ISSN 1936-5330

THE FEDERAL RESERVE BANK of KANSAS CITY
RESEARCH WORKING PAPERS

Location Decisions of Natural Gas Extraction Establishments: A Smooth Transition Count Model Approach

Jason P. Brown and Dayton M. Lambert

April 2014
RWP 14-05

Location Decisions of Natural Gas Extraction Establishments: A Smooth Transition Count Model Approach

Jason P. Brown[1] and Dayton M. Lambert[2]

[1,*]Corresponding author, Federal Reserve Bank of Kansas City
E-mail: jason.brown@kc.frb.org

[2]University of Tennessee Institute of Agriculture
Department of Agricultural & Resource Economics
E-mail: dlamber1@utk.edu

Acknowledgements

The views expressed here are those of the authors' and do not represent the views of the Federal Reserve Bank of Kansas City, the Federal Reserve System or the University of Tennessee. The United States Department of Agriculture Hatch Project NE–1049 partially supported this research. We wish to thank seminar participants at the North American Regional Science Council, the Southern Regional Science Association, the Federal Reserve Bank of Chicago and the Federal Reserve Bank of Kansas City for helpful comments. All remaining errors or omissions are our own.

Abstract

The economic geography of the United States' energy landscape changed rapidly with domestic expansion of the natural gas sector. Recent work with smooth transition parameter models is extended to an establishment location model estimated using Poisson regression to test whether expansion of this sector, as evidenced by firm location decisions from 2005 to 2010, is characterized by different growth regimes. Results suggest business establishment growth of firms engaged in natural gas extraction was faster when the average area of shale and tight gas transition coverage in neighboring counties exceeded 17%. Local agglomeration externalities, access to skilled labor and transportation infrastructure were of more economic importance to location decisions in the high growth regime. Accordingly, growth rates were heterogeneous across the lower 48 States, suggesting potentially different outcomes with respect to local investment decisions supporting this sector.

JEL codes: C21, C25, D21, R12, R30

Keywords: natural gas extraction, location choice, count model, endogenous growth regimes, spatially varying parameters

1. Introduction

Technological advances in hydraulic fracturing and horizontal drilling accelerated the expansion of natural gas production in several regions of the United States. According to the Energy Information Agency (EIA, 2013), the total U.S. recoverable natural gas resources were estimated to be 2,327 trillion cubic feet. This quantity represents an estimated 70 years' worth of supply taking into account the projected annual growth in domestic natural gas consumption. The natural gas supply shock reversed a several decades long downward trend in U.S. natural gas production. In the 1970s the U.S. energy sector seemingly conceded its decline and began investing in global markets. That trajectory reversed in the middle of the last decade.

In the mid-2000s natural gas production increased dramatically in shale and tight gas formations. In recent years, domestic natural gas prices have declined because of increased supply and lack of infrastructure to transport and export to global markets, creating a gap between domestic U.S. and world prices. Natural gas prices have fallen dramatically from $12.76 to $1.95 per million British thermal units since recent peak drilling activity. As a result, the number of active drilling rigs and production has also decreased. Extraction is expected to increase once more favorable prices return. We use the recent period of increased natural gas production to model the location decisions of natural gas extraction establishments at the county level in the lower contiguous 48 states.

We extend recent developments of smooth transition parameter models to an establishment location model estimated using Poisson regression to test whether growth in this sector, as evidenced by extraction establishment growth from 2005 to 2010, is characterized by high and low growth regimes according to the shale resource

endowments of counties. We develop a parsimonious count model that provides a tractable interpretation of parameters specific to spatial units according to endogenous growth regimes. Results suggest that counties with a locally weighted average of more than 17% coverage by shale and tight gas formations transition to a high establishment growth regime where the rate of extraction establishment growth is higher. Local agglomeration externalities, access to skilled labor, and transportation infrastructure appear to be of more economic importance to location decisions in the high growth regime than in other counties lacking comparative advantage in terms of shale and tight gas formation endowments.

The presence of shale or tight gas formations in an area does not guarantee that the region will, ceteris paribus, attract business establishments engaged in the production and distribution of natural gas. However, if the shale and tight gas areas cover a large enough area in a given region, that region may be more likely to attract firms engaged in the natural gas economy. The medium or long term establishment growth trajectories of counties with relatively greater resource endowments, notwithstanding shale and tight gas formations, may be quantitatively different than trajectories characterizing other administrative units, given their resource endowments.

Our contribution to the literature is to provide a method that estimates the threshold level of coverage of shale and tight gas formations where the probability of attracting a business engaged in the natural gas extraction sector is more likely and how being above or below that threshold changes the influence local endowments have on the site selection choice. The factors causing local variation in business establishment growth trajectories may be relevant to communities for planning short to medium term projects

that could impact the allocation of limited financial resources for attracting business ventures that could be spent elsewhere, or decisions that could alter the natural landscape of a region which could affect the non-market amenity value or resource base of a local community.

2. Background of the Shale-Gas Industry

Technologies pursued initially by two independent energy companies, but that were eventually combined, transformed the oil and gas industry. In the early 1980s Mitchell Energy drilled the first well in the Barnett Shale in western Texas (Yergin, 2011). Instead of encountering the typical, highly porous, rock of conventional formations, Mitchell encountered shale. Shale can hold vast reserves of natural gas but is highly nonporous and traps gas. Over a period of 20 years, Mitchell Energy experimented with different extraction techniques and found that by using hydraulic fracturing (commonly referred to as “fracking”) shale layers could be broken to release natural gas. Fracking consists of injecting a mixture of water, chemicals, and sand into wells to create fissures in rock formations, liberating trapped gases.

Over the same period, Devon Energy developed horizontal drilling technologies. Advances in controls and measurement allowed operators to drill to certain depths, and then drill further at angles to expose more of reservoirs, allowing much greater recovery of natural gas. In 2002 Devon acquired Mitchell Energy (Yergin, 2011). Devon combined their horizontal drilling expertise with Mitchell’s fracking techniques. By 2003 Devon discovered a combination of the two technologies that proved successful. Suddenly, natural gas that had been economically inaccessible was now exploitable.

Higher natural gas prices in the mid-2000s and the combination of horizontal drilling with fracking changed the economics of natural gas production. New reserves from unconventional shale and tight gas formations became profitable to extract, and continued development of drilling and hydraulic fracturing techniques enhanced further production efficiencies. Today, shale wells have an extremely low risk of being unproductive (unproductive wells are commonly referred to as “dry-holes”). Prior to the advent of shale gas, total annual natural gas production in the U.S. was flat; at about 19 to 20 trillion cubic feet (figure 1). However, by 2011 total annual production grew nearly 30 percent to 24.6 trillion cubic feet. Over the same period, the amount of proven reserves continued to increase as exploration intensified.

<< Figure 1 >>

Gas well profitability varies from formation to formation and depends on geological attributes that determine how quickly production declines after opening a well (Massachusetts Institute of Technology, 2011). Since the first major shale boom in the Barnett (TX), additional large-scale natural gas extraction has occurred in other shale formations including the Woodford (OK), Fayetteville (AR), Haynesville (LA and TX), Marcellus (PA and WV), and Eagle Ford (TX) plays. Activity has also increased in the Niobrara shale, which extends across portions of Colorado, Kansas, Nebraska, and Wyoming. We explain the pattern of new extraction businesses entering the natural gas extraction and distribution market using location theory models originally developed by Weber (1929), used extensively in numerous applied studies that explain business

establishment location as a function of local comparative advantage, resource availability, and input and product market access.

3. Firm Location Theory

Optimal firm site selection is a trade-off between input transport costs from extraction sites to consumers (Weber, 1929). Firms choose least cost sites to maximize profit (π). Holding transportation costs and other local factors constant, firm-level profits relying extensively on natural resource extraction are more sensitive to the productivity of the site selected, which is typically unknown. The cost structure of the natural gas extraction sector is therefore similar to supply-oriented food processors to the extent that locating near raw materials naturally minimizes cost (Connor and Schiek, 1997). Supply oriented firms have a total industry cost structure dominated by the purchase of a single input, and therefore prefer to locate near raw materials to minimize input procurement costs (Lambert and McNamara, 2009).

The probability firm selects location *i* can be estimated with a conditional logit regression, assuming the stochastic components follow a Weibull distribution and are independent and identically distributed (McFadden, 1974). As an example, a firm chooses site *i* over *n* possible locations when the expected profits associated with site *i* exceed those of *j*; $E[\pi_i] > E[\pi_j]$. Consequently, the probability a firm chooses site *i* is $\Pr(\pi_i > \pi_j) = \Pr(s_i = 1)$. The empirical analysis simplifies greatly when the analysis focuses on a single sector. In this special case locations decisions are typically modeled using standard count regression approaches such as the Poisson or negative binomial (Guimarães, Figueiredo, and Woodward, 2003; 2004). This convention provides the conceptual

framework for a number of empirical studies linking firm location events to location-specific human capital, social, infrastructure, and natural resource endowments (e.g., Fotopoulos and Louri, 2000; Henderson and McNamara, 2000; Guimarães, Figueiredo, and Woodward, 2004; Carod and Antolín, 2004; Davis and Schluter, 2005; Carod, 2005; Chong, 2006; Lambert, Garret, and McNamara, 2006a,b; Lambert and McNamara, 2009; Lambert, Brown, and Florax, 2010). We extend this conventional modeling approach to explore the possibility that site selection determinants of business establishments engaged in the natural gas extraction and provision economy may exhibit (1) spatial heterogeneity as a function of endogenous growth regimes, (2) that the marginal effects of covariates on site selection decisions may vary across spatial units according to regimes, and (3) that regime membership is conditional on local resource endowments, namely the abundance of shale and tight gas formations.

4. Data and Location determinants

The variable descriptions, data sources, and descriptive statistics are shown in Table 1. County Business Pattern (CBP) data measure firm site selection activity (n = 3,078 counties) from 2005 to 2010. Gas extraction establishment location events were measured by counting the number of new extraction establishments following the method used by Davis and Schluter (2005). The positive cumulative change in the number of an establishment type over the sample period in a given county i was enumerated as $s_i = \Sigma_t s_{i,t}$, where $s_{i,t} = C_{i,t} - C_{i,t-1}$ if $C_{i,t} > C_{i,t-1}$, 0 otherwise (C is the number of observed establishments in period t). This measure of gross establishment entry may underestimate the actual number of establishments entering the sector because it is not possible to

identify exiting firms in the annual net count provided by the CBP data (Lambert and McNamara, 2009).

<< Table 1 >>

Unlike prior studies examining firm location decisions, less is known about the local factors determining the site selection decisions of natural gas extraction establishments. Given the recent surge in natural gas production beginning in the mid-2000s, it is likely that the earliest extraction companies selected sites by weighting more heavily the distribution of unconventional (shale and tight) gas formations in a particular region in their decision calculus. Using spatial information provided by the Energy Information Agency (EIA), we calculated the share of a county covered by shale and tight gas formations (figure 2). Counties endowed with more shale and tight gas formations are naturally expected to attract more gas extraction establishments (figure 3). In fact, these counties likely exhibited faster growth in the natural gas extraction sector early in the natural gas rush than counties endowed with less shale and tight gas formations. Similarly, historical production, as measured by natural gas production (billions of cubic feet) in 2000, is expected to be a reasonable site selection predictor as entering firms observe the past success of others.[1]

Infrastructure related to the shipment and storage of natural gas is likely to affect the profitability and therefore firm location choice. Transportation costs increase as distance to pipelines increases, possibly requiring the installation of additional pipelines

[1] County-level gas production data were collected from state agencies and compiled by the USDA Economic Research Service.

before extraction begins. Similarly, proximity to storage facilities at distribution or processing points is expected to affect prices as well as the variation in prices received by extraction firms. A priori, longer distances to storage facilities are expected to be negatively associated with location activity. Using ArcGIS, spatial data on the location of natural gas pipelines and storage facilities from the EIA were used to calculate the distance to the closest pipelines and storage facilities for each county.[2]

<< Figure 2 >>

<< Figure 3 >>

Agglomeration economies are typically the most studied determinant of firm location (Carod, Solis, and Antolin, 2010). There is general agreement that agglomeration economies have a positive impact on firm location decisions via knowledge spillovers between businesses in similar market conditions when groups of firms in the same sector locate near each other (Glaeser and Kohlhase, 2004). For example, natural gas extraction companies may benefit from a common pool of labor skilled in specific drilling techniques. We measure this specialized labor pool by the share of mining employment in the county. We also control for local industrial structure by using the shares of manufacturing, construction, and agricultural employment in the county.

The effect of urban areas and potential urban externalities on resource extraction establishments is ambiguous. On one hand, proximity to demand markets reduces input transportation costs used in production as well as final output prices. However, when it

[2] An ArcGIS file of natural gas pipelines was provided by EIA upon request. Storage capacity data are available at: http://www.eia.gov/naturalgas/storagecapacity/.

comes to natural resources extraction and the effects activities may have on local the appearance and quality of local amenities, environmental issues may be of concern in relatively densely populated areas. Moreover, more densely populated areas may also reflect higher land values and rental rates making resource extraction relatively more expensive. In the broader location literature, urbanization economies also tend to be positively correlated with firm location decisions (Head, Reis, and Swenson, 1995; Guimaraes, Figueiredo, and Woodward, 2000). We control for urban influence using county population density.

Market demand is also an important firm location determinant (Blair and Premus, 1987; List, 2001; Gabe, 2003; Guimaraes, Figueiredo, and Woodward, 2004). Proximity to demand markets reduces transportation costs of final products. Most of the natural gas extracted in the U.S. is consumed domestically; therefore, extraction establishments are assumed to find higher demand for gas from businesses and households in higher income locations. County median household income measures the effect of market demand on firm location decisions.

Labor availability is an important factor in firm location decisions. Areas with slack in the labor force may make it easier for firms to hire workers. At the same time, firms are more likely to find skilled workers in locations with higher levels of educational attainment (Woodward, 1992; Coughlin and Segev, 2000). These two factors are measured by the unemployment rate in the county and the percentage of the population with a high school or associate degree and those with bachelor’s degree or higher, respectively.

Transportation infrastructure contributes to regional economic development by establishing and maintaining connections to other regions. Transportation infrastructure is important for some manufacturing location decisions (Smith and Florida, 1994; Luker, 1998; List, 2001). Natural gas extraction using hydraulic fracturing is truck intensive because water and other drilling equipment must be transported to extraction sites. Areas with better transportation infrastructure are expected have relatively lower input sourcing costs. The total miles of interstate highway in a county measures the impact of transportation infrastructure on firm location decisions.

Policies related to natural gas drilling as well as lease and royalty payments vary across states. Some states have been proactive in attracting natural gas development, while others have placed a ban on drilling, such as New York. State fixed effects are included in the model to control for these differences.

5. Poisson Regression with a Smooth Growth Regime Transition

We hypothesize that growth in the natural gas extraction sector is characterized by high and low growth regimes, with counties on the margin of shale-rich regions exhibiting lower entry rates over time. As such, the local determinants correlated with site selection decisions may exhibit different magnitudes of importance and county-specific heterogeneity as determined by growth regimes.

Recent attention has focused on the challenging issue of incorporating spatial processes into discrete regression models. Examples of approaches that incorporate spatial aspects into count model regression analyses are conditional autoregressive models (Rasmussen, 2004), spatial general linear models (Gotway and Stroup, 1997),

geographically weighted regression (Schabenberger and Pierce, 2002), information theoretic approaches (Bhati, 2005), Bayesian hierarchical methods (Banerjee, Carlin, and Gelfand, 2004; LeSage and Fisher, 2012), and Poisson models with spatially lagged dependent variables (Lambert, Brown, and Florax, 2010). This research extends the Poisson count model to the family of smooth transition models with endogenous spatial regimes developed by Pede (2010), Pede, Florax, and Holt (2009), and Lambert, Xu, and Florax (2013).

The Poisson probability mass function for s random location decisions at location i is,

(1) $$f(s_i) = \frac{\mu_i^{s_i} \exp(-\mu_i)}{s_i!},$$

with μ_i the expected value that site i is selected by s firms in the natural gas extraction sector over some period. The expected conditional mean is typically represented by the inverse of the logarithmic canonical link function (Cameron and Trivedi, 1998);

(2) $$\mu_i = \exp\left(\beta' x_i\right),$$

where x_i is a $k \times 1$ vector of covariates containing measurements on observation $i = 1,\ldots, N$ including a constant.

Maximizing the log-likelihood function yields the mean vector of coefficients,

(3) $$\max_{\beta} L = \sum_{i=1}^{N} s_i \beta' x_i - \exp(\beta' x_i) - \ln \Gamma(1 + s_i),$$

where Γ is the gamma distribution. The Poisson maximum likelihood estimates are consistent and efficient when the conditional mean and variance are equal. This assumption is maintained while extending the Poisson regression model to one with endogenous growth regimes with spatially varying parameters. The mean-variance equality assumption of the Poisson regression is relaxed later and a quasi-maximum likelihood covariance estimator robust to spatial autocovariance and heteroskedasticity is proposed.

At one extreme, the smooth coefficient Poisson model is but an extension of a count regression model with discrete regimes. For example, Lambert and McNamara (2009) used a negative binomial count regression model with discrete regimes according to metropolitan, micropolitan, and noncore counties to examine the location decision of food manufacturing firms. In this case, the conditional mean of (1) is

(4) $$\mathrm{E}(s_i) = \exp\left(\beta' x_i + d_{1i} x_i \delta_1 \ldots + d_{Ri} x_i \delta_R\right),$$

where $r = 1,\ldots R$ spatial regimes are indicated by dummy variables. Intercepts and slopes are specific to each regime. A "spatial Chow" test (Anselin, 1988) can be applied to determine if the conditional means of the response variable vary according to regime membership. Under the null hypothesis $\delta_1 = \ldots \delta_R = 0$, the mean response is not different according to the indicators used to identify regimes.

The smooth transition parameter model with endogenous regimes is different from the discrete regime approach because regimes are not selected a priori and coefficients are permitted to vary across spatial units. Pede, Florax, and Holt (2009) and Pede (2010) modified Lebreton's (2005) spatial version of the time series smooth transition autoregressive model for time series data by including spatial variables in the transition function, thereby admitting a smooth transition process by incorporating spillover effects through spatial multipliers. The Poisson model with a smooth spatial regime transition extends this strand of research. Other approaches such as local regression techniques or other nonparametric or semiparametric estimators could be used to explore spatially heterogeneous parameter variation are not considered in this analysis.

Let $G(v; \gamma, c)$ be a function with (respectively) slope and location parameters γ and c and a transition variable v. In the time series literature, and recent work with endogenous spatial regime models in the spatial econometric literature (Lambert, Xu, and Florax, 2013), use of the logistic function $[1 + \exp(-\gamma[v - c]/\sigma_v)]^{-1}$ is a common parameterization of G. As γ becomes large, G approaches 1; observations sort uniquely into one regime or another. When $\gamma = 0$, $G = 0.5$; and the model coefficients share a common global interpretation. When G is intermediate 0 and 1, observations sort into two growth regimes, but some observations are *in transition*, with the degree of coefficient heterogeneity dependent on the transition variable associated with a spatial unit.

The location parameter c maps to the inflection point (the median) of G when $\gamma > 0$. The parameters are scale-neutral when normalized by the standard deviation of the transition variable (σ_v). Of particular importance is the choice of the transition variable because it is hypothesized to drive the endogenous sorting process. Ideally, v conveys

information about connectivity, distance, or feedback between spatial units by identifying potentially nonlinear structural breaks across space. The transition variable should also be exogenous. In this analysis, we use the weighted average of a location's neighbor's share of the county covered by shale and tight gas formations. This measure is a localized neighborhood average of shale and tight gas coverage. A number of alternative transition variables are conceivable, but the choice of the average area of shale in surrounding counties is appealing to the extent that most of the increase in natural gas extraction over the last decade occurred in unconventional gas formations where the combination of hydraulic fracturing and horizontal drilling have been used. We therefore hypothesize that establishment growth of this sector occurs at different rates, depending on the likelihood that shale and tight formation deposits are relatively dense in a given micro-region (i.e., a county neighborhood).

The conditional mean function of the Poisson model with endogenous spatial regime-switching potential is,

$$\mu_i = \exp(\beta_1' G_i(v_i, \gamma, c) \odot x_i + \beta_2'(1 - G_i(v_i, \gamma, c)) \odot x_i) \tag{5}$$

where "$\odot$" is an element-wise Hadamer multiplication operator and (β_1, β_2) are coefficients corresponding with regimes 1 and 2. Equation 1 is rearranged following Madalla's (1983) restriction;

$$\mu_i = \exp(\beta_2' x_i + (\beta_1' - \beta_2') G_i(v_i, \gamma, c) \odot x_i), \tag{6}$$

or equivalently,

$$\mu_i = \exp(\beta_2' x_i + \delta' G_i(v_i, \gamma, c) \odot x_i)\,. \tag{7}$$

The interaction between the transition function and the covariates permits coefficients to vary nonlinearly across spatial units, with the coefficients of the interaction terms (δ) the difference from the reference group mean response to local determinants (the β_1's) and the alternative regime. Thus, in the context of discrete count models permitting the parameters determining location attractiveness to vary across sites, rejection of the null hypothesis $\delta = 0$ suggests a nonlinear relationship between location factors, shale and tight gas geological formations, and site selection decisions of natural gas extraction establishments. As with the discrete regime approach using indicator variables, regimes are not evident when $\delta = 0$ and the effects of the exogenous variables are geographically invariant. For continuous variables the marginal effects are calculated as;

$$\frac{\partial E(s_i)}{\partial x_{ik}} = (\beta_k + \delta_k G_i(v_i, \gamma, c))\exp(\beta_1' x_i + \delta' G_i(v_i, \gamma, c) \odot x_i)\,. \tag{8}$$

Estimating the Poisson model with spatially varying parameters and endogenous regimes

As suggested above, the log likelihood function of the typical Poisson model is altered to reflect regime switching potential according to the transition function, G;

(9) $$\max_{\beta,\gamma,c} L = \sum_{i=1}^{N} s_i \beta_1' x_i + s_i \delta' G_i(v_i, \gamma, c) \odot x_i$$

$$- \exp(\beta_1' x_i + \delta' G_i(v_i, \gamma, c) \odot x_i) - \ln \Gamma(1 + s_i) \,.$$

Given this specification, the log likelihood function can be maximized using a variety of statistical software packages.

Critical for maximizing the success of convergence is determining a set of starting values for the transition function's location parameter (*c*) and shape and slope parameter (γ). We use a double grid search to determine feasible starting points. The location parameter, *c*, is varied over the 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th, and 99th percentiles of the transition variable, *v* (the average percent of shale deposits in neighboring counties). The shape parameter, γ, is varied from 0 to 100 in increments of 5, noting that at γ = 0 there are no regimes and at γ = 100 there are two distinct regimes (similar to a dummy variable indicator). The modified Poisson log likelihood function is maximized at each starting point combination. The log likelihood is retained after each iteration and Akaike's Information Criteria is calculated. After all combinations are exhausted, the (*c*, γ) combination yielding the lowest AIC score is used as a starting pair in a final maximum likelihood estimation of the modified Poisson log likelihood function.

Quasi-maximum likelihood covariance estimator with spatial autocovariance

We extend covariance estimation of the smooth coefficient Poisson model to a general spatial autocorrelation model suggested by Kelejian and Prucha (2007) to accommodate the interaction between unobserved factors possibly influencing firm site selection, including competition between counties for firm investment, inter-county trade

patterns, or possibly similar resource endowments and the attendant spillover due to the correlated error structure arising from these omitted variables. The covariance model therefore assumes a cross sectional disturbance process allowing for unknown forms of heteroskedasticity and correlation between counties. Kelejian and Prucha (2007) outline the assumptions of their generalized spatial correlation model, which are comparable with robust covariance estimation in general linear model theory (Cameron and Trivedi, 2005). Given a consistent estimator for equation (9), the quasi-maximum likelihood disturbance vector is $u_i = s_i - \mathrm{E}[s_i] = r_{j.}\varepsilon$, where $r_{j.}$ is the jth row of an n by n non-stochastic matrix (R) with unknown elements whose row and column sums are uniformly bounded in absolute value (i.e., correlation between cross sectional units is restricted), and ε is an n by 1 independent and identically distributed vector of disturbance with and expected mean of zero and a constant variance. The exact form of R is unspecified, but Kelejian and Prucha (2007) demonstrate that the Cliff-Ord type spatial error processes models (e.g., Anselin, 1988) are a special case of the generalized spatial autocorrelation model so long as some specific assumptions are maintained (below). The asymptotic distribution of the non-stochastic location determinants is; $\Psi = X'\Sigma X$, where $\Sigma = \mathrm{E}[uu']$. Kelejian and Prucha (2007) define a non-parametric estimator for Ψ, and prove its consistency. As Kelejian and Prucha demonstrate, the asymptotic results extend to nonlinear models as well as a variety of other distributions, of which the Poisson model is a special class in the context of general linear model theory (Cameron and Trivedi, 2005). The spatial HAC estimator is extended to the Poisson robust covariance estimator below.[3]

[3] Cameron and Trivedi (1998) provide details of the time series Newey and West (1987) HAC analogue as applied to count models.

Consider the heteroskedastic robust covariance matrix (V) of the Poisson maximum likelihood estimator; $V = A^{-1}BA^{-1}$, where A is the expected Hessian, B the outer product of the maximum likelihood function gradients, $B = \sum_{i=1}^{N} b_i b_i'$, and b the derivative of the Poisson log likelihood function with respect to relevant coefficients. For example, in the standard Poisson regression, $b_i = x_i(y_i - \exp(\beta' x_i))$, which reduces to the product of a residual" and each covariate for the ith observation; $x_i u_i$.

Consider next the spatial HAC estimator suggested by Kelejian and Prucha (2007), which uses a nonparametric estimator to adjust for covariance between cross sectional units. A kernel function determines the range over which the cross products of the residuals are correlated. The kernel choice is generally not critical so long as K is a bounded, symmetric, real, and continuous function that integrates to one (Mittelhammer et al, 2000).[4] Candidate functions include Gaussian, Parzen, Bartlett, Epanechnikov, or the bi-square. The Epanechnikov kernel (K) is applied here given its relative efficiency (Mittelhammer et al., 2000). Given an appropriate kernel function, and the covariance structure of the robust Poisson estimator, the (r, s) elements of Ψ are estimated as:

$$\hat{\psi}_{rs} = \sum_{i=1}^{N} \sum_{j=1}^{N} x_{ir} x_{js}' \hat{u}_i \hat{u}_j K(d_{ij}, d_{max}), \tag{10}$$

where (r, s) indexes the lag structure between locations (i, j), d_{ij} is the distance between spatial units i and j; and d_{max} is defined below. See Kelejian and Prucha (2007, p. 136) for details.

[4] Two recent applications of the spatial HAC estimator report that standard errors estimated using different kernels (but the same bandwidth) were similar with respect to inference (Lambert et al., 2007; Anselin and Lozano-Gracia, 2008).

In large samples, the kernel bandwidth choice (d_{max}) is more important than the functional form of the kernel (Cameron and Trivedi, 2005). In the spatial context, this amounts to identifying an optimal neighborhood of observations. In their Monte Carlo evaluation of the spatial HAC estimator, Kelejian and Prucha (2007) use a plug-in bandwidth $n^{1/4}$ for identifying the neighbors determining d_{max}. The theoretical idea behind the plug-in estimator is similar to the Newey and West (1987) HAC plug-in estimator determining the number of lags in the time series literature. In this study, n = 3,078, which corresponds with q = 7 neighbors. For county i, the vector of distances between j and all other locations are sorted in ascending order. The q value is therefore a cutoff point in the sorted distance vector identifying d_{max}, the last distance entry in the truncated vector corresponding with county i. The weighting mechanism allows $K(d_{ij}, d_{max})$ to expand or contract across cross-sectional units, thereby re-weighting residual cross-products according to the distance between a set of neighbors. In this study, the distance between population-weighted county centroids was used as the inter-county distance measure.

6. Empirical Results and Discussion

As a reference point we report results from standard Poisson count model in the absence of spatial regimes in addition to the smooth transition count model (table 2). Results from the standard model are generally as expected. Oil and gas companies often rely on information of past production for future location decisions. The historical production of natural gas (*gas*) was positively correlated with location decisions. Both distance to nearest pipeline (*pipeline distance*) and storage facility (*storage distance*) had

the expected negative sign, while only distance to the nearest natural gas storage facility was statistically significant. Local agglomeration, as measured by the share of mining employment in the county (*mining share*), was positively associated with business establishment site selection decisions. Access to a specialized labor pool is often important to drilling operations. Higher educational attainment (*bachelors*) and road infrastructure (*interstate*) were also positively correlated with the extraction establishment count. These results assume that the effects of these local attributes are the same across space.

<< Table 2 >>

The smooth transition count model relaxes the assumption of coefficient spatial homogeneity across space. The null hypothesis that the effects of the covariates on location choices were geographically invariant ($\delta = 0$, from equation 7) was rejected at the 99% level (LR = 162, df = 14), suggesting that the location factors exhibited heterogeneity across counties. Based on this result, we conclude that the smooth transition count model aptly described the data-generating process determining gas extraction establishment location decisions during the study period. The transition function parameters were $\gamma = 100$ (the shape parameter) and $c = 0.17$ (the location parameter, moving neighborhood average of shale and tight gas formation). The value and statistical significance of the shape parameter suggests the presence of two distinct growth regimes (figure 4a) with a transition threshold of 17% coverage of shale or tight gas formations. The regime probabilities suggest that location counts were more common

in regime 1 when $G = 1$ but less common in regime 2 when $G = 0$ (figure 4b). As a result, the effects of the covariates on extraction establishment location decisions are expected to be different once the coverage of a county exceeds 17%.

The coefficient vector (δ) is the difference of effects between regime 1 ($G = 1$) and regime 2 ($G = 0$). Positive and significant differences were observed for the manufacturing, construction, and mining shares of employment as well as for interstate highway infrastructure. Negative and significant differences were found for the unemployment rate and population density.

Marginal effects were calculated for counties sorting into both regimes as well as those in transition (Table 3). In the high growth regime (regime 1), a one unit increase in the share of mining employment in from the base year is associated with an increase of almost 7 additional establishments compared to less than two in the low growth regime (regime 2). Counties in the high growth regime were also more sensitive to educational attainment. Higher levels of educational attainment had stronger positive association with location decisions in the high growth regime. The marginal effect suggests a one percent increase in the unemployment rate decreases the growth in establishments by 0.3. Similarly, an increase of 100 miles of interstate highway network in the high growth regime is associated with two additional natural gas extraction establishments.

<< Table 3 >>

The likelihood of counties sorting into high growth regions was determined by calculating the probability of the predicted value of establishments being greater than or

equal to the 90th percentile of counts; e.g., two establishments locating in a county during the study period as observed in the full sample. This conditional probability is:

$$\Pr(s_i \geq 2) = 1 - \left(\frac{e^{-\mu_i} \mu_i^0}{0!} + \frac{e^{-\mu_i} \mu_i^1}{1!} \right), \tag{11}$$

The mapped probabilities are shown in figure 5. A spatial cluster analysis was then conducted on these estimated location probabilities to determine regions where, ceteris paribus, natural gas extraction business establishment events were more likely to occur. Using the Getis-Ord statistic (Ord and Getis, 1995), figure 6 shows statistically significant clustering of counties with high probabilities of growth exceeding the 90th percentile that have neighboring counties with similar high growth probabilities (shaded red) and counties with low location event probabilities with neighboring counties exhibiting similarly low site selection probabilities (shaded blue). The map highlights several regions where future natural gas extraction is more likely to occur with a return to higher prices or policy changes. These areas are identified by the lighter red color. Several of these counties are in Colorado, Wyoming, Oklahoma, Texas, Louisiana, southern California, eastern Ohio and western Pennsylvania. Activity levels have been high in the Marcellus shale which covers large portions of Pennsylvania and New York. However, New York currently has a moratorium on fracking, which has prevented drilling in the state. California has also been resistant to develop the Monterey Shale in the southern part of the state citing environmental concerns. Given price and/or policy changes, those areas would be the most likely to see more activity in the future.

<< Figure 5 >>

<< Figure 6 >>

7. Conclusion

The exponential growth of natural gas production over the last decade has been largely driven by the combination of horizontal drilling and hydraulic fracturing technologies. Gas reserves that were once unprofitable to extract in shale and tight gas formations have opened up in several regions of the United States. We find that business establishment growth in this sector (as evidenced by location decisions of natural gas extraction establishments from 2005 to 2010) is characterized by high and low growth regimes, with counties on the margin of shale-rich regions exhibiting slower growth. As such, the local determinants correlated with location decisions exhibited different magnitudes of importance and county-specific heterogeneity associated with firm site selection rates. Our results show that counties with a locally weighted average of more than 17% coverage by shale and tight gas transition to a high growth regime where the incidence of extraction establishments is higher. Local agglomeration externalities, access to skilled labor, and transportation infrastructure where of more economic importance to location decisions in the high growth regime.

Presence of shale or tight gas formations in an area does not guarantee that a county will be an ideal location for businesses engaged in natural gas extraction and distribution. However, if the shale and tight gas formations cover a large enough area, it is more likely that attracting firms involved in this sector will occur. Our contribution to the literature is to provide a method that allows for the estimation of the threshold level of

coverage where the probability of development is more likely and how being above or below that threshold affects the location factors relevant to location choice. A priori, one might think that the necessary coverage for development to occur would be higher than what was estimated in the model. Our approach also helps identify areas where future development is more likely. This information could help policy makers and economic development practitioners better plan for infrastructure and other local service needs in the face of resource extraction.

One limitation in the data relates to our measure of gross establishment entry, which may underestimate the actual number of new establishments. It is not possible to identify exiting firms in the annual net count in publically available data. Moreover, we are not able to distinguish between single- and multi-unit establishments, which may face very different location choice problems. Having additional information on the cost of extraction (depth of the shale) and the quality (thickness of the shale) of the natural gas resource would also be useful. Information on water availability and location of water treatment facilities would also aid in determining some of the environmental costs of extraction.

Possible extensions of this work would be to consider other count data models. The Poisson model assumes that the conditional mean and variance are equal. Relaxing this assumption by estimating a negative binomial model would allow for unobserved heterogeneity, which is often interpreted as a location specific random effect. Future work might also consider other semi-parametric models with more flexible functional forms to estimate endogenous regimes.

References

Anselin, L., 1988. Spatial Econometrics: Methods and Models. London, Kluwer Academic Publishers.

Anselin, L. and Lozano-Gracia, N. 2008. Errors in variables and spatial effects in hedonic house price models of ambient air quality. Empirical Economics, 34(1): 5-34.

Banerjee, S., Carlin, B.P., and Gelfand, A.E. 2004. Hierarchical Modeling and Analysis for Spatial Data. Chapman & Hall/CRC.

Bhati, A.S. 2005. Robust spatial analysis of rare crimes: an information-theoretic approach. Sociological Methodology 35(1): 277-289.

Blair, J.P., Premus, R., 1987. Major factors in industrial location: a review. Economic Development Quarterly 1(1), 72–85.

Cameron, A.C., Trivedi, P. K., 1998. Regression Analysis of Count Data. Econometric Society Monograph No.30, Cambridge, Cambridge University Press.

Cameron, A.C., Trivedi, P. K., 2005. Microeconometrics: Methods and Applications. Cambridge, Cambridge University Press.

Carod, J.M.A., 2005. Determinants of industrial location: an application for Catalan municipalities. Papers in Regional Science 82(1), 105 – 120.

Carod, J. M. A., Antolín, M., 2004. Firm size and geographical aggregation: an empirical appraisal in industrial location. Small Business Economics 22, 299 – 312.

Carod, J.M.A., Solis, D. and Antolín, M. 2010. Empirical studies in industrial location: An assessment of their methods and results. Journal of Regional Science 50(3): 685-711.

Chong, S., 2006. The role of labor cost in the location choices of Japanese investors in China. Papers in Regional Science 85(1), 121–138.

Conley, T. 1999. GMM Estimation with Cross Sectional Dependence. Journal of Econometrics 92(1):1–45.

Coughlin, C.C., Segev E., 2000. Location determinants of new foreign–owned manufacturing plants. Journal of Regional Science 40, 323–351.

Cressie, N. 1993. Statistics for spatial data. Wiley, New York.

Davis, D.E., Schluter, G.E., 2005. Labor-force heterogeneity as a source of agglomeration economics an in empirical analysis of county-level determinants of food plant entry. Journal of Agricultural and Resource Economics 30(3), 480–501.

Henderson, J.R., McNamara , K.T., 2000. The location of food manufacturing plant investments in Corn Belt counties. Journal of Agricultural and Resource Economics 25(2), 680–697.

Fotopoulos, G., Louri, H., 2000. Location and survival of new entry. Small Business Economics 14, 311– 321.

Gabe, T. 2003. Local industry agglomeration and new business activity. Growth and Change 34(1): 17-39.

Glaeser, E.L., Kohlhase, J.E., 2004. Cities, Regions, and the Decline of Transport Costs. Papers in Regional Science 83, 197–228.

Gotway, C.A. and Stroup, W.W., 1997. A generalized liner model approach to spatial data analysis and prediction. Journal of Agricultural, Biological, and Environmental Statistics 2(2): 157-178.

Greene, W.H. 2000. Econometric Analysis. Upper Saddle River, NJ: Prentice Hall.

Guimarães, P., Figueiredo, O., Woodward, D., 2003. A tractable approach to the firm location decision problem. The Review of Economics and Statistics 85(1), 201–204.

Guimarães, P., Figueiredo, O., Woodward, D., 2004. Industrial location modeling: extending the random utility framework. Journal of Regional Science 44(1), 1–20.

Hays, J.C. and Franzese, R.J. 2009. A comparison of the small-sample properties of several estimators for spatial-lag count models. Unpublished Working Paper, University of Illinois Champaign-Urbana.

Head, K., Ries, J., and Swenson, D. 1995. Agglomeration benefits and location choice: evidence from Japanese manufacturing investments in the United States. Journal of International Economics 38, 223-247.

Kelejian H.H., Prucha, I.R., 2007. HAC estimation in a spatial framework. Journal of Econometrics 140, 131 – 154.

Lambert, D.L., Brown, J.P., and Florax, R.J.G.M. 2010. "A two-step estimator for a spatial lag model of counts: Theory, small sample performance and an application." Regional Science and Urban Economics, 40(4): 241-252.

Lambert, D. M., C. D. Clark, M. D. Wilcox, and W. M. Park. 2007. Do Migrating Retirees Affect Business Establishment and Job Growth? An Empirical Look at Southeastern Non-metropolitan Counties, 2000 – 2004. Review of Regional Studies 37(2): 251 – 278.

Lambert, D.M., Garret, M.I., McNamara, K.T., 2006a. An application of spatial Poisson models to manufacturing investment location analysis. Journal of Agricultural and Applied Economics 38(1), 105–121.

Lambert, D.M., Garret, M.I., McNamara, K.T., 2006b. Food industry investment flows: implications for rural development. Review of Regional Studies 36(2), 140–162.

Lambert, D.M., McNamara, K.T., 2009. Location determinants of food manufacturers in the U.S., 2000 - 2004: are nonmetropolitan counties competitive? Agricultural Economics, 40(6): 617–630.

Lambert, D. M., Xu, W., and Florax, R. J. G. M. 2012. Partial Adjustment Analysis of Income and Jobs, and Growth Regimes in the Appalachian Region with Smooth Transition Spatial Process Models. International Regional Science Review. In press.

Lebreton, J.-D. 2005. ‘‘Dynamical and Statistical Models for Exploited Populations.’’ Australian and New Zealand Journal of Statistics 47:49–63.

LeSage, J.P. and Fischer, M.M. 2012. Estimates of the Impact of Static and Dynamic Knowledge Spillovers on Regional Factor Productivity. International Regional Science Review 35(1): 103-127.

List, J.A. 2001. US county-level determinants of inbound FDI: evidence from a two-step modified count model. International Journal of Industrial Organization 19, 953-973.

Luker, B. 1998. Foreign investment in nonmetropolitan U.S. South and Midwest: A case of mimetic location behavior? International Regional Science Review 21: 163-184.

Maddala, G.S. 1983. Limited-Dependent and Qualitative Variables in Econometrics. Cambridge, UK: Cambridge University Press.

McFadden, D., 1974. Conditional Logit Analysis of Qualitative Choice Behavior. In: Zarembka, P. (Ed.). Frontiers in Econometrics. New York: Academic Press.

Mittelhammer R., Judge G. and Miller D. 2000. Econometric Foundations, Cambridge University Press, NY.

Newey, W. and West, K. 1987. A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. Econometrica 55:703-708.

Ord, J.K. and Getis, A. 1995. Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. Geographical Analysis 27(4): 286-306.

Pede, V.O. 2010. ‘‘Spatial Dimensions of Economic Growth: Technological Leadership and Club Convergence.’’ Ph.D. Dissertation, Purdue University.

Pede, V.O., R.J.G.M. Florax, and M.T. Holt. 2009. ‘‘A Spatial Econometric STAR Model with an Application to U.S. County Economic Growth, 1969–2003.’’ Working paper, Department of Agricultural Economics, Purdue University.

Rasmussen, S. 2004. Modeling of discrete spatial variation in epidemiology in SAS using GLIMMIX. Computer Methods and Programs in Biomedicine 76(1): 83-89.

Schabenberger, O., and F.J. Pierce. 2002. Contemporary Statistical Models for the Plant and Soil Sciences. New York: CRC Press.

Smith, D.F. and Florida, R. 1994. Agglomeration and industrial location: An econometric analysis of Japanese-affiliated manufacturing establishments in automotive-related industries. Journal of Urban Economics 36, 23-41.

Woodward, D.P., 1992. Location determinants of Japanese manufacturing start–ups in the United States. Southern Economics Journal 58, 690–708.

Yergin, D. 2011. *The Quest*. Penguin Press, New York.

Table 1. Natural Gas Extraction Establishments and Location Factors

Variable	Description	Mean	Std. Dev.
gas establishments	Count of new extraction establishments (2005 - 2010)[1]	0.9	3.4
shale	Share of the county covered by shale or tight gas formations[2]	0.2	0.4
gas production	Natural gas production in 2000 (billions of cubic feet)[3]	5.1	28.0
pipeline distance	Distance to nearest natural gas pipeline (miles)[2]	24.7	35.7
storage distance	Distance to nearest natural gas storage facility (miles)[2]	107.2	92.9
manufacturing share	Manufacturing share of employment[4]	0.13	0.09
construction share	Construction share of employment[4]	0.06	0.03
mining share	Mining share of employment[4]	0.01	0.03
agriculture share	Agricultural share of employment[4]	0.11	0.09
unemployment	Unemployment rate (%)[5]	4.3	1.6
household income	Median household income (Thous. $)[6]	35.2	8.7
population density	Hundred people per square mile[6]	9.4	18.3
associates	% of adult population with high school or associates degree[6]	60.9	7.0
bachelors	% of adult population with bachelor's degree or higher[6]	16.4	7.7
interstate	Total miles of Interstate highway in the county[7]	14.7	25.2

N = 3,078; Sources: [1] County Business Patterns; [2] Author's calculations; [3] USDA ERS; [4] Bureau of Economic Analysis, REIS; [5] Bureau of Labor Statistics; [6] US Census Bureau, 2000 Decennial Census; [7] US DOT.

Table 2. Standard and Smooth Transition Count Model Results

	Standard Count Model		Smooth Transition Count Model			
	β Coefficient	Robust S.E.	β Coefficient	Robust S.E.	δ Coefficient	Robust S.E.
Intercept	-1.104	0.888	1.244	1.245	-3.942***	1.430
gas	0.004***	0.001	0.004***	0.001	-0.0004	0.001
pipeline distance	-0.008	0.005	-0.009	0.005	-0.005	0.006
storage distance	-0.008***	0.001	-0.007***	0.002	0.0004	0.002
manufacturing share	-1.792***	0.575	-3.007***	0.969	2.133*	1.158
construction share	-3.645*	2.168	-8.465***	2.839	6.775**	3.191
mining share	3.908***	0.839	3.857***	1.390	1.134	1.746
agriculture share	-5.716***	0.657	-6.618***	0.969	2.066*	1.117
unemployment	-0.001	0.048	-0.080	0.070	0.127	0.078
household income	0.002	0.006	0.005	0.008	-0.006	0.010
population density	0.003***	0.001	0.007***	0.003	-0.003	0.002
associates	0.007	0.010	-0.014	0.013	0.036**	0.015
bachelors	0.032***	0.009	0.015	0.012	0.025*	0.015
interstate	0.008***	0.002	0.001	0.002	0.012***	0.002
γ	---		100.0			
c	---		0.168***	0.001		
State fixed effects†	yes		yes			
Fit statistics						
Log Likelihood	-3192.4		-3111.1			
AIC	6504.8		6374.1			

Notes: N = 3,078; Statistical significance at the 99th, 95th, and 90th percentile is represented by ***, **, and *, respectively.† The results for the state fixed effects are not shown to conserve space.

Table 3. Average Marginal Effects by Regime

	Regime 1	Regime 2
G (regime probability)	1.000	0.000
Gas	0.006	0.002
pipeline distance	-0.020	-0.004
storage distance	-0.010	-0.003
manufacturing share*	-1.262	-1.374
construction share*	-2.441	-3.868
mining share*	7.212	1.763
agriculture share	-6.576	-3.024
unemployment*	0.068	-0.037
household income	-0.001	0.002
population density	0.005	0.003
associates*	0.032	-0.006
bachelors*	0.058	0.007
interstate*	0.019	0.0004

Notes: The asterisk indicates significant differences between regimes 1 and 2.

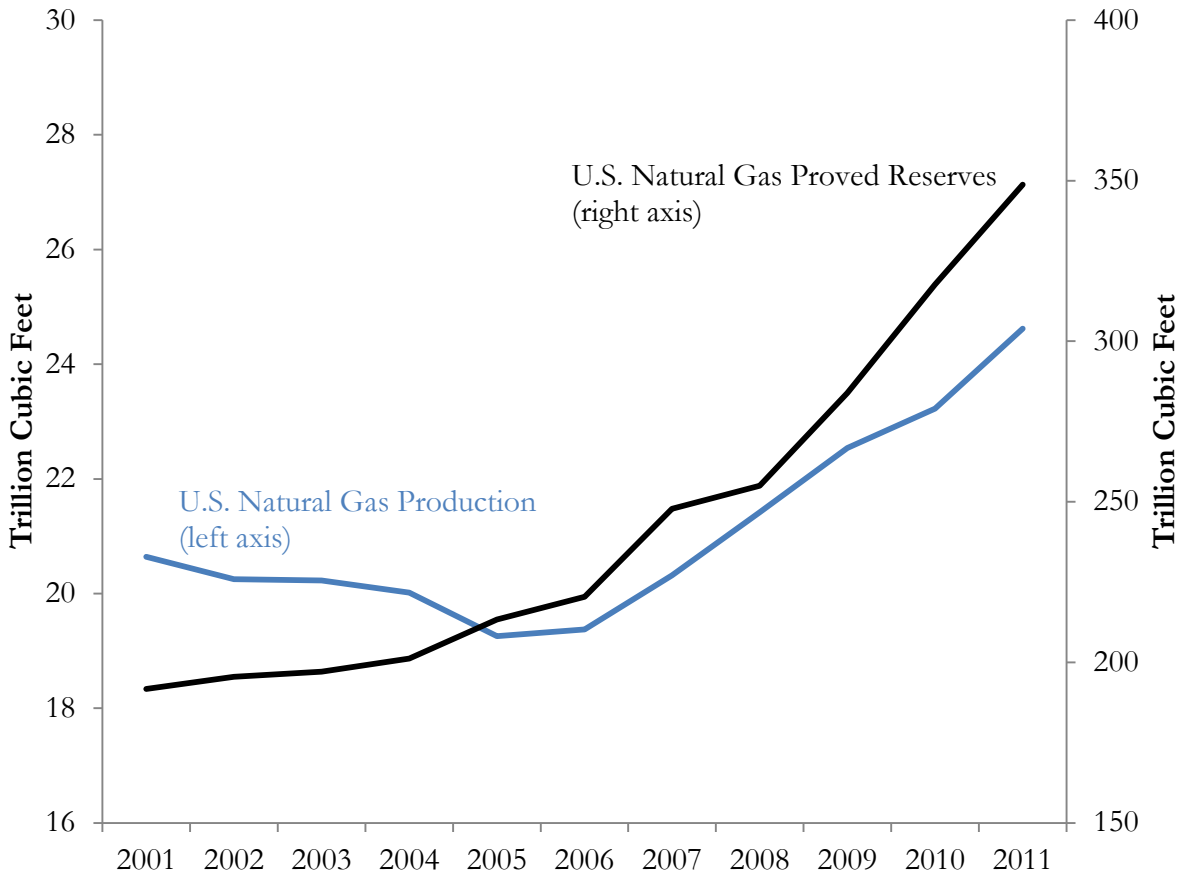


Figure 1. U.S. Natural Gas Production and Proved Reserves, 2001 – 2011

Source: Energy Information Agency

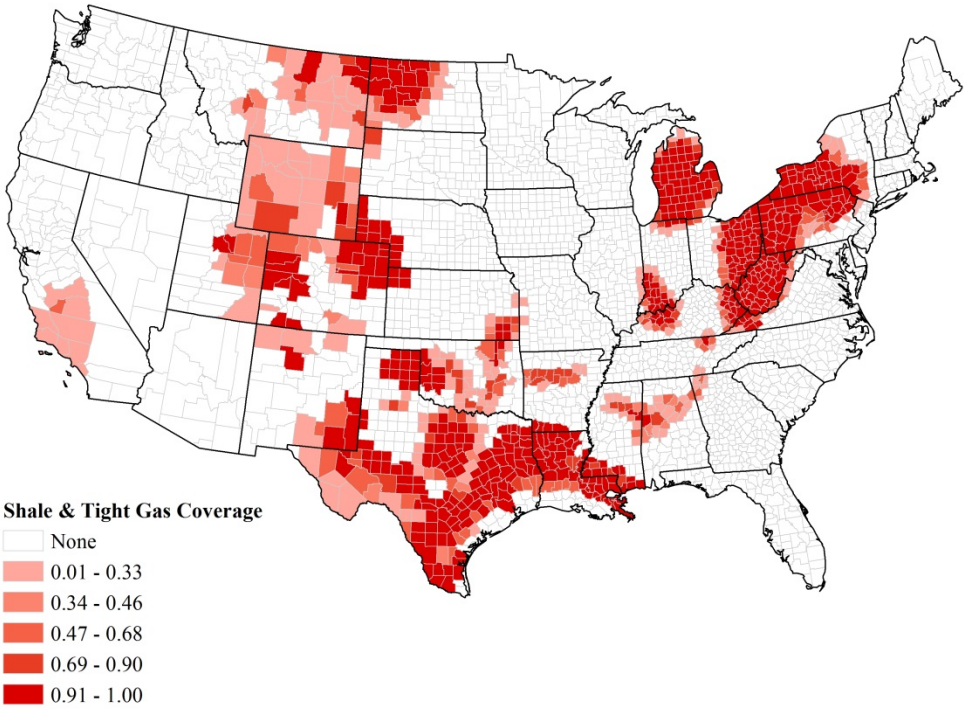


Figure 2. Share of County Covered by Shale and Tight Gas Formations

Source: Energy Information Agency

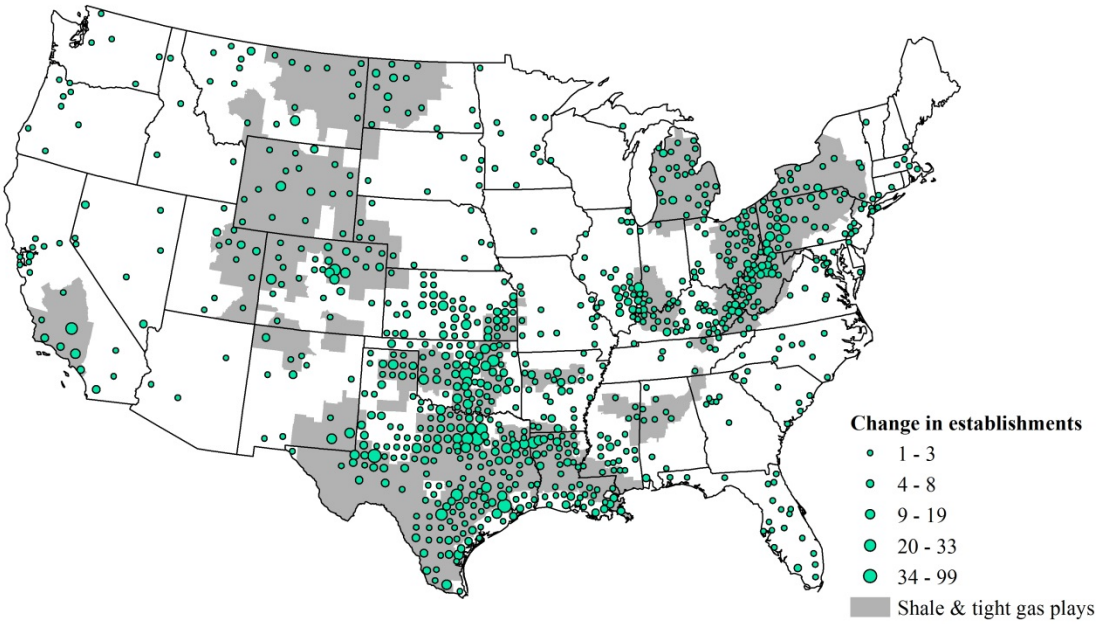
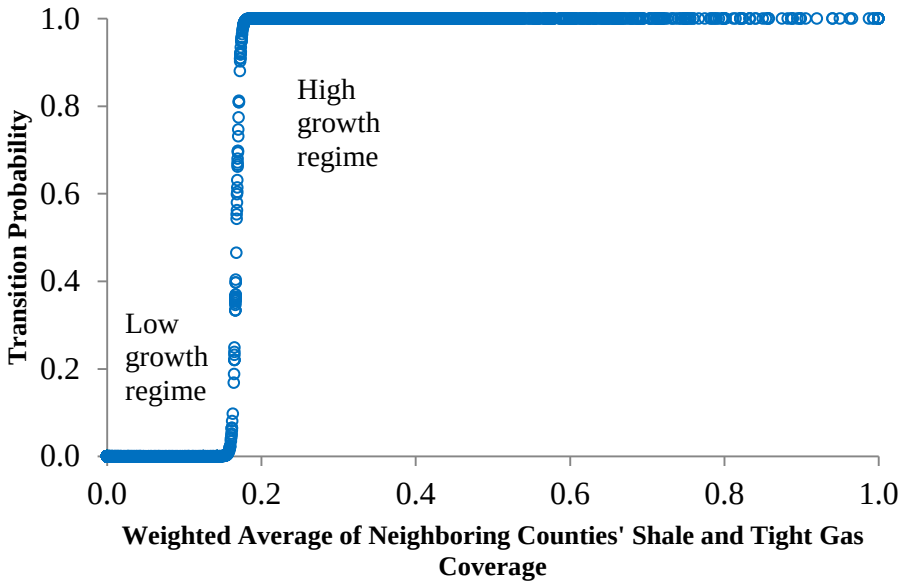
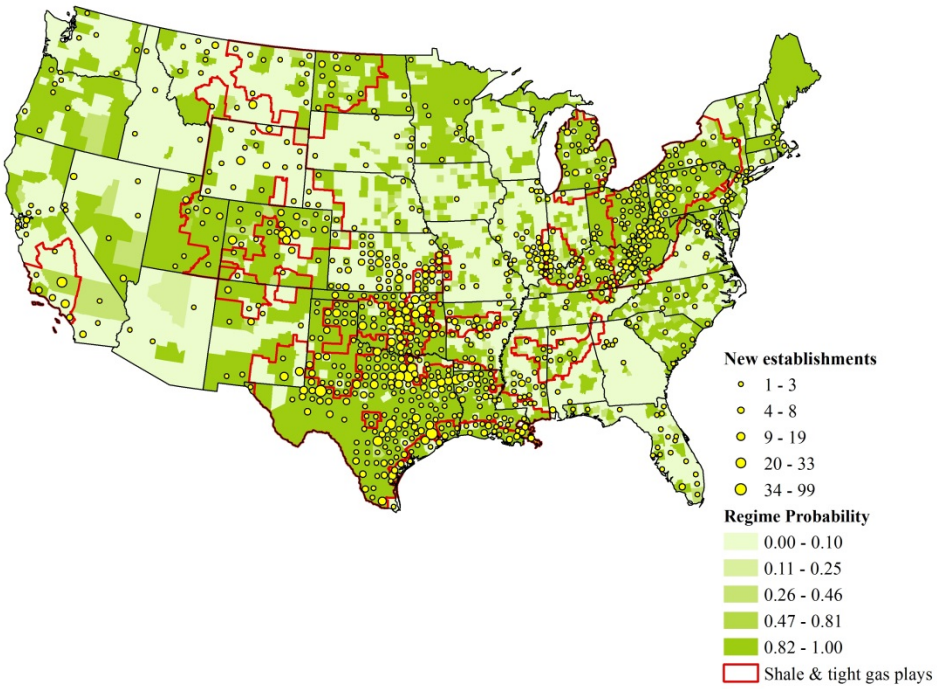


Figure 3. Location of new natural gas extraction establishments, 2005 to 2010.

Source: Authors' calculations



(a)



(b)

Figure 4. Regime Probabilities Across Counties

Source: Authors' calculations

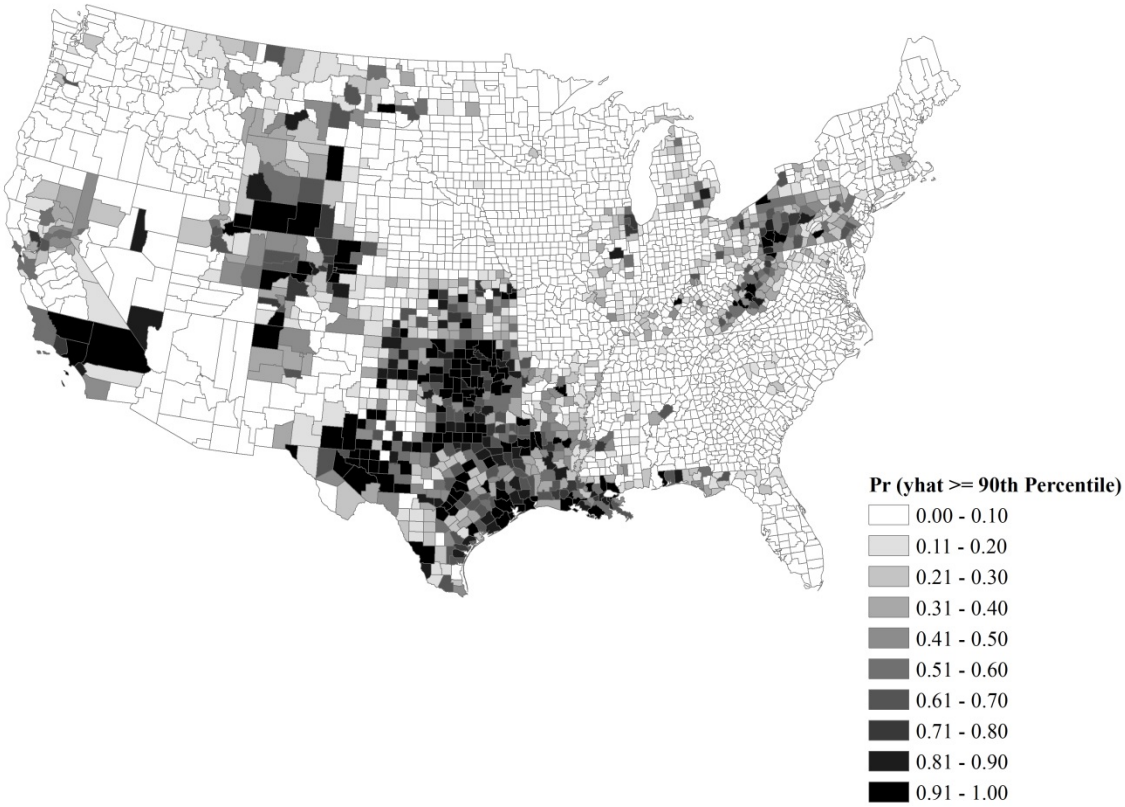


Figure 5. Probability of Establishment Growth Exceeding 90th Percentile

Source: Authors' calculations

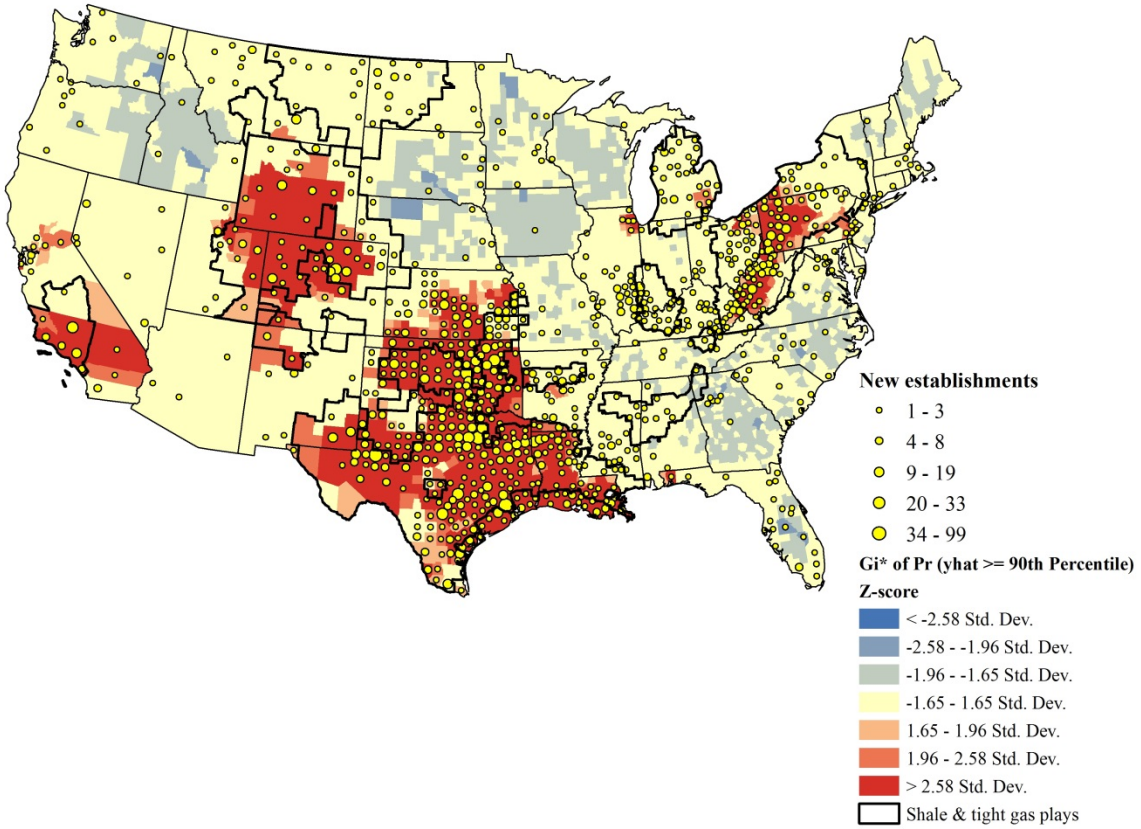


Figure 6. Clusters of High and Low Growth Probabilities

Source: Authors' calculations