U.S. Patent and Trademark Office
OFFICE OF THE CHIEF ECONOMIST

Economic Working Paper Series

Supplementary material

Where are U.S. women patentees? Assessing three decades of growth

Michelle Saksena, PhD, *Senior Research Economist, USPTO*
Nicholas Rada, PhD, *Deputy Chief Economist, USPTO*
Lisa Cook, PhD, *Professor, Michigan State University*

October 2022

UNITED STATES
PATENT AND TRADEMARK OFFICE ®

1. Introduction

The supplement describes the data and empirical methods used in USPTO IP Data Highlight Number 6, "Where are U.S. women patentees? Assessing three decades of growth." The report documents where women are participating as inventor-patentees across U.S. counties and explores some of the county-level factors correlated with their participation. We present the supplemental material in three sections: (1) we describe the data sources and variable construction; (2) we describe our estimation methods; and (3) we discuss results and their implications.

2. Data

We combine four sources to form the main dataset used in the analysis. The first source is USPTO's PatentsView (PV) data spanning years 1990-2019. PV contains information on the gender of inventor-patentees, their locations, the number of inventors per granted patent, and the Cooperative Patent Classification (CPC) that denotes the technologly field of a patent.[1] From the Bureau of Labor Statistics (BLS), Local Area Unemployment Statistics Program, we obtain county-level employment data. The third source, the Bureau of Economic Analysis (BEA), provides county-level income data. The final source is the Census Decennial data and 5-year American Community Survey (ACS).

For our analysis, we use per capita income from BEA and labor force data from BLS. We draw on Census and ACS to construct the metrics for educational attainment by county and year. Last, we use PV to construct all other variables: team size, number of all male teams, and the dependent variable (described below). Table 1 displays comparative statistics for all variables in our analysis. We group the statistics by counties with and without women inventors.

We create the dependent variable—*number of women inventors*—from PatentsView by summing over the number of unique female inventors within each county-year combination.[2] Importantly, we take advantage of the inventor disambiguation (or name harmonization) algorithm provided by PatentsView, which assigns unique inventor IDs, making it possible to identify unique men and women inventor names on patents over time.[3]

We generate variables that capture the volume of patents by technology for each county and year. These variables are in Table 1 under 'CPC technology shares'.[4] They tell us the intensity of

[1] See www.PatentsView.org.

[2] In situations where more than one address with different counties was listed on a patent, we assigned women inventors to both counties.

[3] See https://patentsview.org/disambiguation for thorough documentation of the disambiguation process.

[4] The USPTO classifies patents into at least one technical area using the Cooperative Patent Classification (CPC) system. Within in the CPC system, there are eight top-level sections corresponding to the International Patent Classification (IPC), plus a "Y" section to tag emerging and cross-referenced technologies. (Note: Y classified patents are excluded in this analysis.) Each patent is assigned to a classification that best captures the invention as a whole for

patenting activity for each technology field and how this might differ in the presence of women inventor-patentees. Specifically, these variables measure the percentage of patents granted in each year and county for each of the eight CPC sections, giving us an average county portfolio of patenting by technology field.[5] For example, Table 1 (left side) indicates that for counties with no women patentees, only 11% of patents are in physics. Contrast this with the right side of Table 1, which indicates that 17% of patents are in physics when women patentees are present in the county. This 6-percentage point spread, together with the results from figure 3 in the main report, shows that there is more patenting in physics when women are present, which is surprisingly unique among technology fields.

We also construct indicator variables for each of the eight CPC sections, presented as 'CPC technology indicators' in Table 1, to illustrate the scope of counties participating in each technology field. The indicators tell us whether a certain technology field is concentrated in a few counties or if technology development is geographically dispersed. We assign a value of one to a county-year observation if at least one granted patent in that county-year fell into one of the eight CPC sections. For example, the textiles; paper technology field is the most concentrated. We find only 5% of counties with no women patentees produce textiles; paper patents, whereas 28% of counties with women patentees have patents in this field. Contrast this with human necessities, in which 46% and 86% of county-years have at least one patent in this field in counties without and with women patentees, respectively, indicating wider dispersion among counties.

To develop the education variables, we combine Census decennial data and 5-Year ACS estimates. For the intervening 9-year period between census data for which we do not have education estimates (i.e., 1991-1999), we linearly interpolate to infer educational attainment values between census years 1990 and 2000. Starting in 2005, Census changed their decennial data collection process by replacing it with the ACS. ACS data collection occurs every 5 years. We linearly interpolate to calculate education values for years 2001-2004. For years 2005 and onward, we apply 5-year ACS values. Census publishes a new 5-year ACS dataset annually. Following guidance from Census concerning overlapping 5-year estimates, we use stepwise construction for years 2005-2019.[6] For example, the value from the 2005-2009 ACS education estimates repeats for years 2005 to 2009.

3. Methods

Our analysis uses a zero-inflated negative binomial (zinb) model to assess how county-level economic factors might influence women inventor-patentees, controlling for a number of potential confounders such as the patent technology, time, U.S. states, and patent teams. The model further distinguishes differences in how higher levels of education influence the probability

the patent family; this classification is designated as the "CPC First" classification. See www.uspto.gov/web/patents/classification/cpc/html/cpc.html.

[5] Note, the cpc technology shares sum to 100% in each county.

[6] See https://www.census.gov/programs-surveys/acs/guidance/estimates.html.

that a U.S. county has its first woman inventor-patentee, controlling for time and U.S. states. One advantage of the zinb model is that it accommodates a data generating process that produces excessive zeros (Lambert, 1992; Lord et al., 2004; Raihan et al., 2019). This is particularly applicable to our analysis as the majority of U.S. counties in our sample do not have any women inventors. (Note the swaths of grey counties in Figures 1a and 1b of the report.)

The zinb model assumes two different processes that could result in a county not having a women inventor: (1) counties may not have an environment conducive to women inventors, and thus are defined as structurally zero-women inventor counties; and, (2) an innovation ecosystem is present, but no women inventors have applied for and received a patent in a given year.

To accommodate the two zero-county processes, the zinb model uses a logistic function to estimate how higher education affects the likelihood that a county has no women inventor-patentees. In doing so, the model separates zero-counties into structural zero-women inventor counties (process (1)) and observational zero-counties (process (2)). The observational zero-counties combine with non-zero counties, and a negative binomial function estimates how economic factors and other controls influence the number of women inventors in these counties.

A feature of having many zeros in the data is that it can result in overdispersion of the dependent variable. Overdispersion occurs when the variance of the dependent variable exceeds its mean. The zinb model accounts for this by parametrizing the dispersion. In the event that the data are not overdispersed, a zero-inflated Poisson (zip) model that assumes a Poisson distribution for the number of women inventor-patentees is most efficient (Cameron and Trivedi, 2005). We test the women inventor data for overdispersion in four ways: (1) we calculate whether the mean-to-variance ratio of the dependent variable is greater than unity, (2) we test the statistical significance of the overdispersion parameter alpha (α) and log(α), (3) we adopt a likelihood ratio test to assess the goodness of fit between the zip and zinb models, and (4) we calculate Akaike's information criterion (AIC) and Bayesian information criterion (BIC) to compare maximum likelihood models. All four methods confirm overdispersion and support the zinb model.[7]

[7] For (1), we report a mean-to-variance ratio of 7.4, which is larger than 1. For (2), our model reports an α=0.4 and log α = -0.86 (if there was no dispersion, α = 0 and log(α)= -∞), both have a p-value of 0.000. For (3), we perform the likelihood ratio test where the null hypothesis is α = 0. Our χ^2= 1.5e+05 and is statistically significant at the 99.9% level; thus, we reject the null hypothesis. Finally (4), we calculate AIC and BIC for both models:

Model	**AIC**	**BIC**
Zinb	182,931.5	184,735.6
Zip	328,448.6	330,243.6

Both AIC and BIC are smaller for the zinb model, thus concluding that the zinb model is appropriate.

Table 1. Summary statistics comparing counties with and without women inventors

	Counties with women inventors = 0					**Counties with women inventors > 0**				
			Std.					Std.		
	N	Mean	Dev.	Min	Max	N	Mean	Dev.	Min	Max
Number of women inventors	-	-	-	-	-	29,801	15	67	1	2,956
Labor force	34,145	16,481	16,508	261	416,540	29,801	120,298	250,215	460	5,121,584
Per capita income ($)	34,145	27,615	10,474	7,096	175,998	29,801	33,672	13,570	9,798	230,141
Team size	34,145	2	1	1	17	29,801	3	1	1	23
Number of all male teams	34,145	5	8	0	225	29,801	149	578	0	21,415
CPC technology shares										
Human necessities	34,145	23%	33%	0%	100%	29,801	22%	21%	0%	100%
Performing operations; transporting	34,145	26%	34%	0%	100%	29,801	20%	18%	0%	100%
Chemistry; metallurgy	34,145	8%	21%	0%	100%	29,801	12%	15%	0%	100%
Textiles; paper	34,145	2%	10%	0%	100%	29,801	1%	6%	0%	100%
Fixed constructions	34,145	9%	22%	0%	100%	29,801	5%	11%	0%	100%
Mechanical engineering; lighting; heating; weapons; blasting engines or pumps	34,145	13%	26%	0%	100%	29,801	10%	14%	0%	100%
Physics	34,145	11%	24%	0%	100%	29,801	17%	16%	0%	100%
Electricity	34,145	9%	21%	0%	100%	29,801	13%	15%	0%	100%

Table 1. Summary statistics comparing counties with and without women inventors, cont'd

	Counties with women inventors = 0					Counties with women inventors > 0				
			Std.					Std.		
	N	Mean	Dev.	Min	Max	N	Mean	Dev.	Min	Max
CPC technology indicators										
Human necessities	34,145	46%	50%	0%	100%	29,801	85%	35%	0%	100%
Performing operations; transporting	34,145	51%	50%	0%	100%	29,801	84%	36%	0%	100%
Chemistry; metallurgy	34,145	20%	40%	0%	100%	29,801	68%	47%	0%	100%
Textiles; paper	34,145	5%	21%	0%	100%	29,801	28%	45%	0%	100%
Fixed constructions	34,145	22%	41%	0%	100%	29,801	58%	49%	0%	100%
Mechanical engineering; lighting; heating; weapons; blasting engines or pumps	34,145	31%	46%	0%	100%	29,801	71%	45%	0%	100%
Physics	34,145	30%	46%	0%	100%	29,801	76%	43%	0%	100%
Electricity	34,145	23%	42%	0%	100%	29,801	69%	46%	0%	100%
Number of women with...										
Bachelor's degrees	34,145	1,310	1,572	17	33,362	29,801	14,556	33,026	34	761,572
Master's degrees	34,145	521	664	0	22,797	29,801	6,252	14,227	3	287,419
PhDs	34,145	42	70	0	1,483	29,801	700	1,854	0	40,577

Note: N is the number of observations, Mean is the sample average, Std. Dev. is the standard deviation, Min is the sample minimum, and Max is the sample maximum. Labor force, team size and number of all male teams are specified in counts.

3.1 Empirical Model

We allow counts of women inventors (*y*) in county *i* to be distributed,

$$y_i \sim Poisson(\mu_i), \tag{1}$$

where $\mu_i = \exp(\mathbf{X}_i \mathbf{B} + v_i)$ is the mean woman inventor frequency, and $e^{v_i} \sim Gamma\left(\frac{1}{\alpha}, \alpha\right)$(Stata, 2019, p.1635). Equation (1) explains counts of women inventor-patentees by a vector of independent variables (**X**), estimable parameters (**β**), an unobserved error (v_i), and the overdispersion parameter α. When $\alpha = 1$, y is Poisson distributed and the zip model is most efficient. When $\alpha > 1$, y is distributed by a negative binomial process and the zinb model is more appropriate.

There are three elements to the zinb model. The first is a probability density function of observing county *i* with zero women inventors (*y*):

$$\Pr(y_i = 0) = F_i + (1 - F_i) f(y_i = 0) \tag{2}$$

(Stata, 2019, p.1635, 2859; Raihan et al., 2019). Note that equation (2) models the probability of **not** observing a woman inventor-patentee, which is critical when interpreting the results. Also, there are two terms in equation (2), indicating two possible explanations for observing a county with zero women inventors. The first term, F_i, is the probability of observing a structural-zero county, which follows a logistic distribution function where $F_i = \frac{\lambda_i}{1+\lambda_i}$ and $\lambda_i = \exp(\boldsymbol{Z}_i \Delta)$. Recall that structural-zero counties are those that never had a woman inventor. The second term in (2), $(1 - F_i) f(y_i = 0)$, follows a negative binomial distribution of women inventors. These counties have observed women inventor-patentees. Some of these counties, however, do not consistently have women inventor-patentees every year. Hence, they are termed observational-zero counties.

Once the model determines that a county had or currently has women inventors who patent, a second probability density function calculates the probability of observing the number of such women in non-structural zero counties, given by,

$$\Pr(y_i > 0) = (1 - F_i) f(y_i). \tag{3}$$

Importantly, equation (3) contains the third element of the zinb model, link function $f(y_i)$. The link function provides the functional form of the nested distribution as a function of the mean (μ) and overdispersion (α) parameters. The model estimates μ and α by a regression that assumes a negative binomial distribution, $\boldsymbol{\Gamma}(.)$. The link function is given by,

$$f(y_i) = \Pr(Y = y_i | \mu_i, \alpha) = \frac{\Gamma(m + y_i)}{\Gamma(y_i + 1)\Gamma(m)} p_i^m (1 - p_i)^{y_i}, \tag{4}$$

where $m = 1/\alpha$ and $p_i = 1/(1 + \alpha\mu_i)$. Substituting (4) into (2) and (3), adding (2) and (3) together, and taking logs forms the following log-likelihood function:

$$lnL = \sum_{i \in S} w_i \ln\{F_i + (1 - F_i)\, p_i^m\} + \sum_{i \notin S} w_i\{\ln(1 - F_i) + \mathbf{\Gamma}(m + y_i)$$

$$-\mathbf{\Gamma}(y_i + 1) - \mathbf{\Gamma}(m) + m\ln p_i + y_i(1 - p_i)\}. \qquad (5)$$

The zinb model in equation (5) estimates simultaneously the probabilities given in (2) and (3) in a single log-likelihood function, nesting the negative binomial distribution of non-zero counties, $\mathbf{\Gamma}(.)$, inside the logistic distribution, F_i. The index term S in the summations of equation (5) identifies the set of counties that do not have women inventors ($y_{it} = 0$), and w_i are weights (Stata 2019, p.2859).

3.2 Empirical Application

We first explore how women's educational attainment in a county affects the probability that a county has its first women inventor-patentee. Specifically, we test three categories of female educational attainment: number of women with bachelor's degrees (bachelors), master's degrees (masters), and PhDs (phd). We expect that education has a positive and increasing influence on the likelihood of a county having a women inventor-patentee.

We specify the logistic distribution function shown in equation (2) to account for time (1990-2019):

$$\lambda_{it} = \exp(\boldsymbol{Z}_{it}\Delta) = \exp\left(z'_{itj}\delta\right) = \exp\,(\delta_0 + \delta_{fbachelors} fbachelors_{it} + \delta_{fmasters} fmasters_{it}$$

$$+\delta_{fphd} fphd_{it} + \delta_t YearFE_t + \delta_l StateFE_l), \qquad (6)$$

where subscript *i* refers to counties, *t* refers to time, subscripts *j = bachelors, masters and phd* and *l*=1,...,51 for all U.S. states plus the District of Columbia. Equation (6) tests whether educational attainment influences the likelihood of a county **not** having a woman inventor-patentee after accounting for time-invariant heterogeneity at the state level (*StateFE*), as well as common shocks to all U.S. counties in a given year (*YearFE*). States vary in their policies to incentivize business investment, which we hypothesize contributes to the participation by women in the innovation ecosystem.[8] The model includes intercept term (δ_0) and robust standard errors at the county level.

For non-zero counties, we specify two equations to investigate different facets of a county's economic environment. To that end, we specify

[8] We run a Wald test to determine the joint significance of state fixed effects (FEs). The null hypothesis, H_0, assumes that all state FEs are simultaneously equal to zero. The resulting Wald test gives us a chi-squared value of 3164.76 with 100 degrees of freedom (state FEs appear twice in the model: first in the logit and again in the negative binomial estimation) with p-value of 0.000. As a result, we reject the null hypothesis as the coefficients are not simultaneously equal to zero. Thus, including state FEs results in a statistically significant improvement in the model fit.

$$\mu_{it} = \exp\left(\mathbf{X}_{it}\mathbf{B}\right) = \exp\left(x'_{itj}\beta = \beta_0 + \beta_{LF}LF_{it} + \beta_{PCI}PCI_{it} + (\beta_{TS}TS_{it} + \beta_{TS^2}TS^2_{it}) + \beta_{AM}AM_{it} + \beta_{\%cpc}\%cpc_{it} + \beta_{dcpc}dcpc_{it} + \beta_t Year_t + \beta_{state}State_l\right) \qquad (7)$$

where subscript *i* refers to counties, *t* refers to time (1990-2019), and *j* = *Labor force counts (LF), Per-capita Income (PCI), inventor team size (TS), all male inventor teams (AM), technology field shares (%cpc) and presence of a technology field (dcpc)*. To control for unobserved heterogeneity, we include year and *l*=1,…,51 state fixed effects. We include robust standard errors to account for heteroskedasticity and serial correlation. Equation (7) uncovers systematic empirical relationships between a given county's labor force, per capita GDP, and its number of women inventors, controlling for other important factors.

We estimate equations (6) and (7) simultaneously using the log-likelihood function specified by (5) and the STATA package *zinb*.

4. Discussion

Table 2 provides selected results from the zinb econometric regression.[9]

Table 2. Selected zinb regression results

Variable	IRR*	p-value
Negative Binomal		
Economic variables		
Labor force	1.000002	0.00
Per capita income (USD)	1.000019	0.00
Logit		
Bachelors	0.999537	0.01
Masters	0.999231	0.07
PhD	0.995192	0.00

*Note: incidence-rate ratios (IRR) are equal to the exponentiated beta coefficient.

Results from the logit model indicate that both the number of women with bachelors and those with PhDs in a given county are statistically significant correlates.[10] An increase in each reduces the probability that a county remains structurally zero with no women inventor-patentees. In other words, the presence of highly educated women in a county increases the probability of that county having its first woman inventor-patentee. Doubling the number of women college graduates in a county that never had a woman patentee correlates with increasing a county's likelihood of having its first woman inventor-patentee by 61% = (1,310*(1- 0.999537)). Note that 1,310 reflects the

[9] The full set of parameter estimates are available upon request.

[10] The parameter estimate of the Masters variable is only statistically significant at the 10% level.

county mean of women with a Bachelor's degree. Doubling the number of women with PhDs increases the likelihood of a county having its first woman inventor-patentee by 20% = (42*(1-0.995192). Again, 42 reflects the county mean of women with a PhD. Though the total effect of adding PhDs is smaller, at the margin it is ten times greater, (1- 0.995192) / (1-0.999537) = 10; that is, adding one additional woman with a PhD to a county has the same effect as adding ten women with bachelors' degrees.

Though these numbers are in the expected direction, the magnitude of these effects suggests that educational attainment is a small factor in determining whether a county is observed as a zero-women inventor-patentee county. The results allude to the likelihood that other factors (e.g., working conditions) play a larger part in ensuring that a county has women inventors.

Moving to the negative binomial results in the upper portion of Table 2, labor force and per capita income correlate with the number of women inventors in the expected direction. Counties with relatively large labor forces and high per capita income also tend to have more women inventors. The magnitude of these effects is statistically significant but economically small. While this opens the door to explore other factors that contribute to the proliferation of women inventor-patentees, a possible culprit for the small marginal effects is that the employment and wage data are aggregated across all employment fields, including those not typically associated with patenting. For example, the data groups the 'Professional, Scientific, and Technical Services' sector (NAICS 54) – which includes jobs that are more likely associated with patentees (e.g., computer and math occupations) – with jobs not associated with patenting, such as legal occupations. The lack of disaggregated employment data likely results in the underestimation of income and labor-force effects on the abundance of women inventors.

In summary, our empirical results corroborate the direction of the relationships of educational attainment with regard to the probability of a county having women inventor-patentees, and income and labor force with respect to the number of women inventor-patentees. The fact that the magnitudes of our results were small indicates that other factors likely contribute to women's participation, and that future studies will need disaggregated employment data to more precisely determine the relationship between labor factors and women inventors.

References

Cameron, A. Colin, and Pravin K. Trivedi. Microeconometrics: methods and applications. Cambridge university press, 2005.

Lambert, D., 1992. Zero-inflated Poisson Regression, With an Application to Defects in Manufacturing. *Technometrics* 34(1): 1-14.

Lord, D., Washington, S.P. and Ivan, J.N., 2005. Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accident Analysis & Prevention, 37(1), pp.35-46.

Raihan, A., P. Alluri, W. Wu, and A. Gan, 2019. Estimation of bicycle crash modification factors (CMFs) on urban facilities using zero inflated negative binomial models. *Accident Analysis and Prevention* 123: 303-313.

Stata, 2019. *Stata Base Reference Manual, Release 16*. A Stata Press Publication, College Station, Texas.