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

Thomas R. Covert*

February 22, 2015

Abstract

Little is known about how firms learn to use new technologies. Using novel data on inputs, profits, and information sets, I study how oil companies learned to use hydraulic fracturing technology in North Dakota between 2005-2012. Firms only partially learned to make profitable input choices, capturing just 60% of possible profits in 2012. To understand why, I estimate a model of input use under technology uncertainty. Firms chose fracking inputs with higher expectations but lower uncertainty about profits, consistent with passive learning but not active experimentation. Most firms over-weighed their own information. These results provide evidence of impediments to learning.

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. Theoretical models of learning predict that rational agents (such as firms) efficiently analyze information about new technologies, invest in experiments to create new information, and incorporate information generated by others.[2] However, to test these models, it is necessary to measure the information that firms have, which is challenging 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 input choices, profits, and direct measures of the information firms had when making those choices. The oil companies in this data learn to use

*University of Chicago Energy Policy Institute and Department of Economics; tcovert@uchicago.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.

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 for several reasons. First, the profit maximizing choice of fracking inputs may vary across drilling locations in unpredictable ways, so firms must empirically learn the relationship between inputs and outputs over time and adjust their behavior as they learn. Thus, there is a well defined learning problem that many firms simultaneously face. Second, 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. This information is not disseminated until 6 months after a well is fracked, making it possible to precisely measure differences in knowledge about fracking across firms. Third, the industry is not concentrated, which motivates studying learning as a single agent problem. During the time period I study, there are 82 active firms, the market share of the largest firm is only 11% and the combined share of the five largest firms is under 50%. Fourth, 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. Finally, 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 $8 million, while the average profit for each well's most profitable choice is $16 million. Since the regulator in North Dakota expects that 40,000 wells will eventually be fracked over the next 20 years, the potential for lost profits from inefficient learning is substantial.[3]

Learning-by-doing and social learning are both important in this context. In 2005 and 2006, the average well is fracked by a firm that had fracked only a single well before. By 2012, the average well is fracked by a firm that had previously fracked 171 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 in addition to their own. Between 2005 and 2006, the average well is fracked by a firm that can observe 9 wells previously fracked by other firms, a number which rises to 2,786 in 2012. 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 2012 (82 firms and 4,408 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, which limits the likelihood of 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

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

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.

Existing research measures learning from experience-driven upward trends in *productivity*, or residual production that is not explained by input choices. I measure the extent to which these firms learn to be more productive by estimating the correlation between the production function residuals with direct measures of their experience. Surprisingly, there are no systematic correlations between the production function residuals and measures of experience. If anything, firms with more experience are empirically less productive, but the differences are small.

In contrast, firms do learn to make more *profitable* input choices. Wells fracked in early years capture only 20% of the profits that optimally fracked wells would have produced. However, profit capture grows almost monotonically over time, with firms capturing 61% of maximal profits in 2012. 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.

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 80% of *ex ante* optimal profits in 2005, they capture only 60% in 2012. 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. The firms in this data are more likely to select fracking designs with higher expected profits and *lower* standard deviation of profits. Their choices indicate that they are indifferent between a \$0.34-\$0.63 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 estimate the weight that firms place on their own experience, relative to their competitors' experiences. Most firms place more weight on their own experiences than their competitors' experiences. Even after allowing firms

to learn differently from their own data than their competitors', 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 is 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.

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

[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).

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

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. Much of this evidence is in agricultural settings. Ryan and Gross (1943), 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 material 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.

[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)

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 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 drilling locations.[20] Because the research documenting these results was not publicly available to firms during the time period I study, it is possible that they did not know these facts initially.

[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.

[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)

[20] See Baihly et al. (2012)

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]

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, 2012. 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

[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.

[22] See Larsen (2011)

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. 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 the set (H_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 2005-2012, I have production data through November 2014, making it possible to study the performance of all wells for at least two years. 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 308 townships in North Dakota with Bakken wells during 2005-2012. 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.

In addition to the well characteristics, I also collect geological characteristics of the rock into which each well is drilled. The data comes from the North Dakota Geological Survey maps and GIS shape files published in 2008, and provides *estimates* of the thickness of the upper and lower Bakken shales, their total organic content, and their thermal maturity.[24] 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. For each well, I compute thickness as the sum of the estimated thicknesses for the upper and lower Bakken shales, total organic content as the thickness-weighted average of the estimated total organic content of the upper and lower Bakken shales, and thermal maturity as a composite index of two thermal maturity measures reported in the maps.[25]

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

[24] Specifically, the data come from NDGS maps GI-59 and GI-63.

[25] Organic material is converted into petroleum following long term exposure to high temperatures. The extent of this 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. 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 generating potential. Potential oil production is higher for larger values of the hydrogen index, with thermally mature rock at values as low as 200. The temperature of the highest rate of oil expulsion, called S2-TMAX,

The NDGS developed these maps by analyzing and spatially interpolating data that operators are legally required to submit to the NDIC. This data includes "cuttings", which are the returned rock samples generated during the drilling process, "cores", which are contiguous sections of undrilled rock, and "well logs", which show the underground measurements taken during drilling and completion.[26] The NDIC makes this data available to anyone[27], so the information content in these maps may have been known by firms before they were published in 2008.

Though these maps provide *estimates* of organic content, thickness, and maturity at a given location, opportunities to measure the *actual* geological characteristics of the rock in a specific well are infrequent. The most reliable way to do this is by drilling a core, which is slower and more expensive than simply drilling a well. As a result, there are a limited number of wells that have ever been cored, and of the 4,408 wells studied here, only 97 have been cored. The results of certain geology tests can also reveal useful information about organic content and maturity, but according to the NDGS, these tests were performed on only a third of the wells in this study. Moreover, its not known whether these logs were analyzed before or after these wells were fracked.[28] Though cuttings could be informative about organic content and maturity, geologists have only began to study their use in providing information about well quality.[29] Even if these techniques had been available (and in widespread use) between 2005-2012, they would only provide information about the middle Bakken member, as that is the predominant source rock for cuttings. Thus, this geology data represents only a proxy for the underlying quality of the rock into which a well is drilled.

Table 1 reports the cross-sectional distribution of well characteristics and oil production in the first year.[30] There is substantial variation across wells in both fracking input use and oil production. The 75th percentiles of sand, water and oil production are approximately 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 10 bbl per foot of horizontal length in its first year. Since the price of oil averaged $76 per bbl during 2005-2012, the value of production in the first year for the average

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. For more information, see McCarthy et al. (2011). I combine these thermal maturity measures into a single thermal maturity score for each well, *mature*, defined as the thickness-weighted average of the maturity scores for the lower and upper Bakken shales. The maturity score for either layer is 0.5 if the layer is either mature by the hydrogen index standard or the S2-TMAX standard, 1 if it is mature by both standards, and 0 if its mature by neither standard. Thus the combined score, *mature* ranges from 0 to 1.

[26]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, as well as the results of well logs and geology tests, to the NDGS within 90 days of collection, where they can be publicly observed and analyzed by anyone.

[27]Subject to well confidentiality constraints

[28]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."

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

[30]Because they are self reported and sometimes appear to contain typographical errors, I winsorize the lateral length, sand use and water use variables at the 0.5% level

well is worth $6.5 million. Most wells tend to produce on the majority of days during a month, and though not shown in the table, only 93 wells have fewer than 20 average producing days.

The middle two panels of Table 1 show the distribution of past experience across wells. The average well is fracked by a firm that has previously fracked 115 of its own wells, participated in 243 other wells, and can observe the data on 1,446 wells fracked by others. Among wells that are within 1 township of the well a firm is about to frack, that firm can draw on its own experience for 27 wells, the experience of 25 wells it participated in, and and can observe an additional 33 wells as a result of the public disclosure process.

The bottom panel of Table 1 reports the cross-sectional distribution of the geology covariates. Most wells are drilled into rock that is high in organic content, and there is little variation in organic content across wells. The average well is drilled into rock that is estimated to be 14% organic, by mass, and 75% of wells are drilled into rock with TOC at or above 13%.[31] As noted in Section 2, thicker locations in the Bakken have the potential to contain more oil. Across all wells, the combined thickness of the upper and lower Bakken shales averages 43.74 feet, with a standard deviation of 13.5 feet. Finally, most wells have at least some rock that is considered thermally mature. The average maturity score is 0.64, the median is 0.5, and the interquartile range is (0.5,0.78). 58% of wells have a maturity score of 0.5.

Table 2 shows the distribution of well characteristics, oil production, and geology covariates over time. The number of wells fracked and the number of active firms both increase in every year. Nearly 38% of all wells in the sample are fracked during the last year, and in that year. Over time, firms frack longer wells, using more sand and more water. Firms operating in 2012 use nearly three times as much sand and five times as much water per foot of horizontal length, on average, as firms in 2005. Although not shown in the table, there is also meaningful variation in input use within townships during the same year.[32] However, average oil production does not rise monotonically. It peaks in 2008 and slowly falls in each subsequent year. Geology characteristics are stable across years.

Oil production is correlated with sand and water use. Table 3 reports estimates of a regression of oil production per foot of horizontal length onto township fixed effects and dummy variables for quintiles of sand and water use average oil production per foot by quintiles of sand and water use. Across the first four quintiles of sand and all quintiles of water use, the highest input levels are associated with higher oil production, a result that is consistent with the physical intuition behind fracking. However, at high levels of sand use, the relationship between water use and oil production appears to be non-monotonic, which is also consistent with the petroleum engineering evidence in section 2.2.

[31] For comparison, the rock in Saudi Arabia's Ghawar Field, the most prolific oil field in history, is only 5%. See Fox and Ahlbrandt (2002).

[32] The within-township-year standard deviation of sand and water use are 76 lbs per foot and 86 gals per foot, respectively.

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, and daily spot prices for Bakken crude oil at the Clearbrook, Minnesota hub from Bloomberg. 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. However, given its geographic proximity, the Clearbrook price may be a better representation of the price firms in this data actually receive. Unfortunately, Bloomberg did not start recording Clearbrook prices until October, 2010, so it is necessary to use Cushing prices before then. Between October 2010 and December 2012, the Clearbrook price was \$1.75 per bbl less than the Cushing price, on average. Thus, I assume that between January 2005 and September 2010, firms received the Cushing price, minus \$1.75 per bbl, and after October 2010, they received the Clearbrook price. Figure 3 plots the quarterly averages of these. Between 2005-2012, there was a boom and bust in oil prices, with prices climbing from approximately \$50 per bbl in early 2005, reaching more than \$120 per bbl in mid 2008 and falling to \$45 per bbl in early 2009. In 2010-2012, when more than 78% of the wells are fracked, oil prices average \$89 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 rights.[33] 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 4,408 wells in this dataset, the cost records for 199 are in the public domain for this reason.[34]

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 2005-2012, 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.[35] Finally, starting in the first

[33] 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.

[34] I was also able to find the cost information for an addition 22 wells by contacting the North Dakota Land Trust and by conversations with a private investor in North Dakota.

[35] Spears & Associates surveys independent engineers in North Dakota quarterly, asking them to estimate the cost of a reference well. 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

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. Figure 4 plots the cost index over time.

To estimate the individual components of costs, I assume that costs are the sum of five components: the cost of the vertical portion of the wellbore w, the fixed cost of having any frack job f, the variable costs of drilling the horizontal segment v, and the costs p_s of pumping sand and p_w of pumping water. That is:

$$\text{total cost} = w + f + v \times L + p_s \times S + p_w \times W$$

The Spears & Associates data include estimates for the cost of the vertical wellbore, so I use that data directly for w. To estimate the remaining cost terms, I subtract the cost of the vertical wellbore from reported costs, divide the remainder by the cost index, and regress those values onto a constant, the lateral length of the well, and the sand and water use for the well. The R-squared of this regression is 0.27, and all coefficients are positive and significantly different from zero at the 5% level. I define the fixed cost of fracking as the constant, the variable cost of drilling as the coefficient on lateral length, and the sand and water costs as the coefficients on sand and water use. I generate time-specific costs by multiplying these estimates by the cost index. Finally, since the fixed cost of fracking and the cost of the vertical wellbore are both "fixed", I combine them into a single fixed cost. Figure 5 plots these costs over time. Note that except for the fixed cost, which changes as the cost of vertical drilling changes, the only source of time variation in costs is driven by changes in the cost index. With these cost estimates, the average well in the sample cost $7.2 million.

2.5.1 Information Sets

Firms can learn about fracking from three sets of wells. First, they can observe all wells that the regulator has made public. This public knowledge includes wells that a given firm operated and wells that other firms operated. Second, firms can observe their own wells which are not yet public knowledge, due to well confidentiality. Third, firms can observe other firms' wells in which they are non-operating participants. 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 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

are due to changes in prices, not quantities.

records lease assignments.[36] 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.[37] This dataset has one or more leases for the stated operator of 93% of the wells, and has mineral rights owned by at least one firm for 99% of the wells. 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.[38]

2.5.2 Outside Experience

Throughout the paper, I assume that the only relevant knowledge firms have about fracking comes from the wells fracked in North Dakota during 2005-2012. To assess the validity of this assumption, I collect firm-specific drilling history from IHS International for the 10 most active firms in my data, which I report in Table 4. This data records the locations, both geographic and geologic, well characteristics and timing of every well drilled by these firms in the United States. The data does not indicate whether these wells had a fracture treatment. In the first column, I list the number of wells each firm completed in the Bakken during 2005-2012. These 10 firms frack 64% of the wells in the dataset. During the time period I study, 9 firms are public corporations, either as independent entities (Brigham[39], 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). Petro-Hunt is privately held.

On the right hand side of Table 4, I list the US operating history of these firms outside of North Dakota. In the 10 years prior to the period I study, these firms collectively completed tens of thousands of conventional, non-shale wells.[40] However, they only completed 299 shale wells, suggesting that they collectively had very little experience with the technology necessary to develop wells in the Bakken Shale. Only EOG had previously completed more than 100 shale wells, and four firms had done none. During 2005-2012, all ten firms are active outside North Dakota, with five 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 non-shale wells, though all firms firms do complete non-Bakken shale 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

[36] I also collect additional lease information from NDRIN for counties and time periods not covered by DrillingInfo.

[37] To account for the possibility that firms may have older mineral rights that are not in mineral rights databases yet, I also assume that firms who operate pre-2005 non-shale wells in a section ($\frac{1}{36}^{th}$ of a township) also have leases in that section.

[38] That is, I assume that no mineral rights owners are non-participants. Since only 199 out of 4,408 wells in this time period had risk penalty challenges, and risk penalty cases rarely happen between operators, this is likely a reasonable assumption.

[39] In early 2012, Statoil ASA, a publicly traded Norwegian oil & gas company, purchased Brigham. Prior to 2012, Brigham was publicly traded in the US.

[40] I define a "shale" well as a well with a horizontal segment that is drilled into a formation listed on the US EIA shale map, available here: http://www.eia.gov/oil_gas/rpd/shale_gas.pdf.

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. Because wells with longer horizontal segments have more contact with oil-bearing rock, it is likely that $\eta > 0$. However, it is not obvious that longer wells should proportionately produce more than shorter wells (i.e., that η should equal one). Fracking treatments applied to the *toe* of the well, the point on the horizontal segment furthest away from the vertical segment, may not be as effective as treatments applied to the *heel*, where the wellbore switches from vertical to horizontal. If this is the case, η may be less than one. The intercept α measures average Hicks-neutral baseline productivity and I discuss ϵ_i 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. Current petroleum engineering suggests that this relationship differs across locations and is nonlinear and non-monotonic in its inputs. For this reason, I estimate $f(Z_i)$ non-parametrically, using Gaussian process regression, or GPR. GPR makes kernel regression techniques easily available within a panel data framework and have a natural interpretation in learning settings. 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. This probability distribution is defined by two functions: a *mean function* $m(Z)$ and a positive definite *covariance function* $k(Z, Z')$. The mean function of a Gaussian Process G is the expectation of the value of a function f drawn at random from G evaluated 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 is generated by a function drawn from a Gaussian process with mean $m(Z)$ and covariance $k(Z, Z')$. This likelihood can be used to find the best fitting mean and covariance functions for the data or to compute posterior beliefs about the distribution of $f(Z)$.

Conditional on a dataset $(g_i, Z_i)_{i=1}^N = (\mathbf{g}, \mathbf{Z})$, the posterior distribution of f evaluated at an out-of-sample point $\widetilde{Z}$ is normal, with mean and variance given by:

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

where $k(\widetilde{Z}) = (k(Z_1, \widetilde{Z})...k(Z_N, \widetilde{Z}))^\top$. This is a result of the normality property above, and the Gauss-Markov theorem. Note that the formula for the mean of $f(\widetilde{Z})$ is similar to the formula for the estimated regression function in kernel regression[41] and that the formula for the variance does not depend on the observed values $\mathbf{g}$.

Gaussian processes are commonly used in the artificial intelligence and operations research literatures, though their application in economics is now becoming more common, with recent work by Chetverikov et al. (2013), Kasy (2013), Meagher and Strachan (2014) and others. For a detailed treatment of Gaussian processes, see Rasmussen and Williams (2005).

3.3 Specification of $k(Z, Z')$ and $m(Z)$

To estimate the Gaussian process portion of the production function for fracking, I must make assumptions about the form of the covariance and mean functions. For the covariance function, I

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

assume that $k(Z, Z')$ is a multivariate normal kernel:[42]

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

The first parameter, γ_0, controls the variance of realizations of functions drawn from the Gaussian Process for $f(Z)$. To see this, note that as points (Z_i, Z_j) become arbitrarily close to each other, the covariance function approaches $\exp(2\gamma_0)$. Thus, higher values of γ_0 are associated with Gaussian processes whose functions are less predictable. The remaining parameters $\gamma = (\gamma_S, \gamma_W, \gamma_{lat}, \gamma_{lon})$ measure how smooth realizations of f are in each dimension. Larger values of γ_d are associated with function realizations that are flatter across dimension d.

I assume that the mean function $m(Z)$ is a Cobb-Douglas production function for sand and water use that varies across locations. To parameterize the dependence of the Cobb-Douglas parameters on location, I assume they are linear combinations of the location-specific geology characteristics in the bottom panel of Table 1. Specifically, I assume that:

$$\begin{aligned} m(Z) &= m(S, W, lat, lon) \\ &= \omega_0 + R(lat, lon)\omega_R + \log(S)\left(\omega_S + R(lat, lon)\omega_{R,S}\right) + \log(W)\left(\omega_W + R(lat, lon)\omega_{R,W}\right) \\ &= m(Z, R \mid \omega) \end{aligned}$$

where $R(lat, lon)$ is a row vector of total organic content, total formation thickness and the maturity score specific to (lat, lon), and the ω's are parameters to be estimated. The first parameter, ω_0, represents the average Hicks-neutral productivity common to all wells and fracking input choices. The vector ω_R represents Hicks-neutral productivity shifters which depend on the location of a well through its geology characteristics. If these parameters are nonzero, the mean function predicts that different locations have different amounts of recoverable oil. ω_S and ω_W represent the average Cobb-Douglas productivity of sand and water use in fracking. Finally, $\omega_{R,S}$ and $\omega_{R,W}$ are vectors of Cobb-Douglas productivity shifters for sand and water, respectively. If these parameters are nonzero, then the mean function predicts that the sensitivity of oil production to fracking inputs varies across locations with different geological characteristics.

3.3.1 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.[43] 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

[42] This is the most commonly used covariance function in applied computer science and operations research studies.

[43] 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.

the well pipe.[44] 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 sand and water choices are independent of chemical choices.[45]

The growing petroleum engineering literature on the Bakken emphasizes the importance of spatial variation in explaining both recoverable oil and sensitivity to fracking inputs.[46] Some of this variation has observable proxies, like thickness, TOC and maturity. However, there is much less publicly available data regarding other important rock characteristics, including permeability. If firms have private information about permeability or other geology characteristics, 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.[47] These tests are only possible if firms elect to drill the vertical portion of the wellbore all the way through the entire Bakken formation, and collect a core sample, which they rarely do. To emphasize this point, consider that 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.

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 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.[48] 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.

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

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

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

[47] 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.

[48] 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."

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

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

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.

3.4 Zero Production Events

Empirically, wells occasionally do not producing at all during a month. This is typically caused by logistical delays in maintenance events, as subsequent production is at or above previous production levels. However, since the above production function model is specified in logs, I am implicitly assuming non-zero production always happens. To account for the zero production months in my data, I compute the empirical probability of zero production, conditional on a well's age, using a linear probability model. Later on, I use these probabilities in computing the distribution of present discounted revenues.[49]

3.5 Likelihood

I calculate 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.

3.6 Production Function Estimates

Table 5 shows maximum likelihood estimates of several production function specifications. Column 1 shows a baseline Gaussian process specification, in which the non-constant mean function coefficients are set to zero, while column 2 shows the full mean function. The remaining columns show estimates of a production function which replaces the non-parametric term $f(S_i, W_i, lat_i, lon_i)$ with township fixed effects and the mean function itself. In columns 3 and 4, coefficients for the geology covariates and their interactions with sand and water use are set to zero, similar to column 1. Columns 5 and 6 show unrestricted estimates.

Across all six specifications, the common coefficients $(\alpha, \beta, \delta, \eta)$ have similar values that are precisely estimated. As expected, wells produce less oil per month as they age, with an estimated

[49] The probability of zero production in a month during the well's first year is 0.0171, 0.0226 in the second year, 0.0297 in the third year, 0.0407 in the fourth year, and 0.0697 in the fifth and later years.

log decline rate of approximately -0.56.[50] The coefficient on days producing is approximately 1.17, suggesting that when wells produce for less than a full month, production per day is lower than when wells for the whole month. Wells with longer horizontal segments produce more oil than wells with shorter segments, but the effect is not linear and estimates of its magnitude are different with and without spatial controls. In specifications that include some kind of spatial controls (columns 1, 2, 4 and 6), the estimated return to doubling the horizontal length of a well is an increase in baseline production of 80-85%. In the Cobb-Douglas specifications without township fixed effects (columns 3 and 5), the return is considerably smaller, only 43-48%, suggesting that there may be large differences across locations in baseline production that are correlated with firms' horizontal length choices.

In both Gaussian Process specifications (columns 1 and 2), the estimated values of the smoothness parameters indicate that production varies more across locations than across input choices. To see this, note that the correlation in baseline production between two wells that are identical except for a difference Δ in input k is $\exp\left(-\frac{1}{2}\frac{\Delta^2}{\exp(2\gamma_k)}\right)$.[51] Thus, for a pair of wells in the same township (i.e., nearly identical latitude and longitude) whose sand use differs by 128 lbs/foot (1 standard deviation), the correlation in production is 0.93.[52] By comparison, for wells that are located in vertically adjacent townships (so that they are approximately 0.12 longitude degrees away from each other) and have identical fracking inputs, the correlation is only 0.25.

Overall, the estimated geology data does a poor job of explaining baseline production. Most of the parameter estimates for the mean function in column 2 are not significantly different from zero, and take on values that are inconsistent with geological intuitions.[53] More thermally mature areas have higher production, while areas with higher organic content have lower production, and thickness does not reliably predict production. Most of the remaining mean function coefficients are not significantly different from zero. Moreover, the covariance function smoothing parameter estimates in columns 1 and 2 are statistically indistinguishable, and the R^2 values of the two models differ by less than 0.001. Similar patterns occur in the parametric specifications as well. These results suggest that the estimated geology covariates provide limited explanatory power.

The estimates in columns 2 and 6 both fit the data reasonably well, having R^2 values of 0.74 and 0.69, respectively. Much of this is driven by a strong cross sectional fit, with "between" R^2's, which measure the correlation of predicted baseline production and actual baseline production, are higher, at 0.81 and 0.71, 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.

[50] 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.

[51] This is a direct application of the definition of the covariance function specified in section 3.3

[52] For a one standard deviation difference in water use (138 gals/foot) the correlation is 0.89.

[53] Recall that geology models predict that production should increase with thickness, thermal maturity and organic content.

To visualize estimates of the 4-dimensional baseline production function, Figure 6 shows contour plots over sand and water use, evaluated at average GPS coordinates of the most active township, as well as its northern neighbor.[54] The contour lines are iso-production curves, or combinations of sand and water choices with the same estimated baseline production. The top two panels show Gaussian Process estimates (column 2) at these two locations. In both locations, greater sand use is associated with higher oil production for almost all levels of water use, while greater water use is associated with lower production, except at the highest level of sand use. These contour plots highlight how different the production function can be at nearby locations. In the first panel, the input bundle with highest average production is at the highest levels of sand use and lowest levels of water use. In the second panel, highest production occurs at high levels of sand use and intermediate levels of water use. The "peaks" in the two panels also have different levels, with the left panel attaining a maximum of about 6.60 log points, and the right attaining 6.15 log points. These figures provide strong evidence of non-monotonicity and spatial heterogeneity in the estimated production function.

To demonstrate the limitations of the parametric production function estimates, the bottom two panels show contour plots using Cobb-Douglas estimates (column 6). In these townships, both sand and water use are estimated to have positive returns to scale, so more aggressive frack jobs are always more productive than less aggressive frack jobs. As in the Gaussian Process specifications, the left panel shows higher baseline production levels than the right panel, providing similar evidence for spatial heterogeneity. However, due to the Cobb-Douglas assumption, non-monotonicity is impossible and the sole driver of spatial heterogeneity is differences in the estimated geology covariates.

Figure 6 makes it clear that the Gaussian Process and Cobb-Douglas specifications make fairly different predictions about the impact of fracking inputs on oil production. However, these differences primarily occur in areas of the input space that are "out of sample". For example, near the average sand and water choices for the most active township, 261 lbs and 138 gals per foot, respectively, the Gaussian Process and Cobb-Douglas models both predict approximately 5.80 log points of baseline production. Thus, "in sample", the two models make nearly identical predictions. At input levels further away, they differ starkly. For higher levels of sand use and low to intermediate values of water use, the Cobb-Douglas model predicts 0.25 to 0.75 more log points of baseline production than the Gaussian Process model. In contrast, for higher levels of sand use and lower levels of water use, it is the Gaussian Process model that predicts higher baseline production.

Overall, both production function specifications predict that baseline production is higher under higher sand and water use in most locations. To visualize this, Figure 7 plots the difference between estimated production at "high" and "low" input bundles across space, for both production function specifications. The high input bundle is a frack job with sand use and water use set to their respective 75^{th} percentiles, and the low input bundle sets them to their respective 25^{th} percentiles. Figure 7 is a "heat map", with red areas indicating that high input bundles are more productive

[54] The most active township is 154-92 and its northern neighbor is 154-93

than low input bundles, blue areas indicating the reverse and green areas indicating no productivity differences. The black dots are the locations of wells. The left panel, showing Gaussian Process estimates, indicates that higher inputs are more productive than lower inputs in most locations. In several clusters of wells, the difference is greater than 0.50 log points. One exception is a cluster of townships in the North East quadrant of the play, where higher inputs are significantly less productive than lower inputs. In contrast, the right panel, which shows Cobb-Douglas estimates, indicates that higher inputs are always a bit more productive than lower inputs, with little spatial heterogeneity in the productivity gain.

In order for firms to learn the relationship between baseline oil production, fracking inputs, and location, the relationship must be stable over time. I test for stability by regressing $\widehat{\epsilon_i}$, the estimated random effects from the production function estimates, onto dummy variables for the cohort a well belongs to. Figure 8 graphs the estimates of and confidence intervals for the coefficients on these dummy variables across years. While the 32 wells in the 2005 and 2006 cohorts are indeed less productive than later wells, the remaining 4,376 wells in the later years are equally productive. Though a Wald test rejects the hypothesis that wells in the 2007 through 2012 cohorts are equally productive at the 1% level, this is entirely driven by the wells in the 2009 cohort, which appear to be 0.02-0.05 log points more productive than wells in earlier and later cohorts.

4 Evidence for Learning

I look for three types of evidence that firms are learning. First, I check if wells fracked by firms with more experience produce more oil than similar wells fracked by firms with less experience. That is, I estimate a traditional Learning-by-Doing model, similar to Kellogg (2011) or Benkard (2000). Next, to determine if firms are specifically learning the shape of the fracking production function, I next ask whether firms make more profitable input choices over time. To do this, I calculate profits with respect to information firms had when they were making fracking choices, which I call *ex ante* profits, as well as profits calculated using all of the data in this study, which I call *ex post* profits.

4.1 Learning to be Productive

To measure if firms learn to be more productive, I regress the production function residuals $\widehat{\epsilon_i}$ onto cohort dummies and various measures of experience. "Total" wells is simply the number of wells in a firm's information set, while "operated" wells are those wells in the information set that the firm fracked itself. Because it is likely that the production function is spatially heterogeneous, I also compute the number of "total" and "operated" wells that are within 1 township of a well, which I call "close" experience. Table 6 shows estimates of the coefficient on these experience variables, both in levels and in logs.[55]

Overall, the production function residuals are not systematically correlated with experience. In

[55] Technically, $log(1 + \text{experience})$, for many wells are fracked by firms with no experience.

levels, all measures of experience are negatively correlated with the production function residuals, but only the two "close" experience variables are statistically significantly different from zero. In logs, two of the measures are positively correlated and two are negatively correlated. The only measure of experience with the same precisely estimated sign in both levels and logs is close operated experience, which, surprisingly, is negative. These results suggest that if experience is correlated with production, it is not in the way predicted by a theory of Learning-by-Doing.

4.2 Learning to be Profitable

Though firms may not be learning to be more productive, they might learn to make more profitable choices. This is an important distinction: more productive wells produce more oil, conditional on input choices, but it may still be possible for firms to make more profitable input choices as they better learn the production function for fracking. 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. I call this ratio "profit capture".

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

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

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, M_{it} is a Bernouli random variable for the event that the well has non-zero production, and $c_i(S_j, W_j)$ is the total cost of drilling and fracking that design.[56] The main empirical object needed in the calculation of Π_{ij} is the expected present value of discounted oil production, $\mathbb{E}[DOP_{ij}]$:

$$\begin{aligned}\mathbb{E}[DOP_{ij}] &= \mathbb{E}\left[\sum_{t=1}^{T} \rho^t \widetilde{Y}_{ijt} M_{it}\right] \\ &= \sum_{t=1}^{T} \rho^t \mathbb{E}\left[\widetilde{Y}_{ijt}\right] \Pr(\text{zero production in month } t)\end{aligned}$$

I compute this expectation using the Gaussian Process production function estimates specified in

[56] 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.

column 2 of Table 5 generated by two different information sets: the full data that I have, and the data each firm had when it made a fracking input 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 parameter estimates of the mean function and variance terms from column 2 in Table 5 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 mean function as I do now. This is a strong assumption, and the *ex ante* calculation of expected oil production will be biased if firms had different beliefs than I do. However, it is likely that these biases are small, as decline rates can be predicted using geophysical models[57] and bandwidth and variance parameters do not affect the asymptotic properties the production function estimate.[58] Moreover, most of the impact of fracking inputs and location $f(Z)$ on baseline oil production is computed non-parametrically 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.

4.3 *ex post* Comparisons

Over time, firms choose fracking designs with higher *ex post* expected profits. The top half of Figure 9 plots the *ex post* ratio of actual profits to maximal profits per well.[59] The average fraction of profits captured increases nearly monotonically over time, from 21% in 2005 to 60% in 2012. The bottom half of Figure 9 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 $25.2 million in profits, meaning that in 2008, foregone profits from inefficient fracking choices averaged $14.6 million per well. By 2012, lower oil prices reduced these maximal profits to $15.9 million per well. Combined with the higher fraction of profits captured, firms in 2011 left only $7.0 million per well on the table.

Firms captured more profits by selecting more profitable fracking designs over time. In Figure 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. Through 2008, the average well was fracked with approximately 251 lbs sand per foot less than the profit maximizing level. In the following years, this difference falls monotonically until reaching 96 lbs per foot in 2012. Though the differences in actual and optimal water use start out considerably larger than the differences in sand use, actual water choices get

[57] See Fetkovich (1980).

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

[59] I only include wells in this calculation that have both positive actual profits and positive maximal profits. Over the entire sample, 10% of wells have either negative actual profits or negative maximal profits.

closer to optimal water choices in almost every year. In 2005, firms fracked the average well with 425 gals per foot less water than the water use in the optimal well. By 2012, the difference is only 217 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 more profitable choices.

4.4 *ex ante* Comparisons

Though firms made fracking choices that failed to capture a substantial portion of *ex post* profits, it is possible that, given the information they had, firms may have believed they were making profitable choices. To evaluate this prospect, I repeat the previous analysis, calculating profits using the actual information available to firms when they made fracking input choices. The top half of Figure 11 plots the ratio of actual profits to maximal profits per well using *ex ante* expectations.[60] Firms initially made fracking input choices with expected profits that are close to the optimal choices, capturing 79% of potential *ex ante* profits in 2005. However, *ex ante* profit capture actually falls over time, reaching 60% in 2012, the same level as the *ex post* case in 2012.

While the fraction of profits captured falls, *ex ante* expectations of maximal profits rise from 2005-2008 and again from 2009-2011, as show in the bottom half of Figure 11. 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 $20.6 million per well. Though average oil prices are similar in 2008 ($100 per 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 less than 1 million in 2005 to $8.0 million in 2012.

Firms capture a shrinking fraction of *ex ante* profits over time because actual input use is persistently below the expected profit maximizing level. Figure 12 plots average profit maximizing and actual sand use per well over time. In 2005 and 2006, actual sand use is quite similar to *ex ante* optimal sand use. However, as new data accumulates, optimal sand use increases faster than actual sand use, and by 2012, the difference between optimal and actual sand use is greater than 100 lbs per foot, approximately the standard deviation of sand use in that year. Though this difference is similar to the difference in the *ex post* case during 2012, 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.

The bottom panel of Figure 12 plots average *ex ante* optimal and actual water use per well, showing a similar pattern to the *ex post* case: on average, firms use less than the *ex ante* optimal amount of water in fracking, but make improved water choices over time, especially after 2008. From 2005 to 2008, firms use 303 gals per foot less water than the optimal level, on average. This difference shrinks in each subsequent year, and by 2012, it is only 142 gals per foot, less than the standard deviation of water use in that year.

[60] 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, 12.8% of wells have either negative actual profits or negative maximal profits.

5 Fracking input choice model

Though firms do learn to make more profitable choices 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 on firm preferences. First, firms may care about the uncertainty in their estimates of the profits from fracking with a given choice of inputs. Second, in learning the production function, 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 matter. Firms raise outside capital to finance operations and the presence of debt capital could lead firms to select fracking designs with higher uncertainty, as bond holders 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.

In addition to finance-driven 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, the oil industry as a whole exhibits 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. To answer it, I estimate the preferences firms have over the expectation and standard deviations of the profits to fracking designs. I use a simple logit preference model in which the "utility" a firm has for fracking design j on 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). Previous

empirical work on learning in economic settings has assumed that rational decision makers use UCB learning strategies. For example, Dickstein (2013) estimates the parameters of UCB learning strategies that rationalize prescribing behavior by physicians.

In all of the existing literature on 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. Often, the weight actually increases with the square root of the logarithm of the size of an agent's information set. That is, agents with more information should have stronger tastes for uncertainty than agents with less information. For that reason, I also consider preferences with weight on uncertainty that depends on the size of the information set:

$$u_{ij} = \xi_m \left(\phi P_i \mathbb{E}\left[DOP_{ij}\right] - c_i(S_j, W_j)\right) + \left(\xi_{s,0} + \xi_{s,1}\sqrt{\log \mid I \mid}\right) \phi P_i \left(\mathbb{V}\left[DOP_{ij}\right]\right)^{\frac{1}{2}} + \epsilon_{ij}$$

where $\mid I \mid$ is the number of wells that the firm fracking well i can observe.

Table 7 shows maximum likelihood estimates of these parameters for each of the 10 most active firms, as well as the industry as a whole. For each firm, the first row represents the simplest preference specification (constant preferences for uncertainty over time), while the second row allows for time-varying preferences for uncertainty.[61] Focusing first on the first row, all firms and the pooled industry have positive "taste" for the expectation of profits ($\xi_m > 0$) of a fracking design and negative "taste" for the standard deviation ($\xi_{s,0} < 0$). 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 \$0.34 to \$0.63 for a reduction of \$1 in the standard deviation of profits.

There is some evidence that firms' taste for uncertainty increases with the size of their information sets. For five firms and the industry as a whole, estimates of $\xi_{s,1}$ are positive, though it is not significantly different from zero for two of those firms. In these six situations where $\xi_{s,1}$ is positive, it is not large enough in magnitude to generate positive taste for uncertainty, even when firms have large information sets.[62] For three firms, $\xi_{s,0} > 0$ while $\xi_{s,1} < 0$. However, this is not evidence that these firms have a positive taste for uncertainty. In two of these situations, $\xi_{s,0}$ is not significantly different than zero, and in the third, the combined taste for uncertainty is negative for information sets with 43 or more wells.

Overall, Table 7 provides 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 aversion to uncertainty.

[61] The third and fourth rows are discussed in the next section

[62] To have positive taste for uncertainty, $\xi_{s,0} + \xi_{s,1}\sqrt{\log \mid I \mid} > 0$. When $\xi_{s,0} < 0$ and $\xi_{s,1} > 0$, this happens if $\sqrt{\log \mid I \mid} > \frac{-\xi_{s,0}}{\xi_{s,1}}$. Since the largest possible information set has fewer than 4,408 wells, $\sqrt{\log \mid I \mid} < 2.89$, while $\frac{-\xi_{s,0}}{\xi_{s,1}}$ is never smaller than 3.8 for the six firms where $\xi_{s,0} < 0$ and $\xi_{s,1} > 0$.

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 the ones they hold. There are many ways that a firm's beliefs may be different than the ones I calculate: firms may have different priors about the relationship between fracking inputs, location and oil production, they may have simpler beliefs about the functional form for that relationship, or my price and cost data could be different from the prices and costs firms experience. However, I focus on a simpler explanation: 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 possibility as "own-data bias".

I interpret the "weight" that firms place on a particular well as their beliefs about the variance of the well-specific baseline production shock, ϵ. If firms weigh their own experience more than the experiences of others, then when forming estimates of the production function, they will believe that production outcomes from their own wells have less variance than production outcomes from their competitors' wells. To measure whether this is happening, I decompose the variance of baseline production in well i perceived by the operator of well j into two components:

$$\log \sigma_{\epsilon_{ij}} = \widehat{\log \sigma_\epsilon} + \lambda \mathcal{I}(i \text{ and } j \text{ operated by same firm})$$

The first component, $\widehat{\log \sigma_\epsilon}$, is the estimated value of $\log \sigma_\epsilon$ from column 2 of Table 5. The second component, λ, is the increase or decrease in perceived variance that operators assign to their own data, relative to data they only observed. Given a value of λ, I can compute an own-data-biased production function estimate, and use this estimate to calculate the distribution of own-data-biased discounted oil production, which I write as $DOP \mid \lambda$. Finally, I can use the preference model from the previous section and firms' input choices to estimate the value of λ that best rationalizes firms' preferences.

Table 7 reports maximum likelihood estimates of λ. Overall, the data suggest that firms weigh their own information more heavily than information generated by their competitors. Of the ten most active firms, seven have values of $\lambda < 0$, five of which are statistically significantly different from zero. The estimated λ for the industry as a whole is also significantly negative. Though three firms have estimates which are positive, only one of which is significantly different from zero. The magnitudes of these estimates are large. For example, the estimate for the pooled industry suggests that firms believe that the standard deviation of the shocks to log baseline production for their own wells is about 20% smaller than the standard deviation of shocks for their competitor's wells. Moreover, some firms appear to be especially confident about their own data. For example, Marathon's revealed preferences are consistent with a belief that the standard deviation of shocks to their own wells are 64-74% lower than the shocks to their competitors' wells. Only Whiting's revealed preferences suggest it believes its own wells are *more* variant than its competitors' wells. Importantly, allowing firms to have own-data bias does not change their preferences over uncertainty. No firms have positive taste for uncertainty, even when $\lambda \neq 0$.

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 20% to 60% 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 less uncertainty, and are willing to give up \$0.34-0.63 in expected profits for a reduction of \$1 in the standard deviation of profits. Finally, most firms 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 there is enormous value to better information. 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 4,408 wells in this data, the average well appears to forego \$8.2 million in profits on an *ex post* basis and \$6.5 million on an *ex ante* basis, resulting in \$29-36 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.[63] 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. The second phenomenon raises important questions about the effects of contractual incompleteness on the oil and gas exploration industry.

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.

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

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.

Chetverikov, Denis, Bradley Larsen, and Christopher Palmer, "IV Quantile Regression for Group-level Treatments, with an Application to the Effects of Trade on the Distribution of Wages," *Available at SSRN 2370140*, 2013.

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.

Meagher, Kieron J and Rodney W Strachan, "Gaussian Process: A Smooth and Flexible Approach to Estimating Index Complementarities in Organizational Economics," *Organizational Economics Proceedings*, 2014, *2* (1).

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.

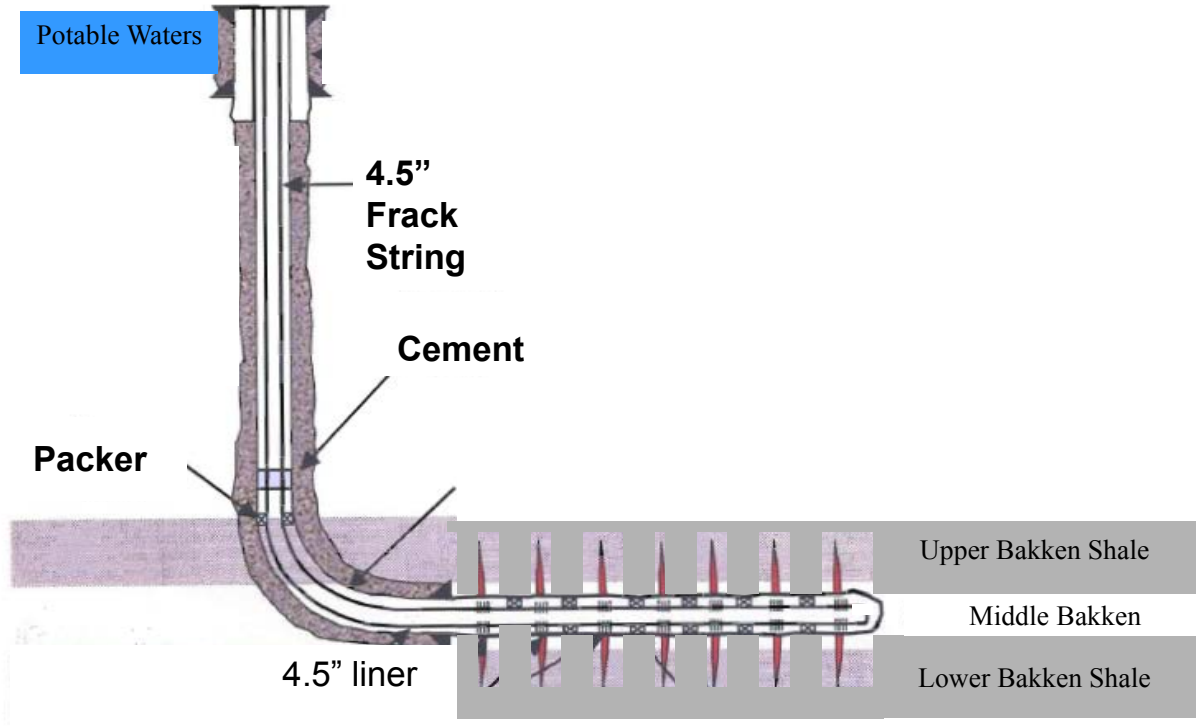
Vidic, RD, SL Brantley, JM Vandenbossche, D Yoxtheimer, and JD Abad, "Impact of Shale Gas Development on Regional Water Quality," *Science*, 2013, *340* (6134).

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.

Figure 1: Diagram of a Hydraulically Fractured Bakken Shale well



Adapted from Hicks (2012)

Figure 2: Diagram of a hypothetical spacing unit

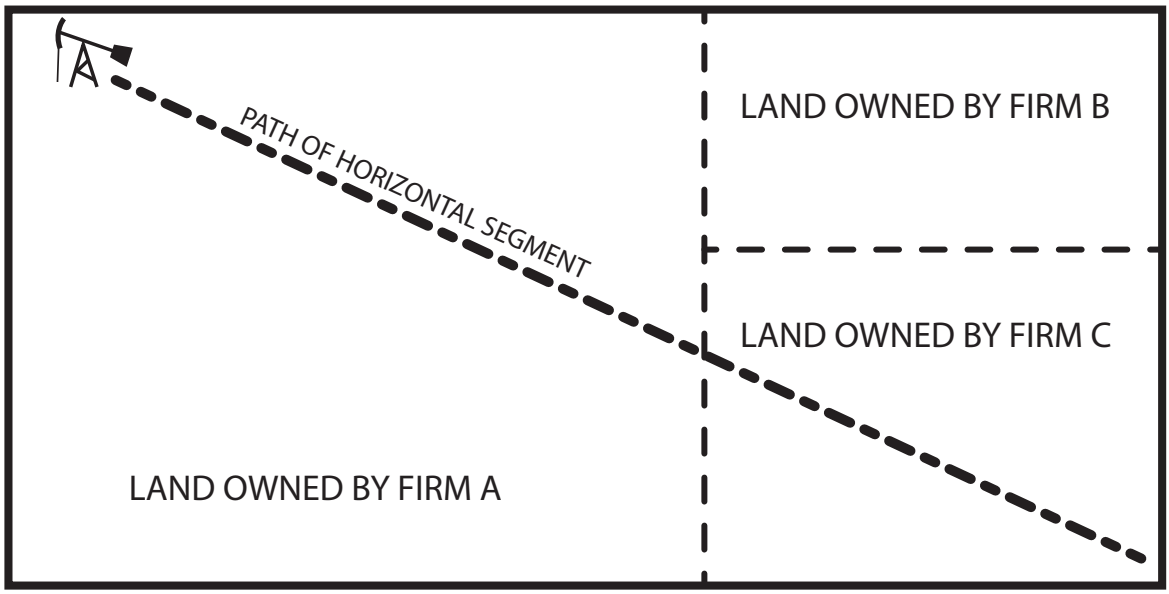
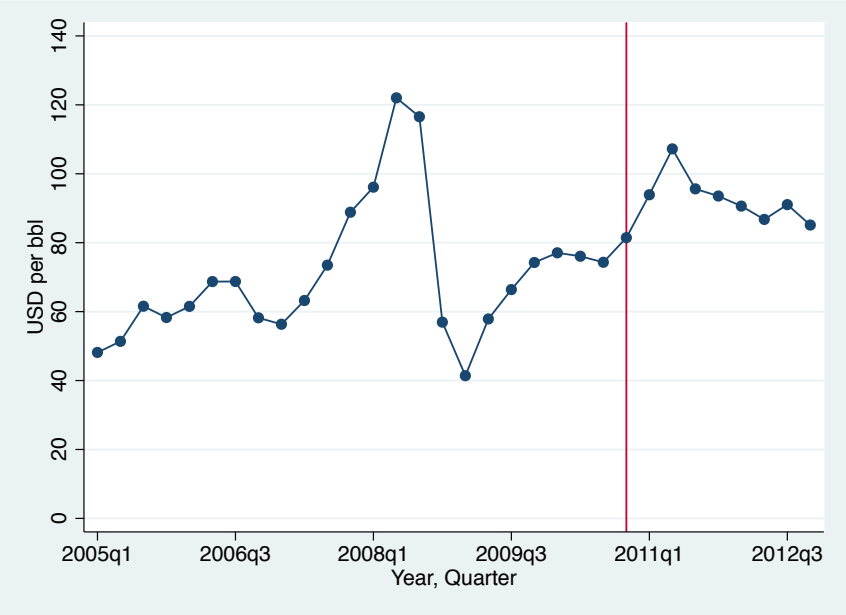
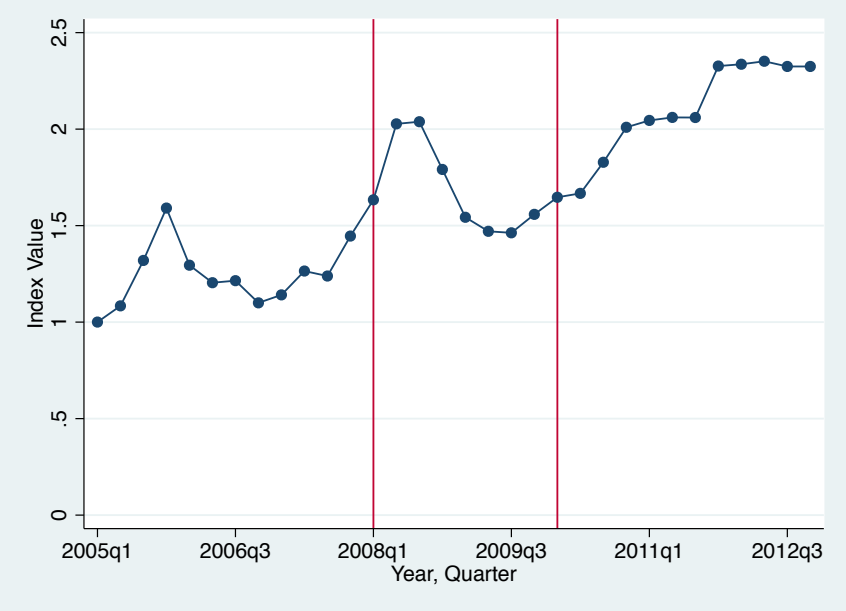


Figure 3: Quarterly Average Oil Prices for Bakken Producers



From January 2005 to October 2010, the figure shows the Cushing price minus a $1.75 Cushing-Clearbrook premium. After October 2010, the figure shows the Clearbrook price.

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 2012 it is calculated from the Spears & Associates data for horizontal wells in North Dakota.

Figure 5: Estimated Costs of Drilling and Fracking

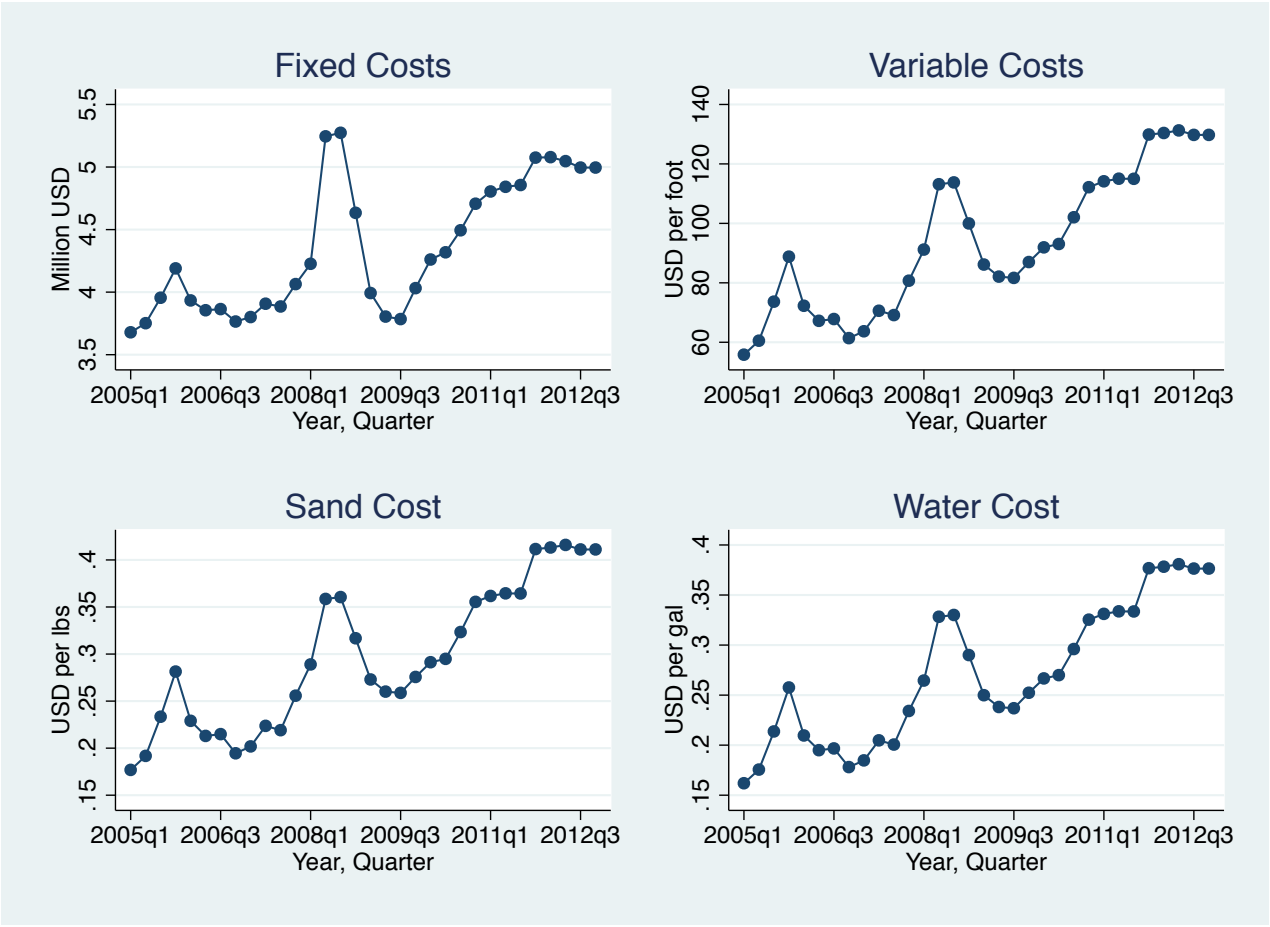
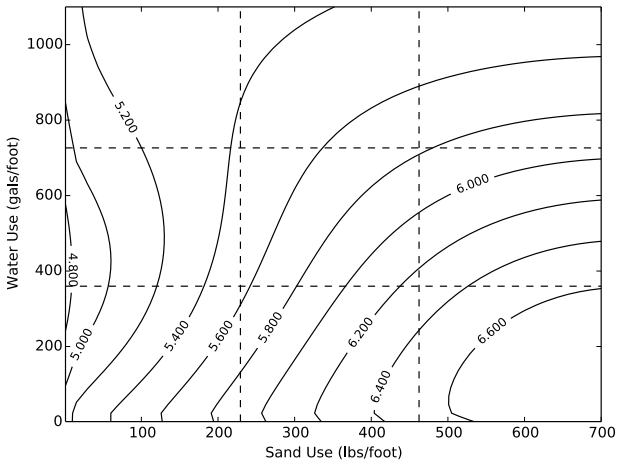
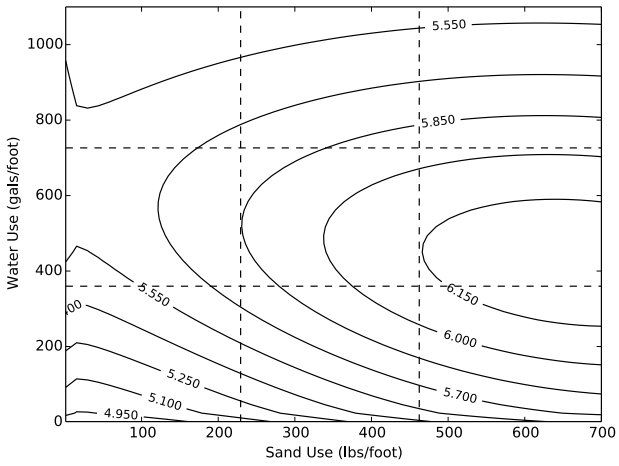


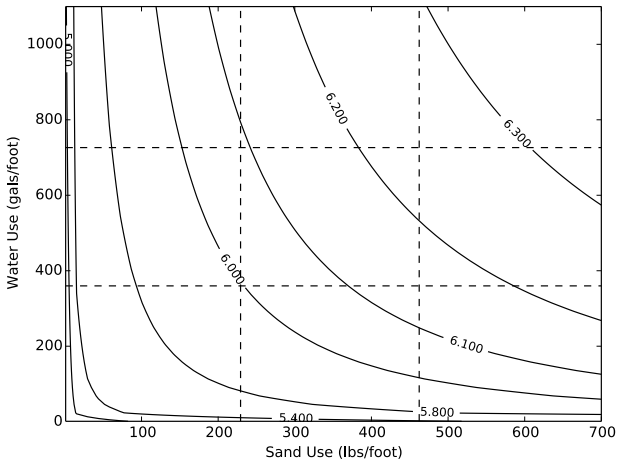
Figure 6: Contour Plots of Production Function Estimates



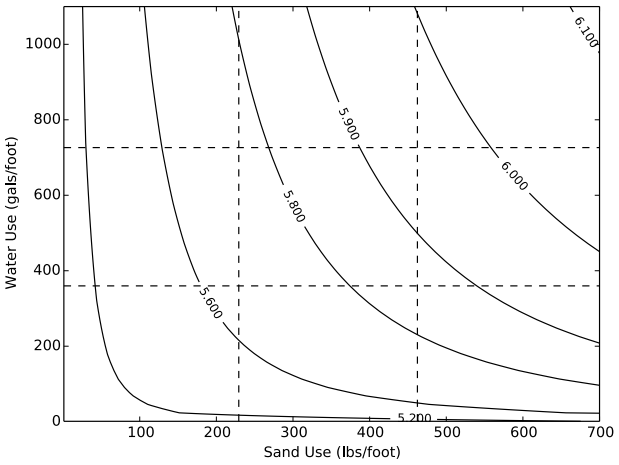
(a) Gaussian Process, Most Active Township



(b) Gaussian Process, Neighboring Township



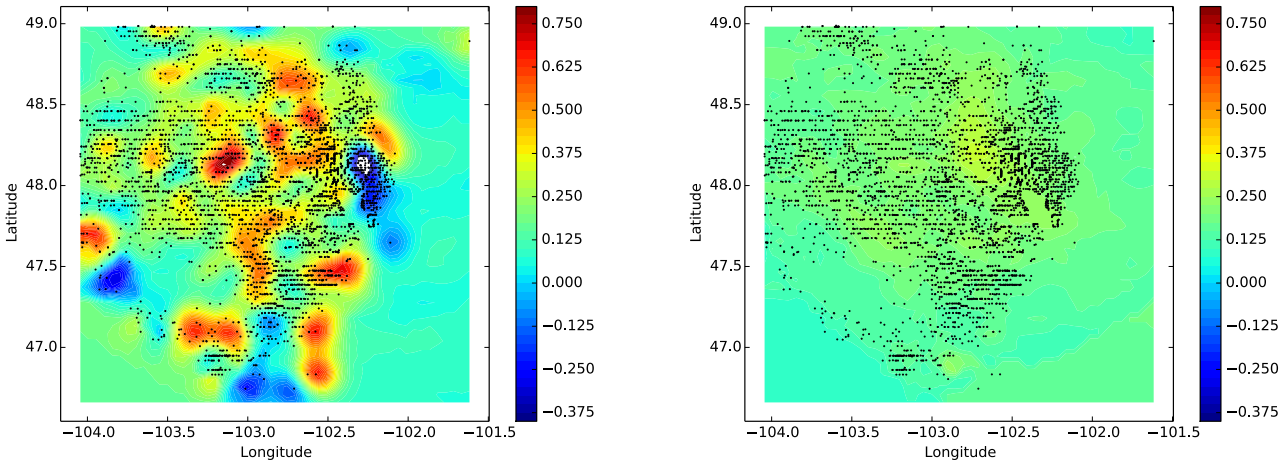
(c) Cobb-Douglas, Most Active Township



(d) Cobb-Douglas, Neighboring Township

Gaussian Process estimates are based off of column 2 in Table 5, while Cobb-Douglas estimates are based off of column 6. The most active township is 154-92, and its neighbor is 154-93. The units in all contour plots are log baseline production.

Figure 7: Estimated Production Differences Between High and Low Inputs

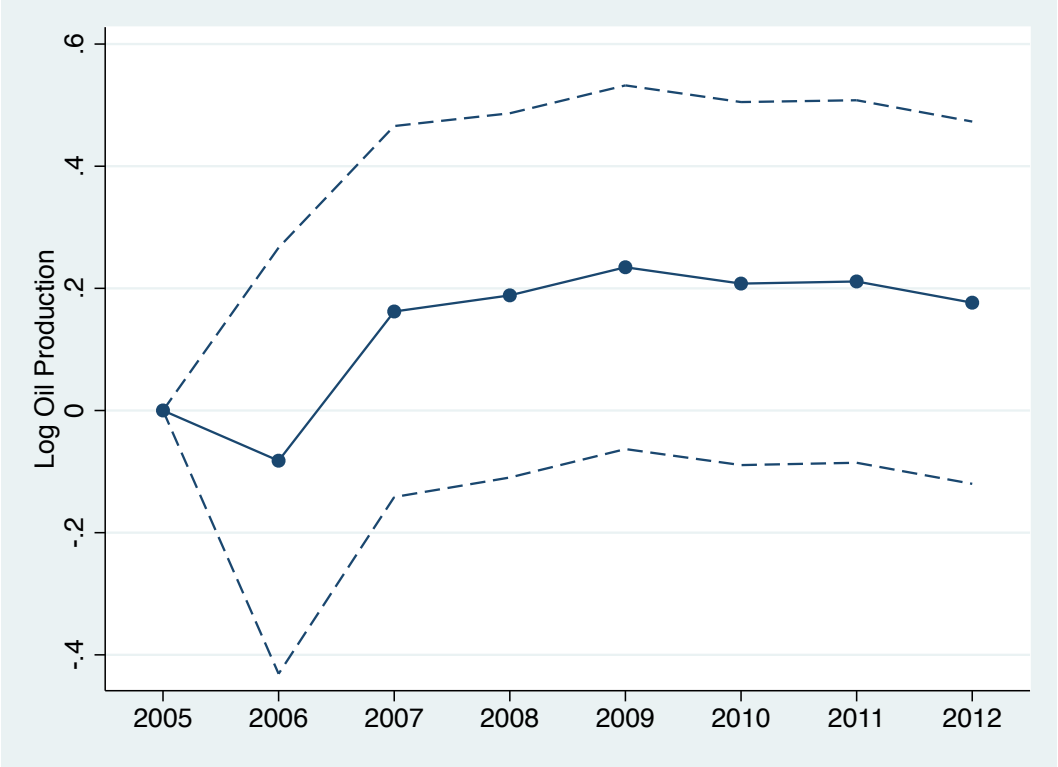


(a) High Minus Low, Gaussian Process

(b) High Minus Low, Cobb-Douglas

Gaussian Process estimates are based off of column 2 in Table 5, while Cobb-Douglas estimates are based off of column 6. The units in all contour plots are log baseline production.

Figure 8: Estimated Productivity Differences Across Years



Estimates and confidence intervals for year-cohort fixed effects in a regression of production function residuals onto year-cohort fixed effects.

Figure 9: Fraction of Positive Profits Captured and Maximal Profits by Year, ex post

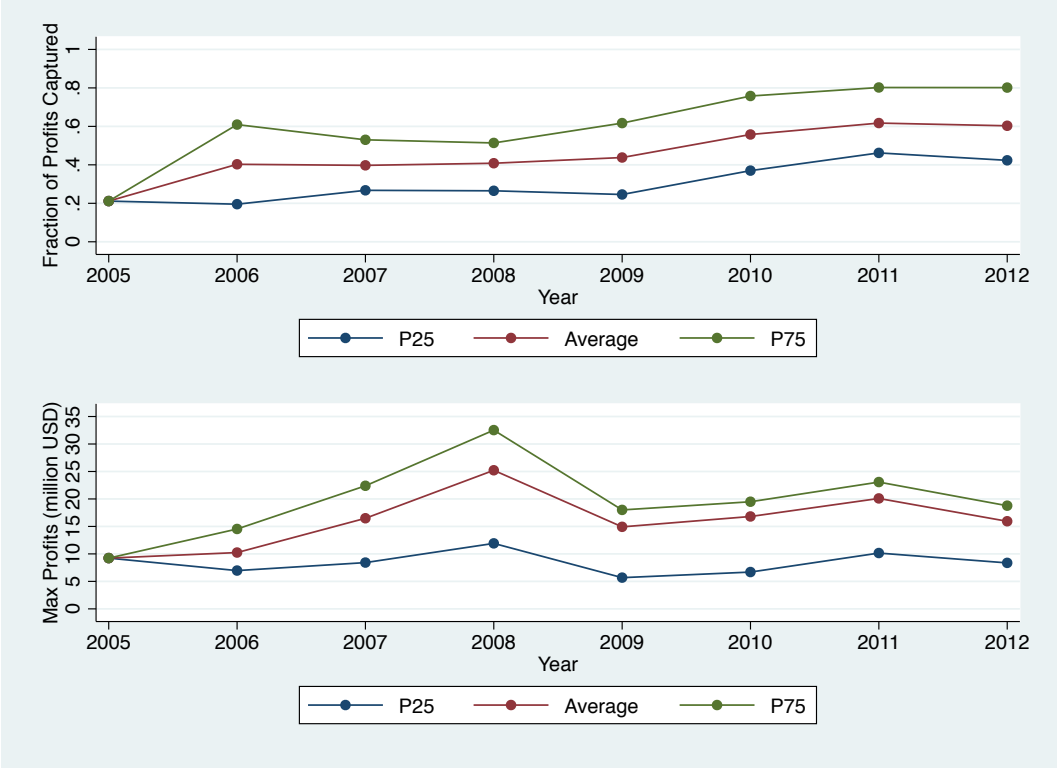


Figure 10: Average Profit Maximizing Input Use and Actual Input Use Per Well, ex post

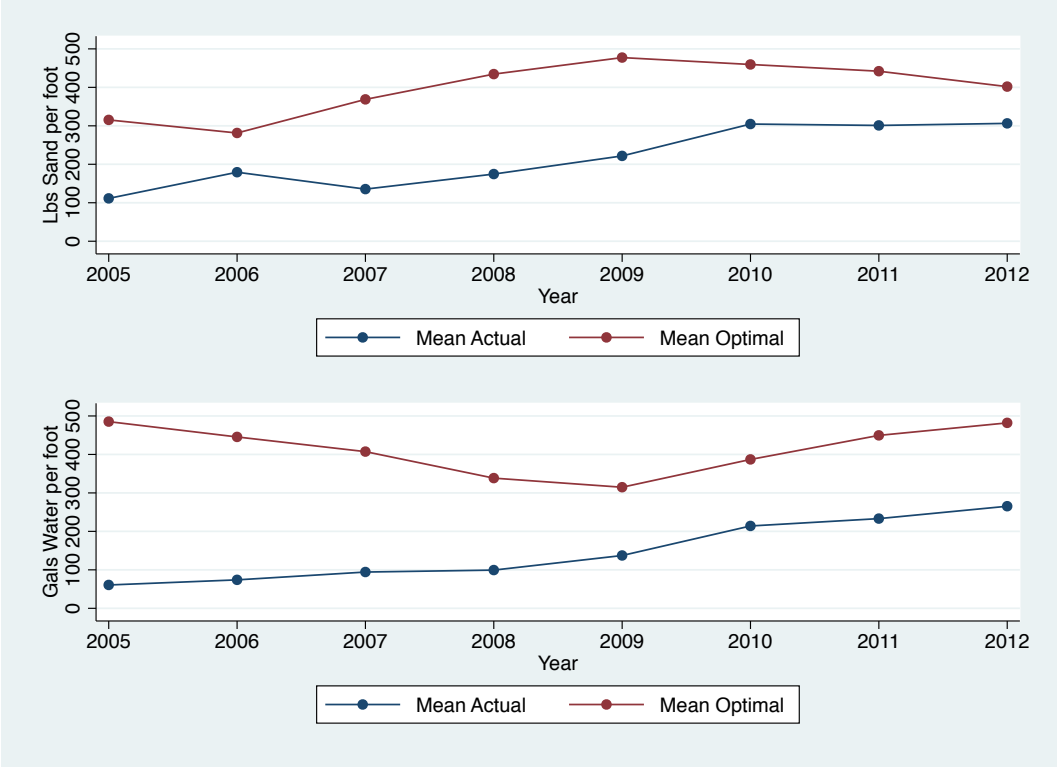


Figure 11: Fraction of Positive Profits Captured and Maximal Profits by Year, ex ante

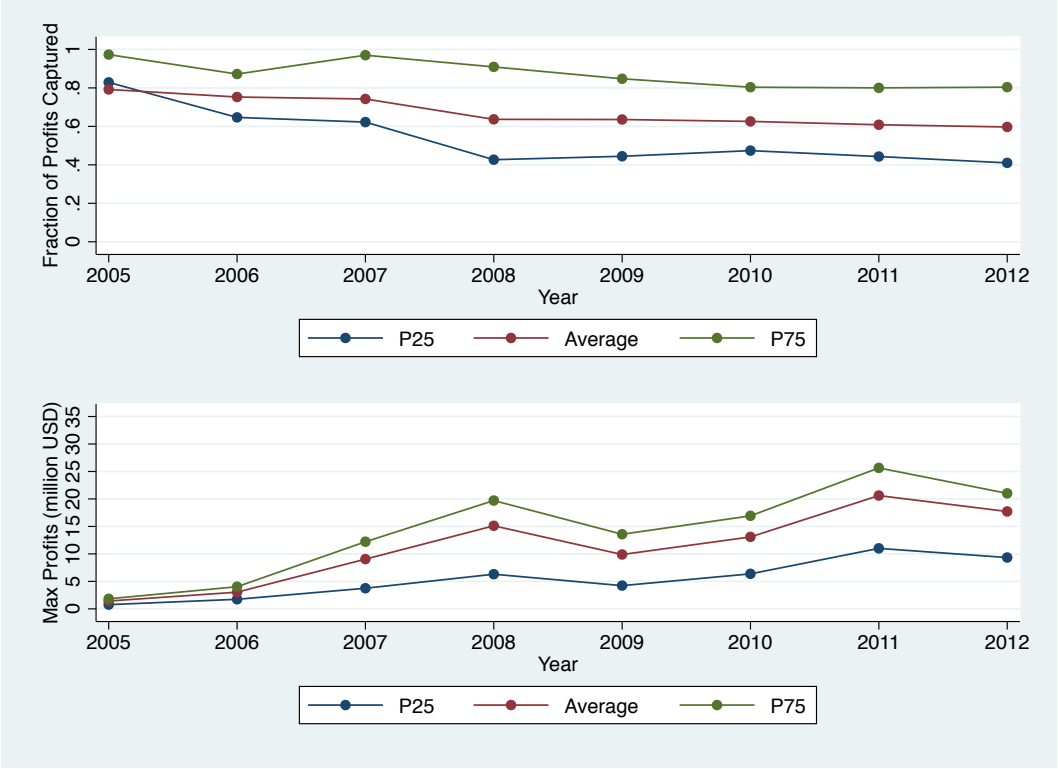


Figure 12: Average Profit Maximizing Input Use and Actual Input Use Per Well, ex ante

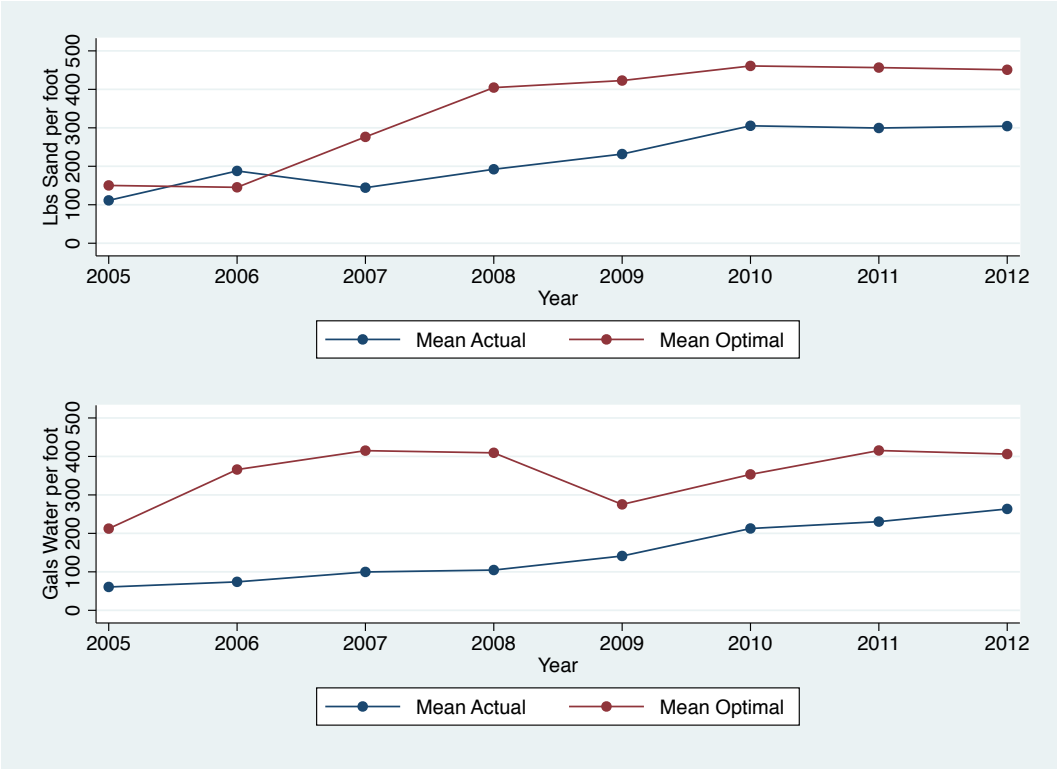


Table 1: Summary Statistics

Variable	Mean	Std. Dev	P25	P50	P75
Sand (lbs per foot)	276.71	127.75	190.99	277.39	374.07
Water (gals per foot)	214.24	137.77	122.04	201.23	271.20
Length (feet)	8,535.07	1,928.66	8,624.00	9,348.50	9,639.00
Oil (bbls per foot)	10.00	7.51	5.42	8.16	12.18
Average Days	25.89	3.41	24.75	26.75	28.08
# Non-operating Participants	4.23	2.66	2.00	4.00	6.00
Information Set: All Wells					
Own Wells	114.89	114.67	25.00	71.00	176.00
Participated Wells	243.41	220.44	67.00	184.50	364.00
Observed Wells	1,446.44	891.96	713.00	1,429.50	2,151.00
Information Set: Close Wells					
Own Wells	27.36	36.56	5.00	15.00	35.00
Participated Wells	25.04	28.62	5.00	15.00	34.00
Observed Wells	33.34	40.47	5.00	19.00	49.00
Estimated Geology Characteristics					
Total Organic Content	0.14	0.02	0.13	0.14	0.15
Thickness (feet)	43.74	13.50	35.50	43.50	52.50
Thermal Maturity	0.64	0.21	0.50	0.50	0.77

Summary statistics computed across all wells in the sample, $N = 4,408$.

Table 2: Summary Statistics by Year

		2005	2006	2007	2008	2009	2010	2011	2012
# Wells		14	18	100	365	444	704	1,107	1,656
# Active Firms		8	10	18	29	34	49	49	48
Sand	Average	104.5	125.4	127.1	177.3	210.4	306.8	298.8	301.0
	Std. Dev	15.7	137.1	130.8	140.8	143.2	120.1	110.4	107.7
Water	Average	50.3	64.3	86.5	106.9	137.8	214.7	230.6	258.0
	Std. Dev	19.2	61.5	47.7	57.6	91.2	98.7	115.2	163.5
Length	Average	6,088.8	6,457.4	7,176.7	7,346.8	7,421.0	8,074.3	8,886.9	9,181.6
	Std. Dev	1,573.0	2,077.9	1,983.9	2,161.3	2,295.1	2,133.9	1,705.4	1,358.6
Oil	Average	2.9	4.9	10.3	12.8	11.4	10.9	9.6	9.0
	Std. Dev	2.0	6.7	13.1	14.4	9.6	6.6	5.6	5.0
TOC	Average	0.14	0.14	0.13	0.14	0.14	0.14	0.14	0.14
Thickness	Average	47.00	46.56	40.64	43.19	46.36	44.62	43.34	43.17
Maturity	Average	0.74	0.70	0.66	0.63	0.65	0.65	0.64	0.64

Table 3: Relationship Between Oil Production and Quintiles of Sand and Water Use

		Quintiles of Water Use				
		First	Second	Third	Fourth	Fifth
Quintiles of Sand Use	First		0.80	1.64	2.86	4.32
			(0.45)	(0.64)	(0.75)	(1.12)
	Second	1.82	1.20	2.58	3.56	4.24
		(0.43)	(0.36)	(0.43)	(0.59)	(0.90)
	Third	3.67	2.29	3.10	3.50	4.44
		(0.56)	(0.42)	(0.41)	(0.42)	(0.59)
	Fourth	4.52	6.66	3.59	4.14	5.45
		(1.25)	(0.58)	(0.41)	(0.40)	(0.41)
	Fifth	4.65	11.10	7.26	6.06	6.94
		(1.78)	(0.74)	(0.52)	(0.45)	(0.36)

Regression of oil production in the first year divided by horizontal length onto sand and water use quintile bins, as well as township fixed effects. The first quintile bin is omitted, so coefficients represent relative changes in average oil production. Standard errors in parentheses.

Table 4: Wells Completed by the 10 Most Active Firms, by Location, Time and Completion Technique

Firm	North Dakota	Outside North Dakota			
	2005-2012	1995-2004		2005-2012	
	Bakken Shale	Conventional	Shale	Conventional	Shale
Continental Resources	490	332	30	410	568
EOG	426	5,118	182	4,456	3168
Whiting	416	122		2,031	56
Hess	316	757		347	35
Marathon	304	3,062	4	1,075	350
Brigham	241	170		101	23
Oasis	164			10	52
Burlington	160	4,216	28	2,531	805
XTO	155	2,643	39	6,496	3,044
Petro-Hunt	145	103	16	109	36
Rest of industry	1,591				

Table 5: Production Function Model Estimates

	(1) Gaussian Process		(2) Gaussian Process		(3) Cobb-Douglas		(4) Cobb-Douglas		(5) Cobb-Douglas		(6) Cobb-Douglas	
Coefficient	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
γ_0	-0.3923	0.0519	-0.4552	0.0558								
γ_S	6.1910	0.1073	6.2347	0.1294								
γ_W	6.0067	0.1175	6.0079	0.1268								
γ_{lat}	-2.3848	0.0458	-2.4227	0.0509								
γ_{lon}	-2.2974	0.0489	-2.3485	0.0560								
α	-2.5177	0.2361	-2.1914	0.8241	0.3877	0.2955	-3.9305	0.3585	1.7730	0.7489	-2.8046	0.6635
β	-0.5592	0.0012	-0.5592	0.0012	-0.5592	0.0012	-0.5593	0.0012	-0.5592	0.0012	-0.5594	0.0012
δ	1.1663	0.0025	1.1663	0.0025	1.1662	0.0025	1.1663	0.0025	1.1664	0.0025	1.1663	0.0025
η	0.8525	0.0251	0.8466	0.0251	0.4335	0.0311	0.7966	0.0257	0.4770	0.0308	0.8041	0.0257
ω_{TOC}			-0.9552	0.4460					-0.8794	0.3961	0.0062	0.3202
ω_{THICK}			-0.2531	0.8717					-0.5705	0.5853	-2.3859	0.5017
ω_{MATURE}			1.1767	0.3846					-0.5281	0.3774	-0.2376	0.2814
ω_S			-0.1951	0.1515	0.2087	0.0165	0.1457	0.0122	-0.3054	0.1632	-0.0528	0.1226
$\omega_{TOC,S}$			0.2093	0.0855					0.0933	0.0900	0.0197	0.0697
$\omega_{THICK,S}$			0.0180	0.1526					0.7271	0.1328	0.4945	0.1022
$\omega_{MATURE,S}$			-0.0886	0.0764					0.0790	0.0863	-0.0587	0.0634
ω_W			0.1718	0.1665	0.0437	0.0168	0.1334	0.0125	0.1856	0.1655	0.1941	0.1292
$\omega_{TOC,W}$			-0.0599	0.0914					0.0499	0.0891	-0.0483	0.0695
$\omega_{THICK,W}$			0.0517	0.1622					-0.4550	0.1394	-0.1389	0.1089
$\omega_{MATURE,W}$			-0.1075	0.0809					0.0136	0.0846	0.0946	0.0646
$\log \sigma_\epsilon$	-1.1807	0.0135	-1.1829	0.0138	-0.5601	0.0109	-1.0264	0.0111	-0.5940	0.0109	-1.0316	0.0111
$\log \sigma_\nu$	-0.7885	0.0016	-0.7885	0.0016	-0.7884	0.0016	-0.7885	0.0016	-0.7884	0.0016	-0.7885	0.0016
Township FE	No		No		No		Yes		No		Yes	
Overall R^2	0.7385		0.7389		0.5046		0.6900		0.5213		0.6915	
Between R^2	0.8093		0.8100		0.2668		0.7056		0.3130		0.7087	

Wells = 4,408, # Well-months = 193,846. "Between" R^2 is the R^2 for the average predicted log baseline production. The R^2 for the predicted time series of production is .6717 for all specifications. Maximum likelihood estimates of Cobb-Douglas production function models:

$$\log Y_{it} = \alpha + \beta \log t + \delta \log D_{it} + \eta \log H_i + m(Z_i, R_i \mid \omega) + \tau_i + \epsilon_i + \nu_{it}$$

and Gaussian process production function models:

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

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, R_i is the vector of organic content, thickness and maturity, 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.

Table 6: Estimates of Learning-by-Doing

	Levels (100s)		Logs	
Information Set	Estimate	Std. Err	Estimate	Std. Err
# Total Wells	-0.0008	0.0010	0.0039	0.0217
# Operated Wells	-0.0073	0.0043	-0.0052	0.0039
# Close Total Wells	-0.0222	0.0070	0.0041	0.0051
# Close Operated Wells	-0.0621	0.0143	-0.0092	0.0038

Estimates of ρ_1 in learning-by-doing production function models:

$$\widehat{\epsilon_i} = \rho_0 + \rho_1 \mathbf{E}_i + \sum_{q=2006}^{2012} \Upsilon_q \mathcal{I}(i \in q) + \kappa_i$$

where $\mathbf{E}$ is a measure of previous experience (either the number of wells in an information set or the log of 1 plus that number), Υ's are coefficients on dummy variables for the year in which a well is completed, and κ_i is an iid mean zero shock. The number of wells is 4,408.

Table 7: Uncertainty Preferences

Firm	# Wells	ξ_m Estimate	ξ_m Std. Err	$\xi_{s,0}$ Estimate	$\xi_{s,0}$ Std. Err	$\xi_{s,1}$ Estimate	$\xi_{s,1}$ Std. Err	λ Estimate	λ Std. Err
Continental	490	15.47	0.77	-30.83	1.50				
		15.62	0.82	-22.30	14.28	-3.21	5.35		
		15.42	0.77	-31.23	1.53			-0.21	0.09
		15.63	0.82	-19.64	14.35	-4.36	5.38	-0.21	0.09
EOG	426	6.42	0.34	-15.64	0.89				
		6.41	0.34	-14.43	4.72	-0.46	1.77		
		6.41	0.34	-16.03	0.97			-0.17	0.16
		6.41	0.34	-14.35	4.61	-0.65	1.74	-0.18	0.16
Whiting	416	10.97	0.62	-32.15	1.80				
		11.81	0.66	16.92	9.42	-18.65	3.57		
		11.16	0.63	-32.20	1.80			0.52	0.13
		11.91	0.67	19.66	10.02	-19.62	3.78	0.54	0.14
Hess	316	9.88	0.63	-21.81	1.34				
		8.09	0.66	-68.10	9.62	17.98	3.58		
		9.90	0.64	-21.81	1.34			0.06	0.17
		8.11	0.66	-69.95	9.93	18.68	3.69	0.19	0.17
Marathon	304	12.06	1.00	-34.52	2.33				
		8.11	1.04	-155.71	19.78	47.1	7.35		
		12.05	1.02	-35.95	2.48			-1.01	0.35
		8.57	1.04	-161.3	19.00	48.44	7.02	-1.36	0.47
Brigham	241	12.45	0.77	-23.78	1.57				
		16.13	1.01	81.45	13.53	-40.97	5.29		
		12.64	0.78	-25.60	1.69			-0.39	0.09
		16.84	1.05	86.75	13.65	-44.29	5.41	-0.50	0.09
Oasis	164	12.38	1.01	-19.71	1.66				
		12.66	1.05	3.67	19.87	-8.64	7.36		
		12.11	1.01	-19.70	1.67			-0.31	0.22
		12.45	1.05	11.89	19.67	-11.7	7.31	-0.38	0.22
Burlington	160	10.20	0.92	-22.41	1.91				
		8.53	0.99	-68.91	14.38	18.14	5.42		
		10.22	0.92	-23.10	1.97			-0.60	0.26
		8.70	0.95	-69.91	14.07	18.19	5.27	-0.64	0.27
XTO	155	14.49	1.29	-28.36	2.44				
		14.18	1.35	-46.27	26.57	6.72	9.89		
		14.54	1.29	-28.29	2.44			0.11	0.21
		14.24	1.36	-45.27	26.57	6.37	9.89	0.10	0.21
Petro-Hunt	145	17.14	1.51	-31.09	2.73				
		16.50	1.52	-67.56	20.98	13.66	7.68		
		17.78	1.57	-33.56	2.97			-0.74	0.19
		17.23	1.59	-64.22	21.95	11.51	8.07	-0.72	0.19
All	4,408	7.37	0.13	-17.72	0.30				
		7.20	0.13	-29.39	2.04	4.39	0.75		
		7.26	0.13	-17.82	0.30			-0.20	0.05
		7.12	0.13	-28.54	2.05	4.04	0.76	-0.16	0.05

Maximum likelihood estimates of the uncertainty preference model:

$$u_{ij} = \xi_m \left(\phi P_i \mathbb{E}\left[DOP_{ij} \mid \lambda\right] - c_i(S_j, W_j)\right) + \left(\xi_{s,0} + \xi_{s,1}\sqrt{\log \mid I \mid}\right) \phi P_i \left(\mathbb{V}\left[DOP_{ij} \mid \lambda\right]\right)^{\frac{1}{2}} + \epsilon_{ij}$$

P_i is the price of oil for well i, $\mathbb{E}\left[DOP_{ij}\right]$ 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, $\mid I \mid$ is the number of wells in the information set, λ is the increase (or decrease) in $\log \sigma_\epsilon$ for wells in I operated by the same firm, and ϵ_{ij} is an iid logit shock.

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

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

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:

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

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

Discounted oil production is the product of two random variables:

$$\mathrm{DOP}_{ij} = \underbrace{\exp\left(\alpha + \eta \log H_i + f\left(S_j, W_j, lat_i, lon_i\right) + \epsilon_i\right)}_{\text{Baseline Production}} \times \underbrace{\sum_{t=1}^{240} \rho^t M_{it} \exp\left(\beta \log t + \delta D_{it} + \nu_{it}\right)}_{\text{Decline, Maintenance and Discounting}}$$

Because ϵ and the ν's are all normal and jointly independent, it is easy to compute the moments of DOP_{ij}. The mean and variance of baseline production are given by the standard formula for normal random variables:

$$\begin{aligned}
\mu_{BP} &= \exp\left(\alpha + \eta \log H_i + \hat{g}\left(S_j, W_j, lat_i, lon_i\right) + \frac{1}{2}\left(\sigma_\epsilon^2 + \sigma_{g,ij}^2\right)\right) \\
\sigma_{BP}^2 &= \left(\exp\left(\sigma_\epsilon^2 + \sigma_{g,ij}^2\right) - 1\right) \exp\left(2\left(\alpha + \eta \log H_i + \hat{g}\left(S_j, W_j, lat_i, lon_i\right)\right) + \sigma_\epsilon^2 + \sigma_{g,ij}^2\right)
\end{aligned}$$

where $\hat{g}\left(S_j, W_j, lat_i, lon_i\right)$ is the posterior mean value of the Gaussian Process for the effect of fracking inputs and location on baseline oil production and $\sigma_{g,ij}^2$ is its posterior variance. The

mean and variance of the effects of decline, maintenance and discounting are:

$$\mu_{DMD} = \sum_{t=1}^{240} \mathbb{E}\left[M_{it}\right] \exp\left(t \log \rho + \beta \log t + \delta D_{it} + \frac{1}{2}\sigma_\nu^2\right)$$

$$\begin{aligned}\sigma_{DMD}^2 = &\sum_{t=1}^{240} \mathbb{E}\left[M_{it}\right] \left(\exp\left(\sigma_\nu^2\right) - 1\right) \exp\left(2\left(t \log \rho + \beta \log t + \delta D_{it}\right) + \sigma_\nu^2\right) \\ &+ \sum_{t=1}^{240} \mathbb{V}\left[M_{it}\right] \exp\left(2\left(t \log \rho + \beta \log t + \delta D_{it} + \frac{1}{2}\sigma_\nu^2\right)\right)\end{aligned}$$

under the assumption that the M_{it}'s and ν_{it}'s are jointly independent. Finally, the mean and variance of discounted oil production are:

$$\begin{aligned}\mathbb{E}\left[DOP_{ij}\right] &= \mu_{BP}\mu_{DMD} \\ \mathbb{V}\left[DOP_{ij}\right] &= \mu_{BP}^2\sigma_{DMD}^2 + \sigma_{BP}^2\mu_{DMD}^2 + \sigma_{BP}^2\sigma_{DMD}^2\end{aligned}$$

I compute $\mathbb{E}\left[DOP_{ij}\right]$ and $\mathbb{V}\left[DOP_{ij}\right]$ for a 10 by 10 grid of possible input choices j for all wells i. For both inputs, the grid is evenly spaced between the minimum and maximum values attained in the sample.