Appendix For Online Publication

A Data and Geoprocessing Steps

This appendix describes the individual steps that we took to attach the $ln(FirstYearProduction_{lct})$ independent variable and the average total drilled depth variable to each lease in our sample, both of which are required to estimate model 3 described in Section 5.

A.1 Sample of Unconventional Oil and Gas Wells

The Unconventional Natural Gas Well Geodatabase published by the Carnegie Museum of Natural History (2022) compiles the various oil and gas well reports published by the Pennsylvania Department of Environmental Protection (DEP) in a clean, easy-to-use format. We downloaded the data published as of the second quarter of 2021, and begin with the shapefile dataset, which includes variables across multiple DEP reports, including permit, spud, production, waste, and compliance report. We exported the attribute table in ArcGIS as a .dbf file and imported it into Stata. The file contains 22,597 rows, each corresponding to a unique DEP permit number. We dropped 17 observations of wells that are classified as observation, coalbed methane, or storage wells, which are outside of the scope of our study. We also dropped 9,524 wells that were permitted by never drilled. This leaves us with data on the locations and attributes of 13,063 unconventional oil and gas wells, uniquely defined by their permit numbers.

A.2 Creation of a Lease and Unconventional Well Near Table

We imported the wells locations back into ArcGIS and created a "near table" using the Generate Near Table tool. The near table calculated the distance between each lease polygon and each well, for all lease and well pairs that are within 3 km of each other. The near-table contains 982,091 unique lease-well pairs within 3 km and 469,080 unique lease-well pairs within 2 km.

A.3 Estimation of First Year Production For Each Well

The Unconventional Natural Gas Well Geodatabase published by the Carnegie Museum of Natural History (2022) also provides a .csv file of 17,099 unique wells drilled between 1974 to 2021. Most of the wells are observed multiple times, which a unique row in the data corresponding to an individual production report submitted by the oil and gas operator to the DEP. Each row (report) includes a unique identifier for the well, as well as the duration (in days) of production and quantity of natural gas and oil produced within the production period. Over the period from 1974 to 2021, operators submitted production reports at least annually, and after 2015 the operators submitted reports up to monthly.

We began with 895,383 reports and dropped 209,390 reports that were (1) missing the number of reported production days, (2) missing the quantity of natural gas produced in MCF, (3) associated with wells that have no production report within their first year of production, (3) associated with wells that were plugged, abandoned, proposed but never materialized, and not drilled (as reported by the operator), (4) associated with wells that are in periods of regulatory inactive status but report positive production, because by definition inactive wells cannot report production. (5) associated with wells that are not categorized as "unconventional," (6) reports for which the production quantity was not for an individual well but was averaged over a group of nearby wells, and (7) reports associated with wells drilled before widespread Marcellus production began in 2008.

With the remaining 685,993 reports, we converted the production variables (volume of oil and natural gas) to millions of British thermal units (mmbtu), and then summed the oil and gas production together to obtain a combined production figure. Next, we created a variable capturing the cumulative number of days that the well had been producing from its first day in production through to the last day of the report. We also created a variable that captures the total cumulative lifetime production in mmbtu for each well from its first day in production through to the last day of the report. We ran a separate regression through the origin for each individual well of total cumulative production on a linear term, a squared term, and a cubed term for the total number of days that the well had been producing through to the end of the reporting period. To obtain an estimate of aggregate first year production for each well, we multiplied the estimated coefficients on the linear, squared, and cubed days by 365 days, and added the three products together. We estimated first year production measures this way because for the vast majority of wells we do not observe production on day 365 (e.g., the first production report of a well may cover the first 60 days of a year, while the second production report of that well may cover the entire following calendar year).

Altogether, were able to project first year production for 10,825 wells that meet our seven sample requirements and have more than three reports (which is needed for the regression to run). The first row of table A1 displays descriptive statistics for our first year production estimates for the 10,825 wells. We compared the estimates to those of Harleman (2021) (who uses a nearly identical methodology) and data from the US Energy Information Administration (2021). These sources indicate that in their first year, the typical Marcellus well produces around 1.5 trillion btu in its first year (see Harleman (2021), Online Appendix Page 9), with a maximum of between 4 and 5 trillion btu (see Harleman (2021), Online Appendix Page 20). Based on these benchmarks, the first row of table A1 suggests that our model vastly overestimates production for some wells, which is largely due to measurement error or misreporting of production in the DEP production reports. Therefore, we drop wells for which our model estimates a first year production greater than the 90th percentile of 4.20 trillion btu, and proceed with 9,743 wells.

The second of Table A1 displays the distribution of average FYP for the 9,743 wells. These wells have a mean first year production of 1.45 trillion btu, and median of 1.18 trillion btu, and a maximum of 4.20 trillion btu. These estimates are similar to the estimates found in Harleman (2021) and from the US Energy Information Administration (2021).

Table A1: Descriptive Statistics of Average First Year Production (Well-Level Dataset)

	Mean	SD	Min.	p25	p50	p75	Max	N
Average FYP (trillion btu)	1.97	3.32	0.00	0.68	1.33	2.57	168.52	10,825.00
Average FYP (trillion btu)	1.45	1.05	0.00	0.62	1.18	2.12	4.20	9,743.00

Note: The first row displays the distribution of average FYP all 10,825 wells for which we are able to project FYP. The second row displays the distribution of average FYP after removing the top 10 percent of observations.

A.4 Depth Instrument

Data on the total depth of each well, our instrument variable, comes from the Pennsylvania Department of Conservation and Natural Resources (2022). The DCNR provides data on the depth of by shale play as part of their EDWIN database, which is a subscription service. We downloaded data on wells and their depth for the Marcellus shale play, as well as the Geneseo, Utica, and Rhinestreet plays (which lie above or beneath the Marcellus). EDWIN contains two variables related to depth, "DeepestProducingDepth" and "TotalDepthft." The former is the deepest depth at which the oil and gas driller produces oil or gas, and is missing for 25 percent of the 8,073 wells in our EDWIN dataset. The latter is the total depth to which the well has been drilled, and is non-missing for 99.5 percent of wells in our EDWIN dataset. The two depth variables are highly correlated, with a Pearson correlation coefficient of .9882. Therefore, we use the total depth variable to retain as many wells as possible in our analysis.

We merged the 8,073 wells with depth data in EDWIN with the 9,743 wells for which we can estimate first year production, and are left with 7,336 wells with complete data for both variables which enable our model 3 analysis described in Section 5.

A.5 Merging the Well Data with the Lease-Well Near Table

Next, we merged well records for 7,336 wells with complete estimated first year production and depth data with the near-table created in appendix section . The near table initially contains 982,091 unique lease-well pairs within 3 km and 469,080 unique lease-well pairs within 2 km. When keeping lease-well pairs using only the 7,336 wells with complete data, we retained 725,816 unique lease-well pairs within 3 km and 343,781 unique lease-well pairs within 2 km.

We collapsed the lease-well dataset into a lease-level dataset, taking lease-level averages of all wells within 2 km (or 3 km) radius for the total depth and average first year production variables. We only retain leases that have at least 10 wells within the radius, to limit the effect of outlying depth or first year production estimates biasing the results. For the within 2 kilometer analysis, this yields a sample of 8,445 leases (the last three rows of Table 2 and columns 1 through 6 of Table 6). For the within 3 kilometer analysis, which additionally only attached wells to leases if the well was drilled before the lease was signed, this yields a sample of 1,367 leases (columns 7 and 8 of Table 6).