How Do Natural Gas Pipeline Networks Affect Emissions From Drilling and Flaring? *

Lauren Beatty †

November 14, 2022

Latest Version Here

Abstract

Most oil wells co-produce natural gas. Producers can choose to burn this valuable co-product on site (known as flaring) if the cost of connecting to the existing natural gas pipeline network is sufficiently high. While flaring is damaging to the climate, there exists surprisingly little research on the economics of flaring. I construct and estimate a dynamic model of producer drilling and flaring decisions which depend on the current state of the pipeline network and expectations over its evolution. My model also allows producers to internalize spillover effects for their neighbors – any pipeline they build will extend the network and weakly decrease their neighbors' future pipeline connection costs. Using my model estimates, I simulate pipeline development and flaring outcomes under counterfactual policies: a flaring tax, a flaring ban, and a gas subsidy. My counterfactual simulations show that flaring abatement costs are higher than previous studies but suggest that a flaring tax could substantially reduce flaring. A \$5/Mcf tax reduces flaring by 39%.

*I am deeply grateful for my advisor, Josh Linn, and committee members, Louis Preonas, James Archsmith, and Rob Williams for helpful comments, feedback and encouragment. This paper was supported by the Alfred P. Sloan Foundation Pre-doctoral Fellowship on Energy Economics, awarded through the NBER.

†Department of Agricultural and Resource Economics, University of Maryland, College Park.
E-mail: lbeatty1@terpmail.umd.edu Website: https://lbeatty1.github.io

1 Introduction

Networks are frequently studied in economics, but typically in static contexts where space is either unimportant or abstracted away from such as communication networks or two-sided markets. When space is abstracted away from networks can be modeled using graph theoretic techniques or even more simple descriptors such as numbers of network participants. For example, Rochet and Tirole (2003), in their classic models of two-sided markets, simply specify buyer and seller gross surplus as a function the of number of participants on each side. Typically network problems are characterized by network externalities. Consider the early development of phone or internet networks where users are positively affected by the entrance of a new user.[1]

However, there are many examples of network problems where space is crucial and where dynamics are important such as electricity grids, electric vehicle charging networks, and pipeline networks. These structures are still characterized by traditional network externalities, but in addition dynamics and space are important. The effect of a new entrant into the network is dependent upon where that entrant is sited, and the future states of the market are dependent upon current and past states. For example, consider the siting and construction of new renewable energy sources. Many of the best places to site renewable energy projects are located in very rural areas with relatively little transmission infrastructure and high costs of connecting to the existing grid. When a new generation source, like a wind or solar farm, is proposed the developer is often on the hook for expensive upgrades to transmission infrastructure that later entrants can use without sharing in the costs (Department of Energy, 2022). Clearly, there are network externalities – one producer's decision to upgrade the grid will affect future potential entrants, but in addition the decision to enter generates spatial and temporal dependencies. Moreover, in all of these examples network externalities interact with more traditional environmental externalities. The interaction of spatial and temporal dependencies have presented challenges in past work modelling these types of problems. This paper will focus on the growth of a natural gas pipeline network, its relationship with producer decisions to connect to it, and its implications for environmental outcomes.

In this paper, I set up and estimate a dynamic model of producer decisions to drill and connect

[1] However, network externalities need not be positive or even monotonic.

to natural gas pipelines in the Permian Basin in Texas, accounting for the fact that the pipeline network grows over time and that producers might consider spillover effects for their neighbor – that when a driller connects, the pipeline network expands, weakly reducing costs for nearby potential wellsites. Oil wells often co-produce natural gas, and producers have the option to burn the gas at the wellhead (flare) if the cost of connecting to pipelines is too high. Flaring is socially costly, so this paper will be primarily focused on investigating the relationship between the pipeline network and producer drilling and flaring decisions. I answer three main questions. First, how do producers choose to drill and flare and how is that decision affected by the pipeline network? Second, how might potential policies aimed at reducing flaring affect flaring and production decisions, and the growth of the pipeline network over time? Lastly, do producers internalize spillovers for their neighbors, and if not, how much does this matter for flaring? I find that distance to pipelines is an important driver of firm drilling and flaring decisions. I also show that flaring abatement costs are relatively high, but that a flaring tax near the social cost of flaring could substantially reduce the practice. Lastly, I estimate that firms do not consider spillovers for their neighbors – leading to less drilling and higher emissions from flaring than in a counterfactual world where these spillovers were internalized.

These questions are important because flaring likely has large social costs. The U.S. EIA estimates that in 2020 about 420 million Mcf of natural gas was flared (or vented) in the U.S (U.S. Energy Information Administration, 2021b). Natural gas primarily consists of methane (CH4) a very potent greenhouse gas, but when it is burned it is mostly converted to CO2. Flaring is less socially costly than simply letting the methane escape into the atmosphere (venting).[2] However, there is likely a wedge between the social cost of flaring gas at the wellhead and the social cost of selling it for end use precisely because flares are oftentimes inefficient or unlit, allowing methane

[2]Assuming a discount rate of 3% the social cost of carbon is around $50 per ton while the social cost of methane is $1500 per ton (Interagency Working Group on Social Cost of Greenhouse Gases, 2021). An Mcf of gas weighs around 50lbs. When burned it is converted into about 121 pounds of CO2. Therefore if 100% is converted to CO2 the social cost of burning an Mcf of gas is around $3. If instead, the same amount of gas is vented, this would result in whopping $37.5 in damages. Thus, it's easy to see why the social cost of flaring is heavily dependent upon the efficiency of the flares – which is still hotly debated. EDF (2022) find that 7% of flared gas is released as methane and a recent study by Plant et al. (2022) estimate that 9% is. In contrast, the EPA assumes this number is 2% (40 CFR 98.233, 2010).

to escape into the atmosphere. A study by the Environmental Defense Fund found that 7% of gas flared is actually released directly as methane. Their estimates imply that flares in the Permian Basin release 300,000 metric tons of uncombusted natural gas per year–resulting in around a half billion dollars in external damages just from methane from flares (EDF, 2022). Flaring is mostly unpriced in the U.S. In fact, most leases do not contain royalty clauses on flared gas, and most states do not tax flared gas. Moreover, pipelines can be quite expensive. The Interstate Gas Association of America (INGAA) reported that in 2016 an 8-inch gathering pipeline costs $\$264,856$ per mile (ICF, 2018). Thus, flaring can be an attractive option for producers – especially when gas prices are low, when the expected level of gas production from a well is low, or when the well is very far away from existing pipelines.

In my model, drillers hold leases with time-limited options to develop parcels of land by drilling a well. As more pipelines are built over time, connection costs for new wells decline. Producers have expectations over the evolution of prices and connection costs, and choose when to drill and whether to flare or connect to pipelines. The state space of my model is too large to compute a full solution method (through backward induction) so I leverage the insights of Hotz and Miller (1993) and Arcidiacono and Miller (2011) and use a two-step estimator with conditional choice probabilities (CCPs) to estimate the parameters of the model. Estimation of the model uses variation in expected revenues, which comes through changing prices and differences in lease-level geology, and variation in distance to pipelines to back out the implied fixed costs of drilling, the implied costs of building pipelines, and whether spillovers are internalized.

This paper makes a few important contributions to the literature. First, this paper is the first to explore the endogenous relationship between drilling and gas pipeline construction. Second, this paper is the first paper which explores how policies aimed at reducing flaring might affect the drilling margin. Third, this paper is the first which estimates pipeline costs directly instead of relying on reported costs. Combined, these contributions make this paper the best model of how potential policies to reduce flaring would affect the drilling and connecting decisions of firms.

The closest paper is perhaps Lade and Rudik (2020). This paper presents a static model of wellpad-level flaring decisions which rules out spillovers and assumes the drilling margin is

fixed. They demonstrate that a policy instituted in North Dakota mandating a firm-level flaring standard is far more costly than a tax which generates an equivalent reduction in flaring due to heterogeneous costs across firms. They do this by constructing firm-level marginal abatement curves using estimates of pipeline costs from an industry report. In contrast my model considers dynamics, allows for producers to internalize spillovers, and allows policy to affect the drilling margin. I also estimate pipeline costs directly rather than relying on industry data. I find much higher abatement costs and demonstrate that these estimates do a better job of rationalizing the observed flaring outcomes.

This paper also relates more generally to the economics literature on oil and gas development. It's common in this literature to treat the decision to drill as a dynamic optimization problem (see Paddock, Siegel, and Smith (1988), Kellogg (2014)). This paper is the first that borrows the use of CCPs to estimate the parameters of a drilling function, demonstrating that credible estimates of the parameters can be estimated with this computationally inexpensive method. Next, the context of this paper is quite similar to that of Covert and Kellogg (2017) who study the choice between shipping crude via rail (which is environmentally costly and has high variable costs) and via pipeline (which has high investment costs, negligible variable costs, and significantly lower environmental damages). They find that pricing the externalities of rail shipping leads to a 12-29% increase in pipeline capacity. In my context the "dirty", low fixed-cost option is flaring, while the clean high fixed-cost option is constructing a pipeline. I find that pricing the externalities of flaring leads to a negligible increase in pipeline investment but a sizable reduction in flaring.

My model estimates indicate that industry reported estimates of pipeline costs underestimate the true costs of pipeline building and that firms do not consider spillovers to their neighbors. I run counterfactual simulations to explore the importance of spillovers and evaluate policy options to reduce flaring. My counterfactual simulations suggest that the benefits of spillover internalization are relatively small. Thus a sensible explanation for the lack of spillover internalization is simply that the costs of contracting or negotiating might simply be higher than the benefits. My policy counterfactuals suggest that a tax on flaring close to its social cost would substantially reduce flaring. A \$5/Mcf tax reduces flaring by 39% while a 3\$/Mcf tax reduces flaring by 29%.

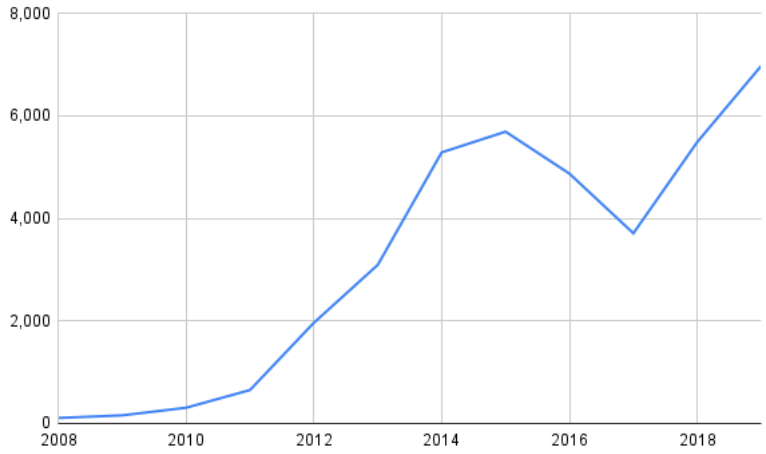


Figure 1: Number of Flaring Exceptions Issued by the Texas RRC by Year

2 Background

Oil and gas are often co-produced in wells. Oftentimes, the driller is primarily interested in the oil and gas is treated as a by-product. This happens when the oil-to-gas ratio at a potential well site is high, when gas prices are very low, or when the well is far away from existing infrastructure. When the costs of collecting the gas are high, it is a common practice for drillers to flare the gas. Texas allows producers to flare for up to 10 days after a new well completion, then requires a permit to continue flaring. The Texas Railroad Commission (RRC) states that to receive a permit, the operator "must provide documentation that progress has been made toward establishing the necessary infrastructure to produce gas rather than flare it" (Texas Railroad Commission). However, the number of flaring authorizations has ballooned over the past decade with the increase in fracking. Figure 1 plots the number of exemptions over time, which I've obtained via a Public Information Act request. In practice, many permits are issued because the gas is "uneconomical to collect." Moreover the Texas RRC has a reputation for "rubber stamping thousands of flaring permits without requiring oil companies to come up with a plan to curb the practice" (Chapa, 2021). In fact, between 05/02/2021 and 05/02/2022 401 applications for permanent exemptions have been approved and zero have been denied.[3]

This paper studies flaring caused by producers choosing to not build pipelines. However, other

[3]Beginning on 05/02/2021, the Texas RRC implemented online applications for flaring permits, and has made this data publicly available at https://webapps.rrc.state.tx.us/swr32/publicquery.xhtml.

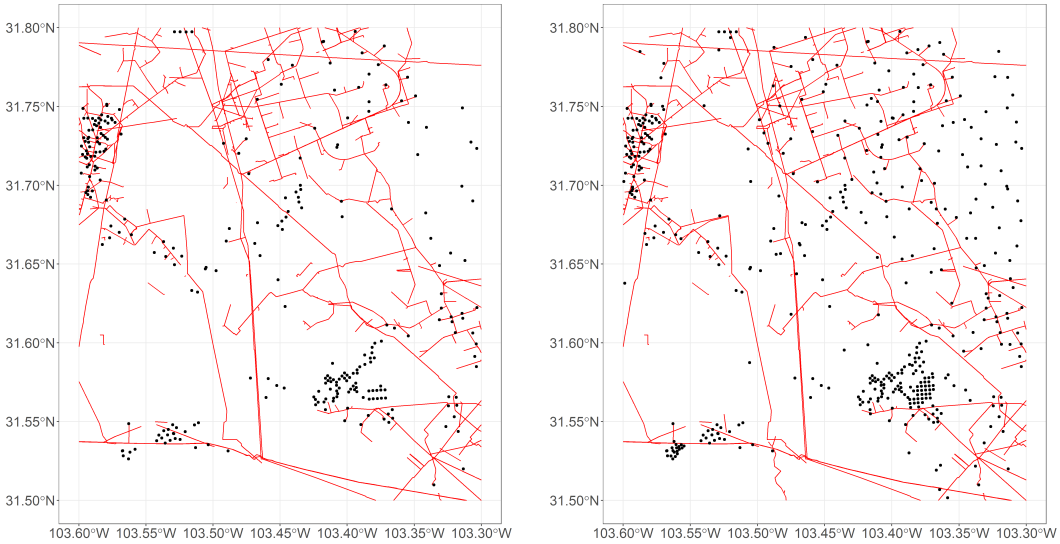
causes of flaring do exist. Agerton and Upton (2020) show that substantial portions of flaring come from wells that have previously sold gas–implying that the wells are indeed connected but must flare due to pipeline capacity or processing constraints. In this paper, I abstract away from processing and capacity constraints since I cannot directly observe processing and pipeline capacity utilization rates.[4]

Producers drilling in new fields often drill first, then wait to build infrastructure until they are sure the well will be productive, or until gathering and processing capacity becomes available in the area. This leads to any produced gas being flared during the interim. In figure 2 I show completed wells and permitted pipelines at four points in time. In the first two panels, there are many wells that are unconnected (and therefore flaring any associated gas), but will eventually go on to be connected by the last panel.

The pipeline networks which collect gas likely exhibit some network properties. Once a line is built, more connections can be added to that line at a later date (subject to capacity constraints). This means that the costs of connection at a given site likely decrease over time as new infrastructure creeps closer to that site. In figure 3 you can see that between 2005 and 2020, the pipeline network in the Permian Basin, especially the westernmost parts, fills out substantially. In theory, contracting would likely be able to internalize this network externality. However, little research has been done on whether this happens and to what degree.

Typically drillers will contract out with mid-stream gas gathering companies to have their wells connected. These gathering companies are one way that network externalities might be internalized. For example, a gathering company might be willing to charge below marginal cost to connect a well that would generate lots of spillovers (and hence contracts for the gathering company in the future). However, the industry is mixed between independent well operators and gatherers and larger producers which are vertically integrated and have the ability to both drill and build pipelines. In this paper, like many others in the literature, I abstract away from these

[4]Moreover, the analysis of Agerton and Upton (2020) seems to complement the main idea of this paper. They show that the proportion of gas from unconnected wells in three out of four basins decreases over the 2008-2020 period, consistent with the idea that the practice of flaring without connecting to a pipeline will be more prevalent earlier in a basin's productive life and decrease as the network get built out and as infrastructure becomes available.



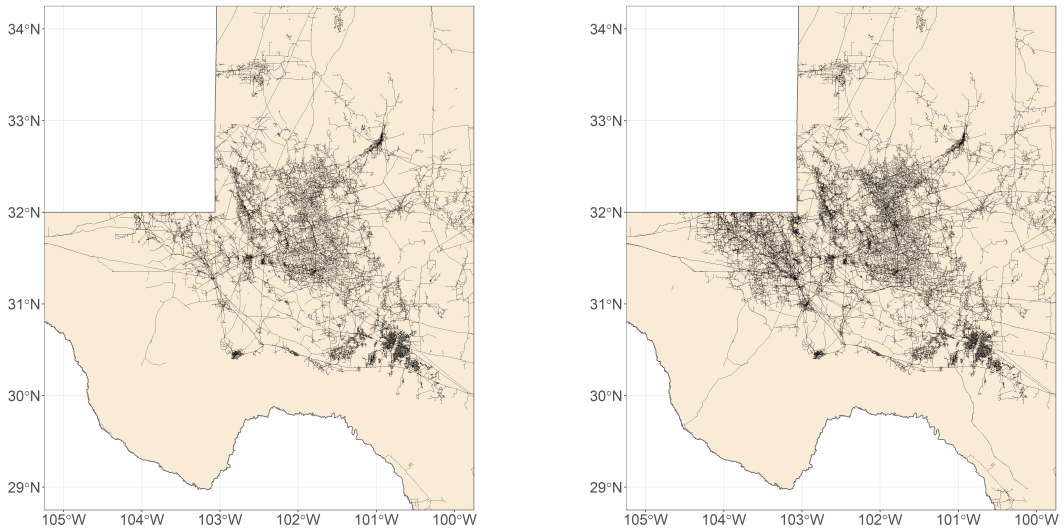
(a) 2005 Snapshot

(b) 2010 Snapshot

(c) 2015 Snapshot

(d) 2020 Snapshot

Figure 2: Snapshots of the development of a field from 2005-2020. Black dots are the locations of well surface holes, and red lines are natural gas pipelines permitted by that date.



(a) 2005 Pipeline Network

(b) 2020 Pipeline Network

Figure 3: 15 Years of Network Expansion

firm-type heterogeneities.

3 Data

3.1 Lease data

I use lease data collected by Enverus (formerly DrillingInfo). This lease data provides the spatial polygon of the lease as well as observables such as the grantor, grantee, effective date, expiration of primary term, and royalty rate. I process the raw lease data through two layers of clustering, described more in the appendix. The first issue in the lease data is that there are often multiple rows identifying what is clearly the same lease. This seems to be primarily a result of undivided mineral interests.[5] To deal with this, I follow many of the lease processing choices of Herrnstadt, Kellogg, and Lewis (2020). Namely, I use a hierarchical clustering algorithm on the observables of leases to collapse these into single rows.

Next, I identify lease amendments by also using a hierarchical clustering algorithm, this time, only on the spatial data. For example, there might be two rows of data corresponding with the same polygon. One has a primary term from 12-01-2005 through 12-01-2008, the other has a

[5]For example, multiple members of a family might each show up as a distinct grantor with her own row in the data.

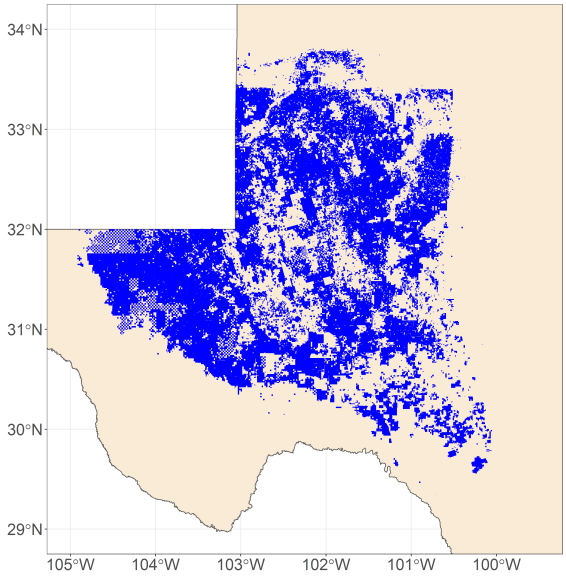


Figure 4: Leases in the Sample

primary term running from 12-01-2007 through 12-01-2010. I interpret this as a lease amendment whereby the primary term is extended mid-lease. More details on the processing of the lease data can be found in the Appendix.

In my analysis I only use leases from the Permian Basin.[6] The reason for this is primarily computational. Different basins have vastly different geological characteristics, and time-profiles of drilling. In my computational model, I would be unable to compute basin-time fixed effects, so I simply select the Permian Basin, the most active basin in Texas for my time sample (2010-2019). Figure 4 plots the leases in my estimation sample.

3.2 Production data

I use well-level monthly production data obtained from Enverus to predict expected production for leases (both drilled and never drilled). Enverus provides data on well-level monthly production of both oil and gas. It also contains information on the observables of the well such as the spud and completion dates, and the depth, lateral length, and location. To construct expected oil and gas production from lease i at t, which I call ξ^o_{it} and ξ^g_{it}, respectively, I calculate the inverse-distance weighted mean of expected production from all wells completed within the last 18 months of t

[6] I use U.S. EIA boundaries to define the Permian basin. These boundaries can be found at the EIA.

within 20km of i. Section 6.4 provides more detail on how I construct expected production.

Unfortunately, the Enverus data does not provide information on flared versus sold volumes at the well-level so I gather this from the Texas Railroad Commission (RRC). The RRC provides monthly production data at the lease-level (not the same lease definitions as Enverus), with gas volumes broken down into flared, sold, and vented volumes. I infer a connection date for each lease as the earliest date when gas is sold. I merge that connection date into the Enverus production dataset using an RRC lease-Well API crosswalk.

3.3 Pipeline data

I obtain geospatial pipeline data from the Texas Railroad Commission (RRC). I also collect data on when individual pipelines were permitted from 2010 through 2019, which allows me to construct the evolution of the gathering network over those 9 years. Unfortunately, these permitting dates are not necessarily construction dates and construction typically happens after permitting. Moreover, I am unable to observe expansions to a line under the same permit, so these data contain errors where the direction of bias is ambiguous.

3.4 Data processing and panel construction

I begin by turning the lease data into a quarter-lease panel dataset with a start date of 2010-01-01 and an end-date of 2019-12-01. I spatially join the lease data with the Enverus well data. A lease becomes "drilled" if I observe a well spudded between the expiration and effective dates and intersect the lease polygon. I call a lease "unconnected" if either zero gas is sold for the first year of the lease or if there are at least 200 days between the first production dates and first connection date (inferred from Texas RRC data) and over 70% of the RRC lease's first-year of produced gas is flared.

Before dropping any data, there are 64,948 leases and 617,846 lease-quarter observations. I drop 18,989 lease-quarter observations (3%) due to missing distance or expected productivity. Another 37,319 are dropped because of missing neighbor observables (which are needed to calculate the spillover effects).

In addition I am missing observations on whether a drilled lease becomes connected. There are 24,159 drilled leases and I cannot determine how much 851 of those flared. I drop all lease-quarter observations for those wells. I also randomly drop a proportionate amount of never-drilled leases since my estimation strategy hinges on being able to estimate the probability of drilling and connecting. I am left with 57,776 leases and 528,883 lease-quarter observations. Summary statistics for the leases are presented in Table 1.

Table 1: Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Primary Term Length	57776	38.775	15.252	-2	36	36	360
EffectiveYear	57776	2013.026	3.405	1995	2011	2016	2020
ExpirationofPrimaryTerm	57776	2016.288	3.398	2010	2014	2019	2045
Unconnected	23575	0.099	0.299	0	0	0	1
Expected Discounted Oil	57776	0.005	0.005	0	0.001	0.008	0.024
Expected Discounted Gas	57776	0.02	0.033	0	0.002	0.024	0.432
Start Distance	57776	1.478	3.522	0	0	1.345	34.239
End Distance	57776	1.402	3.508	0	0	1.127	34.239
Royalty	33514	0.232	0.032	0.01	0.2	0.25	3.1

Most leases have a primary term length of three years. Both the 25th percentile and 75th percentile lease have primary term lengths of 36 months. About 40% of leases become drilled and of those about 10% of leases that are drilled are unconnected. The average lease is about a kilometer and a half away from the nearest pipeline, though many leases have a pipeline that intersects with the lease polygon.

4 Descriptive Results

To motivate the setup of my model and confirm that the variation in the observables has the predicted effect on drilling and flaring outcomes I estimate some descriptive regressions. The first set of regressions explores how flaring outcomes conditional on drilling vary across leases. The second set of regressions explores how the lease observables affect the timing of drilling – regardless of flaring outcomes.

First, I divide the spatial extent of the Permian basin into 0.05 longitude $\times$ 0.05 latitude grids.[7]

[7]In the Permian, these grids are approximate 10 square miles – smaller than Covert and Sweeney (2019) use in their paper with a similar research design.

I use this grid to include grid-square fixed-effects in all regressions. The underlying logic is that geology should be similar within a given grid-square. This approach is similar to Covert and Sweeney (2019) who use a 10 mile by 10 mile grid to control for underlying geology to study revenues received by landowners under auctions versus informal negotiations.

Specifically, the equations I estimate are:

$$y_{it} = \beta_1 d_{it} + \beta_2 d_{it}^2 + \beta_3 \xi_{it}^o + \beta_4 \xi_{it}^g + \beta_5 p_t^o + \beta_6 p_t^g + \beta_7 N_{it}^{\text{benefit}} + \gamma_{g(i)} + \epsilon_{it} \tag{1}$$

where y_{it} is the outcome of interest, d_{it} is the distance from lease i at time t to the nearest pipeline, ξ^g, and ξ^o are expected discounted gas and oil production, p_t are oil and gas spot prices during quarter t, N^{benefit} measures how many leases would have their distance reduced by i connecting, and $\gamma_{g(i)}$ is a grid fixed-effect. The outcomes of interest are quantity flared, quantity of gas sold, the percentage of gas flared, and a dummy for whether the lease is unconnected during its first year. I expect distance to be correlated with increased flaring, high natural gas prices to be correlated with decreased flaring, and high oil prices to be correlated with increased flaring. The coefficients on N^{benefit} will be negative if leases which could generate higher spillovers for their neighbors are more likely to connect.

Results from these regressions are presented in table 2. All dependent variables are standardized to have a mean of zero and a standard deviation of one. The first three columns present OLS results with standardized dependent variables, the last column presents the coefficients of a logistic regression with the binary dependent variable, 1(Unconnected). The last column is the primary column of interest since it maps to the outcomes in the full dynamic discrete choice model. Focusing on column (4), the effect of distance is quite large. A one standard deviation increase in distance increases the odds of a well being unconnected by 68% ($\exp(0.576 - 0.057)$). More productive wells tend to be far less likely to remain unconnected. However, looking at columns (1), (2) and (3) suggest that wells with higher expected oil productivity tend to both flare more and sell more gas, while wells with higher expected gas productivity do not flare more gas than their less productive counterparts. The effect of prices on all of the dependent variables appears relatively small, though the coefficients have the expected sign and many are significant. Higher oil prices are associated

with higher percentages of gas flared, while high gas prices are associated with lower percentages of gas flared. Lastly, none of these regressions present any evidence that leases are more likely to connect if they have neighbors that might benefit from that choice.

Table 2: Effect of Observables on Flaring Outcomes Conditional on Drilling

	Percent Flared (1)	Q Flared (2)	Q Sold (3)	Unconnected (4)
Distance	0.238***	0.006	-0.013	0.576*
	(0.073)	(0.048)	(0.054)	(0.350)
Distance^2	-0.025*	-0.003	0.008*	-0.057
	(0.014)	(0.008)	(0.004)	(0.050)
Expected Oil Productivity (ξ^o)	0.005	0.141***	0.116***	-1.24***
	(0.023)	(0.021)	(0.026)	(0.245)
Expected Gas Productivity (ξ^g)	-0.096***	-0.019	0.234***	-0.452**
	(0.035)	(0.034)	(0.043)	(0.205)
Oil Price (p_t^o)	0.056***	0.002	-0.051***	0.082
	(0.015)	(0.015)	(0.012)	(0.107)
Gas Price (p_t^g)	-0.019*	-0.007	0.011	-0.005
	(0.011)	(0.014)	(0.010)	(0.083
N Benefiting Leases	0.008	-0.007	-0.013	0.121
	(0.014)	(0.011)	(0.019)	(0.153)
Family	OLS	OLS	OLS	Logit
Observations	18,846	19,737	19,737	5,135
R^2	0.627	0.503	0.698	
Within R^2	0.015	0.012	0.088	
Squared Correlation				0.288
Pseudo R^2				0.268
BIC				8,562.2
Grid fixed effects	✓	✓	✓	✓

Notes: Each column displays estimates of equation 1. Standard-errors are clustered at the grid-level. All independent variables are standardized.

Next I show how the observables are associated with drilling probabilities using a similar grid fixed-effect strategy. I run a logit of:

$$\mathbb{1}(i \text{ drilled during } t) = \beta_1 d_{it} + \beta_2 d_{it}^2 + \beta_3 \xi_{it}^o + \beta_4 \xi_{it}^g + \beta_5 p_t^o + \beta_6 p_t^g + \beta_7 N_{it}^{\text{benefit}} + \gamma_{g(i)} + \epsilon_{it} \quad (2)$$

Results are presented table 3. Again, distance has a strong negative effect on the likelihood of drilling while expected oil production has a strong positive effect. Here, the characterization of drillers being primarily interested in oil seems true. Higher expected oil productivity and higher oil prices are both strongly associated with increases in the propensity to drill, while neither expected

gas productivity or gas prices have significant coefficients. A one standard deviation increase in ξ^o is associated with a 127% increase in the likelihood of drilling while a one standard deviation increase in the price of oil is associated with a 38% increase in the likelihood of drilling.

The coefficient on the number of benefitting leases is negative. If firms internalized benefits for their neighbors, we would expect that coefficient be positive. However, N benefitting leases is likely correlated with a lease's number of undrilled adjacent leases and firms might prefer to wait to see if adjacent tracts are productive before drilling or might be interested in the productive inputs used by neighboring firms.[8] These issues have been previously explored by Décaire, Gilje, and Taillard (2019), Covert (2015), and Covert and Sweeney (2022).

Table 3: Fixed-Effect Logit of Observables on Drilling Outcomes

	Drill	
	(1)	(2)
Distance	-0.617***	-0.763***
	(0.187)	(0.134)
Distance^2	-0.001	0.025*
	(0.022)	(0.014)
Expected Oil Productivity (ξ_o)	0.417	0.819***
	(0.394)	(0.104)
Expected Gas Productivity (ξ_g)	-0.216	-0.093
	(0.510)	(0.085)
N Benefiting Leases	-0.433***	-0.199***
	(0.068)	(0.065)
Gas Price (p^g)		0.039
		(0.045)
Oil Price (p^o)		0.325***
		(0.059)
Standard-Errors	Grid-date	Grid
Observations	30,087	200,341
Squared Correlation	0.315	0.091
Pseudo R^2	0.338	0.212
BIC	44,158.5	56,027.4
Grid-date fixed effects	✓	
monthstilexp fixed effects	✓	✓
Grantee fixed effects	✓	✓
Grid fixed effects		✓

Note: Each column displays estimates of equation 2. Standard errors are clustered at the grid level. All independent variables are standardized.

[8] When a well is drilled, the depth is usually observable by neighbors. The inputs in the fracking job are also required to be disclosed on fracfocus.org, so firms can learn about the best inputs for production by learning from their neighbors.

5 A Dynamic Model of Drilling and Flaring

I formulate the drillers' problem as a discrete-time optimal stopping problem. In each quarter up until and including the time of the leases expiration, the lessee observes a vector of the state variables, $\vec{x}_{it}$ and makes the decision to either wait, drill and flare, or drill and gather. I denote this set $j \in \{W, F, G\}$. Once the lessee decides to drill, either with $j = G, F$, that period becomes the terminal period. The operator then receives the expected sum of discounted revenue from production and pays the drilling and connecting costs.[9] This means the drillers' problem is a finite-horizon optimal stopping problem with two potential actions for stopping, and can be formulated as a dynamic programming problem with a Bellman equation.

Thus, the firm's problem is

$$V(\vec{x}_{it}, \epsilon_{it}) \equiv \max_j E\left[\sum_{t=1}^{T} \beta^{t-1}(\pi(j, \vec{x}_{it}) + \epsilon_{itj})\right] \tag{3}$$

where T is expiration quarter of the lease, $\pi(j, \vec{x}_{it})$ is the per-period profit of taking action j, with observable lease state variables, $\vec{x}_{it}$. ϵ_{itj} is an unobserved shock assumed to be i.i.d. Type-I Extreme Value with scale parameter σ. Unlike discrete choice estimation of consumer preferences where the outcome is scale-less indirect utility, σ is identified since expected revenues are being measured directly in dollars. The set of state variables, $\vec{x}_{it}$, include i's distance to the nearest gathering line at t, d_{it}, the price of oil and gas, p_t^o, p_t^g, the expected sum of discounted production, ξ_{it}^o, ξ_{it}^g, the effect of building a pipeline on i's neighbors $\Delta\tilde{V}_{it}$, the time t (and hence the time until i's expiration date), and the royalty rate of the lease, r_i.[10]

[9] An underlying assumption here is that each lease can only be drilled once. This is clearly not true - many parcels could contain multiple wellsites. However, this is a simplifying assumption also made by Herrnstadt, Kellogg, and Lewis (2020).

[10] I only consider pipeline construction spillovers. I do not, for example, think about common pool problems that can arise in some natural resource extraction problems (see **?** for an exploration of common pool problems and lease option exercise). The common pool problem is not a major issue in shale formations, since hydrocarbons are only released after fracturing the rocks under high pressures, causing little or no impact on neighboring parcels (Décaire, Gilje, and Taillard, 2019).

The firm's maximization problem can be expressed in Bellman form as:

$$V(\vec{x}_{it}, \epsilon_{it}) = \max_j \{\pi(j, \vec{x}_{it}) + \epsilon_{itj} + E[V(\vec{x}_{it+1}|\vec{x}_{it})]\} \tag{4}$$

The deterministic part of per period payouts are given by:

$$\pi_i(j, \xi^o_{it}, \xi^g_{it}, p_t, d_{it}, \Delta\tilde{V}) = \begin{cases} R^o_i(p_t, \xi^o_{it}) - C^F_t & j = F \\ R^o_i(p_t, \xi^o_{it}) + R^g_i(p_t, \xi^g_{it}) - C^G_t(d_{it}, \Delta\tilde{V}) & j = G \\ 0 & j = W \end{cases}$$

where R^o and R^g are the expected revenue from oil an gas, respectively, and C^F, C^G are the the costs of drilling and flaring and drilling and gathering. Revenues are calculated as expected discounted production (ξ) multiplied by the spot price less royalties and state severance taxes (4.6% for oil and 7.5% for gas). Royalties often depend upon the lessor and for private land, are often negotiated along with the up-front price of the lease.

$$R^o_i(p_t, \xi^o_{it}) = \xi^o_{it} p^o_t (1 - r_i - 0.046)$$

$$R^g_i(p_t, \xi^g_{it}) = \xi^g_{it} p^g_t (1 - r_i - 0.075)$$

I assume costs can be decomposed into:

$$C^F_t = \delta^F + \delta_{y(t)}$$

$$C^G_t = \delta^G + \delta_{y(t)} + \alpha_d d_{it} - \alpha_{\tilde{V}} \sum_k \Delta\tilde{V}_{ik}$$

δ^F and δ^G are the fixed costs associated with flaring and gathering, respectively. $\delta_{y(t)}$ are year fixed effects that allow the price of drilling to move over time. $\delta_{y(t)}$ capture changes in the rental rate of drilling rigs as well as changes over time in average well depth or complexity. $\Delta\tilde{V}_{ik}$ is a measurement (in the same units $\tilde{V}$ is denoted in) of how much building a pipeline to i increases the value of a neighboring lease, k, and will be discussed in greater detail in the next section. The logic

behind this cost function is that pipeline connection costs will be a function of overall distance, but that there might be a discount to the drilling firm if the pipeline is expected to be valuable for the gathering firm in the future. $\alpha_{\tilde{V}}$ will measure the proportion of spillover benefits internalized by firm i.

Define the ex-ante value function as the expected value of V_{it} before the revelation of ϵ_{it}:

$$\bar{V}_{it}(\vec{x}_{it}) \equiv \int V_{it}(\vec{x}_{it}, \epsilon) f(\epsilon) d\epsilon \tag{5}$$

where $f(\cdot)$ is the distribution of the error term, ϵ.

Finally, let $\Omega(\cdot)$ be the transition probabilities of the state variables and define the conditional value function as the value of making a choice j in period t (less the error term), then behaving optimally from $t+1$ onward:

$$v_{it}(j, \vec{x}_{it}) \equiv \pi_i(j, \vec{x}_{it}) + \beta \int \bar{V}_{it+1}(\vec{x}_{it+1}) \Omega(\vec{x}_{it+1} | \vec{x}_{it}) d\vec{x}_{it+1} \tag{6}$$

Since the error terms are assumed to be type-I Extreme Value, the choice probabilities are given by the familiar logit form:

$$Pr(i \text{ chooses } j | \vec{x}_{it}) = \frac{\exp(v_{it}(j, \vec{x}_{it})/\sigma)}{\sum_k \exp(v_{it}(k, \vec{x}_{it})/\sigma)} \tag{7}$$

6 Estimation

6.1 Dynamic Discrete Choice Estimation

Because of the assumption on the error term, ϵ, Hotz and Miller (1993) and Arcidiacono and Miller (2011) show that the ex-ante value function admits the closed form:

$$\bar{V}_{it}(\vec{x}_{it}) = -\sigma \ln(Pr(j = \tilde{j} | \vec{x}_{it})) + v_{it}(\tilde{j}, \vec{x}_{it}) + \sigma\gamma \qquad \forall \tilde{j} \tag{8}$$

where γ is Euler's constant. The intuition here is that the ex-ante value function can be found by evaluating the conditional value function, v, at any $\tilde{j}$, then applying a correction which is larger when the probability of choosing $\tilde{j}$ is smaller. The constant term at the end, $\sigma\gamma$ is simply the expected value of ϵ. I omit this term for the remainder of the paper since choice probabilities are invariant to an additive constant in each v_j.

Since $\bar{V}(\vec{x}_{it})$ can be evaluating using any choice $\tilde{j}$, I can select a terminating action $\tilde{j} = G$, then

$$\bar{V}_{it}(\vec{x}_{it}) = -\sigma \ln(Pr(j = G)|\vec{x}_{it}) + \pi(G, \vec{x}_{it}) \tag{9}$$

Equation 9 can be substituted into the conditional value function, equation 6 to yield:

$$v_{it}(j, \vec{x}_{it}) = \pi(j, \vec{x}_{it}) + \beta \int [-\sigma \ln(Pr(j_{t+1} = G|\vec{x}_{it+1})) + \pi(G, \vec{x}_{it+1})]\Omega(\vec{x}_{it+1}|\vec{x}_{it})d\vec{x}_{it+1} \tag{10}$$

All of the pieces of equation 10 can be constructed from the data. $\pi(\cdot)$ is simply a linear function of the parameters and observables. $Pr(j_{t+1=G}|\vec{x}_{it+1})$ and $\Omega(\vec{x}_{it+1}|\vec{x}_{it})$ can be directly estimated from the data. Since π is linear in the stochastic state variables, it can be brought outside the integral and evaluated at $E[\vec{x}_{it+1}|\vec{x}_{it}]$. To construct $\int -\sigma \ln(Pr(j_{t+1} = G|\vec{x}_{it+1})\Omega(\vec{x}_{it+1}|\vec{x}_{it})d\vec{x}_{it+1}$, I estimate the transition function for $\vec{x}_{it}$, $\Omega(\vec{x}_{it}|\vec{x}_{it+1})$, then take simulation draws of $\vec{x}_{it+1}$, calculate the log probabilities associated with the simulation draws, then average.

For each j, you can express the v as:

$$v(j, \vec{x}_{it}) = \begin{cases} \xi^o_{it} p^o_t - (\delta^F + \delta_t) & j = F \\ \xi^o_{it} p^p_t + \xi^g_{it} p^g_t - (\delta^G + \delta_t + \alpha_d d_{it} - \alpha_V \sum_k \Delta\tilde{V}_{ik}) & j = G \\ \beta[\xi^o_{it} E[p^o_{t+1}] + \xi^g_{it} E[p^g_{t+1}] - (\delta^G + E[\delta_{t+1}] + \alpha_d E[d_{it+1} - \alpha_V \sum_k \Delta\tilde{V}_{ik}) + & j = W \\ \sigma \int -\ln(Pr(j_{t+1} = G|\vec{x}_{it+1}))\Omega(\vec{x}_{it+1}|\vec{x}_{it})d\vec{x}_{it+1} & \end{cases} \tag{11}$$

The parameters of to be estimated are $\alpha_d, \alpha_{\tilde{V}}, \delta_C, \delta_G, \delta_t$, and σ. In theory, β is identified also, but its common practice in the dynamic discrete choice literature to simply pick a reasonable value for β. I select 0.98, implying a yearly rate of about 0.92. δ_t and $E[\delta_t]$ are not all identified without

some restrictions. The same issue appears in Gowrisankaran et al. (2012) and Murphy (2018). Both papers fit δ_{t+1} and δ_t to an AR(1), though their exact estimation process varies. I choose to estimate

$$\delta_{t+1} = \omega_0 + \omega_1 \delta_t + \eta_t$$

within the likelihood function. Then replace $E[\delta_{t+1}]$ with $\hat{\delta}_{t+1}$.

Finally, with all of the v's in hand, I can estimate the parameters of of the model via maximum likelihood by constructing a likelihood function from the choice probabilities given in equation 7.

6.2 Transition Probabilities - $\Omega(\vec{x}_{it+1}|\vec{x}_{it})$

Some of the state variables, such as the time fixed effects, transition deterministically. Other state variables, namely $\xi^o_{it}, \xi^g_{it}, \Delta\tilde{V}_{it}$, transition over time. However, I do not want to think of the firm as "waiting for good draws" of these state variables so I assume that their expected value in $t+1$ is equal to their value in t. Kellogg (2014) essentially makes the same modelling assumption.[11]

That leaves transitions in d_{it}, p^o_t, and p^g_t. To simulate changes in d_{it} I simply take draws from the observed distribution in Δd_{it} for different bins of d. Figure 5 plots these distribution of these distance changes.

To follow estimate the transition of p^o_t and p^g_t I follow Hernnstadt et al (2020) and Kellogg (2014) and estimate the following Markov process for both oil and gas:

$$\ln(p_{t+1}) = \ln(p_t) + \kappa_o + \kappa_1 p_t + \sigma\eta_{t+1} \tag{12}$$

In this price process κ_0 and κ_1 are drift and mean reversion parameters, respectively. Coefficients from the estimation of equation 12 are found table 4. I've also plotted the time series of prices along with some simulations based off of these parameters in figure 6. The estimated coefficients indicate that both price series are process with positive expected drift and mean reversion,

[11] Kellogg(2014) does not so explicitly try to construct expected production. Instead, he assumes production rationalizes costs (whereas I essentially do the opposite – construct expected revenue then estimate costs to rationalize production decisions.). However, in his dynamic model he makes the assumption that producers do not integrate over their expectation of new realizations in expected production to calculate their continuation value of the lease.

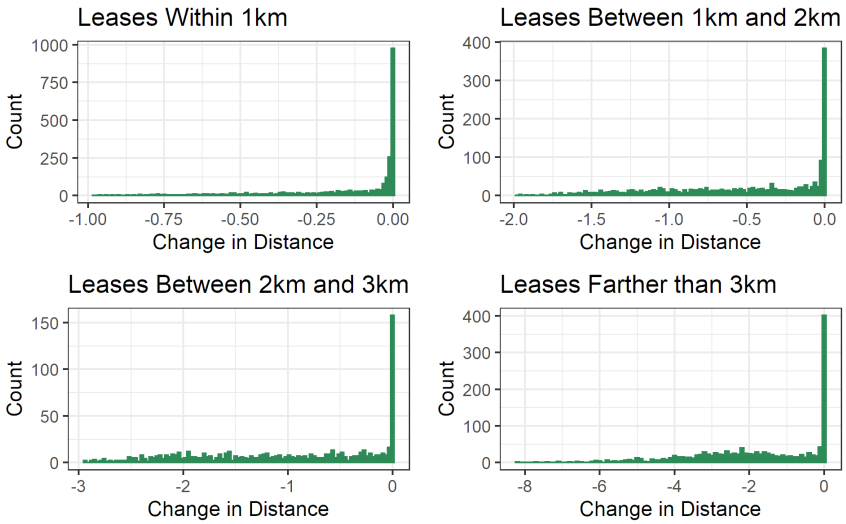


Figure 5: Positive changes in distance (m) for leases at varying distances

but that they are predominantly driven by the error term. You can see that for both processes, the residual standard error is much larger than the constant (drift term).

Table 4: Estimated Markov Process Parameters

	Dependent variable:	
	Δ log oil price	Δ log gas price
	(1)	(2)
Price	-0.002^{**}	-0.022^{**}
	(0.001)	(0.010)
Constant	0.107^{**}	0.096^{*}
	(0.048)	(0.049)
Observations	83	83
R^2	0.059	0.055
Adjusted R^2	0.047	0.043
Residual Std. Error (df = 81)	0.168	0.193
F Statistic (df = 1; 81)	5.047^{**}	4.697^{**}
Note:	*p<0.1; **p<0.05; ***p<0.01	

6.3 Construction of the spillover term - $\tilde{V}$

Consider the example in figure 7. There are nearby leases, A and B. A is considering building and connecting a well. If A chooses to connect, the future connection costs for B will decrease in the next period by $\alpha_d(|AB| - |B|)$. B should be willing to compensate A up to the difference

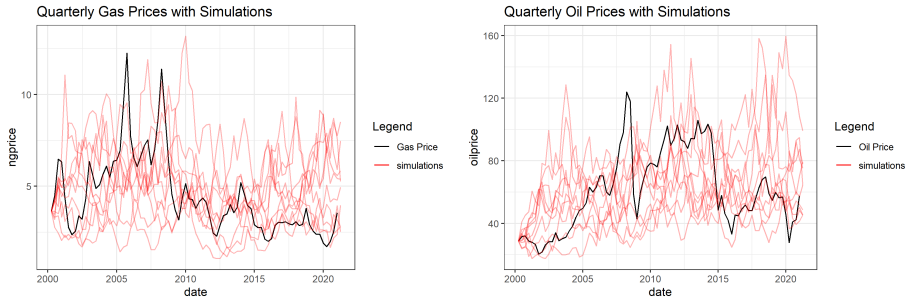


Figure 6: EIA Spot Prices and Examples of Simulated Price Paths

in B's ex-ante value: $\overline{V_{Bt}}(p_t, d_{Bt} = |AB|) - \overline{V_{Bt}}(p_t, d_{Bt} = |B|)$ to connect.[12] In practice, the construction of pipes and wells are typically completed by separate companies. However, the general principle still applies. A gathering company might be willing to offer a discount to A (to entice it to connect), with the hopes of decreasing its future costs and increasing its expected number of future customers. However, the gathering company is unlikely to be able to capitalize the entire change in $\overline{V}_B$, and therefore will be unable (and potentially unwilling) to pass that entire change on to A. Moreover, there might be other price distortions such as market power that would lead the gathering companies to not pass along a discount to A. My estimation strategy will estimate the proportion of this change in value captured by driller when the he chooses to connect (which should be between zero and one).

I can use the Hotz-Miller identity, equation 9, to simplify this difference to become the sum of the difference in expected log probabilities and the difference in pipeline costs. The parts of equation 13 can be constructed by calculating the counterfactual distance for B if A connects, then constructing the predicted probabilities (as discussed later in section 6.5).

$$\overline{V}_{Bt}(p_t, |AB|) - \overline{V}_{Bt}(p_t, |B|) = \ln(pr(j_t = G||B|, p_t)) - \ln(pr(j_t = G||AB|, p_t) + \alpha_d(|AB| - |B|) \quad (13)$$

[12] Some readers might notice that if the goal is to connect both A and B at minimum cost, the best way to do it would be to construct $|B|$, then draw a line orthogonal to $|B|$ to A. I believe that this is a *planning* issue as opposed to a spillover issue. For example, you would expect if A and B were owned by the same firm they would be more likely to select this minimum distance solution. Méndez-Ruiz (2019) studies whether contracting costs lead to more flaring by estimating a structural model of flaring in the Bakken, and addresses this planning issue more directly. He assumes that contracting costs are minimized within a firm, and estimates the degree to which a group of firms can transact as a single firm by exploiting variation in firm concentration across markets.

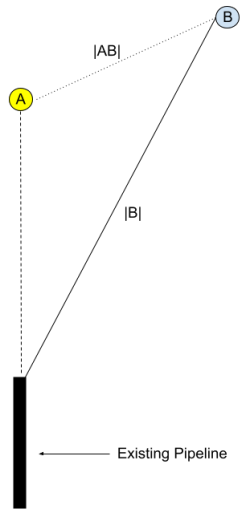
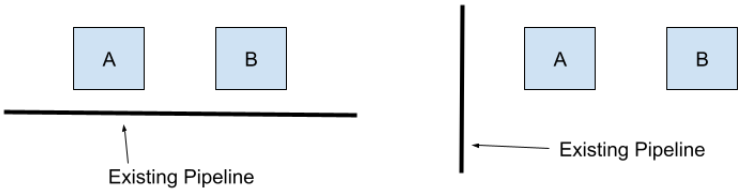


Figure 7: Hypothetical example showing that lease B's connection distance depends upon the choice of lease A.



(a) The spillover for B of A connecting is zero

(b) The spillover for B of A connecting is positive

Figure 8: Variation in $\tilde{V}$ resulting from geometrical differences in the configuration of leases and pipelines.

To construct $\tilde{V}$, for each lease-quarter I draw the shortest line from that lease to already existing pipes. For all leases $k \in K$ within 2km of that hypothetical line, I save their actual and counterfactual distances. Then $\tilde{V}_{it} = \sum_{k \in K} \Delta \overline{V}_{ik}$ is calculated during each iteration of the likelihood function. Variation in $\tilde{V}$ comes from differences in the likelihood of neighbor lease exercise (which is affected by, among other things, lease productivity, lease distance, and time until lease expiration) as well as geometrical variation in how much A choosing to connect affects B. Consider the simple example below in figure 8. In the first panel, A connecting has no effect on the distance B must traverse to connect. In the second panel, if A connects, that generates a spillover for B.

6.4 Construction of expected revenue

To measure expected productivity at each lease at time t, I take an inverse-distance weighted mean of expected production, ξ_i, from all wells completed within the past 18 months of t within 20km of the lease. In the dynamic model ξ_{it} enters as a state variable. Since the time window used to compute ξ moves with t, this means that expected productivity will change over time, despite the fact that the actual geology of the lease is time-invariant. This seems reasonable given that the outcome of neighboring wells likely leads to firms updating their beliefs.

Expected production from each well is an estimated term. To construct estimated production, I make a common assumption in the literature – to follow the Arps Model and assume production decays exponentially from a well (Lade and Rudik (2020), Méndez-Ruiz (2019)).[13] Oil and gas production in each time period, t, is given by:

$$o_{jt} = O_{j0} t^{\beta o} \exp(\nu^o_{jt})$$

$$g_{jt} = G_{j0} t^{\beta g} \exp(\nu^g_{jt})$$

where O_{j0}, G_{j0} are the initial rates of production, β are the decline rates, and ν are the model errors.

I estimate the decline rates, β^o and β^g, separately for wells drilled by year by estimating the well fixed-effect equations:

$$\log(g_{jt}) = \beta^g \log(t) + \delta_j + \nu^g_{jt}$$

$$\log(o_{jt}) = \beta^o \log(t) + \delta_j + \nu^o_{jt}$$

Figure 9 plots the estimated $\hat{\beta}^g$ and $\hat{\beta}^o$. For context, a decline parameter of -0.5 implies that

[13] An assumption used throughout the construction of expected production is that producers take expected production from their leases as a given. This is both technically false and a good simplifying assumption. Producers can affect their production by varying well depth, horizontal length, and amount of fluid used in fracturing. However by constructing expected production using nearby wells, I am essentially assuming that well-design questions are already enveloped-in and allowing expected production to reflect changes in technology over time and differences in optimal well construction across space.

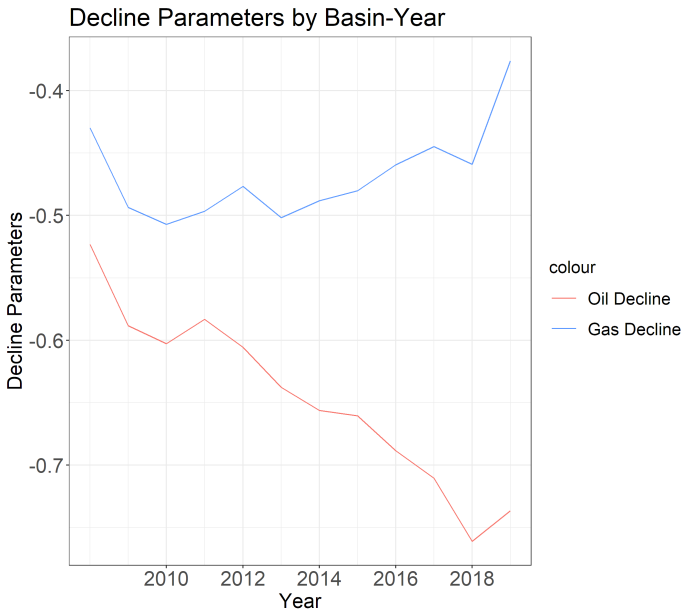


Figure 9: Plot of estimated Permian basin decline parameters by drilling year. Oil decline parameters are plotted in red while gas decline parameters are plotted in blue.

46% of discounted production over the first five years of a well's life will occur during the first year. With the estimated decline parameters in hand I can calculate expected monthly production by observing the well's initial production rate, which I obtain from Enverus. I use second-month production instead of first-month, since the first-month values are often truncated by not having a full month of production. Finally, I obtain expected well-level lifetime production by summing and discounting expected monthly production for each well over a five year period.

6.5 Estimation of conditional choice crobabilities

Ideally, I would be able to construct conditional choice probabilities non-parametrically. However, the relatively large state space makes this impossible. Following other papers in the literature such as Murphy (2018), I estimate conditional choice probabilities by estimating a logit with b-spline expansions. Specifically I estimate:

$$\hat{Prob}\left(j = G|\vec{x}_{it}\right) = \Lambda\left(\sum_{s=1}^{5}\phi_s^{\xi^o}S_s^{\xi^o}(\xi_{it}^o)+\sum_{s=1}^{5}\phi_s^{\xi^g}S_s^{\xi^g}(\xi_{it}^g)+\sum_{s=1}^{5}\phi_s^d S_s^d(d_{it})+\sum_{y=2011}^{2019}\phi_y+\sum_{\tau=0}^{8}\psi_\tau\right) \quad (14)$$

where Λ denotes the logistic CDF, $S(\cdot)$ are the spline basis functions of the argument in the superscript, ϕ_y are year fixed-effects, and ψ_τ are quarter-until-expiration fixed-effects.

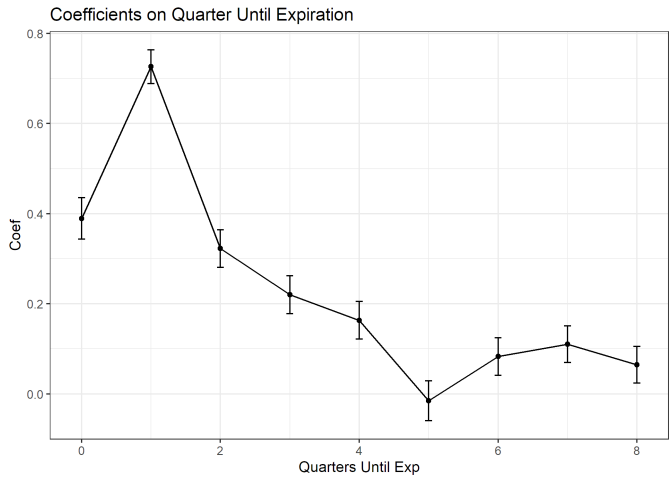


Figure 10: Coefficients on Quarter Until Expiration

7 Results

7.1 Conditional choice probabilities

The coefficients from the logit match the expected intuition from the underlying dynamic model. Figure 10 plots the coefficients on quarter until expiration from the choice probability logit. In the dynamic model the probability of drilling and gathering should increase exponentially until the time of expiration. The shape of the estimated fixed effects is mostly consistent with this, with the main exception being that the coefficient at the quarter of expiration is lower than the coefficient for one quarter away from the time of expiration. These coefficients are an important source of variation in the expected probability of drilling since they do not directly affect the payout associated with drilling.

7.2 Dynamic discrete choice cesults

Results for the dynamic discrete choice can be found in table 5 and the estimated year fixed effects are plotted in figure 12. Standard errors are calculated with 50 bootstrap replications where I sample leases with replacement. Expected revenue was entered into the likelihood as $\$10,000,000$ in revenue, so each parameter can be interpreted as tens of millions of dollars. The scale parameter of the error term is estimated to be \$7 million. This implies a standard deviation of the error term of \$7 million $\cdot \pi/\sqrt{6} = \$9$ million.

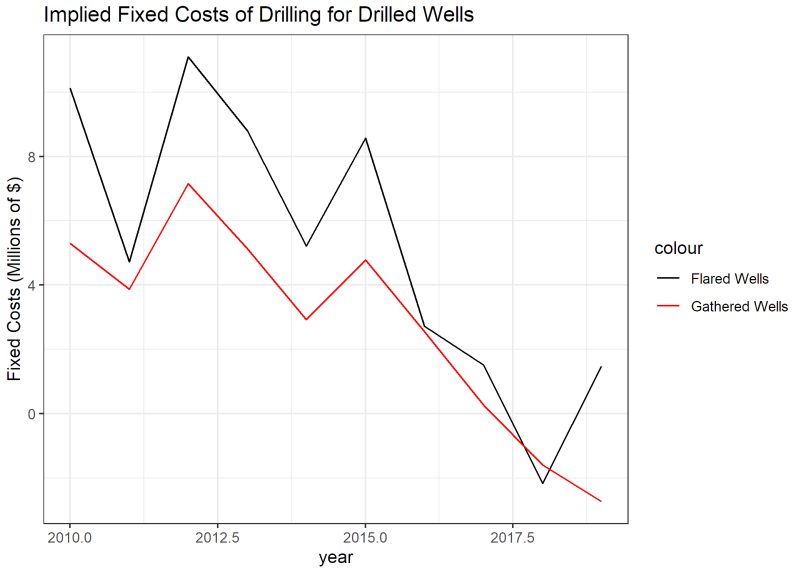


Figure 11: Expected value of fixed costs for drilled wells by year. For flared wells this is calculated as $C_f + \delta_y(t) + E[\epsilon_{iFt}|i \text{ chose } F \text{ during } t]$. It is calculated analogously for gathered wells.

The estimated fixed costs are $\$38,000,000$ and $\$21,000,000$ for F and G, respectively, which are quite high. However, these are the implied costs for a randomly selected lease – most of which are not observed drilling. The implied cost for drilled wells is the estimated fixed cost plus the error term (which is structurally a cost shock). Conditional on drilling the error term is expected to be positive and relatively large. Thus, the expected cost for wells actual observed drilling is far lower. The same issue appeared in Murphy (2018) in his study of parcel development in San Francisco using very similar methods. His estimated cost of development were very high, but the fixed costs of development for parcels which actually developed were far more believable. I find that when the expectation of the error term for drilled wells is included with the fixed costs, most of the implied fixed costs are positive, suggesting that my estimates of revenue are biased downward. Figure 11 plots these implied fixed costs. You can also see that after including the expectation of the conditional error, the fixed costs for gathered wells are greater for gathered wells than flared wells.

The estimated cost of distance is also quite high. The estimated cost per kilometer of pipeline for less than 1km is almost $\$10,000,000$. I estimate that the cost per km is decreasing in overall distance. However, these cost estimates are biased upwards since my counterfactual simulations show that predicted pipeline distance is less than actual distance built – eg if a lease is 1km from

the nearest pipeline, in expectation more than 1km will be built. This gap can be explained by two factors. First, pipelines often do not traverse the shortest distance between two points and must go around obstacles or land where a right-of-way could not be secured. Second, I observe many pipelines being constructed in parallel with each other. This implies that some pipelines are at capacity and that the nearest connection for a new well is not necessarily available. Both of these reasons bias my cost estimates upward and will be discussed in greater detail in the following section.

Lastly, the parameter on the spillover term is 0.03, implying that producers do not internalize the vast majority of spillovers for their neighbors. The estimate has a very tight confidence interval that does not overlap with zero. However, the magnitude of the parameter is very close to zero. In order to assess the empirical magnitudes of the cost parameters and of the spillover internalization parameter, I use a forward simulation procedure described in the next section.

	Estimate	sd	Interval95
Error Scale Parameter (σ)	0.70	0.13	(0.442, 0.962)
Fixed Costs of Flaring (C_f)	-3.87	0.86	(-5.55, -2.183)
Fixed Costs of Gathering (C_g)	-2.13	0.54	(-3.179, -1.083)
$Spillovers\Delta\tilde{V}$	0.03	0.00	(0.02, 0.035)
Distance (d)	-0.96	0.19	(-1.336, -0.582)
$d \times \mathbb{1}(1 \leq d < 2)$	0.01	0.09	(-0.168, 0.194)
$d \times \mathbb{1}(2 \leq d < 3)$	0.36	0.11	(0.146, 0.579)
$d \times \mathbb{1}(3 \leq d)$	0.81	0.17	(0.49, 1.139)

Table 5: DDC Results

8 Simulations and Policy Counterfactuals

Having estimated my model, I now proceed to using my model estimates for simulations. The first set of simulations I discuss involve constructing a close analogue to the static marginal abatement curves constructed by Lade and Rudik (2020). Then, I proceed to a full dynamic forward simulation which will allow me to asses the performance of my model estimates, examine the estimated magnitudes of the parameters, and make predictions about drilling and flaring outcomes under varying policy counterfactuals.

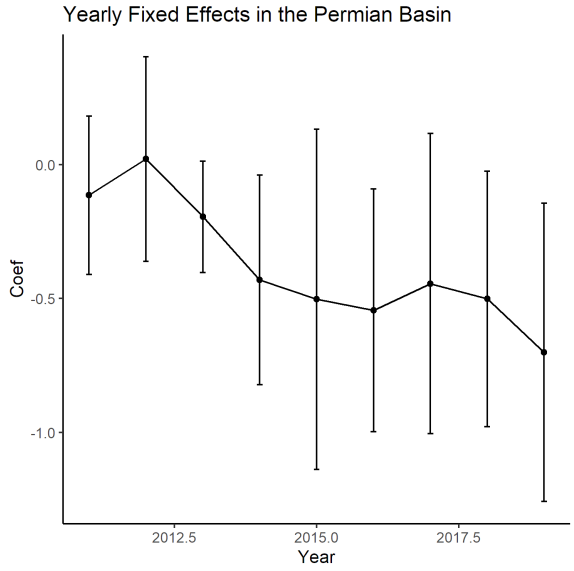


Figure 12: Year Fixed Effects

8.1 Estimated costs perform better than reported costs

To compare my results with previous work I follow the methodology of Lade and Rudik (2020) and construct static marginal flaring abatement cost curves on my own sample using first, the reported cost estimates used in that paper, then my own cost estimates.

Lade and Rudik (2020) calculate the marginal abatement cost for wellpad i as:

$$MAC_{it} = \frac{FC + \alpha_d d_i}{G_{it}}$$

where G_i is the sum of discounted expected production from well i from t onwards. The industry-wide abatement curve is calculated by calculating MAC_{it} for each well, i, then ordering them by cost. Since my unit of analysis is the lease-level I use leases as opposed to wellpads. Lade and Rudik (2020) use a fixed cost estimate of $\$202,000$ lifted from a survey result. They predict α_d by using an ordered probit to predict necessary pipeline width, and pipeline cost estimates lifted from the same survey results.

I construct a simplified version of their cost curves. First I use the same costs they did, lifted from the report from INGAA. However, I do not do an ordered probit and instead make the

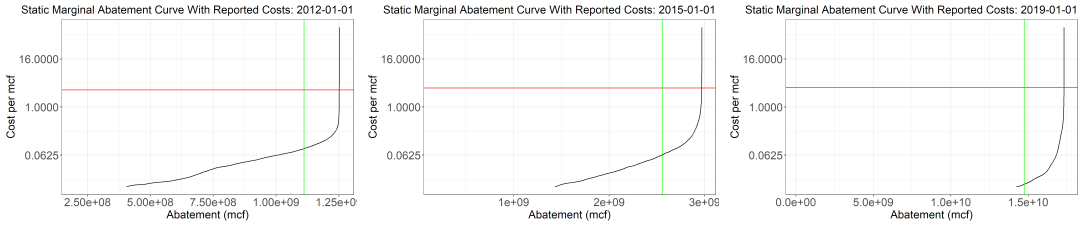


Figure 13: Static Marginal Abatement Curve Calculated with Engineering Estimates of Cost. The red line gives the observed spot price of gas at the time, while the green line gives the observed level of abatement.

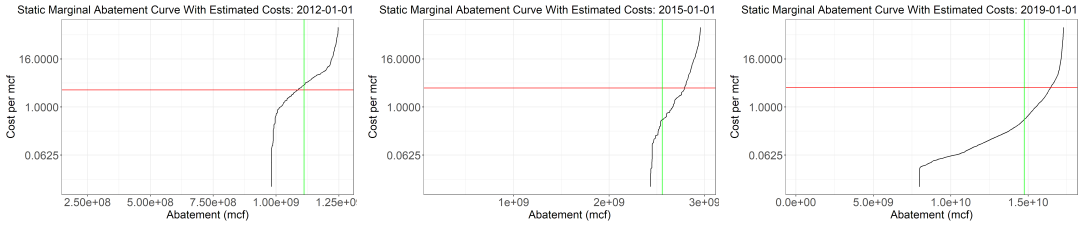


Figure 14: Static Marginal Abatement Curve Calculated with Estimated Costs. The red line gives the observed spot price of gas at the time, while the green line gives the observed level of abatement.

conservative calculation of assuming all pipelines are 8 inches.[14] Next, I use my own distance and cost measures lifted from my model. The fixed cost estimates are taken as the difference in fixed costs between flared and gathered wells. I use the estimates from figure 11, since they reflect differences in fixed costs for drilled wells as opposed to values from table 5, which reflect estimated fixed costs for randomly selected leases.

Figure 13 plots the marginal abatement curves following Lade and Rudik (2020) and figure 14 plots the marginal abatement curves with my estimated costs. I've also plotted the spot price of gas in red, and the observed level of abatement in green. When using engineering estimates, the observed level of abatement is too low – implying that the cost curves are also too low. When using my own cost estimates, the abatement cost curve shifts upwards, decreasing the wedge between expected and observed abatement.

8.2 Full dynamic simulations: computational procedure

To calculate a full dynamic forward simulation, I can no longer rely on having $\log(pr(j = G))$ in hand to calculate $E[V_{t+1}]$. Therefore, I take simulation draws of prices and distance changes,

[14] This is conservative because 8-inch pipelines are the most expensive in their model, and I find that generally these engineering estimates over predict abatement.

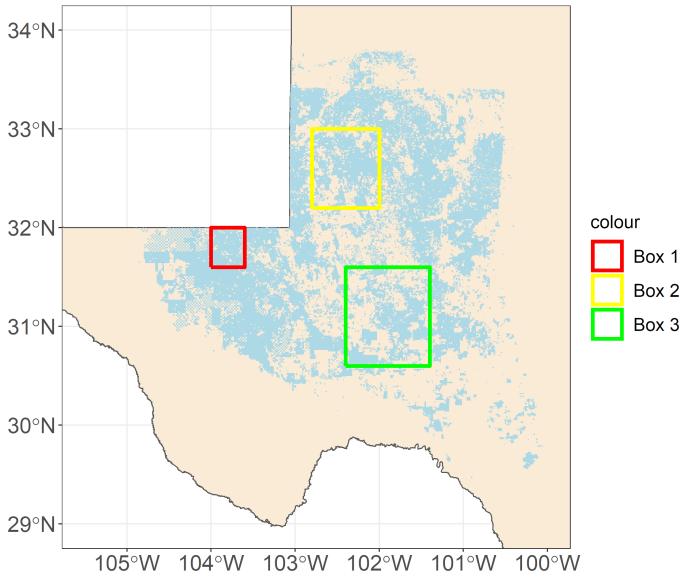


Figure 15: Plot of the counterfactual boxes. Leases are plotted behind in light blue.

then use backwards induction to calculate the expectation of V. I take draws of ϵ_{it} to determine the action of each lease in each period. For each period, after the gathered leases are determined, I draw a line from the existing network to each lease. Then distances for remaining leases are re-computed for $t+1$.

Since this process relies on taking draws of distance changes and the distribution of distance changes varies with the policy counterfactual, this routine iterates until the expectation of predicted distance changes is close to the expectation of the distribution of distance changes used to compute the counterfactual. Specifically, I specify that the mean predicted distance change must be within 10m of the mean of the distance changes used to compute the counterfactual. In practice, most counterfactual runs only iterate once. All converge within two iterations. This implies two things – first that the counterfactuals do a very good job of predicting observed distance changes. Second that the network is not particularly sensitive to changes in policy. This will be demonstrated in the following subsections.

Due to computational constraints, I can only compute counterfactuals on samples of the Permian Basin. Figure 15 plots all of the leases in my sample. These boxes were selected since they are from disparate regions of the Permian basin.

Box	Actual Distance Constructed	Simulation Distance
Box 1	1129.875	93.63384
Box 2	3133.516	116.0214
Box 3	814.2046	268.3943

Table 6: Actual and Predicted Pipeline Distances (km)

8.3 Performance of counterfactuals against observed activity

To assess the fit of the simulations I've plotted the predicted versus actual lease outcomes below in figure 16 for each box. Expected productivity is spatially correlated, so when counterfactuals are computed on only a small part of the basin it's unsurprising that predicted flaring or gathering might be systematically above or below the observed level. For example, actual flaring is higher than the baseline counterfactual in Box 2 and lower in Box 1. However, the counterfactual simulation on Box 3 looks quite good. Overall, you would expect that if I could run counterfactuals on the basin as a whole, the errors across smaller areas would average out and the fit for the overall basin would be good.

As noted in the discussion of the computation of my counterfactuals, the counterfactuals match predicted distance changes with observed distance changes quite well. However, the overall prediction of pipelines built tends to be far too low. A comparison of predicted versus observed pipeline construction over my sample period (01/01/2010-12/01/2019) is in table 6. The discrepancies are likely due to the fact that my model ignores capacity and processing constraints. In practice, I observe many parallel pipelines, which could not be rationalized in my model. However, it does resolve why the estimated per-mile costs of pipelines in my model are so high. Recall that my estimate of pipeline costs for pipeline costs per kilometer was \$9.6 million for distances less than 1km, \$6 million for distances between 2km and 3km, and \$1.5 million for distances greater than 3km. In Box 2, predicted pipeline construction was off by a factor of 27 – implying that for every well connected that was 1 km away from pipelines, 27km of pipelines was actually built. Adjusting my pipeline costs by this factor would produce a per kilometer estimate of $\$355,555$ for distances less than 1km and $\$55,555$ per km for distances greater than 3km. Predicted construction in box 3 was three times lower than the actual length constructed, implying a per kilometer cost of \$3.2 million for distances less than 1km and $\$500,000$ for distances greater than 3km.

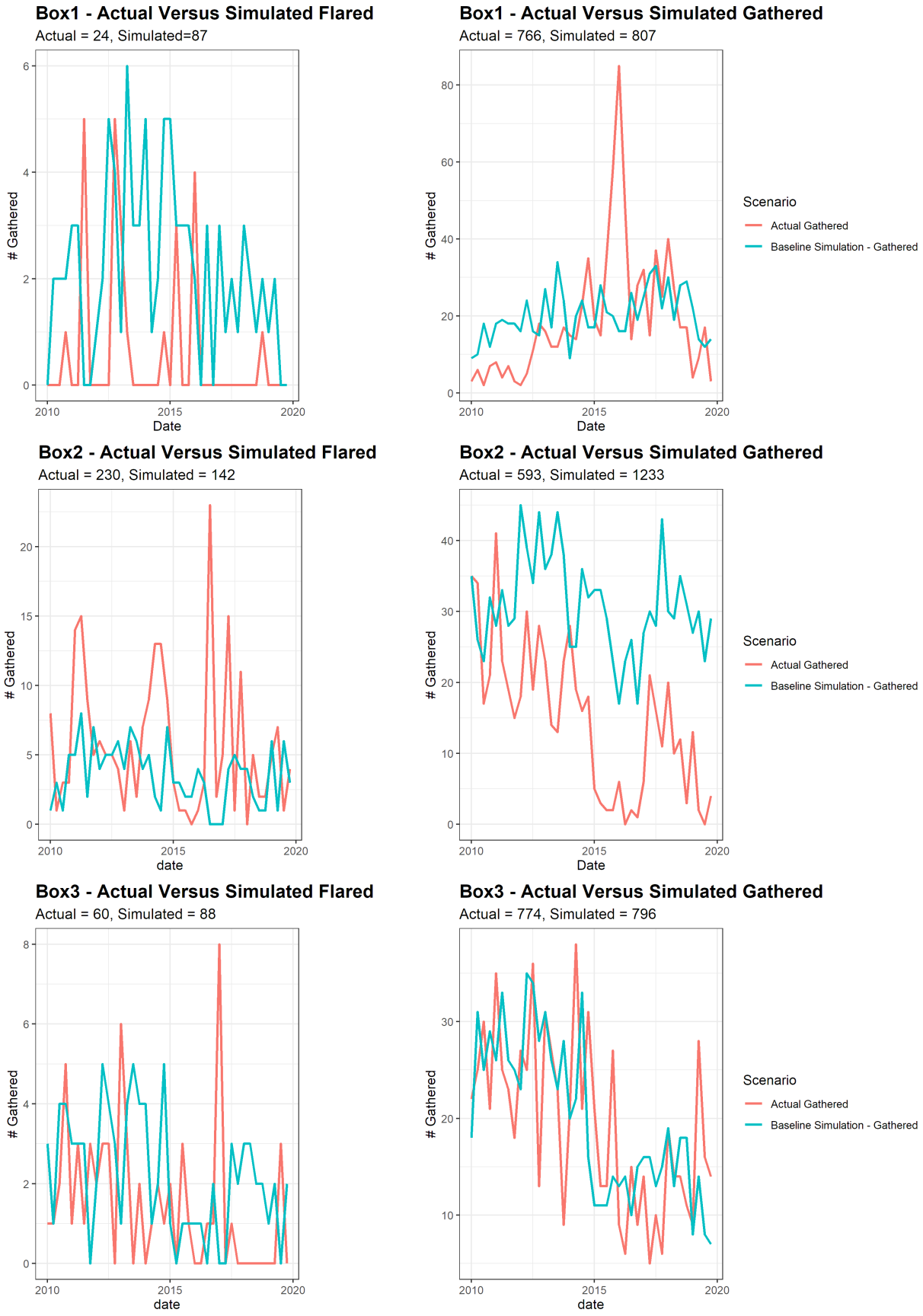


Figure 16: Actual Versus Simulated Outcomes

8.4 Spillover internalization does not appear to be a major driver of flaring

To asses how the importance of the spillover internalization parameter, I run counterfactual simulations in Box 1 where I switch the parameter to 1 – implying perfect internalization. These are plotted in figure 17. This figure plots the number of wells of each type (gathered or flared) drilled in each time period, as well as the percentage of production relative to the baseline simulation. Perfect internalization does seem to induce some wells that had previously been flaring to gather, leading to a small increase in gathered gas volumes and a small decrease in flared gas volumes. I also present total production figures in table 7. Perfect internalization leads to a 5% increase in sold gas volumes and a 6% decrease in flared volumes. Overall, the effects of spillover internalization seem to be relatively small. Nonetheless, given the magnitude of the social costs of flaring, it's plausible that perfect internalization could decrease social costs by tens of millions of dollars per year.

One explanation for the seemingly small impact of the internalization parameter is that the costs of gathering appear to be relatively small compared to the very high fixed costs of gathering. I run a comparison of perfect versus zero internalization where I double the costs of pipelines. These results are also plotted in figure 17. This seems to increase the wedge between the two internalization scenarios ($\alpha_{\tilde{V}} = 0$ and $\alpha_{\tilde{V}} = 1$), but not dramatically.

8.5 The effect of a flaring tax, flaring ban, and gas subsidy on drilling and flaring

To assess the performance of potential policies I run simulations in Box 1 with various levels of flaring taxes, a gas subsidy, and a flaring ban. These results are plotted in figure 18, and aggregate production figures can also be found in table 7. A subsidy poorly targets flaring. A \$1 per Mcf gas subsidy increases total gas production by 4% but only decreases flaring by 2% at a cost (in just this box) of \$600 million. A tax targets flaring quite well. A \$5/Mcf tax reduces flared volumes by 39%, increases gathered volumes by 5%, decreases oil produced by 2%, and raises \$150 million in revenue.

	Gas Gathered (10 Bcf)	Oil From Gathered Wells (10 million barrels)	Gas Flared (10 Bcf)	Oil From Flared Wells (10 million barrels)	Total Oil Production (10 million barrels)
Baseline Simulation	57.85	7.58	4.91	0.69	8.27
Subsidy of \$1	60.30	7.89	4.82	0.67	8.56
Flaring Ban	59.20	7.78	0.00	0.00	7.78
Tax of \$0.55 per Mcf	58.03	7.61	4.68	0.66	8.27
Tax of \$1 per Mcf	58.07	7.62	4.45	0.63	8.26
Tax of \$2 per Mcf	58.04	7.61	3.90	0.55	8.16
Tax of \$3 per Mcf	58.19	7.63	3.48	0.50	8.13
Tax of \$5 per Mcf	58.22	7.63	3.00	0.43	8.07
Spillovers Internalized	60.73	7.91	4.62	0.65	8.55
Gathering Costs Doubled	49.53	6.49	5.16	0.72	7.21
Spillovers Internalized & Gathering Costs Doubled	54.73	7.16	4.74	0.67	7.82
Network Fixed at $t = 0$	48.32	6.29	5.11	0.72	7.01

Table 7: Total oil and gas production in Box 1 in various simulations

Flaring volumes seem to respond linearly to taxes at low levels of the tax, then less than linearly as the tax get higher. A \$1/Mcf tax reduces flared volumes by 9% while a \$2/Mcf tax reduces flared volumes by 20%. More than doubling that tax to \$5/Mcf less than doubles the reduction in flared volumes to 39%. Lastly, a ban eliminates all flaring, leads to a small increase in gas production (and hence only a small increase in pipeline investment), and causes a small decrease in oil production. Theoretically, a ban might be attractive since there are multiple market failures (spillover effects and the external costs of flaring). It might be the case that a ban could simply induce flaring wells to wait to drill. However, my empirical results suggest that a ban is a very blunt policy tool in this context, leading to an overall decrease in oil production.

8.6 Endogenous network growth has a substantial effect on drilling and flaring

Lastly, I show that allowing the network to grow endogenously is actually quite important for simulation outcomes. I run the simulation but fix lease distances to their values at t =2010-01-01. Flared wells increase modestly, but connected wells decrease drastically. Results from this simulation are plotted in figure 19, and aggregate production in this simulation can be found in table 7. Total gas production is 16% lower if the network cannot grow, demonstrating that growth over time is an important determinant of drilling outcomes.

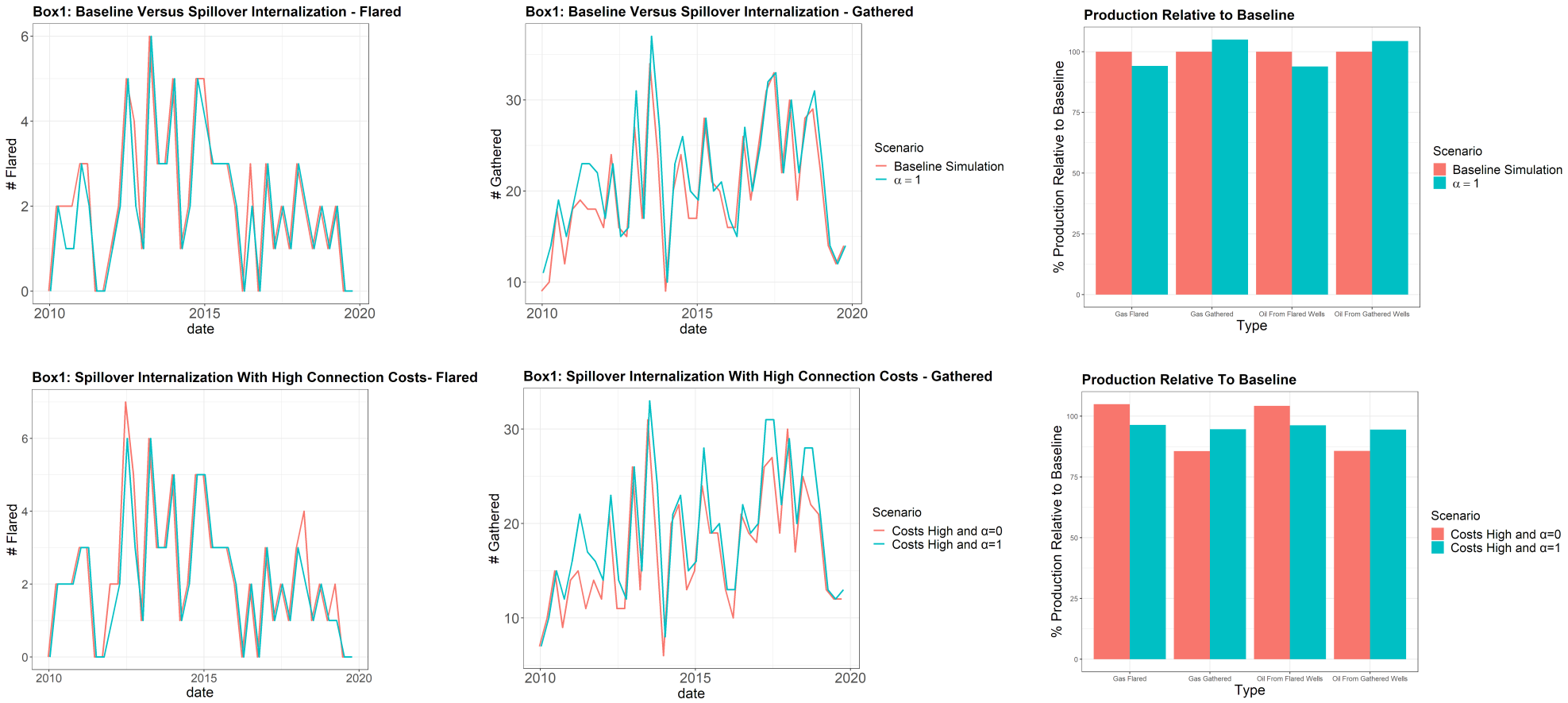


Figure 17: Simulated outcomes with the spillover internalization parameter switched to 1. The line plots present number of wells drilled in each category over time, while the bar charts present total production from each well type relative to the baseline. The graphs in the top panel of this figure only alter the spillover internalization parameter. The graphs in the lower panel multiply the costs of pipelines by 2 to determine whether this would increase the wedge between $\alpha_{\tilde{V}} = 0$ and $\alpha_{\tilde{V}} = 1$.

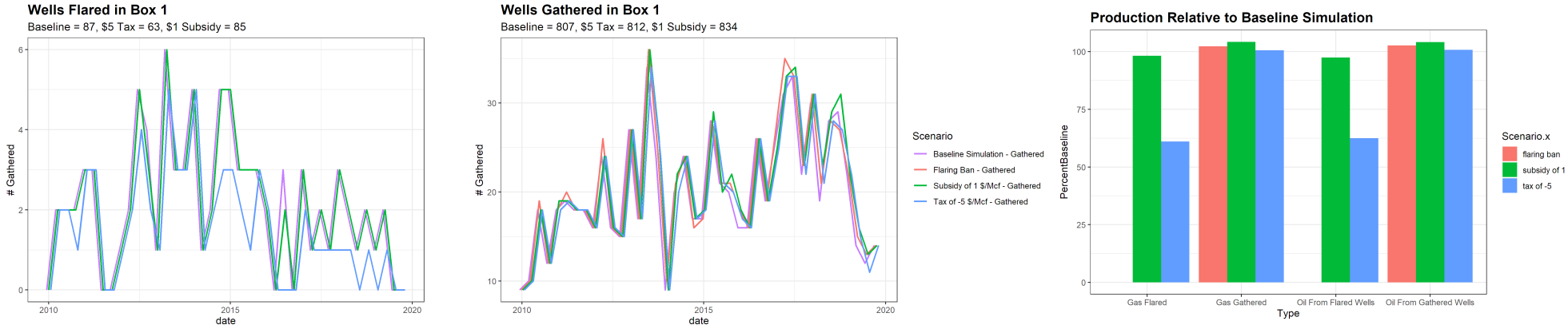


Figure 18: Outcomes of Flaring Ban, Flaring Tax, Gas Subsidy

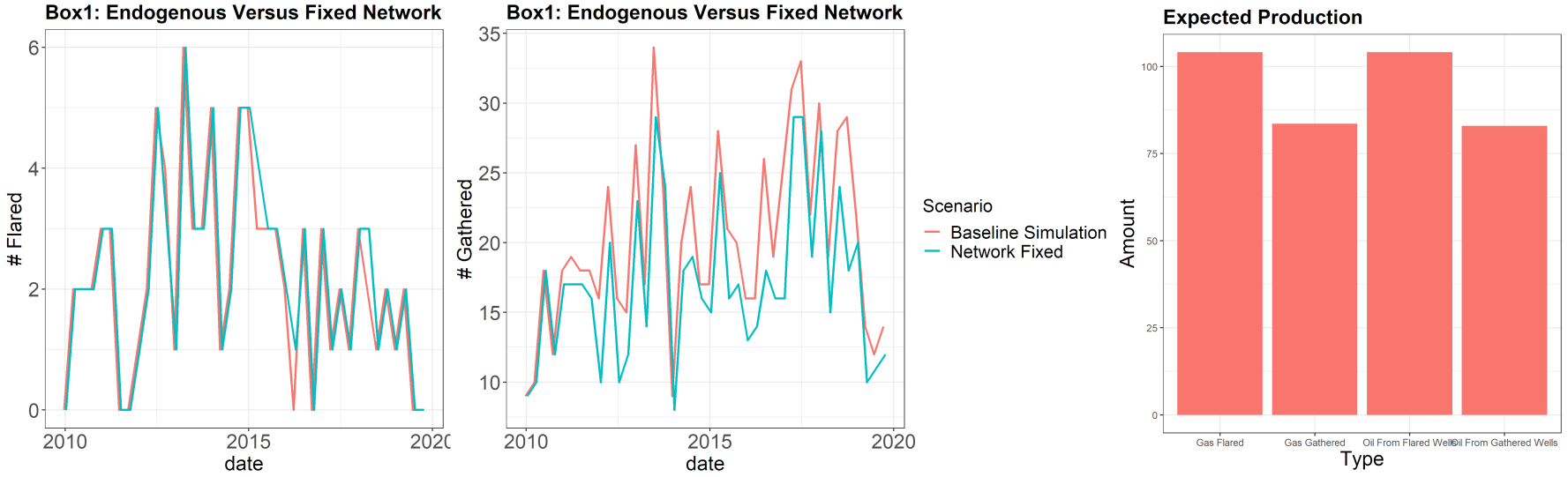


Figure 19: Outcomes with network fixed at t =2010-01-01 versus endogenous

9 Conclusion

I show that distance to pipeline infrastructure matters for flaring outcomes and that as the network grows over time, connections become less costly. My model indicates that flaring abatement is relatively costly, but that a tax near the likely social cost could substantially reduce flaring. A $5 per Mcf tax would reduce flared volumes due to wells not connecting by 39%. Moreover, my results find that the majority of the decrease in flaring wells comes from flaring wells switching to not drilling rather than switching to connecting. In addition, I also find that subsidies poorly target flaring, and that bans only lead to a small increase in pipeline investment. My model estimates also suggest that spatial spillovers are not internalized. However, my counterfactual simulations suggest that the lack of spillover internalization does not have a substantial effect on flared volumes and that the lack of internalization might simply be due to the benefits of internalization being relatively small.

This paper makes a few important contributions to the literature. First, this paper is the first to explore the endogenous relationship between drilling and gas pipeline construction. Second, this paper is the first paper which models flaring *and* allows policies to affect the drilling margin. Third, this paper estimates pipeline costs directly instead of relying on industry reported costs. Combined, these contributions make this paper the most accurate model of the drilling and connecting decisions of firms.

These decisions have large consequences for U.S. greenhouse gas emissions, making this paper a timely contribution for policymakers interested in addressing climate change. The U.S. EIA estimates that in 2020 about 420 million Mcf of natural gas was vented or flared in the U.S. (U.S. Energy Information Administration, 2021b). Assuming a social cost of flaring between $3 and $6 implies that flaring accounts for billions of dollars in climate damages annually.

References

Agerton, Mark and Gregory B Upton. 2020. "The Economics of Natural Gas Flaring in U.S. Shale: An Agenda for Research and Policy." URL https://www.energy.gov/sites/prod/files/2019/08/f65/.

Arcidiacono, Peter and Robert A Miller. 2011. "Conditional Choice Probability Estimation of Dynamic Discete Choice Models with Unobserved Heterogeneity." *Econometrica* 79 (6):1823–1867. URL https://www.jstor.org/stable/41336537.

Chapa, Sergio. 2021. "Texas Oil Body Makes It a Little Harder for Wells to Flare Gas - Bloomberg." URL https://www.bloomberg.com/news/articles/2021-09-14/texas-oil-body-makes-it-a-little-harder-for-wells-to-flare-gas.

Covert, Thomas and Ryan Kellogg. 2017. "Crude by Rail, Option Value, and Pipeline Investment." *Becker Friedman Institute Working Paper* URL https://bfi.uchicago.edu/working-paper/crude-by-rail-option-value-and-pipeline-investment/.

Covert, Thomas R. 2015. "Experiential and Social Learning in Firms: The Case of Hydraulic Fracturing in the Bakken Shale." *SSRN Electronic Journal* URL https://papers.ssrn.com/abstract=2481321.

Covert, Thomas R and Richard L Sweeney. 2019. "NBER WORKING PAPER SERIES RELINQUISHING RICHES: AUCTIONS VS INFORMAL NEGOTIATIONS IN TEXAS OIL AND GAS LEASING." URL https://github.com/rlsweeney/public_cs_texas.

———. 2022. "Secrecy Rules and Exploratory Investment: Theory and Evidence from the Shale Boom *." *Working Paper* URL http://www.richard-sweeney.com/pdfs/cs_disclosure.pdf.

Décaire, Paul H, Erik Gilje, and Jérôme Taillard. 2019. "Real Option Exercise: Empirical Evidence." *SSRN Electronic Journal* URL https://papers.ssrn.com/abstract=3342565.

Department of Energy. 2022. "Queued Up... But in Need of Transmission Unleashing the Benefits of Clean Power with Grid Infrastructure." Tech. rep. URL https://www.energy.gov/sites/default/files/2022-04/Queued%20Up%E2%80%A6But%20in%20Need%20of%20Transmission.pdf.

EDF. 2022. "Methodology for EDF's Permian Methane Analysis Project (PermianMAP) Data Collection and Analysis." Tech. rep. URL https://www.edf.org/sites/default/files/documents/PermianMapMethodology_1.pdf.

Gowrisankaran, Gautam, Marc Rysman, Dan Ackerberg, Victor Aguirregabiria, Ana Aizcorbe, Rabah Amir, Lanier Benkard, Steve Berry, Sofronis Clerides, Tim Erickson, Simon Gilchrist, and Avi Goldfarb. 2012. "Dynamics of Consumer Demand for New Durable Goods." *Source: Journal of Political Economy* 120 (6):1173–1219.

Herrnstadt, Evan, Ryan Kellogg, and Eric Lewis. 2020. "The Economics of Time-Limited Development Options: The Case of Oil and Gas Leases." *SSRN Electronic Journal* URL https://papers.ssrn.com/abstract=3604162.

Hlavac, Marek. 2018. "stargazer: Well-Formatted Regression and Summary Statistics Tables." URL https://CRAN.R-project.org/package=stargazer.

Hotz, Joseph A. and Robert A. Miller. 1993. "Conditional choice probabilities and the estimation of dynamic models." *Review of Economic Studies* 60 (3):497–529.

ICF. 2018. "North America Midstream Infrastructure through 2035: Significant Development Continues." Tech. rep. URL https://www.ingaa.org/File.aspx?id=34658.

Interagency Working Group on Social Cost of Greenhouse Gases. 2021. "Technical Support Document: Social Cost of Carbon, Methane, and Nitrous Oxide Interim Estimates under Executive Order 13990 Interagency Working Group on Social Cost of Greenhouse Gases, United States Government With participation by Council of Economic Advisers Council on Environmental Quality." .

Kellogg, Ryan. 2014. "The effect of uncertainty on investment: Evidence from texas oil drilling." *American Economic Review* 104 (6):1698–1734.

Lade, Gabriel E. and Ivan Rudik. 2020. "Costs of inefficient regulation: Evidence from the Bakken." *Journal of Environmental Economics and Management* 102:102336.

Maechler, Martin, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. 2021. "cluster: Cluster Analysis Basics and Extensions." URL https://CRAN.R-project.org/package=cluster.

Méndez-Ruiz, Andrés. 2019. "Coase on Fire: Natural Gas Flaring Abatement Technology Adoption and its Regulation in North Dakota." URL https://www.perc.org/wp-content/uploads/2019/08/Andres_Workshop_2_Paper.pdf.

Murphy, Alvin. 2018. "A Dynamic Model of Housing Supply." *American Economic Journal: Economic Policy* 10 (4):243–67.

Paddock, James L., Daniel R. Siegel, and James L. Smith. 1988. "Option valuation of claims on real assets: The case of offshore petroleum leases." *Quarterly Journal of Economics* 103 (3):479–508.

Plant, Genevieve, Eric A. Kort, Adam R. Brandt, Yuanlei Chen, Graham Fordice, Alan M. Gorchov Negron, Stefan Schwietzke, Mackenzie Smith, and Daniel Zavala-Araiza. 2022. "Inefficient and unlit natural gas flares both emit large quantities of methane." *Science* 377 (6614):1566–1571. URL https://www.science.org/doi/10.1126/science.abq0385.

Rochet, Jean Charles and Jean Tirole. 2003. "Platform Competition in Two-Sided Markets." *Journal of the European Economic Association* 1 (4):990–1029. URL https://academic.oup.com/jeea/article/1/4/990/2280902.

Texas Railroad Commission. ???? "Flaring Regulation FAQs." URL https://www.rrc.texas.gov/about-us/faqs/oil-gas-faqs/flaring-regulation/.

U.S. Energy Information Administration. 2021a. "Maps: Oil and Gas Exploration, Resources, and Production - Energy Information Administration." URL https://www.eia.gov/maps/maps.htm.

———. 2021b. "Natural Gas Vented and Flared." URL https://www.eia.gov/dnav/ng/ng_prod_sum_a_EPG0_VGV_mmcf_a.htm.

10 Appendix

10.1 Processing the Leases

I run the raw lease data through two rounds of clustering. The first round is intended to identify and remove duplicates. The second round is intended to identify lease amendments and changes to lease ownership.

For the first round of clustering I follow many of the lease processing choices of Herrnstadt, Kellogg, and Lewis (2020). In the Enverus data on leases, there are many likely duplicates that need to be dealt with. There are many cases where the same plot is leased to the same grantee by multiple grantors. These observations are likely a result of undivided mineral interests. For example, multiple members of a family might each show up as a distinct grantor with her own row in the data. I keep all observations labeled as "Leases," "Lease amendments," "memos of leases," and "lease extensions." I drop observations on mineral rights assignments, lease ratification, royalty deeds, and mineral deeds. Next, I construct a dissimilarity matrix which I pass along to an agglomerative hierarchical clustering method implemented by the agnes function in the `cluster` package, version 2.2.1, in R.

The entries of the dissimilarity matrix are calculated as which I calculate as:

$$d_{ij} = \sum_k m_k(x_i^k, x_j^k)$$

where $m_k(\cdot)$ is equal to 0 if the kth attribute of lease i and j are identical, and positive but less than or equal to 1 otherwise. The attributes and calculation of m_k are as follows:

- $m_k(x_i^k, x_j^k) = |x_i^k - x_j^k|$ for various continuous variables. I use this on both the start and end date of the leases, and the Demerau-Levenshtein string distance between the grantors and grantees. Once I calculate all the pairwise combinations, I then rescale all the observations by dividing through by the max (for each characteristic), so all the date differences lie between 0 and 1.
- $m_A(x_i^A, x_j^A) = \frac{|x_i^A \cap x_j^A|}{2(|x_i^A| + |x_j^A|)}$ - where x_i^A is the spatial polygon of i. This measures the average

percentage overlap in area between i and j, and will be bound between 0 (if the polygons do not intersect) and 1 (if the polygons are exactly the same).

Before clustering, there are over 500,000 observations in the data. Constructing a 500,000×500,000 matrix would be infeasible. Even when accounting for the fact that the matrix would be symmetric it would have over 100 billion elements. To deal with this, I run the clustering algorithm at the year of expiration date level.

After this initial round of clustering, primarily intended to identify rows of duplicates, I run this data through one more layer of clustering. This time, I just consider the average percentage overlap between two leases and group them together if the percentage overlap is over 70%. Once two leases are clustered together, I consider the lease with the greatest Instrument Date active at time t to be the active lease and throw out the older ones. Consider the hypothetical scenario below: In this example, for all dates before 12-01-2009, the expiration date would be treated as

Effective Date	Expiration Date	Instrument Date	Polygon
12-01-2007	12-01-2010	12-01-2007	A
12-01-2009	12-01-2012	12-01-2009	A

12-01-2010. So for example at 12-01-2008, there would be 24 months until the lease expiration. Once the new lease becomes active after 12-01-2009, the time until expiration would resent to 36 months.