1 Literature

2 Ideas put forward by the literature

Explanations for the gender gap

- Gender-biased technological change (Black and Spitz-Oener, 2010).
- Differences in commuting patterns and transportation (Black et al., 2014; Liu and Su, 2020).
- Differences on job flexibility (Goldin, 2014).
- Structural change (Olivetti and Petrongolo, 2014, 2016)
- Changes in selection patterns (Mulligan and Rubinstein, 2008)

3 Main findings

Fact 1: there are substantial differences in the level of the gender gap across CZ Figure 1 illustrates variation of the gender gap across US CZ in 2020. The figure restricts to CZ with population densities above 1 person per square kilometer in 1950. In 2020, men had an unconditional average wage 19 (se 5) log-points larger than women's. The map however, shows that there are wide variations from this average across CZ. Men's wage advantage is below 14 log-points in the Northeast and most of the West Coast, while it is above 24 log-points in the parts of the South West. The standard deviation in gender-wage gap across the 625 CZ shaded in the graph is of 5 log-points, which represents 26% of the average national gap.

[make an argument here that these differences are economically significant]

Fact 2: there are substantial differences in the evolution of the gender gap across CZ

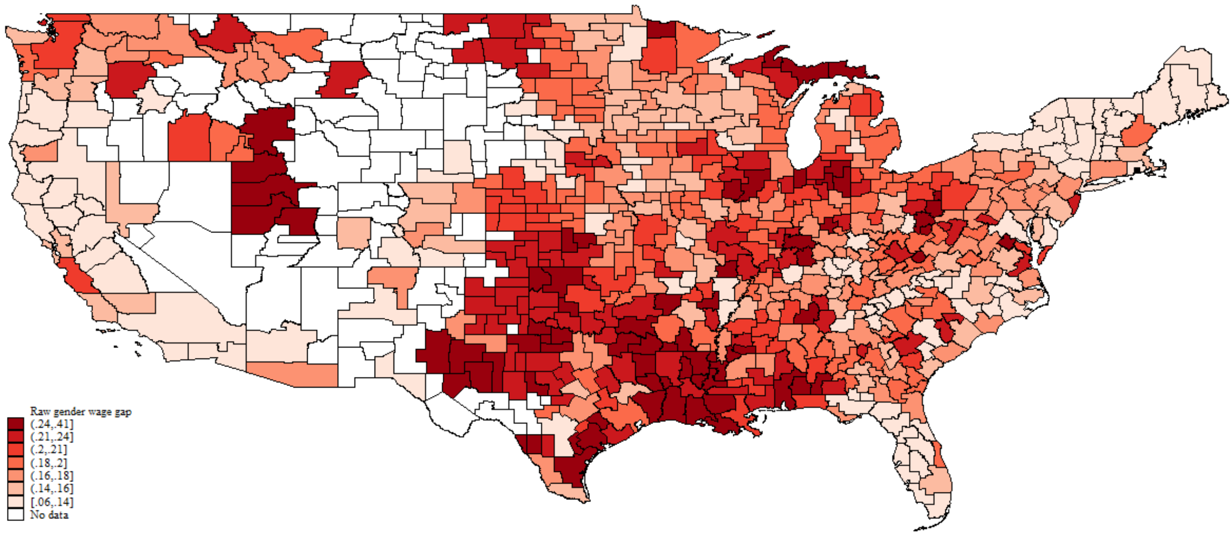
- Take two CZ as an example and show the evolution of the gender gap in these two places.
- Then show statistics on the change of the gender gap across places.

Fact 3: the gender gap has decreased the most in the densest CZ [add residualization at the individual level]

Fact 4: the relationship between population density and the gender gap has inverted over the period [write regression I am writing here]

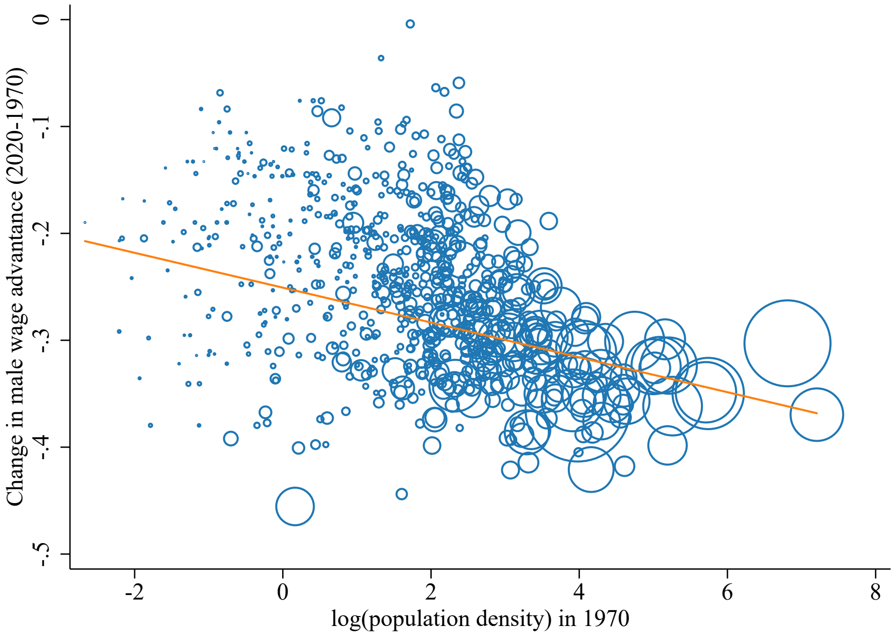
[graph of cross-sectional slope goes here]

Figure 1: The gender gap in the US in 2020



Note: darker colors denote higher relative wages for men. Figure restricts to czones with population densities above 1 person per km^2 and full-time year-round workers.

Figure 2: Change in male wage advantage in US CZ



4 Robustness of facts 3 and 4

4.1 Composition of the sample

- Results are robust to including all male and female workers.
- Results are also results for controlling for basic demographics

4.2 Weighting of regressions

Both facts are robust to alternative weighting mechanisms. Weighting only changes the timing of the decline in the population density gradient.

- This happens because decline of the gap happens first in denser places and then it decelerates.
- Places at medium levels of density speed the decline from 1990 on. This would explain the difference between between weighted and unweighted estimates.

4.3 Alternative measures of density

- Results are robust to using population as a measure of density.

5 Possible explanations

5.1 Increased sorting

- Regressions above do not control for observable characteristics.
- If women with stronger ability sort themselves:
  - Increasingly in denser cities.
  - This sorting is stronger than men.
- This would generate faster decrease of the gender gap in denser places.

Suppose wages are determine as follows:

$$y_{igr} = X_{igr}\gamma + \varepsilon_{igr}$$

taking averages by gender at the CZ level we have:

$$\bar{y}_{gr} = \bar{X}_{gr}\gamma + \bar{\varepsilon}_{gr}$$

therefore:

$$\bar{y}_{mr} - \bar{y}_{fr} = (\bar{X}_{mr} - \bar{X}_{fr})\gamma + \bar{\varepsilon}_{mr} - \bar{\varepsilon}_{fr}$$

so by running the regression:

$$\bar{y}_{mr} - \bar{y}_{fr} = \beta \log(density)_r + \bar{\varepsilon}_{mr} - \bar{\varepsilon}_{fr}$$

β would be reflecting the correlation between CZ population density and the average gap between male and female characteristics. This omitted variable problem is easily resolved by running the regression:

$$\bar{y}_{mr} - \bar{y}_{fr} = (\bar{X}_{mr} - \bar{X}_{fr})\gamma + \beta \log(density)_r + \bar{\varepsilon}_{mr} - \bar{\varepsilon}_{fr} \tag{1}$$

Things to have in mind

- These regressions impose the same return to observable characteristics for men and women in all CZ. Differential returns across CZ will go into the residual.

Procedure

Aggregate level data I run the regression

$$\bar{y}_{mr} - \bar{y}_{fr} = \alpha_t + (\bar{X}_{mr} - \bar{X}_{fr})\gamma_t + \beta_t \log(density)_r + u_{rt}$$

where I allow the return to observable characteristics to vary by year. The main interest is looking at the resulting evolution of β_t.

Individual level data This just allows for a more flexible variation on the returns of age birth place. Here the estimation is done in two steps:

1. Estimate the regression:

$$y_{igr} = X_{igr}\gamma + \lambda_{gr} + \varepsilon_{igr} \tag{2}$$

2. Compute CZ-adjusted wage gap:

$$\tau_r = \lambda_{mr} - \lambda_{fr}$$

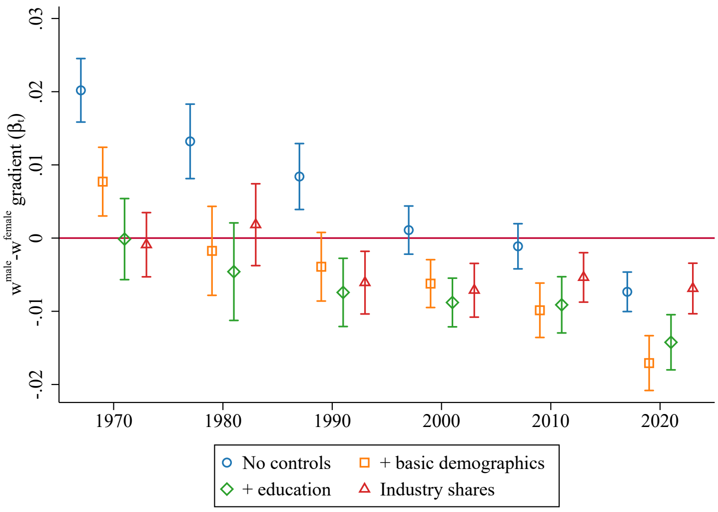
3. Run the regression:

$$\tau_r = \alpha_t + \beta_t \log(density)_r$$

I prefer this method as it exploits the individual level data in a richer way.

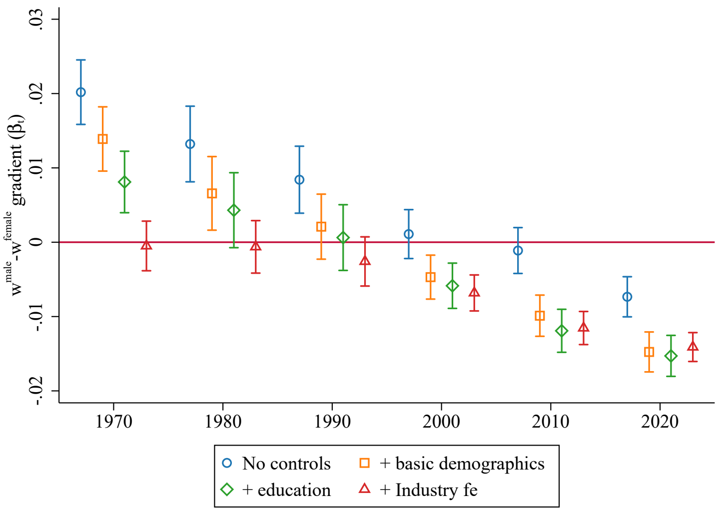
Results Overall individual level characteristics have limited value in accounting for the cross-sectional gradient and time-variation. Industry-level dummies are much more successful in accounting for the 1970-1990 period.

Figure 3: Coefficient on population density β_t controlling for worker characteristics



Note: figure restricts to CZ with more than 1 people per km^2. The regressions are done on data aggregated at the CZ level. Bars show 95% robust confidence intervals.

Figure 4: Coefficient on population density β_t controlling for worker characteristics



Note: figure restricts to CZ with more than 1 people per km^2. The regressions are done on data aggregated at the CZ level. Bars show 95% robust confidence intervals.

5.2 Changes in CZ industrial structure

Figure 4 suggests that changes in the CZ industrial structure can go a long way in accounting for the cross-sectional variation. Here I do several exercises to explore this possibility.

5.2.1 Some national level facts

Women are initially concentrated in low-pay industries See here => regions specialized on these high-pay industries will show a higher gender gap.

Getting a workable definition of a high-wage 70s industry A high wage industry is one that has a high worker-adjusted average pay in 1970. To be more precise, using individual level data in 1970 I run:

$$y_i = X_i\beta + \lambda_s \tag{3}$$

where s denotes the industry. I define an industry as having high-pay if they are in the top quartile of the λ_s distribution. When computing the quartiles, industries are weighted by employment share so that in 1970 each quartile accounts for 25% of the national level.

High pay industries were disproportionately concentrated in denser places in 1970 (graph)

High pay industries are in decline at the national level (graph) decline at the national level starts in 1990

High pay industries belong mainly to manufacturing (log file) the rest are mostly oil or utilities.

Employment share in highly paid industries accounts for most cross-sectional gradient on density during 1970-90 (graph) it also accounts for the time variation from 1970-90s. There's still something going on from 90 to 20

What can be happening:

- High wage industries are in decline in denser places... then the decline decline in male advantage comes from employment reallocation.
- It can be that at the start of the period, women are getting better access these industries. *I think for the 90's this seems to be the case.*

These industries continue to be highly paid industries See here

Denser CZ are more specialized in 70s high pay industries See here

70s high-pay industries decline disproportionately more in denser places

Women here

A On weighting

Here the basic question I want to answer is, can I have a good answer as to why I am not weighting.

Suppose wages are determined according to the model:

$$w_{ir}^{g} = \beta X_{ir}^{g} + \varepsilon_{ir}^{g} \tag{4}$$

where $\varepsilon_{ir}^{g} = \gamma_r^g + u_{ir}$ where γ_r^g and u_{ir} are independent.

$$\begin{aligned} \bar{w}_r^m - \bar{w}_r^f &= \beta(X_r^m - X_r^f) + \bar{\varepsilon}_r^m - \bar{\varepsilon}_r^f \\ &= \beta(X_r^m - X_r^f) + v_r \end{aligned}$$

note that if we assume that $var(\gamma_r^g) = \sigma_\gamma^2$ and $var(u_{ir}) = \sigma_u^2$.

$$\begin{aligned} var(v_r) &= var(\bar{\varepsilon}_r^m) + var(\bar{\varepsilon}_r^f) - cov(\bar{\varepsilon}_r^m, \bar{\varepsilon}_r^f) \\ &= 2\sigma_\gamma^2 + \sigma_u^2 \left(\tfrac{1}{N_m} + \tfrac{1}{N_f}\right) \end{aligned}$$

so, in the end I can test whether heteroskedasticity is a problem by running the regression (5) by OLS, extract the residuals and then run the regression:

$$\hat{u}_r = \alpha + \beta \left(\frac{1}{N_m} + \frac{1}{N_f}\right)$$

The results from this exercise give little justification for weighting the regressions. See this log-file.

B On the interpretation of the coefficients

The main findings come from regressions of the form:

$$\ln\left(w^{male} - w^{female}\right) = \alpha_t + \beta_t \ln(pop\_density)_{rt} + \epsilon_{rt}$$

without giving any causal interpretation to the coefficients, what is the interpretation of β_t?

B.1 Mathematical interpretation

First note that:

$$\beta_t = \frac{\partial \ln\left(w^{male}/w^{female}\right)}{\partial \ln pop\_density_{rt}}$$

thus β_t can be interpreted as the elasticity of the male wage advantage with respect to population density. Table 1 shows the estimated elasticities.

Table 1: Elasticities of male wage advantage to population density

Regression specification	**1970**	**1980**	**1990**	**2000**	**2010**	**2020**
Unweighted OLS	0.046***	0.030***	0.019***	0.003	-0.003	-0.017***
	(0.005)	(0.006)	(0.005)	(0.004)	(0.004)	(0.003)
Weighted by population	0.034**	0.007	-0.019*	-0.026**	-0.024***	-0.026***
	(0.013)	(0.013)	(0.008)	(0.008)	(0.006)	(0.005)
Observations	625	625	625	625	625	625

Note: Robust standard errors in parenthesis. Sample restricts to full-time year-round workers.. Table generated on 14 Aug 2020 at 15:39:24.

B.2 Economic interpretation

My basic results show that this elasticity:

- Has continually declined since the 1970s.
- It went from positive in 1970, to negative by 2020.

These elasticities are big and economically significant. Tables and 3 shows two ways of putting these numbers into perspective. Table 2 shows that moving form a CZ with a population density in the 25th percentile, to one in the 75th percentile translates into an increase of 4 p.p. in the male wage advantage in 1970. However, this same movement is associated with a decrease of 1.5 p.p. in the male advantage in 2020. These movements are equivalent to:

- A change of ∼10% relative to the average male advantage in each respective year.
- If we translate these figures to dollar amounts using as reference the annual wage income of the average full-time female worker, moving from the 25th to the 75th percentile in population density translates into $1.1 k relative gain for men in 1970, but a $0.6k relative loss in 2020.

Additionally, table 3 shows the estimated elasticities when I transform both the wage gaps and the CZ population density. If we focus on the unweighted OLS estimates, an increase of 1 sd in the log of population density is associated with 0.3 sd in the male wage advantage in 1970, but with a 0.2 sd deviations decrease in 2020.

Table 2: Male advantange changes implied by estimated elasticities

p.p. change in male advantage	**1970**	**1980**	**1990**	**2000**	**2010**	**2020**
Average male advantage	0.44	0.41	0.33	0.26	0.20	0.19
p75-p25	0.04	0.03	0.02	0.00	-0.00	-0.01
Relative male gain ($ USD)	1,116	793	534	73	-80	-559
p85-p15	0.08	0.05	0.03	0.00	-0.00	-0.02
Relative male gain ($ USD)	1,898	1,313	877	122	-134	-910
p90-p10	0.09	0.06	0.04	0.00	-0.00	-0.03
Relative male gain ($ USD)	2,379	1,644	1,088	153	-167	-1,159

Note: changes based on unweighted estimated elasticities in table 1. Sample restricted to full-time year-round workers. I compute the dollar figures using the wage of the average full-time year-round woman in my sample, assuming she worked 40 hrs a week during 40 weeks. All figures are in 2018 dollars. Table generated on 14 Aug 2020 at 18:35:47. Table generated with do file 2_analysis/code_files/create_IC_table.do

Table 3: β_t on standardized data

Regression specification	**1970**	**1980**	**1990**	**2000**	**2010**	**2020**
Unweighted OLS	0.330***	0.205***	0.176***	0.029	-0.030	-0.192***
	(0.036)	(0.040)	(0.048)	(0.044)	(0.042)	(0.036)
Weighted by population	0.244**	0.047	-0.173*	-0.301**	-0.283***	-0.300***
	(0.089)	(0.089)	(0.076)	(0.091)	(0.074)	(0.057)
Observations	625	625	625	625	625	625

Note: Robust standard errors in parenthesis. Sample restricts to full-time year-round workers.. Table generated on 14 Aug 2020 at 16:48:21.

References

Black, D. A., Kolesnikova, N., and Taylor, L. J. (2014). Why do so few women work in New York (and so many in Minneapolis)α Labor supply of married women across US cities. *Journal of Urban Economics*, 79:59–71.

Black, S. E. and Spitz-Oener, A. (2010). Explaining women's success: Technological change and the skill content of women's work. *Review of Economics and Statistics*, 92(1):187–194.

Goldin, C. (2014). A grand gender convergence: Its last chapter. *American Economic Review*, 104(4):1091–1119.

Liu, S. and Su, Y. (2020). The Geography of Jobs and the Gender Wage Gap. Technical report.

Mulligan, C. B. and Rubinstein, Y. (2008). Selection, investment, and women's relative wages over time.

Olivetti, C. and Petrongolo, B. (2014). Gender gaps across countries and skills: Demand, supply and the industry structure. *Review of Economic Dynamics*, 17(4):842–859.

Olivetti, C. and Petrongolo, B. (2016). The Evolution of Gender Gaps in Industrialized Countries. *Annual Review of Economics*, 8(1):405–434.