HARVARD Kennedy School
MOSSAVAR-RAHMANI CENTER
for Business and Government 30 YEARS

Experiential and Social Learning in Firms: The Case of Hydraulic Fracturing in the Bakken Shale

Thomas R. Covert

May 2014

M-RCBG Associate Working Paper Series | No. 27

Mossavar-Rahmani Center for Business & Government
Weil Hall | Harvard Kennedy School | www.hks.harvard.edu/mrcbg

Experiential and Social Learning in Firms: The Case of Hydraulic Fracturing in the Bakken Shale

Thomas R. Covert*

May 17, 2014

Abstract

Learning how to utilize new technologies is a key step in innovation, yet little is known about how firms actually learn. This paper examines firms' learning behavior using data on their operational choices, profits, and information sets. I study companies using hydraulic fracturing in North Dakota's Bakken Shale formation, where firms must learn the relationship between fracking input use and oil production. Using a new dataset that covers every well since the introduction of fracking to this formation, I find that firms made more profitable input choices over time, but did so slowly and incompletely, only capturing 67% of possible profits from fracking at the end of 2011. To understand what factors may have limited learning, I estimate a model of fracking input use in the presence of technology uncertainty. Firms are more likely to make fracking input choices with higher expected profits and lower standard deviation of profits, consistent with passive learning but not active experimentation. Most firms over-weight their own information relative to observable information generated by others. These results suggest the existence of economically important frictions in the learning process.

1 Introduction

New technologies are important contributors to economic growth[1], but little is known about how firms learn to profitably use them. While there is longstanding evidence that firms learn from their own experiences (learning-by-doing), and from others (social learning), the specific actions that firms actually take in learning are not well understood. Models of learning predict that firms efficiently analyze information about new technologies, invest in experiments to create new information, and incorporate information generated by other firms.[2] However, to test these models,

*Harvard Business School and Harvard University Department of Economics; tcovert@hbs.edu. I am grateful to Paul Asquith, Bharat Anand, Greg Lewis, Ariel Pakes and Parag Pathak for their guidance and encouragement. I thank Michael Luca, Thomas G. Wollman, Richard Sweeney, Bryce Millet-Steinberg, Stephanie Hurder, Alex Peysakhovich, Evan Herrnstat, Joseph Shapiro, Hugh Daigle, Heath Flowers, Chris Wright, Leen Weijers and the Harvard Industrial Organization, Environment Economics, and Work-In-Progress lunches for their helpful comments and discussions. Funding from the Harvard Business School Doctoral Programs, a Harvard University Dissertation Completion Fellowship and the Sandra Ohrn Family Foundation is gratefully acknowledged.

[1] See, for example, Arrow (1962), Romer (1986) and Kogan et al. (2012)

[2] See Aghion et al. (1991) in the single agent context and Bolton and Harris (1999) in the multi-agent context.

it is necessary to observe data on the information that firms have, which is difficult to acquire in many empirical settings. This paper tests predictions of learning models for the first time, using data on oil companies that employ hydraulic fracturing (fracking) in the North Dakota Bakken Shale. The data covers operational choices, profits, and measures of the information firms had when making choices. The oil companies in this data learn to use fracking more profitably over time, but are slow to respond to new information, avoid experiments and underutilize data provided by their competitors.

Fracking is a useful context to study learning behavior in firms. The profit maximizing choice of fracking inputs may vary across drilling locations in unpredictable ways, so firms must empirically learn this relationship over time and change their behavior accordingly. In North Dakota, firms can learn about fracking from a wealth of publicly available information. Regulators collect and publicly disseminate unusually detailed, well-specific information about oil production and fracking input choices. Moreover, regulators delay dissemination until 6 months after a well is fracked, making it possible to measure differences in knowledge about fracking across firms. The industry is not concentrated, which motivates studying learning as a single agent problem. During the time period I study, there are 70 active firms, the market share of the largest firm is only 13% and the combined share of the five largest firms is under 50%. The two main inputs to fracking, sand and water, are commodities, as is the output of fracking, crude oil. The unique regulation and industry structure make fracking in the Bakken shale an unusually compelling setting for studying learning in firms. Moreover, the stakes in fracking are large. Using a production function, I estimate that the average NPV of profits per well for actual fracking choices is about $12.8 million, while the average profit for each well's most profitable choice is $24.5 million. Since the regulator in North Dakota expects that 40,000 wells will eventually be fracked over the next 18 years, the potential for lost profits from inefficient learning is substantial.[3]

Learning-by-doing and social learning are both important in this context. Between 2005 and 2006, the average well is fracked by a firm that had fracked only a single well before. By 2011, the average well is fracked by a firm that had previously fracked 117 wells. Thus, firms can learn from an increasing amount of their own experience. However, North Dakota's disclosure laws make it possible for firms to study their competitors' data. Between 2005 and 2006, the average well is

[3]See https://www.dmr.nd.gov/oilgas/presentations/NDOGCPC091013.pdf

fracked by a firm that can observe 10 wells previously fracked by other firms, a number which rises to 1,783 in 2011. As a result, most of the information firms have comes from others, and firms have the ability to socially learn.

The data I collect from the regulator in North Dakota is well suited to estimate the relationship between location, fracking, and oil production. I observe the complete operating history of every firm and every well they frack in the Bakken Shale between January 2005 and December 2011 (70 firms and 2,699 wells), so there is no possibility for survivorship bias. The data contains precise measurements of a well's production, location and most important fracking inputs, so there are no endogenous omitted variables. Moreover, the engineering requirements for wells drilled into the Bakken prevent firms from selecting observed fracking inputs on the basis of information I do not observe. Thus, the standard endogeneity problem in production function estimation is unlikely to be a concern.

Using the data I collect, I semi-parametrically estimate a production function for fracking which represents what firms need to learn. These estimates show that amount of oil in the ground and the sensitivity of its production to fracking both vary over space, a result that is consistent with geological theory and data. Estimates made using subsets of the data that were available to firms when they were fracking have qualitatively similar results, suggesting that firms could have used this data to make informed fracking decisions. The estimated production function fits the data well and is stable across robustness tests.

I use this production function to measure how quickly firms learn. Wells fracked in 2005 capture only 16% of the profits that optimally fracked wells would have produced. However, profit capture grows almost monotonically over time, with firms capturing 68% of maximal profits in 2011. This growth is driven by improved fracking input choices, with firms gradually increasing their use of sand and water towards optimal levels over time. I interpret this upward trend in the profitability of fracking input choices as evidence for learning.

Existing research measures learning from upward trends in *productivity*, or residual production that is not explained by input choices. I test for productivity based learning by analyzing the growth of estimated production function residuals over time. Wells fracked in 2011 are 34% more productive than wells fracked in 2005, suggesting some role for productivity-driven learning. However, the majority of the growth in productivity occurs by 2008, and there is no statistically significant difference in productivity between 2008 and 2011. This contrasts with the fraction of profits captured, which increases monotonically over time, and from 44% to 67% between 2008 and 2011. Thus, during 2008-2011, when 95% of wells in my data are fracked, there is little productivity growth, even though there is substantial growth in the fraction of profits captured. These results help clarify the difference between models of learning in which knowledge is a direct input in the production function, and a model of learning about the production function itself.

To see if firms are using their information to make better fracking choices over time, I estimate *ex ante* production functions for each well, using the subset of the data that firms had when they were making choices. I use these estimates to compute *ex ante* profits. Though firms capture 76% of *ex ante* optimal profits in 2007, they capture only 68% in 2011. The fraction of *ex ante* profits falls because initial fracking input choices are close to the (then) estimated optimal levels, but optimal levels subsequently change more quickly than choices do.

Theory predicts that firms may sacrifice estimated profits in the current period by experimenting in order to generate information for the future. To test if experimenting behavior can rationalize the decline in the fraction of estimated *ex ante* optimal profits captured, I estimate a simple model of fracking input choice under technology uncertainty. In this model, firms have preferences over the expectation and standard deviation of their *ex ante* estimates of profits for a fracking input choice. If firms are experimenting, they should be empirically more likely to choose inputs with higher standard deviations of profit. I do not find support for this theory. Firms are more likely to select fracking designs with higher expected profits and *lower* standard deviation of profits. Firms are indifferent between a \$0.60-\$0.98 increase in expectation of profits and a \$1 reduction in the standard deviation of profits.

My calculation of the expectation and standard deviation of profits assumes that firms equally learn from their own and others' experiences. However, firms may treat the social portion of their data differently than the data they directly experience, and in the process form different estimates of profits than what I calculate. To account for this possibility, I modify my fracking input choice model to allow for weighted production function estimates estimates. I use this model and data on firms' choices to estimate the weight they place on their own experiences relative to their competitors' experiences. Most firms place more weight on their own experiences than their competitors' experiences. Even after controlling for weighted estimates, firms still prefer fracking choices with lower standard deviations and higher means.

This paper finds that firms are reluctant to experiment and ignore valuable data generated by their competitors. These firms are not unsophisticated or under-incentivized. They have access to capital markets, are managed by executives with engineering and business education and are the primary equity holders in the wells they frack. These findings stand in contrast to some theories of efficient learning behavior by rational agents, which predict that firms will take experimental risk and learn from all the information they have.

In addition to its usefulness as a laboratory to study learning, fracking plays a prominent role in current public policy debates about growing oil production and its effects on the environment. The US EIA reports that fracking has caused national oil production to grow 22% since 2009, reversing almost two decades of declines.[4] There is early evidence that fracking-driven resource booms have affected housing prices[5] and local banking markets.[6] However, there are growing concerns about the potential for fracking to negatively affect the quantity and quality of local ground water supplies,[7] which the US EPA is currently studying.[8] In response to these concerns, federal regulators have proposed significant increases to disclosure requirements for fracking operations.[9] Though this push for increased transparency around fracking is driven by environmental concerns, new disclosure regulations may also have an impact on learning by increasing the availability of data.

Finally, the Bakken Shale unlikely to be the last oil and gas formation where fracking and the learning it requires play an important role. Fracking is currently in use in the Eagle Ford and Barnett Shales in Texas, the Woodford Shale in Oklahoma, and several locations in Canada. International oil companies are now developing shale resources in Argentina, Poland and China. The results of this paper may be useful to both policy makers and oil & gas companies alike in regulating access to information and understanding the benefits of more efficient learning behavior.

[4] http://www.eia.gov/todayinenergy/detail.cfm?id=13251

[5] Muehlenbachs et al. (2012) find that housing prices increase after the introduction of fracking to a community, except for houses that depend on groundwater.

[6] See Gilje (2012)

[7] See Vidic et al. (2013) for an overview

[8] See http://www2.epa.gov/hfstudy

[9] See Deutsch (2011).

1.1 Related literature

Firms in many industries and time periods have become more productive by learning from their own experiences. Researchers studying the manufacturing of World War II ships (Thornton and Thompson 2001), aircraft (Benkard 2000) and automobiles (Levitt et al. 2012) have documented an important empirical regularity: with the same inputs, firms are able to produce more output as they accumulate experience in production.[10] That is, they learn by doing (LBD). The LBD result that productivity is correlated with experience suggests that the knowledge embedded in this experience is a direct input to the production function. Changes over time in capital, labor and materials are thus interpreted as profit-maximizing responses to increases in productivity, not changes in specific knowledge. In this paper, I instead assume that the production technology itself is initially unknown and that experience has no direct impact on production. As firms accumulate experience in fracking, they acquire more data about the fracking production function, perform inference on this data, and make more profitable input choices on the basis of their inference. This is similar to the approach taken by Foster and Rosenzweig (1995) and Conley and Udry (2010) in the development literature.

Economic theory predicts that when firms are learning about a new technology, they face a tradeoff between "exploration" and "exploitation" (or experimentation). Firms may actively learn by experimenting with fracking input choices that have highly uncertain profits or passively learn by exploiting choices with high expected profits. Except in the simplest theory models, the optimal amount of experimentation and exploitation is a challenging problem to solve. However, most models of learning predict that forward-looking firms will always do some experimenting. In the single agent context, Aghion et al. (1991) show that forward-looking firms will almost always do some exploration. Bolton and Harris (1999) find a similar result in the multi-agent context. Wieland (2000) employs computational methods to characterize the costs and benefits of exploration, finding that firms who only exploit can get stuck, and repeatedly choose suboptimal actions. To my knowledge, this paper is the first to empirically measure the amount of experimenting that firms perform in a learning situation.

This paper adds to a wide literature documenting the existence and importance of social learning between firms. Most of this evidence is in agricultural settings. Ryan and Gross (1943),

[10]This phenomenon has also been observed by Anand and Khanna (2000) in the corporate strategy setting.

Griliches (1957) and Foster and Rosenzweig (1995) demonstrate that farmers learn about the benefits of adopting new technologies from the experiences of their neighbors. Conley and Udry (2010) show that farmers in Ghana learn about the efficient use of fertilizer from other farmers in their social networks, demonstrating that social learning in agriculture is not limited to the adoption decision. Social learning has also been observed in manufacturing. During the construction of WWII ships, Thornton and Thompson (2001) find that firms benefited from accumulated experience by other firms. Similarly, Stoyanov and Zubanov (2012) find evidence that firms in Denmark became more productive after hiring workers away from their more productive competitors.

Finally, this paper is complementary to the existing literature on learning behavior by oil and gas companies. Levitt (2011) shows that the observed temporal and spatial patterns of the oil exploration process match the predictions of a forward-looking learning model. In a study of offshore drilling, Corts and Singh (2004) show that as oil companies gain experience with their service contractors, they learn to trust them and tend to select low-powered contracting terms. Kellogg (2011) studies this phenomenon in the on-shore setting and shows that oil companies and their service contractors jointly learn to be more productive in drilling as they accumulate shared operating experience.

The remainder of the paper is as follows. In Section 2, I provide institutional background on fracking in North Dakota and describe the data I have on operational choices, production results and information sets. Next, in Section 3, I estimate a production function model of fracking and evaluate its ability to predict oil production. In Section 4, I use the production function estimates to test if firms learned to make more profitable fracking choices over time. In Section 5, I specify and estimate the model of fracking input choice under technology uncertainty. Finally, I conclude in Section 6.

2 Institutional Background and Data

2.1 Fracking and US Oil Production

The hydraulic fracturing of shale formations, like the Bakken, has had a profound impact on the fortunes of energy producing states and the US as a whole. In 2009, the US Energy Information Administration reported that national oil production grew 6.8% year-over-year, the first increase in over two decades.[11] This trend has continued and between 2009 and 2012, national oil production increased 21.7%. Three states represent the majority of this growth: Texas, Oklahoma and North Dakota. This paper focuses on what has happened in North Dakota.

In March 2012, North Dakota surpassed Alaska to become the second most prolific oil producing state in the US, after Texas. Between January 2005 and July 2013, oil production in North Dakota increased from 93,000 barrels (bbl) per day to 874,000 bbl per day. During the same time period, total US oil production increased from 5.63 million bbl per day to 7.48 million bbl per day, meaning that increased production in North Dakota amounted to 42% of the net increase in total production. Though production increased in Texas and Oklahoma as well, it is striking that North Dakota went from producing less than 2% of national oil production to almost 12% in the span of 8 years.[12] This vast expansion in North Dakotan oil production coincided with the introduction of fracking to the Bakken Shale formation.

2.2 The Bakken Shale and Hydraulic Fracturing

The Bakken Shale spans 200,000 square miles in North Dakota, Montana and Saskatchewan.[13] It lies 10,000 feet underground and contains 3 distinct layers: the upper Bakken member (a shale layer), the middle Bakken member (a layer of sandstone and dolomite), and the lower Bakken member (also a shale layer). The US Geological Survey estimates that the upper and lower shales together contain 4.6 billion bbl of recoverable oil.[14] Though the middle Bakken member is not formed from organic materal and as such does not generate any oil of its own, firms typically drill horizontally through it and use hydraulic fracturing, or "fracking", to make contact with the oil bearing shales above and below, as shown in Figure 1.

Fracking is the process of pumping a mix of water, sand and chemicals into a well at high pressures. The high pressure of the mix fractures the surrounding rock and the sand in the mix props those fractures open.[15] The fractures created by fracking the middle Bakken radiate outwards

[11] See the EIA Annual Energy Review, 2009. http://www.eia.gov/totalenergy/data/annual/archive/038409.pdf

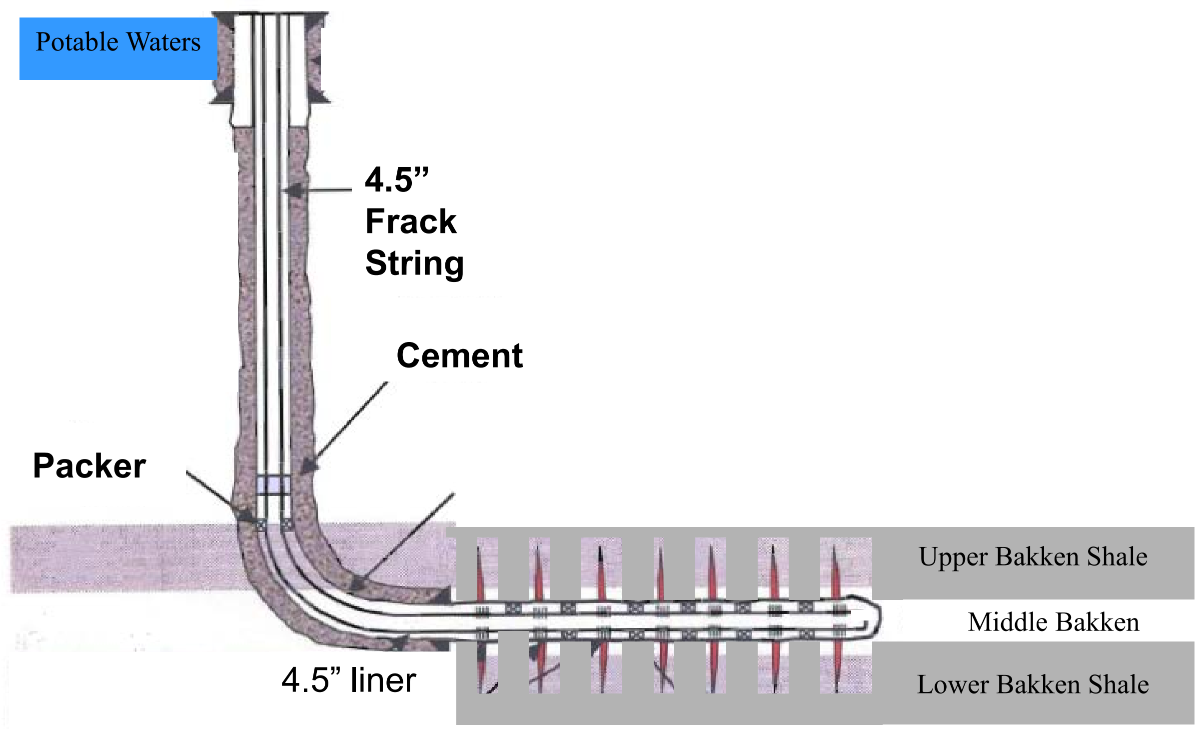
[12] Texas also experienced production significant production increases during that same time period, though from a much higher base level (from 1.08 million bbl per day to 2.62 million bbl per day, a 143% increase). Much of this increase can also be attributed to the technology changes described here. Operators applied fracking technology successfully to the Eagle Ford, Permian and Barnett shales.

[13] See Gaswirth (2013)

[14] See Gaswirth (2013)

[15] Chemicals reduce mineral scaling, inhibit bacterial growth, reduce wear and tear on fracking hardware and increase the buoyancy of sand in the fracking mixture. See http://www.fracfocus.org for an overview.

Figure 1: Diagram of a Hydraulically Fractured Bakken Shale well



Adapted from Hicks (2012)

into the upper and lower Bakken shales, as shown in Figure 1. These fractures both serve as a conduit between the wellbore in the middle Bakken and the upper and lower shales, and also increase the permeability of the upper and lower shales.

Permeability is a geological measure of the ease at which oil naturally flows through rock. The upper and lower shales are unusually impermeable, making it impossible for the oil they contain to naturally reach a wellbore drilled through the middle member. Without fracking, wells drilled into the middle member will not produce profitable quantities of oil.[16] After fracking, oil inside the lower and upper shales can more easily travel through the new fractures into the wellbore in the middle member.

Firms choose how much water and sand to use in fracking and this choice can have a large impact on the profitability of a well. Wells fracked with more sand and water may produce more oil than wells fracked with less, but fracking is expensive, and water and sand represent the bulk of this expense. In 2013, the reported costs of fracking range from $2-5 million per well, out of total well costs of $9 million.[17] Thus, to maximize profits, firms must balance the benefits of sand and water use in fracking with their costs. This requires firms to understand the relationship between oil production and fracking inputs, and it is unlikely that firms initially knew this relationship. The first Bakken wells to be developed with fracking were not drilled until 2005, and at the time, the firms developing those wells had limited experience in fracking shale formations.[18] Without prior experience, firms had to learn how to use fracking by doing it themselves or by studying their competitors.

There is now a growing literature about best practices in fracking. Petroleum engineers have found that wells fracked with more water and sand are often more productive than similar wells with less aggressive fracking treatments.[19] However, there is also evidence that the relationship between oil production and fracking inputs is not necessarily monotonic and that it varies over

[16] See Hicks (2012)

[17] See Hicks (2012)

[18] Fracking was first successfully used in shale formations in the 1990s. Under the hunch that permeability issues could eventually be resolved through the use of fracking, Mitchell Energy worked for years on its own and with the help of the US Department of Energy to learn how to apply fracking technology to the Barnett shale in Texas. They succeeded in 1997. See Michael Shellenberger and Jenkins (2012). Two firms active in North Dakota, EOG and XTO, were active in the Barnett as well. However, the Barnett Shale is different from the Bakken. Barnett wells are drilled directly into the shale layer, and produce natural gas instead of oil. It is unlikely that any knowledge that these firms may have had about fracking in the Barnett was useful in the Bakken.

[19] See Shelley et al. (2012)

drilling locations.[20] Research documenting these results was not publicly available to firms during the time period I study, which means that firms faced a complicated learning problem.

2.3 The Information Environment in North Dakota

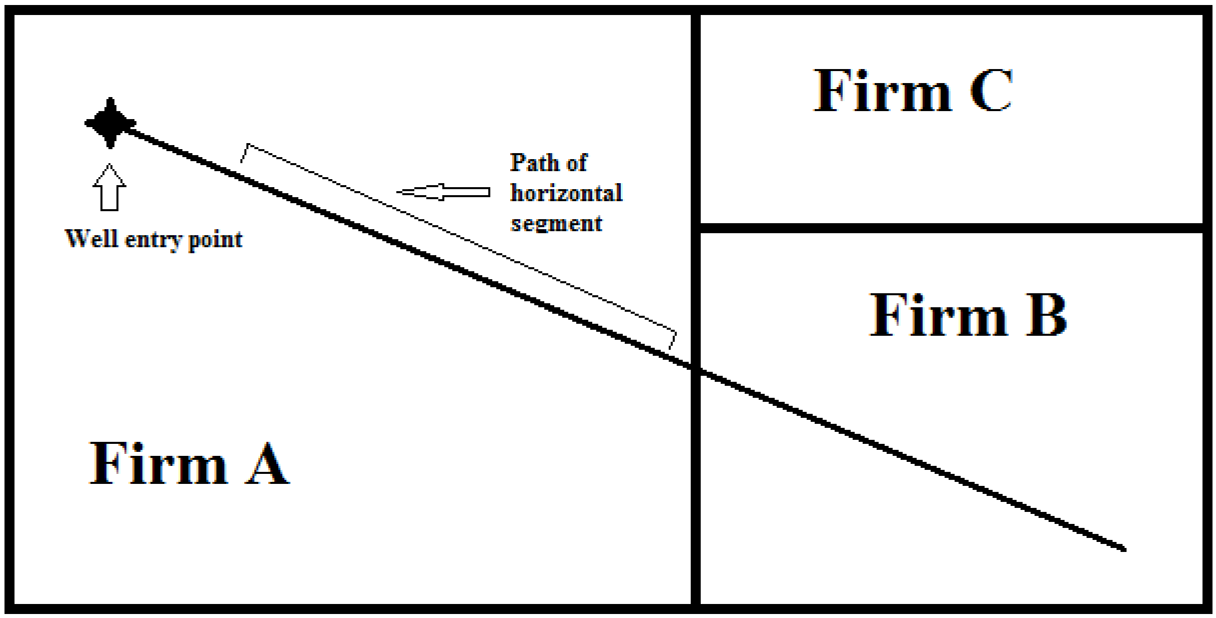
Firms in North Dakota can learn about the relationship between oil production, location, and fracking inputs from the past experiences of other firms. After a firm fracks a well, the oil and gas regulator in North Dakota requires the firm to submit a well completion report, detailing the well's horizontal length, location and fracking inputs. Additionally, the regulator and tax authorities require the firm to submit audited production records on a monthly basis. The regulator publishes this information on the internet, making it easy for firms to learn information about every previously fracked well in the state, including information about wells that they took no part in developing.

North Dakota's well confidentiality laws generate a 6 month delay between when firms submit well completion reports and when the regulator makes them public. This delay creates differences across firms in what wells they can learn from at each point in time, as the operating firm of a well has a temporary knowledge advantage over other firms. However, the ownership structure of mineral rights in a well mitigates some of these differences. Mineral rights for a well are often owned by many separate firms. Every firm that owns mineral rights in the area spanned by a well is entitled to pay a share of the capital expenditures needed to develop the well in exchange for a share of the revenue generated by the well. The firm with the largest mineral rights claim in a well is called the "operator", and it retains all control rights, including the choice of the well's fracking inputs. The remaining owners of mineral rights are called "non-operating participants". Figure 2 depicts a hypothetical ownership situation for a well in the Bakken. The land spanned by the well is a 2 mile by 1 mile rectangle, called a "spacing unit". Within this spacing unit, Firm A has the largest mineral rights claim, followed by firms B and C. The wellhead enters the ground in A's claim and the horizontal segment passes through B's claim. Though the well does not directly pass through C's claim, it is close enough to C's claim that it may be drawing oil from the claim. While A retains control rights, B and C must pay their respective share of capital expenditures.[21]

[20] See Baihly et al. (2012)

[21] Firms can choose to opt out of a spacing unit, but that does not allow them to operate another well within the spacing unit, so opt outs are rare.

Figure 2: Diagram of a hypothetical spacing unit



Non-operating participants have immediate access to a well's completion report.[22] This means that non-operating participants in a well are not subject to well confidentiality rules and thus observe information regarding a well before the public does.

2.4 Data

2.4.1 Well Characteristics and Production History

I have collected operating and production data for every well targeting the Bakken shale formation in North Dakota that was fracked between January 1, 2005 and December 31, 2011. This data is reported by oil companies to the North Dakota Industrial Commission (NDIC), and the NDIC publishes their submissions on the internet. For each well i, I observe the location of its wellhead in latitude lat_i and longitude lon_i coordinates, its horizontal length H_i, the mass of sand S_i and volume of water W_i per foot of horizontal length used in fracking and the identity of the operating firm f_i. Additionally, I observe oil production Y_{it} for well i in it's t-th month of existence and the number of days D_{it} during that month that the well was actually producing. Let X_{it} denote

[22] See Larsen (2011)

the set (H_i, f_i, D_{it}) and let Z_i denote the set (S_i, W_i, lat_i, lon_i). Then the dataset (Y_{it}, X_{it}, Z_i) has a panel structure, where i indexes wells and t indexes well-specific timing. Though I only study wells fracked during 2005-2011, I have production data through February 2013, making it possible to study the performance of all wells for at least a year. While the production history is reported electronically on the NDIC website, the static well characteristics are stored in PDF format, so much of this dataset was entered into the computer manually. I also observe the "township" τ_i that the wellhead lies in. Townships are 6 mile by 6 mile squares, defined by the US Geological Survey and are a standard measure of location in the oil & gas business. There are 272 townships in North Dakota with Bakken wells during 2005-2011. I have also collected the geographic boundaries of the spacing units for every well. This data comes from various portions of the NDIC website.

Though most of the data I collect from the NDIC is self reported by firms, there are two reasons why it is likely to be truthfully reported. First, oil and gas regulations in North Dakota specify explicit penalties for failure to report required information and false reporting, including fines of up to $12,500 per day per offense and felony prosecution.[23] Second, because operators wish to collect payment for capital expenditures from their non-operating partners, they must share the documentation and billing they receive from their service contractors. If operators were to report data to the NDIC that was at odds with what they had shared with their non-operating partners, they might jeopardize their ability to collect payment.

Table 1 reports the cross-sectional distribution of well characteristics and oil production in the first year. There is substantial variation across wells in both fracking input use and oil production. The 75th percentiles of sand, water and oil production are more than double their respective 25th percentiles. This variation will be important later on in estimating the relationship between oil production and fracking inputs. Most wells have horizontal segments that are 9,000 feet or longer. The length of a well's horizontal segment is determined by the size of its spacing unit. Though not shown in the table, approximately 75% of wells have rectangular spacing units that are two miles wide and one mile tall. The remaining 25% have 1 mile square spacing units. The average well produces almost 11 bbl per foot of horizontal length in its first year. Since the price of oil averaged $76 per bbl during 2005-2011, the value of production in the first year for the average well is worth $6.6 million. Most wells tend to produce on the majority of days during a month,

[23]See Section 38-08-16 in the NDIC Rulebook.

Table 1: Summary Statistics

Variable	Mean	Std. Dev	P25	P50	P75	N
lbs sand per foot	265.02	138.68	158.27	264.53	378.66	2,699
gals water per foot	188.87	110.73	100.31	181.70	249.52	2,699
horizontal feet in length	8,040	2,138	5,600	9,135	9,518	2,699
avg producing days per month	26.80	2.99	25.90	27.56	28.67	2,699
oil production per foot in first year	10.86	8.95	5.38	8.39	12.99	2,699
# non-operating participants	3.00	2.50	1.00	3.00	4.00	2,699
# past wells fracked by operator	80	82	16	49	125	2,699
# past wells fracked by others	1,089	658	511	1,062	1,698	2,699

Table 2: Summary Statistics by Year

		2005	2006	2007	2008	2009	2010	2011
# wells fracked		10	20	94	352	463	691	1,069
# active townships		9	17	37	102	132	179	231
# active firms		5	11	17	28	34	47	49
Sand	Average	94.50	136.88	134.64	180.00	212.75	308.82	302.79
	Std. Dev	22.01	152.43	143.15	146.79	145.32	121.68	110.85
Water	Average	49.53	64.29	95.67	108.28	137.08	215.14	232.68
	Std. Dev	25.03	61.87	83.72	59.90	88.36	99.88	111.28
Length	Average	6,883	6,062	7,017	7,283	7,238	8,006	8,795
	Std. Dev	1,679	2,001	2,048	2,233	2,316	2,144	1,715
Oil	Average	3.08	4.85	10.76	13.41	11.55	11.15	9.73
	Std. Dev	1.94	7.59	13.72	15.16	9.83	6.72	5.78

and though not shown in the table, only 93 wells have fewer than 20 average producing days. The bottom rows of Table 1 show the distribution of non-operating participants and past experience across wells. In the average well, 3 other firms obtain knowledge about a well at the same time as the well's operator. The average well is fracked by a firm that has previously fracked 80 of its own wells, and can observe the data on 1,089 wells fracked by others.

Table 2 shows the distribution of well characteristics and oil production. The number of wells fracked and the number of active townships and firms all increase over time. More than 65% of all wells are fracked during the last two years, and in 2011, wells are fracked in 85% of townships by 70% of all firms. Over time, firms frack longer wells, using more sand and more water. Firms operating in 2011 use more than three times as much sand and four times as much water per foot of horizontal length, on average, as firms in 2005. However, average oil production does not rise monotonically, reaching its peak in 2008 and then falling thereafter.

Table 3: Average First Year's Oil Production per Foot of Horizontal Length by Quintiles of Sand and Water Use

		Quintiles of Water Use				
		First	Second	Third	Fourth	Fifth
Quintiles of Sand Use	First	8.09	8.27	6.77	8.82	10.16
		(0.33)	(0.44)	(0.73)	(2.20)	(0.81)
	Second	9.53	10.50	9.27	10.50	9.37
		(0.37)	(0.35)	(0.33)	(0.56)	(1.37)
	Third	10.25	11.51	10.91	10.56	10.81
		(0.52)	(0.36)	(0.29)	(0.35)	(0.76)
	Fourth	10.71	10.48	13.24	11.46	11.87
		(0.52)	(0.58)	(0.55)	(0.40)	(0.48)
	Fifth	10.80	12.24	13.37	13.19	13.85
		(1.06)	(0.83)	(0.98)	(0.52)	(0.37)

Net of township fixed effects. Standard errors in parentheses.

Table 3 reports average oil production per foot by quintiles of sand and water use per foot.[24] Across both sand and water use, the highest input levels are associated with higher oil production. For every quintile of water use (columns), the top quintile of sand use has higher production than the bottom quintile. For all but the second quintile of sand use (rows), the top quintile of water use has higher production that the bottom quintile. Thus the data shows that sand and water use affect oil production, though not strictly monotonically.

To verify the importance of spatial heterogeneity in the relationship between fracking inputs and oil production, I estimate a simple Cobb-Douglas production function for fracking, with and without township fixed effects. I regress the log of first years oil production per foot of horizontal length on the well's log sand use and log water use:

$$\log \text{oil per foot}_i = \alpha_0 + \alpha_S \log S_i + \alpha_W \log W_i + \tau_i + \epsilon_i$$

Table 4 reports coefficient estimates for this regression. The first column shows estimates without fixed effects, and the second column shows estimates with fixed effects. Consistent with the results in Table 3, higher sand and water use are associated with higher production. This is true with and without fixed effects. However, the inclusion of township fixed effects decreases the coefficient on sand use and increases the coefficient on water use, suggesting the existence of spatial heterogeneity

[24] To control for the effects of location, I first subtract the average levels of oil production and input use per township from actual production and input use. Then, I add back the overall average levels, creating township fixed effects.

Table 4: Spatial Heterogeneity in the Relationship Between Sand, Water and Oil Production

	(1) Log Oil per foot	(2) Log Oil per foot
α_0	-0.0280	0.319
	(0.104)	(0.0948)
α_S	0.352	0.208
	(0.0211)	(0.0183)
α_W	0.0512	0.137
	(0.0228)	(0.0185)
Township FE		X
N	2,698	2,698
R^2	0.159	0.618

Standard errors in parentheses. OLS estimates of

$$\log \text{oil per foot}_i = \alpha_0 + \alpha_S \log S_i + \alpha_W \log W_i + \tau_i + \epsilon_i$$

in oil production and the possibility that firms make different input choices in different locations.

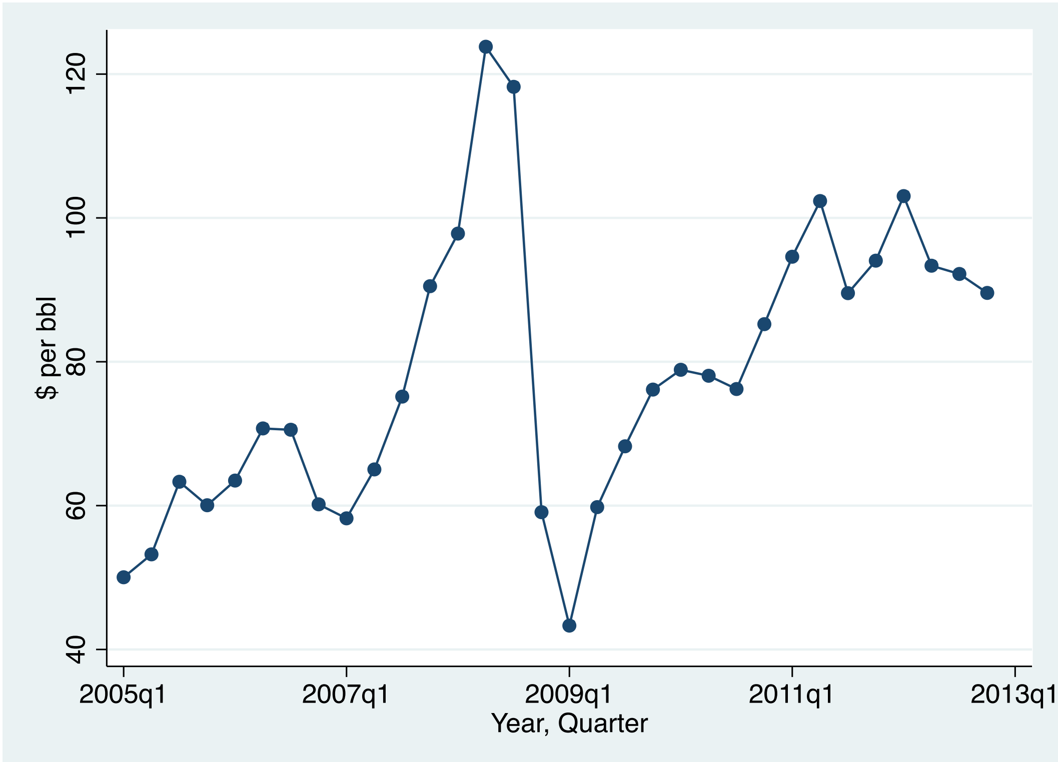
2.4.2 Oil Prices

I collect the daily spot prices for West Texas Intermediate crude oil at the Cushing, Oklahoma oil trading hub from the US Energy Information Administration. The Cushing price is the reference price for oil futures traded on the NYMEX commodity exchange, and the Cushing hub is connected to North Dakota through the Keystone and Enbridge pipeline systems. Figure 3 plots quarterly average oil prices at the Cushing hub. Between 2005-2011, there was a boom and bust in oil prices, with prices climbing from approximately \$60 per bbl in early 2007, reaching more than \$120 per bbl in mid 2008 and falling to \$45 per bbl in early 2009. In 2010-2011, when more than 65% of the wells are fracked, oil prices average \$87 per bbl.

2.5 Drilling and Fracking Costs

Though the NDIC does not require firms to report their costs, the legal process in North Dakota occasionally makes this information public. In particular, when a non-operating mineral rights owner decides not to participate in a well, the operator can ask the NDIC to impose a “risk penalty”, which temporarily prevents the non-participant from earning revenue from its mineral

Figure 3: Quarterly Average Cushing Oil Prices



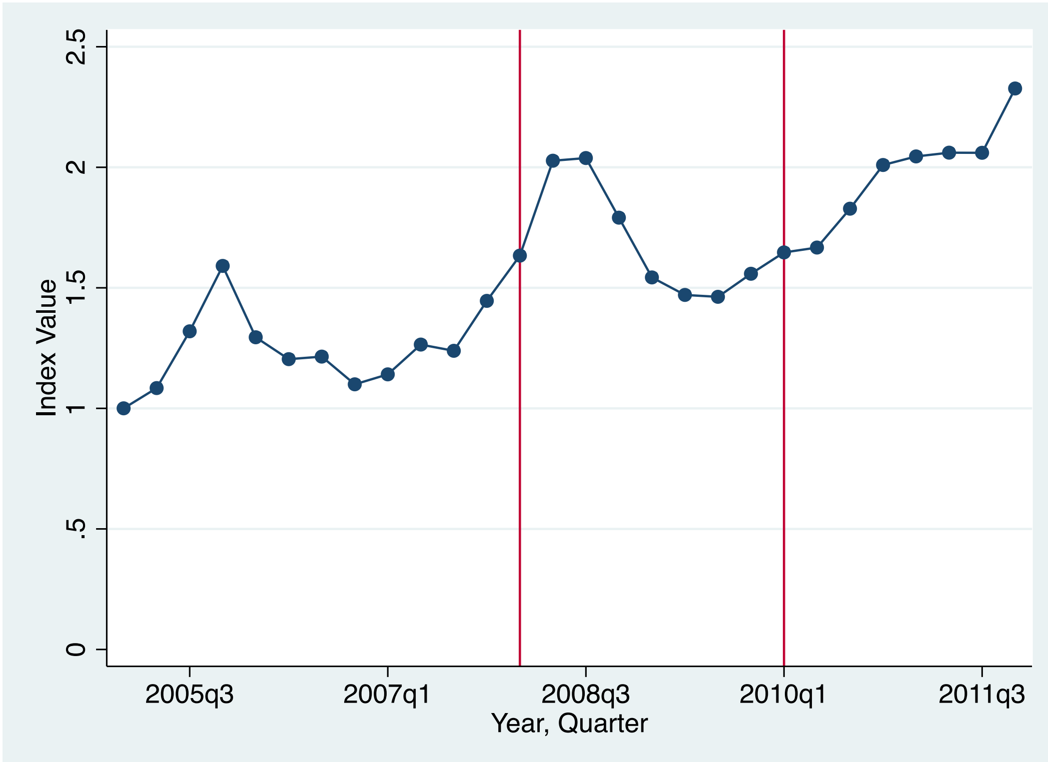
rights.[25] In order to make this request, the operator must legally submit its estimate of the cost of drilling and fracking the well, and this information is publicly recorded by the NDIC. Of the 2,699 wells in this dataset, the cost records for 90 are in the public domain for this reason.

These wells span several years, so to make their costs comparable, I normalize them using a cost index. There is no single publicly available cost index that is both specific to the Bakken and available for all of 2005-2011, so I construct one by combining several other indices. Between the first quarter of 2005 and the fourth quarter of 2007, the index grows at the rate of the BLS Producer Price Index for oil & gas extraction. Between the the first quarter of 2008 and the fourth quarter of 2009, the index grows at the rate of a cost index for vertical wells drilled in North Dakota, published by Spears & Associates, a private consulting firm.[26] Finally, starting in the first

[25] A non-participating mineral rights owner faced with a risk penalty forfeits a significant portion of its share of the well's revenue. In North Dakota, risk penalties are set to 200% of a non-participant's share of capital expenditures. This means that non-participants do not earn any revenue from a well in which they own mineral rights until the well has generated 200% of its capital expenditures in oil production.

[26] Spears & Associates surveys independent engineers in North Dakota quarterly, asking them to estimate the cost of a reference well. The cost estimates are divided into 14 categories, of which 4 are fracking related and 10 are drilling related. The data is separately available for a vertical reference well design, which begins in the first quarter of 2008 and a horizontal reference well design, which begins in the first quarter of 2010. The vertical reference design does not include a fracking treatment. The characteristics of the reference wells stay constant over time, so the changes in estimated costs are due to

Figure 4: Fracking Cost Index



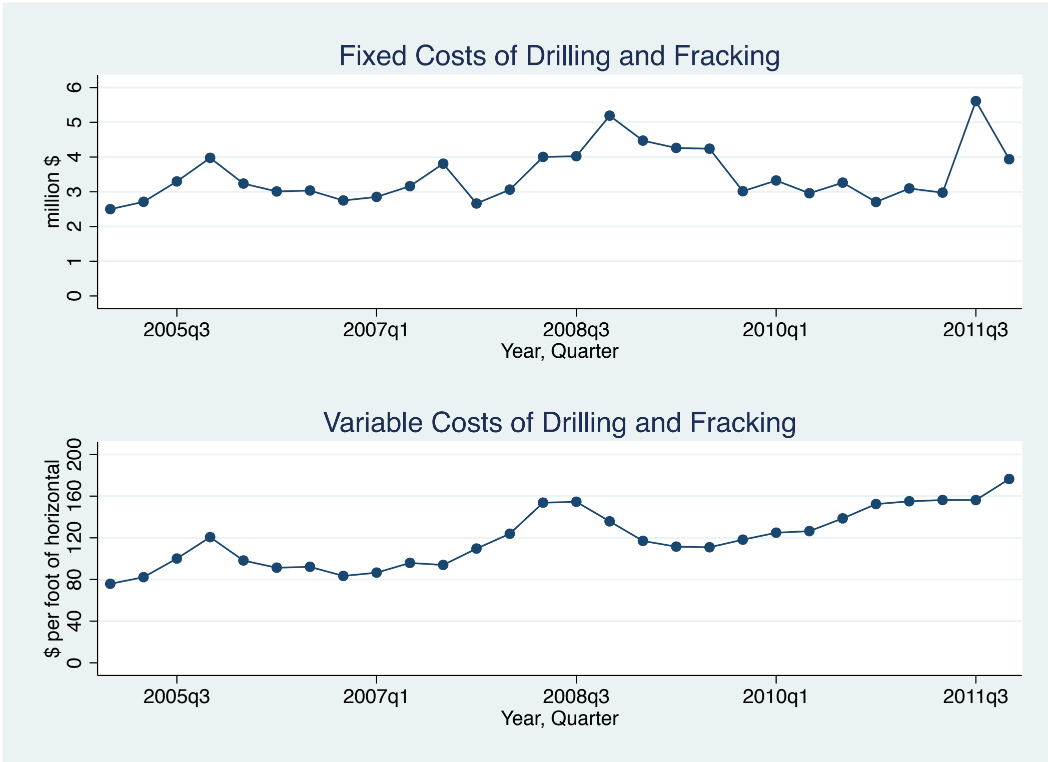
The cost index is computed from the BLS Producer Purchasing Index (PPI) for the Oil & Gas Extraction industry from the first quarter of 2005 to the fourth quarter of 2007. Then, from the first quarter of 2008 to the fourth quarter of 2009, it is calculated from the Spears & Associates data for vertical wells in North Dakota. Finally, from the first quarter of 2010 to the fourth quarter of 2011 it is calculated from the Spears & Associates data for horizontal wells in North Dakota.

quarter of 2010, the index grows at the rate of the Spears & Associates cost index for horizontal wells drilled in North Dakota. I fix the cost index to 1 in the first quarter of 2005 and define "normalized costs" as reported costs divided by the cost index. Figure 4 plots the cost index over time.

To estimate the individual components of costs, I regress normalized costs for these 90 wells onto a constant, lateral length, total sand use, total water use and year-quarter fixed effects. The adjusted R-squared of this regression is 0.54, and the coefficients on lateral length, sand and water are all significantly different from zero at the 5% level. I define the fixed drilling and fracking cost as the sum of the constant and the year-quarter fixed effects, the variable drilling and fracking cost as the coefficient on lateral length, and the sand and water costs as the coefficients on sand and water use. Finally, I generate time-specific costs by multiplying these estimates by the cost index.

changes in prices, not quantities.

Figure 5: Fixed and Variable Costs of Drilling and Fracking



The variable costs of using sand and water in fracking are estimated from a regression of the normalized total drilling and fracking costs for 90 wells with cost data in the public domain onto a constant, lateral length, total sand use, total water use and year-quarter fixed effects. The estimated fixed cost of drilling and fracking is equal to the constant plus the year-quarter fixed effect, divided by the cost index. The estimated variable cost of drilling and fracking is equal to the coefficient on lateral length, divided by the cost index.

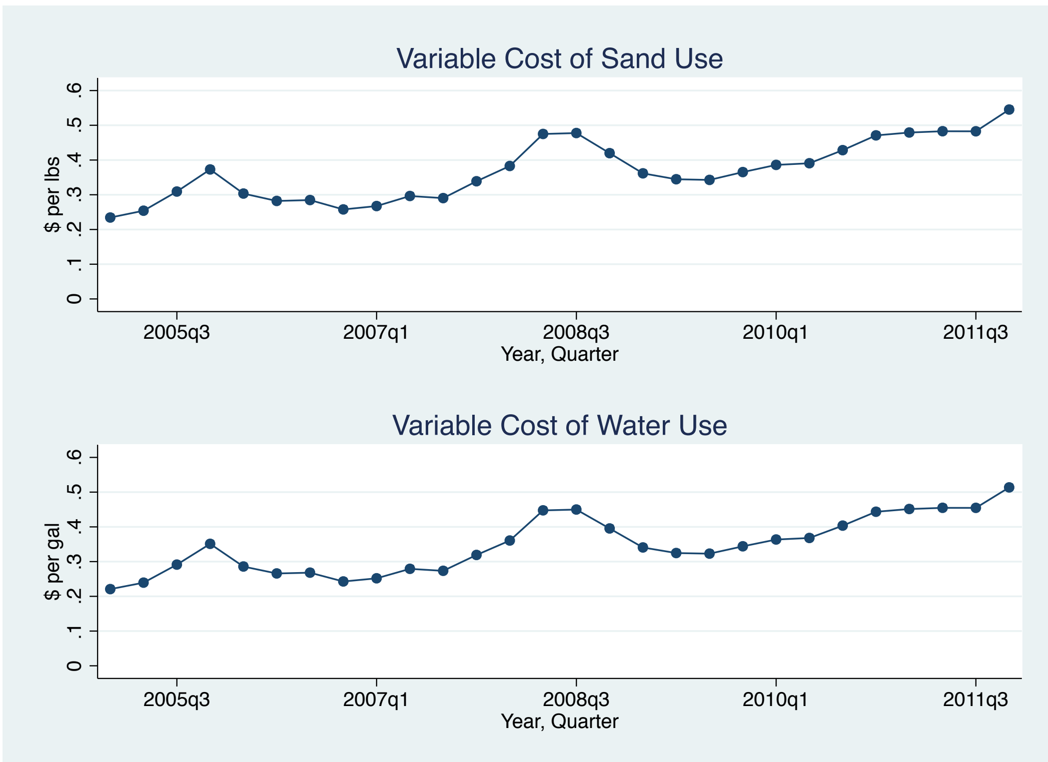
Figures 5 and 6 plot these costs over time.

2.5.1 Information Sets

At time t, firm f can learn about fracking from three sets of wells. First, f can observe all wells that the regulator has made public by time t. This public knowledge includes wells that f operated and wells that other firms operated. Second, f can observe its own wells which are not yet public knowledge, due to well confidentiality. Third, f can observe other firms' wells in which it is a non-operating participant. I can compute the first two sets of information from well completion reports alone. To compute the third set, I must identify the mineral rights owners in each well's spacing unit.

I collect mineral rights lease data from DrillingInfo.com, which digitally records the universe of mineral rights transactions filed in county registries of deeds. These leases are often between a

Figure 6: Variable Costs of Using Sand and Water in Fracking



The variable costs of using sand and water in fracking are estimated from a regression of the normalized total drilling and fracking costs for 90 wells with cost data in the public domain onto a constant, lateral length, total sand use, total water use and year-quarter fixed effects. The estimated cost of pumping 1 pound of sand is equal to the coefficient on sand use, divided by the cost index, while the estimated cost of pumping 1 gallon of water is equal to the coefficient on water use, divided by the cost index.

Table 5: Wells Completed by the 8 Most Active Firms, by Location, Time and Well Characteristics

Firm	North Dakota 2005-2011 Bakken Shale	Outside North Dakota 1995-2004 Vertical	Outside North Dakota 1995-2004 Horizontal	Outside North Dakota 2005-2011 Vertical	Outside North Dakota 2005-2011 Horizontal
Brigham	113	161	0	93	0
Burlington	105	3,826	26	2,792	532
Continental Resources	313	597	3	657	167
EOG	354	4,659	91	6,566	2,914
Hess	165	639	2	219	15
Marathon	223	2,221	4	813	87
Whiting	247	131	0	1,150	11
XTO	101	2,349	53	7,749	2,801
Rest of industry	1,078				

surface owner and an intermediary lease broker operating on behalf of an oil company. Once the broker acquires a lease, it assigns this lease back to its client, a transaction which is not recorded by DrillingInfo.com. To capture the information in the lease assignment process, I also scrape the website of the North Dakota Registry Information Network (www.ndrin.com), which electronically records lease assignments. I combine this lease and lease assignment data into a single dataset identifying the names of any firm that has mineral rights in a spacing unit. I assume that all firms with mineral rights in a well's spacing unit that are not the well's operator are non-operating participants.[27]

2.5.2 Outside Experience

Throughout the paper, I assume that the only knowledge firms have about fracking comes from the wells fracked in North Dakota during 2005-2011. To assess the validity of this assumption, I collect firm-specific drilling history from IHS International for the 8 most active firms in my data, which I report in Table 5. In the first column, I list the number of wells each firm completed in the Bakken during 2005-2011. These 8 firms frack 60% of the wells in the dataset. During the time period I study, these firms are all publicly held, either as independent firms (Brigham, Continental Resources, EOG, Hess, Marathon and Whiting) or as subsidiaries of larger oil companies (Burlington is owned by Conoco Phillips, XTO is owned by Exxon Mobil).

On the right hand side of Table 5, I list the US operating history of these firms outside of

[27] That is, I assume that no mineral rights owners are non-participants. Since only 90 out of 2,699 wells in this time period had risk penalty challenges, this is a reasonable assumption.

North Dakota. In the 10 years prior to the period I study, these firms collectively completed tens of thousands of vertical wells, which are typically drilled into conventional formations, without frack jobs. However, they only completed 179 horizontal wells, suggesting that they had very little experience with the technology necessary to develop wells in the Bakken Shale. Only three firms had previously completed more than ten horizontal wells, and two had done none. During 2005-2011, all eight firms are active outside North Dakota, with four firms completing more than a thousand wells each. Except for EOG and XTO, the vast majority of contemporaneous operational experience outside North Dakota is in vertical wells, though seven of the eight firms do complete horizontal wells. Thus, there is limited scope for these firms to learn about fracking from experience outside of the Bakken.

3 The Fracking Production Function

To quantify what knowledge firms learn about fracking, it is necessary to measure the empirical relationship between oil production, location and fracking input choices. I do this by estimating a production function for fracking. This production function accounts for variation in oil production across a well's life and variation between wells in average production levels.

A well's production changes over time due to age and maintenance-driven downtime. I measure the impact of these factors on oil production using a simple model common in the petroleum engineering literature. Because a well's age is outside the firm's control and because maintenance needs are both similar across wells and scheduled in advance, I argue that the time-varying error in production is plausibly exogenous.

Wells have different average production levels due to differences in their horizontal lengths, locations and fracking inputs. Location and fracking inputs may nonlinearly affect production, so I measure their impact non-parametrically, using Gaussian process regression (GPR), which I describe in detail below. The well-specific error in average production includes the effects of unobserved inputs, such as chemicals, the unobserved amount of oil that can be recovered and its sensitivity to fracking. I argue that chemical choices are independent of sand and water choices for engineering reasons, and that the information which only firms observe about the well's specific geological properties while drilling is unlikely to be correlated with production outcomes.

In the next two sections, I explain this production function model in further detail.

3.1 The Time Series of Oil Production

Per unit of time, wells of all kinds (including non-fracked wells in conventional formations) tend to produce more oil when they are younger and less oil when they are older. This decline in performance over time is not surprising, because the amount of oil that can be recovered is finite and as more of it is pumped out of the ground, the rest becomes more difficult to recover. For nearly 70 years, petroleum engineers have used the simple "Arps" model to illustrate this basic phenomenon (see Fetkovich 1980). The Arps model states that oil production in the t-th month of well i's life is:

$$Y_{it} = Q_i t^{\beta} \exp(\nu_{it})$$

where Q_i is the *baseline* level of production, $\beta < 0$ is a constant governing the production decline of the well and ν_{it} is a mean-zero production shock. In log terms, this is

$$\log Y_{it} = \log Q_i + \beta \log t + \nu_{it}$$

meaning that a 1% increase in a well's age should decrease per period production by $-\beta$%, on average.

The operator of a well chooses D_{it}, the number of days during month t that well i is producing. Unless the well needs maintenance, there is no reason the operator would choose to produce for fewer than the full number of days during a month. All wells experience two routine maintenance events: the installation of external pumping hardware, and the connection of the well to a gas pipeline network. During maintenance, the operator must shut the well down, reducing D_{it}. My data does not indicate whether maintenance occurs in a month, but it does report the number of producing days D_{it}, which I incorporate in the model:

$$\log Y_{it} = \log Q_i + \beta \log t + \delta \log D_{it} + \nu_{it}$$

The time-varying shock to log production, ν_{it}, is the result of unobserved geological variation and deviations from the Arps model. Firms cannot control t, the age of a well, and it is unlikely that firms observe anything correlated with ν before choosing to do maintenance. Even if they did, firms would rather have the well producing on more days than fewer days, independent of ν.

Moreover, firms cannot predict ν when fracking the well, which happens before production starts. For these reasons, I assume that ν is exogenous:

$$\mathbb{E}\left[\nu_{it} \mid t, H_i, D_{it}, S_i, W_i, lat_i, lon_i\right] = 0$$

3.2 The Cross Section of Oil Production

I specify a semi-parametric model for $\log Q$, the log of baseline production:

$$\log Q_i = \alpha + \eta \log H_i + f(S_i, W_i, lat_i, lon_i) + \epsilon_i$$

The parametric part of this model, $\alpha + \eta \log H_i$, is a Cobb-Douglas production function relating the horizontal length of a well to its baseline production. Though it may seem natural that η should equal one, there are practical reasons why this may not be true. Fracking applied to the furthest away points of the horizontal segment of a well may not always perform as well as fracking applied to the closest points. If this decline in effectiveness is nonlinear, wells with longer horizontal segments may not proportionally outperform wells with shorter horizontal segments. The Hicks-neutral productivity α measures the average log baseline production across wells. I discuss the well-specific productivity shock ϵ_i in more detail below.

The function $f(S_i, W_i, lat_i, lon_i) = f(Z_i)$ captures the relationship between baseline production, location and fracking choices. Table 4 in the data section suggests that this relationship differs across locations, and current petroleum engineering suggests that it may be nonlinear. For this reason, I estimate $f(Z_i)$ non-parametrically, using Gaussian process regression, or GPR. GPR makes kernel regression techniques available within a panel data framework. Because there are few examples of GPR in applied economic settings, I provide a basic overview of its application here.

3.2.1 Gaussian process regression

A *Gaussian process* G is a probability distribution over continuous real functions. Gaussian processes are defined by two functions: a mean function $m(Z)$ and a positive definite covariance function $k(Z, Z')$. The mean function is the expectation of the value of a function f drawn at random from G at the point Z. The covariance function is the covariance between $f(Z)$ and $f(Z')$.

In mathematical terms, the mean and covariance functions satisfy:

$$m(Z) = \int f(Z)dG(f)$$
$$k(Z, Z') = \int (f(Z) - m(Z))(f(Z') - m(Z'))dG(f)$$

A Gaussian process is "Gaussian" because the joint distribution of the values $f(Z_1)...f(Z_N)$ is multivariate normal, with a mean vector μ and covariance matrix Σ given by:

$$\mu = (m(Z_1)...m(Z_N))^\top$$
$$\Sigma_{i,j} = k(Z_i, Z_j)$$

This implies that the distribution of $f(Z)$ is also normal with mean $m(Z)$ and variance $k(Z, Z)$. The normality property makes it easy to compute the likelihood that a dataset $(g_i, Z_i)_{i=1}^N$ is generated by the relationship $g = f(Z)$ for a function f drawn from a Gaussian process with mean $m(Z)$ and covariance $k(Z, Z')$. By selecting mean and covariance functions from parametric families, the parameters that best fit the dataset can be estimated using maximum likelihood.

To estimate the function $f(Z_i)$ above, I assume $m(Z) = 0$ due to the presence of the constant term, α, in the parametric portion of the production function. I assume that $k(Z, Z')$ takes the form of a multivariate normal kernel:

$$k(Z_i, Z_j \mid \gamma) = \exp(2\gamma_0) \exp\left(-\frac{1}{2} \sum_{d \in S,W,lat,lon} \frac{(Z_{i,d} - Z_{j,d})^2}{\exp(2\gamma_d)}\right)$$

The first parameter, γ_0, measures the variance of the unknown function $f(Z)$. As points (Z_i, Z_j) become arbitrarily close to each other, the covariance function approaches the variance of f, and its formula collapses to $\exp(2\gamma_0)$. The remaining parameters $\gamma = (\gamma_S, \gamma_W, \gamma_{lat}, \gamma_{lon})$ measure how smooth f is in each dimension.

If the mean function is 0 and the covariance function parameters are γ, then the log likelihood of the data $(g_i, Z_i)_{i=1}^N$ is:

$$\log \mathcal{L}(\gamma) = -\frac{1}{2} g^\top K(\gamma)^{-1} g - \log |K(\gamma)| - \frac{N}{2} \log (2\pi)$$

where $g = (g_1 ... g_N)^\top$ and $K(\gamma)_{i,j} = k(Z_i, Z_j \mid \gamma)$. The process of maximizing this likelihood over γ is called *Gaussian process regression*, or GPR. Conditional on γ and the data $(g, \mathbf{Z})$, the distribution of f evaluated at an out-of-sample point $\widetilde{Z}$ is normal, with mean and variance given by:

$$\mathbb{E}\left[f(\widetilde{Z}) \mid g, \mathbf{Z}, \gamma\right] = k(\widetilde{Z} \mid \gamma)^\top K(\gamma)^{-1} g$$
$$\mathbb{V}\left[f(\widetilde{Z}) \mid g, \mathbf{Z}, \gamma\right] = k(\widetilde{Z} \mid \gamma)^\top K(\gamma)^{-1} k(\widetilde{Z} \mid \gamma)$$

where $k(\widetilde{Z} \mid \gamma) = (k(Z_1, \widetilde{Z} \mid \gamma) ... k(Z_N, \widetilde{Z} \mid \gamma))^\top$. Note that the formula for the mean of $f(\widetilde{Z})$ is similar to the formula for the estimated regression function in kernel regression.[28] However, the additional assumptions about the distribution of possible regression functions in GPR make it possible to select smoothing parameters γ using likelihood techniques, which is not possible in kernel regression. Moreover, since GPR can be defined in terms of a likelihood function, it can easily be incorporated into panel data methods, something which is challenging in standard kernel regression.

Gaussian processes are commonly used in the artificial intelligence and operations research literatures, though their application in economics is so far limited to econometric theory.[29] For a detailed treatment of Gaussian processes, see Rasmussen and Williams (2005).

3.2.2 The Well-Specific Shock ϵ_i

The well-specific shock to log baseline production, ϵ_i, contains unobserved inputs to the fracking process and unobservable variation in geology. Fracking chemicals are the main unobserved input.[30] Firms primarily use chemicals to inhibit bacterial growth in the fracking mixture, to provide lubrication for the pumping units used in fracking and to prevent corrosion and mineral scaling in the well pipe.[31] There is evidence in the petroleum engineering literature that an operator's choice of chemicals does not directly affect the efficiency of its sand and water choices, so I assume that

[28] In kernel regression, the term $k(\widetilde{Z} \mid \gamma)^\top K(\gamma)^{-1}$ in the estimated regression function is replaced with $\frac{k(\widetilde{Z}|\gamma)^\top}{\sum_i k(Z_i, \widetilde{Z}|\gamma)}$. However, the estimates of variance in kernel regression are are not directly comparable to the variance formulas in GPR.

[29] See Kasy (2013) for a recent example.

[30] Another unobserved input is the characteristics of the piping and fracking hardware that firms use to implement frack jobs. This hardware determines the number of fracture initiation points, their distribution across the lateral segment and the level of pressure inside the wellbore.

[31] See http://www.fracfocus.org for further details on the chemicals used in fracking.

sand and water choices are independent of chemical choices.[32]

The petroleum engineering literature predicts that different parts of the Bakken contain different amounts of oil and respond to fracking inputs differently.[33] In particular, wells that are drilled into parts of the Bakken which are thicker, contain more organic material or are more thermally mature have more oil to draw from, and as a result, fracking inputs may be more productive. Similarly, fracking inputs may generate more extensive fracture networks in wells drilled into more permeable parts of the Bakken than wells in less permeable parts. However, aside from the location-specific nature of the production function, I do not have data to control for geological variation in the Bakken.[34] If firms have geological data that may be indicative of how much oil a well contains or how amenable it is to fracking, they may adjust their fracking inputs in response and ϵ_i will not be independent of these choices. Unfortunately, I do not have instruments for fracking input choices, so it is important to consider what addition information firms could have about the wells they are fracking and whether they use it to make fracking decisions.

For the vast majority of wells, firms do not have well-specific information about the thickness, organic content, thermal maturity or permeability of the rock they drill into. To get this information, firms must perform expensive and time-consuming geological tests, the results of which are publicly documented by the NDIC.[35] These tests are only possible if firms elect to drill the vertical portion of the wellbore all the way through the entire Bakken formation, which they rarely do.[36]

Firms do have a potentially useful source of information about well quality in the samples of rock that they collect during drilling, called "cuttings". As the drill bit passes through the upper Bakken shale on its way into the middle Bakken, firms can analyze the returned rock, which may be indicative of the amount of the oil and the level of permeability in the upper Bakken shale at the location where the horizontal segment starts. However, since the goal in horizontal drilling is to stay inside the middle Bakken, firms receive no additional information about the upper Bakken shale and receive no information at all about the lower Bakken shale during the course

[32]See, for example, Jabbari et al. (2012)

[33]See Baihly et al. (2012), Jabbari et al. (2012) and Saputelli et al. (2014)

[34]In the appendix, I analyze the (limited) publicly available data on thickness, organic content and thermal maturity. Broadly speaking, this data is not well-specific (it is spatially interpolated from a small number of wells) and does not explain much variation in production after conditioning on location.

[35]Specifically, firms use gamma ray well logs to determine thickness, rock evaluation pyrolysis of cuttings or well cores to measure organic content and thermal maturity and drill stem tests or MRI/NMR tests to measure permeability.

[36]For example, Sitchler et al. (2013), a recent petroleum engineering study of well performance, fracking inputs, and geology characteristics, has the necessary data for just seven wells.

of drilling. Moreover, the characteristics of the upper Bakken shale can change over the length of the horizontal segment, and there is no guarantee that the lower Bakken shale has the same characteristics at a point as the upper Bakken shale. During the time period I study, laboratory tools to infer rock properties like permeability from cuttings data had not yet been developed.[37] Thus, the information firms can acquire during drilling is unlikely to be helpful in choosing fracking inputs, and in practice may not be used at all.

For these reasons, I argue that ϵ_i is exogenous to firm choices and other well characteristics:

$$\mathbb{E}\left[\epsilon_i \mid t, H_i, D_{it}, S_i, W_i, lat_i, lon_i\right] = 0$$

Combining everything together, the whole production function model is:

$$\log Y_{it} = \alpha + \beta \log t + \delta \log D_{it} + \eta \log H_i + f(Z_i) + \epsilon_i + \nu_{it}$$

Since Gaussian process regression generates a normal likelihood for $f(Z_i)$, I assume that ν_{it} and ϵ_i are both normal, with zero mean and variances σ_ν^2 and σ_ϵ^2, respectively.

3.3 Likelihood

I compute the likelihood function in two steps. In the first step, I treat the unobserved effect of fracking and location $f(Z_i)$ as observed and compute the likelihood of (Y_{it}, X_{it}) conditional on $f(Z_i)$ and the parameters. In the second step, I integrate out the unobserved values of $f(Z_i)$ using the likelihood function for $f(Z_i)$ generated by GPR. I describe the likelihood calculation in detail in the appendix.

3.4 Production Function Estimates

Table 6 shows maximum likelihood estimates of the semi-parametric production function described above in addition to a simpler parametric specification. The parametric specification replaces $f(S_i, W_i, lat_i, lon_i)$ with township fixed effects, τ_i, and a Cobb-Douglas production technology in sand and water, $\kappa_S \log S_i + \kappa_W \log W_i$.

[37] See, for example, Ortega et al. (2012), who note that "Cuttings have not been used in the past quantitatively for optimization of hydraulic fracturing jobs."

All of the parametric model coefficients are statistically significantly different from zero in both specifications and the coefficients common to both have similar estimates. As expected, wells produce less oil per month as they age, with an estimated log decline rate of -0.56.[38] The coefficient on days producing is 1.75, suggesting that when wells undergo maintenance, production per day is lower than when wells do not have maintenance issues. Wells with longer horizontal segments produce more oil than wells with shorter segments, but the effect is not linear. Doubling the horizontal length of a well increases production by 80% in the Cobb-Douglas specification and 85% in the Gaussian process. The variance of ϵ is larger in the Cobb-Douglas specification than in the Gaussian process, suggesting that the flexibility of the Gaussian process explains more of the variation in baseline oil production than Cobb-Douglas and location fixed effects do. The estimated Cobb-Douglas marginal productivities of sand and water are precisely estimated and are smaller than the preliminary estimates in Table 4. Sand and water both increase oil production, with decreasing returns to scale.

The estimated GPR smoothing parameters do not have an intuitive interpretation, so I illustrate the estimated production relationships graphically in Figure 7. The top panel is a contour plot of the non-parametrically estimated function $f(S_i, W_i, lat_i, lon_i)$, evaluated at the geographic centroid of the most active township during this time period. The lines are iso-production curves, which are combinations of sand and water choices with the same estimated value of f. Across all levels of water use, greater sand use is associated with higher oil production, while greater water use is only associated with higher production at the highest level of sand use, and only in a limited range. The middle panel shows contour lines for the Cobb-Douglas specification. The Gaussian process and Cobb-Douglas specifications make starkly different predictions about the impact of fracking inputs and location on oil production. At the average sand and water choices for this township, 266 lbs and 131 gals per foot, respectively, the Gaussian process predicts -3.5 log points of baseline production, while Cobb-Douglas predicts -3.1, meaning that the predictions of the two models differ by 40%. Additionally, the non-parametric specification makes different predictions in different locations. The bottom panel shows contour lines for the production function evaluated at the centroid of a nearby township. The location of the most productive sand and water choices differ across the two townships. In the top panel, the maximal choice is approximately 600 lbs

[38] Current geophysics research on the Bakken has found similar decline rates. Hough and McClurg (2011), for example, estimates the decline rate to be -0.5.

Table 6: Production Function Model Estimates

Coefficient	Cobb-Douglas Estimate	Cobb-Douglas Std. Error	Gaussian Process Estimate	Gaussian Process Std. Error
α			-4.4152	(0.3278)
β	-0.5576	(0.0024)	-0.5570	(0.0024)
δ	1.7543	(0.0035)	1.7549	(0.0035)
η	0.7977	(0.0363)	0.8479	(0.0357)
γ_0			-0.3945	(0.0572)
γ_S			6.1757	(0.1343)
γ_W			5.9467	(0.1232)
γ_{lat}			-2.4702	(0.0539)
γ_{lon}			-2.2376	(0.0609)
κ_S	0.1582	(0.0157)		
κ_W	0.1148	(0.0159)		
$\log \sigma_\epsilon$	-0.9086	(0.0147)	-1.0591	(0.0187)
$\log \sigma_\nu$	-0.4898	(0.0024)	-0.4897	(0.0024)
Township Fixed-effects	X			
Overall R^2	0.783		.811	
Between R^2	0.813		.882	
Within R^2	0.764		.764	
# Wells	2,699			
# Well-months	91,783			

Maximum likelihood estimates of the Cobb-Douglas production function model:

$$\log Y_{it} = \beta \log t + \delta \log D_{it} + \eta \log H_i + \kappa_S \log S_i + \kappa_W \log W_i + \tau_i + \epsilon_i + \nu_{it}$$

and the Gaussian process production function model:

$$\log Y_{it} = \alpha + \beta \log t + \delta \log D_{it} + \eta \log H_i + f(Z_i \mid \gamma) + \epsilon_i + \nu_{it}$$

Y_{it} is oil production for well i when it is t months old, D_{it} is the number of days producing, H_i is the horizontal length, and Z_i is the vector of sand use S_i, water use W_i, latitude lat_i and longitude lon_i. τ_i is a set of township fixed effects. "Between" R^2 is the R^2 for the average predicted log baseline production. "Within" R^2 is the R^2 for the predicted time series of production.

sand and 200 gals water, per foot, while in the bottom panel it is 400 lbs sand and 500 gals water, per foot. This variation across townships in the relationship between oil production and inputs is not possible with the Cobb-Douglas specification, so for the rest of the paper, I focus on the Gaussian process specification.

The fit of both models is high, with R^2's of 78% for the Cobb-Douglas model and 81% for the Gaussian process model. The "between" R^2's, which measure the correlation of predicted baseline production and actual baseline production, are higher, at 81% and 88%, respectively. The production function models fit the data well for several reasons. Both the inputs to fracking, sand and water, and the single output of fracking, crude oil production, are precisely measured. The main unobserved input, fracking chemicals, does not directly affect production or observed input choices, and Gaussian process regression flexibly controls for spatial heterogeneity. Moreover, the production function for fracking is an approximation to a true physical relationship between sand, water, location and oil production. However, since I estimate this approximation non-parametrically, there is the possibility that the estimated smoothing parameters are too narrow, leading to over-fitting.

To check for this, I perform a cross-validation test of the model estimates. For each of 25 test runs, I randomly split the wells into two separate datasets: a training dataset containing 90% of the wells, and a validation dataset containing the remaining 10%. I re-estimate the production function on the training dataset and use the estimates to predict production in the validation dataset. I save the estimated production function coefficients, the R^2 values generated by the training data and the R^2 values generated by the validation data, and report their distribution across test runs in Table 7. The parametric components of the production function model are quite stable across runs, with the average model estimates being similar to the full dataset maximum likelihood estimates. The standard deviations across runs are smaller than the maximum likelihood standard errors for the full dataset. Though the R^2 values for validation samples are lower than for training samples, they are still quite high, with the average overall R^2 for validation samples at approximately 78%, compared to 81% in the training samples. To complement these checks, I provide a series of robustness checks of the stability of the production function across well cohorts in the appendix.

The consistency of the coefficient estimates across cross-validation tests and the high goodness-

Figure 7: Contour Plots of Production Function Estimates

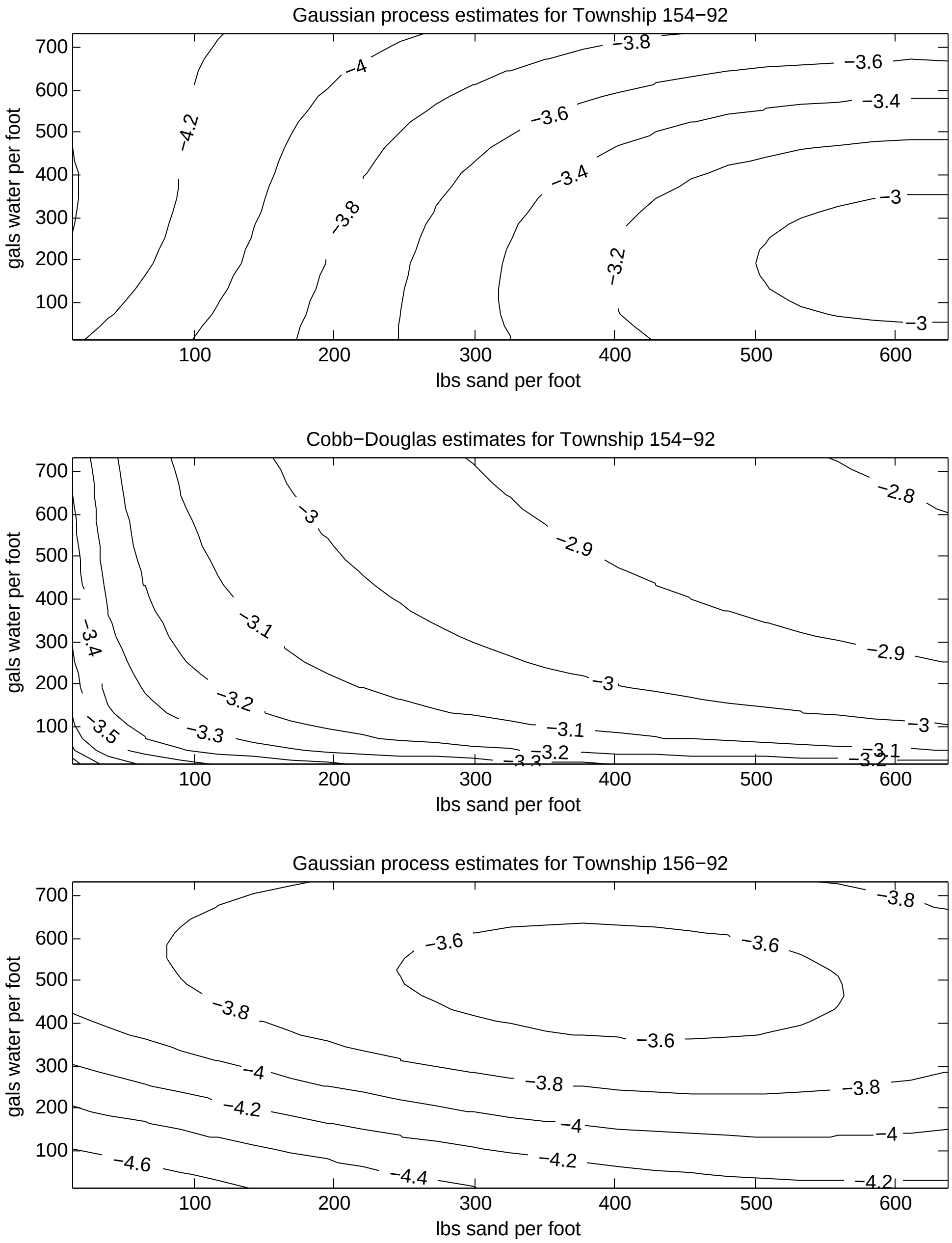


Table 7: Production Function Model Cross-Validation Statistics

Coefficient	Average Estimate	Std. Dev. of Estimate
α	-4.3388	0.1295
β	-0.5570	0.0016
δ	1.7533	0.0055
η	0.8404	0.0136
γ_0	-0.4046	0.0290
γ_S	6.1659	0.0532
γ_W	5.9278	0.0616
γ_{lat}	-2.4454	0.0236
γ_{lon}	-2.2211	0.0392
$\log \sigma_\epsilon$	-1.0545	0.0109
$\log \sigma_\nu$	-0.4916	0.0063
	R^2 comparisons	
R^2 type	Avg. in training	Avg. in validation
Overall R^2	0.8116	0.7835
Between R^2	0.8826	0.8098
Within R^2	0.7652	0.7594
# Wells	2,699	
# Well-months	91,783	
# Cross validation samples	25	

Maximum likelihood estimates of the production function model:

$$\log Y_{it} = \alpha + \beta \log t + \delta \log D_{it} + \eta \log H_i + f(Z_i \mid \gamma) + \epsilon_i + \nu_{it}$$

Y_{it} is oil production for well i when it is t months old, D_{it} is the number of days producing, H_i is the horizontal length, and Z_i is the vector of sand use S_i, water use W_i, latitude lat_i and longitude lon_i.

of-fit measures in validation samples suggest that the maximum likelihood estimates in Table 6 do not suffer from over-fitting and represent a stable and causal relationship between inputs and production.

4 Evidence for Learning

As firms learn to use fracking technology more efficiently, they should make more profitable fracking design choices. If oil prices, input costs and the quality and size of drilling locations were constant over time, I could test this prediction by extrapolating future production from current production and simply check if average expected discounted profits per well increased over time. However, oil prices, input costs and locations do vary over time, so I control for this variation by examining trends in the ratio of actual profits to counterfactual maximal profits. That is, I compute a profitability measure which compares the profits firms earned with the highest amount of profits they could have earned with the best fracking design for each well.

I use the fracking production function to compute these profits. The profits to well i fracked using design j are

$$\Pi_{ij} = \phi P_i \mathbb{E}\left[\sum_{t=1}^{T} \rho^t \widetilde{Y}_{ijt}\right] - c_i(S_j, W_j)$$

where ϕ is the fraction of oil production the firm keeps for itself, P_i is the price the firm will receive for its oil production, T is the number of periods the well is expected to produce for, ρ is the per-period discount rate, $\widetilde{Y}_{ijt}$ is the realization of the level of oil production for well i under fracking design j at age t, and $c_i(S_j, W_j)$ is the total cost of drilling and fracking that design.[39] The main empirical object needed in the calculation of Π_{ij} is the expected present value of discounted

[39] I assume firms believe oil prices follow a martingale process, and thus use a single price, P_i for all future revenues. Additionally, I assume that the fraction of oil revenue that accrues to the firms is 70%, based on typical royalty rates of 16.5%, state taxes of 11.5% and ongoing operating costs of 2%. I set $T = 240$ months, though the NDIC expects Bakken wells to produce for 540 months, making these profit calculations an underestimate. I set $\rho = .9$, which is the standard discount rate use in oil & gas accounting. At this rate, the difference between 540 months and 240 months is only 2.6% in present value terms.

oil production, $\mathbb{E}\left[DOP_{ij}\right]$:

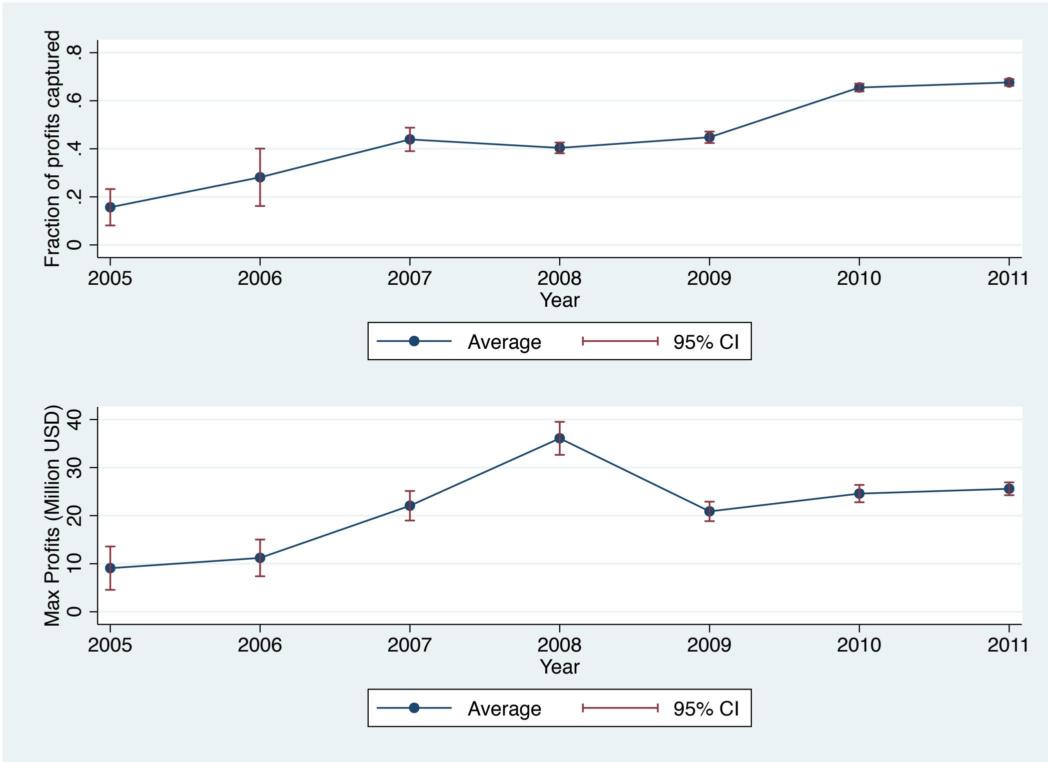
$$\begin{aligned}\mathbb{E}\left[DOP_{ij}\right] &= \mathbb{E}\left[\sum_{t=1}^{T} \rho^t \widetilde{Y}_{ijt}\right] \\ &= \sum_{t=1}^{T} \rho^t \mathbb{E}\left[\widetilde{Y}_{ijt}\right]\end{aligned}$$

I compute this expectation conditional on two different information sets: the full data that I have, and the data each firm had when it made a fracking design decision. The first case represents an *ex post* expectation, and provides a way of asking whether firms made better fracking design decisions over time, given today's knowledge. The second case represents an *ex ante* expectation, and provides a way of asking whether firms' choices were consistent with static profit maximization, given my measures of their information sets.

In both cases, I combine the production function parameter estimates in Table 6 with the normality assumptions on the unobserved terms to compute a probability distribution over oil production. Since the production function estimates depend on the full dataset, this means that I am computing *ex ante* expectations under the assumption that firms had the same beliefs about the production function parameters as I do now. This is a strong assumption. The *ex ante* calculation of expected oil production will be biased if firms had different beliefs than I do about the decline rate β, the productivity of producing days δ and horizontal length η, the bandwidth parameters γ and the variances σ of the unobservable production shocks. I assume that these biases are small, as decline rates and productivity parameters can be predicted using geophysical models[40], and bandwidth and variance parameters do not affect the asymptotic properties the production function estimate.[41] Moreover, the impact of fracking design and location $f(Z)$ is computed nonparametrically from both the bandwidth parameters γ and the information set. Thus firms with different information sets will have different beliefs about $f(Z)$, and these beliefs will differ from the *ex post* beliefs as well.

I present the full calculation of expected discounted oil production in the appendix.

Figure 8: Fraction of Positive Profits Captured and Maximal Profits by Year, ex post



4.1 *ex post* Comparisons

Over time, firms choose fracking designs with higher *ex post* expected profits. The top half of Figure 8 plots the *ex post* ratio of actual profits to maximal profits per well.[42] The average fraction of profits captured increases nearly monotonically over time, from 15.7% in 2005 to 67.6% in 2011. Much of this growth happens in two phases. Between 2005 and 2007, the fraction increases from 15.7% to 43.9%, and between 2009 and 2010, the fraction increases from 44.8% to 65.5%. By 2011, firms earn an average of 67.6% of the maximum profits they could have earned with optimal fracking input choices.

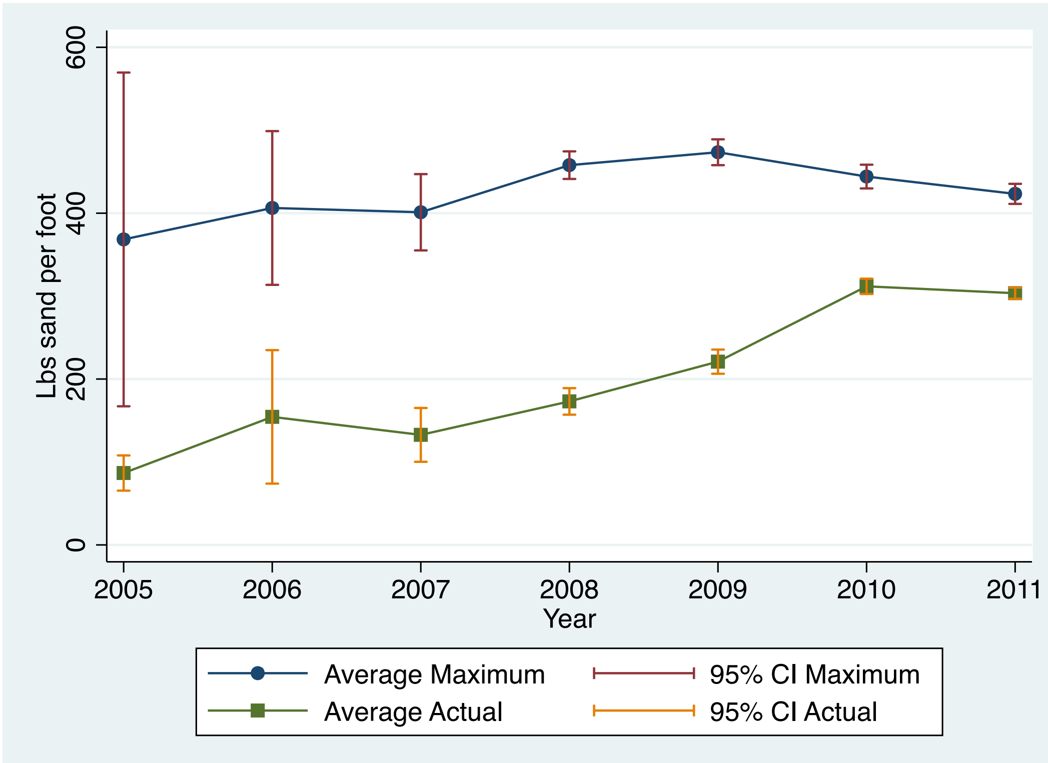
The bottom half of Figure 8 shows how these maximal profits evolve over time. When oil prices were at their peak in 2008, the profit maximizing input choice for the average well would have generated \$36.1 million in profits, meaning that in 2008, foregone profits from inefficient fracking choices averaged \$21.3 million per well. By 2011, lower oil prices reduced these maximal profits

[40]See Fetkovich (1980).

[41]See section 7.1 in Rasmussen and Williams (2005).

[42]I only include wells in this calculation that have both positive actual profits and positive maximal profits. Over the entire sample, 5.2% of wells have either negative actual profits or negative maximal profits.

Figure 9: Average Profit Maximizing Sand Use and Actual Sand Use Per Well, ex post

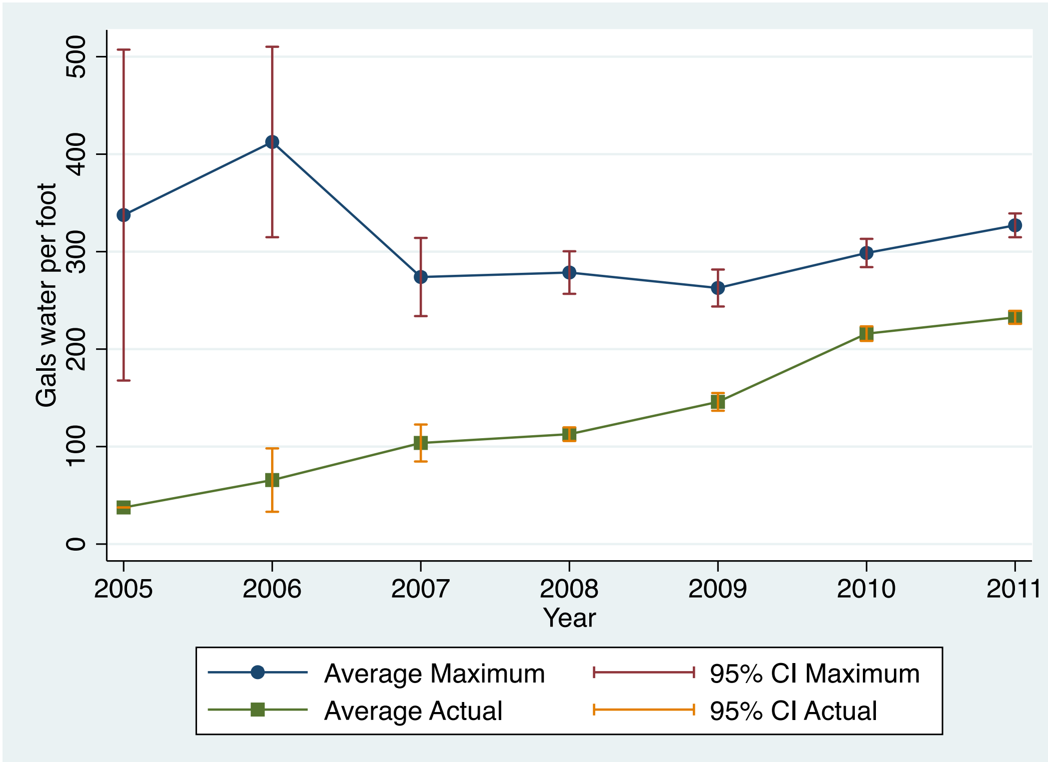


to \$25.6 million per well. Combined with the higher fraction of profits captured, firms in 2011 left only \$9.9 million on the table.

Firms captured more profits by selecting more profitable fracking designs over time. In Figures 9 and 10, I plot average profit maximizing and actual input use per well over time. Though firms use less sand in fracking than the estimated profit maximizing levels, starting in 2009, actual choices approach optimal choices. In 2005 and 2006, the average well was fracked with approximately 275 lbs sand per foot less than the profit maximizing level. This difference in sand use doesn't meaningfull fall until reaching 132 lbs per foot in 2010. By 2011, the difference between optimal sand use and actual sand use is only 120 lbs per foot.

Though the differences in actual and optimal water use start out considerably larger than the differences in sand use, actual water choices get closer to optimal water choices in almost every year. In 2005, firms fracked the average well with 300 gals per foot less water than the water use in the optimal well. By 2011, the difference is only 98 gals per foot. These trends in actual input use towards optimal input use are consistent with the idea that firms are learning about the efficient use of fracking inputs as they observe more data, and with this knowledge they make

Figure 10: Average Profit Maximizing Water Use and Actual Water Use Per Well, ex post



more profitable choices.

4.2 Profitability vs. Productivity

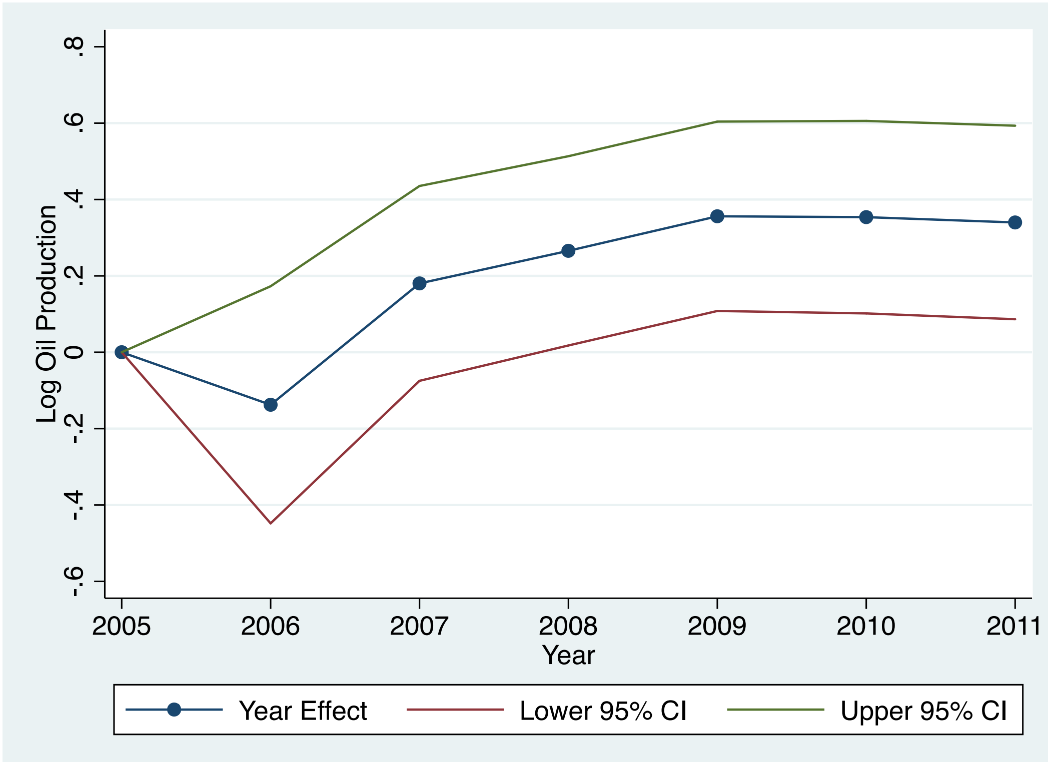
The existing literature on learning in firms focuses on *productivity* instead of *profitability*. Scholars in this literature measure learning by comparing estimates of the time-varying component of Hicks-neutral productivity with the amount of experience a firm has in producing.[43] This approach to studying learning does not treat the production function as an object for firms to learn. Rather, the knowledge from accumulated experience serves as an *input* to the firm's production function, in the same way that labor, capital and materials do.

To determine if firms in this dataset became more productive, in addition to more profitable, I add year fixed effects to the Gaussian process production function specification, and plot their estimated values and confidence intervals in Figure 11.

Wells fracked in 2005 are actually 13.7% *more* productive than wells fracked in 2006. However,

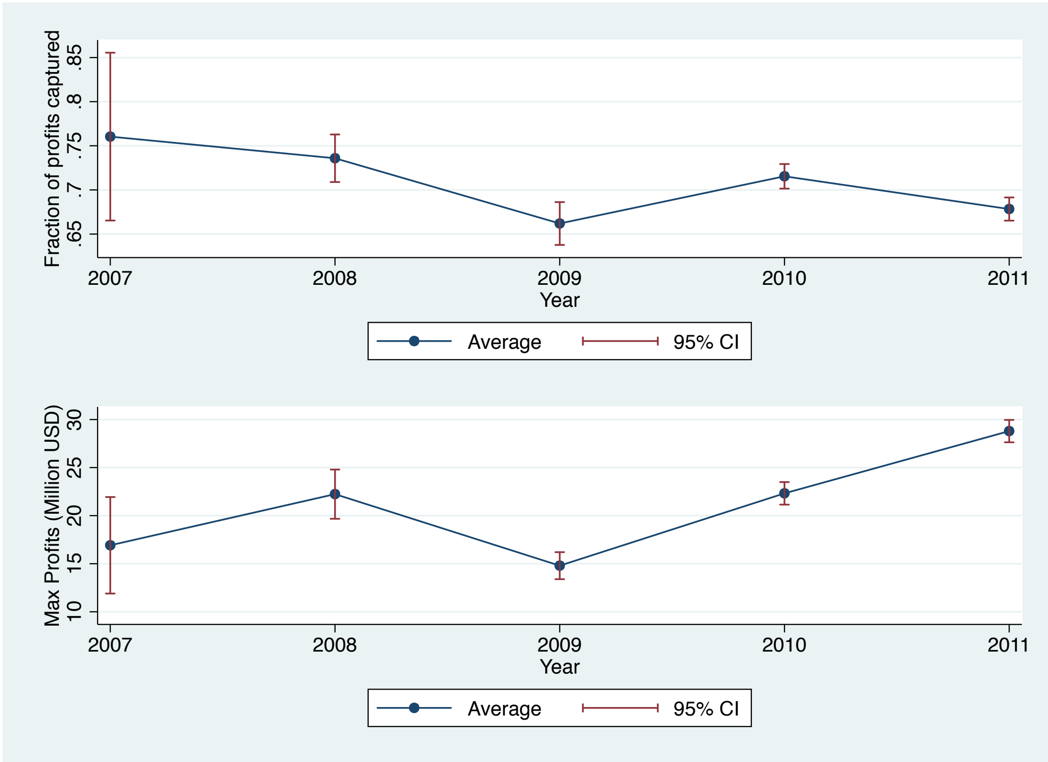
[43] For example, Benkard (2000) correlates log labor requirements per unit of production with measures of experience (and forgetting), and Thornton and Thompson (2001) estimate a semi-parametric production function model in which various measures of experience are direct inputs to production.

Figure 11: Gaussian Process Year Effects



the confidence interval around this estimate is wide enough to include zero, as there are only 10 wells in 2005 and 20 wells in 2006. Wells fracked in later years are more productive than wells fracked in 2005 or 2006. For example, wells fracked in 2009 are 35.6% more productive than those in 2005, and 49.3% more productive than those in 2006. Again, the confidence intervals around these estimates are wide, and I cannot reject the hypothesis that there is no change in productivity between 2006 and 2009. In each of the next 2 years, productivity falls slightly, though the differences are not statistically significant. Overall, wells fracked between 2008-2011 cohorts are more productive than the earliest wells, but there is no productivity growth during 2008-2011. Since this time period covers 95% of the wells studied in this paper, I interpret this as evidence that firms learned to be more productive only in the earliest years. In contrast, the results in the previous section show that firms learned to be more profitable in all years, and especially during 2008-2011.

Figure 12: Fraction of Positive Profits Captured and Maximal Profits by Year, ex ante



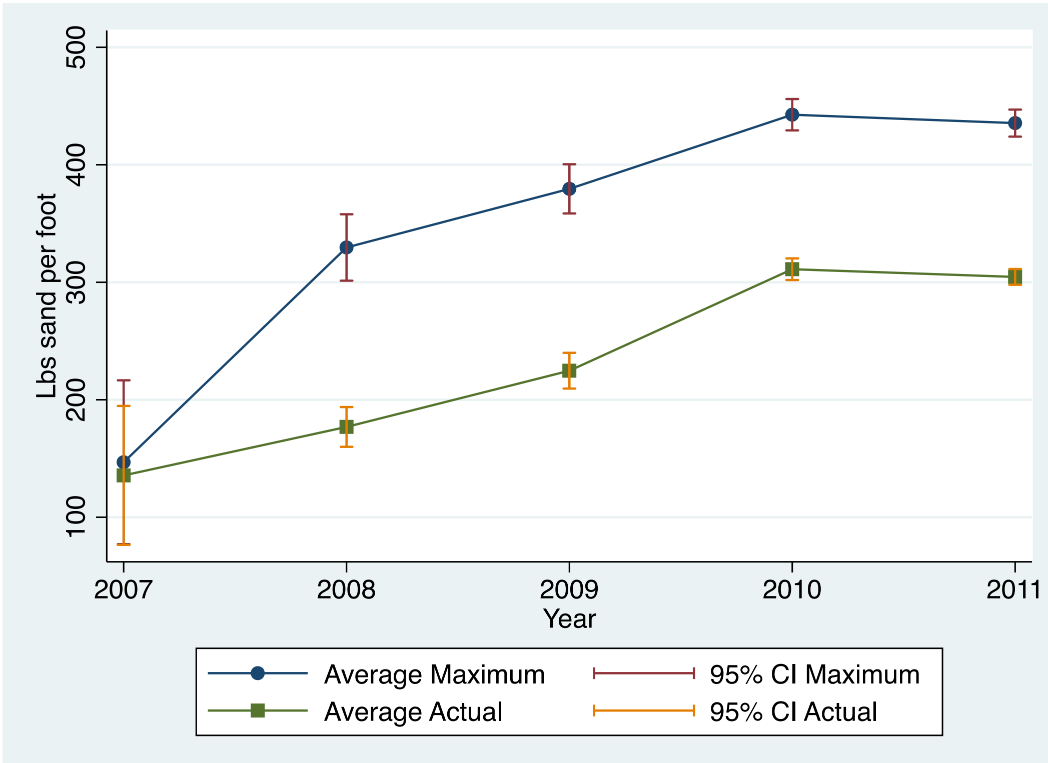
4.3 *ex ante* Comparisons

Though firms make choices which approach the *ex post* estimates of optimal choices over time, those choices do not always maximize the *ex ante* estimates of expected profits. The top half of Figure 12 plots the ratio of actual profits to maximal profits per well using *ex ante* expectations.[44] Firms initially make fracking input choices with expected profits that are close to the optimal choices, capturing 76.0% of potential *ex ante* profits in 2007. However, profit capture actually falls over time, reaching 67.8% in 2011, approximately the same level as the *ex post* case in 2011.

While the fraction of profits captured falls, *ex ante* expectations of maximal profits rise from 2009-2011, as show in the bottom half of Figure 8. Unlike the *ex post* case, where the highest level of maximal profits coincides with the 2008 peak in oil prices, *ex ante* maximal profits are highest in 2011, reaching $28.8 million per well. Though average oil prices are similar in 2008 ($100 per

[44] As in the *ex post* case, I only include wells in this calculation that have both positive actual profits and positive maximal profits. Over the entire sample, 6.1% of wells have either negative actual profits or negative maximal profits. Half of these wells are fracked in 2009. Moreover, I further limit the set of wells by computing expected profits for the subset of wells that are fracked by firms which can observe 50 wells and 300 well-months of production history. The first wells that satisfy this criteria are not fracked until 2007.

Figure 13: Average Profit Maximizing Sand Use and Actual Sand Use Per Well, ex ante

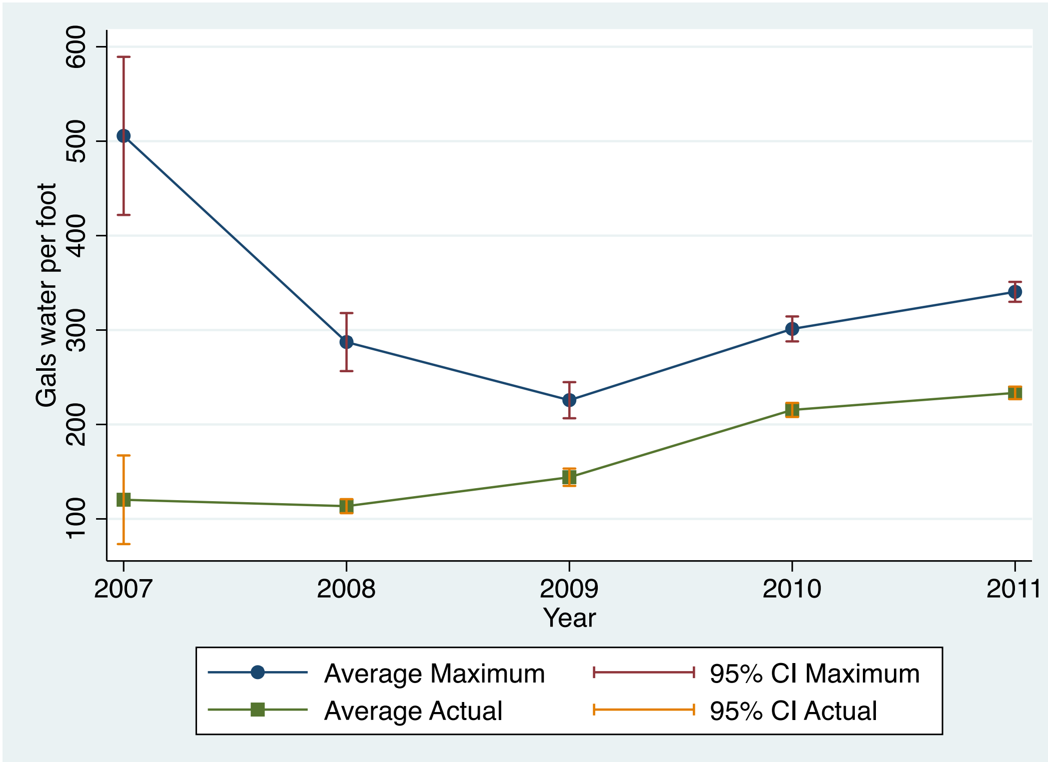


bbl) and 2011 ($95 per bbl), firms have much more information about fracking in 2011 and this information generates more optimistic expectations. The combined effect of falling *ex ante* profit capture and rising maximal profits increases foregone *ex ante* profits from $3.1 million in 2007 to $10.6 million in 2011.

Firms capture a shrinking fraction of *ex ante* profits over time because their actual sand use grows more slowly than the expected profit maximizing sand use does. Figure 13 plots average profit maximizing and actual sand use per well over time. In 2007, actual sand use is quite similar to *ex ante* optimal sand use. However, as the data firms have to learn from accumulates, optimal sand use increases faster than actual sand use, and by 2011, the difference between optimal and actual sand use reaches 131 lbs per foot. Though this difference is similar to the difference in the *ex post* case during 2011, it is striking that the differences in actual and optimal sand use increase over time in the *ex ante* case while decreasing in the *ex post* case.

Figure 14 plots average *ex ante* optimal and actual water use per well is similar to the *ex post* case in Figure 10: on average, firms use less than the *ex ante* optimal amount of water in fracking, but make improved water choices over time. In 2007, firms use 385 gals per foot less water than

Figure 14: Average Profit Maximizing Water Use and Actual Water Use Per Well, ex ante



the optimal level. This difference shrinks in each year, and by 2011, it is only 107 gals per foot.

5 Fracking input choice model

Though firms do learn over time, many of their choices do not coincide with the predicted optimal choices, even on an *ex ante* basis. I consider two possible explanations for this phenomenon based firm preferences. First, firms may care about the uncertainty in their estimates of the profits of a fracking design. Second, in estimating the profits of a fracking design, firms may weigh their own data differently than the data generated by their competitors.

5.1 Preferences Over Uncertainty

In comparing the expected profits a firm earned to the maximal expected profits a firm could have earned, I have implicitly assumed that the *correct* strategy is for firms to select fracking designs solely on the basis of expected profits, without regard to the uncertainty of profits across designs. There are two potential problems with this assumption. First, viewing fracking design as an investment project selection problem, there may be financial or organizational factors that cause firms to have preference over uncertainty. Second, when learning about the performance of different fracking designs, firms may care about uncertainty through the *explore vs. exploit* tradeoff that exists in all learning problems.

Though it is appropriate for firms to ignore uncertainty in simple and frictionless models of investment project selection, there are practical reasons why uncertainty may also matter. Firms raise outside capital to finance operations and the presence of debt capital can lead firms to select fracking designs with higher uncertainty, as bond holders will bear the downside risk. On the other hand, capital constrained firms may not necessarily have the option of selecting fracking designs with higher uncertainty if they are more expensive to implement. Financial considerations can thus push firms towards or away from fracking designs with more uncertain profits. Firms must also hire and incentivize potentially risk averse engineers, who select fracking designs. Depending on the extent of their career concerns and the structure of their compensation, engineers themselves may have preferences over uncertainty.

The prescribed learning strategies in most theoretical models of learning involve uncertainty seeking behavior. Analyses of the *explore vs. exploit* tradeoff in learning predict that agents should always do some amount of exploration, by selecting actions with more uncertain payoffs. This tradeoff will frequently require agents to sacrifice expected payoffs in the present in order to acquire uncertainty resolution in the future. Since actions with the more uncertain payoffs can resolve more future uncertainty, experimenting agents should have a positive taste for uncertainty.

Most theory models predict that agents will experiment, at least initially. In most of the settings studied by Aghion et al. (1991), a fully rational, expected present discounted value maximizing agent will do some amount of exploring forever and a similar result obtains in the multi-agent context studied by Bolton and Harris (1999). The implied preferences for uncertainty in both of these models arise out of the natural dynamics of learning problems. Agents are still risk neutral over their payoffs, but because there is present value to better information in the future, they prefer those actions with uncertain payoffs which can produce more future information.

Empirically, oil companies exhibit both risk seeking and risk averse behavior. The process of acquiring mineral rights for new drilling prospects and establishing the existence of oil within those prospects is an especially risky one (see, for example Walls and Dyer 1996 and Reiss 1989).

However, oil companies are price takers in the world market for oil, and many use financial markets to hedge some or all of their future oil production, suggesting that firms may wish to avoid risks associated with future price fluctuations (see Haushalter 2000).

Whether the companies I study here prefer fracking input choices with more or less uncertain production is an empirical question. I estimate firm preferences over expectations and variance of fracking designs by analyzing realized choices. To do this, I fit a multinomial logit preference model of fracking design choice in which the "utility" a firm has for fracking design j applied to well i is:

$$\begin{aligned} u_{ij} &= \xi_m \left(\phi P_i \mathbb{E}\left[DOP_{ij}\right] - c_i(S_j, W_j)\right) + \xi_s \phi P_i \left(\mathbb{V}\left[DOP_{ij}\right]\right)^{\frac{1}{2}} + \epsilon_{ij} \\ &= \widetilde{u}_{ij}(\xi_m, \xi_s) + \epsilon_{ij} \end{aligned}$$

where ϕ is the fraction of oil revenues firms keep, P_i is the price of oil for well i, $c_i(S_j, W_j)$ is the cost of fracking design j for well i, and ϵ_{ij} is an iid logit error. The parameters (ξ_m, ξ_s) represent the firm's preference over expected present discounted revenues and the standard deviation of present discounted revenues, conditional on the data they have. Under this preference specification, the probability that a firm selects design j for well i is given by the standard logit formula:

$$p_{ij} = \frac{\exp(\widetilde{u}_{ij})}{\sum_k \exp(\widetilde{u}_{ik})}$$

The mean utilities in this preference model are linear in the expectation and standard deviation of profits to a fracking design. Preferences of this type have precedence in the theoretical learning literature. Brezzi and Lai (2002) show that a linear combination of the expectation and standard deviation of the payoff to a choice can represent a simple and efficient approximation to the Gittins index value for the choice, if the choices have independently distributed payoffs. Since Gittins and Jones (1979) show that ordinal preferences over Gittins indices result in dynamically efficient learning behavior, agents that utilize these linear approximations attain near-optimal learning. Though the profits to fracking input choices are not distributed independently, authors in the computer science and operations research literatures have found that these learning strategies also perform well in the general case. In those literatures, learning strategies which select the choice with the highest value of a linear combination of the expectation and standard deviation of payoffs are called "upper confidence bound", or UCB strategies. Rusmevichientong and Tsitsiklis (2010) and Srinivas et al. (2012) have established that UCB strategies quickly identify the highest performing choice, and do so in a way which minimizes an agent's *ex post* cumulative regret over its past choices. UCB strategies are also reported to be in use at major technology companies, like Yahoo, Microsoft and Google (see Chapelle and Li 2011, Graepel et al. 2010 and Scott 2010). In all of the existing literature which utilizes UCB learning strategies, the weight on the standard deviation of the payoffs to a choice is positive, hence the "upper" in upper confidence bound strategies. This paper is not the first in economics to utilize UCB learning strategies in an empirical context. Dickstein (2013) estimates the parameters of a UCB learning strategy in a study of learning behavior by physicians.

With data on the choices firms made, expectation and standard deviation calculations made using their information sets, and oil price and fracking cost data, I estimate the parameters (ξ_m, ξ_s) using maximum likelihood. I estimate separate values of (ξ_m, ξ_s) for each of the 8 most active firms, and also estimate a pooled value of (ξ_m, ξ_s) for the industry as a whole. Table 8 reports these coefficient estimates, standard errors, and several measures of goodness-of-fit. All firms and the pooled industry have positive "taste" for the expectation of profits of a fracking design and negative "taste" for the standard deviation. That is, every firm appears to avoid fracking input choices with high uncertainty. I can reject risk-neutrality for all firms and for the pooled industry. In dollar terms, firms make choices as if they are willing to accept a reduction in expected profits of \$0.60 to \$0.98 for a reduction of \$1 in the standard deviation of profits.

I report three goodness-of-fit statistics. The likelihood based pseudo-R^2, which I refer to as $LLPR$, is defined as 1 minus the ratio of the optimized log-likelihood over the log-likelihood evaluated at the null hypothesis:

$$LLPR = 1 - \frac{\log \mathcal{L}(\widehat{\xi}_m, \widehat{\xi}_s)}{\log \mathcal{L}(0, 0)}$$

This statistic is similar to a real R^2 in that it varies between 0 and 1, with 0 indicating that the model does not fit any better than no model and 1 indicating that the model fits the data perfectly (see Train 2009). This measure of fit indicates how far from "perfect" the fit actually is, but it does not have a "fraction of variance explained" interpretation the way a true R^2 does. I also compute the correlation between the expected input use implied by the model's estimated choice

Table 8: Uncertainty Preference Model Estimates

Firm	$\widehat{\xi_m}$	$se(\widehat{\xi_m})$	$\widehat{\xi_s}$	$se(\widehat{\xi_s})$	# Wells	$LLPR$	ρ_S	ρ_W
Brigham	11.05	1.05	-11.30	1.16	111	0.24	0.00	0.15
Burlington	12.02	1.25	-15.39	1.57	102	0.34	0.55	0.47
Continental	13.54	0.83	-17.26	1.05	313	0.33	0.53	0.50
EOG	5.88	0.39	-7.75	0.57	339	0.17	-0.18	0.33
Hess	10.69	0.96	-13.10	1.10	143	0.30	0.60	0.45
Marathon	15.52	1.24	-21.99	1.67	209	0.44	0.61	0.30
Whiting	10.25	0.74	-16.97	1.21	247	0.36	-0.02	0.05
XTO	11.56	1.22	-14.36	1.45	101	0.32	0.50	0.51
All	7.46	0.17	-10.39	0.23	2,605	0.23	0.50	0.40

Maximum likelihood estimates of the uncertainty preference model:

$$u_{ij} = \xi_m \left(\phi P_i \mathbb{E}\left[DOP_{ij}\right] - c_i(S_j, W_j)\right) + \xi_s \phi P_i \left(\mathbb{V}\left[DOP_{ij}\right]\right)^{\frac{1}{2}} + \epsilon_{ij}$$

P_i is the price of oil for well i, $\mathbb{E}\left[DOP_{ij}\right]$ is the expectation of the present discounted value of oil production for well i when it is fracked using design j, $\mathbb{V}\left[DOP_{ij}\right]$ is the variance of the present discounted value of oil production for i under design j, $c_i(S_j, W_j)$ is the cost of implementing design j on well i, and ϵ_{ij} is an iid logit shock. $LLPR$ is a likelihood-based pseudo-R^2:

$$LLPR = 1 - \frac{\log \mathcal{L}(\widehat{\xi}_m, \widehat{\xi}_s)}{\log \mathcal{L}(0,0)}$$

where $\mathcal{L}(\widehat{\xi}_m, \widehat{\xi}_s)$ is the likelihood of the model evalauted at the MLE and $\mathcal{L}(0,0)$ is the likelihood of the model evaluated at the null hypothesis. ρ_S and ρ_W are the correlations of actual sand and water use decisions with their predicted values from the model.

probabilities and actual input use, for both sand and water. If expected input use is similar to what is observed in the data, these correlations should be positive and (ideally) close to 1.

The fit of this model varies a fair amount across firms, but is generally modest. The pseudo-R^2 measures are less than 50% for all firms and for the pooled industry, suggesting that the best fitting values of the model's parameters still require a lot of support from the logit errors to rationalize firm behavior. For 6 of the 8 firms, the correlation between predicted sand use and realized sand use is positive, and for 5 it is at least 50%. The correlations between predicted and realized water use are smaller, with only 2 firms having correlations at or above 50%, but no firms have negative correlations. Though the coefficient estimates are all significantly different from zero, the low fit statistics suggest that preferences that are linear in the mean and standard deviation of profits only explain a small portion of observed behavior.

I also estimate a version of this model which includes an interaction term between expected profits and the standard deviation of profits. While learning rules which are nonlinear in the mean and standard deviation do not appear in the existing learning literature, it is possible that true firm preferences over risk and reward are more complicated than a linear model can capture. By including an interaction between expected profits and the standard deviation of profits, I allow for risk preferences that may vary with the mean. Table 9 reports estimates of these models. The results are qualitatively the same as Table 8, with all firms showing risk aversion and all but one firm showing increasingly negative taste for risk as reward increases. Goodness-of-fit measures are slightly better for these models than for the standard mean/variance models, though this is to be expected from the inclusion of an additional covariate.

Overall, Tables 8 and 9 provide evidence that firms tend to select fracking designs with higher expected profits and avoid fracking designs with higher standard deviation of profit. This behavior is not consistent with the notion that firms are actively exploring uncertain fracking designs, but it is consistent with passively learning firms that are constrained by organizational or financially motivated variance aversion.

5.2 Own-data bias

A different explanation for firms' apparent unwillingness to select the fracking design with the largest expected profits is that I am computing expectations with respect to different beliefs than

Table 9: Uncertainty Preference Model Estimates, With Interaction

Firm	$\widehat{\xi_m}$	$se(\widehat{\xi_m})$	$\widehat{\xi_s}$	$se(\widehat{\xi_s})$	$\widehat{\xi_I}$	$se(\widehat{\xi_I})$	# Wells	$LLPR$	ρ_S	ρ_W
Brigham	20.12	1.95	-8.37	1.27	-2.86	0.44	111	0.31	-0.03	0.16
Burlington	14.08	1.66	-14.81	1.58	-0.70	0.36	102	0.35	0.57	0.51
Continental	21.26	1.29	-17.25	1.12	-2.25	0.27	313	0.37	0.60	0.53
EOG	6.10	0.41	-7.49	0.58	-0.07	0.03	339	0.17	-0.17	0.35
Hess	9.80	1.16	-13.37	1.12	0.38	0.29	143	0.30	0.60	0.45
Marathon	20.55	1.72	-22.36	1.75	-1.44	0.33	209	0.45	0.67	0.25
Whiting	11.58	0.86	-17.03	1.24	-0.24	0.07	247	0.37	0.03	0.05
XTO	13.87	1.70	-14.59	1.48	-0.50	0.24	101	0.32	0.48	0.53
All	9.06	0.21	-10.59	0.23	-0.29	0.02	2,605	0.24	0.53	0.44

Maximum likelihood estimates of the uncertainty preference model:

$$u_{ij} = \xi_m \left(\phi P_i \mathbb{E}\left[DOP_{ij}\right] - c_i(S_j, W_j)\right) + \xi_s \phi P_i \left(\mathbb{V}\left[DOP_{ij}\right]\right)^{\frac{1}{2}} + \xi_I \left(\phi P_i \mathbb{E}\left[DOP_{ij}\right] - c_i(S_j, W_j)\right) \left(\phi P_i \left(\mathbb{V}\left[DOP_{ij}\right]\right)^{\frac{1}{2}}\right) + \epsilon_{ij}$$

P_i is the price of oil for well i, $\mathbb{E}\left[DOP_{ij}\right]$ is the expectation of the present discounted value of oil production for well i when it is fracked using design j, $\mathbb{V}\left[DOP_{ij}\right]$ is the variance of the present discounted value of oil production for i under design j, $c_i(S_j, W_j)$ is the cost of implementing design j on well i, and ϵ_{ij} is an iid logit shock. $LLPR$ is a likelihood-based pseudo-R^2:

$$LLPR = 1 - \frac{\log \mathcal{L}(\widehat{\xi_m}, \widehat{\xi_s}, \widehat{\xi_I})}{\log \mathcal{L}(0,0,0)}$$

where $\mathcal{L}(\widehat{\xi_m}, \widehat{\xi_s}, \widehat{\xi_I})$ is the likelihood of the model evalauted at the MLE and $\mathcal{L}(0,0,0)$ is the likelihood of the model evaluated at the null hypothesis. ρ_S and ρ_W are the correlations of actual sand and water use decisions with the predicted values from the model.

those held by firms. There are many ways that a firm's beliefs may be different than the ones I calculate: firms may have biased prior beliefs about the role of fracking design and location, they may have simpler beliefs about the functional form relating fracking design and location to production, or my fracking cost and oil price data could be different from the costs and prices firms experience. However, using the data that I have, I am only able to test a simpler explanation. I assume that firms do have the belief structure I have described here, but do not necessarily treat all of the data available to them equally. In particular, firms may weigh data from their own experiences differently than data from the experiences of other firms that they observe through the public disclosure process. I refer to this explanation as "own-data bias".

To test for this phenomenon, I introduce a new parameter, $\lambda \in (0, 1)$, which represents the firm's relative weighting scheme. If $\lambda = 0$, the firm places no weight on the data generated by other firms and if $\lambda = 1$, the firm places no weight on its own data, relying entirely on outside data to learn. At $\lambda = \frac{1}{2}$, the firm puts equal weight on its own data and the data generated by others, which gives the preference model described in the previous section. For each value of λ, I can compute the expectation and standard deviation of *weighted* discounted profits for well i with fracking design j, for which I provide a calculation in the appendix. I then use these weighted profits in the same multinomial logit choice model described in the previous section, and refer to the choice model with weighted estimates as the weighted preference model.

In Table 10, I report maximum likelihood estimates of λ, as well as the other preference model coefficients, for the same specification in Table 8. The estimated value of λ is less than $\frac{1}{2}$ for all individual firms, and for 5 firms, the 95% confidence intervals do not include $\frac{1}{2}$. The pooled estimate is also less than $\frac{1}{2}$ and its 95% confidence interval does not include $\frac{1}{2}$. Comparing Tables 8 and 10, the preference model coefficients do change slightly, but allowing for weighted beliefs does not affect the previous conclusion that all firms dislike uncertainty in the profits of a fracking input choice. Firms are willing to trade \$0.61 to \$0.83 in expected profits for a reduction of \$1 in the standard deviation of profits, which is a similar range to the model estimated in Table 8. The fit of the model in Table 10 is somewhat better than the model in Table 8, but it is still modest.

Table 10: Weighted Uncertainty Preference Model Estimates

Firm	$\widehat{\xi_m}$	$se(\widehat{\xi_m})$	$\widehat{\xi_s}$	$se(\widehat{\xi_s})$	$\widehat{\lambda}$	$se(\widehat{\lambda})$	# Wells	$LLPR$	ρ_S	ρ_W
Brigham	14.37	1.33	-17.27	1.71	0.12	0.04	111	0.30	0.04	0.16
Burlington	13.95	1.46	-17.98	1.84	0.41	0.05	102	0.36	0.55	0.50
Continental	18.26	1.10	-22.38	1.36	0.36	0.02	313	0.37	0.63	0.55
EOG	7.01	0.45	-10.65	0.75	0.15	0.04	313	0.20	-0.12	0.36
Hess	10.87	1.00	-14.13	1.19	0.46	0.05	143	0.31	0.62	0.46
Marathon	21.77	1.72	-27.79	2.12	0.34	0.03	209	0.47	0.68	0.24
Whiting	9.82	0.70	-16.00	1.14	0.00		247	0.36	0.04	0.06
XTO	12.98	1.52	-15.84	1.63	0.44	0.04	101	0.32	0.48	0.53
All	8.22	0.19	-11.74	0.26	0.38	0.01	2,605	0.24	0.54	0.45

Maximum likelihood estimates of the uncertainty preference model:

$$u_{ij} = \xi_m \left(\phi P_i \mathbb{E}\left[DOP_{ij} \mid \lambda\right] - c_i(S_j, W_j)\right) + \xi_s \phi P_i \left(\mathbb{V}\left[DOP_{ij} \mid \lambda\right]\right)^{\frac{1}{2}} + \epsilon_{ij}$$

P_i is the price of oil for well i, $\mathbb{E}\left[DOP_{ij} \mid \lambda\right]$ is the expectation of the present discounted value of oil production for well i when it is fracked using design j, $\mathbb{V}\left[DOP_{ij} \mid \lambda\right]$ is the variance of the present discounted value of oil production for i under design j, λ is the weighting parameter, $c_i(S_j, W_j)$ is the cost of implementing design j on well i, and ϵ_{ij} is an iid logit shock. $LLPR$ is a likelihood-based pseudo-R^2:

$$LLPR = 1 - \frac{\log \mathcal{L}(\widehat{\xi_m}, \widehat{\xi_s}, \widehat{\lambda})}{\log \mathcal{L}(0, 0, \frac{1}{2})}$$

where $\mathcal{L}(\widehat{\xi_m}, \widehat{\xi_s}, \widehat{\lambda})$ is the likelihood of the model evalauted at the MLE and $\mathcal{L}(0, 0, \frac{1}{2})$ is the likelihood of the model evaluated at the null hypothesis. ρ_S and ρ_W are the correlations of actual sand and water use decisions with the predicted values from the model. Because Whiting's estimate of λ is at the boundary, standard errors are computed with respect to ξ_m and ξ_s only.

6 Conclusion

This paper provides one of the first empirical analyses of learning behavior in firms using operational choices, realized profits, and information sets. Oil companies in the North Dakota Bakken Shale learned to more efficiently use fracking technology between 2005-2011, increasing their capture of possible profits from 15.7% to 67.6% by making improved fracking design choices over time. Contrary to the predictions of most theoretical models of learning, I do not find evidence that firms actively experiment in order to learn. Instead, firms prefer fracking input choices with lower variance, and are willing to give up \$0.60-0.98 in expected profits for a reduction of \$1 in the standard deviation of profits. Finally, firms in my data appear to overweight data from their own operations relative to the data they observe from their competitors.

From a neoclassical economics perspective, it is surprising that these firms do not experiment, even though it is valuable to do so. They operate in an industry known for its appetite for risk and use of advanced technology and have access to a wealth of data to learn from. However, they leave money on the table. Across the 2,699 wells in this data, the average well appears to forego \$12.1 million in profits on an *ex post* basis and \$7.6 million on an *ex ante* basis, resulting in \$20-33 billion in lost profits.

These results complement recent work by petroleum engineers on their own failures to learn to use to new technologies in a variety of contexts. Authors in this literature note that explicit learning efforts like experiments do happen, but less frequently and later in the development of a formation than they should.[45] Much of this research cites two hurdles to learning: a tendency by operators to prematurely focus their optimization efforts on cost reductions instead of improvements in operational choices, and the absence of incentive contracts between operators and their service contractors. The first phenomenon suggests that operators *believe* they know the production function with high certainty, but later discover their beliefs were wrong. In future work, I plan to incorporate this possiblity into my model of input choice under uncertainty. The second phenomenon raises important questions about the effects of contractual incompleteness on the oil and gas exploration industry that I hope to study in future work.

[45] For a detailed overview of this literature, see Vincent (2012)

References

Aghion, Philippe, Patrick Bolton, Christopher Harris, and Bruno Jullien, "Optimal Learning by Experimentation," *The Review of Economic Studies*, 1991, *58* (4), pp. 621–654.

Anand, Bharat N and Tarun Khanna, "Do firms learn to create value? The case of alliances," *Strategic management journal*, 2000, *21* (3), 295–315.

Arrow, Kenneth J., "The Economic Implications of Learning by Doing," *The Review of Economic Studies*, 1962, *29* (3), pp. 155–173.

Baihly, Jason, Raphael Altman, and Isaac Aviles, "Has the Economic Stage Count Been Reached in the Bakken Shale?," in "SPE Hydrocarbon Economics and Evaluation Symposium" 2012.

Benkard, C.L., "Learning and Forgetting: The Dynamics of Aircraft Production," *The American Economic Review*, 2000.

Bolton, Patrick and Christopher Harris, "Strategic Experimentation," *Econometrica*, 1999, *67* (2), pp. 349–374.

Brezzi, Monica and Tze Leung Lai, "Optimal learning and experimentation in bandit problems," *Journal of Economic Dynamics and Control*, 2002, *27* (1), 87–108.

Chapelle, Olivier and Lihong Li, "An empirical evaluation of thompson sampling," in "Advances in Neural Information Processing Systems" 2011, pp. 2249–2257.

Conley, T.G. and C.R. Udry, "Learning about a new technology: Pineapple in Ghana," *The American Economic Review*, 2010, *100* (1), 35–69.

Corts, K.S. and J. Singh, "The effect of repeated interaction on contract choice: Evidence from offshore drilling," *Journal of Law, Economics, and Organization*, 2004, *20* (1), 230–260.

Deutsch, John, "Secretary of Energy Advisory Board Shale Gas Production Subcommittee Second Ninety Day Report," Technical Report November 2011.

Dickstein, M.J., "Efficient provision of experience goods: Evidence from antidepressant choice," Working Paper 2013.

Fetkovich, MJ, "Decline curve analysis using type curves," *Journal of Petroleum Technology*, 1980, *32* (6), 1065–1077.

Foster, A.D. and M.R. Rosenzweig, "Learning by doing and learning from others: Human capital and technical change in agriculture," *Journal of Political Economy*, 1995, pp. 1176–1209.

Fox, J.E. and T.S. Ahlbrandt, *Petroleum geology and total petroleum systems of the Widyan Basin and Interior Platform of Saudi Arabia and Iraq*, US Department of the Interior, US Geological Survey, 2002.

Gaswirth, Stephanie B., "Assessment of Undiscovered Oil Resources in the Bakken and Three Forks Formations, Williston Basin Province, Montana, North Dakota, and South Dakota, 2013," Technical Report, U.S. Geological Survey April 2013.

Gilje, E., "Does Local Access To Finance Matter?: Evidence from US Oil and Natural Gas Shale Booms," Working Paper November 2012.

Gittins, John C and David M Jones, "A dynamic allocation index for the discounted multi-armed bandit problem," *Biometrika*, 1979, *66* (3), 561–565.

Graepel, Thore, Joaquin Q Candela, Thomas Borchert, and Ralf Herbrich, "Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine," in "Proceedings of the 27th International Conference on Machine Learning (ICML-10)" 2010, pp. 13–20.

Griliches, Zvi, "Hybrid corn: An exploration in the economics of technological change," *Econometrica, Journal of the Econometric Society*, 1957, pp. 501–522.

Haushalter, G. David, "Financing Policy, Basis Risk, and Corporate Hedging: Evidence from Oil and Gas Producers," *The Journal of Finance*, 2000, *55* (1), 107–152.

Hicks, Bruce E., "North Dakota Oil & Gas Update," in "Presented for Dunn County Oil Day in Killdeer, ND on February 21, 2012" North Dakota Industrial Commission 2012.

Hough, E and T McClurg, "Impact of Geological Variation and Completion Type in the U.S. Bakken Oil Shale Play Using Decline Curve Analysis and Transient Flow Character," in "AAPG International Conference and Exhibition, Milan, Italy, October 23-26, 2011" 2011.

Jabbari, Hadi, Zhengwen Zeng et al., "Hydraulic Fracturing Design For Horizontal Wells In the Bakken Formation," in "46th US Rock Mechanics/Geomechanics Symposium" American Rock Mechanics Association 2012.

Kasy, Maximilian, "Why experimenters should not randomize, and what they should do instead," *Working Paper*, 2013.

Kellogg, Ryan, "Learning by Drilling: Interfirm Learning and Relationship Persistence in the Texas Oilpatch," *The Quarterly Journal of Economics*, 2011, *126* (4), 1961–2004.

Kogan, Leonid, Dimitris Papanikolaou, Amit Seru, and Noah Stoffman, "Technological Innovation, Resource Allocation, and Growth," Working Paper 17769, National Bureau of Economic Research January 2012.

Larsen, Lamont C, "Horizontal Drafting: Why Your Form JOA May Not Be Adequate for Your Companys Horizontal Drilling Program," *Rocky Mtn. Min. L. Found. J.*, 2011, *48*, 51.

Levitt, Clinton J., "Learning through Oil and Gas Exploration," *Working Paper*, November 2011.

Levitt, Steven D., John A. List, and Chad Syverson, "Toward an Understanding of Learning by Doing: Evidence from an Automobile Assembly Plant," Working Paper 18017, National Bureau of Economic Research April 2012.

McCarthy, Kevin, Katherine Rojas, Martin Niemann, Daniel Palmowski, Kenneth Peters, and Artur Stankiewicz, "Basic petroleum geochemistry for source rock evaluation," *Oilfield Review*, 2011, *23* (2), 32–43.

Muehlenbachs, Lucija, Elisheba Spiller, and Christopher Timmins, "Shale Gas Development and Property Values: Differences across Drinking Water Sources," Working Paper 18390, National Bureau of Economic Research September 2012.

Nordhaus, Alex Trembath Michael Shellenberger Ted and Jesse Jenkins, "Where the Shale Gas Revolution Came From," Technical Report, The Breakthrough Institute May 2012.

Ortega, Camilo Ernesto, Roberto Aguilera et al., "Use of Drill Cuttings for Improved Design of Hydraulic Fracturing Jobs in Horizontal Wells," in "SPE Americas Unconventional Resources Conference" Society of Petroleum Engineers 2012.

Pimmel, A and G Claypool, *Introduction to shipboard organic geochemistry on the JOIDES Resolution*, http://www-odp.tamu. edu, 2001.

Rasmussen, Carl Edward and Christopher KI Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*, The MIT Press, 2005.

Reiss, Peter C., "Economic and Financial Determinants of Oil and Gas Exploration Activity," Working Paper 3077, National Bureau of Economic Research August 1989.

Romer, Paul M., "Increasing Returns and Long-Run Growth," *Journal of Political Economy*, 1986, *94* (5), pp. 1002–1037.

Rusmevichientong, Paat and John N Tsitsiklis, "Linearly parameterized bandits," *Mathematics of Operations Research*, 2010, *35* (2), 395–411.

Ryan, Bryce and Neal C Gross, "The diffusion of hybrid seed corn in two Iowa communities," *Rural sociology*, 1943, *8* (1), 15–24.

Saputelli, Luigi Alfonso, Mohamed Y Soliman, Carlos Manuel Lopez, Alejandro Javier Chacon et al., "Design Optimization of Horizontal Wells With Multiple Hydraulic Fractures in the Bakken Shale," in "SPE/EAGE European Unconventional Resources Conference and Exhibition" Society of Petroleum Engineers 2014.

Scott, Steven L, "A modern Bayesian look at the multi-armed bandit," *Applied Stochastic Models in Business and Industry*, 2010, *26* (6), 639–658.

Shelley, Robert, Nijat Guliyev, and Amir Nejad, "A Novel Method to Optimize Horizontal Bakken Completions in a Factory Mode Development Program," in "SPE Annual Technical Conference and Exhibition" 2012.

Sitchler, Jason Cale, Bilu Verghis Cherian, Maraden Luigie Panjaitan, Christopher Michael Nichols, Jayanth Krishna Krishnamurthy et al., "Asset Development

Drivers in the Bakken and Three Forks," in "SPE Hydraulic Fracturing Technology Conference" Society of Petroleum Engineers 2013.

Srinivas, Niranjan, Andreas Krause, Sham M Kakade, and Matthias Seeger, "Information-theoretic regret bounds for gaussian process optimization in the bandit setting," *Information Theory, IEEE Transactions on*, 2012, *58* (5), 3250–3265.

Stoyanov, Andrey and Nikolay Zubanov, "Productivity spillovers across firms through worker mobility," *American Economic Journal: Applied Economics*, 2012, *4* (2), 168–198.

Thornton, R.A. and P. Thompson, "Learning from experience and learning from others: An exploration of learning and spillovers in wartime shipbuilding," *American Economic Review*, 2001, pp. 1350–1368.

Train, Kenneth, *Discrete choice methods with simulation*, Cambridge university press, 2009.

Vidic, RD, SL Brantley, JM Vandenbossche, D Yoxtheimer, and JD Abad, "Impact of Shale Gas Development on Regional Water Quality," *Science*, 2013, *340* (6134).

Vincent, MC, "The Next Opportunity to Improve Hydraulic-Fracture Stimulation," *Journal of Petroleum Technology*, 2012, *64* (3), 118–127.

Walls, Michael R. and James S. Dyer, "Risk Propensity and Firm Performance: A Study of the Petroleum Exploration Industry," *Management Science*, 1996, *42* (7), 1004–1021.

Wieland, Volker, "Learning by doing and the value of optimal experimentation," *Journal of Economic Dynamics and Control*, 2000, *24* (4), 501–534.

A Likelihood Calculation

A.1 Step 1

Let $\theta = (\alpha, \beta, \delta, \eta)$ represent the vector of the non-fracking parameters and let $\phi = (\sigma_\epsilon, \sigma_\nu)$ represent the vector of the variance parameters. I compute the pseudo-observation g_i from (Y_{it}, X_{it}), conditional on θ as

$$
\begin{aligned}
g_i &= \frac{1}{N_i}\sum_{t=1}^{N_i}(\log Y_{it} - X_{it}\theta) \\
&= \frac{1}{N_i}\sum_{t=1}^{N_i}(g(Z_i) + \epsilon_i + \nu_{it}) \\
&= f(Z_i) + \epsilon_i + \frac{1}{N_i}\sum_{t=1}^{N_i}\nu_{it}
\end{aligned}
$$

g_i is the sum of the "true" effect of fracking and location on oil production and a normally distributed error with zero mean and variance $\sigma_\epsilon^2 + \frac{1}{N_i}\sigma_\nu^2$.

A.2 Step 2

Conditional on the pseudo-observations g_i, the likelihood of (Y_{it}, X_{it}) follows the standard formula for panel data with a random effect on each well. Let $\psi(\cdot \mid \mu, \sigma)$ denote the normal likelihood with mean μ and standard deviation σ and let $e_{it} = \log Y_{it} - X_{it}\theta$. Finally, let bolded capital letters represent vectors of the time series of a variable. The log-likelihood of observing $(\mathbf{Y}_i, \mathbf{X}_i)$ conditional on the parameters (θ, ϕ) and the unobserved impact of fracking g_i is

$$
\begin{aligned}
\log \mathcal{L}(\mathbf{Y}_i, \mathbf{X}_i \mid g_i, \theta, \phi) &= \log\left[\int \psi(\epsilon_i \mid 0, \sigma_\epsilon)\prod_{t=1}^{T_i}\psi(e_{it} - g_i - \epsilon_i \mid 0, \sigma_\nu)d\epsilon_i\right] \\
&= -\frac{1}{2}\left[\frac{1}{\sigma_\nu^2}\left(\sum_{t=1}^{T_i}(e_{it} - g_i)^2 - \frac{\sigma_\epsilon^2}{T_i\sigma_\epsilon^2 + \sigma_\nu^2}\left(\sum_{t=1}^{T_i} e_{it} - g_i\right)^2\right)\right] \\
&\quad - \frac{1}{2}\left[\log\left(T_i\frac{\sigma_\epsilon^2}{\sigma_\nu^2} + 1\right) + T_i \log\left(2\pi\sigma_\nu^2\right)\right], \text{ which simplfies to} \\
&= -\frac{1}{2}\left[\log T_i + \frac{\sum_t e_{it}^2 - \frac{1}{T_i}\left(\sum_t e_{it}\right)^2}{\sigma_\nu^2} + (T_i - 1)\left(2\log\sigma_\nu + \log(2\pi)\right)\right] \\
&\quad + \log\psi\left(g_i \mid \frac{1}{T_i}\sum_{t=1}^{T_i} e_{it}, \sigma_\epsilon^2 + \frac{1}{T_i}\sigma_\nu^2\right) \\
&= \log J(\mathbf{Y}_i, \mathbf{X}_i, T_i \mid \theta, \phi) + \log\psi\left(g_i \mid \frac{1}{T_i}\sum_{t=1}^{T_i} e_{it}, \sigma_\epsilon^2 + \frac{1}{T_i}\sigma_\nu^2\right)
\end{aligned}
$$

The first term does not depend on g_i and the second term is simply a normal log-likelihood, evaluated at g_i, the effect of fracking and location for well i. Though g_i is unobserved, by the properties of GPR, the vector $\mathbf{g}$ of g_i's for all N wells is distributed multivariate normal with mean zero and variance $K(\mathbf{Z} \mid \gamma)$. Thus, I can integrate over the values of g_i to obtain the likelihood in terms of observable data and parameters. Let $\mathbf{T}$ denote the vector of values of T_i, $\Sigma(\mathbf{T}, \phi)$ be an N by N matrix with $\sigma_\epsilon^2 + \frac{1}{T_i}\sigma_\nu^2$ in the i-th diagonal position and zeros elsewhere and let $\mu(\mathbf{Y}, \mathbf{X}, \mathbf{T}, \theta)$ be a vector with $\frac{1}{T_i}\sum_{t=1}^{T_i} e_{it}$ in the i-th position. Then the full log-likelihood is:

$$
\begin{aligned}
\log \mathcal{L}(\mathbf{Y}, \mathbf{X}, \mathbf{Z}) &= \log \int \psi(\mathbf{g} \mid \mathbf{0}, K(\mathbf{Z} \mid \gamma)) \prod_{i=1}^{N} \mathcal{L}(\mathbf{Y}_i, \mathbf{X}_i \mid g_i, \theta, \phi) dgi \\
&= \sum_{i=1}^{N} \log J(\mathbf{Y}_i, \mathbf{X}_i, T_i \mid \theta, \phi) + \log \int \psi(\mathbf{g} \mid \mathbf{0}, K(\mathbf{Z} \mid \gamma)) \psi(\mathbf{g} \mid \mu(\mathbf{Y}, \mathbf{X}, \mathbf{T}, \theta), \Sigma(\mathbf{T}, \phi)) d\mathbf{g} \\
&= \sum_{i=1}^{N} \log J(\mathbf{Y}_i, \mathbf{X}_i, T_i \mid \theta, \phi) + \log \psi\left(\mu(\mathbf{Y}, \mathbf{X}, \mathbf{T}, \theta) \mid 0, \Sigma(\mathbf{T}, \phi) + K(\mathbf{Z} \mid \gamma)\right)
\end{aligned}
$$

where the last line comes as a result of equations $A.7$ and $A.8$ from Rasmussen and Williams (2005). Having integrated out the unobserved values g_i, the full log-likelihood is completely in terms of the observed data $(\mathbf{Y}, \mathbf{X}, \mathbf{T})$, the parameter vectors θ and ϕ, and the covariance matrix $K(\mathbf{Z} \mid \gamma)$ of the nonparametric effect of fracking and location on oil production.

B Expected Present Discounted Value of Oil Production

I compute *ex post* expectations for all wells, and I compute *ex ante* expectations for wells fracked by firms with sufficiently large information sets. I require that a firm's information set has at least 50 wells and at least 300 well-months of production. This limits the set of wells I can analyze, and the earliest wells with information sets this large do not appear until the fourth quarter of 2007.

I compute $\mathbb{E}\left[DOP_{ij}\right]$ using both expectation operators for a 10 by 10 grid of possible frack designs j, with sand use between 0 and 650 lbs per foot and water use between 0 and 750 gals per foot. These grid points cover 95% of observed sand choices and 99% of observed water choices. By the normality assumptions in the production function model, the joint distribution of log-production for well i under fracking design j over T months of existence (call this $\log \widetilde{Y}_{ij}$) is multivariate normal, with mean μ_{ij} and covariance Σ_{ij} given by:

$$\begin{aligned}
\mu_{ij} &= \widetilde{\mathbf{X}}_i\theta + \widetilde{g}(Z_{ij}) \\
\Sigma_{ij} &= \widetilde{\mathbf{X}}_i\Sigma^\theta\widetilde{\mathbf{X}}_i^\top + \left(\sigma_\epsilon^2 + \sigma_{g,ij}^2\right)\mathbf{1}_T + \sigma_\nu^2\mathbf{I}_T
\end{aligned}$$

where $\widetilde{\mathbf{X}}_i$ is a matrix of well i's static characteristics and a vector of log-age values from 1 month to T months, $\widetilde{g}(\cdot)$ is the estimated GPR, Z_{ij} is the a vector of design (S_j, W_j) and latitude and longitude for well i, Σ^θ is the covariance matrix for the estimates of θ, $\sigma_{g,ij}^2$ is the estimated variance of the GPR at Z_{ij}, $\mathbf{1}_T$ is a T by T matrix of ones, and $\mathbf{I}_T$ is a T by T identity matrix. With this construction, I am assuming that the variances for ϵ and ν are estimated perfectly (i.e., there is no term in Σ_{ij} that accounts for variance in those estimates).

Because $\log \widetilde{Y}_{ij}$ is multivariate normal, the distribution of the *level* of production over time, $\widetilde{Y}_{ij}$, is multivariate log-normal with the same parameters. The mean vector and covariance matrix of this distribution are:

$$\begin{aligned}
\widetilde{\mu}_{ij} &= \exp\left(\mu_{ij} + \frac{1}{2}\mathcal{D}\left(\Sigma_{ij}\right)\right) \\
\left[\widetilde{\Sigma}_{ij}\right]_{kl} &= \exp\left(\left[\mu_{ij}\right]_k + \left[\mu_{ij}\right]_l + \frac{1}{2}\left(\left[\Sigma_{ij}\right]_{kk} + \left[\Sigma_{ij}\right]_{ll}\right)\right)\left(\exp\left(\left[\Sigma_{ij}\right]_{kl}\right) - 1\right)
\end{aligned}$$

where $\mathcal{D}(\cdot)$ represents the diagonal vector of a square matrix and $[M]_{xy}$ is the (x,y)-th entry of a matrix M.[46] Finally, $\mathbb{E}\left[DOP_{ij}\right]$ is:

$$\mathbb{E}\left[DOP_{ij}\right] = \sum_{t=1}^{T}\rho^t\widetilde{\mu}_{ijt}$$

A similar calculation is available for the variance of present discounted oil production:

$$\begin{aligned}
\mathbb{V}\left[DOP_{ij}\right] &= \mathbb{V}\left[\sum_{t=1}^{T}\rho^t\widetilde{Y}_{ijt}\right] \\
&= \sum_{t_1=1}^{T}\sum_{t_2=1}^{T}\rho^{t_1+t_2}\left[\widetilde{\Sigma_{ij}}\right]_{t_1,t_2}
\end{aligned}$$

[46] By the properties of the log-normal distribution, the mean and standard deviation of production are closely related, with the standard deviation equal to the mean times the exponent of the variance minus 1. This means that the "correlation" between the mean and standard deviation of production, computed across designs j will be positive by construction.

C Weighted Gaussian Process Estimates

Recall that the mean and variance of the Gaussian process estimates of f at the point $\widetilde{Z}$ are given by:

$$\mathbb{E}\left[f(\widetilde{Z}) \mid g, \mathbf{Z}, \gamma\right] = k(\widetilde{Z} \mid \gamma)^\top K(\gamma)^{-1} g$$
$$\mathbb{V}\left[f(\widetilde{Z}) \mid g, \mathbf{Z}, \gamma\right] = k(\widetilde{Z} \mid \gamma)^\top K(\gamma)^{-1} k(\widetilde{Z} \mid \gamma)$$

where $k(\widetilde{Z} \mid \gamma) = (k(Z_1, \widetilde{Z} \mid \gamma)...k(Z_N, \widetilde{Z} \mid \gamma))^\top$, $K(\gamma)$ is the matrix of pairwise kernel distances for each point in $\mathbf{Z}$ and $g = (g_1...g_N)^\top$. To compute a *weighted* mean and variance, I introduce a weighting matrix function, $L(\lambda)$, and compute a weighted estimate of the mean and variance:

$$\mathbb{E}\left[f(\widetilde{Z}) \mid g, \mathbf{Z}, \gamma, \lambda\right] = k(\widetilde{Z} \mid \gamma)^\top L(\lambda)^\top K(\gamma)^{-1} g$$
$$\mathbb{V}\left[f(\widetilde{Z}) \mid g, \mathbf{Z}, \gamma, \lambda\right] = k(\widetilde{Z} \mid \gamma)^\top L(\lambda)^\top K(\gamma)^{-1} L(\lambda) k(\widetilde{Z} \mid \gamma)$$

The weighting matrix function $L(\lambda)$ biases these estimates towards a firm's own experiences when λ is closer to 0 and towards other firms' experiences when λ is closer to 1. In particular, if $(k_0(\gamma), K_0(\gamma), g_0)$ are the subsets of $k(\gamma), K(\gamma), g$ computed using only the firm's own wells, and $(k_1(\gamma), K_1(\gamma), g_1)$ are the subsets computed using only other firms' wells, then the weighted estimates satisfy 3 relationships:

1. At $\lambda = 0$, the weighted estimates are equal to the estimates computed using the subset of wells the firm operated:

$$k(\widetilde{Z} \mid \gamma)^\top L(0)^\top K(\gamma)^{-1} g = k_0(\widetilde{Z} \mid \gamma)^\top K_0(\gamma)^{-1} g_0$$
$$k(\widetilde{Z} \mid \gamma)^\top L(0)^\top K(\gamma)^{-1} L(0) k(\widetilde{Z} \mid \gamma) = k_0(\widetilde{Z} \mid \gamma)^\top K_0(\gamma)^{-1} k_0(\widetilde{Z} \mid \gamma)$$

2. At $\lambda = \frac{1}{2}$, the weighted estimates are equal to the unweighted estimates:

$$k(\widetilde{Z} \mid \gamma)^\top L\left(\frac{1}{2}\right)^\top K(\gamma)^{-1} g = k(\widetilde{Z} \mid \gamma)^\top K(\gamma)^{-1} g$$
$$k(\widetilde{Z} \mid \gamma)^\top L\left(\frac{1}{2}\right)^\top K(\gamma)^{-1} L\left(\frac{1}{2}\right) k(\widetilde{Z} \mid \gamma) = k(\widetilde{Z} \mid \gamma)^\top K(\gamma)^{-1} k(\widetilde{Z} \mid \gamma)$$

3. At $\lambda = 1$, the weighted estimates are equal to the estimates computed using the subset of wells the firm did not operate:

$$k(\widetilde{Z} \mid \gamma)^\top L(1)^\top K(\gamma)^{-1} g = k_1(\widetilde{Z} \mid \gamma)^\top K_1(\gamma)^{-1} g_0$$
$$k(\widetilde{Z} \mid \gamma)^\top L(1)^\top K(\gamma)^{-1} L(1) k(\widetilde{Z} \mid \gamma) = k_1(\widetilde{Z} \mid \gamma)^\top K_1(\gamma)^{-1} k_1(\widetilde{Z} \mid \gamma)$$

At intermediate values of λ, $L(\lambda)$ interpolates between these extremes. To accomplish this, $L(\lambda)$ takes this form:

$$L(\lambda) = \begin{bmatrix} L_1(\lambda)\mathbf{I}_{n_0} & L_2(\lambda) K_{01}(\gamma) K_{11}(\gamma)^{-1} \\ L_3(\lambda) K_{10}(\gamma) K_{00}(\gamma)^{-1} & L_4(\lambda)\mathbf{I}_{n_1} \end{bmatrix}$$

where n_0 is the number of wells in the firm's information set that it operated, n_1 is the number of wells that other firms operated, the matrices $K_{00}(\gamma), K_{01}(\gamma), K_{10}(\gamma), K_{11}(\gamma)$ are submatrices of $K(\gamma)$:

$$K(\gamma) = \begin{bmatrix} K_{00}(\gamma) & K_{01}(\gamma) \\ K_{10}(\gamma) & K_{11}(\gamma) \end{bmatrix}$$

and the functions L_1, L_2, L_3, L_4 are

$$L_1(\lambda) = 1 + \lambda - 2\lambda^2$$
$$L_2(\lambda) = -\lambda + 2\lambda^2$$
$$L_3(\lambda) = 1 - 3\lambda + 2\lambda^2$$
$$L_4(\lambda) = 3\lambda - 2\lambda^2$$

Thus, $L(\lambda)$ is a quadratic interpolation between $L(0)$, which selects out the firm's own wells, and $L(1)$, which selects out all other firms' wells.

D Geology Covariates

In the production function defined in Section 3, the only spatially varying observable characteristics are the well's location and the fracking choices its operator makes. However, the North Dakota

Geological Survey (NDGS) has published maps of potentially relevant geological information. In this appendix, I describe this data and evaluate its ability to explain oil production. The geology data explains a small, but statistically significant amount of variation in production, even after conditioning on a well's location. However, compared to production function models with location fixed effects, the explanatory power of geology data is small and the coefficients do not always have the signs that would be predicted by geology theory.

D.1 Available Data

The quantity of oil that a well draws from depends broadly on three geological factors: the thickness of the upper and lower Bakken shales, their total organic content, and their thermal maturity. These three factors describe the quantity of rock in the formation, the fraction of the rock that can generate oil, and the likelihood that oil generation has occurred, respectively. Fortunately, in 2008, the North Dakota Geological Survey (NDGS) published maps and GIS shape files documenting the spatial variation in these characteristics over the area covered by the wells in this paper.[47] I summarize this data in Table 11.

As noted in Section 2, thicker locations in the Bakken have the potential to contain more oil. Using data from NDGS map GI-59, Panel A of Table 11 shows the mean, standard deviation and within-township standard deviations of the thickness of the upper, middle, and lower Bakken members across the wells in this paper. The overall Bakken formation averages 86 feet thick, about half of which is the middle member. There is large variation in each of the thickness measures across wells, with the coefficient of variation ranging from 23-40%. However, within a township, the standard deviations of thickness measures are only 22-31% of the overall standard deviations.

In the upper and lower shales, oil can be generated from the fraction of mass that is organic (i.e., containing mostly carbon and hydrogen). All else equal, shale that has a higher organic content has the ability to generate more oil than shale with less organic content. Using data from NDGS map GI-63, Panel B of Table 11 shows the distribution of organic content in the upper and lower shales. In the average well, approximately 14% of the mass is organic in both members. There is limited variation in organic content, overall and within a township. Though not shown in the table, 99% of wells have 9% or more organic content in the upper shale and, 99% have 9.5%

[47] These maps are freely available in PDF format at https://www.dmr.nd.gov/ndgs/bakken/bakkenthree.asp. The shape files are available for purchase from the NDGS.

Table 11: Geology Covariates Summary Statistics

Variable	Mean	Std. Dev	Min	Max	Within Std. Dev
Panel A: Thickness (ft)					
Bakken Formation	86.05	24.03	5.00	155.00	5.26
Upper Shale	16.50	3.74	1.00	31.00	1.16
Middle Member	42.39	13.10	2.50	82.50	2.86
Lower Shale	27.64	10.93	2.50	57.50	2.90
Panel B: Total Organic Content (%)					
Upper Shale	13.80	2.44	3.00	27.00	1.01
Lower Shale	14.01	2.32	8.50	22.50	1.04
Panel C: Thermal Maturity - Hydrogen Index					
Upper Shale	358.10	179.30	75.00	775.00	32.96
Lower Shale	343.27	181.90	25.00	1125.00	65.72
Panel D: Thermal Maturity - S2-TMAX (degrees celsius)					
Upper Shale	435.68	5.60	417.50	447.50	1.92
Lower Shale	433.89	10.12	387.50	447.50	2.81

$N = 2,699$. Reported Bakken Formation thickness does not exactly add up to the sum of the thickness of the three members in the data. "Within Std. Dev" is the standard deviation of the data after subtracting mean values within townships. Source: NDGS Maps GI-59 and GI-63.

or more in the lower shale. For comparison, the organic content in the Ghawar Field of Saudia Arabia, the most prolific oil field in history, is only 5%.[48]

Long term exposure to high temperatures converts organic material into oil. The extent of exposure is called thermal maturity, and geologists use three categories to describe the thermal maturity of a rock sample. Thermally immature rock has less exposure than is necessary for the conversion of organic material into oil. Thermally mature rock has enough exposure for the conversion of its organic content into oil. Thermally over-mature rock has too much exposure, and its organic content is converted into natural gas.

In map series GI-63, the NDGS provides two measures of the thermal maturity of the Bakken: hydrogen index and S2-TMAX. Both measures are collected by heating a rock sample to high temperatures and measuring the rate of oil expulsion across temperatures. The maximum rate at which oil is expelled, divided by organic content, gives the hydrogen index. Since hydrogen is one of the two elements contained in all hydrocarbons, more hydrogen indicates higher hydrocarbon

[48]See Fox and Ahlbrandt (2002)

generating potential. Potential oil production is higher for larger values of the hydrogen index, with thermally mature rock at values as low as 200.[49] The temperature of the highest rate of oil expulsion, called S2-TMAX, is the other laboratory measure of thermal maturity. Thermally mature rock corresponds to S2-TMAX values between 435 and 460, with higher values in that range corresponding to higher oil production. Above 460 degrees celsius, oil production is decreasing, and the rock is thermally over-mature.[50]

Panel C of Table 11 shows the distribution of the hydrogen indices across wells. The average well has a hydrogen index suggestive of thermal maturity for both the upper and lower shales, though approximately 25% of wells are thermally immature. Within a township, the standard deviations of the hydrogen indices are 18-30% of the overall standard deviations. Panel D shows the distributions of S2-TMAX. The average well is just at the start of thermal maturity and no wells are thermally over-mature. Over 80% of wells have S2-TMAX in the range of thermal maturity in the upper shale, and 53% in the lower shale. Within a township, the standard deviations of the S2-TMAX values are 29-34% of the overall standard deviations.

The NDGS developed these maps using the cuttings, cores and well logs that operators are legally required to submit for every well they drill to the NDIC.[51] Since the NDIC makes these samples and logs available to anyone, the information content in these maps may have been known by market participants before they were published.

Opportunities to measure the thickness, total organic content or thermal maturity of the rock in a specific well are infrequent.[52] Furthermore, only in the last few years have geologists began to study the use of these cuttings in providing information about well quality.[53] Even if these techniques had been available (and in widespread use), they would only provide information about the middle Bakken member, as that is the predominant source rock for cuttings.

[49]For more information, see McCarthy et al. (2011)

[50]For more information, see McCarthy et al. (2011)

[51]Recall that "cuttings" are the returned rock samples generated during the drilling process. Occaisionally operators also preserve contiguous sections of undrilled rock, called "cores". By North Dakota Century Code 38-08-04, Section 43-02-03-38.1, operators are required to send physical samples of cuttings and cores to the NDGS within 90 days of collection, where they can be publically observed and analyzed by anyone. Additionally, operators are required to submit copies of all well logs and geology tests they perform.

[52]For example, Pimmel and Claypool (2001) notes that "rock eval pyrolysis is not normally used to make real-time drilling decisions because of the lengthy sample preparation, running, and interpretation time."

[53]See, for example, Ortega et al. (2012)

D.2 Explanatory Power

To evaluate the ability of these geology covariates to explain oil production, I estimate Cobb-Douglas production function models with and without them. Table 12 shows these results. Column 1 is a specification with no township fixed effects and no geology covariates (i.e., it is a simplification of the results in Colum 2 of Table 6). In Column 2, I add the township fixed effects, increasing the between R-squared from 0.600 to 0.813, suggesting that location-specific factors explain a large portion of variation in production. Next, column 3 shows a specification with the geology covariates but no township fixed effects. Compared to Column 1, the between R-squared increases by 0.083 to 0.683. The coefficients on 6 of the 8 geology covariates are significantly different from zero and a Wald test rejects the hypothesis that the coefficients on the geology covariates are jointly equal to 0 at the 1% level. However, after conditioning on location, the geology covariates have considerably less explanatory power. Column 4 shows a specification with both township fixed effects and geology covariates. The increase in R-squared values from Column 2 to Column 4 is only 0.003, and only 2 of the 8 coefficients on the geology covariates are statistically significant. Again, a Wald test rejects the hypothesis that the geology covariates are jointly equal to zero. These results show that geology covariates do explain some of variation in oil production, but very little compared to the location fixed effects.

Geology theory predicts that the coefficients on each of these covariates should be positive, as greater thickness, organic content and thermal maturity are all thought to be associated with higher oil production. However, the coefficient on organic content in the upper Bakken shale is negative and statistically significant in both specifications.

The inclusion of geology covariates does not meaningfully change the Cobb-Douglas estimates of the productivity of lateral length, sand or water, as the coefficients in columns 2 and 4 are nearly identical.

E Stability of the Production Function Relationship

In order for firms to empirically learn the production function for fracking, the true relationship between oil production, fracking inputs and location must be stable over time. To verify whether the data is consisent with a stable production function, I examine the performance of wells in

Table 12: Explanatory Power of Geology Covariates

Coefficient	(1) Log Oil	(2) Log oil	(3) Log Oil	(4) Log Oil
β	-0.557	-0.557	-0.557	-0.557
	(0.00237)	(0.00237)	(0.00237)	(0.00237)
δ	1.755	1.754	1.755	1.754
	(0.00355)	(0.00354)	(0.00355)	(0.00354)
η	0.436	0.798	0.549	0.795
	(0.0370)	(0.0373)	(0.0358)	(0.0374)
κ_S	0.233	0.158	0.200	0.155
	(0.0187)	(0.0161)	(0.0172)	(0.0161)
κ_W	0.0521	0.115	0.106	0.115
	(0.0200)	(0.0163)	(0.0182)	(0.0163)
κ_{TU}			0.0400	0.0000848
			(0.00333)	(0.00744)
κ_{TL}			-0.00205	-0.00218
			(0.00114)	(0.00299)
κ_{CU}			-0.0494	-0.0350
			(0.00530)	(0.00915)
κ_{CL}			0.0531	0.00718
			(0.00524)	(0.00960)
κ_{HU}			0.00129	0.000441
			(0.000115)	(0.000268)
κ_{HL}			0.0000825	0.0000323
			(0.0000991)	(0.000134)
κ_{SU}			0.00875	0.0139
			(0.00317)	(0.00446)
κ_{SL}			0.0153	0.00212
			(0.00144)	(0.00337)
Overall R^2	0.690	0.784	0.728	0.785
Between R^2	0.600	0.813	0.683	0.816
Within R^2	0.765	0.765	0.765	0.765
Township Fixed-effects		X		X

Standard errors in parentheses. GLS random efffects estimates of the production function model:

$$\log Y_{it} = \alpha + \beta \log t + \delta \log D_{it} + \eta \log H_i + \kappa Z_i + \tau_i + \epsilon_i + \nu_{it}$$

Y_{it} is oil production for well i when it is t months old, D_{it} is the number of days producing, H_i is the horizontal length, and Z_i is the vector of log sand use S_i, log water use W_i, upper Bakken thickness (TU), lower Bakken thickness (TL), upper Bakken organic content (CU), lower Bakken organic content (LU), upper Bakken hydrogen index (HU), lower Bakken hydrogen index (HL), upper Bakken S2-TMAX (SU) and lower Bakken S2-TMAX (SL), and τ_i is a set of township fixed effects. "Between" R^2 is the R^2 for the average predicted log baseline production. "Within" R^2 is the R^2 for the predicted time series of production. Estimated off of all $2,699$ wells and $91,783$ well-months.

similar locations that are fracked with similar designs but in different time periods. If similar wells fracked in different time periods have different performance, on average, then its possible that the production function is not stable over time.

To implement this test, I estimate two time varying production function specifications and conduct Wald tests of the hypothesis that the time effects are jointly equal to zero. In the first specification, I assume that baseline production is Cobb-Douglas with time-varying coefficients. If the true production function is both Cobb-Douglas and stable, the coefficients should not vary over time. In the second specification, I assume that baseline production is the sum of a year fixed effect and a fixed effect for wells with similar input choices and locations. To do this, I form groups of wells that have the same deciles of sand and water use that are also in the same township. Thus, this specification allows for a non-parametric relationship between baseline production, location and inputs. If the true production function is not Cobb-Douglas, but still constant over time, the time fixed effects should equal zero.

Table 13 shows the results of these tests. The specifications in columns 1-3 are Cobb-Douglas in lateral length, sand use and water use, with time fixed effects and time fixed effects interacted with the sand and water use coefficients.[54] Column 1 shows estimates computed from the whole sample. Wells in the 2009 and 2010 cohorts are significantly more productive than wells in the earlier cohorts, and wells in the 2009 cohort are significantly less sensitive to water use than wells in earlier cohorts. Column 2 shows estimates computed from the set of wells that are in bins with 2 or more wells. In this specification, wells in 2009 and 2010 are also more productive than earlier wells, while wells in 2011 are less productive. Wells in 2010 and 2011 are less sensitive to sand use than ealier wells, and wells in 2011 are more sensitive to water use. Column 3 shows estimates computed from the set of wells that are in bins with 2 or more wells fracked in two or more years. Wells in 2010 are more productive than earlier wells, while wells in 2011 are less productive. In this specification, none of the interaction terms are significantly different from zero. In all 3 specifications, a Wald test of the hypothesis that the year effects and their interactions are jointly equal to zero is rejected at the 1% level. These parametric results suggest that if the true production technology is similar to Cobb-Douglas, then it's parameters may not be constant over time.

[54]Because there are only 124 wells fracked between 2005 and 2007, I include them in the 2008 cohort, and specify year dummies for the 2009, 2010 and 2011 cohorts.

Table 13: Stability of Production Function Estimates Over Time

Coefficient	(1) Log Oil	(2) Log Oil	(3) Log Oil	(4) Log Oil	(5) Log Oil	(6) Log Oil
β	-0.557	-0.557	-0.573	-0.557	-0.557	-0.573
	(0.00237)	(0.00297)	(0.00405)	(0.00237)	(0.00297)	(0.00405)
δ	1.754	1.797	1.846	1.753	1.796	1.846
	(0.00354)	(0.00452)	(0.00633)	(0.00355)	(0.00453)	(0.00633)
η	0.761	0.734	0.708	0.850	0.847	0.801
	(0.0399)	(0.0561)	(0.0728)	(0.0680)	(0.0683)	(0.0825)
κ_S	0.159	0.177	0.171			
	(0.0318)	(0.0390)	(0.0621)			
κ_W	0.150	0.114	0.107			
	(0.0386)	(0.0446)	(0.0690)			
κ_{09}	0.785	0.721	0.645	-0.0236	-0.0224	-0.0232
	(0.218)	(0.282)	(0.354)	(0.0473)	(0.0475)	(0.0476)
κ_{10}	0.706	0.732	1.005	-0.124	-0.121	-0.119
	(0.263)	(0.343)	(0.412)	(0.0589)	(0.0592)	(0.0597)
κ_{11}	0.0933	-0.892	-1.602	-0.161	-0.155	-0.153
	(0.227)	(0.334)	(0.490)	(0.0645)	(0.0648)	(0.0657)
$\kappa_{S,09}$	0.0285	0.0253	-0.0186			
	(0.0408)	(0.0496)	(0.0644)			
$\kappa_{S,10}$	-0.0621	-0.137	-0.0851			
	(0.0456)	(0.0670)	(0.0910)			
$\kappa_{S,11}$	-0.0190	-0.0235	0.147			
	(0.0423)	(0.0776)	(0.110)			
$\kappa_{W,09}$	-0.182	-0.177	-0.121			
	(0.0518)	(0.0690)	(0.0920)			
$\kappa_{W,10}$	-0.0577	0.000315	-0.118			
	(0.0523)	(0.0830)	(0.114)			
$\kappa_{W,11}$	0.00842	0.177	0.116			
	(0.0469)	(0.0838)	(0.116)			
# Well-months	91,783	50,866	25,939	91,783	50,866	25,939
# Wells	2,699	1,399	708	2,699	1,399	708
Overall R^2	0.785	0.805	0.808	0.836	0.827	0.823
Between R^2	0.816	0.828	0.846	0.952	0.894	0.888
Within R^2	0.765	0.792	0.800	0.765	0.792	0.800
Fixed-Effects	Township	Township	Township	Bins	Bins	Bins
Sample	All	Bins 1	Bins 2	All	Bins 1	Bins 2

Standard errors in parentheses. GLS random efffects estimates of the production function model:

$$\log Y_{it} = \alpha + \beta \log t + \delta \log D_{it} + \eta \log H_i + \kappa Z_i + \tau_i + \epsilon_i + \nu_{it}$$

Y_{it} is oil production for well i when it is t months old, D_{it} is the number of days producing, H_i is the horizontal length, and Z_i is the vector of log sand use S_i, log water use W_i, dummies for the 2009, 2010 and 2011 cohorts, and interactions between the dummies and log sand use and log water use. τ_i is a set of fixed effects for townships or bins. "Between" R^2 is the R^2 for the average predicted log baseline production. "Within" R^2 is the R^2 for the predicted time series of production. "Bins 1" is the sample of wells in bins with 2 or more wells, while "Bins 2" is wells in bins with 2 or more wells, fracked in 2 or more years.

Next, columns 4-6 show estimates for the non-parametric specification. Again, column 4 is estimated on the entire sample, column 5 is estimated on the sample of wells in bins with 2 or more wells, and column 6 is estimated on the sample of wells in bins with 2 or more wells fracked in 2 or more years. In these specifications, the later cohorts tend to be less productive than the earlier cohorts. Wells in the 2010 and 2011 cohorts are significantly less productive than wells in the 2008. However, a Wald test of the hypothesis that the year effects are jointly equal to zero cannot be rejected at the 5% level in any of the nonparametric specifications, providing some support to the idea that the production function is stable over time.

Since the true production function is unlikely to be spatially homogenous or monotonic in sand and water or, the non-parametric results here may be more relevant.