280	1.145	0.269	0.952	0.130
2010	−1.532	0.985	−0.057	0.887	−0.247	0.393
2011	0.345	0.361	1.022	0.155	0.597	0.077*
2012	0.625	0.393	0.793	0.297	0.585	0.141
2013	1.562	0.207	1.272	0.029**	1.551	0.116
2014	1.509	0.075*	0.820	0.039**	1.105	0.100*
2015	1.394	0.036**	0.734	0.520	1.208	0.023**
2016	0.640	0.184	−0.180	0.671	0.036	0.087*
2017	1.661	0.073*	1.086	0.457	1.082	0.094*
2018	1.986	0.006***	1.355	0.159	1.947	0.002***
2019	1.522	0.052*	0.661	0.369	0.860	0.192

The test statistic, described in Kneip et al. (2016), is computed using dimension-reduced data, such that $p = q = 1$. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 2001–2019. The p-values are developed using 100 sample-splits and 1,000 bootstrap replications using techniques described in Simar and Wilson (2020). Sample-splits are done unevenly to apply more data to the estimator with the slower rate of convergence, as described by Kneip et al. (2016). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table 2.9: Returns to Scale Test (KS Test) Conditional on Z

	Input–Orientation		Output–Orientation		Hyperbolic	
Year	$\widehat{T}_{KSW}$	p-Value	$\widehat{T}_{KSW}$	p-Value	$\widehat{T}_{KSW}$	p-Value
2001	0.514	0.064*	0.468	0.999	0.426	0.577
2002	0.309	0.973	0.208	1.000	0.298	1.000
2003	0.473	0.496	0.430	1.000	0.295	0.999
2004	0.535	0.025**	0.439	0.977	0.509	0.253
2005	0.497	0.127	0.310	1.000	0.449	0.600
2006	0.459	0.300	0.287	1.000	0.361	0.951
2007	0.314	0.975	0.365	0.999	0.225	1.000
2008	0.399	0.831	0.375	0.998	0.438	0.904
2009	0.378	0.893	0.315	1.000	0.376	0.970
2010	0.928	0.000***	0.886	0.000***	0.853	0.085*
2011	0.325	0.979	0.356	1.000	0.236	0.998
2012	0.462	0.450	0.320	0.998	0.398	0.885
2013	0.475	0.206	0.403	0.446	0.462	0.422
2014	0.370	0.475	0.342	0.542	0.341	0.786
2015	0.415	0.533	0.205	1.000	0.408	0.882
2016	0.245	0.989	0.177	1.000	0.222	0.998
2017	0.437	0.540	0.428	0.983	0.393	0.862
2018	0.552	0.003***	0.415	0.350	0.570	0.010***
2019	0.415	0.408	0.279	0.998	0.285	0.979

The test statistic is from the KS test as described in Simar and Wilson (2020). In this case, I use the KS test to test for uniformity of the p-values estimated from the multiple sample-splits used in the tests presented in Table 2.8. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 2001–2019. The p-values are developed using 100 sample-splits and 1,000 bootstrap replications using techniques described in Simar and Wilson (2020). Sample-splits are done unevenly to apply more data to the estimator with the slower rate of convergence, as described by Kneip et al. (2016). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table 2.10: Test For Equivalency of Mean Efficiency: FDH Estimator Conditioned on Z and T

	Input–Orientation		Output–Orientation		Hyperbolic	
Period	Statistic	p-value	Statistic	p-value	Statistic	p-value
2001–2002	2.688	0.007***	2.006	0.045**	2.221	0.026**
2002–2003	−1.317	0.188	−0.921	0.357	−2.145	0.032**
2003–2004	1.058	0.290	−0.733	0.463	1.116	0.264
2004–2005	0.044	0.965	−0.129	0.898	0.256	0.798
2005–2006	−0.481	0.631	0.150	0.881	−0.877	0.380
2006–2007	1.791	0.073*	2.062	0.039**	2.490	0.013**
2007–2008	1.138	0.255	0.229	0.819	0.758	0.449
2008–2009	1.310	0.190	0.261	0.794	0.475	0.635
2009–2010	0.244	0.808	−1.188	0.235	−0.053	0.957
2010–2011	−2.767	0.006***	−0.320	0.749	−3.077	0.002***
2011–2012	0.248	0.804	−0.188	0.850	0.765	0.444
2012–2013	0.269	0.788	0.209	0.835	0.248	0.804
2013–2014	1.275	0.202	1.547	0.122	1.130	0.258
2014–2015	−0.645	0.519	−0.311	0.756	0.096	0.923
2015–2016	2.119	0.034**	1.138	0.255	1.444	0.149
2016–2017	1.538	0.124	2.570	0.010***	1.441	0.150
2017–2018	−3.407	0.001***	−2.048	0.041**	−2.249	0.024**
2018–2019	−1.253	0.210	−2.691	0.007***	−2.498	0.012**
2001–2019	4.261	0.000***	1.098	0.272	2.705	0.007***
2002–2019	1.169	0.242	−1.186	0.236	−0.064	0.949
2003–2019	2.878	0.004***	−0.181	0.857	2.643	0.008***
2004–2019	1.522	0.128	0.630	0.529	1.533	0.125
2005–2019	1.680	0.093*	0.812	0.417	1.250	0.211
2006–2019	2.218	0.027**	0.661	0.508	2.261	0.024**
2007–2019	0.231	0.817	−1.576	0.115	−0.615	0.538
2008–2019	−1.193	0.233	−1.983	0.047**	−1.674	0.094*
2009–2019	−2.625	0.009***	−2.212	0.027**	−2.260	0.024**
2010–2019	−2.810	0.005***	−1.013	0.311	−2.102	0.036**
2011–2019	0.804	0.422	−0.664	0.507	1.810	0.070*

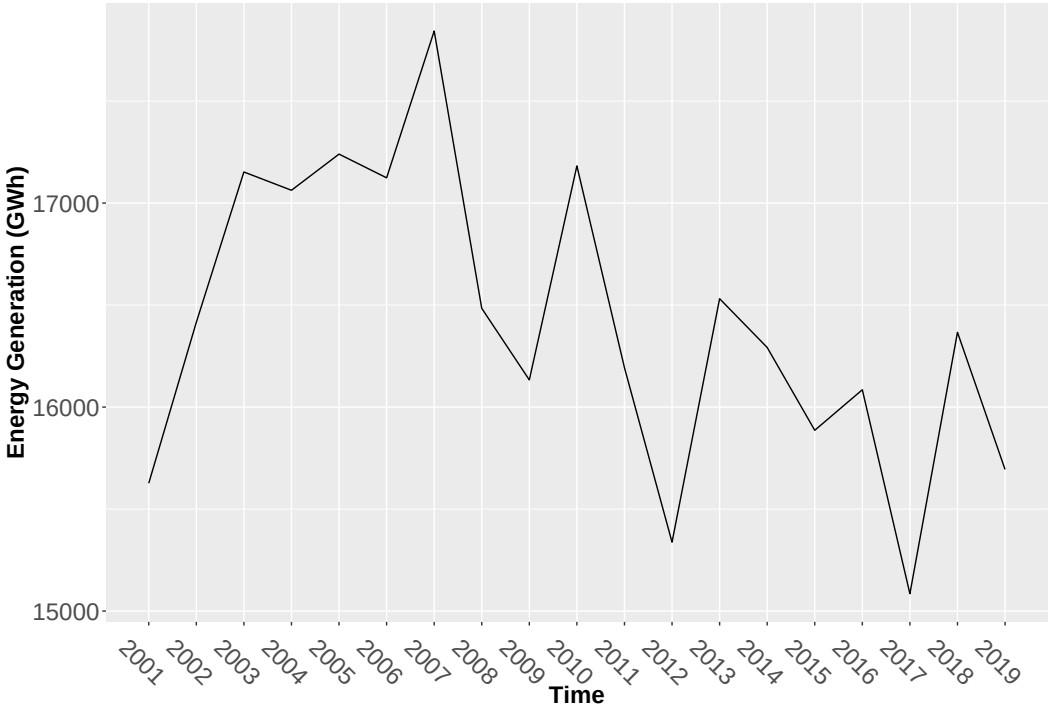
The test statistic, described in Kneip et al. (2016), is computed using the FDH estimator with dimension-reduced data, such that $p = q = 1$. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 2001–2019. One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table 2.11: Productivity Estimates

Period	Statistic	*p*value
2001–2002	−0.470	0.638
2002–2003	0.183	0.855
2003–2004	0.379	0.705
2004–2005	−0.980	0.327
2005–2006	−0.324	0.746
2006–2007	1.393	0.164
2007–2008	−2.152	0.031**
2008–2009	−3.080	0.002***
2009–2010	1.838	0.066*
2010–2011	−2.073	0.038**
2011–2012	−3.017	0.003***
2012–2013	2.801	0.005***
2013–2014	−1.942	0.052*
2014–2015	−1.535	0.125
2015–2016	0.012	0.990
2016–2017	−0.229	0.819
2017–2018	2.144	0.032**
2018–2019	−1.070	0.285
2001–2019	−3.977	0.000***
2002–2019	−4.004	0.000***
2003–2019	−4.000	0.000***
2004–2019	−4.022	0.000***
2005–2019	−3.501	0.000***
2006–2019	−3.752	0.000***
2007–2019	−5.086	0.000***
2008–2019	−4.533	0.000***
2009–2019	−2.695	0.007***
2010–2019	−3.494	0.000***
2011–2019	−2.184	0.029**

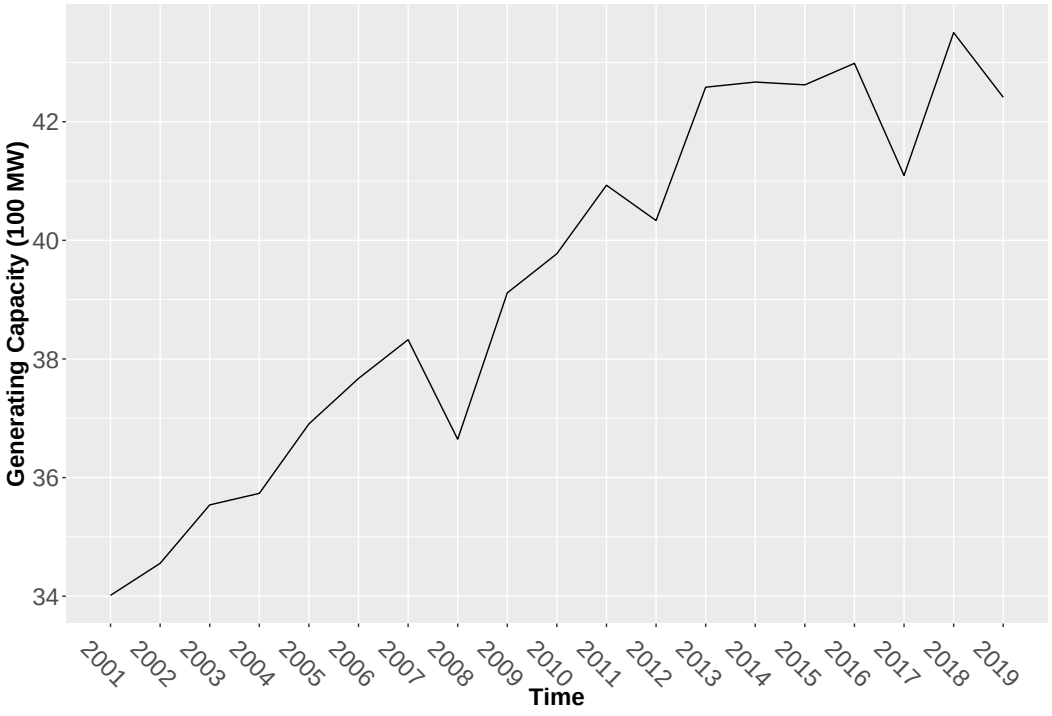
Change in productivity test statistics are computed with dimension-reduced data, such that $p = q = 1$. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 2001–2019. One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Figure 2.1: Average Net Generation of Electric Utilities: 2001–2019



This figure presents data from the FERC Form 1 on average net generation of electric utilities by year. These data are exclusive of energy purchased by utilities then re-sold to customers.

Figure 2.2: Average Generating Capacity of Electric Utilities: 2001–2019

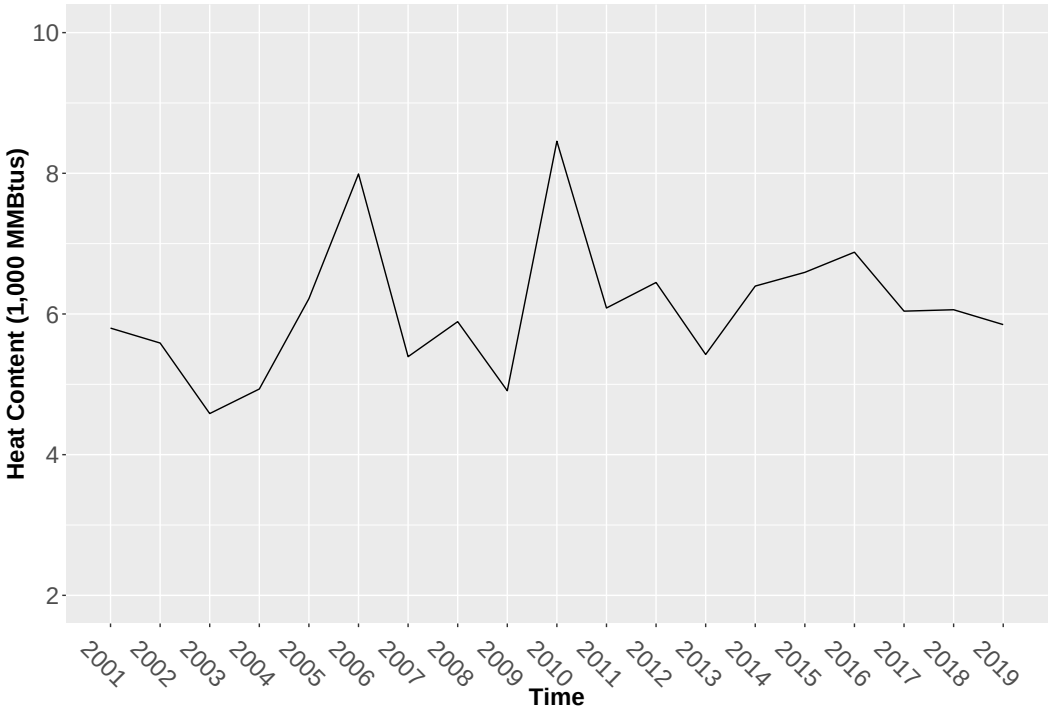


This figure presents data from the FERC Form 1 on average generating capacity of electric utilities by year.

Figure 2.3: Average Number of Employees at Electric Utilities: 2001–2019

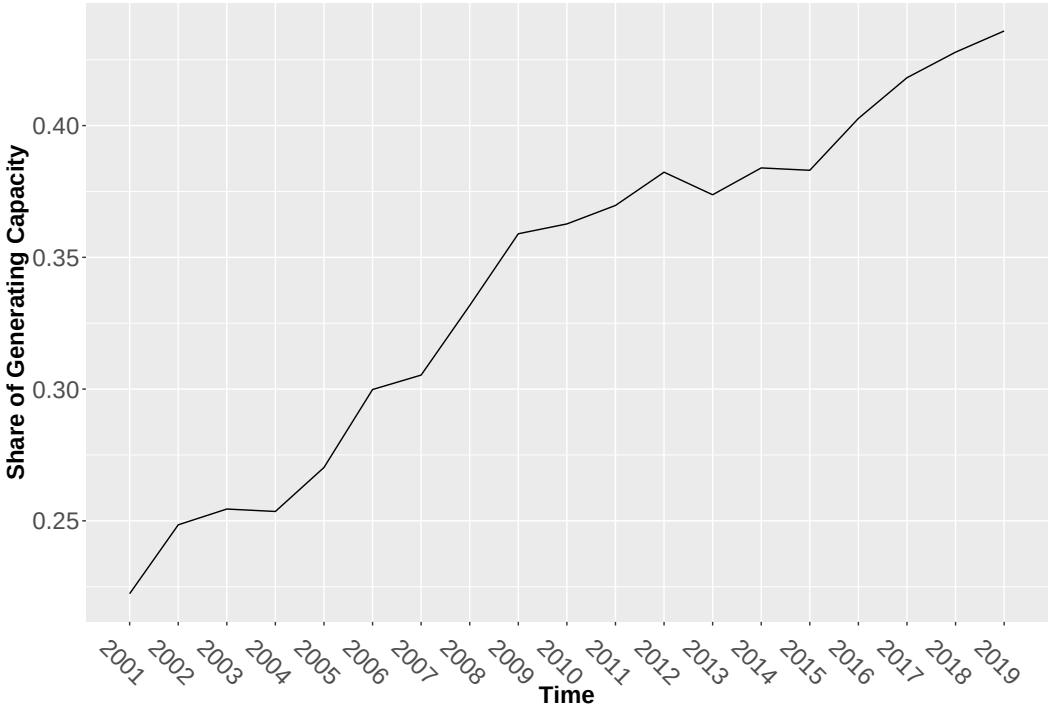
This figure presents data from the FERC Form 1 on average number of power plant employees at electric utilities.

Figure 2.4: Average Heat Content of Fuel Used by Electric Utilities: 2001–2019



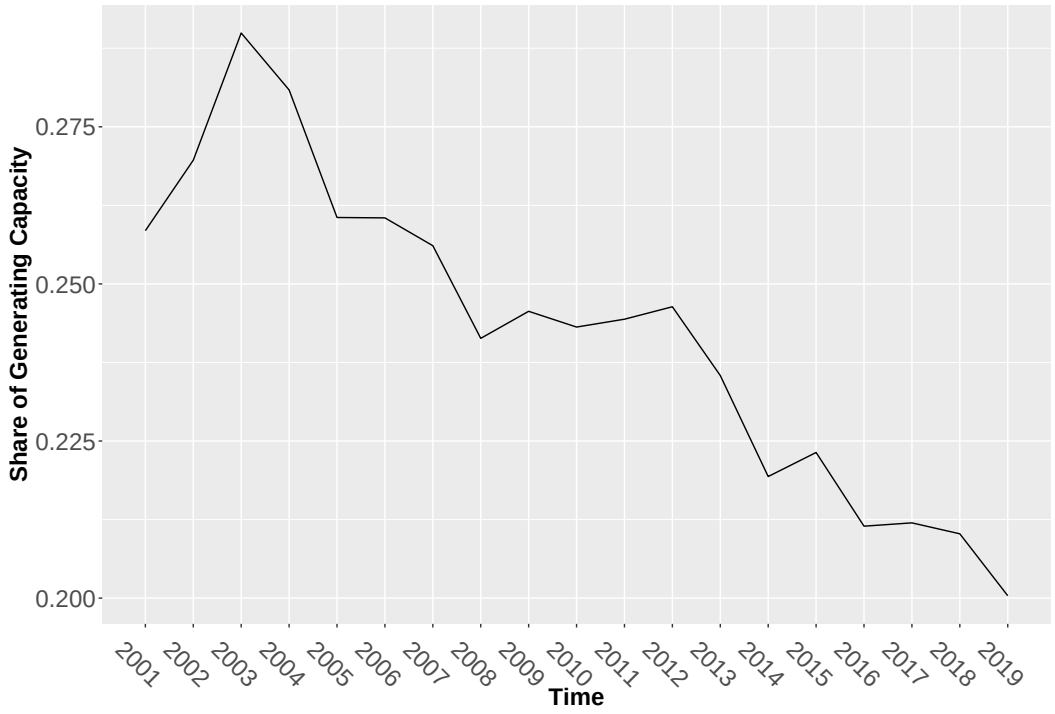
This figure presents data from the FERC Form 1 on average heat content of fuel used by electric utilities.

Figure 2.5: Average Share of Natural Gas Fueled Generating Units: 2001–2019



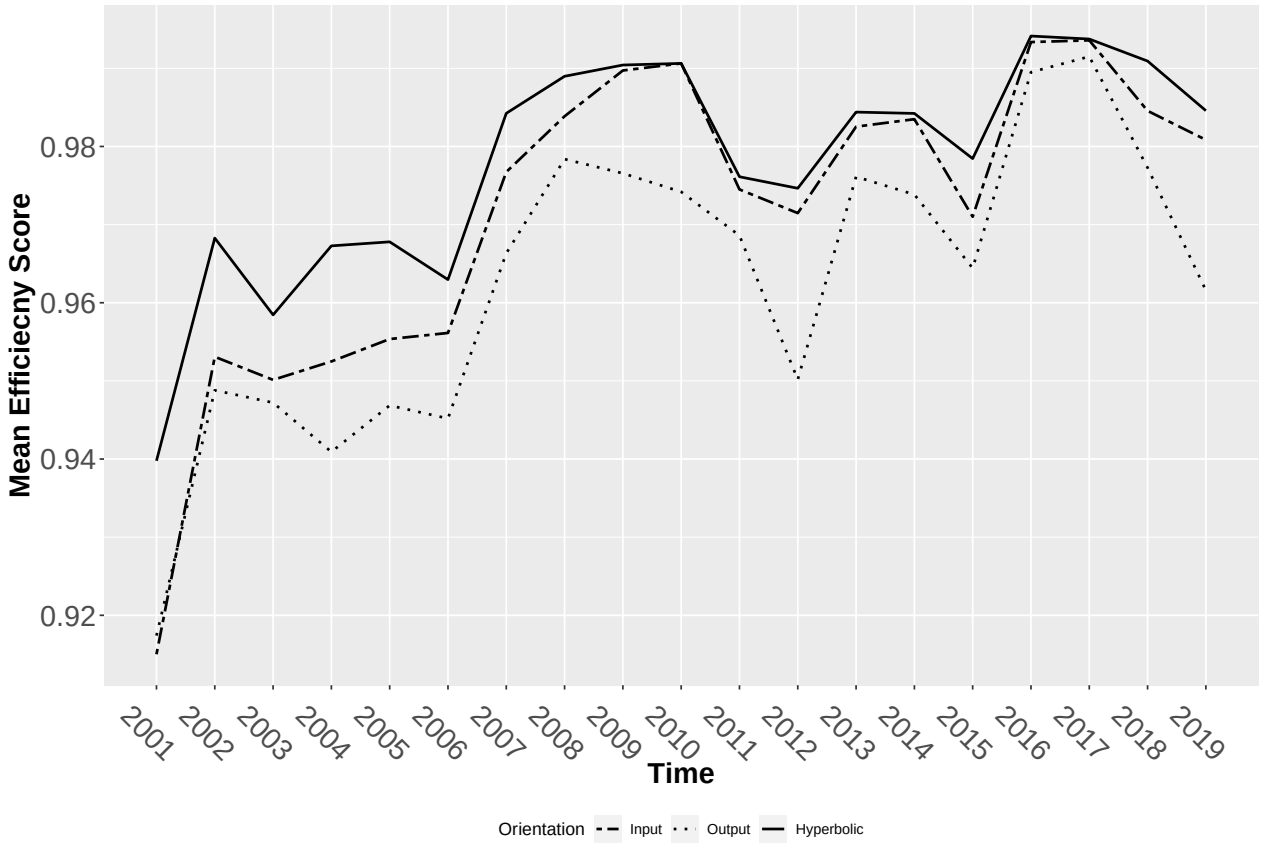
This figure presents data from the FERC Form 1 on average share of capacity attributed to natural gas powered generating units.

Figure 2.6: Average Share of Coal Fueled Generating Units: 2001–2019



This figure presents data from the FERC Form 1 on average share of capacity attributed to coal powered generating units.

Figure 2.7: Mean Efficiency Scores by Orientation Conditional on Z and T: FDH Estimator



The table shows mean efficiency scores by year conditional on Z for the input, output, and hyperbolic orientations. The input and hyperbolic orientations are expressed in the Farrell metric, and the output orientation is expressed in the Shephard metric. The estimation is done using dimension-reduced data such that ($p = q = 1$), and principal components are estimated using the pooled sample of data from 2001–2019.

Chapter 3

Price Convergence Across Natural Gas Markets During The Shale Revolution

3.1 Introduction

In the early 1980s the Federal Energy Regulatory Commission (FERC) began to implement a series of reforms to restructure how wholesale natural gas was transported and delivered to markets from gas producing regions in the U.S. These reforms were developed with the intention of creating a national spot market for natural gas. In order to do this FERC imposed measures to decouple the production of gas from the trading of the commodity itself. In 1992 FERC created a regulatory order, Order 636, designed to "improve the competitive structure of the natural gas industry" where the primary goal was to "ensure that all shippers have meaningful access to the pipeline transportation grid so that willing buyers and sellers can meet in a competitive, national market to transact the most efficient deals possible" (FERC, 1992).

The stated goals of Order 636 imply that geographically dispersed gas markets were not integrated into a national spot market prior to the gas market restructuring that occurred in the

early 1990s. In order to infer if FERC's objectives have been achieved, I test whether regional gas markets in the U.S. became integrated into the same market from July 1996 through December 2019. In order to do this I examine the daily differences (or price-gaps) between pairs of 71 geographically dispersed price hubs in the U.S. and Canada. More specifically, I use several unit root tests to test for stationarity of the price-gaps, where stationary price-gaps would suggest the presence of an equilibrium price-gap. This would provide evidence of the corresponding price hubs existing in the same market.

Previous studies examining the U.S. natural gas markets suggest that Order 636 played a role in integrating some geographically dispersed natural gas markets in the U.S. into the same market. For example, King and Cuc (1996), Doane and Spulber (1994), and Cuddington and Wang (2006) use various time series methods to assess the convergence of U.S. natural gas prices at different gas price hubs across the U.S. Their results suggest that different regional gas markets became more integrated through FERC's market restructuring. In particular, Doane and Spulber (1994) discuss how FERC's restructuring allowed for unbundling of the merchant and transportation services pipeline's traditionally provided. Prior to Order 636, pipeline firms acted as both a transporter and merchant (or seller) of natural gas. Pipeline firms that also acted as sellers of natural gas had an incentive to restrict access to their pipeline network, creating market power in the selling of natural gas. In response to this, Order 636 relegated pipeline firms to only the transportation of gas and required pipelines to become an open access resource available to merchants of natural gas.[1] This resulted in a larger set of natural gas merchants gaining access to pipelines for transporting gas. Moreover, local distribution companies, electric utilites, industrial plants, and other customers began to enter the wholesale market with more options for purchasing and receiving gas. Open access meant that purchase agreements between customers, producers, and pipelines were no longer tied to a specific pipeline and producer pair. This coincided with the duration of shipping contracts becoming shorter with more flexible terms. Thus, a wholesale natural gas customer could substitute between a wider variety of delivery routes to ship gas, allowing for more arbitrage across different regional gas markets. This resulted in the integration of many geographically dispersed natural gas trading hubs into the same market.

However, King and Cuc (1996) and Cuddington and Wang (2006) argue that natural gas

[1]While firms that own pipelines may also have a natural gas merchant function, the merchant operations of gas must be done at an "arms-length" from transporting gas under Order 636.

markets were only imperfectly integrated by the 1990s. King and Cuc (1996) examine monthly U.S. natural prices from 1986 through 1995, and observe an east-west split in U.S. natural gas markets during this time. Moreover, Cuddington and Wang (2006) examine U.S. natural gas markets from 1993 through 1997 with daily natural gas spot price data, and provide more evidence of a separate western U.S. gas market during this period. In order to do this, Cuddington and Wang (2006) estimate the proportion of price-gaps between prices in different regions of the U.S. that are unit root (or non-stationary) processes. The presence of a unit root in the price-gap may suggest that the corresponding prices may respond to different market shocks, indicating arbitrage between the price hubs may not be possible. Cuddington and Wang (2006) estimate that 84 percent of price-gaps between east and west pricing hubs were not integrated into the same market during 1993–1997.

Since 1997, U.S. natural gas markets have changed. One might expect that U.S. natural gas markets became more integrated through the late 1990s and early 2000s for several reasons. First, the increase in shale production, starting in the early 2000s (a period known as the Shale Revolution), diversified natural gas supply regions in the U.S. Second, this expansion of shale gas production coincided with increasing pipeline capacity in the U.S. In addition, the largest pipeline in the U.S., Rockies Express Pipeline (REX), was commissioned in 2009 and runs from the rocky mountains to eastern Ohio. One might expect that the development of east-west pipeline capacity would integrate eastern and western natural gas markets. While I use an approach similar to that of Cuddington and Wang (2006) and use unit root tests on the price-gaps between geographically dispersed price hub pairs, my analysis differs from Cuddington and Wang (2006), and other previous papers, in that I examine U.S. natural gas markets with a later sample period of July 1996 through December 2019. Moreover, while Cuddington and Wang (2006) use only a Ng-Perron unit root test on price-gaps, I use several other unit root tests on price-gaps as robustness checks.

The remainder of the paper is organized as follows. Section 3.2 discusses the statistical model and the method for testing for stationarity of price-gaps between price hubs. Section 3.3 describes the data. Section 3.4 reports the empirical results and findings. Finally, a summary and overview of the conclusions are given in Section 3.5.

3.2 Statistical Model

To test for market integration between price hubs I use an autoregressive (AR) model, similar to that used by Cuddington and Wang (2006), to examine natural gas price-gaps from July 1996 through December 2019. Let $x_t^{ij} = p_t^i - p_t^j$ denote the price-gap between price hub i and price hub j at time t. The idea behind using an AR approach is that if the price series that generate p_t^i and p_t^j are both $I(1)$ processes (i.e., if the price series p^i and p^j are integrated of order 1), they could lie within the same market if price shocks at either hub i or j cause a similar price change at the other hub. In this case, arbitrage will be possible if two price hubs lie within the same market, resulting in an equilibrium price-gap $\bar{x}$. This is consistent with the definition of a market provided by Stigler and Sherwin (1985).

Of course price-gaps are not expected to continuously reflect just $\bar{x}$. Cuddington and Wang (2006) present the partial adjustment mechanism or

$$x_t^{ij} - \bar{x} = \lambda(x_{t-1}^{ij} - \bar{x}) + \epsilon_t^{ij}, \tag{3.1}$$

which was derived from the Enke-Samuelson-Takayama-Judge model of spatial equilibrium in the presence of transport costs and capacity constraints. Expression (3.1) suggests that within a market, deviations from an equilibrium price-gap in period t is posited to be a fraction, λ, of the deviation in $t-1$. In this case λ measures the persistence of disequilibrium where $\lambda \in [0, 1]$. Rearranging terms in (3.1) implies the $AR(1)$ specification

$$x_t^{ij} = c + \lambda x_{t-1}^{ij} + \epsilon_t^{ij}, \tag{3.2}$$

where $c = (1 - \lambda)\bar{x}$. I use the specification in (3.2) to test whether $\lambda = 1$. If $\lambda = 1$, a shock to the price-gap, x_t, is permanent, suggesting that the price hubs of the price-gap to not lie within the same market. Testing $H_0 : \lambda = 1$ versus $H_1 : \lambda \neq 1$ amounts to a unit root test. See Enke (1951), Samuelson (1952), Takayama and Judge (1964, 1971), and Cuddington and Wang (2006) for technical details.

Cuddington and Wang (2006) also discuss persistent serial correlation found in their price-

gaps, and suggest a higher-order $AR(q)$ process

$$x_t^{ij} = c + \lambda x_{t-1}^{ij} + ... + \lambda x_{t-q}^{ij} + \epsilon_t \tag{3.3}$$

to model price-gap dynamics. In this case, I determine the number of lags, q, using two information criteria, i.e., the Akaike information criterion (AIC) and the Bayes information criterion (BIC). For unit root tests that rely on an $AR(q)$ model to deal with serial correlation, I report results corresponding to both information criteria as robustness checks.

It is well-known that the power of unit root tests can be low. Enders (2009) notes that these tests will often fail to reject the presence of a unit root, when in fact the unit root hypothesis should be rejected. Many unit root tests have been proposed in an attempt to increase the power of unit root testing. Because of this I use four different tests to make inference on the stationarity of the price-gaps. These tests are the Augmented Dickey-Fuller (ADF) test, as described by Said and Dickey (1984), the Phillips-Perron test, as described by Phillips and Perron (1988), the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test, as described by Kwiatkowski et al. (1992), and the modified Dickey–Fuller test (ADF-GLS) as described by Elliott et al. (1996). While I expect that each test will vary in terms of the proportion of price-gaps identified as unit root processes, an increasing trend of stationary price-gaps across all tests will provide stronger evidence of increasing market integration, as opposed to relying on a single unit-root test. Moreover, for tests where unit roots are the null hypothesis, any rejection of the null should provide strong evidence of stationarity of the price-gaps due to the low power of these tests.

3.3 Data

The sample is comprised of daily volume-weighted average natural gas prices at 71 geographically dispersed price hubs in the U.S. The sample contains data on daily prices from July 1993 through December 2019 collected from Natural Gas Intelligence (NGI), a trade publication that produces natural gas price indices. NGI natural gas pricing data are developed from information collected from market participants and the Intercontinental Exchange.[2]

[2]The Intercontinental Exchange operates a marketplace for financial assets and commodities, including natural gas.

The selection of the 71 price hubs is based on a review of previous studies examining U.S. natural gas prices and encompasses major trading points of natural gas at all major producing regions and citygates. Table 3.1 presents the price hub counts by region in my sample. Moreover, Figure 3.1 presents a map with the locations of the pricing hubs.[3] NGI categorizes price hubs by natural gas region. For example, in my sample NGI categorizes seven hubs in the Appalachia natural gas region. Consistent with Cuddington and Wang (2006, Table 2), I categorize each NGI region into larger geography of east, central, and western U.S. regions. In addition, NGI provides major Canadian price hubs, which Cuddington and Wang (2006, Table 2) do not include. Scarcioffolo and Etienne (2019) discuss how these Canadian price hubs have become major trading points for supplying natural gas to the western U.S.; therefore I include these Canadian price hubs in my sample.

To reduce the skewness and kurtosis of the price series, the natural logarithm of the price series are used for statistical testing as opposed to prices in levels. This standard practice in the literature examining natural gas prices as well the prices of other commodities. It is also common in research studying natural gas prices to examine the price series and first differences of prices at major price hubs in the sample. Consistent with Cuddington and Wang (2006), I examine the price series at Henry Hub, Southern Natural, and Kern River. Figure 3.2 presents the daily price series in logs as well as the first difference in log prices. While the log prices suggest that the three price series follow a similar pattern, there are some periods where the price movements differ. Henry Hub and Southern Natural are U.S. Gulf Coast pricing points in Louisiana, while Kern River is a major pricing point in California. While Southern Natural and Henry Hub prices, unsurprisingly, appear similar throughout all time periods, they differ in certain periods with Kern River. Notably, Kern River's log price during 2002–2004, 2005–2009, and 2017–2019 appears more volatile than prices at Henry Hub and Southern Natural. This is particularly apparent when examining the first difference in log prices. Consequently, one might expect that Kern River's prices might follow market shocks that are distinct from Southern Natural and Henry Hub. However, formal tests are needed to make inference on this.

[3]In some cases natural gas price indices are based on sales from sections of pipeline or within a larger region. In this case the location of the hub presented in Figure 3.1 is an approximate location.

3.4 Estimation and Results

To determine if an equilibrium price-gap between two price hubs exists (i.e. the price-gap is stationary), or if the price-gap is a unit root process (such as a random walk), I first examine whether the price series are themselves stationary processes. If the price series are stationary, the test of market integration would indicate that price hubs are in the same market regardless of whether customers could substitute between markets, as argued by Werden and Froeb (1993). In this case, examining price-gaps for market integration would be misleading. As explained in Cuddington and Wang (2006, footnote 15), two independent random walks, such as price gaps between non-market integrated price hubs with price series that are $I(1)$, will have a non-stationary price-gap. Thus, the price series themselves need to be non-stationary in order to test for market integration between the price hub pairs. Using an ADF test I find that 4 out of the 71 price series are stationary.[4] Thus I remove these 4 price series from my sample. The remaining 67 price hubs had price series that were stationary in their first difference, suggesting they are $I(1)$ processes and valid for testing for equilibrium price-gaps. My remaining sample of price series from 67 price hubs amounts to a total of 2,211 bilateral price-gaps.

With my remaining sample I first use an ADF test on the price-gaps to test for market integration. In my application, I parse my sample period into three-year intervals and calculate the proportion of stationary price-gaps in each interval.[5] I define a price-gap as stationary if the ADF test rejects the the presence of a unit root. Figure 3.3 shows the proportion of stationary price-gaps over time by lag selection procedure.[6] Figure 3.3 also shows the proportion of rejections at the 1, 5, and 10 percent levels as a robustness check. While the results may vary depending on the lag selection procedure and chosen level of significance, some distinct trends in the stationarity of the price-gaps are apparent across all tests. These include the increase in the proportion of stationary gaps during 1993–2001, and the decline in the proportion of stationary gaps during 2002–2010. The increase in stationary price-gaps is consistent with results obtained by Doane and Spulber (1994) and Cuddington and Wang (2006), who discuss how FERC's restructuring of gas markets played a

[4] As noted above I perform these tests on the natural log of prices.

[5] In some cases, price data are not reported for certain periods at some hubs. For example, the El Paso South Mainline/Noth Baja price hub was not tracked as a major price index until September of 2004. Thus, In each three-year time interval, I only calculate price-gaps (and thus test for unit roots of price-gaps) from price hubs with at least 100 days of observed prices in that time interval.

[6] It is clear from Figure 3.3 that different lag selection procedures produce different results in determining if a price-gap is stationary. The number of lags chosen by each procedure differed. The average lag length selected under the AIC and BIC procedures was 11.3 and 4.6 days, respectively.

role in integrating different gas markets in the U.S.

While one might expect that stationarity of the price-gaps would have persisted, this was not the case. From 2002 through 2010, market integration appears to have declined as suggested by the decreasing proportion of stationary price-gaps. Others have speculated as to why market integration may have declined during this period. Scarcioffolo and Etienne (2019) examine seven natural gas spot markets in the U.S. and Canada during 1994–2016 and argue that market integration declined towards the end of their sample period and cite three potential reasons why this may have occurred. To begin, market power and market manipulation are issues in U.S. natural gas markets. Despite the efforts of FERC to curb market power and prevent market manipulation, some major traders of natural gas have become more sophisticated in their ability to control prices. The second issue is the reduced trading volume at some price hubs stemming from a reduction in the number of traders in the natural gas market. After Enron's bankruptcy in 2001, the credit requirements required by trading counter-parties increased while the number of market intermediaries in many regional markets diminished. Finally, Scarcioffolo and Etienne (2019) argue that pipeline constraints during the shale gas boom played a major role in reducing the contentedness of gas markets across eight different price hubs.

While the results suggest that market integration declined from 2002 through 2010, price-gaps become more stationary from 2011 through 2019. This coincides with the development of major pipeline systems and restructuring of existing systems to accommodate large increases in shale gas production that occurred during this period. For example, REX became fully operation in 2009. It was initially designed to carry gas west to east but, in 2015 its flow was reversed to bring shale gas produced in the east to western markets. During this time, shale production in the the eastern U.S. put eastern U.S. gas prices at a discount to western gas prices. REX and other developed pipelines may have played a role in equating prices across western and other regions. Previous studies have found that western gas markets were not integrated into other regional gas markets in the U.S. prior to the expansion of shale gas. Figure 3.4 presents the proportion of stationary price-gaps between western hubs and hubs in other regions. In this case we see similar trends to the results from examining all price-gaps. This suggests the western U.S. natural gas markets eventually became integrated with other regional markets as well.

In addition to examining the proportion of price-gaps that are stationary processes, I

inspect the distribution of the p-values from the ADF tests for each time period. Figure 3.5 shows empirical distribution functions of the p-values from the ADF test results shown in Figure 3.3. To visually inspect the frequency with which the null is rejected, vertical lines are located at 0.1 on the horizontal axes. It is clear that the distribution of the p-values changes over time with a lower frequency of rejecting the null the during 1993–1995 and 1996–1998, relative to later time periods. Moreover, Figure 3.6 shows empirical distribution functions of the p-values from the ADF test results shown in Figure 3.4, where the analysis was limited to price-gaps between western hubs and hubs in other regions. In this case, the change in the distribution of the p-values is similar. The increasing frequency with which unit roots of price-gaps are rejected over time suggests that price-gaps became more stationary over time and responsive to similar market shocks. This provides some evidence of U.S. gas markets becoming more integrated over time and the western markets becoming increasingly integrated with other U.S. gas markets as well.

As noted above, I use several other unit root tests to test for stationarity of the price-gaps. I perform the same procedure, as outlined above, with the ADF-GLS, Phillips-Perron, and KPSS unit root tests on the price-gaps. Figure 3.7 shows the proportion of stationary price-gaps over time by lag selection procedure, where stationarity is defined as a rejection of the unit root (or null) hypothesis of the ADF-GLS test. As with the ADF test I also examine empirical distribution functions of the corresponding p-values, as shown in Figure 3.8. Comparing the results from ADF test and ADF-GLS tests illustrates the varying degrees of power associated with unit root tests. More specifically, the two tests differ in the proportions of price-gaps determined to be stationary. Notably, the proportion of stationary price-gaps appears to fluctuate each time period under the ADF-GLS test. However, the trends and the qualitative conclusions of the tests are similar as both tests indicate an increasing trend in the proportion of price-gaps that are stationary. Moreover, these trends remain true when examining price gaps between western and other regions, as shown in Figures 3.9 and 3.10. This provides further evidence that not only have U.S. gas markets become more integrated over time, but the distinction between western gas markets and other regions in the U.S. has become less apparent.

Figure 3.11 shows the results the Phillips-Perron tests. As with the ADF and ADF-GLS test, the null hypothesis of the Phillips-Perron test is the presence of a unit root. In this case, the results depict a sharp increase in the proportion of price-gaps that are stationary with nearly 100

percent of price-gaps remaining stationary for the remainder of the sample period after 1999.[7] This becomes increasingly apparent when examining the empirical distribution functions of the p-values which are shown in Figure 3.12. The range of the p-values become more narrow and closer to 0 in the time periods during 1999–2019. For example, the largest p-value from the Phillips-Perron test during the sample period of 1993 through 1995 was 0.997, while during 2017–2019 the maximum p-value was 0.028. Similar to the ADF and ADF-GLS tests noted above, the Phillips-Perron tests largely differs from the other tests in the proportion of price gaps determined to be stationary each time period. However, the qualitative conclusion of increasing stationarity of the price-gaps still holds. Moreover, Figures 3.13 and 3.14 show the same analysis, but limited to price gaps between western and other regions. As with the ADF and ADF-GLS tests, these results provide strong evidence of market integration of western gas markets to a larger U.S. gas market.

Finally, the analogous results of the KPSS tests are shown in Figures 3.15–3.18. While the ADF, ADF-GLS, and Phillips-Perron tests have the presence of unit roots as the null hypothesis, the KPSS test has the presence of a unit root as the alternative hypothesis. In this case, we would expect to see decreasing rejections of the null as some evidence of increasing market integration under the KPSS test on price-gaps. As with all the other unit root tests, the KPSS test largely differs in the quantitative result of the proportion of price-gaps indicated as stationary versus unit root processes, as shown in Figures 3.15 and 3.16. As with the analysis presented above, Figure 3.15 shows the proportion of stationary price-gaps for the whole sample using a KPSS test, while Figure 3.16 shows the corresponding empirical distribution functions of the p-values from the tests. While I observe decreasing rejections of the null hypothesis under the KPSS test, failure to reject the null is not proof that the null hypothesis of stationarity is true. However, the qualitative results of the test are largely consistent with the results of the other tests. The KPSS test alone may not provide strong evidence for market integration, however these results along with the results from the other tests all suggest market integration has increased. As shown in Figures 3.17 and 3.18 these results also hold when examining the gaps between western and other regions as well.

[7] As noted above, unit root tests, such as the Phillips-Perron test, will often fail to reject a false null hypothesis, due to their low power. Thus any rejections of the null in this case are strong results in favor of stationarity of the price gaps.

3.5 Conclusion

In the analyses outlined above I use unit root tests to determine if the price-gaps between geographically dispersed price hubs were stationary. These tests were done to determine if the price hubs corresponding to a price-gap responded to the same market shocks. Consistent with previous studies, I find that from 1993 through 2001 the proportion of price-gaps that were stationary increased, which suggests that a larger set of dispersed price hubs were integrated into the same market. This coincides with FERC's restructuring of gas markets in the U.S. allowing greater entry of natural gas marketers. However, from 2002 through 2010 I find that the level integration declined. This coincides with a period of increased pipeline capacity constraints and reduced competition in natural gas trading. Finally, from 2011 through 2019 I find the integration increased as the proportion of stationary price-gaps increased. The overall trend of increasing stationarity of the price-gaps is robust across four different unit root tests. Moreover, previous work has not considered the integration of western U.S. gas markets into other U.S. gas market regions. I find that western markets may have been integrated into the same market as other regions in the U.S. This coincides with large capacity increases and pipeline expansions running east to west.

Table 3.1: Count of Price Hubs by Region

Canada	
Total:	2
Central Region	
Midcontinent	3
Midwest	10
South Texas	3
West Texas SE New Mexico	3
Total:	19
East	
Appalachia	7
East Texas	4
North Louisiana Arkansas	2
Northeast	12
South Louisiana	6
Southeast	5
Total:	36
West	
Arizona Nevada	2
California	4
Rocky Mountains	8
Total:	14

Figure 3.1: Natural Gas Pricing Hubs

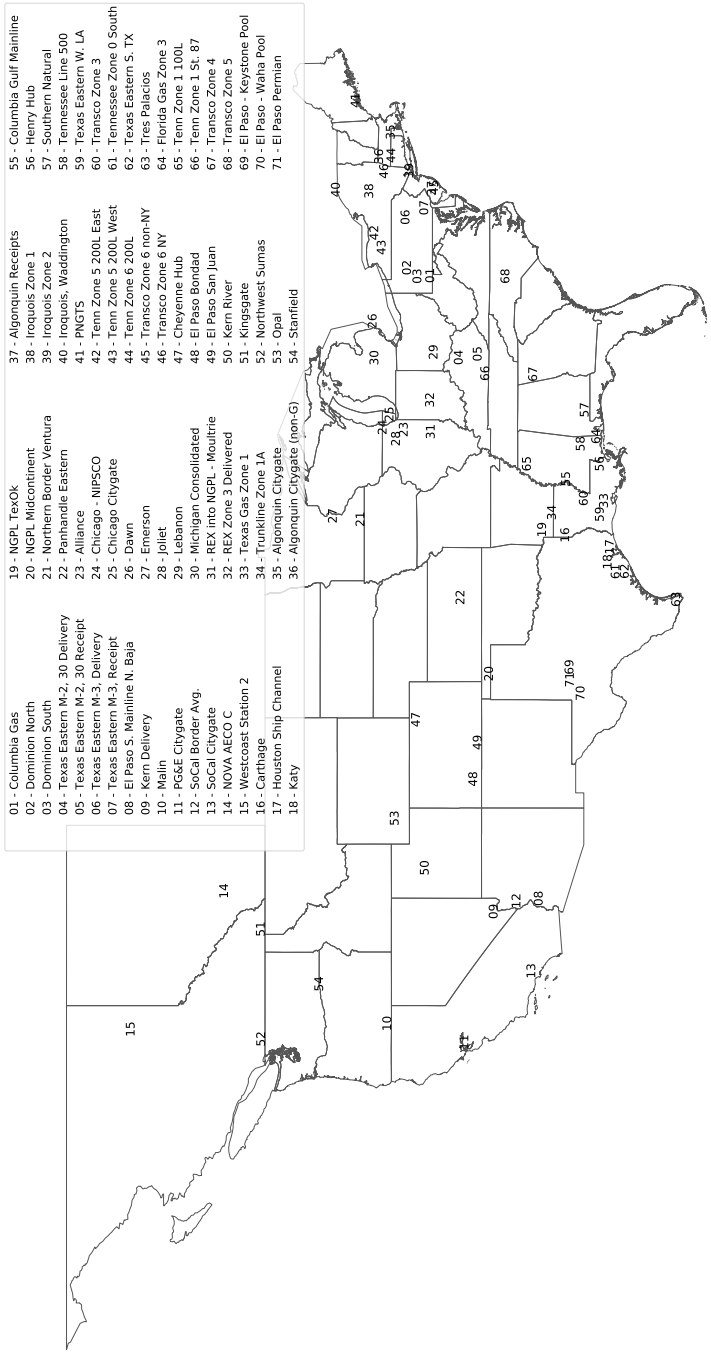


Figure 3.2: Daily Log Prices & First Differences of Log Prices - Kern River, Henry Hub, and Southern Natural

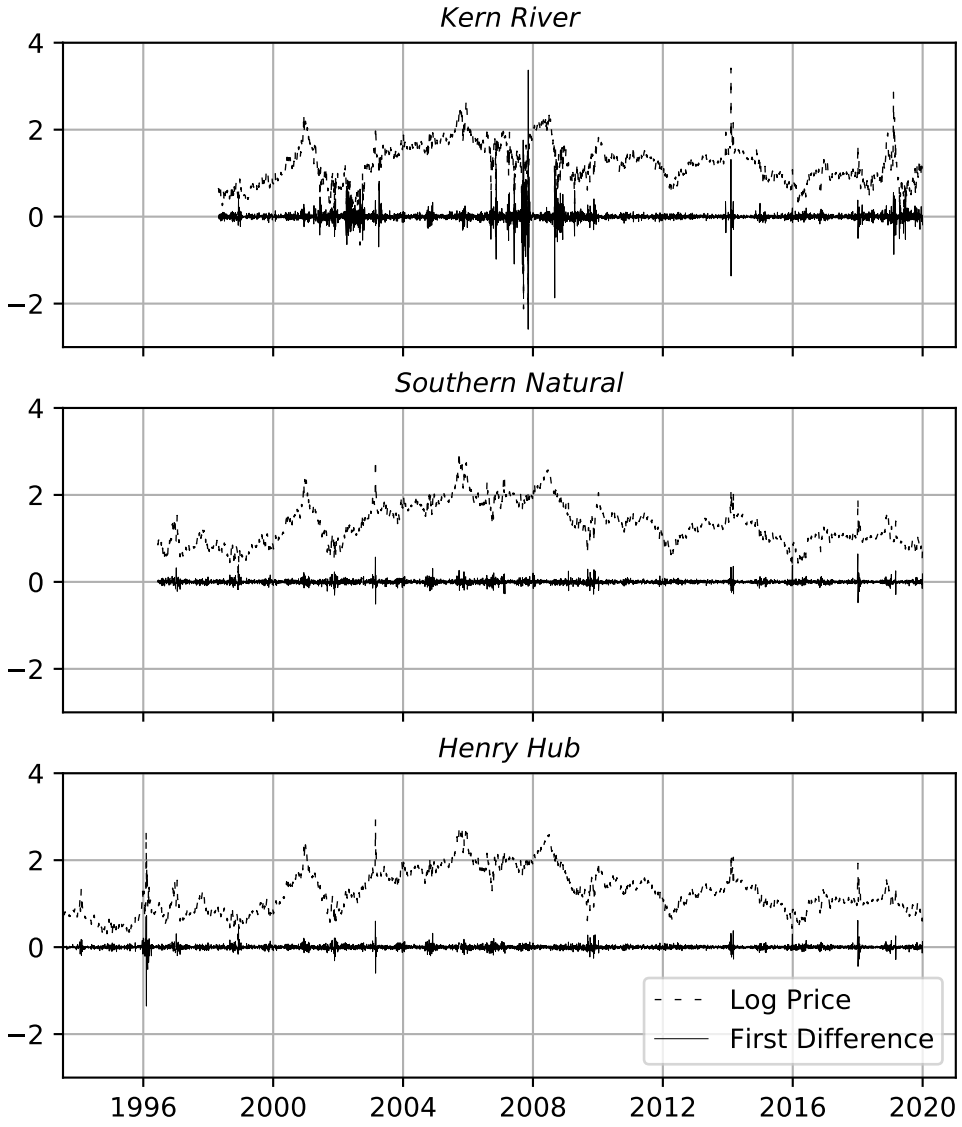
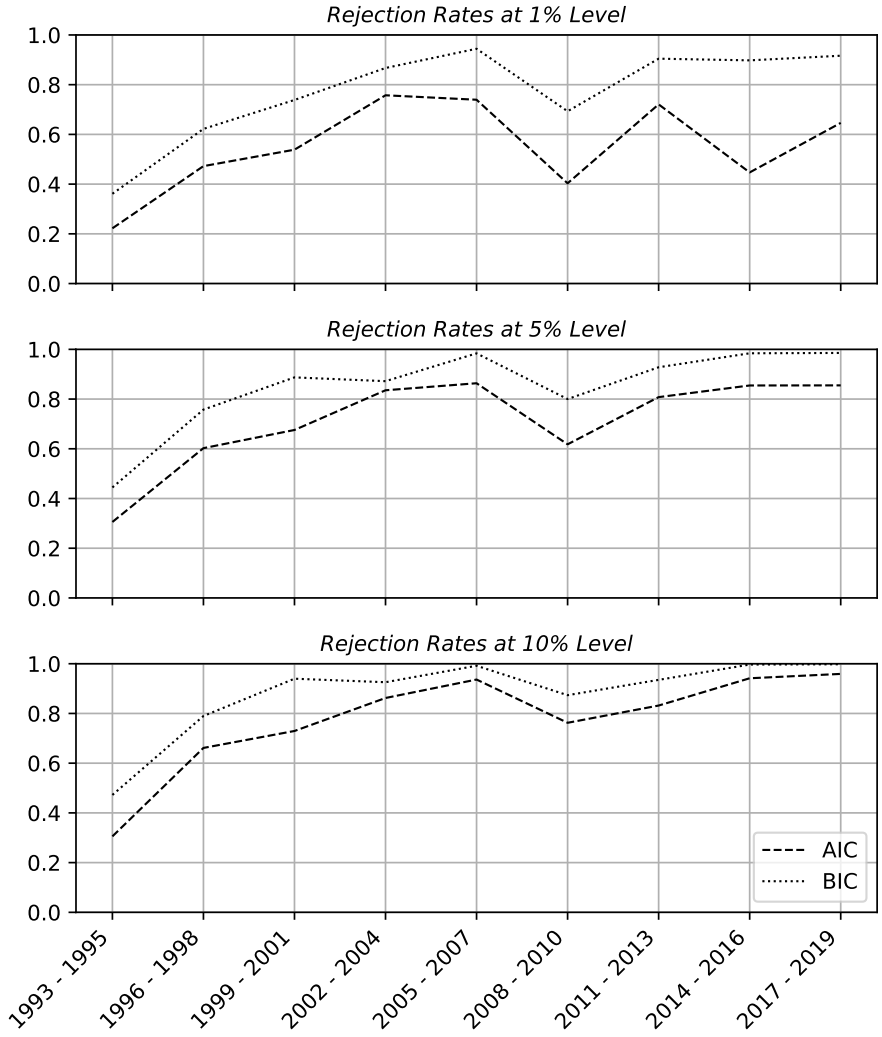
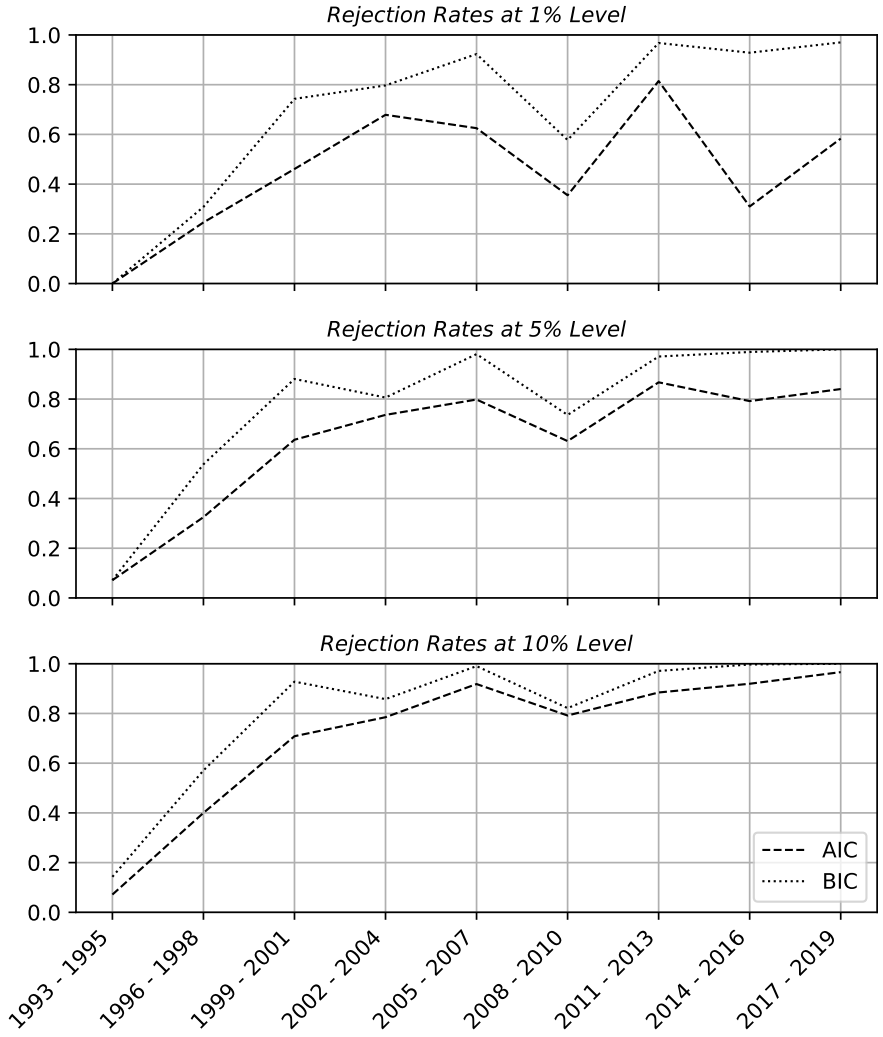


Figure 3.3: Proportion of Stationary Price-Gaps (ADF Test)



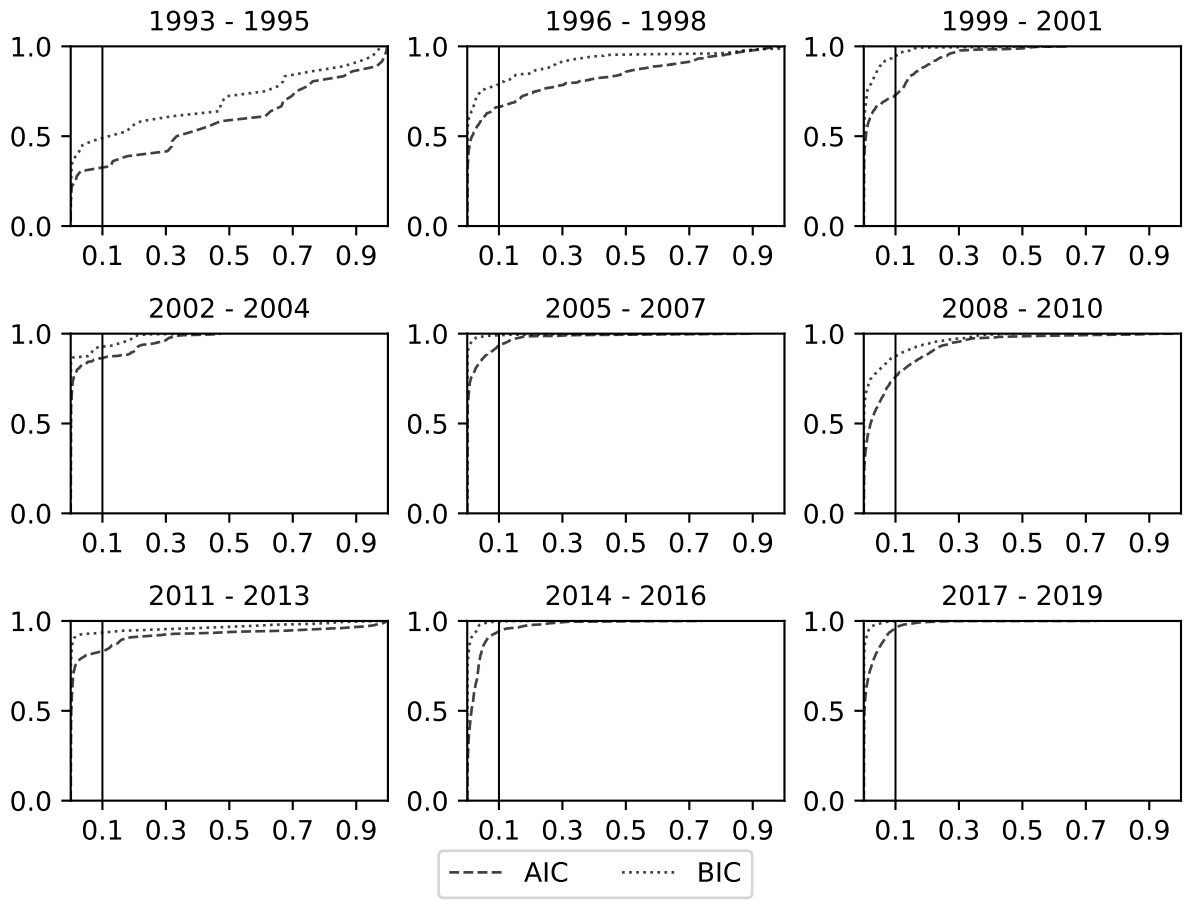
This figure shows the proportion of stationary price-gaps between all combinations of price hub pairs available in each time period. Here I define stationarity as a rejection of the null hypothesis in the ADF test. I report rejection rates at the 1, 5, and 10 percent levels as robustness checks. Four price hubs with stationary price series were excluded from this analysis. Thus, this analysis includes 67 price hubs with price series that are $I(1)$ processes.

Figure 3.4: Proportion of Stationary Western and Other Region Price-Gaps (ADF Test)



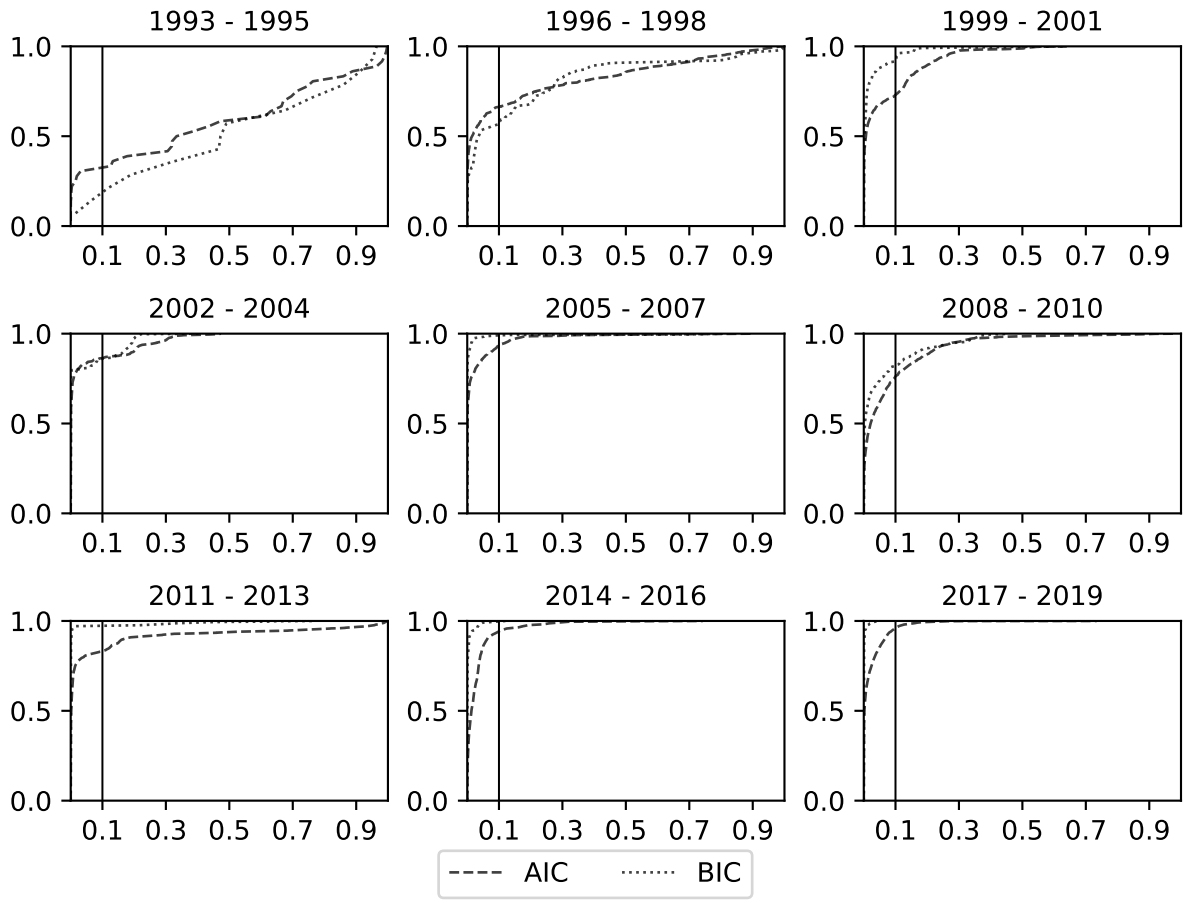
This figure shows the proportion of stationary price-gaps between the price hubs in western and other regions available in each time period. Here I define stationarity as a rejection of the null hypothesis in the ADF test. I report rejection rates at the 1, 5, and 10 percent levels as robustness checks. Four price hubs with stationary price series were excluded from this analysis. Thus, this analysis includes 67 price hubs with price series that are $I(1)$ processes.

Figure 3.5: Empirical Distribution Functions Corresponding to the Distribution of p-values from ADF Tests (All Hubs)



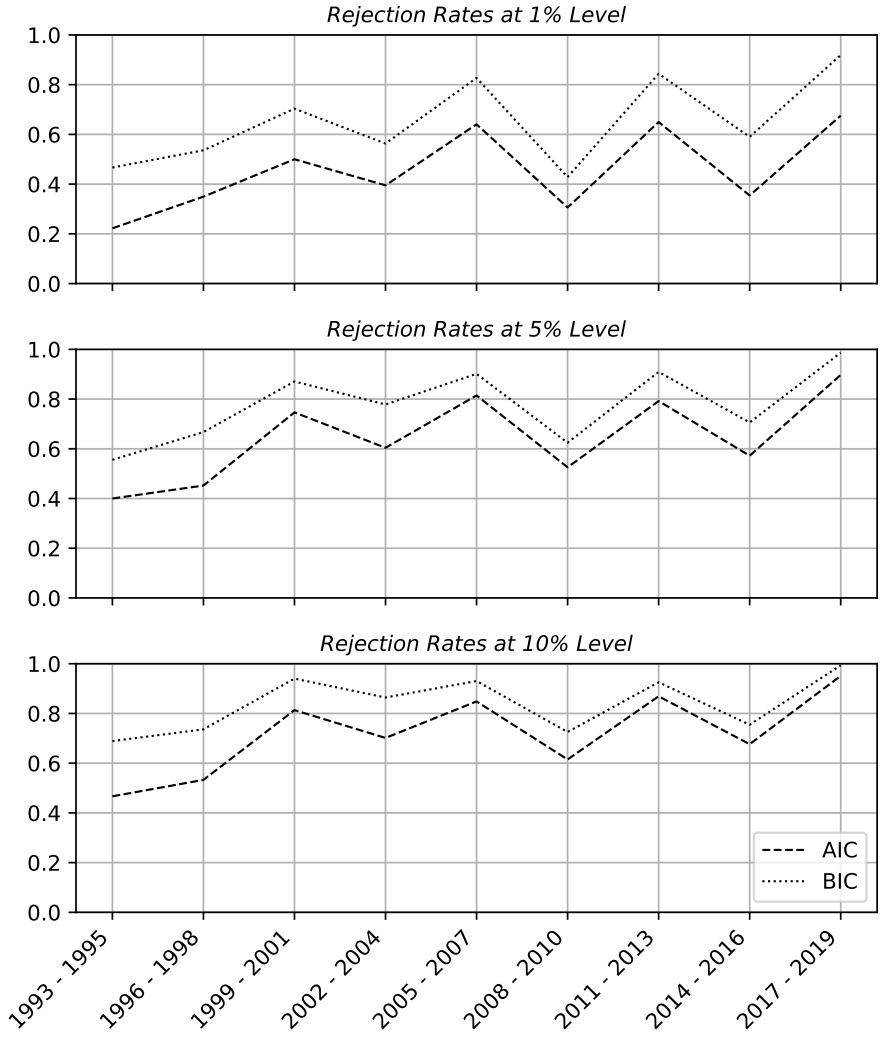
This figure shows the empirical distribution functions of the p-values from the ADF tests shown in Figure 3.3. A vertical line is drawn at 0.1 on the horizontal axis.

Figure 3.6: Empirical Distribution Functions Corresponding to the Distribution of p-values from ADF Tests (Western and Other Region)



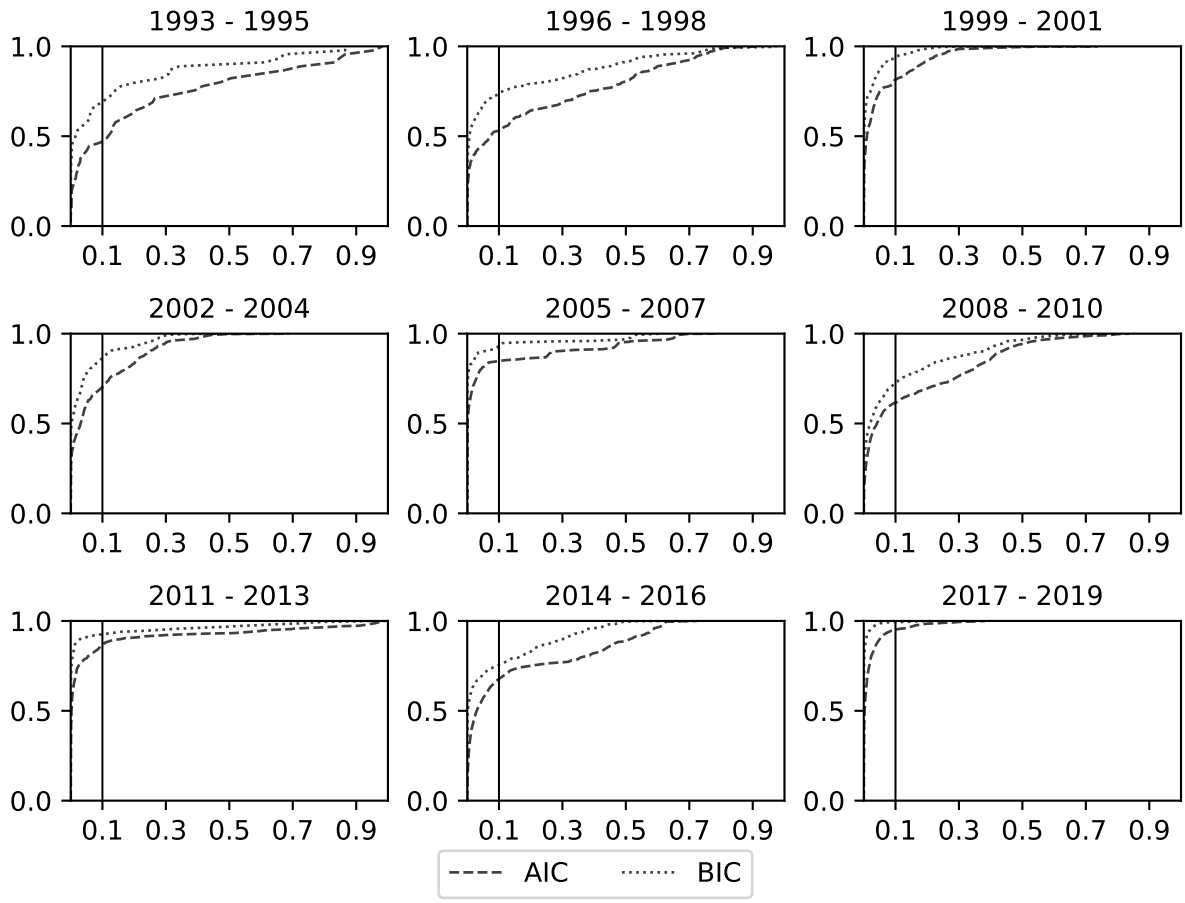
This figure shows the empirical distribution functions of the p-values from the ADF tests shown in Figure 3.4. A vertical line is drawn at 0.1 on the horizontal axis.

Figure 3.7: Proportion of Stationary Price-Gaps (ADF-GLS Test)



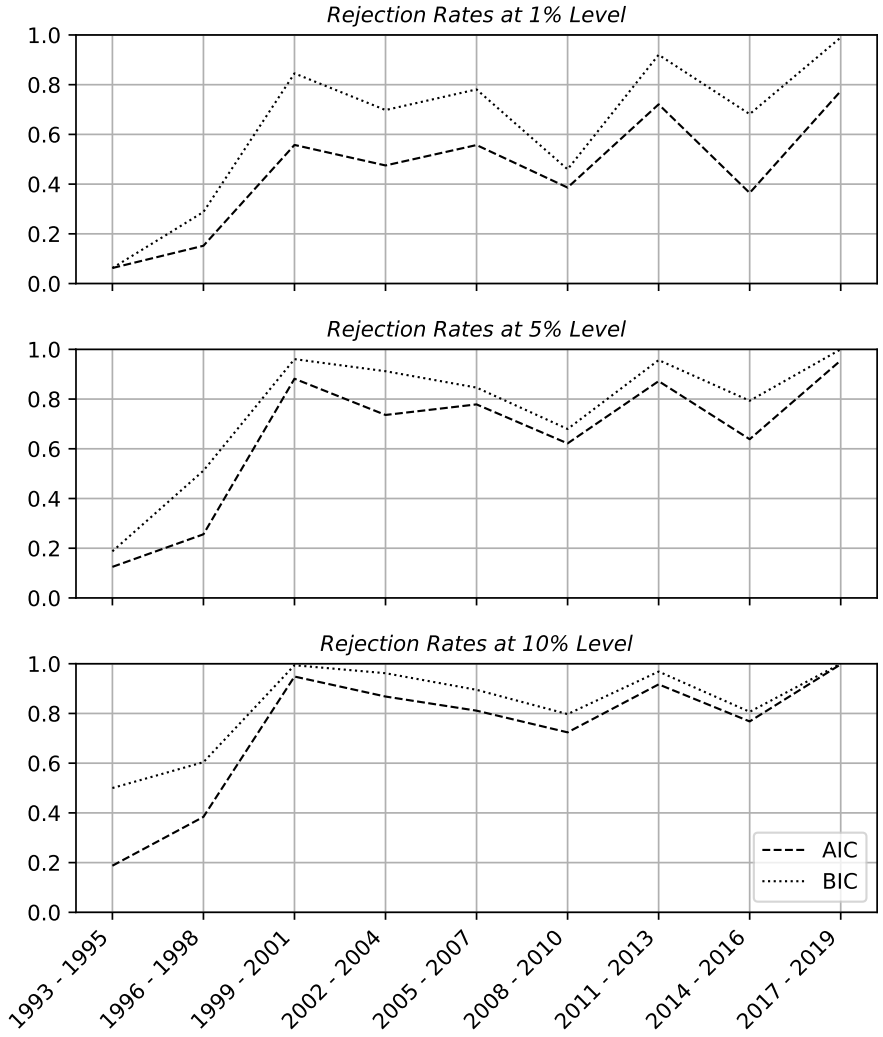
This figure shows the proportion of stationary price-gaps between all combinations of price hub pairs available in each time period. Here I define stationarity as a rejection of the null hypothesis in the ADF-GLS test. I report rejection rates at the 1, 5, and 10 percent levels as robustness checks. Four price hubs with stationary price series were excluded from this analysis. Thus, this analysis includes 67 price hubs with price series that are $I(1)$ processes.

Figure 3.8: Empirical Distribution Functions Corresponding to the Distribution of p-values from ADF-GLS Tests (All Hubs)



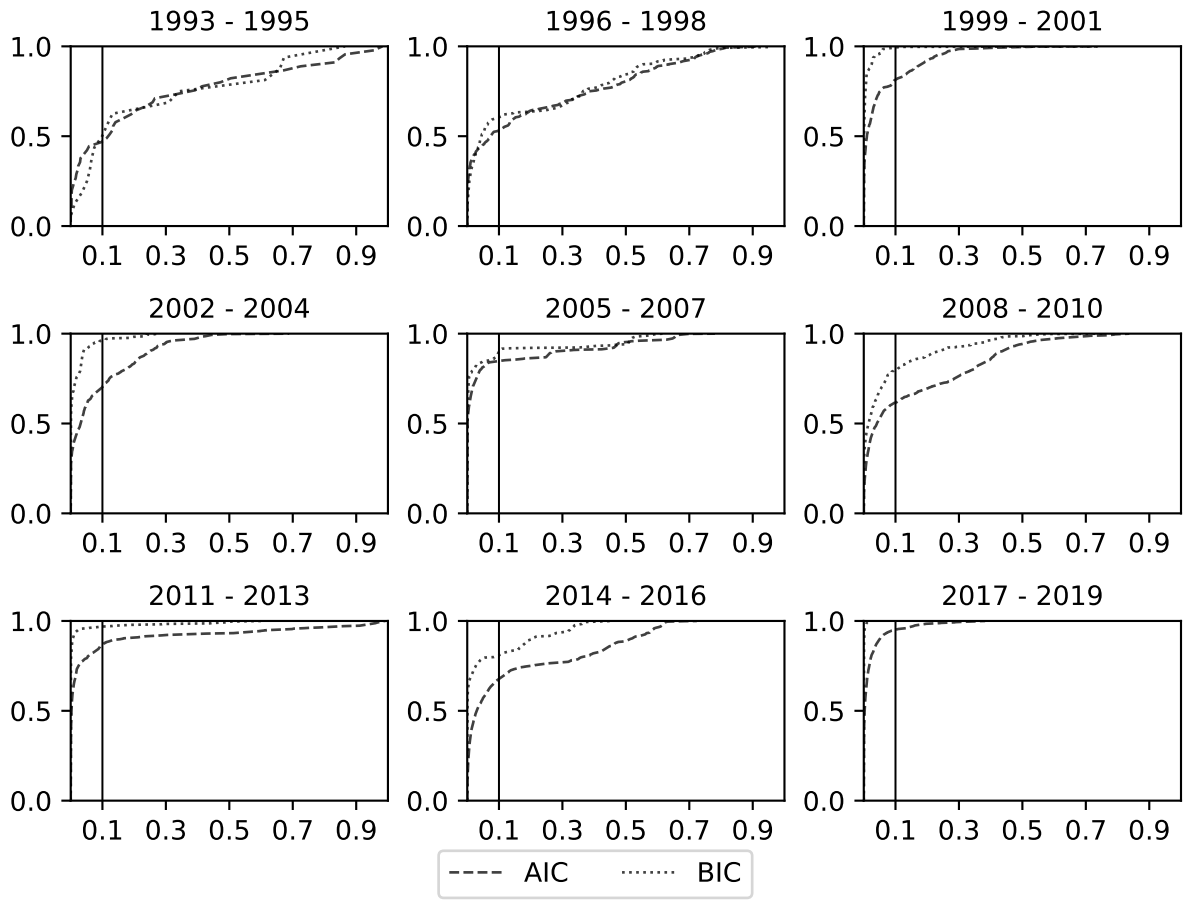
This figure shows the empirical distribution functions of the p-values from the ADF-GLS tests shown in Figure 3.7. A vertical line is drawn at 0.1 on the horizontal axis.

Figure 3.9: Proportion of Stationary Western and Other Region Price-Gaps (ADF-GLS Test)



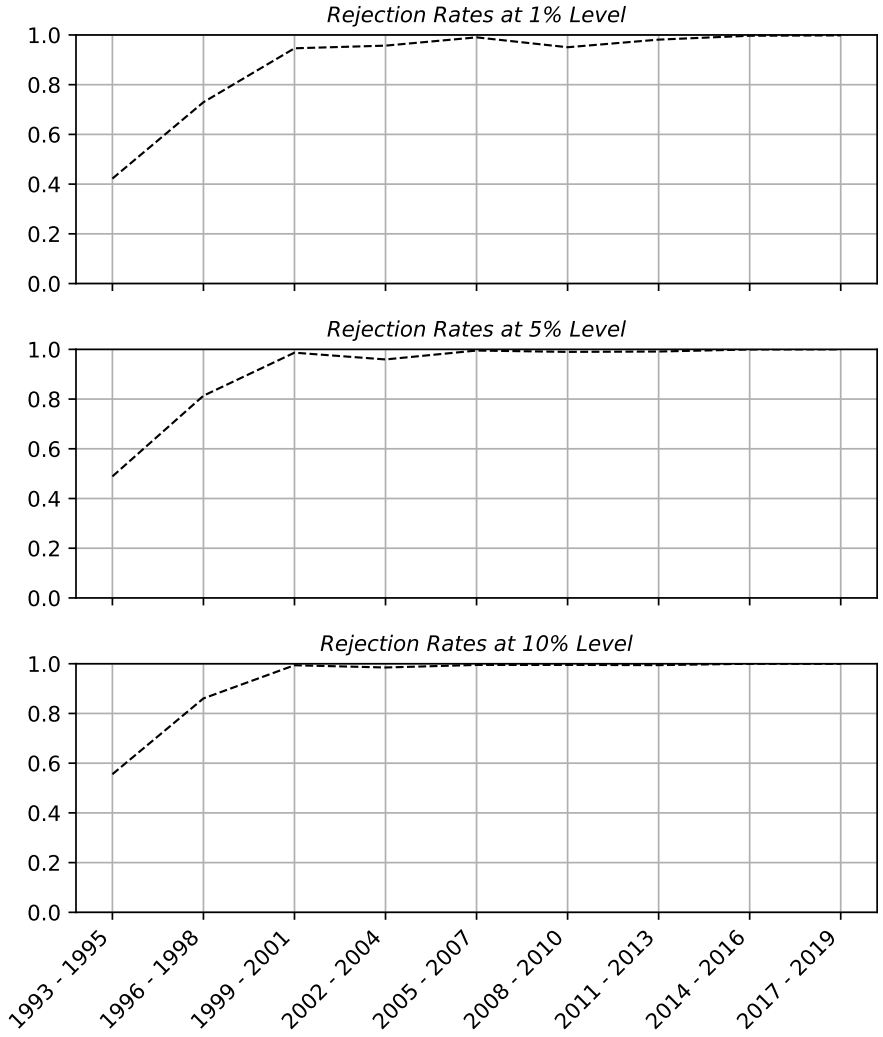
This figure shows the proportion of stationary price-gaps between the price hubs in western and other regions available in each time period. Here I define stationarity as a rejection of the null hypothesis in the ADF-GLS test. I report rejection rates at the 1, 5, and 10 percent levels as robustness checks. Four price hubs with stationary price series were excluded from this analysis. Thus, this analysis includes 67 price hubs with price series that are $I(1)$ processes.

Figure 3.10: Empirical Distribution Functions Corresponding to the Distribution of p-values from ADF-GLS Tests (Western and Other Region)



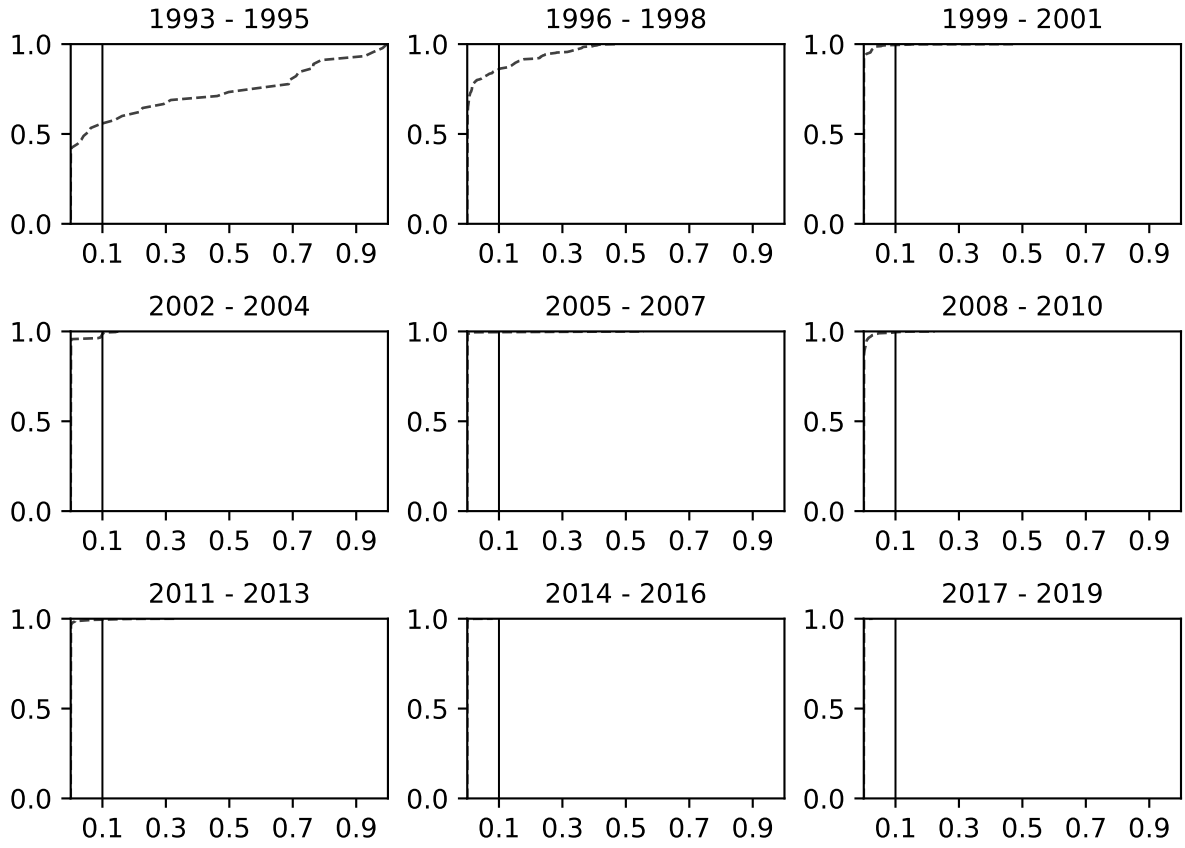
This figure shows the empirical distribution functions of the p-values from the AD-GLS tests shown in Figure 3.9. A vertical line is drawn at 0.1 on the horizontal axis.

Figure 3.11: Proportion of Stationary Price-Gaps (Phillips-Perron Test)



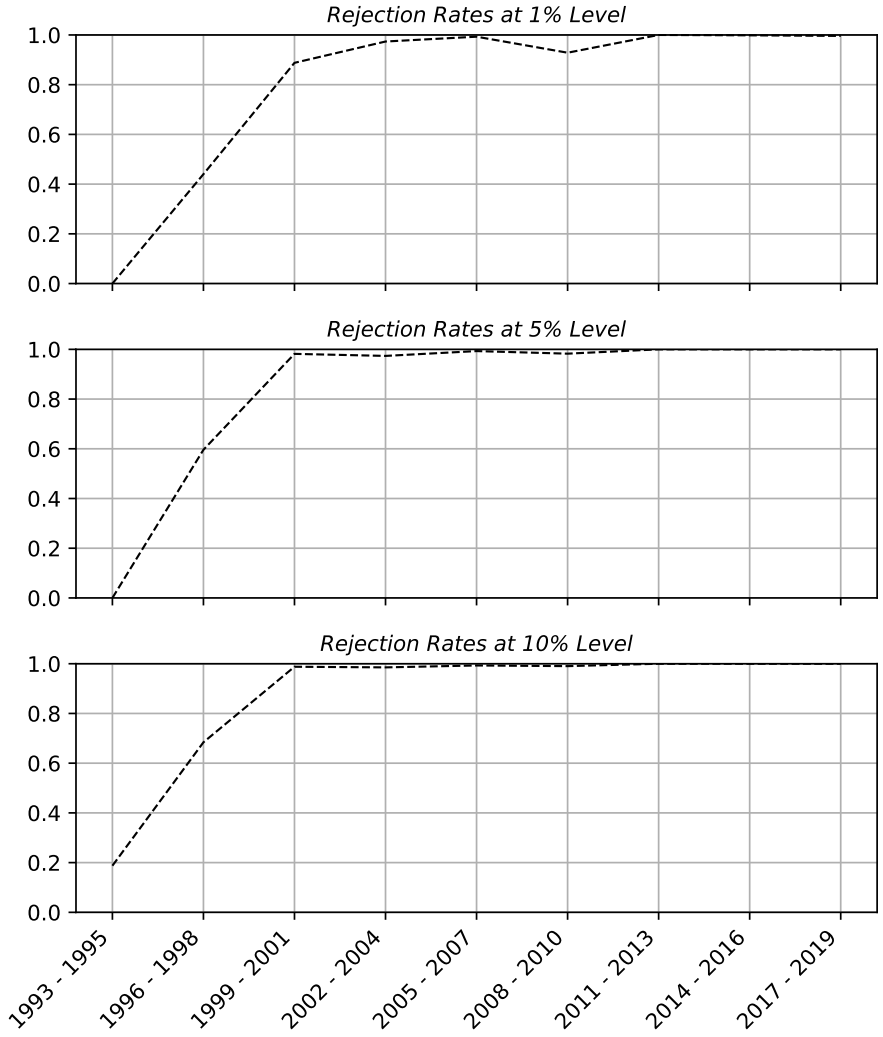
This figure shows the proportion of stationary price-gaps between all combinations of price hub pairs available in each time period. Here I define stationarity as a rejection of the null hypothesis in the Phillips-Perron test. I report rejection rates at the 1, 5, and 10 percent levels as robustness checks. Four price hubs with stationary price series were excluded from this analysis. Thus, this analysis includes 67 price hubs with price series that are $I(1)$ processes.

Figure 3.12: Empirical Distribution Functions Corresponding to the Distribution of p-values from Phillips-Perron Tests (All Hubs)



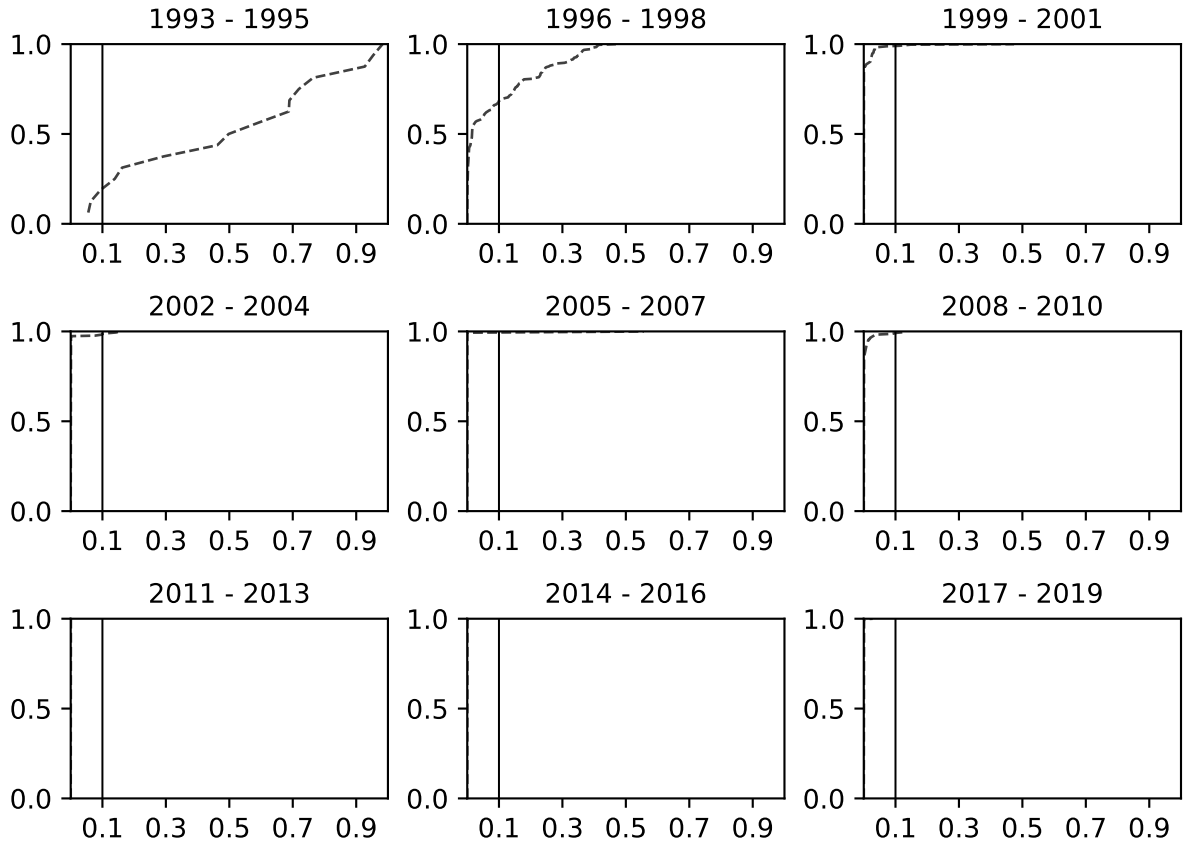
This figure shows the empirical distribution functions of the p-values from the Phillips-Perron tests shown in Figure 3.11. A vertical line is drawn at 0.1 on the horizontal axis. The empirical distribution functions for the periods during 2005–2007, 2011–2013, 2014–2016, and 2017–2019 are not visible as they closely overlap the vertical axis is each plot. This that nearly all bilateral price gaps in the sample are stationary when using a Phillips-Perron test.

Figure 3.13: Proportion of Stationary Western and Other Region Price-Gaps (Phillips-Perron Test)



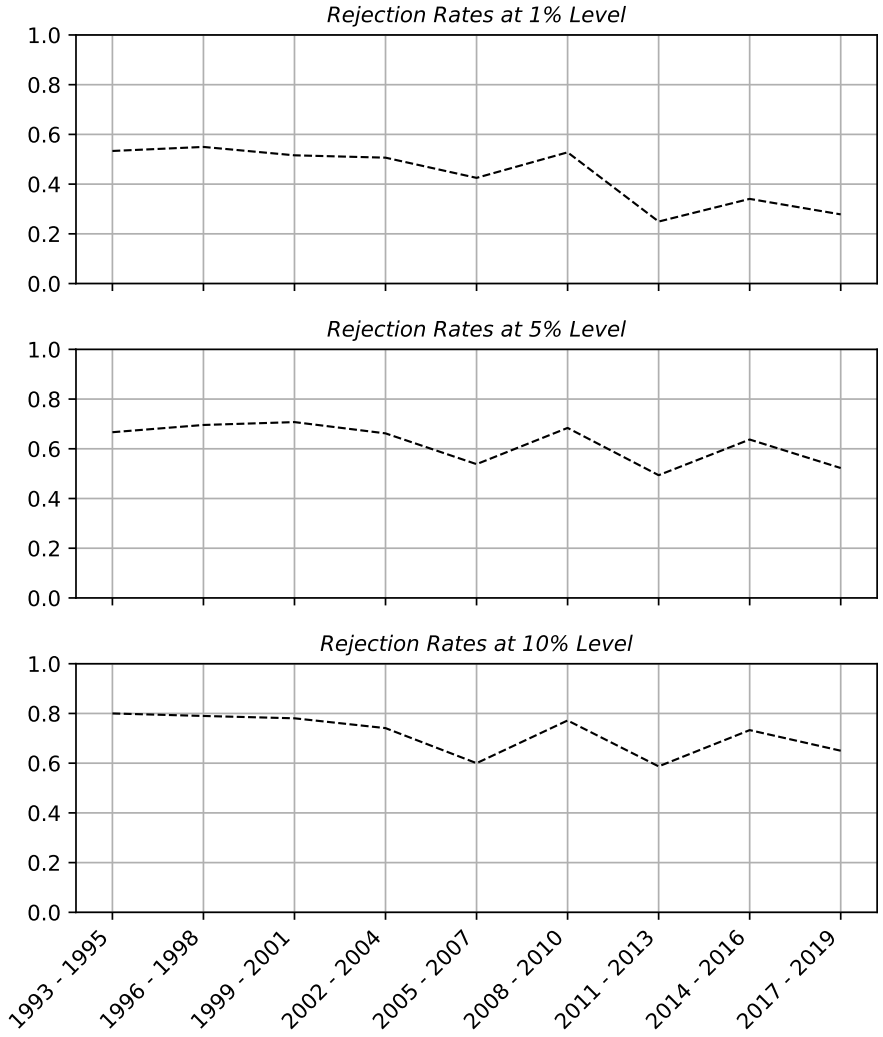
This figure shows the proportion of stationary price-gaps between the price hubs in western and other regions available in each time period. Here I define stationarity as a rejection of the null hypothesis in the Phillips-Perron test. I report rejection rates at the 1, 5, and 10 percent levels as robustness checks. Four price hubs with stationary price series were excluded from this analysis. Thus, this analysis includes 67 price hubs with price series that are $I(1)$ processes.

Figure 3.14: Empirical Distribution Functions Corresponding to the Distribution of p-values from Phillips-Perron Tests (Western and Other Region)



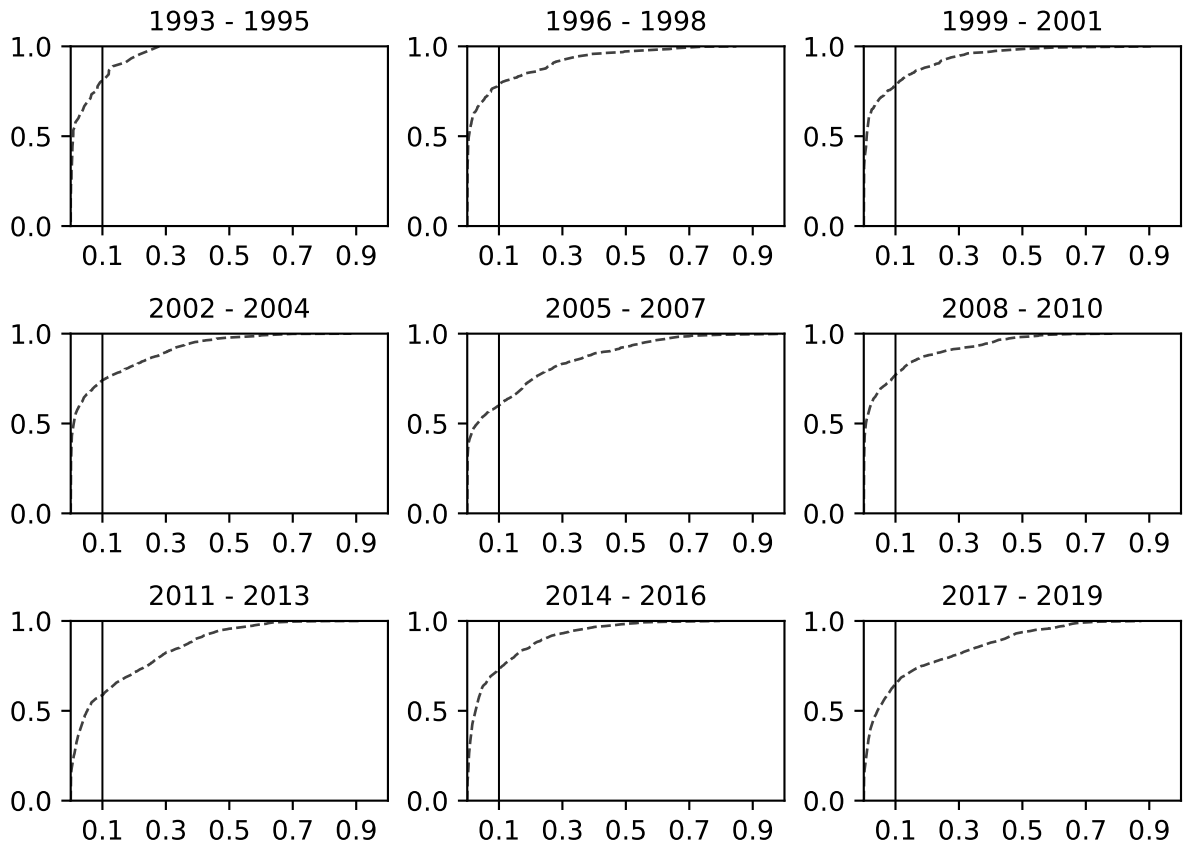
This figure shows the empirical distribution function of the p-values from the Phillips-Perron tests shown in Figure 3.13. A vertical line is drawn at 0.1 on the horizontal axis. The empirical distribution functions for the periods during 2005–2007, 2011–2013, 2014–2016, and 2017–2019 are not visible as they closely overlap the vertical axis is each plot. This that nearly all bilateral price gaps in the sample are stationary when using a Phillips-Perron test.

Figure 3.15: Proportion of Non-Stationary Price-Gaps (KPSS Test)



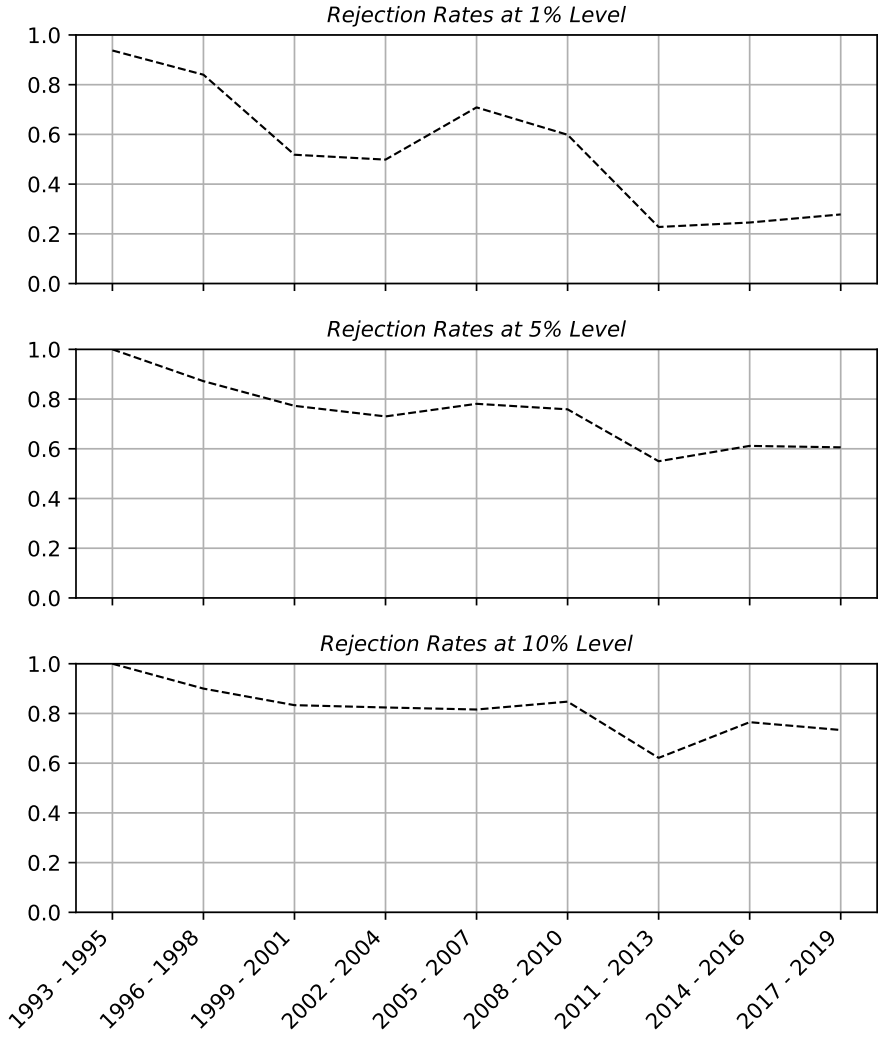
This figure shows the proportion of non-stationary price-gaps between all combinations of price hub pairs available in each time period. Here I define stationarity as a rejection of the null hypothesis in the KPSS test. I report rejection rates at the 1, 5, and 10 percent levels as robustness checks. Four price hubs with stationary price series were excluded from this analysis. Thus, this analysis includes 67 price hubs with price series that are $I(1)$ processes.

Figure 3.16: Empirical Distribution Functions Corresponding to the Distribution of p-values from KPSS Tests (All Hubs)



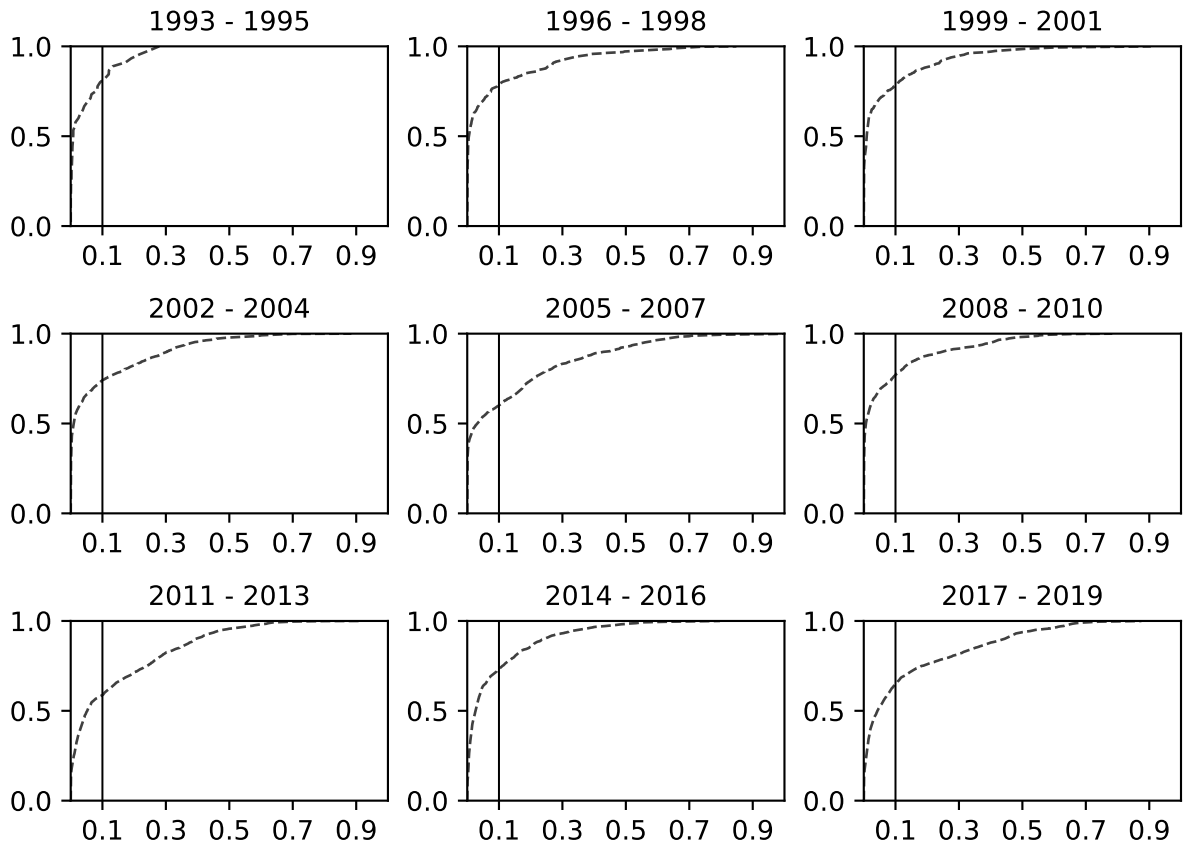
This figure shows the empirical distribution function of the p-values from the KPSS tests shown in Figure 3.15. A vertical line is drawn at 0.1 on the horizontal axis.

Figure 3.17: Proportion of Stationary Western and Other Region Price-Gaps (KPSS Test)



This figure shows the proportion of non-stationary price-gaps between the price hubs in western and other regions available in each time period. Here I define stationarity as a rejection of the null hypothesis in the KPSS test. I report rejection rates at the 1, 5, and 10 percent levels as robustness checks. Four price hubs with stationary price series were excluded from this analysis. Thus, this analysis includes 67 price hubs with price series that are $I(1)$ processes.

Figure 3.18: Empirical Distribution Functions Corresponding to the Distribution of p-values from KPSS Tests (Western and Other Region)



This figure shows the empirical distribution function of the p-values from the KPSS tests shown in Figure 3.17. A vertical line is drawn at 0.1 on the horizontal axis.

Appendices

Appendix A

Estimates Excluding Pipeline Entry and Exit

To remove the effects of pipeline entry and exit on mean technical efficiency and productivity, I limit the sample to only pipelines that are observed in every production year in the sample (old pipelines).[1] Both pipeline entry and exit can positively affect mean technical efficiency and productivity. More specifically, this will affect the proportion of pipelines in the sample using better technology. While inefficient an unproductive pipelines will exit the market, newer pipelines will utilize better technology and locate where they can exhibit the greatest return on investment, subject to regulatory constraints. Examining old pipelines offers insight on how pipelines in existence before the Shale Revolution may have benefited or not benefited from the expansion of shale gas supply. While one would expect that new pipelines would use new technology to transport gas, Dreskin and Boss (2010) explain that this may not be feasible for older pipelines in a rate regulated environment where approval from FERC for increased rates is required to cover the costs of upgrades. As a result, newer pipelines are more likely to incorporate improved technology to transport gas. Thus, any improvements (or declines) in the technical efficiency of old pipelines removes some of the effect of technological improvements.[2]

[1] I observe 41 pipelines that are present in every year in the sample.

[2] I do not observe the specific upgrades or technical specifications of the pipeline steel, compressor stations, or other operating characteristics of the pipelines. As a result, I cannot limit the sample to pipelines of certain technical specifications and test how these specifications impact efficiency and productivity.

To model the technical efficiency of old pipelines, I use the same estimation procedure outlined in Section 1.2 and the same specified inputs and outputs noted in Section 1.3.1. Moreover, all estimation is done with dimension reduced data, where eigensystem methods described in Wilson (2018) are used to reduce the dimensions of the output matrix. In this application, the $q = 4$ unique outputs of total delivery volume, total horsepower, length, and peak delivery volume are reduced to a single principal component which contains 97.9 percent of the independent linear information contained in the original four columns of the output matrix. Thus, dimension reduction in this application is feasible as little information is lost.

As with the primary analysis, I test for separability with respect to production year, convexity, and returns to scale to guide my choice of estimator. Tables A.1 and A.2 report the results of the tests for separability with respect to production year. Table A.1 reports the test statistics associated with the differences in conditional and unconditional efficiency described in Daraio et al. (2018), while Table A.2 reports the associated results from the KS tests. In this case, both sets of tests provide some evidence against separability with respect to production year. Specifically, the bottom row of both tables report the results of testing separability across all production years simultaneously. It is clear that separability is strongly rejected.

Tables A.3 and A.4 report the results of the tests for convexity of the production set. Table A.3 presents the test statistics corresponding to the differences in mean FDH and VRS-DEA estimates by year, as described in Kneip et al. (2016). While there is only mild evidence against convexity under these sets of tests, with a rejection rate of $7/69 \approx .101$, the corresponding KS test statistics, reported in A.4 provide more evidence against convexity with a rejection rate of $15/69 \approx .217$. Thus, I assume that the production set is non-convex. Moreover, given the strong evidence against returns to scale as shown in Tables A.5 and A.6, I estimate technical efficiency using the FDH estimator conditional on production year.

Table A.7 presents the results of the test for changes in technical efficiency over time for old pipelines. While year-over-year technical efficiency increases and decreases with statistical significance quite frequently, comparing the efficiency of pipelines in 2007, 2008, 2009, 2010, and 2011 to 2018 provides similar results to Table 10 in the primary analysis. More specifically, I observe positive changes in technical efficiency during 2007–2018 and 2008–2018. This suggests that, for old pipelines, technical efficiency was greater in 2018 than in 2007 and 2008. Although I

find negative test statistics associated with technical efficiency change during 2009–2018, 2010–2018, and 2011–2018, these estimates are not statistically significant.

Finally, change in productivity estimates, presented in Table A.8, indicate that productivity increased when comparing mean productivity in 2007, 2008, 2009, 2010, and 2011 to 2018. The changes in mean productivity over all these time periods are statistically significant at the 1 percent level, indicating stronger evidence for productivity improvements for old pipelines over the entire sample of pipelines. In this case, if old pipelines are constrained in adopting better technology to transport gas, as suggested by Dreskin and Boss (2010), these pipelines may have benefited purely from increased throughput as pipeline utilization increased during the Shale Revolution.

Table A.1: Separability Test With Respect to Time (FDH Estimator): Old Pipelines Only

	Input–Orientation		Output–Orientation		Hyperbolic	
Period	Statistic	p-value	Statistic	p-value	Statistic	p-value
1996–1997	1.276	0.384	3.451	0.155	4.120	0.450
1997–1998	1.036	0.509	5.016	0.011**	5.036	0.175
1998–1999	1.123	0.423	4.036	0.043**	6.377	0.014**
1999–2000	1.283	0.451	2.970	0.471	5.458	0.091*
2000–2001	1.232	0.440	4.360	0.148	2.849	0.881
2001–2002	1.231	0.355	3.997	0.095*	3.506	0.532
2002–2003	0.245	0.550	3.967	0.048**	5.042	0.067*
2003–2004	1.876	0.036**	4.226	0.110	8.136	0.000***
2004–2005	0.301	0.804	1.905	0.769	7.839	0.003***
2005–2006	0.939	0.498	1.570	0.837	3.028	0.604
2006–2007	−0.176	0.962	3.040	0.319	5.005	0.182
2007–2008	1.614	0.281	1.920	0.832	3.889	0.641
2008–2009	1.891	0.045**	4.891	0.030**	5.847	0.035**
2009–2010	1.690	0.042**	2.694	0.180	0.678	0.970
2010–2011	1.228	0.267	3.248	0.079*	2.622	0.660
2011–2012	3.458	0.002***	1.998	0.718	3.428	0.417
2012–2013	1.767	0.210	1.436	0.782	6.378	0.027**
2013–2014	2.691	0.020**	5.840	0.004***	3.404	0.711
2014–2015	1.250	0.265	−0.268	0.961	5.328	0.072*
2015–2016	1.511	0.183	2.036	0.410	4.198	0.242
2016–2017	0.839	0.534	0.842	0.845	3.811	0.392
2017–2018	1.132	0.511	−0.061	0.942	1.612	0.895
[96, 00]–[01–06]	0.080	0.315	5.727	0.043**	11.003	0.011**
[01, 06]–[07–12]	1.938	0.014**	11.741	0.000***	15.343	0.000***
[07, 12]–[13–18]	2.796	0.000***	12.095	0.000***	2.600	0.978
[96, 00], [01, 06], [07, 12], [13, 18]	7.980	0.000***	7.250	0.000***	8.620	0.000***
96, 97,...,18	14.900	0.000***	18.100	0.000***	17.100	0.000***

The test statistic, described in Daraio et al. (2018), is computed using the FDH estimator with dimension-reduced data, such that $p = q = 1$. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 1996–2018. The p-values are developed using 100 sample-splits and 1,000 bootstrap replications using techniques described in Simar and Wilson (2020). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table A.2: Separability Test (KS Test) With Respect to Time (FDH Estimator): Old Pipelines Only

	Input–Orientation		Output–Orientation		Hyperbolic	
Period	Statistic	p-value	Statistic	p-value	Statistic	p-value
1996–1997	0.530	0.353	0.550	0.612	0.799	0.471
1997–1998	0.532	0.330	0.800	0.066*	0.600	0.876
1998–1999	0.478	0.448	0.684	0.231	0.800	0.332
1999–2000	0.521	0.415	0.646	0.522	0.906	0.093*
2000–2001	0.436	0.562	0.784	0.224	0.599	0.906
2001–2002	0.393	0.588	0.500	0.691	0.700	0.573
2002–2003	0.379	0.329	0.764	0.095*	0.776	0.338
2003–2004	0.460	0.224	0.600	0.503	0.850	0.193
2004–2005	0.347	0.694	0.498	0.822	0.900	0.090*
2005–2006	0.548	0.246	0.500	0.801	0.556	0.871
2006–2007	0.335	0.749	0.623	0.471	0.700	0.644
2007–2008	0.393	0.665	0.502	0.866	0.700	0.671
2008–2009	0.666	0.034**	0.596	0.558	0.600	0.811
2009–2010	0.556	0.118	0.688	0.157	0.365	0.969
2010–2011	0.594	0.141	0.528	0.467	0.672	0.579
2011–2012	0.794	0.034**	0.497	0.830	0.656	0.653
2012–2013	0.577	0.305	0.380	0.937	0.975	0.067*
2013–2014	0.672	0.106	0.740	0.245	0.666	0.839
2014–2015	0.473	0.358	0.260	0.979	0.700	0.607
2015–2016	0.551	0.237	0.596	0.380	0.690	0.607
2016–2017	0.440	0.458	0.558	0.514	0.827	0.276
2017–2018	0.554	0.331	0.372	0.863	0.461	0.921
[96, 00]–[01–06]	0.367	0.341	0.474	0.935	1.000	0.006***
[01, 06]–[07–12]	0.421	0.276	0.822	0.267	0.895	0.508
[07, 12]–[13–18]	0.395	0.311	0.800	0.197	0.495	0.995
[96, 00], [01, 06], [07, 12], [13, 18]	0.900	0.008***	0.700	0.138	0.900	0.002***
96, 97,...,18	0.999	0.000***	0.969	0.000***	0.900	0.000***

The test statistic is from the KS test as described in Simar and Wilson (2020). In this case, I use the KS test to test for uniformity of the p-values estimated from the multiple sample-splits used in the tests presented in Table A.1. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 1996–2018. The p-values are developed using 100 sample-splits and 1,000 bootstrap replications using techniques described in Simar and Wilson (2020). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table A.3: Convexity Test: Old Pipelines Only

	Input–Orientation		Output–Orientation		Hyperbolic	
Year	Statistic	p-value	Statistic	p-value	Statistic	p-value
1996	−1.693	0.997	0.358	0.292	−0.801	0.910
1997	−0.957	0.867	0.000	0.418	−1.199	0.918
1998	0.592	0.125	−0.566	0.677	−2.953	0.999
1999	−1.925	0.962	−0.044	0.310	0.791	0.054*
2000	0.080	0.229	−0.324	0.448	−0.924	0.720
2001	−2.252	0.960	−2.433	0.969	0.846	0.048**
2002	−0.790	0.767	−1.593	0.963	−1.174	0.901
2003	−0.726	0.783	−1.288	0.939	−2.641	0.999
2004	−0.674	0.804	−1.301	0.923	0.583	0.237
2005	−2.521	0.993	−0.542	0.711	−1.713	0.968
2006	−1.032	0.867	−1.319	0.912	0.348	0.296
2007	−0.020	0.461	−1.211	0.901	0.479	0.283
2008	−0.748	0.806	−4.162	0.999	−1.424	0.941
2009	1.234	0.166	1.398	0.101	0.995	0.250
2010	1.101	0.135	−0.450	0.839	0.843	0.213
2011	1.379	0.152	2.219	0.022**	0.014	0.816
2012	−0.557	0.919	1.617	0.099*	1.545	0.098*
2013	0.629	0.224	−1.572	0.989	−0.873	0.910
2014	−0.775	0.783	0.105	0.324	−1.256	0.915
2015	1.407	0.044**	0.122	0.507	−0.081	0.649
2016	0.901	0.269	0.713	0.367	1.184	0.152
2017	1.355	0.087*	0.476	0.516	1.190	0.135
2018	−0.164	0.779	0.329	0.528	1.122	0.163

The test statistic, described in Kneip et al. (2016), is computed using dimension-reduced data, such that $p = q = 1$. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 1996–2018. The p-values are developed using 100 sample-splits and 1,000 bootstrap replications using techniques described in Simar and Wilson (2020). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table A.4: Convexity Test (KS Test): Old Pipelines Only

	Input–Orientation		Output–Orientation		Hyperbolic	
Year	Statistic	p-value	Statistic	p-value	Statistic	p-value
1996	0.395	0.166	0.277	0.484	0.264	0.497
1997	0.295	0.476	0.141	0.957	0.379	0.232
1998	0.452	0.124	0.310	0.446	0.596	0.018**
1999	0.392	0.260	0.282	0.578	0.526	0.062*
2000	0.316	0.527	0.252	0.708	0.333	0.455
2001	0.378	0.353	0.540	0.093*	0.579	0.056*
2002	0.500	0.065*	0.535	0.038**	0.453	0.086*
2003	0.385	0.173	0.354	0.219	0.585	0.013**
2004	0.389	0.333	0.393	0.314	0.499	0.111
2005	0.593	0.027**	0.381	0.263	0.560	0.038**
2006	0.428	0.212	0.353	0.387	0.300	0.531
2007	0.418	0.244	0.300	0.569	0.470	0.175
2008	0.454	0.212	0.798	0.004***	0.481	0.179
2009	0.519	0.174	0.438	0.292	0.615	0.073*
2010	0.478	0.172	0.304	0.559	0.406	0.291
2011	0.595	0.096*	0.493	0.218	0.159	0.967
2012	0.296	0.705	0.522	0.191	0.593	0.106
2013	0.270	0.580	0.541	0.038**	0.460	0.134
2014	0.300	0.459	0.306	0.449	0.269	0.543
2015	0.535	0.075*	0.288	0.545	0.249	0.678
2016	0.451	0.271	0.458	0.272	0.502	0.171
2017	0.532	0.118	0.366	0.430	0.468	0.211
2018	0.184	0.915	0.426	0.307	0.430	0.293

The test statistic is from the KS test as described in Simar and Wilson (2020). In this case, I use the KS test to test for uniformity of the p-values estimated from the multiple sample-splits used in the tests presented in Table A.3. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 1996–2018. The p-values are developed using 100 sample-splits and 1,000 bootstrap replications using techniques described in Simar and Wilson (2020). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table A.5: Returns to Scale Test: Old Pipelines Only

	Input–Orientation		Output–Orientation		Hyperbolic	
Year	Statistic	p-value	Statistic	p-value	Statistic	p-value
1996	3.943	0.020**	3.872	0.008***	4.030	0.032**
1997	1.789	0.549	5.642	0.002***	4.180	0.087*
1998	2.879	0.167	4.377	0.008***	4.653	0.034**
1999	1.276	0.635	3.231	0.041**	4.207	0.080*
2000	1.875	0.448	3.845	0.010***	3.864	0.088*
2001	1.470	0.459	2.660	0.047**	2.454	0.211
2002	1.491	0.498	0.423	0.719	1.842	0.411
2003	2.654	0.194	2.205	0.144	3.075	0.174
2004	0.995	0.636	1.932	0.130	1.278	0.494
2005	2.021	0.253	3.225	0.019**	2.155	0.227
2006	0.580	0.758	2.456	0.050**	1.435	0.283
2007	1.503	0.489	2.752	0.068*	1.443	0.546
2008	1.987	0.375	3.850	0.018**	1.979	0.459
2009	1.718	0.135	1.991	0.067*	2.858	0.014**
2010	0.775	0.518	2.213	0.014**	0.737	0.556
2011	1.620	0.279	1.449	0.281	2.164	0.130
2012	1.804	0.443	3.757	0.032**	5.422	0.008***
2013	0.886	0.688	2.900	0.025**	3.654	0.037**
2014	3.080	0.091*	3.200	0.033**	3.935	0.059*
2015	2.074	0.147	3.179	0.009***	2.991	0.048**
2016	1.458	0.196	3.357	0.005***	2.722	0.021**
2017	1.552	0.278	2.691	0.029**	3.527	0.008***
2018	2.205	0.098*	2.656	0.027**	1.928	0.187

The test statistic, described in Kneip et al. (2016), is computed using dimension-reduced data, such that $p = q = 1$. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 1996–2018. The p-values are developed using 100 sample-splits and 1,000 bootstrap replications using techniques described in Simar and Wilson (2020). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table A.6: Returns to Scale Test (KS Test): Old Pipelines Only

	Input–Orientation		Output–Orientation		Hyperbolic	
Year	Statistic	p-value	Statistic	p-value	Statistic	p-value
1996	0.892	0.033**	0.759	0.087*	0.799	0.081*
1997	0.525	0.655	0.900	0.012**	0.899	0.026**
1998	0.763	0.180	0.894	0.013**	0.855	0.069*
1999	0.399	0.820	0.668	0.152	0.790	0.132
2000	0.645	0.434	0.893	0.017**	0.976	0.007***
2001	0.817	0.084*	0.731	0.061*	0.886	0.049**
2002	0.499	0.569	0.300	0.806	0.311	0.850
2003	0.741	0.177	0.493	0.511	0.597	0.463
2004	0.410	0.718	0.597	0.240	0.391	0.745
2005	0.594	0.411	0.738	0.080*	0.600	0.357
2006	0.414	0.632	0.660	0.133	0.588	0.269
2007	0.675	0.264	0.688	0.159	0.565	0.507
2008	0.637	0.401	0.976	0.007***	0.528	0.609
2009	0.535	0.320	0.837	0.007***	0.787	0.046**
2010	0.341	0.650	0.774	0.013**	0.341	0.670
2011	0.630	0.283	0.556	0.364	0.776	0.083*
2012	0.564	0.499	0.810	0.045**	0.813	0.062*
2013	0.413	0.708	0.801	0.037**	0.911	0.018**
2014	0.935	0.016**	0.787	0.053*	0.900	0.029**
2015	0.748	0.118	0.768	0.065*	0.834	0.057*
2016	0.470	0.372	0.888	0.006***	0.879	0.018**
2017	0.546	0.379	0.697	0.113	0.889	0.018**
2018	0.687	0.142	0.868	0.016**	0.624	0.264

The test statistic is from the KS test as described in Simar and Wilson (2020). In this case, I use the KS test to test for uniformity of the p-values estimated from the multiple sample-splits used in the tests presented in Table A.5. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 1996–2018. The p-values are developed using 100 sample-splits and 1,000 bootstrap replications using techniques described in Simar and Wilson (2020). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table A.7: Test For Equivalency of Mean Efficiency: FDH Estimator With Old Pipelines Only

	Input–Orientation		Output–Orientation		Hyperbolic	
Period	Statistic	p-value	Statistic	p-value	Statistic	p-value
1996–1997	−3.017	0.003***	−2.509	0.012**	−2.839	0.005***
1997–1998	−0.406	0.685	−0.860	0.390	−1.060	0.289
1998–1999	−0.837	0.403	0.477	0.633	1.282	0.200
1999–2000	−0.280	0.780	−1.111	0.267	−1.140	0.254
2000–2001	−1.347	0.178	−1.789	0.074*	−0.542	0.588
2001–2002	3.182	0.001***	2.852	0.004***	2.136	0.033**
2002–2003	−1.961	0.050**	−2.580	0.010***	−2.008	0.045**
2003–2004	2.746	0.006***	1.998	0.046**	2.574	0.010***
2004–2005	−1.309	0.191	−1.720	0.086*	−2.315	0.021**
2005–2006	1.629	0.103	3.006	0.003***	2.131	0.033**
2006–2007	−1.188	0.235	−1.910	0.056*	0.250	0.802
2007–2008	−1.790	0.073*	−1.476	0.140	−2.380	0.017**
2008–2009	8.973	0.000***	8.000	0.000***	7.190	0.000***
2009–2010	−1.918	0.055*	−2.174	0.030**	−0.879	0.379
2010–2011	−0.803	0.422	0.945	0.344	0.661	0.509
2011–2012	−0.193	0.847	0.015	0.988	−0.802	0.422
2012–2013	−2.149	0.032**	−5.562	0.000***	−3.804	0.000***
2013–2014	−3.222	0.001***	0.245	0.806	−2.372	0.018**
2014–2015	5.029	0.000***	4.940	0.000***	4.713	0.000***
2015–2016	0.750	0.453	0.044	0.965	1.570	0.116
2016–2017	1.484	0.138	0.515	0.607	0.017	0.986
2017–2018	−2.964	0.003***	0.178	0.858	−1.412	0.158
2007 & 2018	3.898	0.000***	5.649	0.000***	3.135	0.002***
2008 & 2018	5.609	0.000***	6.802	0.000***	5.033	0.000***
2009 & 2018	−0.480	0.631	−0.772	0.440	−1.328	0.184
2010 & 2018	−1.372	0.170	1.410	0.159	−0.490	0.624
2011 & 2018	−0.823	0.411	0.595	0.552	−0.808	0.419

The test statistic, described in Kneip et al. (2016), is computed using the FDH estimator with dimension-reduced data, such that $p = q = 1$. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 1996–2018. Since separability tests suggest that separability with respect to T does not hold, the estimator allows for a different frontier every production year. One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table A.8: Productivity Estimates: With Old Pipelines Only

Period	Statistic	p-value
1996–1997	0.971	0.331
1997–1998	0.921	0.357
1998–1999	0.155	0.877
1999–2000	−1.219	0.223
2000–2001	−0.136	0.892
2001–2002	−0.935	0.350
2002–2003	−0.081	0.936
2003–2004	−1.377	0.168
2004–2005	−0.578	0.563
2005–2006	1.101	0.271
2006–2007	−0.687	0.492
2007–2008	0.218	0.827
2008–2009	2.330	0.020**
2009–2010	−1.292	0.196
2010–2011	0.011	0.991
2011–2012	2.618	0.009***
2012–2013	1.214	0.225
2013–2014	0.222	0.824
2014–2015	−0.037	0.971
2015–2016	−0.097	0.923
2016–2017	−0.717	0.474
2017–2018	0.073	0.942
2007 & 2018	5.075	0.000***
2008 & 2018	3.457	0.001***
2009 & 2018	2.824	0.005***
2010 & 2018	3.212	0.001***
2011 & 2018	3.785	0.000***

Productivity tests are conducted with dimension-reduced data ($p = q = 1$). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Appendix B

Estimates Excluding Fuel Costs

To model technical efficiency change without fuel costs as suggested in Jamasb et al. (2008), I specify two models. For both models, pipelines are defined as having $q = 4$ unique outputs of total delivery volume, total horsepower, length, and peak delivery volume. The models differ on the defined inputs. One model has a single $p = 1$ monetary input of OPEX that excludes fuel costs, and the second has $p = 2$ inputs of the physical quantity of fuel used for compressor stations and OPEX excluding fuel costs.

All estimates described below are conducted with dimension reduced data. In this case the eigensystem methods described in Wilson (2018) are also applied to the input matrix of physical quantity of fuel and OPEX excluding fuel costs. In this case the first principal component of the input matrix contains 98.7 percent of the independent linear information from the $p = 2$ columns of the input matrix. Thus, the input matrix with both the physical quantity of fuel and OPEX excluding fuel costs can be replaced with it's first principal component without much information loss.

Tables B.1 through B.4 present the results of the tests for separability. While Tables B.1 and B.3 present the estimates of the test statistics described by Daraio et al. (2018), Tables B.2 and B.4 present the associated KS tests.[1] As with the primary results of this paper, I find significant evidence that separability does not hold with respect to time under both models. This provides

[1] I also report the associated KS test statistics for the tests on convexity and CRS versus VRS as described below.

further evidence that the frontier changes from year to year, and efficiency estimates should be conditional on production year, T.

Tables B.5 through B.8 present the results of the tests for convexity. When quantity of fuel is excluded as an input, the tests for convexity reject the null at a rate of $15/69 \approx 0.217$ when examining the differences between the mean FDH and mean VRS-DEA estimates which I report in Table B.5. However, the associated KS test statistics, reported in Table B.6, show an increased rejection rate of $28/69 \approx 0.405$. Similar results are shown when physical quantity of fuel is included as an additional input as shown in Table B.7 and Table B.8. Given the evidence against convexity under both models, I use the FDH estimator for all efficiency estimates. Moreover, given the results from Wilson (2018), this approach should yield less estimation error.

Tables B.9 and B.12 present the results of the tests for CRS versus VRS. In both models there is strong evidence against CRS. The results of these tests provide additional evidence of pipelines not exhibiting CRS. Although, Jamasb et al. (2008) also include capital costs in their specified TOTEX input, they present CRS-DEA efficiency estimates without evidence to suggest CRS holds. It should be noted that they also present VRS-DEA estimates, which would achieve consistent estimates of efficiency under VRS or CRS. In addition, the VRS-DEA estimator would also achieve the faster CRS convergence rate if the frontier is globally CRS, according to Theorem 1 of Kneip et al. (2016). However, the results of the tests for CRS versus VRS in this analysis suggest that CRS does not hold. Thus, the CRS-DEA estimates from Jamasb et al. (2008) may not be consistent.

Tables B.13 and B.14 present the results of the test for changes in technical efficiency over time. When fuel is not included as a part of operating cost, the changes in estimated technical efficiency does not change rapidly from year to year. Under both models there is strong evidence of technical efficiency drops during 2007–2008 and 2010–2011. In addition, there is evidence of technical efficiency improvements during 2008–2009. The net technical efficiency change estimates in both models suggest that technical efficiency change during 2008–2018 was positive, but was negative from 2009 and 2010 through 2018. These results suggest that the technical efficiency gains shown in Table 10, from the primary analysis of this paper, result from pipelines exploiting a changing cost set over time and expanding operations while facing lower fuel costs.

While the efficiency change estimates differ from the primary model used in this analysis,

the productivity estimates under these additional models are consistent with Table 8 in the primary analysis of this paper. Tables B.15 and B.16 present estimates for changes in productivity. The results of these additional models suggest that the net change in productivity was positive over the years of 2007, 2009, and 2010 through 2018.

Table B.1: Separability Test With Respect to Time (FDH Estimator): Without Quantity of Fuel as Input

	Input–Orientation		Output–Orientation		Hyperbolic	
Period	Statistic	p-value	Statistic	p-value	Statistic	p-value
1996–1997	1.516	0.210	2.481	0.486	1.946	0.910
1997–1998	−1.430	1.000	1.379	0.807	3.094	0.666
1998–1999	0.339	0.867	4.136	0.200	5.118	0.044**
1999–2000	−0.076	0.873	6.320	0.005***	3.980	0.264
2000–2001	0.221	0.681	2.207	0.589	4.711	0.127
2001–2002	0.810	0.559	6.141	0.000***	2.772	0.740
2002–2003	1.825	0.105	5.946	0.005***	4.145	0.371
2003–2004	−0.216	0.953	2.614	0.679	4.125	0.375
2004–2005	0.550	0.813	3.912	0.324	3.372	0.842
2005–2006	0.994	0.540	3.381	0.475	6.744	0.066*
2006–2007	0.095	0.830	3.651	0.168	4.392	0.357
2007–2008	5.006	0.000***	5.100	0.007***	5.626	0.030**
2008–2009	3.381	0.000***	1.942	0.463	5.821	0.006***
2009–2010	−0.093	0.881	3.836	0.077*	3.656	0.347
2010–2011	1.201	0.072*	1.957	0.596	7.104	0.002***
2011–2012	0.503	0.362	2.794	0.193	2.457	0.565
2012–2013	0.152	0.578	1.313	0.607	3.258	0.277
2013–2014	0.302	0.487	4.284	0.009***	3.805	0.147
2014–2015	−0.250	0.845	1.578	0.567	1.384	0.919
2015–2016	0.644	0.225	0.948	0.757	3.256	0.229
2016–2017	0.130	0.551	1.565	0.630	4.874	0.071*
2017–2018	0.295	0.545	2.869	0.170	0.295	0.997
[96, 00]–[01, 06]	1.127	0.048**	0.011	0.993	5.374	0.774
[01, 06]–[07, 12]	0.721	0.156	17.050	0.000***	12.957	0.018**
[07, 12]–[13, 18]	4.461	0.000***	6.926	0.004***	13.386	0.000***
[96, 00], [01, 06], [07, 12], [13, 18]	16.200	0.000***	18.200	0.000***	16.500	0.000***
96, 97,...,18	15.200	0.000***	16.300	0.000***	21.300	0.000***

The test statistic, described in Daraio et al. (2018), is computed using the FDH estimator with dimension-reduced data, such that $p = q = 1$. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 1996–2018. The p-values are developed using 100 sample-splits and 1,000 bootstrap replications using techniques described in Simar and Wilson (2020). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table B.2: Separability Test (KS Test) With Respect to Time (FDH Estimator): Without Quantity of Fuel as Input

	Input–Orientation		Output–Orientation		Hyperbolic	
Period	Statistic	p-value	Statistic	p-value	Statistic	p-value
1996–1997	0.520	0.292	0.677	0.390	0.593	0.850
1997–1998	0.300	0.799	0.534	0.683	0.643	0.734
1998–1999	0.360	0.763	0.800	0.177	0.834	0.191
1999–2000	0.197	0.931	0.822	0.132	0.800	0.249
2000–2001	0.301	0.624	0.499	0.740	0.800	0.258
2001–2002	0.491	0.370	0.800	0.115	0.620	0.776
2002–2003	0.485	0.375	0.800	0.135	0.830	0.323
2003–2004	0.308	0.760	0.515	0.860	0.967	0.051*
2004–2005	0.372	0.720	0.698	0.552	0.706	0.789
2005–2006	0.398	0.645	0.847	0.170	0.928	0.237
2006–2007	0.193	0.928	0.882	0.047**	0.745	0.624
2007–2008	0.777	0.006***	0.586	0.491	0.700	0.442
2008–2009	0.590	0.062*	0.400	0.794	0.773	0.287
2009–2010	0.115	0.991	0.796	0.075*	0.870	0.132
2010–2011	0.532	0.066*	0.400	0.889	0.800	0.253
2011–2012	0.383	0.273	0.599	0.338	0.592	0.655
2012–2013	0.330	0.403	0.458	0.696	0.593	0.622
2013–2014	0.208	0.768	0.838	0.016**	0.694	0.374
2014–2015	0.148	0.949	0.500	0.571	0.583	0.761
2015–2016	0.232	0.644	0.396	0.806	0.492	0.817
2016–2017	0.324	0.383	0.435	0.785	0.799	0.221
2017–2018	0.302	0.487	0.775	0.078*	0.499	0.900
[96, 00]–[01–06]	0.489	0.094*	0.534	0.820	0.798	0.832
[01, 06]–[07–12]	0.331	0.562	0.800	0.344	0.700	0.887
[07, 12]–[13–18]	0.866	0.003***	0.447	0.952	0.780	0.723
[96, 00], [01, 06], [07, 12], [13, 18]	0.700	0.083*	0.600	0.316	0.800	0.007***
96, 97,...,18	0.998	0.000***	0.789	0.000***	0.900	0.000***

The test statistic is from the KS test as described in Simar and Wilson (2020). In this case, I use the KS test to test for uniformity of the p-values estimated from the multiple sample-splits used in the tests presented in Table B.1. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 1996–2018. The p-values are developed using 100 sample-splits and 1,000 bootstrap replications using techniques described in Simar and Wilson (2020). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table B.3: Separability Test With Respect to Time (FDH Estimator): With Quantity of Fuel as Input

	Input–Orientation		Output–Orientation		Hyperbolic	
Period	Statistic	p-value	Statistic	p-value	Statistic	p-value
1996–1997	1.128	0.419	1.659	0.751	1.715	0.936
1997–1998	−0.219	0.956	4.624	0.034**	0.101	1.000
1998–1999	1.348	0.440	4.883	0.081*	2.973	0.621
1999–2000	0.795	0.536	4.239	0.189	5.847	0.008***
2000–2001	1.244	0.182	3.956	0.160	5.763	0.017**
2001–2002	0.486	0.771	1.848	0.740	7.047	0.001***
2002–2003	0.478	0.720	3.438	0.290	4.918	0.196
2003–2004	0.316	0.812	3.688	0.405	3.916	0.459
2004–2005	1.148	0.546	4.353	0.167	2.595	0.939
2005–2006	0.698	0.743	4.073	0.288	4.634	0.554
2006–2007	1.374	0.238	1.910	0.733	5.888	0.047**
2007–2008	−1.549	0.996	5.493	0.003***	4.504	0.112
2008–2009	2.383	0.004***	2.509	0.319	4.733	0.060*
2009–2010	0.945	0.257	2.277	0.467	4.421	0.168
2010–2011	1.905	0.005***	4.850	0.018**	4.276	0.179
2011–2012	0.431	0.372	2.648	0.217	5.429	0.008***
2012–2013	1.345	0.036**	2.676	0.190	5.005	0.010***
2013–2014	0.560	0.316	1.560	0.470	−0.090	0.993
2014–2015	0.318	0.528	3.437	0.056*	5.239	0.013**
2015–2016	−0.330	0.834	1.789	0.432	4.537	0.033**
2016–2017	1.282	0.035**	1.270	0.718	3.642	0.297
2017–2018	0.462	0.376	4.550	0.006***	4.033	0.190
[96, 00]–[01, 06]	−0.938	0.792	6.763	0.006***	14.884	0.000***
[01, 06]–[07, 12]	0.882	0.123	8.618	0.022**	11.644	0.049**
[07, 12]–[13, 18]	4.413	0.000***	10.262	0.000***	18.797	0.000***
[96, 00], [01, 06], [07, 12], [13, 18]	5.015	0.000***	5.477	0.000***	5.286	0.000***
96, 97,...,18	12.600	0.000***	14.000	0.000***	18.900	0.000***

The test statistic, described in Daraio et al. (2018), is computed using the FDH estimator with dimension-reduced data, such that $p = q = 1$. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 1996–2018. The p-values are developed using 100 sample-splits and 1,000 bootstrap replications using techniques described in Simar and Wilson (2020). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table B.4: Separability Test (KS Test) With Respect to Time (FDH Estimator): With Quantity of Fuel as Input

	Input–Orientation		Output–Orientation		Hyperbolic	
Period	Statistic	p-value	Statistic	p-value	Statistic	p-value
1996–1997	0.447	0.470	0.554	0.667	0.576	0.872
1997–1998	0.196	0.935	0.792	0.115	0.368	0.994
1998–1999	0.540	0.387	0.818	0.179	0.481	0.939
1999–2000	0.484	0.350	0.820	0.180	0.861	0.143
2000–2001	0.417	0.370	0.500	0.793	0.900	0.070*
2001–2002	0.343	0.720	0.634	0.501	0.800	0.300
2002–2003	0.315	0.722	0.652	0.508	0.778	0.504
2003–2004	0.504	0.350	0.562	0.834	0.798	0.416
2004–2005	0.410	0.669	0.700	0.497	0.695	0.840
2005–2006	0.392	0.670	0.883	0.113	0.846	0.548
2006–2007	0.426	0.463	0.611	0.566	0.988	0.040**
2007–2008	0.402	0.340	0.600	0.462	0.695	0.465
2008–2009	0.683	0.019**	0.506	0.564	0.822	0.129
2009–2010	0.448	0.276	0.659	0.355	0.899	0.069*
2010–2011	0.713	0.005***	0.699	0.275	0.630	0.724
2011–2012	0.492	0.083*	0.600	0.328	0.900	0.019**
2012–2013	0.617	0.018**	0.636	0.256	0.897	0.027**
2013–2014	0.287	0.507	0.578	0.361	0.391	0.940
2014–2015	0.508	0.099*	0.602	0.271	0.797	0.170
2015–2016	0.248	0.620	0.600	0.274	0.620	0.494
2016–2017	0.524	0.049**	0.472	0.718	0.689	0.512
2017–2018	0.375	0.263	0.772	0.073*	0.865	0.126
[96, 00]–[01–06]	0.287	0.563	0.600	0.650	1.000	0.035**
[01, 06]–[07–12]	0.474	0.181	0.600	0.821	0.689	0.955
[07, 12]–[13–18]	0.706	0.009***	0.600	0.721	0.785	0.737
[96, 00], [01, 06], [07, 12], [13, 18]	0.871	0.001***	0.685	0.180	0.683	0.040**
96, 97,...,18	0.897	0.000***	0.797	0.000***	0.899	0.000***

The test statistic is from the KS test as described in Simar and Wilson (2020). In this case, I use the KS test to test for uniformity of the p-values estimated from the multiple sample-splits used in the tests presented in Table B.3. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 1996–2018. The p-values are developed using 100 sample-splits and 1,000 bootstrap replications using techniques described in Simar and Wilson (2020). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table B.5: Convexity Test: Without Quantity of Fuel as Input

	Input–Orientation		Output–Orientation		Hyperbolic	
Year	Statistic	p-value	Statistic	p-value	Statistic	p-value
1996	−1.084	0.987	0.120	0.265	−0.100	0.437
1997	−1.823	0.991	0.254	0.137	−0.206	0.369
1998	−0.321	0.491	0.300	0.118	0.038	0.242
1999	−1.012	0.956	0.196	0.250	−0.003	0.380
2000	−0.592	0.767	0.889	0.022**	0.262	0.197
2001	−0.629	0.873	−0.383	0.680	−0.641	0.848
2002	−1.060	0.984	0.101	0.427	1.049	0.039**
2003	0.222	0.322	0.456	0.199	0.713	0.095*
2004	−0.045	0.507	0.731	0.072*	0.922	0.034**
2005	−0.187	0.485	0.211	0.213	0.827	0.035**
2006	−0.144	0.442	0.250	0.143	−0.108	0.430
2007	−0.029	0.451	0.241	0.216	0.223	0.231
2008	−3.584	1.000	−0.767	0.692	−1.850	0.991
2009	0.122	0.496	0.734	0.055*	0.895	0.032**
2010	0.792	0.060*	1.275	0.008***	1.868	0.001***
2011	−1.530	0.999	0.198	0.120	−0.302	0.558
2012	−2.082	1.000	0.117	0.147	−1.670	0.999
2013	−1.661	0.995	0.861	0.004***	−0.638	0.721
2014	−0.786	0.832	0.991	0.002***	−0.532	0.677
2015	−1.505	0.979	0.311	0.055*	−0.211	0.313
2016	−2.360	1.000	0.400	0.063*	−0.667	0.813
2017	−0.576	0.920	0.060	0.366	−0.150	0.592
2018	−1.249	0.994	−0.042	0.401	−0.588	0.864

The test statistic, described in Kneip et al. (2016), is computed using dimension-reduced data, such that $p = q = 1$. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 1996–2018. The p-values are developed using 100 sample-splits and 1,000 bootstrap replications using techniques described in Simar and Wilson (2020). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table B.6: Convexity Test (KS Test): Without Quantity of Fuel as Input

	Input–Orientation		Output–Orientation		Hyperbolic	
Year	Statistic	p-value	Statistic	p-value	Statistic	p-value
1996	0.452	0.007***	0.140	0.863	0.218	0.391
1997	0.453	0.024**	0.258	0.417	0.247	0.436
1998	0.278	0.308	0.238	0.450	0.261	0.362
1999	0.350	0.082*	0.214	0.434	0.208	0.479
2000	0.338	0.148	0.329	0.169	0.306	0.206
2001	0.275	0.147	0.235	0.269	0.249	0.240
2002	0.419	0.030**	0.175	0.723	0.359	0.092*
2003	0.263	0.308	0.229	0.412	0.355	0.086*
2004	0.235	0.392	0.251	0.290	0.383	0.043**
2005	0.307	0.189	0.172	0.753	0.396	0.053*
2006	0.190	0.552	0.193	0.531	0.166	0.672
2007	0.273	0.167	0.214	0.344	0.233	0.300
2008	0.522	0.013**	0.372	0.172	0.538	0.012**
2009	0.201	0.371	0.310	0.064*	0.338	0.045**
2010	0.366	0.030**	0.469	0.005***	0.584	0.001***
2011	0.478	0.003***	0.174	0.522	0.329	0.068*
2012	0.613	0.000***	0.191	0.477	0.595	0.000***
2013	0.492	0.010***	0.327	0.144	0.385	0.090*
2014	0.383	0.055*	0.322	0.127	0.369	0.081*
2015	0.534	0.002***	0.195	0.609	0.315	0.252
2016	0.620	0.001***	0.193	0.510	0.357	0.092*
2017	0.262	0.136	0.117	0.835	0.171	0.474
2018	0.392	0.011**	0.087	0.965	0.318	0.060*

The test statistic is from the KS test as described in Simar and Wilson (2020). In this case, I use the KS test to test for uniformity of the p-values estimated from the multiple sample-splits used in the tests presented in Table B.5. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 1996–2018. The p-values are developed using 100 sample-splits and 1,000 bootstrap replications using techniques described in Simar and Wilson (2020). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table B.7: Convexity Test: With Quantity of Fuel as Input

	Input–Orientation		Output–Orientation		Hyperbolic	
Year	Statistic	p-value	Statistic	p-value	Statistic	p-value
1996	−0.926	0.973	0.172	0.192	0.066	0.305
1997	−1.024	0.892	−0.160	0.384	−0.502	0.635
1998	−0.646	0.784	0.218	0.162	−0.495	0.697
1999	−1.081	0.986	0.128	0.325	0.566	0.081*
2000	−0.846	0.867	0.226	0.229	0.237	0.217
2001	−1.523	0.996	−0.164	0.471	−0.676	0.854
2002	−0.901	0.971	0.561	0.178	0.686	0.116
2003	−0.105	0.630	0.344	0.294	0.959	0.059*
2004	−0.015	0.499	0.502	0.162	0.525	0.150
2005	−0.449	0.704	0.134	0.254	0.677	0.054*
2006	0.117	0.246	−0.032	0.356	−0.210	0.518
2007	−0.141	0.585	0.002	0.450	0.374	0.176
2008	−3.104	1.000	−0.525	0.515	−1.353	0.931
2009	0.049	0.577	0.882	0.036**	0.683	0.085*
2010	0.946	0.044**	1.277	0.009***	1.674	0.001***
2011	−1.679	0.999	0.636	0.011**	−0.640	0.864
2012	−1.505	0.999	−0.369	0.639	−0.689	0.888
2013	−1.155	0.971	0.535	0.029**	−0.421	0.606
2014	−1.568	0.999	0.796	0.010***	−0.437	0.628
2015	−2.171	0.999	0.662	0.010***	−0.310	0.430
2016	−1.863	1.000	0.750	0.008***	−0.889	0.927
2017	−0.880	0.985	0.337	0.182	−0.159	0.665
2018	−0.788	0.964	−0.077	0.461	−0.931	0.974

The test statistic, described in Kneip et al. (2016), is computed using dimension-reduced data, such that $p = q = 1$. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 1996–2018. The p-values are developed using 100 sample-splits and 1,000 bootstrap replications using techniques described in Simar and Wilson (2020). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table B.8: Convexity Test (KS Test): With Quantity of Fuel as Input

	Input–Orientation		Output–Orientation		Hyperbolic	
Year	Statistic	p-value	Statistic	p-value	Statistic	p-value
1996	0.385	0.030**	0.133	0.881	0.201	0.472
1997	0.342	0.152	0.216	0.586	0.304	0.263
1998	0.284	0.253	0.222	0.505	0.329	0.158
1999	0.378	0.035**	0.183	0.621	0.334	0.085*
2000	0.405	0.062*	0.272	0.328	0.225	0.523
2001	0.428	0.008***	0.166	0.627	0.283	0.172
2002	0.394	0.044**	0.317	0.179	0.373	0.070*
2003	0.282	0.234	0.148	0.807	0.394	0.048**
2004	0.210	0.465	0.242	0.337	0.282	0.191
2005	0.312	0.155	0.131	0.921	0.345	0.091*
2006	0.219	0.419	0.118	0.927	0.209	0.445
2007	0.288	0.139	0.117	0.904	0.257	0.205
2008	0.466	0.032**	0.304	0.335	0.451	0.066*
2009	0.197	0.382	0.365	0.024**	0.280	0.101
2010	0.381	0.037**	0.404	0.019**	0.503	0.002***
2011	0.509	0.000***	0.323	0.062*	0.344	0.045**
2012	0.569	0.000***	0.246	0.213	0.433	0.007***
2013	0.429	0.024**	0.304	0.163	0.372	0.065*
2014	0.533	0.001***	0.303	0.184	0.289	0.205
2015	0.617	0.001***	0.304	0.211	0.279	0.310
2016	0.528	0.002***	0.345	0.081*	0.352	0.072*
2017	0.369	0.012**	0.225	0.223	0.159	0.536
2018	0.296	0.071*	0.096	0.931	0.339	0.034**

The test statistic is from the KS test as described in Simar and Wilson (2020). In this case, I use the KS test to test for uniformity of the p-values estimated from the multiple sample-splits used in the tests presented in Table B.7. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 1996–2018. The p-values are developed using 100 sample-splits and 1,000 bootstrap replications using techniques described in Simar and Wilson (2020). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table B.9: Returns to Scale Test: Without Quantity of Fuel as Input

	Input–Orientation		Output–Orientation		Hyperbolic	
Year	Statistic	p-value	Statistic	p-value	Statistic	p-value
1996	3.038	0.036**	4.529	0.000***	4.753	0.000***
1997	2.693	0.133	4.475	0.000***	3.992	0.012**
1998	2.373	0.210	3.771	0.001***	3.884	0.015**
1999	1.893	0.284	2.858	0.015**	2.639	0.073*
2000	1.624	0.423	3.180	0.011**	2.560	0.120
2001	2.554	0.088*	3.855	0.003***	3.247	0.048**
2002	2.913	0.142	4.553	0.000***	3.784	0.053*
2003	2.190	0.205	3.831	0.001***	3.270	0.023**
2004	1.923	0.240	3.532	0.000***	2.898	0.021**
2005	2.683	0.110	4.685	0.000***	3.609	0.035**
2006	2.690	0.040**	4.479	0.000***	4.276	0.004***
2007	2.849	0.073*	5.014	0.000***	4.653	0.003***
2008	2.157	0.349	4.932	0.000***	5.523	0.003***
2009	2.538	0.054*	4.491	0.000***	4.485	0.000***
2010	0.990	0.347	2.773	0.000***	2.470	0.009***
2011	2.929	0.034**	5.776	0.000***	5.185	0.002***
2012	3.271	0.028**	5.828	0.000***	5.315	0.001***
2013	2.229	0.089*	5.417	0.000***	4.727	0.000***
2014	2.475	0.070*	5.476	0.000***	4.458	0.002***
2015	2.068	0.166	4.443	0.000***	4.553	0.001***
2016	3.155	0.039**	6.366	0.000***	6.097	0.000***
2017	2.998	0.013**	4.928	0.000***	4.258	0.000***
2018	2.676	0.050**	4.779	0.000***	4.497	0.002***

The test statistic, described in Kneip et al. (2016), is computed using dimension-reduced data, such that $p = q = 1$. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 1996–2018. The p-values are developed using 100 sample-splits and 1,000 bootstrap replications using techniques described in Simar and Wilson (2020). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table B.10: Returns to Scale Test (KS Test): Without Quantity of Fuel as Input

	Input–Orientation		Output–Orientation		Hyperbolic	
Year	Statistic	p-value	Statistic	p-value	Statistic	p-value
1996	0.615	0.183	0.924	0.000***	0.824	0.008***
1997	0.720	0.090*	0.911	0.000***	0.794	0.016**
1998	0.592	0.297	0.816	0.003***	0.762	0.023**
1999	0.592	0.284	0.790	0.007***	0.747	0.034**
2000	0.474	0.509	0.635	0.040**	0.487	0.480
2001	0.682	0.089*	0.854	0.002***	0.679	0.105
2002	0.567	0.409	0.880	0.000***	0.709	0.114
2003	0.634	0.179	0.867	0.000***	0.767	0.015**
2004	0.635	0.154	0.831	0.000***	0.698	0.037**
2005	0.700	0.087*	0.877	0.000***	0.736	0.062*
2006	0.756	0.028**	0.939	0.000***	0.891	0.002***
2007	0.587	0.205	0.947	0.000***	0.813	0.010***
2008	0.615	0.281	0.788	0.011**	0.772	0.038**
2009	0.581	0.200	0.814	0.001***	0.830	0.001***
2010	0.400	0.325	0.871	0.000***	0.801	0.002***
2011	0.813	0.009***	0.982	0.000***	0.954	0.000***
2012	0.624	0.165	0.942	0.000***	0.687	0.113
2013	0.579	0.179	0.956	0.000***	0.896	0.001***
2014	0.663	0.100*	0.951	0.000***	0.875	0.000***
2015	0.506	0.398	0.856	0.000***	0.746	0.031**
2016	0.635	0.202	0.900	0.000***	0.782	0.033**
2017	0.835	0.003***	0.944	0.000***	0.903	0.000***
2018	0.543	0.266	0.922	0.000***	0.743	0.027**

The test statistic is from the KS test as described in Simar and Wilson (2020). In this case, I use the KS test to test for uniformity of the p-values estimated from the multiple sample-splits used in the tests presented in Table B.9. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 1996–2018. The p-values are developed using 100 sample-splits and 1,000 bootstrap replications using techniques described in Simar and Wilson (2020). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table B.11: Returns to Scale Test: With Quantity of Fuel as Input

	Input–Orientation		Output–Orientation		Hyperbolic	
Year	Statistic	p-value	Statistic	p-value	Statistic	p-value
1996	2.780	0.054*	4.916	0.000***	4.482	0.001***
1997	2.667	0.155	4.409	0.002***	4.325	0.009***
1998	2.557	0.159	4.143	0.001***	3.516	0.049**
1999	1.901	0.302	3.315	0.007***	2.726	0.102
2000	1.664	0.439	3.547	0.003***	3.157	0.056*
2001	2.616	0.104	4.204	0.000***	3.669	0.025**
2002	2.678	0.204	4.194	0.000***	4.000	0.040**
2003	2.395	0.151	3.860	0.002***	3.150	0.045**
2004	1.891	0.294	3.732	0.001***	3.308	0.015**
2005	2.855	0.118	4.598	0.003***	4.421	0.010***
2006	2.681	0.062*	4.797	0.000***	4.633	0.003***
2007	2.793	0.073*	4.728	0.000***	4.699	0.003***
2008	2.544	0.200	5.071	0.000***	5.001	0.010***
2009	2.878	0.026**	4.714	0.000***	4.061	0.000***
2010	1.176	0.289	3.248	0.000***	2.734	0.004***
2011	3.064	0.028**	5.862	0.000***	5.069	0.001***
2012	3.165	0.060*	6.676	0.000***	5.855	0.000***
2013	2.931	0.016**	5.304	0.000***	4.877	0.000***
2014	2.403	0.095*	5.177	0.000***	4.735	0.002***
2015	1.694	0.355	4.721	0.000***	4.435	0.004***
2016	3.146	0.041**	5.972	0.000***	6.415	0.000***
2017	3.014	0.010***	4.856	0.000***	4.498	0.000***
2018	2.727	0.061*	4.635	0.000***	5.068	0.001***

The test statistic, described in Kneip et al. (2016), is computed using dimension-reduced data, such that $p = q = 1$. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 1996–2018. The p-values are developed using 100 sample-splits and 1,000 bootstrap replications using techniques described in Simar and Wilson (2020). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table B.12: Returns to Scale Test (KS Test): With Quantity of Fuel as Input

	Input–Orientation		Output–Orientation		Hyperbolic	
Year	Statistic	p-value	Statistic	p-value	Statistic	p-value
1996	0.593	0.239	0.924	0.000***	0.820	0.011**
1997	0.705	0.127	0.849	0.000***	0.811	0.016**
1998	0.624	0.240	0.893	0.001***	0.698	0.106
1999	0.538	0.404	0.825	0.000***	0.650	0.141
2000	0.421	0.683	0.662	0.035**	0.576	0.215
2001	0.677	0.108	0.862	0.000***	0.738	0.042**
2002	0.561	0.421	0.784	0.009***	0.661	0.165
2003	0.695	0.101	0.885	0.001***	0.765	0.022**
2004	0.620	0.203	0.873	0.000***	0.712	0.046**
2005	0.740	0.076*	0.848	0.001***	0.859	0.005***
2006	0.759	0.033**	0.963	0.001***	0.911	0.002***
2007	0.656	0.128	0.927	0.000***	0.828	0.008***
2008	0.678	0.139	0.846	0.001***	0.777	0.029**
2009	0.640	0.118	0.879	0.000***	0.771	0.011**
2010	0.462	0.269	0.855	0.000***	0.830	0.000***
2011	0.856	0.006***	0.966	0.000***	0.901	0.000***
2012	0.612	0.225	0.931	0.000***	0.734	0.049**
2013	0.711	0.036**	0.962	0.000***	0.836	0.006***
2014	0.670	0.094*	0.945	0.000***	0.880	0.003***
2015	0.391	0.702	0.849	0.000***	0.698	0.081*
2016	0.646	0.207	0.941	0.000***	0.838	0.009***
2017	0.771	0.021**	0.950	0.000***	0.919	0.000***
2018	0.546	0.280	0.888	0.000***	0.796	0.008***

The test statistic is from the KS test as described in Simar and Wilson (2020). In this case, I use the KS test to test for uniformity of the p-values estimated from the multiple sample-splits used in the tests presented in Table B.11. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 1996–2018. The p-values are developed using 100 sample-splits and 1,000 bootstrap replications using techniques described in Simar and Wilson (2020). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table B.13: Test For Equivalency of Mean Efficiency: FDH Estimator Without Quantity of Fuel as Input

	Input–Orientation		Output–Orientation		Hyperbolic	
Period	Statistic	p-value	Statistic	p-value	Statistic	p-value
1996–1997	−0.571	0.568	0.325	0.745	−1.044	0.297
1997–1998	0.006	0.995	−0.086	0.932	0.372	0.710
1998–1999	0.749	0.454	−0.482	0.629	1.282	0.200
1999–2000	0.288	0.773	0.675	0.500	−0.576	0.565
2000–2001	−0.424	0.671	−1.314	0.189	−1.268	0.205
2001–2002	−0.083	0.934	0.792	0.428	1.825	0.068*
2002–2003	0.65	0.516	−0.305	0.760	−0.369	0.712
2003–2004	−0.119	0.905	0.508	0.611	0.039	0.969
2004–2005	−1.528	0.126	−1.327	0.184	−2.105	0.035**
2005–2006	1.657	0.098*	1.165	0.244	−0.131	0.896
2006–2007	−0.082	0.934	−0.133	0.894	1.211	0.226
2007–2008	−6.945	0.000***	−3.761	0.000***	−5.092	0.000***
2008–2009	8.172	0.000***	5.416	0.000***	6.614	0.000***
2009–2010	0.161	0.872	0.533	0.594	0.706	0.480
2010–2011	−3.512	0.000***	−2.494	0.013**	−4.795	0.000***
2011–2012	0.818	0.413	−0.869	0.385	0.113	0.910
2012–2013	−1.028	0.304	1.507	0.132	0.758	0.448
2013–2014	0.726	0.468	−0.996	0.319	−0.46	0.645
2014–2015	−0.472	0.637	−0.212	0.832	−0.066	0.948
2015–2016	−0.092	0.927	0.484	0.628	0.169	0.866
2016–2017	2.134	0.033**	−0.467	0.641	0.164	0.870
2017–2018	−0.538	0.590	−0.321	0.748	−0.75	0.453
2007 & 2018	−1.077	0.281	−0.850	0.396	−2.182	0.029**
2008 & 2018	6.097	0.000***	2.999	0.003***	2.978	0.003***
2009 & 2018	−2.349	0.019**	−2.660	0.008***	−3.675	0.000***
2010 & 2018	−2.795	0.005***	−2.867	0.004***	−4.489	0.000***
2011 & 2018	0.996	0.319	−0.445	0.656	0.396	0.692

The test statistic, described in Kneip et al. (2016), is computed using the FDH estimator with dimension-reduced data, such that $p = q = 1$. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 1996–2018. Since separability tests suggest that separability with respect to T does not hold, the estimator allows for a different frontier every production year. One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table B.14: Test For Equivalency of Mean Efficiency: FDH Estimator With Quantity of Fuel as Input

	Input–Orientation		Output–Orientation		Hyperbolic	
Period	Statistic	p-value	Statistic	p-value	Statistic	p-value
1996–1997	−0.713	0.476	0.104	0.917	−0.887	0.375
1997–1998	0.3	0.764	0.473	0.636	0.574	0.566
1998–1999	0.146	0.884	−1.401	0.161	0.695	0.487
1999–2000	0.396	0.692	0.373	0.709	−0.569	0.570
2000–2001	−0.491	0.623	−0.628	0.530	−0.868	0.385
2001–2002	0.029	0.977	0.709	0.478	1.655	0.098*
2002–2003	0.526	0.599	−0.26	0.795	0.248	0.804
2003–2004	−0.019	0.985	0.767	0.443	−0.051	0.959
2004–2005	−1.447	0.148	−1.34	0.180	−2.057	0.040**
2005–2006	1.016	0.310	1.117	0.264	−0.178	0.859
2006–2007	0.124	0.902	−0.346	0.730	1.454	0.146
2007–2008	−6.785	0.000***	−2.863	0.004***	−4.261	0.000***
2008–2009	7.854	0.000***	4.843	0.000***	5.888	0.000***
2009–2010	0.453	0.650	0.139	0.890	0.927	0.354
2010–2011	−3.705	0.000***	−2.581	0.010***	−4.873	0.000***
2011–2012	1.068	0.286	−0.675	0.500	0.086	0.932
2012–2013	−0.954	0.340	1.28	0.200	1.029	0.304
2013–2014	0.301	0.764	−0.3	0.764	−0.757	0.449
2014–2015	−0.504	0.614	−0.543	0.587	0.177	0.859
2015–2016	0.434	0.664	0.468	0.640	0.909	0.363
2016–2017	1.690	0.091*	−0.434	0.664	−0.155	0.877
2017–2018	−0.771	0.440	−0.402	0.688	−0.794	0.427
2007 & 2018	−0.760	0.447	−0.803	0.422	−2.116	0.034**
2008 & 2018	5.869	0.000***	2.077	0.038**	2.133	0.033**
2009 & 2018	−1.938	0.053*	−2.960	0.003***	−3.552	0.000***
2010 & 2018	−2.524	0.012**	−3.044	0.002***	−4.391	0.000***
2011 & 2018	1.267	0.205	−0.253	0.800	0.442	0.658

The test statistic, described in Kneip et al. (2016), is computed using the FDH estimator with dimension-reduced data, such that $p = q = 1$. The dimensions are reduced using eigensystem methods as described in Wilson (2018), and the principal components are estimated using the pooled sample of data from 1996–2018. Since separability tests suggest that separability with respect to T does not hold, the estimator allows for a different frontier every production year. One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table B.15: Productivity Estimates: Without Quantity of Fuel as Input

Period	Statistic	p-value
1996–1997	−0.367	0.714
1997–1998	0.130	0.897
1998–1999	−0.536	0.592
1999–2000	1.625	0.104
2000–2001	2.444	0.015**
2001–2002	−0.404	0.686
2002–2003	0.333	0.739
2003–2004	−0.306	0.760
2004–2005	0.472	0.637
2005–2006	0.218	0.827
2006–2007	−0.803	0.422
2007–2008	1.224	0.221
2008–2009	−1.107	0.268
2009–2010	0.156	0.876
2010–2011	1.636	0.102
2011–2012	−0.140	0.889
2012–2013	1.546	0.122
2013–2014	−0.814	0.416
2014–2015	−1.256	0.209
2015–2016	0.891	0.373
2016–2017	−1.165	0.244
2017–2018	0.877	0.380
2007–2018	2.467	0.013**
2008–2018	0.717	0.473
2009–2018	2.625	0.008***
2010–2018	1.982	0.047**
2011–2018	0.125	0.899

Productivity tests are conducted with dimension-reduced data ($p = q = 1$). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Table B.16: Productivity Estimates: With Quantity of Fuel as Input

Period	Statistic	p-value
1996–1997	−0.398	0.691
1997–1998	0.065	0.948
1998–1999	−0.474	0.635
1999–2000	1.640	0.101
2000–2001	2.570	0.010***
2001–2002	−0.424	0.671
2002–2003	0.319	0.750
2003–2004	−0.363	0.716
2004–2005	0.404	0.686
2005–2006	0.276	0.783
2006–2007	−0.867	0.386
2007–2008	1.220	0.222
2008–2009	−1.011	0.312
2009–2010	−0.003	0.998
2010–2011	1.671	0.095*
2011–2012	0.018	0.986
2012–2013	1.492	0.136
2013–2014	−0.773	0.440
2014–2015	−1.264	0.206
2015–2016	0.884	0.377
2016–2017	−1.169	0.243
2017–2018	0.875	0.381
2007–2018	2.559	0.010***
2008–2018	1.110	0.266
2009–2018	2.603	0.009***
2010–2018	2.048	0.040**
2011–2018	0.189	0.849

Productivity tests are conducted with dimension-reduced data ($p = q = 1$). One, two or three asterisks denote significance at the 10, 5 and 1 percent levels, respectively.

Appendix C

FDH From 1996–2018

Plots A.1–A.22 show observed pipeline first principal components, X_t^* (OPEX) and Y_t^*, and the estimated FDH frontier, $\widehat{\Psi}_{FDH}^t$, for $t \in [1996, 2018]$, using the specified inputs and outputs described in Section 1.3.1. Each plot shows $\widehat{\Psi}_{FDH}^t$ for two consecutive year pairs. Line segments trace pipelines on the frontier in year t to their position in year $t+1$, and vise versa. In addition, dotted line segments trace positions for select pipelines, noted below, over time. The plots show that the estimated FDH does change from year to year and steadily moves upward during the sample period. Moreover, firms on the frontier in year t are likely to remain on the frontier in year $t+1$. Finally, the plots do not provide any strong evidence for net technical regress or persistent drop in the frontier over time.

Figure C.1: FDH: 1996 Versus 1997

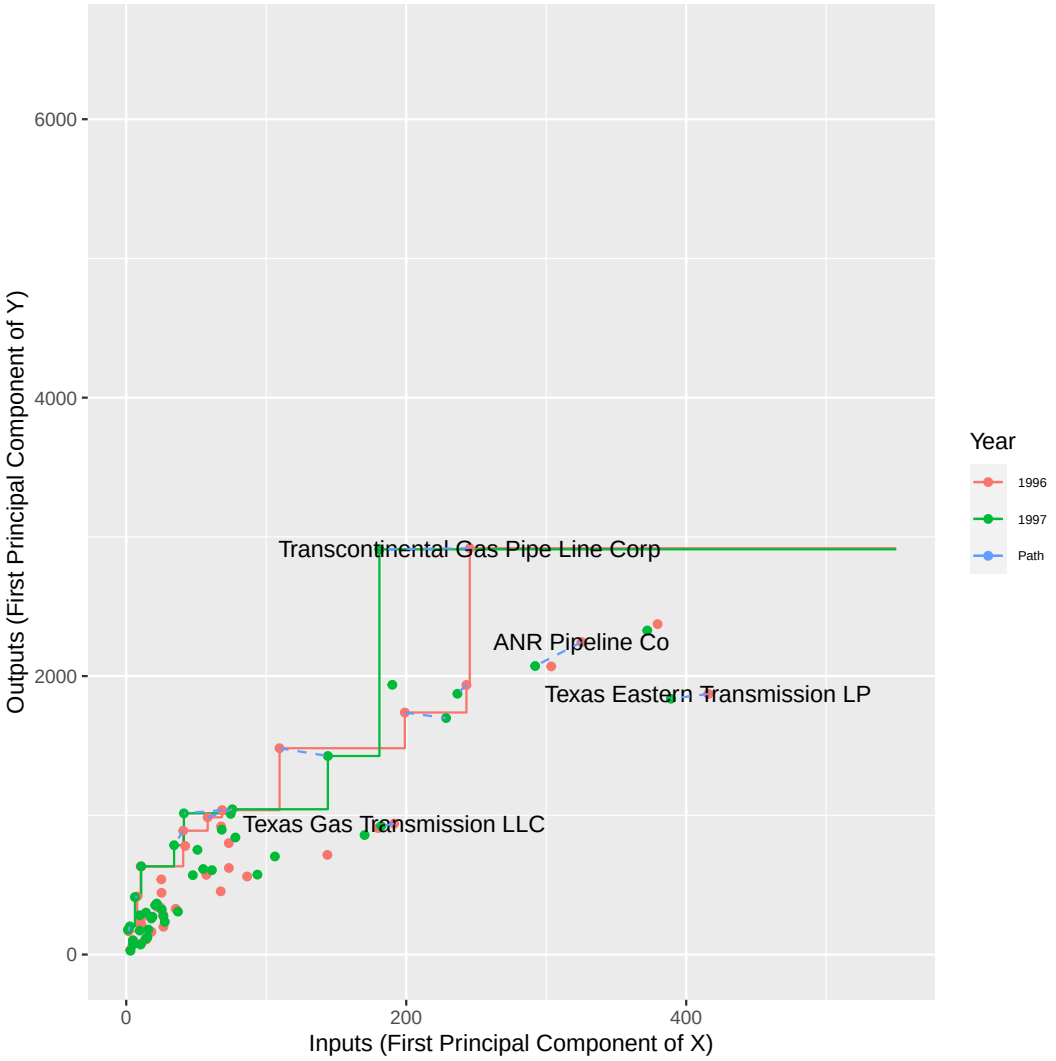


Figure C.2: FDH: 1997 Versus 1998

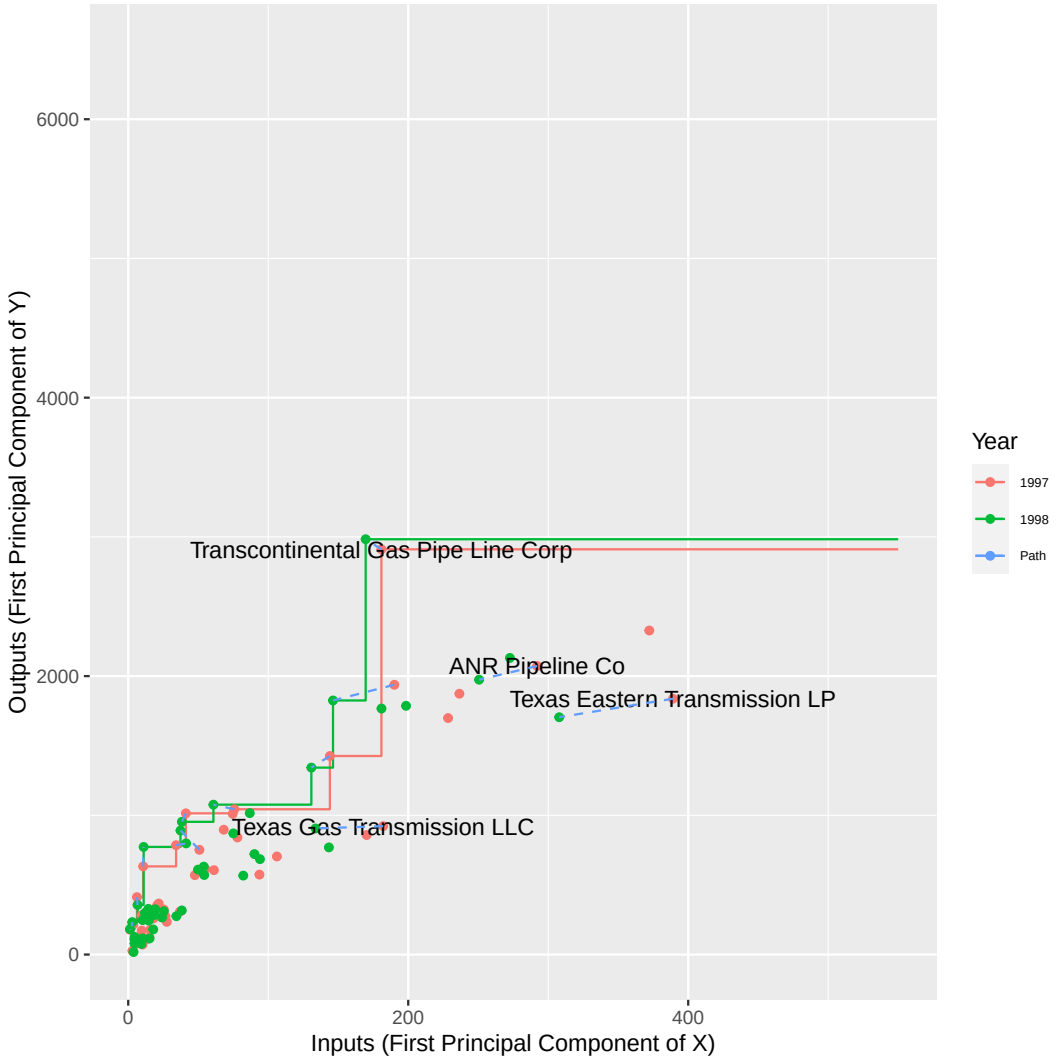


Figure C.3: FDH: 1998 Versus 1999

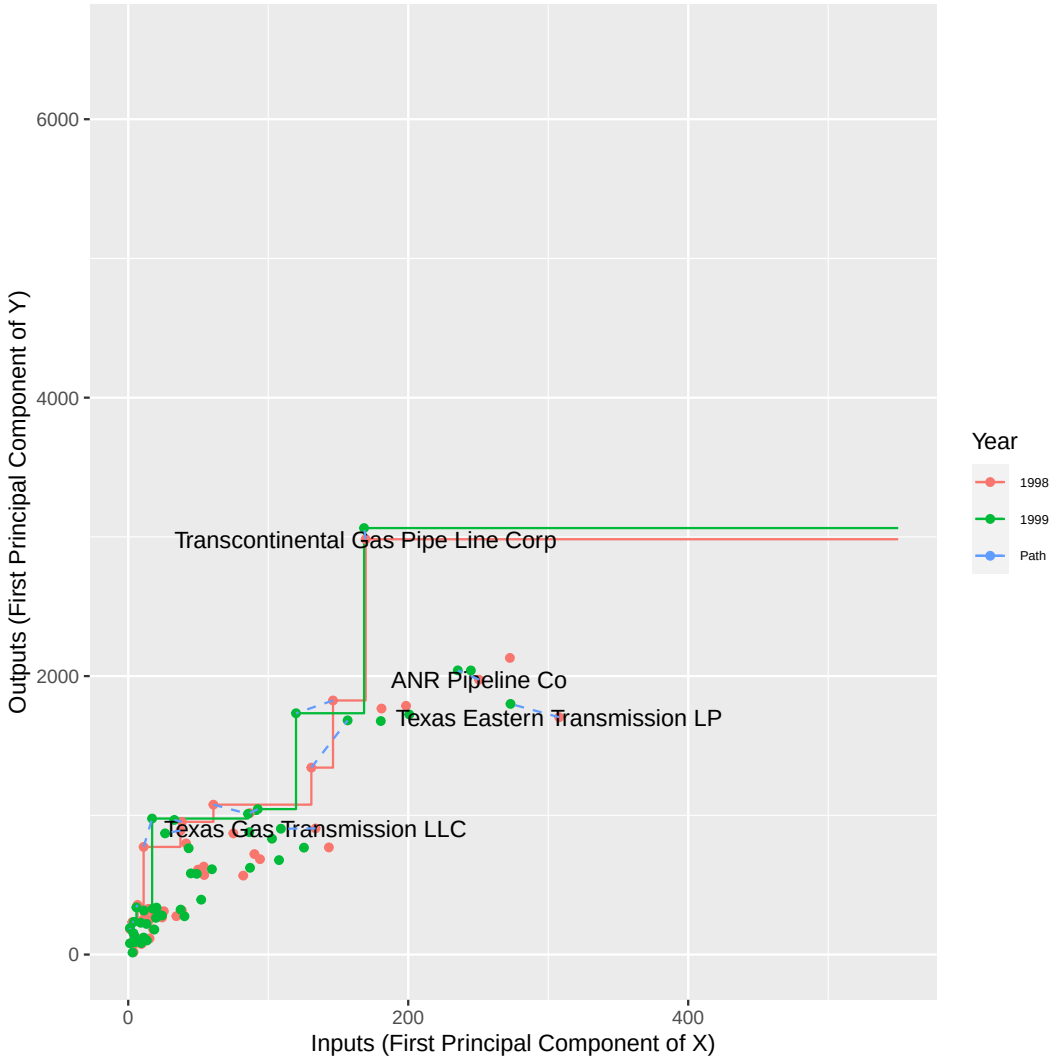


Figure C.4: FDH: 1999 Versus 2000

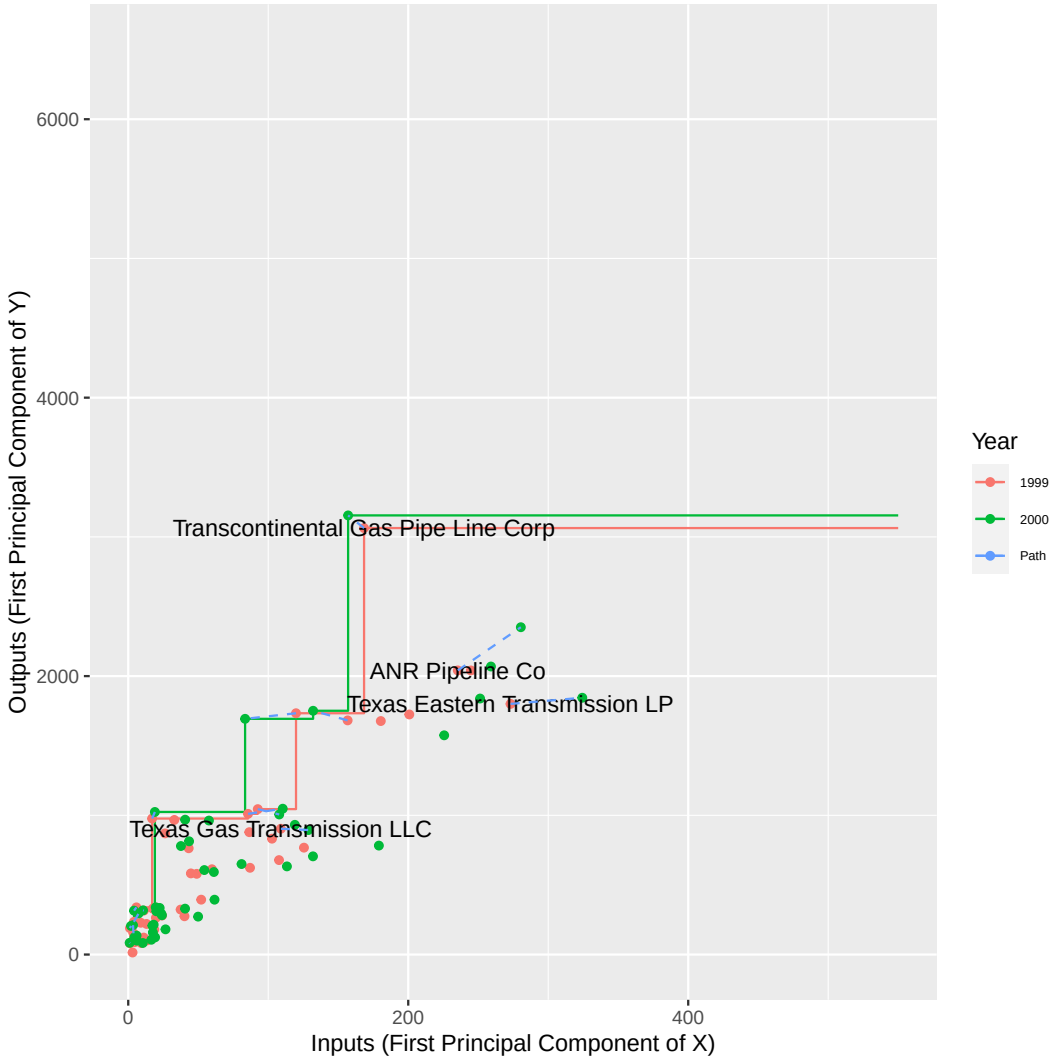


Figure C.5: FDH: 2000 Versus 2001

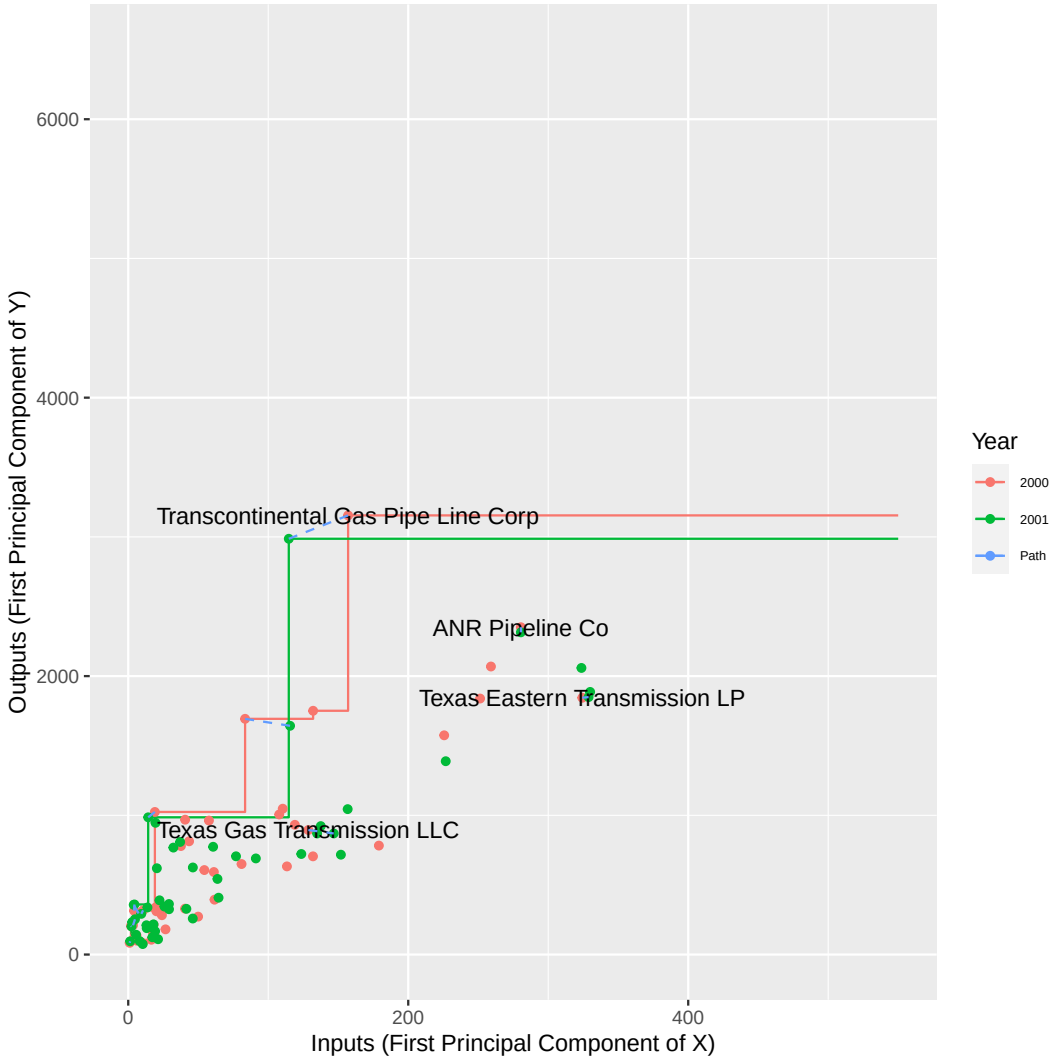


Figure C.6: FDH: 2001 Versus 2002

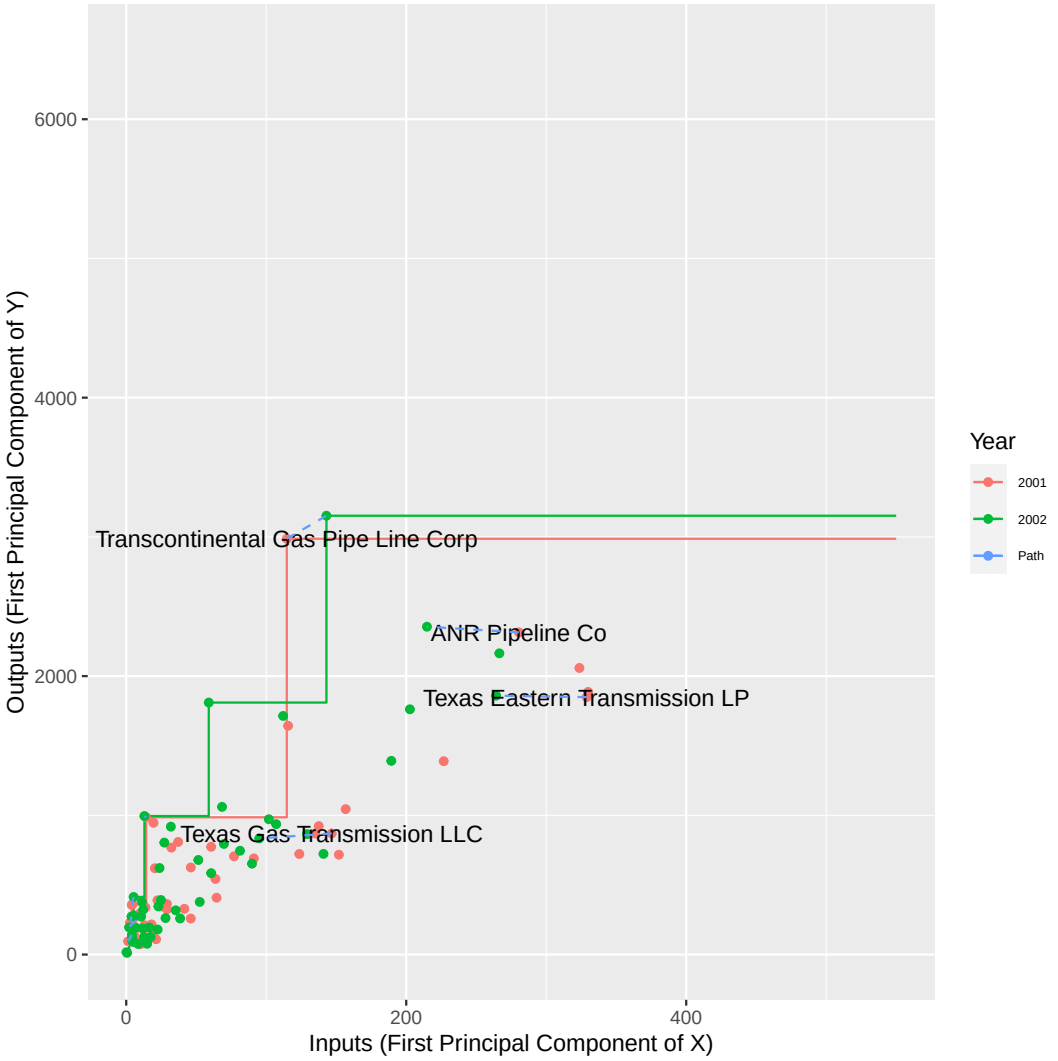


Figure C.7: FDH: 2002 Versus 2003

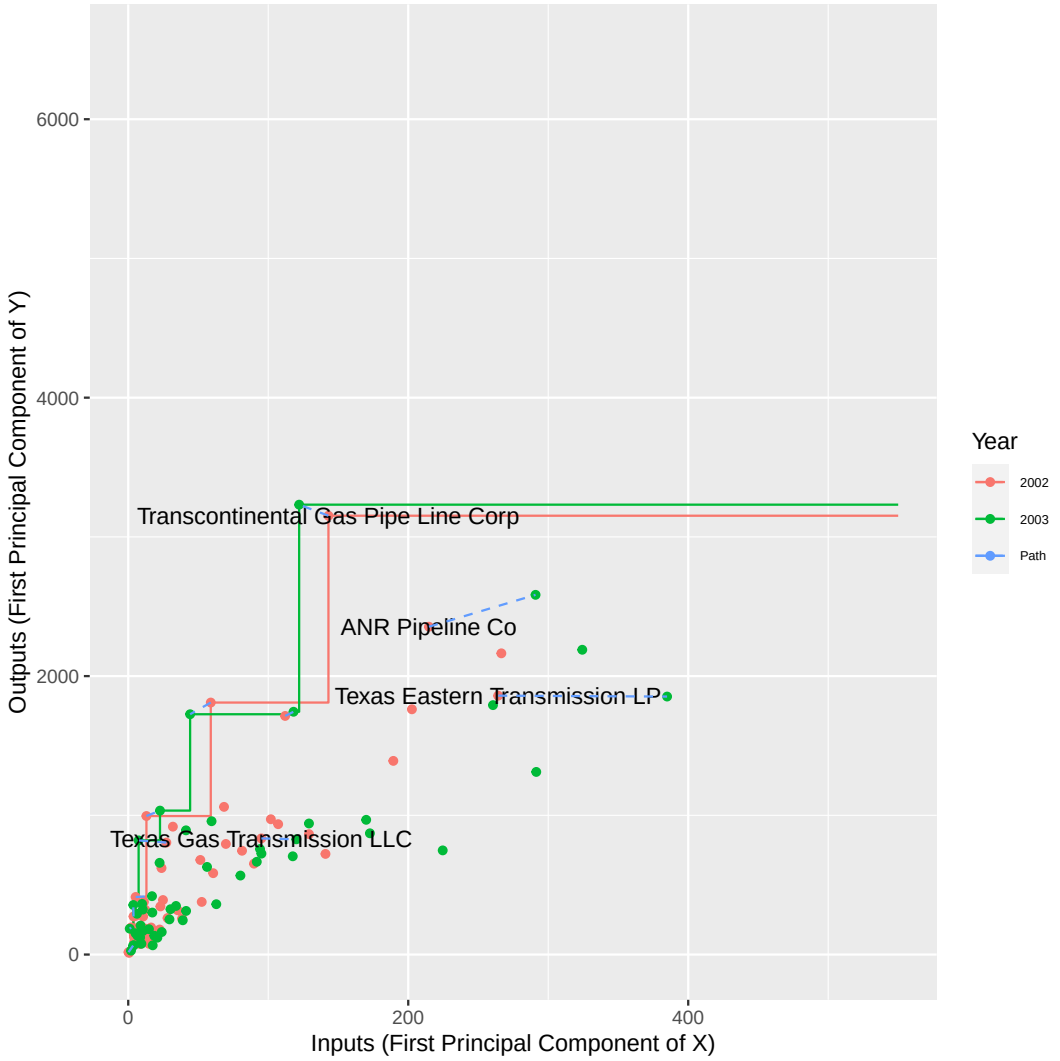


Figure C.8: FDH: 2003 Versus 2004

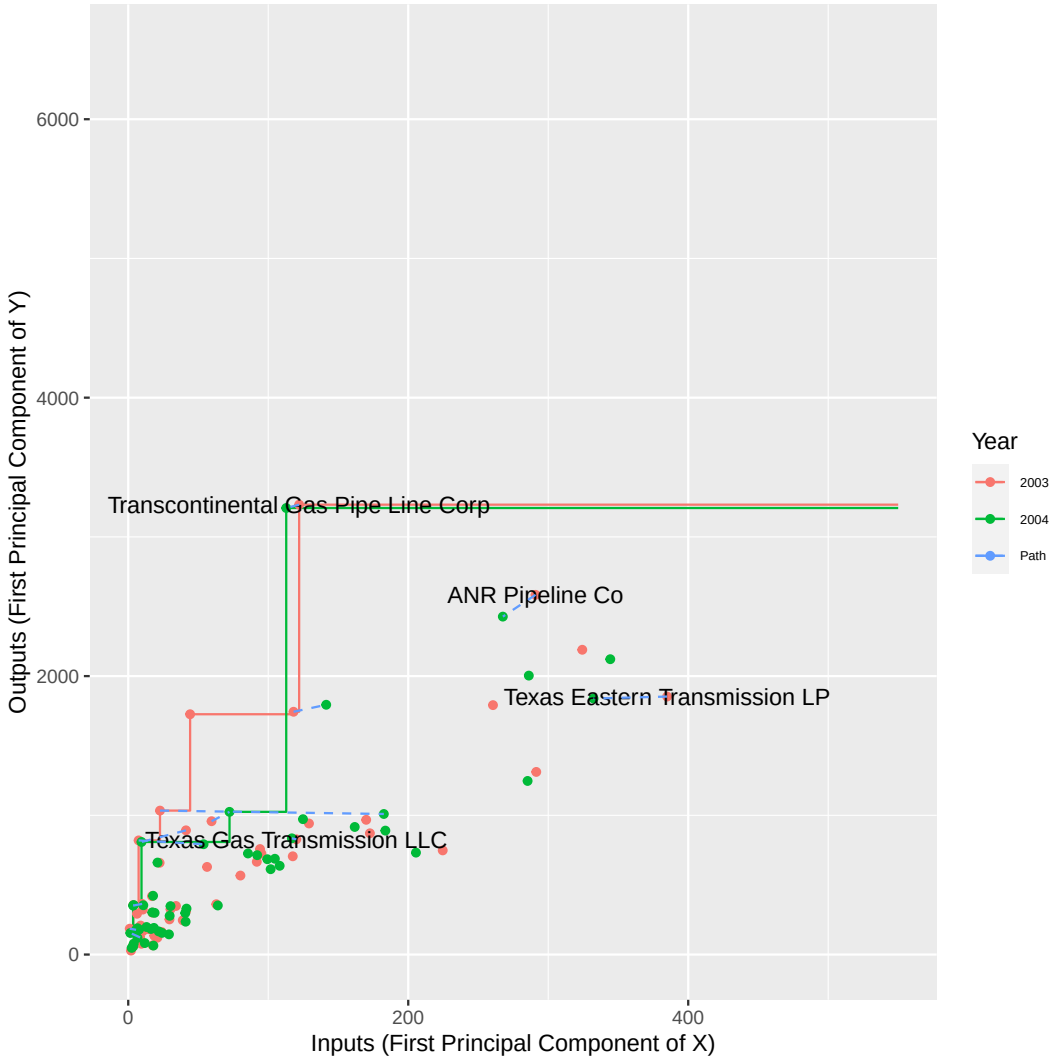


Figure C.9: FDH: 2004 Versus 2005

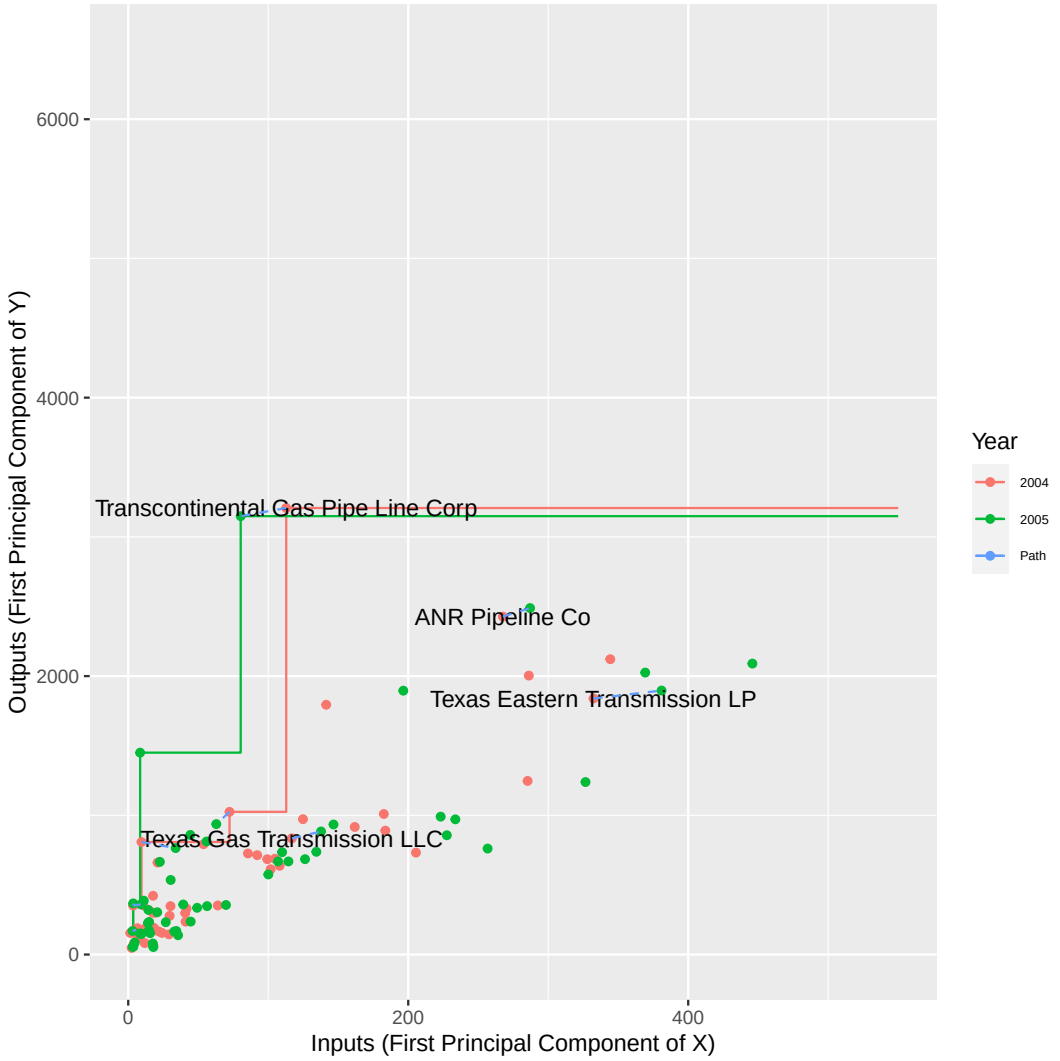


Figure C.10: FDH: 2005 Versus 2006

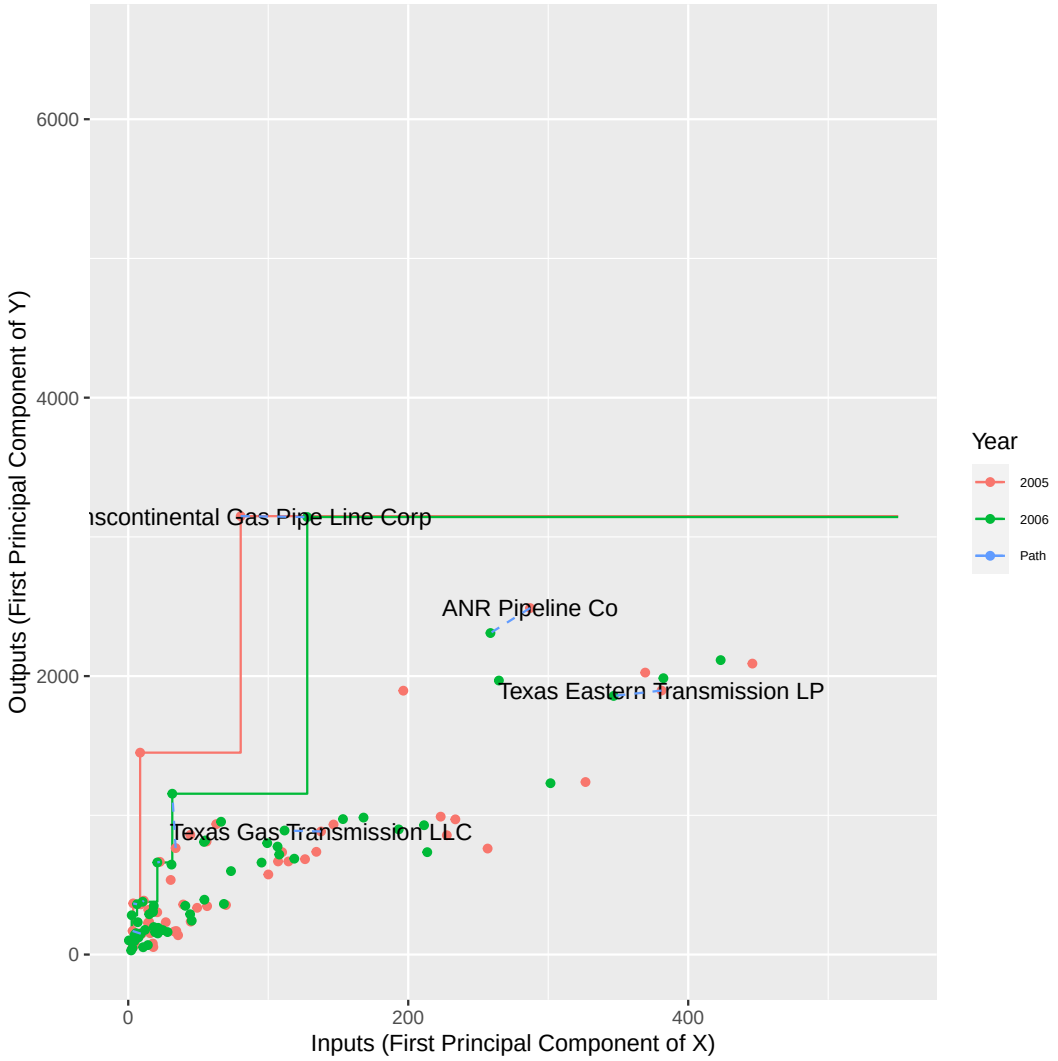


Figure C.11: FDH: 2006 Versus 2007

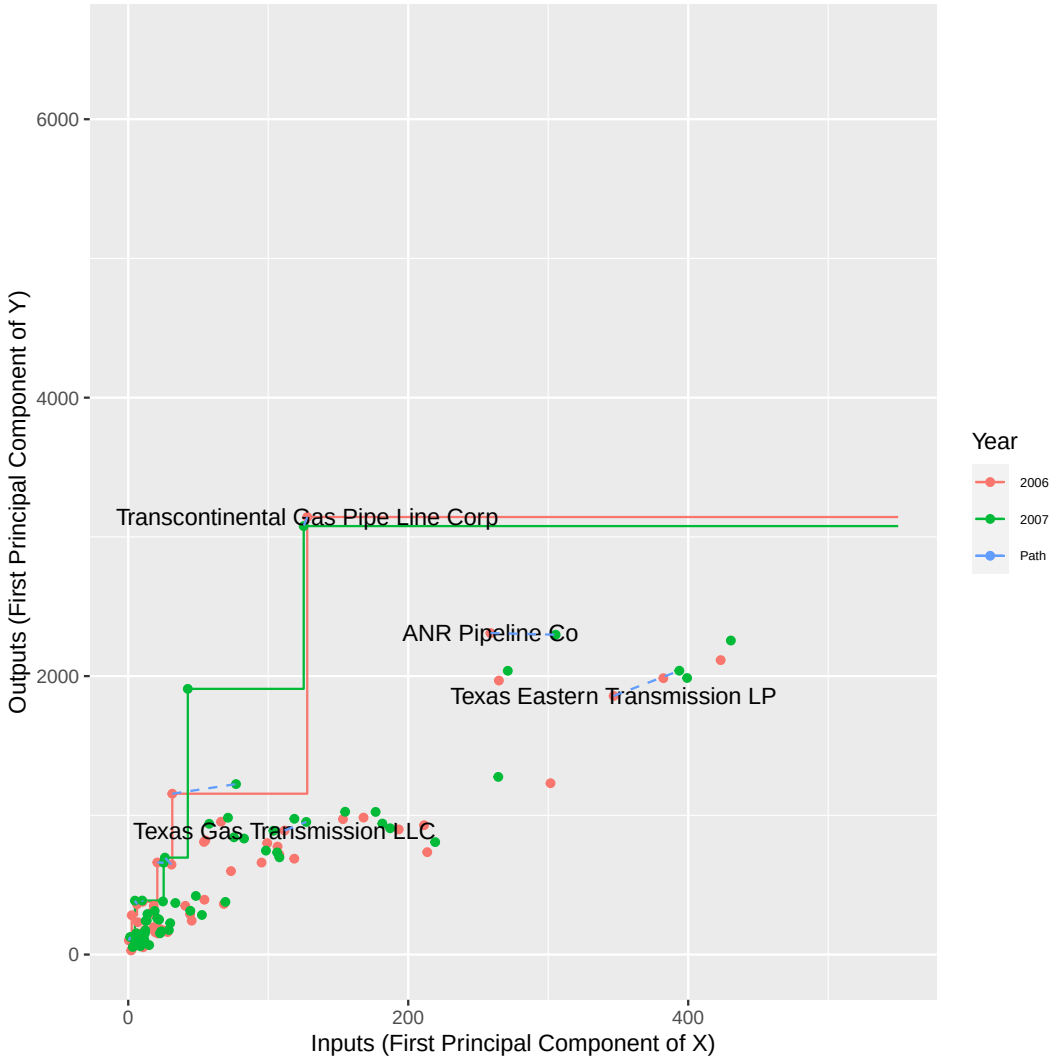


Figure C.12: FDH: 2007 Versus 2008

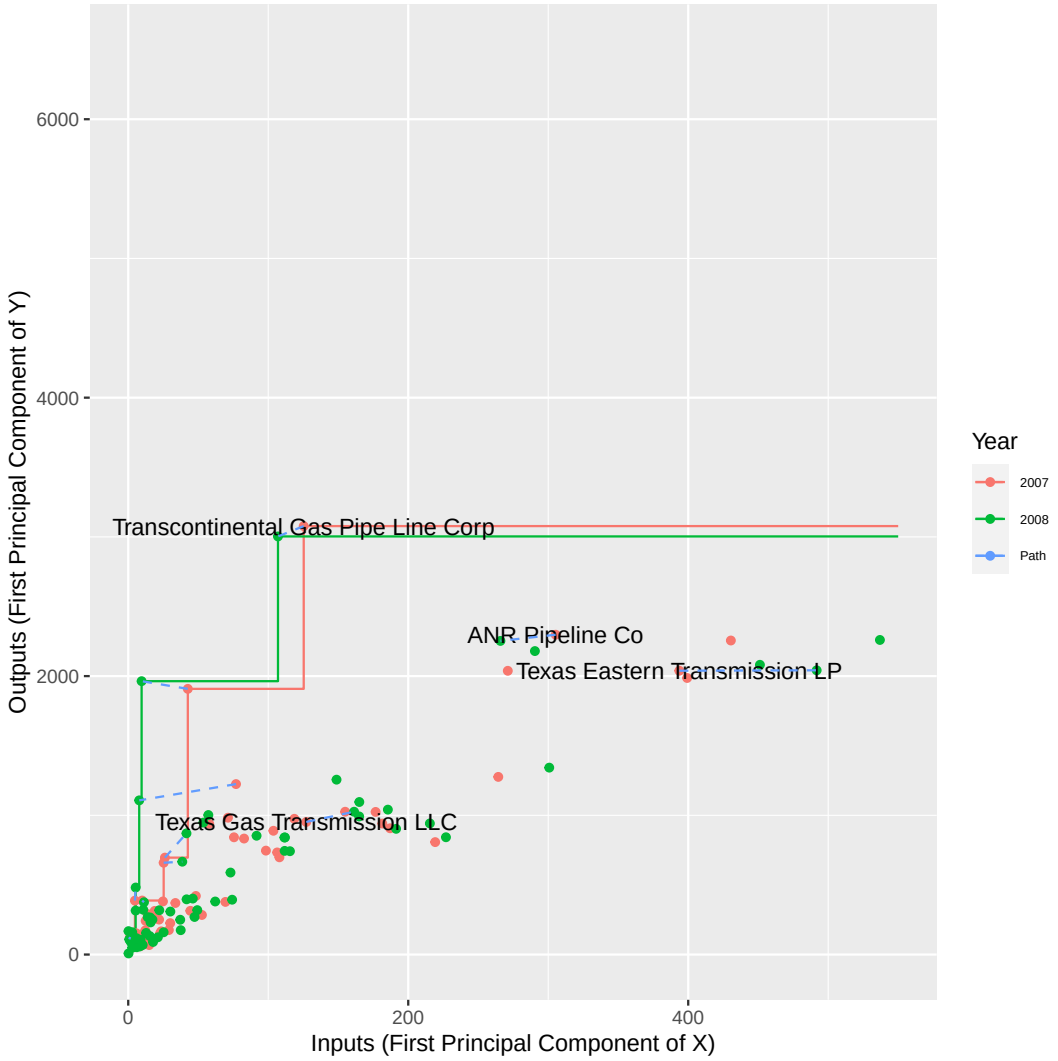


Figure C.13: FDH: 2008 Versus 2009

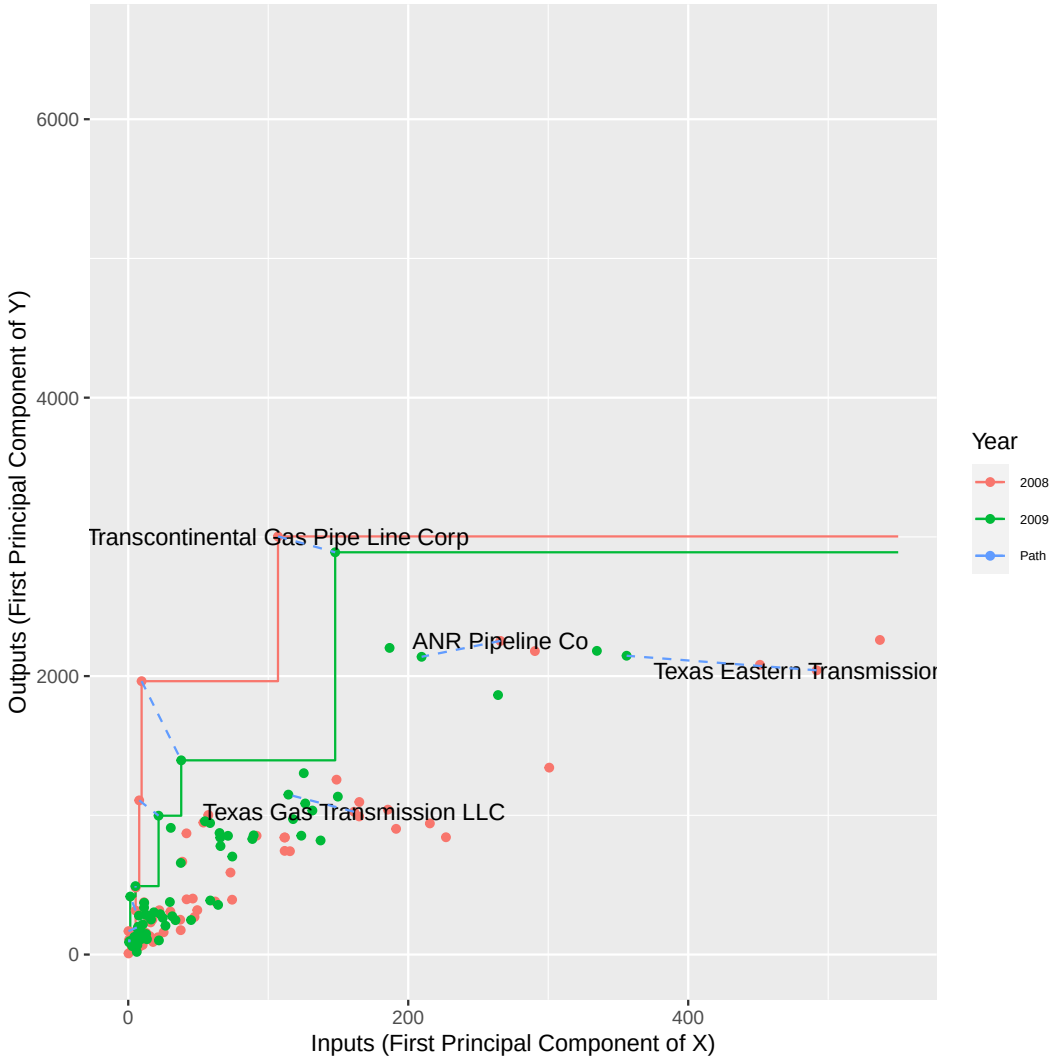


Figure C.14: FDH: 2009 Versus 2010

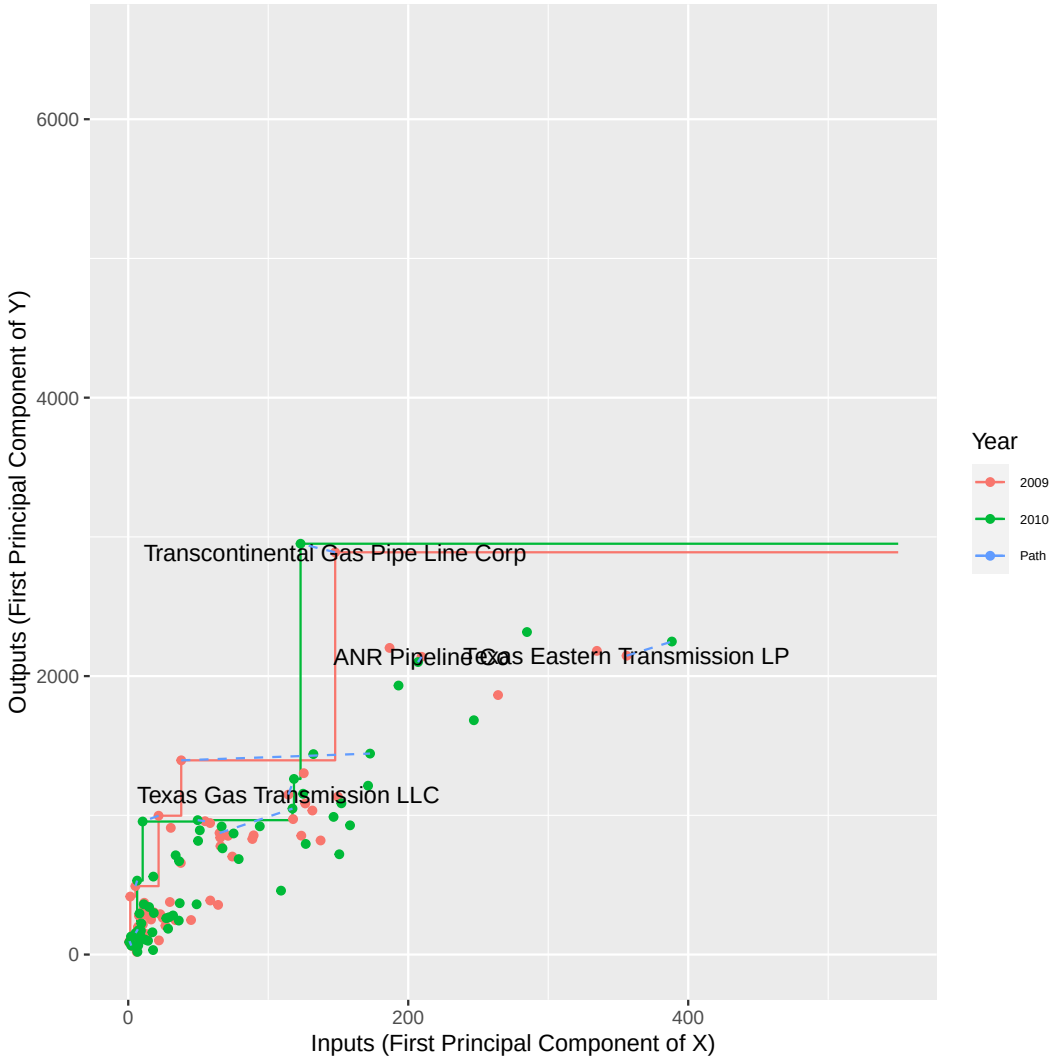


Figure C.15: FDH: 2010 Versus 2011

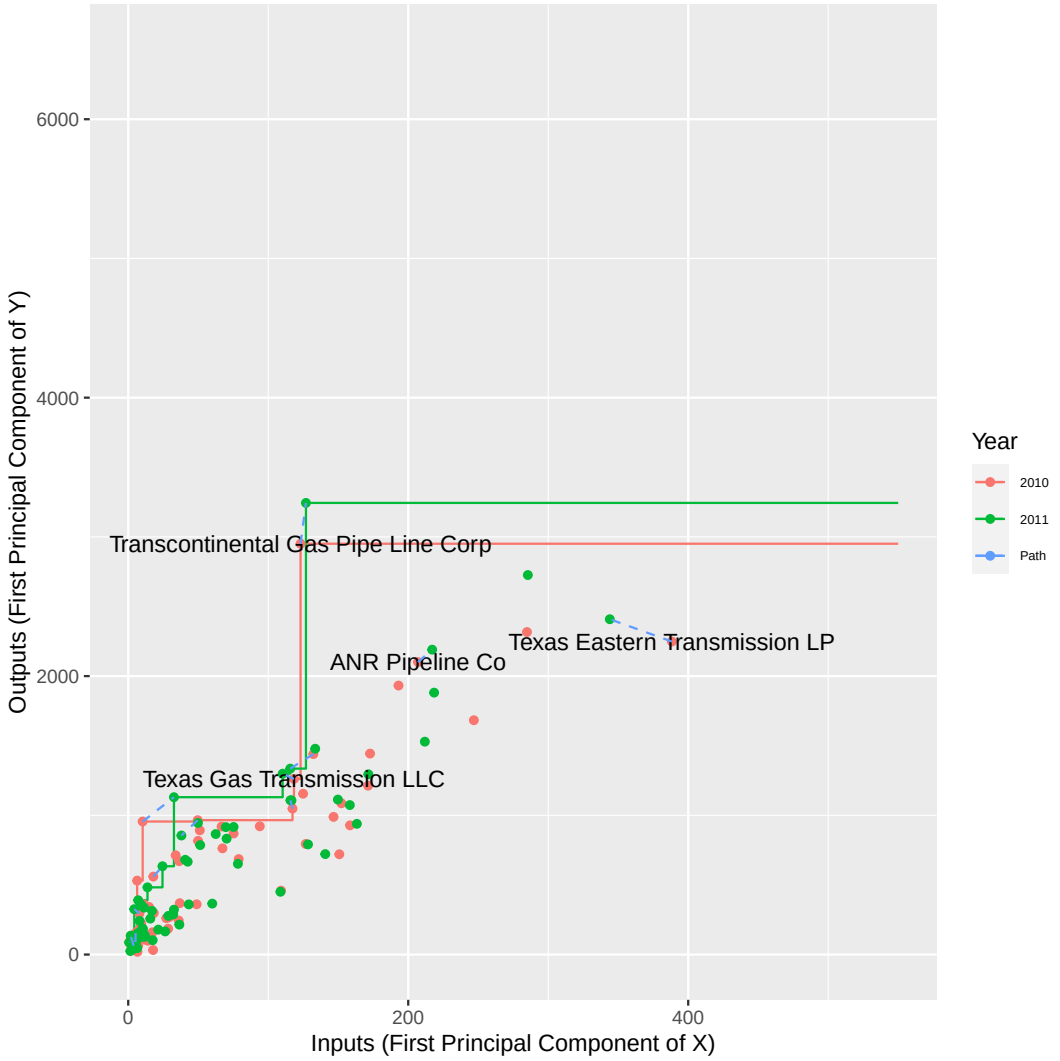


Figure C.16: FDH: 2011 Versus 2012

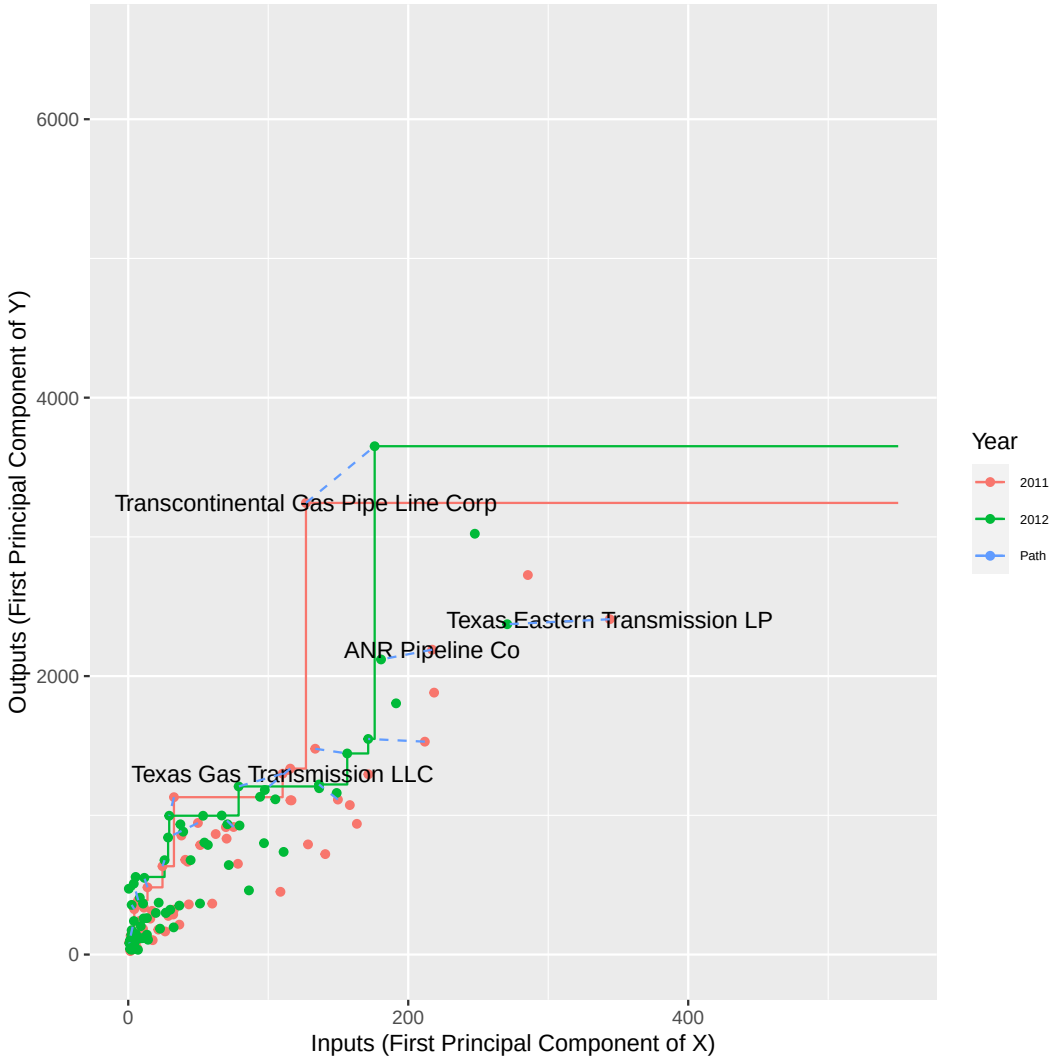


Figure C.17: FDH: 2012 Versus 2013

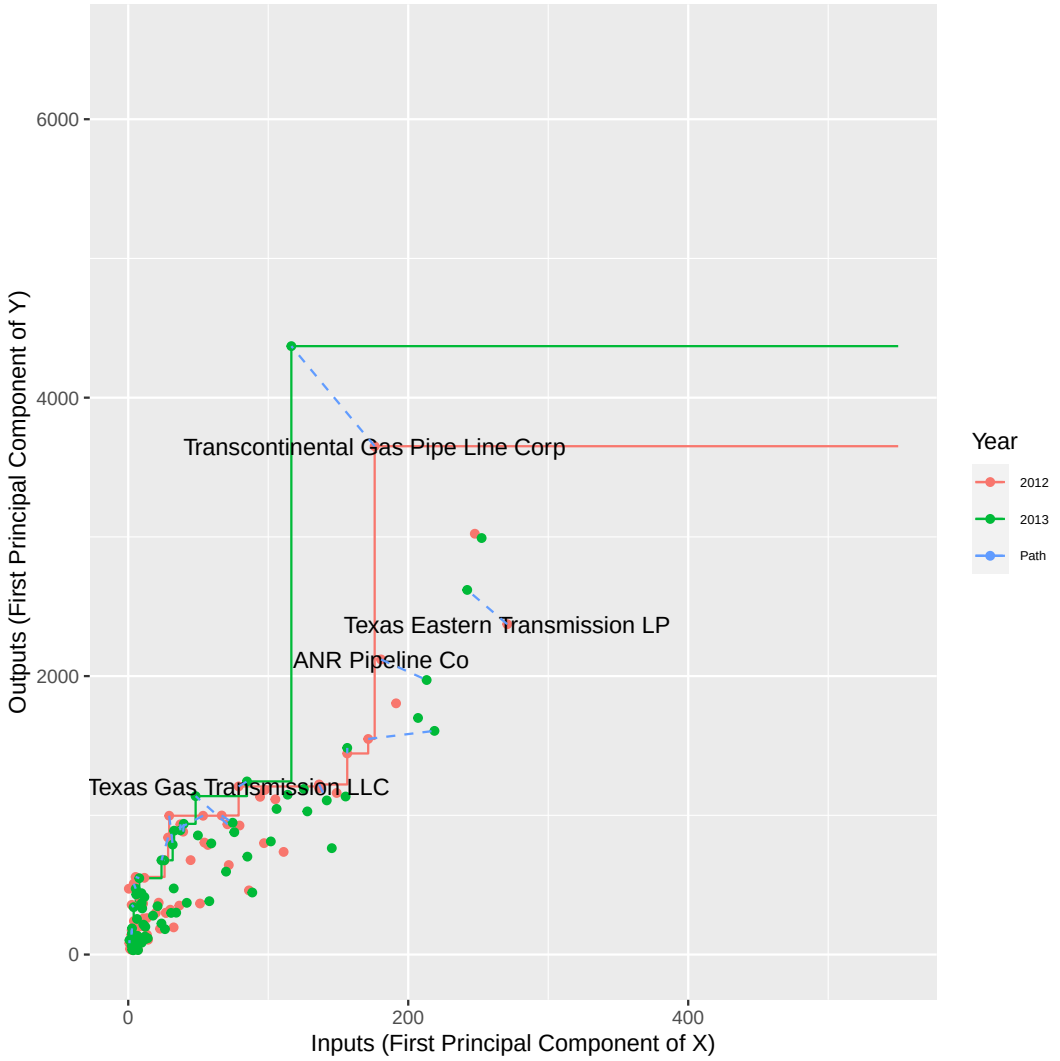


Figure C.18: FDH: 2013 Versus 2014

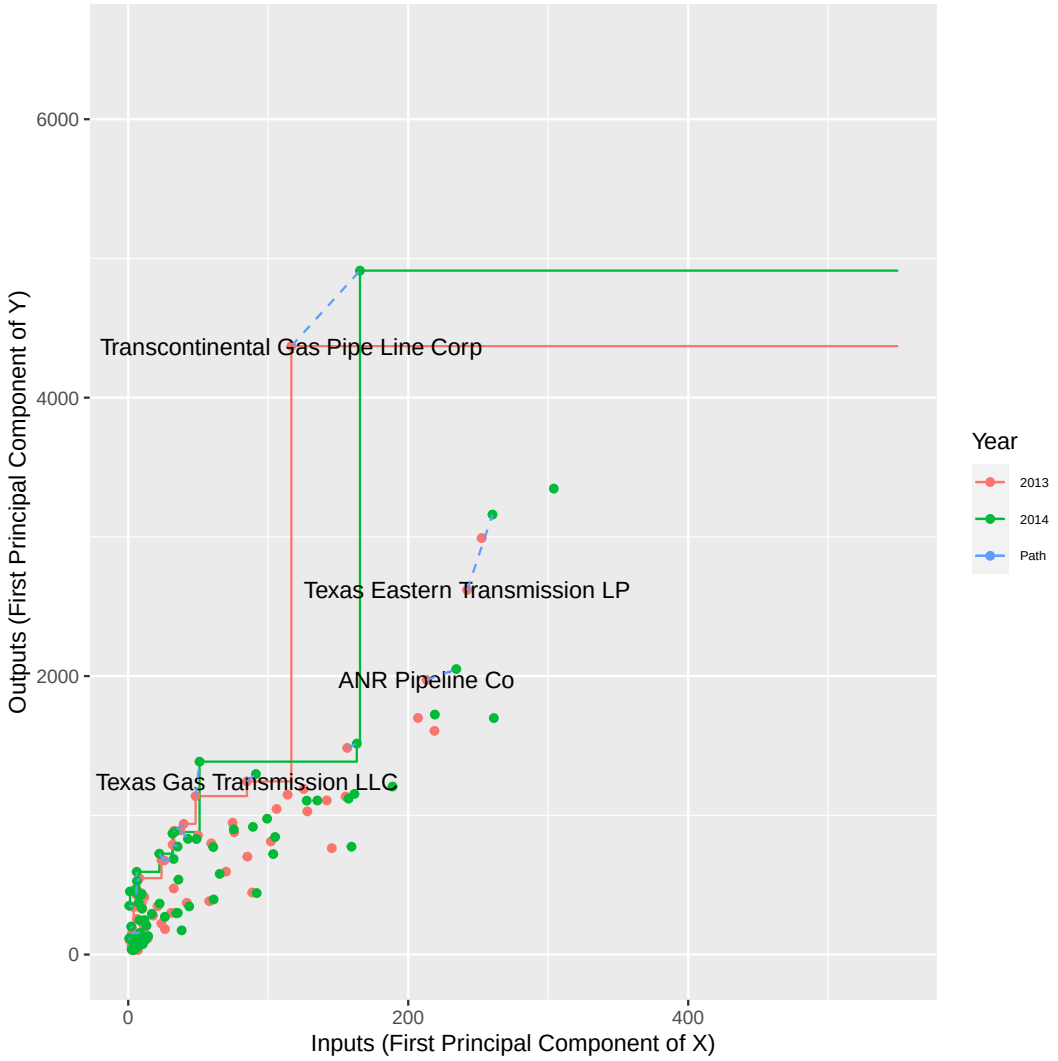


Figure C.19: FDH: 2014 Versus 2015

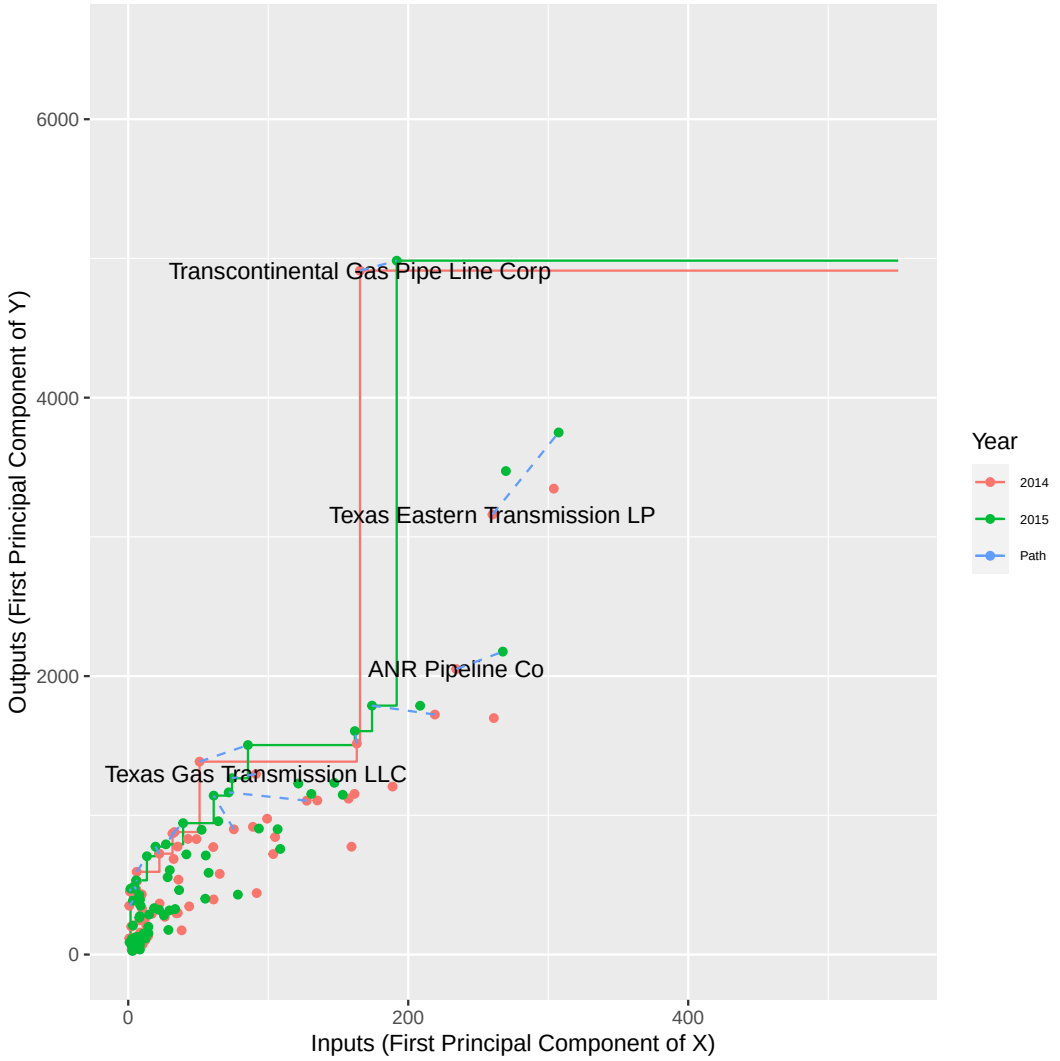


Figure C.20: FDH: 2015 Versus 2016

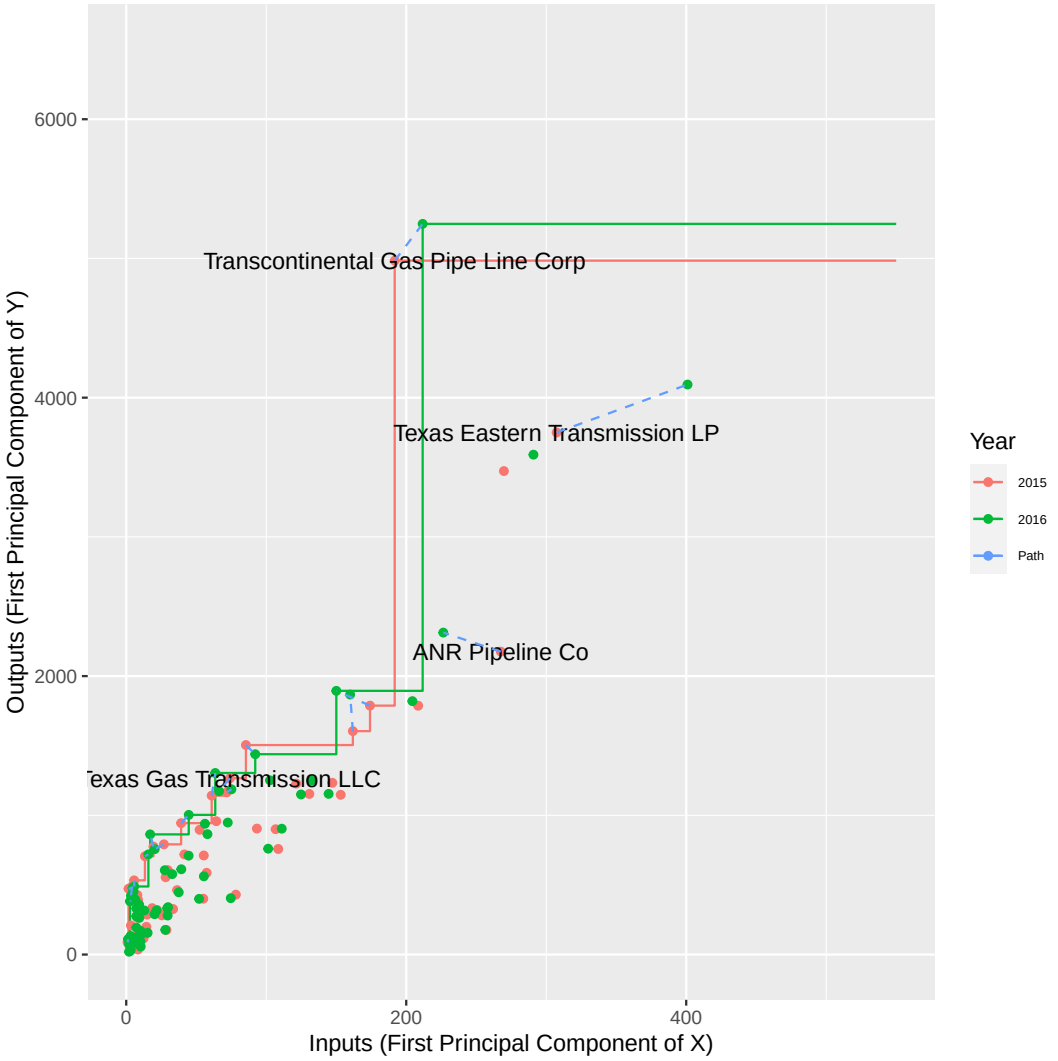


Figure C.21: FDH: 2016 Versus 2017

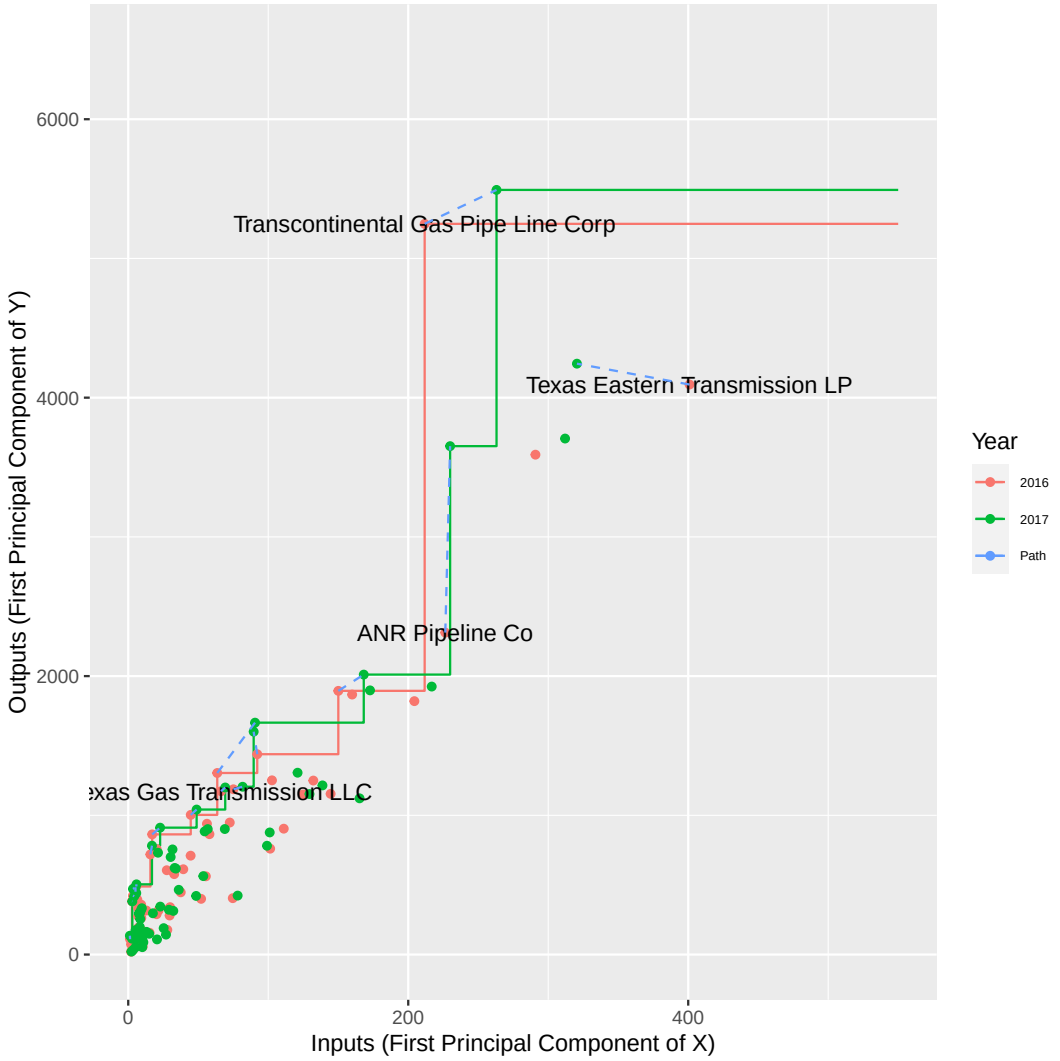
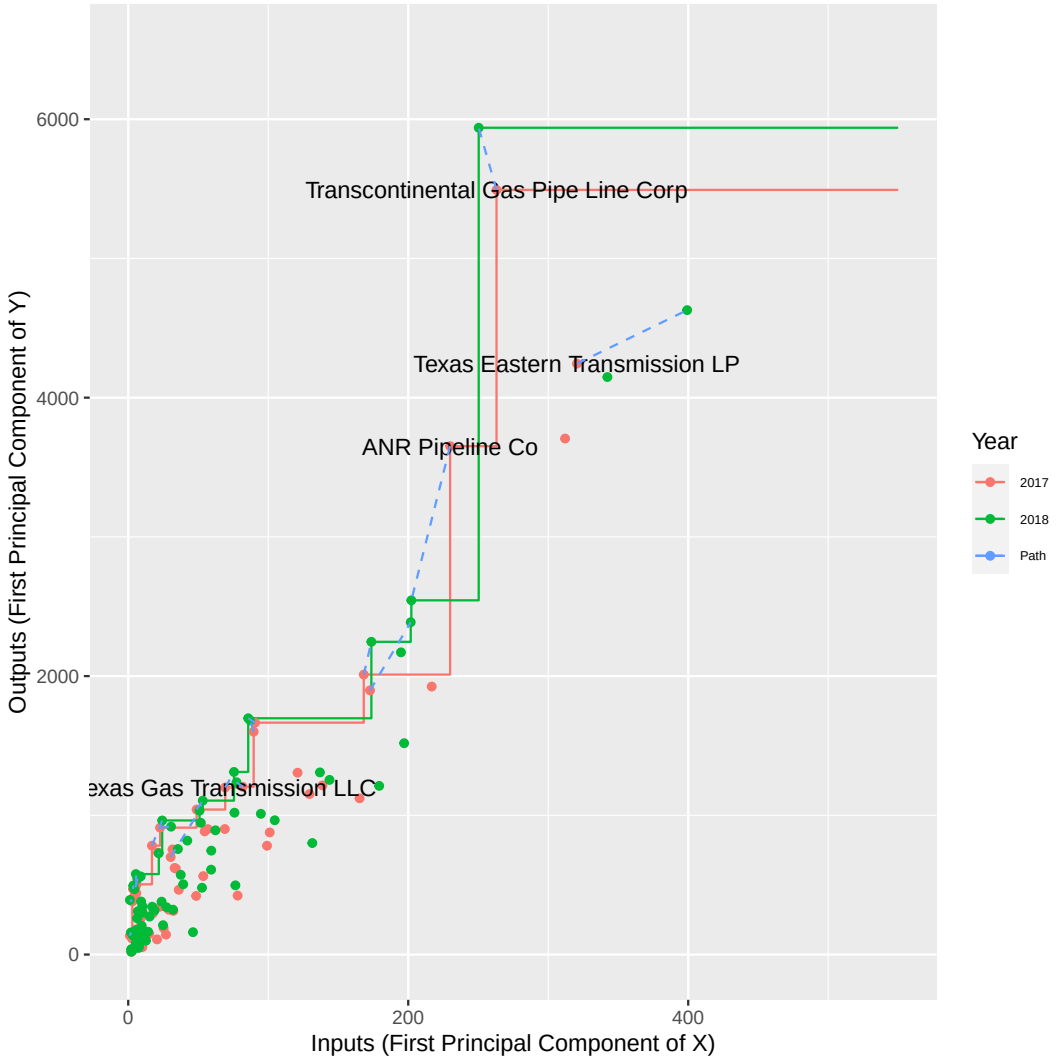


Figure C.22: FDH: 2017 Versus 2018



Appendix D

Discussion on Specific Pipelines

Despite the strong evidence of overall technical efficiency improvements during the Shale Revolution, the model described in this paper does not provide a lot of contextual information on how pipelines changed operations during this time period to become more technically efficient, or what factors prevented some pipelines from achieving greater efficiency. As noted in Appendix C, pipelines that are operating on $\Psi^{t\partial}$ in year t are most likely on $\Psi^{t\partial}$ in year $t+1$. However, a couple pipelines are an exception to this. ANR Pipeline Company (ANR) and Texas Gas Transmission (TGT) were operating in positions far away from $\Psi^{t\partial}$ in periods prior to the Shale Revolution. ANR is a large system that covers a region that includes Louisiana, Mississippi, Tennessee, Kentucky, Indiana, Ohio, Illinois, Iowa, Missouri, Oklahoma and Texas. TGT is also a large system and covers Louisiana, Arkansas, Mississippi, Tennessee, Kentucky, Illinois, Indiana, and Ohio. Both TGT and ANR were initially designed to receive gas produced in the Gulf of Mexico. However, the Shale Revolution made the U.S. a net exporter of natural gas, and the lower cost of shale production reduced the share of U.S. gas produced by offshore rigs. In 2014 both ANR and TGT announced they will restructure the pipeline to reverse flow and take gas produced in the Marcellus and Utica beds (in Pennsylvania, New York, West Virginia, and Ohio) to LNG terminals on the Gulf Coast. While TGT and ANR initially operated off of the frontier they eventually would operate at or near full efficiency in the years following their restructuring in 2014.

Transcontinental Gas Pipeline (TCO) offers an example of how initial pipeline geography

can dictate the scale and efficiency of a pipeline. Like TGT and ANR, TCO is a large system that was initially designed to receive gas produced off the U.S. Gulf Coast and deliver gas across the eastern seaboard. TCO covers an area that includes the Gulf Coast of Texas, Louisiana, Mississippi, and Alabama, as well as Georgia, South Carolina, North Carolina, Virginia, Maryland, Pennsylvania, New Jersey and New York City. TCO occupies a position on $\Psi^{t\partial}$ every year in the sample, and operates as the pipeline with the largest scale on the frontier each year. In addition, TCO's output quantities increase every year, and delivery volume more than doubles from 2007 to 2018. This could be because of TCO's route through many major east coast metropolitan areas. More importantly though, their position on the eastern seaboard allows TCO to pass through every state on the east coast with an existing or proposed LNG terminal. Thus, as shale production increases, TCO remains a highly productive pipeline because of increasing LNG exports.

Finally, pipeline age and initial design can inhibit increases in technical efficiency. Texas Eastern Pipeline Company (TETCO) is similarly located to TGT and transports gas from Marcellus shale deposits to the gulf coast of Louisiana. Along with ANR and TGT, TETCO was one of the first pipelines to become bidirectional and reverse flow to export gas from the Midwest for LNG export. However, the pipeline dates back to World War II and was initially designed as a crude oil pipeline. The pipeline frequently experiences outages due to accidents and pipeline ruptures. This in turn increases OPEX inhibiting TETCO from achieving full efficiency.

Bibliography

Aivazian, V., Callen, J., Chan, M.W.L., Mountain, D., 1987. Economies of scale versus technological change in the natural gas transmission industry. The MIT Press 69, 556–561.

Annibali, V., Burchill, J., Cabrales, R.O., Ellsworth, C., Gillespie, R., Hartman, D., Haymes, A., Hinrichs, L., Kohut, K., Ly, J., Michals, S., Pollonais, S., Primosch, E., Rapp, A., Royster, R., Schaub, P., Sillin, J., Stertz, R., Russo, T., White, C., Zakaria., R., 2012. 2012 state of the markets report. URL: https://www.ferc.gov/sites/default/files/2020-05/2012-som-final.pdf.

Apon, A.W., Ngo, L.B., Payne, M.E., Wilson, P.W., 2015. Assessing the effect of high performance computing capabilities on academic research output. Empirical Economics 48, 283–312.

Bernstein, D.H., 2020. An updated assessment of technical efficiency and returns to scale for U.S. electric power plants. Energy Policy 147.

Callen, J., 1978. Production, efficiency, and welfare in the natural gas transmission industry. American Economic Review 68, 311–323.

Cazals, C., Florens, J.P., Simar, L., 2002. Nonparametric frontier estimation: a robust approach. Journal of Econometrics 106, 1–25.

Cuddington, J.T., Wang, Z., 2006. Assessing the degree of spot market integration for U.S. natural gas: evidence from daily price data. Journal of Regulatory Economics 29, 195–210.

Daouia, A., Simar, L., Wilson, P.W., 2017. Measuring firm performance using nonparametric quantile-type distances. Econometric Theory 14, 783–793.

Daraio, C., Simar, L., 2005. Introducing environmental variablesin nonparametric frontier models: A probabilistic approach. Journal of Productivity Analysis 24, 93–121.

Daraio, C., Simar, L., 2007a. Advanced Robust and Nonparametric Methods in Efficiency Analysis. New York, Springer.

Daraio, C., Simar, L., 2007b. Conditional nonparametric frontier models for convex and nonconvex technologies: a unifying approach. Journal of Productivity Analysis 28, 13–32.

Daraio, C., Simar, L., Wilson, P., 2018. Central limit theorems for conditional efficiency measures and tests of the 'separability' condition in non-parametric, two-stage models of production. Econometrics Journal 21, 170–191.

Davis, L., 2017. Evidence of a decline in electricity use by U.S. households. Energy Institute at Haas Working Paper 279 37, 1098–1105.

Deprins, D., Simar, L., Henry, T., 1984. Measuring Labor-Efficiency in Post Offices. pp. 285–309. doi:10.1007/978-0-387-25534-7_16.

Doane, M.J., Spulber, D.F., 1994. Open access and the evolution of the U.S. spot market for natural gas. The Journal of Law Economics 37, 477–517.

Dreskin, J., Boss, T., 2010. Interstate natural gas pipeline efficiency. URL: https://www.ingaa.org/file.aspx?id=10929.

Ellig, J., Giberson, M., 1993. Scale, scope, and regulation in the texas gas transmission industry. Journal of Regulatory Economics 5, 79–90.

Elliott, G., Rothenberg, T., Stock, J., 1996. Efficient tests for an autoregressive unit root. Econometrica 64, 813–836.

Enders, W., 2009. Applied Econometric Time Series, 3rd Edition. New York, John Wiley & Sons.

Enke, S., 1951. Equilibrium among spatially separated markets: Solution by electric analogue. Econometrica 19, 40–47.

Färe, R., Grosskop, S., Pasurka, C., 1989. The effect of environmental regulations on the efficiency of electric utilities: 1969 versus 1975. Applied Economics 21, 235–235.

Färe, R., Grosskopf, S., Lovell, C.A.K., 1985. The Measurement of Efficiency of Production. Boston: Kluwer-Nijhoff Publishing.

Farrell, M., 1957. The measurement of productive efficiency. Journal of the Royal Statistical Society A, 253–281.

FERC, 1992. Order 636. URL: https://www.ferc.gov/sites/default/files/2020-05/rm91-11-000.txt.

von Geymueller, P., 2009. Static versus dynamic DEA in electricity regulation: the case of us transmission system operators. Central European Journal of Operations Research volume 17, 398–413.

Gilbert, A., Wilson, P.W., 1998. Effects of deregulation on the productivity of korean banks. Journal of Economics and Business 50, 133–155.

Granderson, G., 2000. Regulation, open-access transportation, and productive efficiency. Review of Industrial Organization 16, 251–266.

Growitsch, C., Hess, B., 2009. On the sensitivity of us electric utilities' efficiency estimates – a distance function approach. Applied Economics Letters 16, 847–851.

Hess, B., 2000. Evaluating the efficiency effects of industry consolidation: Evidence from us interstate pipeline companies. International Journal of Energy Sector Management 4, 462–481.

Jamasb, T., Newbery, D., Pollitt, M., Triebs, T., 2006. International benchmarking and regulation of european gas transmission utilities. URL: https://www.acm.nl/sites/default/files/old_publication/bijlagen/6418_ERGEG_Cost_benchmark_final.pdf.

Jamasb, T., Pollitt, M., Triebs, T., 2008. Productivity and efficiency of us gas transmission companies: A european regulatory perspective. Energy Policy 36, 3398–3412.

Kahn, A.E., 1971. The Economics of Regulation: Principles and Institutions, vol. II. Wiley.

King, M., Cuc, M., 1996. Price convergence in North American natural gas spot markets. The Energy Journal 17, 17–42.

Kneip, A., Park, B.U., Simar, L., 1998. A note on the convergence of nonparametric DEA efficiency measures. Econometric Theory 14, 783–793.

Kneip, A., Simar, L., Wilson, P., 2021. Conical hull estimators of general technologies, with applications ot returns to scale and Malmquist productivity indices. In progress.

Kneip, A., Simar, L., Wilson, P.W., 2015. When bias kills the variance: Central limit theorems for DEA and FDH efficiency scores. Econometric Theory 31, 394–422.

Kneip, A., Simar, L., Wilson, P.W., 2016. Testing hypotheses in nonparametric models of production. Journal of Business & Economic Statistics 34, 435–456.

Kwiatkowski, D., Phillips, P.C., Schmidt, P., Shin, Y., 1992. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? Journal of Econometrics 54, 159–178. doi:https://doi.org/10.1016/0304-4076(92)90104-Y.

Li, Q., Lin, J., Racine, J.S., 2013. Optimal bandwidth selection for nonparametric conditional distribution and quantile functions. Journal of Business and Economic Statistics 31, 57–65.

Moss, D.L., 2008. Competition Policy and Merger Analysis in Deregulated and Newly Competitive Industries. Edward Elgar Publishing. chapter Natural Gas Pipelines: Can Merger Enforcement Preserve the Gains from Restructuring?

Nadel, S., Elliott, N., Langer, T., 2015. Energy efficiency in the united states: 35 years and counting. URL: https://www.aceee.org/sites/default/files/publications/researchreports/e1502.pdf.

Nadel, S., Herndon, G., 2014. The future of the utility industry and the role of energy efficiency. URL: https://www.aceee.org/sites/default/files/publications/researchreports/u1404.pdf.

Nadel, S., Young, R., 2014. Why is electricity use no longer growing? URL: https://www.naesco.org/data/news/documents/ACEEE%20White%20Paper,%20Electricty%20Use%20Declining,%202-25-14.pdf.

Nieswand, M., Cullmannt, A., Neumann, A., 2010. Overcoming data limitations in nonparametric benchmarking: Applying pca-dea to natural gas transmission. Review of Network Economics 9, 3398–3412.

Omrani, H., Beiragh, R.G., Kaleibari, S.S., 2015. Performance assessment of iranian electricity distribution companies by an integrated cooperative game data envelopment analysis principal component analysis approach. Electrical Power and Energy Systems 64, 617–625.

O'Neill, R.P., 2005. Network Access, Regulation And Antitrust. Routledge. chapter Natural Gas Pipelines.

Park, B.U., Simar, L., Weiner, C., 2000. FDH efficiency scores from a stochastic point of view. Econometric Theory 16, 855–877.

Phillips, P., Perron, P., 1988. Testing for a unit root in time series regression. Biometrika 75, 335–346.

Ray, S.C., Desli, E., 1997. Productivity growth, technical progress, and efficiency change in industrialized countries: Comment. The American Economic Review 87, 1033–1039.

Said, S.E., Dickey, D.A., 1984. Testing for unit roots in autoregressive-moving average models of unknown order. Biometrika 71, 599–607.

Samuelson, P.A., 1952. Spatial price equilibrium and linear programming. The American Economic Review 42, 283–303.

Sarkis, J., Cordeiro, J.J., 2012. Ecological modernization in the electrical utility industry: An application of a bads–goods DEA model of ecological and technical efficiency. European Journal of Operational Research 219, 386–395.

Scarcioffolo, A.R., Etienne, X.L., 2019. How connected are the U.S. regional natural gas markets in the post-deregulation era? evidence from time-varying connectedness analysis. Journal of Commodity Markets 15, 1–18.

Shephard, R., 1970. Theory of Cost and Production Functions. Princeton: Princeton University Press.

Sickles, R.C., Streitwieser, M.L., 1992. Technical inefficiency and productive decline in the u.s. interstate natural gas pipeline industry under the natural gas policy act. Journal of Productivity Analysis 3, 119–133.

Simar, L., Wilson, P., 2011a. Estimation and inference in two-stage, semi-parametric models of production processes. Journal of Econometrics 136, 31–64.

Simar, L., Wilson, P., 2011b. Two-stage DEA: Caveat emptor. Journal of Productivity Analysis 36, 205–2018.

Simar, L., Wilson, P., 2020. Hypothesis testing in nonparametric models of production using multiple sample splits. Journal of Productivity Analysis 53, 287–303.

Simar, L., Wilson, P.W., 1998. Productivity growth in industrialized countries. discussion paper 9810, Institut de Statistique, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Simar, L., Wilson, P.W., 1999. Estimating and bootstrapping malmquist indices. European Journal of Operational Research 115, 459–471.

Simar, L., Wilson, P.W., 2019a. Central limit theorems and inference for sources of productivity change measured by nonparametric malmquist indices. European Journal of Operational Research 277, 756–769.

Simar, L., Wilson, P.W., 2019b. Technical, allocative and overall efficiency: Estimation and inference. European Journal of Operational Research 282, 1164–1176.

Stigler, G.J., Sherwin, R.A., 1985. The extent of the market. The Journal of Law Economics 28, 555–585.

Takayama, T., Judge, G.G., 1964. Equilibrium among spatially separated markets: A reformulation. Econometrica 32, 510–524.

Takayama, T., Judge, G.G., 1971. Spatial and temporal price and allocation models. North-Holland, Amsterdam Pub. Co.

US Energy Information Administration, 2009. Electric power annual 2007.

US Energy Information Administration, 2020. Electric power annual 2019.

US Energy Information Administration, March 2021. Monthly energy review report.

Werden, G.J., Froeb, L.M., 1993. Correlation, causality, and all that jazz: The inherent shortcomings of price tests for antitrust market delineation. Review of Industrial Organization 8, 329–353.

Wilson, P., 2011. Exploring research frontiers in contemporary statistics and econometrics, Berlin: Springer-Verlag. chapter Asymptotic Properties of Some Non-Parametric Hyperbolic Efficiency Estimators, pp. 115–150.

Wilson, P., O'Loughlin, C., 2021. Benchmarking the performance of U.S. municipalities. Empirical Economics 60, 2665–2700.

Wilson, P.W., 2008. Fear 1.0: A software package for frontier efficiency analysis with R. Socio-Economic Planning Sciences 42, 247–254.

Wilson, P.W., 2018. Dimension reduction in nonparametric models of production. European Journal of Operational Research 267, 349–367.