Moving to Economic Opportunity: The Migration Response to the Fracking Boom

Riley Wilson

Online Appendix

Appendix A. Additional Tables and Figures

Appendix Table A.1. Simulated Production Summary Statistics for Fracking Counties

	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
# Counties with Simulated New Production	368	413	413	489	521	614	631	681	726	720
	Simulated Annual New Production Value by Percentile (Millions 2010$)									
10th	0.02	0.01	0.02	0	0.1	0	0.06	0.03	0.09	0.12
50th	0.66	0.6	0.61	0.75	0.86	0.36	0.62	0.76	0.73	0.92
90th	15.56	18.06	16.91	17.75	24.95	14.82	23.52	27.71	25.88	26.19
	Average Simulated Annual New Production Value (Millions 2010$)									
Mean	7	7.81	7.93	7.9	11.13	7.17	9.79	11.42	11.65	12.88
Mean Among Top 150 Counties	15.05	19.02	19.25	22.94	34.37	26.39	36.74	46.25	50.27	54.81
Mean Among Top 100 Counties	21.71	27.32	27.57	32.87	49.26	37.73	52.38	65.85	71.36	77.7
Mean Among Top 50 Counties	37.97	47.47	47.93	57.04	85.89	65.42	88.8	109.64	119.13	131.86

Notes: The statistics are for all counties with any simulated production in the given year. Simulated production is reported in millions of dollars, deflated to 2010 dollars.

Appendix Table A.2. Reduced Form Impact of Simulated Production on Labor Market Measures by Gender and Education

	Log Average Earnings$_{t-1}$				Log Jobs to Pop. Ratio$_{t-1}$			
	Men		Women		Men		Women	
	No College Degree (1)	College Degree (2)	No College Degree (3)	College Degree (4)	No College Degree (5)	College Degree	No College Degree	College Degree
Sim. New Prod. Value in Cty$_{t-1}$	0.010***	0.006***	0.005***	0.004***	0.020***	0.017***	0.003**	0.003
(10 Millions 2010$)	(0.002)	(0.001)	(0.001)	(0.001)	(0.004)	(0.004)	(0.001)	(0.002)
				Regional Heterogeneity				
Sim. New Prod. Value in Cty$_{t-1}$	0.023***	0.014***	0.013***	0.009***	0.050***	0.042***	0.010***	0.007**
*(10 Millions 2010$)*North Dakota*	(0.003)	(0.002)	(0.001)	(0.001)	(0.006)	(0.007)	(0.002)	(0.003)
Sim. New Prod. Value in Cty$_{t-1}$	0.010***	0.007***	0.005***	0.003***	0.011***	0.010***	0.004**	0.003
*(10 Millions 2010$)*West*	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.004)	(0.002)	(0.003)
Sim. New Prod. Value in Cty$_{t-1}$	0.005***	0.002*	0.001	0.002**	0.010**	0.010*	-0.001	0.002
*(10 Millions 2010$)*South*	(0.002)	(0.001)	(0.001)	(0.001)	(0.004)	(0.005)	(0.001)	(0.004)
Sim. New Prod. Value in Cty$_{t-1}$	0.147***	0.062***	0.036***	0.003	0.174***	0.130***	0.012	-0.005
(10 Millions 2010$) Northeast*	(0.018)	(0.022)	(0.013)	(0.018)	(0.035)	(0.027)	(0.032)	(0.033)
Sim. New Prod. Value in Cty$_{t-1}$	0.210**	0.196	-0.013	-0.135**	-0.093	0.125	0.067	0.160
(10 Millions 2010$) Midwest*	(0.106)	(0.131)	(0.045)	(0.060)	(0.123)	(0.261)	(0.108)	(0.188)
Dependent Mean	37,055	60,556	23,300	37,065	0.544	0.667	0.553	0.649
Observations	31,094	31,157	31,062	31,157	31,094	31,157	31,062	31,157

Notes: Data compiled from the QWI, ACS, and DrillingInfo. Each column in each panel is a separate regression. Observation at the county by year level. All regressions include county and state by year fixed effects, which make this a comparison between counties in the same state. Standard errors are corrected for clustering at the county level. p<0.01 ***, p<0.05 **, p<0.1 *.

Appendix Table A.3. Falsification Test: Effect of Simulated Production on Migration Year T-10

	Number of In-Migrants, as Percent of 2000 Population		
Year of migration outcome	Current (1)	Current (2)	10 Years Prior (3)
Sim. New Prod. Value in Cty$_{t-1}$	0.30***	0.32***	0.01
(10 Millions 2010$)	(0.09)	(0.10)	(0.01)
	Regional Heterogeneity		
Sim. New Prod. Value in Cty$_{t-1}$	0.95***	0.96***	0.03*
*(10 Millions 2010$)*North Dakota*	(0.06)	(0.07)	(0.02)
Sim. New Prod. Value in Cty$_{t-1}$	0.21***	0.16**	0.01
*(10 Millions 2010$)*West*	(0.05)	(0.07)	(0.03)
Sim. New Prod. Value in Cty$_{t-1}$	0.06***	0.07***	0.00
*(10 Millions 2010$)*South*	(0.01)	(0.01)	(0.01)
Sim. New Prod. Value in Cty$_{t-1}$	0.48***	0.48***	-0.47***
(10 Millions 2010$) Northeast*	(0.13)	(0.11)	(0.15)
Sim. New Prod. Value in Cty$_{t-1}$	0.38	0.37	-0.19
(10 Millions 2010$) Midwest*	(0.64)	(0.43)	(0.52)
Years in Sample	1999-2013	2005-2013	2005-2013
Dependent Mean	5.17	5.20	5.07
Observations	31,157	20,850	20,840

Notes: Migration data from IRS SOI, and simulated production constructed from DrillingInfo. Analysis at the county by year level. In the bottom panel, simulated production is interacted with a binary indicator for each of the five regions: North Dakota, West, South, Northeast, and the Midwest. The impact across regions are estimated jointly. All regressions include county and state by year fixed effects, which make this a comparison between counties in the same state. Column (1) presents the baseline estimates from Column (2) of Table 3. Column (2) restricts the baseline estimate to the same years available for the falsification test. Column (3) assigns the annual in-migration rate for the period ten years earlier. For example, this regression considers how simulated production in 2013 affects migration in 2003. Standard errors are corrected for clustering at the county level. p<0.01 ***, p<0.05 **, p<0.1 *.

Appendix Table A.4. State Specific Migration Response to Earnings

	Outcome: Number of In-migrants as a Percent of 2000 Population					
Labor Market Measure	Reduced Form: Sim. Prod.	Average Earnings	Average Earnings Controlling for Housing Price	Housing Adjusted Earnings	Jobs to Population Ratio	Average Earnings per capita
	(1)	(2)	(3)	(4)	(5)	(6)
Log Measure$_{t-1}$	0.95***	38.02***	40.35***	35.03***	32.96***	17.54***
	(0.06)	(5.82)	(6.32)	(5.25)	(5.40)	(2.69)
			Western States			
*Log Measure$_{t-1}$*MT*	-0.60***	-8.53	-10.76	-9.11	7.95	-0.61
	(0.06)	(6.21)	(6.68)	(5.62)	(7.76)	(3.04)
*Log Measure$_{t-1}$*NM*	-0.75***	-27.27***	-29.01***	-25.59***	-17.51***	-11.17***
	(0.08)	(5.97)	(6.47)	(5.38)	(6.68)	(2.86)
*Log Measure$_{t-1}$*CO*	-0.87***	-7.73	-3.66	-11.65	-362.4	15.77
	(0.08)	(24.74)	(20.80)	(14.69)	(3,469)	(60.96)
*Log Measure$_{t-1}$*CA*	-0.88***	-22.63	-25.13	-47.95***	-19.73*	-10.46
	(0.07)	(20.19)	(20.46)	(18.36)	(10.53)	(6.58)
			Southern States			
*Log Measure$_{t-1}$*TX*	-0.88***	-17.96	-18.64	-16.30	-15.91	-8.41
	(0.06)	(12.88)	(14.67)	(11.97)	(16.33)	(7.02)
*Log Measure$_{t-1}$*OK*	-0.93***	-31.91***	-33.88***	-29.38***	-26.52*	-14.43**
	(0.06)	(11.60)	(12.87)	(10.64)	(15.87)	(6.34)
*Log Measure$_{t-1}$*AR*	-0.90***	-34.08***	-36.50***	-31.71***	-17.83	-14.42***
	(0.06)	(6.12)	(6.59)	(5.49)	(18.64)	(2.87)
*Log Measure$_{t-1}$*LA*	-0.93***	-26.33	-28.70	-23.37	-43.99**	-1,047
	(0.06)	(25.62)	(27.85)	(27.79)	(21.96)	(143,500)
			Northeastern States			
*Log Measure$_{t-1}$*PA*	-0.42***	-32.81***	-35.01***	-30.25***	-27.86***	-14.98***
	(0.14)	(6.06)	(6.57)	(5.50)	(5.74)	(2.83)
			Other States			
*Log Measure$_{t-1}$*Other*	-0.79***	-12.53*	-14.58**	-16.60***	-3.73	-3.96
	(0.06)	(6.40)	(6.80)	(6.05)	(7.42)	(3.23)
Independent Mean	0.178	34,516	34,516	28,688	0.538	19,363
Observations	31,143	31,143	31,143	31,141	31,143	31,143

Notes: Data compiled from the IRS SOI, QWI, Federal Housing Finance Agency, and DrillingInfo. Each column is a separate regression. The direct effect of log average earnings represent the impact for North Dakota, and all interactions are deviations from this base. In column (2), I directly control for log housing prices. In column (3) earnings are adjusted to account for differences in housing prices following the method of Ganong and Shoag (2017). All regressions include county and state by year fixed effects, which make this a comparison between counties in the same state. Standard errors are corrected for clustering at the county level. p<0.01 ***, p<0.05 **, p<0.1 *.

Appendix Table A.5. Reduced Form Effect of Simulated Production on Housing Prices

		Characteristic		
	Baseline	Share Vacant in 2000	Geography Constraint	Share Own Water in 2000
	(1)	(2)	(3)	(4)
Sim. Prod. Value in Cty$_{t-1}$	0.004***	0.002**	0.005***	0.003***
*(10 Millions 2010$)*North Dakota*	(0.001)	(0.001)	(0.001)	(0.001)
Sim. Prod. Value in Cty$_{t-1}$	0.001	0.002	0.002	0.002
*(10 Millions 2010$)*West*	(0.001)	(0.003)	(0.003)	(0.003)
Sim. Prod. Value in Cty$_{t-1}$	0.001	-0.005**	-0.001	0.001
*(10 Millions 2010$)*South*	(0.001)	(0.002)	(0.002)	(0.001)
Sim. Prod. Value in Cty$_{t-1}$	0.030**	0.027*	0.013	0.027*
*(10 Millions 2010$)*Northeast*	(0.015)	(0.015)	(0.017)	(0.015)
Sim. Prod. Value in Cty$_{t-1}$	0.036	-0.003	-0.033	0.033
*(10 Millions 2010$)*Midwest*	(0.065)	(0.072)	(0.084)	(0.066)
Sim. Prod. Value in Cty$_{t-1}$		0.007***	0.001	0.002**
**North Dakota*Characteristic*		(0.002)	(0.001)	(0.001)
Sim. Prod. Value in Cty$_{t-1}$		-0.004	0.001	-0.003
**West*Characteristic*		(0.005)	(0.001)	(0.004)
Sim. Prod. Value in Cty$_{t-1}$		0.066***	0.007	0.005
**South*Characteristic*		(0.021)	(0.006)	(0.010)
Sim. Prod. Value in Cty$_{t-1}$		0.127*	-0.090*	2.665**
**Northeast*Characteristic*		(0.074)	(0.050)	(1.277)
Sim. Prod. Value in Cty$_{t-1}$		0.546	-0.175	0.041*
**Midwest*Characteristic*		(0.397)	(0.110)	(0.023)
F-statistic	4.764	3.972	3.345	3.625
Observations	31,157	31,155	31,155	31,155

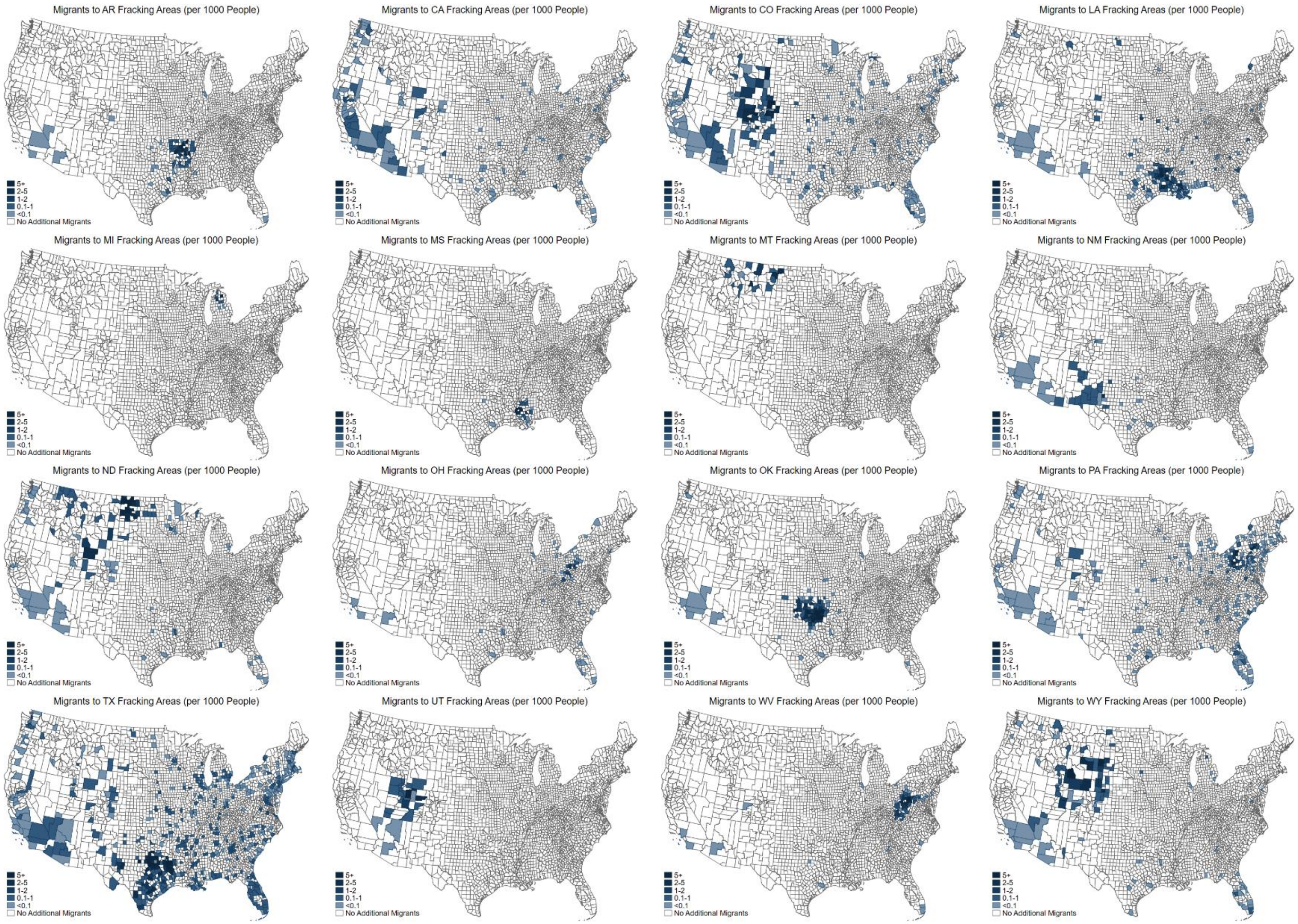
Notes: Housing price constructed from the housing price index provided by the Federal Housing Finance Agency and converted to dollars using county median house prices in 2000. Simulated production is interacted with a binary indicator for each of the five regions. The impact across regions are estimated jointly, to test for differences. In columns (2) through (4) region specific production is then interacted with various characteristics prior to the boom that could possibly affect pricing but otherwise be exogenous to migration. All regressions include county and state by year fixed effects, to control for time invariant county characteristics as well as state specific shocks, making this a comparison between counties in the same state. Standard errors are corrected for clustering at the county level. $p<0.01$ ***, $p<0.05$ **, $p<0.1$ *.

Appendix Table A.6. Robustness of Regional Migration Elasticities

	Outcome: Number of In-migrants as a Percent of 2000 Population							
Specification:	Baseline (1)	Weighted by 2000 Population (2)	Shorter Sample (≤2011) (3)	Actual Prod. as Instrument (4)	Play by Year Interacts as Instruments (5)	Sim. New Wells as Instrument (6)	Sim. Prod. Per Capita as Instrument (7)	Control for Median 2 Bedroom Rent (8)
Log Average Earnings$_{t-1}$	38.02***	37.14***	28.69***	40.81***	36.51***	35.45***	31.06***	37.66***
**North Dakota*	(5.82)	(3.09)	(1.78)	(7.14)	(6.37)	(5.85)	(7.58)	(6.70)
Log Average Earnings$_{t-1}$	24.20***	20.19	20.14***	21.44***	0.74	19.05**	24.85***	16.80**
**West*	(3.81)	(30.82)	(4.08)	(4.48)	(2.15)	(7.76)	(2.65)	(6.56)
Log Average Earnings$_{t-1}$	15.67**	8.77	17.17	10.83*	2.86	14.47**	10.42**	12.04***
**South*	(7.14)	(14.00)	(15.35)	(6.40)	(1.88)	(7.24)	(5.01)	(4.45)
Log Average Earnings$_{t-1}$	4.71***	3.17*	17.01**	5.31***	5.03	4.33***	2.02	5.13***
**Northeast*	(1.61)	(1.63)	(7.59)	(1.78)	(3.62)	(1.58)	(1.73)	(1.67)
Log Average Earnings$_{t-1}$	3.65	-6.99	28.60	-5.89	3.62**	17.76	3.14	10.28
**Midwest*	(7.04)	(22.85)	(44.64)	(6.00)	(1.81)	(33.54)	(5.90)	(13.28)
P-values:								
North Dakota equals West	0.05	0.58	0.05	0.02	<0.01	0.09	0.44	0.03
North Dakota equals South	0.02	0.05	0.46	<0.01	<0.01	0.02	0.02	<0.01
North Dakota equals Northeast	<0.01	<0.01	0.13	<0.01	<0.01	<0.01	<0.01	<0.01
North Dakota equals Midwest	<0.01	0.06	0.99	<0.01	<0.01	0.60	<0.01	0.07
Observations	31,157	31,157	26,533	31,157	31,157	31,157	31,155	26,249

Notes: Data compiled from the IRS SOI, QWI, and DrillingInfo. Each column is modified as specified. All regressions include county fixed effects. All regressions include state by year fixed effects, to control for time invariant county characteristics as well as state specific shocks, making this a comparison between counties in the same state. Standard errors are corrected for clustering at the county level. $p<0.01$ ***, $p<0.05$ **, $p<0.1$ *

Appendix Figure A.1. Additional Migrants to Fracking Areas by Origin County



Notes: For each fracking state, the average annual number of migrants from 2008-2012 minus the average annual number of migrants from 2000-2003 per 1,000 people at the origin county is plotted. The number of migrants is aggregated over all fracking counties in the state, meaning the migration flows to Texas are capturing the flows to more counties than the flows to other states.

Appendix B. Data Appendix

Below I describe each of the key datasets used in my analysis, as well as important characteristics of data construction

I Internal Revenue Service Statistics of Income County Flows

The Internal Revenue Service (IRS) Statistics of Income (SOI) division provides annual counts of county-to-county flows. This provides the raw number of tax returns and exemptions that were filed in one county in year $t-1$ and in another county in year t. Each year, the IRS provides county-to-county flows of exemptions in a file with two years (e.g., 2002to2003). This represents exemptions that were in one county when filing in 2002 and in another county when filing in 2003. As most people file in the beginning of the year before April, I assign this flow to the year 2002.

Using exemptions to approximate people in a household, I collapse each county, year to a single observation of the total number of exemptions.[1] The in-migration rate can be constructed by dividing the number of exemptions by the county population. Throughout my analysis, I divide exemptions by the baseline county population in 2000, in order to provide a common base across all years. Unfortunately, the IRS county to county flows only provide aggregate numbers, and do not break up the migration levels by demographic characteristics (gender, marital status, education). As such, I am unable to use the IRS measures to look at differences across demographics. The only measure provided is the total adjusted gross income for all of the moved-

[1] The IRS censors county pairs that have fewer than ten returns move in each year. However, all of these returns are listed in a separate category as “from same state” or “from different state”. As such, when I collapse to the county level, I will capture the total number of returns, regardless of where they originated.

returns. This is the adjusted gross income in the earlier year, but only the average for all movers in the county pair is provided.

The IRS data does not capture every move from one county to another. Low income households are not required to file a tax return, and thus might be under represented in the data. It is likely that individuals that move to fracking areas will earn well beyond the filing threshold after moving, but they might not have been required to file in the previous year. If there are individuals that did not file in the first year, but moved in response to fracking and filed in the second year, my estimates would be attenuated. In order for the gap across geography to be biased upward, these individuals would have to be sorting into North Dakota. This systematic sorting would provide further evidence that people responded differently to the fracking boom in North Dakota.

The IRS data also does not capture temporary moves. Individuals who moved after filing in year t, but move back before filing in t+1 will not be counted as a move. Anecdotal evidence suggests that there was also large scale short-term relocation in North Dakota. My estimates will not fully capture this, but rather capture long-term adjustments. This measure likely seems more relevant when considering economic mobility, although it would be useful to test and see if individuals are responding by short term relocation rather than long term moving.

II American Community Survey

To explore demographic differences and understand who moves, I use the American Community Survey (ACS) between 2005 and 2011. The ACS is an annual survey ran by the Census Bureau of approximately a one percent sample of households and has replaced the Census long for. All participants are asked where they lived one year ago, and both the previous state and local migration public use microdata area (MIGPUMA) are recorded. These MIGPUMA usually correspond to PUMA, but are enlarged to encompass the entire county. For rural areas MIGPUMA can often cover multiple counties or large portions of the state. When looking at fracking regions this can be problematic, as the MIGPUMA covering fracking areas also cover many surrounding counties. I identify the fracking status of a MIGPUMA, by simply indicating if it has any county with simulated production in it. I also do this separately for different plays (Bakken region) to look at heterogeneity within fracking. To the extent that I am capturing untreated areas as well, this will attenuate my estimates toward zero. Unfortunately, the boundaries for MIGPUMA changed in 2012. In many of the states the uniquely identifiable areas between 2005 and 2013 encompasses most of the state. For this reason I choose to focus on the ACS from 2005 to 2011. As such, I am not able to capture demographic characteristics in the later years, which might be important given the steep rise in North Dakota.

In all of my estimation using the ACS microdata, I collapse my observations from the individual level to unique cells. These cells are defined by demographics (e.g., gender, marital status, race, age group, education), migration status, fracking destination, and state of previous residence. When collapsing to these cells, I sum the individual weights provided by the Census Bureau and then use these weights in my regression analysis. These estimates are population representative and are identical to estimates obtained using weights at the individual level. Unfortunately, the migration questions from the 2000 Census ask about migration in the previous 5 years, and are thus not comparable to migration in the ACS.

III U.S. Census Bureau Quarterly Workforce Indicators

The Quarterly Workforce Indicators (QWI) are constructed by the Census from the Longitudinal-Employer Household Dynamics (LEHD) Program and use firm level employment to construct aggregate employment and earnings reports. The QWI is aggregated from the

Longitudinal Employer-Household Dynamics (LEHD) micro-level data collected from unemployment insurance earnings data from participating states and several other sources.[2] The QWI is aggregated to the county level, and can be tabulated by firm characteristics (industry, size) or worker characteristics (gender, age, education).[3] When tabulating by worker characteristics, only two levels of tabulation are feasible (gender by age or gender by education). The educational attainment measures in the LEHD are imputed based on a state-specific logistic regression among individuals from the 2010 Census long form and uses individual measures (such as age, earnings, and industry) to predict education level (LEHD, 2005). Because of this potential measurement error, I focus on earnings and employment across the entire population in my main analysis, rather than separately by education. The QWI data is constructed through a state sharing process, and as such, only states that have made agreements with the Census have reported data. Many of the states began participating in 2000 with most participating by 2003. As such, some states and counties are missing wage information in the early years. Most of these were not involved in fracking.

The main measure I use is the beginning of quarter earnings for all jobs. This measures the quarterly earnings for all jobs that existed at the beginning of the quarter. I choose this measure rather than stable jobs (spanning multiple quarters) and total jobs (employed at any time during quarter). I take the implied average annual earnings across all four quarters weighting by the quarter specific employment to construct the group specific average earnings for each year.

Because the QWI is constructed from firm employment, all measures are constructed for the job count. This means that average quarterly earnings are the average earnings of all jobs in a

[2] Most states began participating prior to 2000. However, during the years of the fracking boom South Dakota and Massachusetts did not participate in the data submission.
[3] I take the implied average annual wage across all four quarters weighting by the quarter specific employment to construct the group specific average wage for each year.

given quarter. Individuals who are unemployed are not considered, and individuals who hold two jobs will be treated as two separate individuals. In general, average earnings levels in the QWI are higher than those calculated elsewhere, as it records average earnings conditional on working. Also, because some workers might hold jobs for less than the full year, the average annual earnings constructed from the QWI will be higher, because my construction implicitly assumes the job lasts the entire year. This measure of earnings can be interpreted as the potential earnings if an individual was to move to the region.

IV DrillingInfo Well Database

Well level information on drilling date, lease agreements, location, direction, and geological formation as well as other characteristics are provided through a restricted use data agreement from DrillingInfo. This data is proprietary, and obtained through an academic use agreement with DrillingInfo, available through their academic outreach initiative. These well level characteristics are then merged to well level quarterly oil and gas production reports also provided by drilling info. Oil and gas production are reported in barrels and thousands of cubic feet respectively. Using the annual West Texas Intermediate crude oil price and the Henry Hub Natural Gas national prices provided by the Energy Information Administration (EIA), I convert these into dollar amounts and deflate to 2010 dollars.

DrillingInfo does not indicate if a well is a fracking well, as fracking is a means of stimulating production. To infer wells that are affected by the technological innovation associated with fracking, I use details on drilling direction and well location. Localized fracking booms occurred in part because of the combination of horizontal (directional) drilling and hydraulic fracturing. The DrillingInfo data reports whether a well is horizontally or vertically drilled. In addition, fracking was particularly impactful over shale plays, as these resources were not extractable previously. For this reason I assign non-vertical wells drilled in counties that intersect with shale plays as fracking wells.

V Shale Play Boundary Shapefiles

Shale play boundary shapefiles are provided by the EIA in order to map the estimated boundaries of shale formations. These shapefiles have been updated over the years as new formations and reserves have been discovered. Prior to the shale boom, these formations had not be systematically mapped because they did not have economic value. I use the latest shapefile available at the time from 2015 to map shale play boundaries. These shapefiles are then overlaid by county shapefiles provided by the U.S. Census Bureau, and with the help of two research assistants I calculate the area of shale play and county intersections. This intersection measure is used when simulating production.

VI Housing Price Index

The Housing Price Index is constructed by the Federal Housing Finance Agency at the three digit zip code. Three digit zip codes span the entire country, allowing me to construct a measure for rural counties. To construct the county level measure I assign each county the average housing price index of all three digit zip codes that intersect the county, weighted by the share of the county in that zip code. For some three digit zip codes there is insufficient data, so the zip code is assigned the index from a larger geographic unit (such as the MSA or the state). I then adjust the housing price index baseline to be equal to 100 in 2000. Using the county level median house value from the 2000 Census, I convert the housing price index to dollars. A similar developmental index is available at the county level but does not include all counties. I find that both indices follow similar patterns for the available counties.