

The Spatial Clustering of Occupations

by

Henry Spencer Swift

B.A. Economics and Statistics
Swarthmore College, 2007

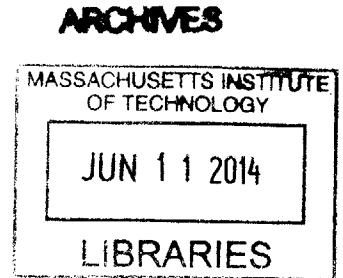
SUBMITTED TO THE DEPARTMENT OF ECONOMICS IN PARTIAL
FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2014

©Henry Swift. All rights reserved.

The author hereby grants to MIT permission to reproduce
and to distribute publicly paper and electronic
copies of this thesis document in whole or in part
in any medium now known or hereafter created.



Signature redacted

Signature of Author: _____

Department of Economics
May 28, 2014

Signature redacted

Certified by: _____

Michael Greenstone
3M Professor of Environmental Economics
Thesis Supervisor

Signature redacted

Accepted by: _____

Michael Greenstone
3M Professor of Environmental Economics
Chairman, Departmental Committee on Graduate Studies

The Spatial Clustering of Occupations

by

Henry Spencer Swift

Submitted to the Department of Economics on May 28, 2014
in partial fulfillment of the requirements for the degree of
Master of Science in Economics

Abstract:

Workers in similar occupations cluster, much like firms in similar industries. This may be due to firm clustering, but I propose a supply-side mechanism that may also provide an explanation. When workers face a risk of separation from a particular job, they will consider the other jobs available in a particular area in their location decision. Based on this theory I make three predictions. Workers will tend to cluster in areas where their skills are in high demand. They will be paid less in these areas, *ceteris paribus*. And demand shocks will affect workers' wages less, and employment more, in areas where their skills are in high demand. I test this mechanism using data from the decennial U.S. Census. I use O*NET data on occupational tasks to construct a measure of occupational distance. I then estimate labor supply curves to determine to test the predictions of the theory. I do not find substantial evidence for this mechanism.

Thesis Supervisor: Michael Greenstone

Title: 3M Professor of Environmental Economics

1 Introduction¹

Why do workers in similar occupations cluster and what does this mean for labor markets? Firms tend to cluster geographically, both in general and by industry (Ellison, Glaeser, and Kerr (2010)). Marshall (1920) famously proposed three explanations for this. First, firms in the same industry are more able to share technical knowledge if they are in the same location. Second, clustering of firms may reduce transportation costs to and from shared suppliers and customers. Third, firms may select from a larger pool of potential workers. A literature has empirically tested these propositions.

These mechanisms work primarily through the demand side of the labor market. They explain why it may be profitable for firms to co-locate by industry. Then workers in similar occupations tend to cluster because the firms that employ them tend to cluster. In this paper, I propose a mechanism that works through the *supply* side of the market and provides an additional explanation for the co-location of workers in similar occupations. Thicker markets provide outside opportunities to workers should their firm or industry face a negative shock. If there are other firms and industries that employ their skills in the same location, they have a better chance of being re-employed if they lose their job without having to incur moving costs. The entire local labor market available to a worker, not just the offered job or occupation, must therefore be considered when a worker makes a relocation decision.

Consider an example of this mechanism. A worker is deciding whether to move in order to take a job at a new technology start-up in Silicon Valley or Detroit. Start-ups have a high rate of failure, so the worker must take into account her expected utility if she loses the job. If moving costs are sufficiently high, the worker prefers to stay in that location if she loses her job. She will have a better chance of finding another firm that will pay for her skills if there is a large market for such jobs in that area. Other things equal, the worker prefers to take a start-up job in San Francisco to a start-up job in Detroit.

¹This material is based upon work supported by the National Science Foundation Graduate Research Fellowship.

This has a few implications. First, workers will tend to cluster where occupations similar to theirs are employed, even if firms do not. In the example, more workers will tend to choose San Francisco, where there are already many start-ups. Second, all else equal, wages will be higher at start-ups in Detroit, because the worker must be compensated for a lack of outside opportunities. However, firms may be clustered in San Francisco because productivity is higher there. Third, a demand shock will have a smaller effect on wages and a larger effect on employment in San Francisco than in Detroit. This is because workers can more easily substitute to alternative jobs and occupations in San Francisco.

In this paper, I will test these three predictions empirically. For any two occupations, I construct a measure of occupational distance using task data from O*NET. For each MSA-occupation pair I then construct a measure of *average* occupational distance (AOD) based on the occupations that are prevalent in that MSA. I use this as a measure of the quality of the outside options in that MSA for a worker in that occupation. I then estimate the effect of AOD on the intercept and slope of the labor supply curve using Bartik-style instruments.

There is a small literature on the effects of thick labor markets. These papers typically posit that there are increasing returns to search—larger labor markets result in more matches per worker. Bleakley and Lin (2007) demonstrate in Census microdata that workers are less likely to switch occupations or industries when they live in densely populated areas. They also find that this result is weaker for younger workers, who may be learning about their preferences and human capital. These workers switch occupations *more* in dense areas. Ellison, Glaeser, and Kerr (2010) find that pooling in shared occupational labor markets is an important determinant of coagglomeration of industries. However, they do not consider spillovers across similar occupations. Papageorgiou (2011) finds a result similar to Bleakley and Lin and provides a model that also takes into account moves across MSAs as well as across occupations.

This paper is also closely related to the literature on demand shocks. This

literature is largely interested in how workers are reallocated after a negative demand shock and what the incidence of a labor demand shock is. Bartik (1991) introduced the idea of using simulated employment instruments to measure responses to local labor demand shocks and Blanchard and Katz (1992) use this instrument in their exploration of labor market dynamics. More recently, Notowidigdo (2011) studied the incidence of local labor demand shocks, in part to answer the question of why low-skill workers are less likely than high-skill workers to move after a negative shock. He finds that the effect of a negative shock is largely to decrease housing prices and increase transfers, which leads to a relatively small effect on low-skill workers. Autor, Dorn, and Hanson (2011) use local industry shares and Chinese import penetration to construct a measure of labor demand shocks based on trade with China. As one would expect, they find that this reduces wages and employment. Similarly to Notowidigdo, they also find evidence of increased transfer payments. A number of papers in the development literature use agricultural labor demand shocks (e.g. rainfall) to study labor markets. Jayachandran (2006) finds that when workers are more constrained in their ability to migrate or borrow, they face larger wage losses in response to negative shocks.

I will also draw on the growing literature on tasks. Autor, Levy, and Murnane (2003) hypothesized that technological change caused routine tasks to be automated, leading to less demand for non-college workers. This was followed up by Autor and Dorn (2012) who propose that labor market “polarization” (decrease in middle-skill jobs, combined with increase in low- and high-skill jobs) can be attributed to exactly this automation of routine tasks. Acemoglu and Autor (2010) pull together much of the research on tasks and technology and formulate a model of selection of skill groups into tasks. Autor and Handel (2009) also uses unique data to explore the relationship between tasks and occupations.

Gathmann and Schönberg (2010) consider task-specific human capital by constructing a measure of distance between occupations based on tasks. They use German social security data to show that workers who switch occupations tend to switch to similar occupations. Tenure in the previous occupation de-

creases the average distance of a move. Wages before and after switching are more highly correlated if the two occupations are more similar. They also introduce the idea that average occupational similarity in a local labor market affects the ability of workers to move to similar occupations. In particular, they use average occupational similarity as an instrument for “task tenure”, essentially the match between a worker’s skills and her job. They also suggest that wage losses due to displacement may depend on labor market thickness. Poletaev and Robinson (2008) find similar results using factor analysis to construct a measure of occupation distance.

There is a large parallel literature on the effects of job loss on displaced workers. In the seminal paper in the literature, Jacobson, LaLonde, and Sullivan (1993) find that high-tenure workers suffer a long-term loss of 25% of earnings. They find that workers do worse when their local labor market has lower employment growth and when there are more adverse cyclical conditions. Finally, they show that manufacturing workers have much higher losses if they are not re-employed in the manufacturing sector, although the particular industry does not matter much. This evidence suggests that local labor market conditions matter for worker outcomes and the *kind* of re-employment opportunities that are available is also important. Many others have followed up this landmark paper.

In section 2, I formalize the intuition described above in a simple model of location and occupation choice. I derive three qualitative predictions from this model. In section 3, I lay out the empirical strategy I will use to test these predictions and the instrumental variables I will use. In section 4, I describe the data and the construction of the main variables. Section 5 presents the results. Section 6 concludes.

2 Theory

In this section I introduce a simple model of mobility and occupation choice. The model produces qualitative predictions that I will test empirically in the following sections.

The model focuses on a group of workers with a particular set of skills and human capital. Workers choose an occupation and location based on firms' wage offers, after which there may be a negative labor demand shock that causes the worker to find a new job. Workers are able to switch occupations, but not locations, and so their initial occupation-location choice must take into account the wages offered in the labor market beyond their immediate employment.

I find that workers tend to choose a location in which their occupation is similar to those in the rest of the labor market. As a result, these locations will have lower wages because the worker is being compensated through her outside options in the state of the world in which she loses her job. Finally, negative labor demand shocks have a smaller effect on the wages of those in markets with more similar occupations.

2.1 Set-up

There are three occupations, 0, 1, 2, and two locations, 0, 1. Let d_0 be the distance between occupations 0 and 2. Let d_1 be the distance between occupations 1 and 2. I will define what occupational distance means empirically in section 4, but it can be thought of as an index of how well human capital transfers between two occupations. If d_0 is high, that means a worker trained for occupation 2 will earn little in occupation 0. If d_0 is low, that means a worker trained for occupation 2 will earn a lot in occupation 0.

I consider a continuum of homogeneous workers with measure L , trained in occupation 2. Each location has a large perfectly competitive home industry with a linear production function. Each industry employs a single occupation. The home industry in location l employs workers in occupation l and so produces $y_h(d_l) > 0$ for every unit of labor employed, where $y'_h < 0$. The home industry in location l pays a wage of w_{hl} . I assume that $d_0 > d_1$. This assumption means that the workers we are considering are a better match for the home industry in location 1 than for the home industry in location 0 and therefore that they are more productive in the home industry at location 1.

A continuum of new firms with measure $N < L$ decides where to locate. Each firm hires exactly one worker in occupation 2. Each new firm i is endowed with technology that produces y_{nli} in location l . I assume that $y_{n1i} - y_{n0i} \sim F$, where F is a symmetric distribution function with infinite support. This ensures that some new firms will always want to locate in each location, no matter what wage they must pay. Firm i in location l pays a wage of w_{nli} .

The timing of the model is as follows:

1. New firms decide whether to locate at location 0 or location 1.
2. New firms then commit to wage offers w_{hli} .
3. Workers decide whether to locate at 0 or 1 and choose an industry and, if they choose the new industry, a firm i .
4. Each new firm shuts down with probability p .
5. Workers that were employed by firms that shut down are reallocated to the home industry.
6. Firms that did not shut down produce, firms receive profits, and workers receive wages.

2.2 Solution

To solve the model, I begin by calculating stage 3 expected wages for workers who choose each location-industry pair:

- If a worker chose industry 0 in location 0, she receives w_{h0} .
- If a worker chose industry 1 in location 1, she receives w_{h1} .
- If a worker chose firm i in industry 2 in location 0, she receives $(1 - p)w_{n0i} + pw_{h0}$.
- If a worker chose firm i in industry 2 in location 1, she receives $(1 - p)w_{n1i} + pw_{h1}$.

First, since the home industry is large, competition drives the wage down to marginal productivity, so $w_{hl} = y_h(d_l)$ for each l . Second, since workers are homogeneous, $w_{nli} = w_{nl}$ for each l and each i . Third, note that no worker will choose in stage 3 to work in industry 0 in location 0 because $w_{h1} > w_{h0}$. Assuming that all new firms will hire a worker, it must be the case in equilibrium that workers are indifferent between the new firms in each location and the home industry in location 1:

$$(1-p)w_{n1} + py_h(d_1) = (1-p)w_{n0} + py_h(d_0) = y_h(d_1). \quad (1)$$

We can now solve for the wage offers of the new firms in each location:

$$w_{n1} = y_h(d_1) \quad (2)$$

$$w_{n0} = \frac{y_h(d_1) - py_h(d_0)}{1-p}. \quad (3)$$

We can also solve for quantities of firms at each location. Firms earn $y_{n0i} - w_{n0}$ in location 0 and $y_{n1i} - w_{n1}$ in location 1, so we have that the proportion of new firms locating at 0 is

$$\frac{N_0}{N} = P(y_{n0i} - w_{n0} > y_{n1i} - w_{n1}) = F(w_{n1} - w_{n0}) = F\left(-\left(\frac{p}{1-p}\right)(y_h(d_1) - y_h(d_0))\right). \quad (4)$$

The proportion of new firms locating at 1 is

$$\frac{N_1}{N} = F\left(\left(\frac{p}{1-p}\right)(y_h(d_1) - y_h(d_0))\right). \quad (5)$$

The quantity of labor locating at 0 is

$$L_0 = N_0 = NF\left(-\left(\frac{p}{1-p}\right)(y_h(d_1) - y_h(d_0))\right), \quad (6)$$

while the quantity of labor locating at 1 is

$$L_1 = L - N_0 = L - NF\left(-\left(\frac{p}{1-p}\right)(y_h(d_1) - y_h(d_0))\right). \quad (7)$$

2.3 Predictions

This model produces three predictions that will be tested in this paper.

First, workers tend to cluster where the home industry allows them to be more productive, that is, where their occupational distance to the home industry's occupation is smaller, even though they are not initially employed by that industry. We can see this in the model through comparing equations 4 and 5. Since F is a CDF with support the whole real line, it is strictly increasing, so $\frac{N_1}{N} > \frac{N_0}{N}$. Workers will tend to locate where other occupations in that location offer them favorable wages even if the industry in which they choose to work is the same.

Second, wages tend to be higher in locations where the worker is a *worse* match to the home industry. We see this in the model through comparing equations 3 and 2, which shows that $w_{n0} > w_{n1}$. This comes about because new firms in location 0 must pay a premium to attract workers. This premium compensates workers for the state of the world in which the firm shuts down and the workers are reallocated to the home industry. Wages will tend to be lower in an occupation when the distance to the prevalent local occupations is lower.

Third, demand shocks have greater effects on the wages of workers when those workers are a worse match for the home industry. In the model, a demand shock is a firm shutting down. When a firm shuts down, the worker it employed must relocate to the home industry, which leads to a fall in wages. Comparing equations 3 and 2 again, we see that $w_{n0} - w_{h0} > w_{n1} - w_{h1} = 0$. Therefore, in the location where workers' occupations are further from the occupation employed by the home industry, they suffer a fall in wages, while in the location where workers are more productive in the home industry, they suffer no fall in wages at all.

2.4 Key Assumptions

There are several key assumptions that drive these results.

First, I assume an extreme form of moving costs. Workers are not allowed

to move after demand shocks are resolved. In other words, moving costs at that stage are infinite. If workers were allowed to move costlessly, then if they lost their job they would simply move to the location with the best matched home industry. In particular, it is important for the results that moving costs be large relative to occupational switching costs.

Second, there must be diminishing returns to labor in the new industry. Suppose that there were not and that new firms had a linear production function. Then, after a negative shock to some firms, other firms in the same location could soak up the newly free labor with no change in wages. In this model, I employ a highly diminishing return at the firm level by allowing each firm to employ only one unit of labor and fixing the total number of firms.

Third, there must be variation in home industries. That is, the productivity of workers must vary across locations due to the different industries that exist in each location. If any industry could locate in any location, then workers could simply reallocate to those industries that they are well-matched to.

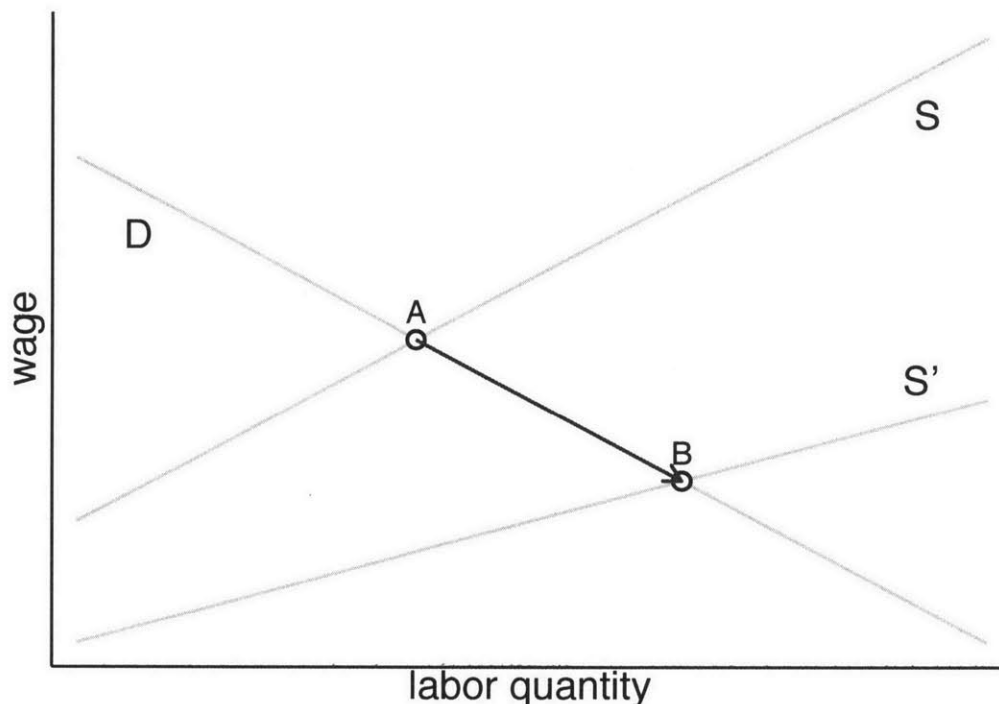
Fourth, firms must commit to a wage before workers make moving decisions. If workers made the moving decisions before firms committed to wages, firms would not have to make workers indifferent between the two locations. In each location, the wage paid by new firms would be equal to the wage paid by the home industry, $w_{nl} = w_{hl}$. This assumption fits better a growing industry that must attract workers to that location rather than a shrinking one. In reality, some workers are making moving decisions while others are constrained. It may be possible to study a dynamic model in which these assumptions are somewhat relaxed, but the qualitative predictions will still hold.

Finally, note that no risk-aversion is required in this model. New firms in the poorly-matched location are not paying a risk premium, but are compensating workers for a bad state of the world.

3 Research Design

In this section, I lay out a research design that will allow me to test all three predictions of the model.

Figure 1: Model predictions: supply shift.



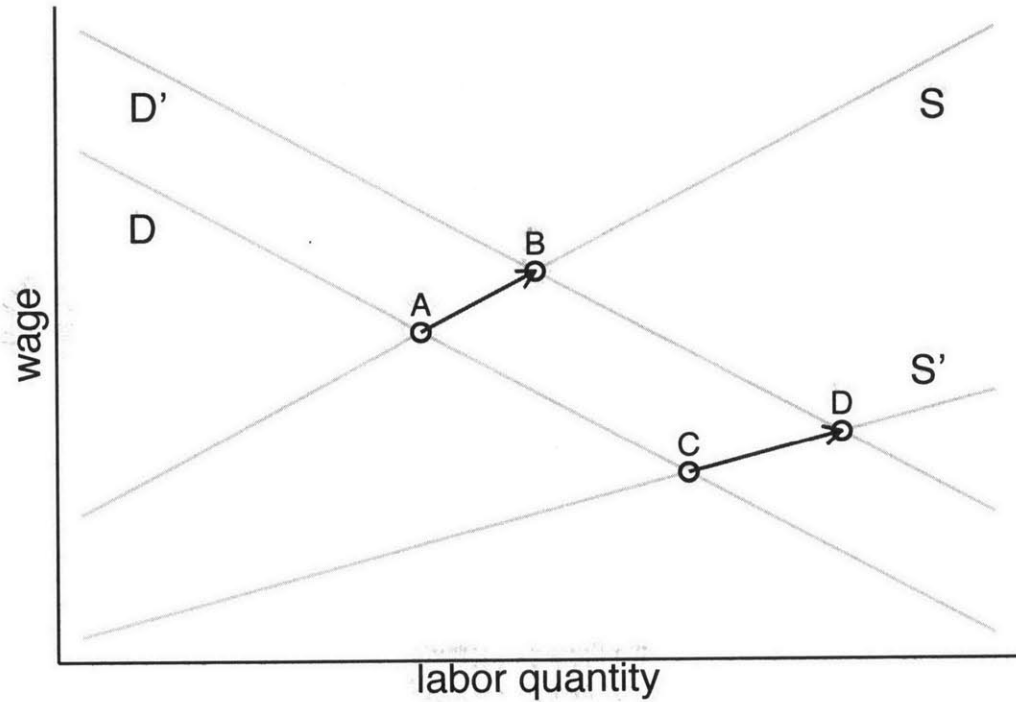
3.1 Supply and Demand Framework

Figures 1 and 2 demonstrate that the qualitative predictions of the model can be summarized in a simple supply and demand framework.

In Figure 1, S is a supply curve with high average occupational distance. S' is the supply curve after a decrease in average occupational distance. The equilibrium moves along the demand curve from A to B . As average occupational distance decreases, wages fall and quantity employed increases. The predictions for wages and labor quantity in this framework match the qualitative predictions of the model. This can be summarized by a shift *out* of the supply curve.

In Figure 2, S and S' remain the same, but we now consider shifts in the demand curve due to a shock to labor demand. The demand shock shifts the demand curve from D to D' . At S , the supply curve with high occupational

Figure 2: Model predictions: supply flattening.



distance, we move from point A to point B . At S' , the supply curve with low occupational distance, we move from point C to point D . The result is that at S' , relative to S , the demand shock results in a smaller effect on wages and a larger effect on employment. The supply curve is more elastic at S' than at S .

The supply and demand framework summarizes the predictions of the model. The first two predictions can be summarized as a shift out of the supply curve. The second two predictions can be summarized as a flattening of the supply curve. So testing the predictions of the model simply requires estimating the effect of average occupational distance on the position and slope of the labor supply curve.

3.2 Notation

I now introduce notation to be used in the empirical analysis. Let i be an industry, j be an occupation, l a location, and t a decade. Let the employment of industry i in occupation j , location l be L_{ijlt} . Let $L_{jlt} = \sum_i L_{ijlt}$. Let the industry share of industry i in occupation j and location l at time t be $Z_{ijlt} = \frac{L_{ijlt}}{L_{jlt}}$. Let w_{jlt} be the average wage in occupation j , location l , at time t . Let \bar{d}_{jlt} be the average occupational distance (AOD) of occupation j in location l at time t . I will discuss the construction of these variables in section 4.

3.3 Specification

I specify the following inverse log labor supply curve:

$$\ln w_{jlt} = \alpha_{jt} + \mu_{lt} + \theta_0 \ln N_{jlt} + \theta_1 \bar{d}_{jlt} + \theta_2 \ln N_{jlt} \cdot \bar{d}_{jlt} + \varepsilon_{jlt}. \quad (8)$$

Here α_{jt} and μ_{lt} are time-varying occupation and location effects. Together, these determine an occupation-location-year-specific intercept for the supply curve. These will serve an important role in identifying θ_1 and θ_2 , as I will discuss later.

Suppose that we can estimate this equation. How would we test for the predictions of the model? As discussed above, the first two predictions of the model can be summarized as an increase in AOD shifting in the supply curve, so that a given level of employment leads to a higher wage. This is reflected in Equation 23 by the hypothesis that

$$\theta_1 < 0. \quad (9)$$

The third prediction of the model can be summarized as an increase in AOD decreasing the elasticity of the supply curve. Since this equation is in logs, $\theta_0 + \theta_2 \bar{d}_{jlt}$ is the inverse elasticity of supply. Because we predict that when AOD is higher, the supply curve will be more inelastic, the prediction

of the model in Equation 23 that

$$\theta_2 > 0. \tag{10}$$

These two hypotheses summarize the qualitative predictions of the theory and so θ_1 and θ_2 are our main coefficients of interest.

Due to the strategy I will use to construct instruments, I will estimate 23 in changes rather than in levels. Taking changes in equation 23:

$$\Delta \ln w_{jlt} = \Delta \alpha_{jt} + \Delta \mu_{lt} + \theta_0 \Delta \ln N_{jlt} + \theta_1 \Delta \bar{d}_{jlt} + \theta_2 \Delta (\ln N_{jlt} \cdot \bar{d}_{jlt}) + \Delta \varepsilon_{jlt}. \tag{11}$$

We can still estimate θ_1 and θ_2 in this equation, but we will take advantage of multiple periods of data to construct instruments.

3.4 Instruments

In order to estimate this equation, I will have to overcome the potential endogeneity of $\Delta \ln N_{jlt}$ and $\Delta \bar{d}_{jlt}$. In particular, $\ln N_{jlt}$ is endogenous because we are estimating the supply curve in a system of simultaneous supply and demand equations. Changes in $\Delta \ln N_{jlt}$ may reflect either shifts in the supply curve or shifts in the demand curve. Change in AOD $\Delta \bar{d}_{jlt}$ may also be endogenous. For example, if workers in similar occupations also tend to value similar amenities, then an increase in those amenities will increase both $\Delta \varepsilon_{jlt}$ and $\Delta \bar{d}_{jlt}$, so they may be correlated.

Therefore, I will need instruments that are uncorrelated with shocks to supply ε_{jlt} . Following the literature beginning with Bartik (1991), I will construct a demand shock based on local industry shares that will serve as an instrument for the change in log employment $\Delta \ln N_{jlt}$ in the supply equation. In a unique departure from previous papers, I will use the same strategy to construct instruments for the change in AOD $\Delta \bar{d}_{jlt}$ as well as for the change in the interaction between labor quantity and AOD $\Delta (\ln N_{jlt} \cdot \bar{d}_{jlt})$.

To form instruments, I assume that the MSA-occupation industry share Z_{ijlt} is orthogonal to ε_{jlt} for all industries i , so that $Cov(Z_{ijlt}, \varepsilon_{jlt}) = 0$ for all i .

We can then form Bartik-style instruments as follows. Note that

$$\Delta \ln L_{jlt} \approx \sum_i Z_{ijlt} \Delta \ln L_{ijlt}. \quad (12)$$

Letting $L_{it} = \sum_{j,l} L_{ijlt}$ be the employment in the national industry i at time t , we replace the local industry growth rates with national industry growth rates to form the instrument:

$$\Delta \ln \hat{L}_{jlt} = \sum_i Z_{ijlt} \Delta \ln L_{it}. \quad (13)$$

Goldsmith-Pinkham, Sorkin, and Swift (2014) show that this is a valid instrument as long as Z_{ijlt} is orthogonal to ε_{jlt} and other technical assumptions hold. Intuitively, we can treat $d \ln L_{it}$ as a constant when the number of industries is not large relative to the sample size; then in the limit, this is just a linear combination of exogenous variables, which is itself exogenous. This shift-share strategy has been employed by a large literature, which Goldsmith-Pinkham, Sorkin, and Swift (2014) show has implicitly assumed that Z_{ijlt} is orthogonal to ε_{jlt} for all industries i even if not explicitly stated.

We can construct similar instruments for the two other terms in the supply curve: $\Delta \bar{d}_{jlt}$ and $\Delta (\ln N_{jlt} \cdot \bar{d}_{jlt})$. In particular, we have

$$\Delta \bar{d}_{jlt} = \sum_i Z_{ijlt} \Delta \bar{d}_{ijlt} \quad (14)$$

so we can construct an instrument

$$\Delta \hat{\bar{d}}_{jlt} = \sum_i Z_{ijlt} \Delta \bar{d}_{it}. \quad (15)$$

This instrument also relies on the assumption that Z_{ijlt} is orthogonal to ε_{jlt} for all industries i .

Finally,

$$\Delta (\ln N_{jlt} \cdot \bar{d}_{jlt}) = \Delta \ln N_{jlt} \cdot \bar{d}_{jlt} + \ln N_{jlt} \cdot \Delta \bar{d}_{jlt} \quad (16)$$

$$= \left(\sum_i Z_{ijlt} \Delta \ln L_{ijlt} \right) \cdot \bar{d}_{jlt} \quad (17)$$

$$+ \ln N_{jlt} \cdot \left(\sum_i Z_{ijlt} \Delta \bar{d}_{ijlt} \right) \quad (18)$$

$$= \sum_i Z_{ijlt} (\Delta \ln L_{ijlt} \cdot \bar{d}_{jlt} + \ln N_{jlt} \cdot \Delta \bar{d}_{ijlt}) . \quad (19)$$

Therefore we can use the following instrument for the interaction:

$$\Delta (\ln N_{jlt} \cdot \bar{d}_{jlt}) = \sum_i Z_{ijlt} \sum_{j',l'} (\Delta \ln L_{ij'l't} \cdot \bar{d}_{j'l't} + \ln N_{j'l't} \cdot \Delta \bar{d}_{ij'l't}) . \quad (20)$$

The key assumption is that industry shares are not correlated with supply factors. The rationale for this assumption is that industry share in a given occupation-location cell is determined entirely by demand-side factors like natural productivity differences, variation in agglomeration across locations, and so forth. However, there are reasons to suspect that this may not be a strong assumption. For example, industry shares may be correlated with location-based amenities. A greater share in the hospitality industry may indicate that the location is a particularly nice place to live. I attempt to deal with this issue in the following section.

3.5 MSA and Occupation Effects

I have two main reasons for using MSA and occupation effects in my preferred specifications. First, including MSA effects will absorb any correlation between the instruments and location-based amenities. Second, following Goldsmith-Pinkham, Kovak, Sorkin, and Swift (2014), an important way in which local labor demand shocks Z_{ijlt} may be endogenous is through wages at other locations. The inverse labor supply curve specified above omits wages at other locations and in other occupations. However, wages in other occupation-

locations cells affect labor supply for jl . If the constructed demand shock $\ln \hat{L}_{jlt}$ is correlated with wages at other locations (because demand shocks are geographically correlated, for example), then it will be correlated with the error term and induce bias.

In particular, if demand shocks are correlated between those occupation-locations that workers are likely to choose between, then a negative demand shock at jl is correlated with a decline in the wages of all the options in the worker's choice set. These correlated shocks do not lead workers to migrate or switch occupations, but still count as large shocks at the jl -level. As a result, the supply elasticity is estimated to be more inelastic than it really is.

In this paper, there is substitution across *both* locations and occupations, unlike in most demand shock papers that use only locations. Note that the error term includes wages at all other location-occupation pairs: $\Delta \varepsilon_{jlt} = f(\Delta w_{11t}, \dots, \Delta w_{JLt}, \xi_{jlt})$, where ξ_{jlt} is some underlying shock. Therefore if Z_{jlt} is correlated with wage changes at other locations, it is correlated with the error term.

Demand shocks may well be correlated across similar occupations, just as demand shocks are correlated across geographically close locations. However, this also presents a unique solution. Goldsmith-Pinkham, Kovak, Sorkin, and Swift (2014) shows that controlling for region fixed effects corrects for this bias if the underlying choice problem is a nested logit with nests at the region level. In this paper, I treat occupations and locations as choice nests and include occupation and location fixed effects. This may outperform region fixed effects because the worker's outside utility is allowed to vary at the MSA level instead of the region level. Similarly, it is also allowed to vary at the occupation level.

4 Data Overview

4.1 Labor Market Data

I use data on wages and full-time employment quantities from the 1980 – 2000 decennial Census via IPUMS (Ruggles, et al (2010)). The details of

the construction of my sample and these variables can be found in the Data Appendix.

4.2 Construction of Occupational Distance

In order to construct the average occupational distance (AOD) variable, I use occupational task data from O*NET, a survey of firms and workers in the United States. More details on O*NET can be found in the Data Appendix. Here I discuss the construction of variables from this data.

First, I demonstrate the variation in a selected sample of task ratings in order to convince the reader that these are in fact picking up important aspects of occupations. I have normalized the ratings here so that they have mean 0 and standard deviation 1. I display a sample of occupations and their values for selected ratings in Figures 3 and 4. In Figure 3, the y -axis shows the rating for “Analyzing Data or Information” and the x -axis shows the rating for “Operating Vehicles, Mechanized Devices, or Equipment”. As one would expect, actuaries, social scientists, and psychologists tend to rate high on the former and very low on the latter. Airline pilots, bus drivers, cab drivers, and mail carriers have high ratings for operation of vehicles. However, airline pilots are differentiated by having a relatively high rating on analyzing data as well as a higher rating for operation of vehicles, reflecting that this is an occupation requiring a greater overall degree of skill. Waitstaff and janitors rate low in both areas. Figure 4 shows similarly intuitive patterns, with social workers and psychologists rating high on care of others, while actuaries rate high of use of computers.

From the raw task ratings, of which I use 41, there are many ways to construct a measure of distance between occupations. Essentially, this comes down to defining a metric of the distance between any two vectors of task ratings, r_j and r_k . I construct a general and intuitive measure by using the L^p -norm between these two vectors:

$$d_{jk}^p = \left(\frac{1}{S} \sum_s |f_{sj} - f_{sk}|^p \right)^{1/p}. \quad (21)$$

Figure 3: Comparison of Task Ratings for Sample of Occupations

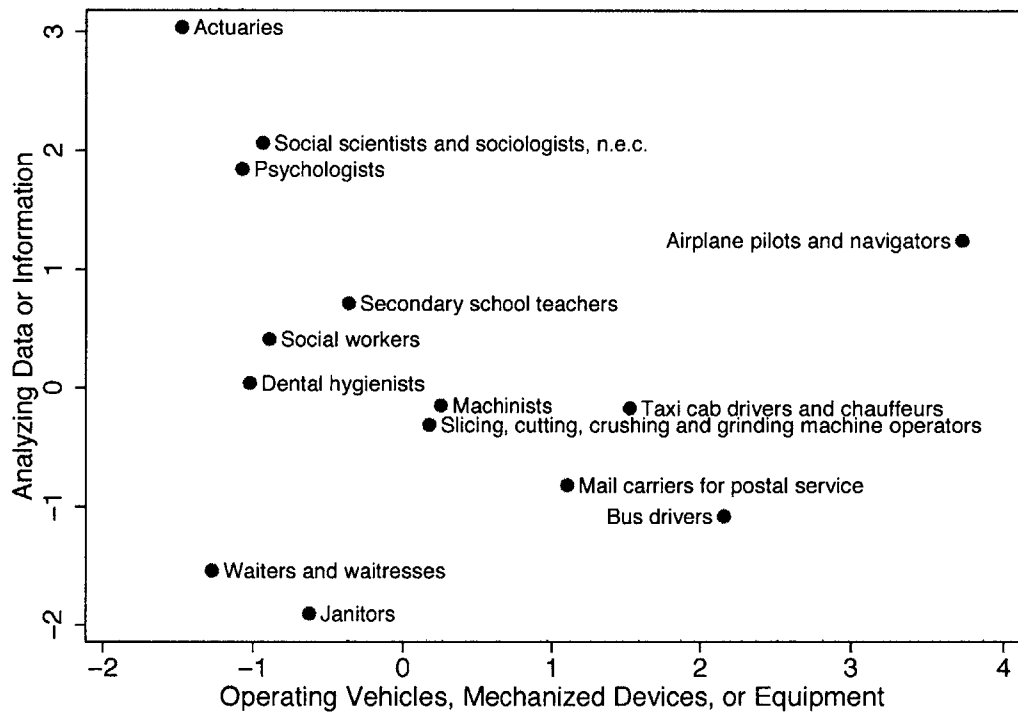
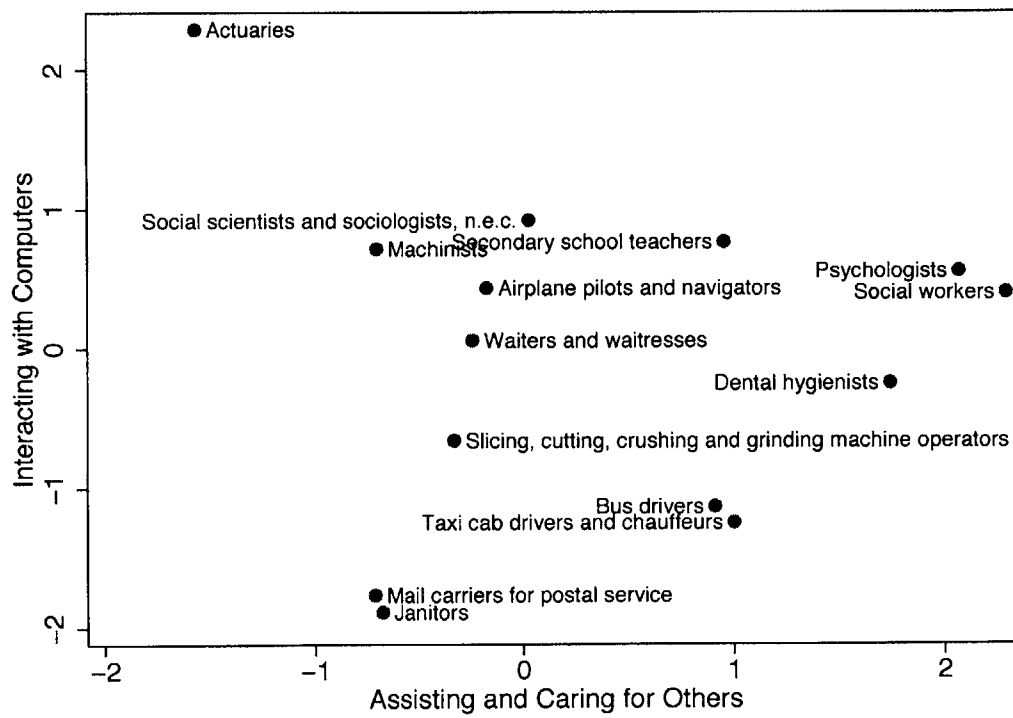


Figure 4: Comparison of Task Ratings for Sample of Occupations



The primary measure that I use is the root mean squared deviation generated by setting $p = 2$. Alternative measures I use are the mean absolute deviation generated by setting $p = 1$ and $p = 10$.

An important difference between this measure and the measure of occupational similarity used in Gathmann and Schönberg (2010) is that this allows for differences in the *skill level* of tasks. Gathmann and Schönberg’s measure is the angular separation between two task vectors. However, this throws away information contained in the level of tasks. For example, suppose there are two tasks, “vehicle operation” and “recording data”. Consider two occupations: “bus driver”, which has task ratings (3, 1), and “pilot”, which has task ratings (6, 2). The angular separation measure would report that these are perfectly similar occupations. In fact, the human capital required to operate an airplane is more advanced, and this is reflected in the level of task used. The construction of the O*NET survey is such that there is information encoded in the level and not just the direction of the task vector. In particular, the “level” measure in O*NET is anchored by different examples so that higher ratings in fact do represent different levels of human capital.

I now demonstrate the properties of the occupational distance index. By and large, these correspond quite well with our intuitions about what occupations are similar. Figures 5 and 6 display distance measures over a broad range of occupations. In Figure 5, similarity with “Economists, market researchers, and survey researchers” is on the y -axis and similarity with “Bus drivers” is on the x -axis. As we would expect, taxi drivers are very similar to bus drivers, while actuaries actuaries and financial managers are much more similar to economists. Figure 6 displays similarly intuitive results.

Finally, I construct the MSA-level average occupational distance (AOD) for occupation j and MSA l by taking the mean of occupational distance across all occupations, weighted by the fraction of occupational employment in that location:

$$\bar{d}_{jlt} = \sum_k \frac{L_{klt}}{L_{lt}} d_{jk}. \quad (22)$$

Figures 7 and 8 show AOD for these two occupations over a range of

Figure 5: Comparison of Occupational Distance for Sample of Occupations

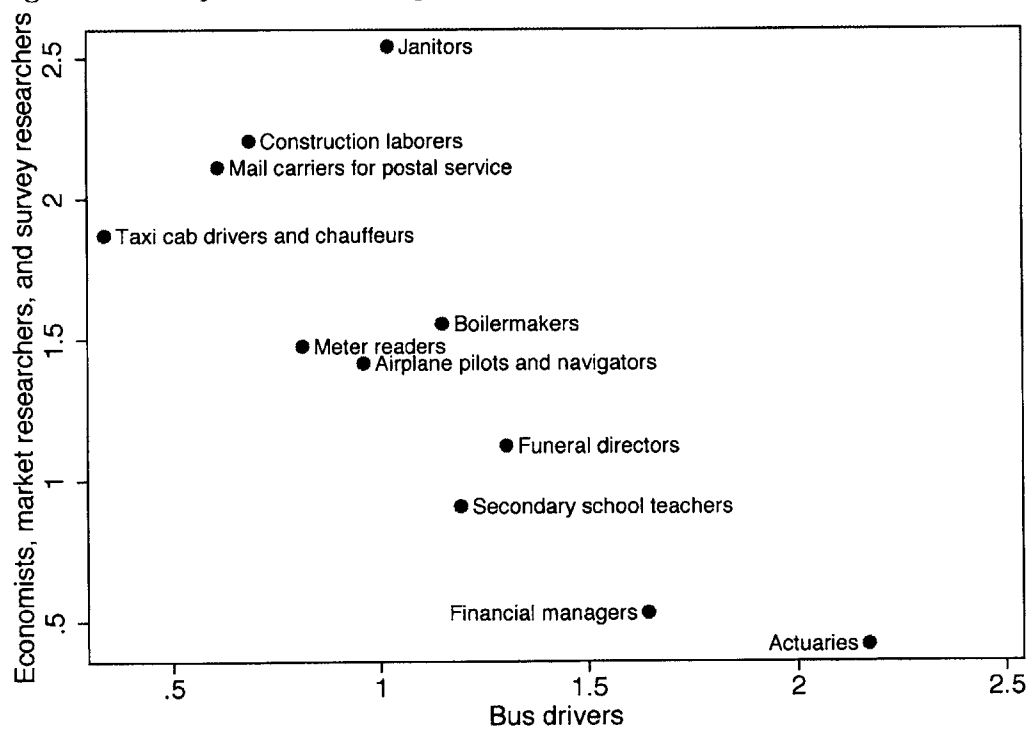


Figure 6: Comparison of Occupational Distance for Sample of Occupations

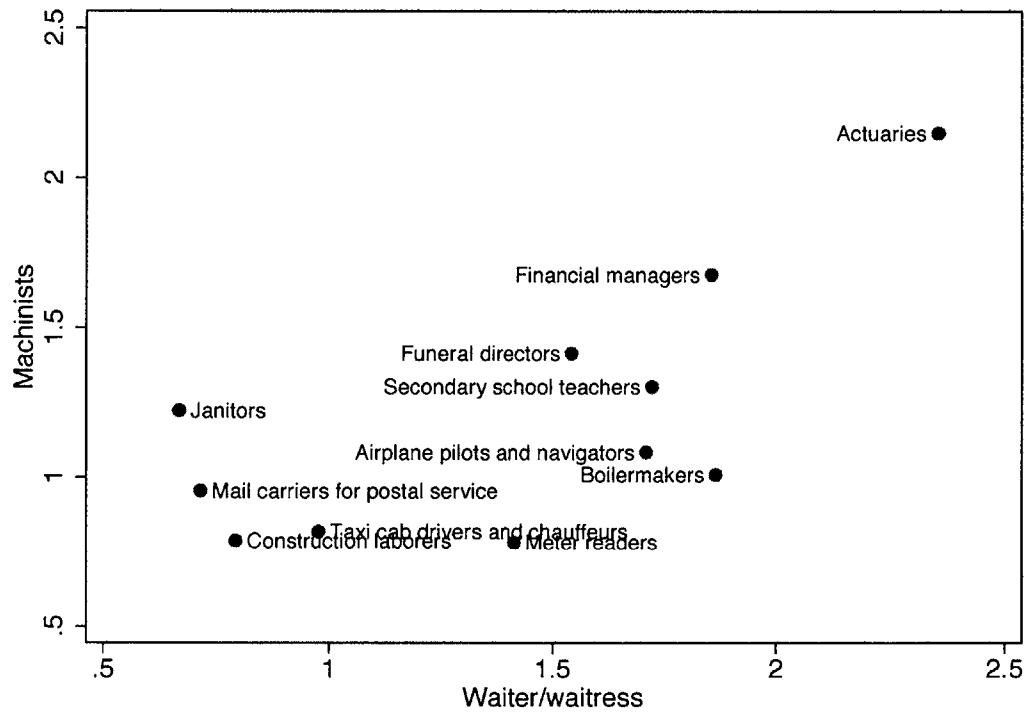
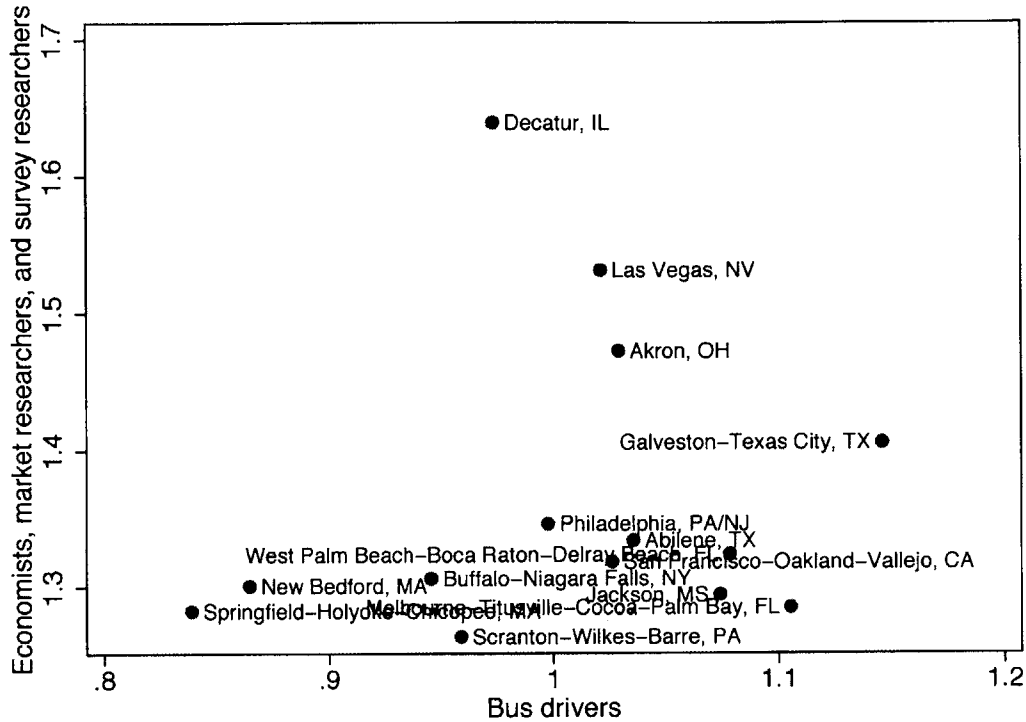
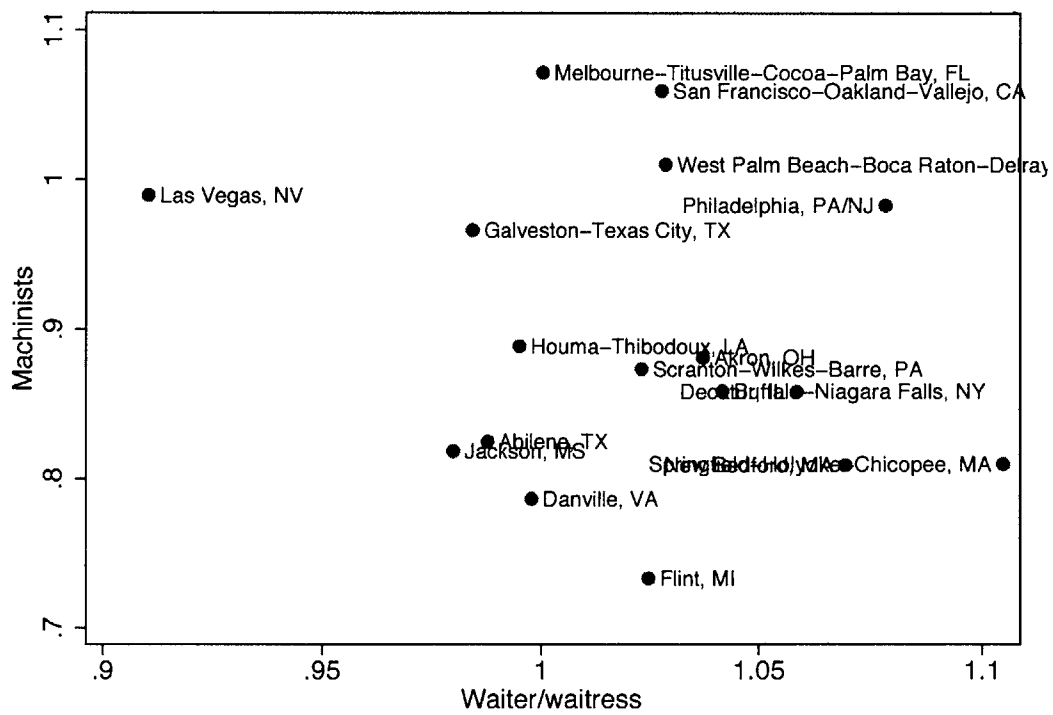


Figure 7: Comparison of AOD for Sample of MSAs



MSAs. In Figure 7, AOD for “Economists, market researchers, and survey researchers” is on the y -axis and AOD for “Bus drivers” is on the x -axis. The patterns shown in the figures correspond loosely to our intuitions. Note as well another key feature of the data: AOD tends to have quite different ranges across occupations. That is, most of the variation is across occupation and not within occupation. In Figure 8, AOD for “Machinists” is on the y -axis and AOD for “Waiter/waitress” is on the x -axis. The patterns shown in the figures correspond loosely to our intuitions. Machinists tend to have low AOD in traditionally industrial cities like Flint, MI. Waitstaff have low AOD in cities known for their service industries, like Las Vegas, NV.

Figure 8: Comparison of AOD for Sample of MSAs



5 Results

I now proceed to estimate the model specified in section 3. I provide summary statistics for all variables included in the regressions in Table 1.

Table 1: Summary statistics

	1980-1990		1990-2000	
	Mean	SD	Mean	SD
Base-year wages w_{jlt}	17.16	6.20	17.85	7.74
End-year wages w_{jlt}	17.37	7.31	19.07	9.41
Change in log wages $\Delta \ln w_{jlt}$	-0.001	0.169	0.052	0.173
Base-year employment (unweighted) N_{jlt}	918.7	4000.2	1062.2	4730.8
End-year employment (unweighted) N_{jlt}	1127.7	4900.1	1283.6	4863.6
Change in log employment $\Delta \ln N_{jlt}$	0.101	0.464	0.063	0.503
Simulated change in log employment $\Delta \ln \hat{N}_{jlt}$	0.189	0.121	0.180	0.080
Base-year AOD \bar{d}_{jlt}	5.979	1.000	6.272	1.000
End-year AOD \bar{d}_{jlt}	5.995	0.954	6.278	1.032
Change in AOD $\Delta \bar{d}_{jlt}$	0.016	0.276	0.006	0.240
Simulated change in AOD $\Delta \hat{\bar{d}}_{jlt}$	0.004	0.107	-0.074	0.060
N	53532	53532	57679	57679

Statistics (except employment) weighted by MSA-occupation employment.

As a first pass, I assume that \bar{d}_{jlt} is exogenous and estimate the equation

$$\Delta \ln w_{jlt} = \Delta \alpha_{jt} + \Delta \mu_{lt} + \theta_0 \Delta \ln N_{jlt} + \theta_1 \bar{d}_{jlt} + \theta_2 \bar{d}_{jlt} \Delta \ln N_{jlt} + \Delta \varepsilon_{jlt} \quad (23)$$

with and without occupation and location effects. As instruments for $\Delta \ln N_{jlt}$ and $\bar{d}_{jlt} \Delta \ln N_{jlt}$, I use the Bartik demand shock $\Delta \ln \hat{N}_{jlt}$ and this shock interacted with AOD $\bar{d}_{jlt} \cdot \Delta \ln \hat{N}_{jlt}$. If \bar{d}_{jlt} is exogenous, this will simply give the relative supply elasticities at various levels of \bar{d}_{jlt} . Here I have demeaned \bar{d}_{jlt} .

The reduced-form regressions of $\Delta \ln N_{jlt}$ on the instruments are given in columns (1)-(4) of Table 2. Each regression is run in all four combinations of these two variants: with and without MSA and occupation effects and using the periods 1980 – 1990 and 1990 – 2000. The Bartik demand shock seems to be a strong predictor of changes in labor quantity. The third row indicates that occupation-location cells with greater AOD tend to have a larger effect of demand shocks on changes in labor quantity. Note that this opposes the prediction of the model.

The reduced-form regressions of $\Delta \ln w_{jlt}$ on the instruments are given in columns (5)-(8) of Table 2. The Bartik demand shock is also a strong predictor of changes in wages. The third row shows evidence in the 1980 – 1990 period that occupation-location cells with greater AOD tend to have a larger effect of demand shocks on changes in wages. This is consistent with the model.

The 2SLS results are given in Table 3. The first row shows estimates of the inverse supply elasticity at the mean of \bar{d}_{jlt} varying between 0.193 and 0.801. The evidence for an effect of base period match quality on the inverse supply elasticity is mixed at best.

Also note that including MSA and occupation effects in columns (2) and (4) increases the estimate of the inverse supply elasticity. This is not what is predicted by Goldsmith-Pinkham, Kovak, Sorkin, and Swift (2014), who suggest that including such effects should reduce upward bias in the estimate of the inverse supply elasticity.

Table 2: First pass reduced-form regressions, assuming exogenous \bar{d}_{jlt} .

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\Delta \ln \hat{N}_{jlt}$	1.411^{**} (0.061)	0.432^{**} (0.074)	1.329^{**} (0.050)	0.400^{**} (0.094)	0.270^{**} (0.219)	0.214^{**} (0.035)	0.193^{**} (0.018)	0.313^{**} (0.038)
\bar{d}_{jlt}	0.009 (0.006)	0.043^* (0.017)	-0.059^{**} (0.011)	0.031^* (0.016)	-0.003 (0.003)	-0.021^{**} (0.008)	0.002 (0.003)	-0.025^{**} (0.005)
$\bar{d}_{jlt} \cdot \Delta \ln N_{jlt}$	0.103^{**} (0.022)	0.124 (0.067)	-0.008 (0.047)	0.154^* (0.061)	0.044^{**} (0.012)	0.079^{**} (0.018)	0.009 (0.016)	-0.029 (0.021)
N	53532	53532	53532	53532	57679	57679	57679	57679
Year	1980	1980	1990	1990	1980	1980	1990	1990
MSA Effects		X		X		X		X
Occupation Effect		X		X		X		X

Regressions are at the MSA-occupation level. They are weighted by MSA-occupation employment. Standard errors clustered by MSA. Stars indicate significance level. * : 5%, ** : 1%.

Table 3: First pass IV regressions, assuming exogenous \bar{d}_{jlt} .

	(1)	(2)	(3)	(4)
	Δw_{jlt}	Δw_{jlt}	Δw_{jlt}	Δw_{jlt}
$\Delta \ln N_{jlt}$	0.193** (0.013)	0.496** (0.089)	0.146** (0.016)	0.801** (0.206)
\bar{d}_{jlt}	-0.002 (0.002)	-0.038** (0.014)	0.012** (0.001)	-0.074** (0.013)
$\bar{d}_{jlt} \cdot \Delta \ln N_{jlt}$	0.027* (0.013)	0.016 (0.037)	0.009 (0.012)	-0.182* (0.092)
N	53532	53532	57679	57679
Year	1980	1980	1990	1990
MSA Effects		X		X
Occupation Effect		X		X

Regressions are at the MSA-occupation level. They are weighted by MSA-occupation employment. Standard errors clustered by MSA. Stars indicate significance level. * : 5%, ** : 1%.

I now move on to the main estimates. First, Tables 4, 5, 6, and 7 show the reduced-form OLS estimates of the endogenous variables regressed on the instruments. Table 8 then presents the 2SLS instrumental variables results.

Table 4 shows the regression of $\Delta \ln N_{jlt}$ on the instruments. As before, we note that the Bartik-style demand shock is a strong predictor of the change in labor quantity and the F -statistics are high. The results from 1980 – 1990 suggest that an increase in Bartik AOD reduces employment, as predicted from the model. However, this result is reversed in the 1990 – 2000 period.

Table 4: Main estimates, reduced-form regressions of change in employment on instruments.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	$\Delta \ln N_{jlt}$	$\Delta \ln N_{jlt}$	$\Delta \ln N_{jlt}$	$\Delta \ln N_{jlt}$	$\Delta \ln N_{jlt}$	$\Delta \ln N_{jlt}$	$\Delta \ln N_{jlt}$	$\Delta \ln N_{jlt}$
$\Delta \ln \hat{N}_{jlt}$	1.060** (0.056)	0.468** (0.086)	0.734** (0.055)	0.205* (0.097)	1.277** (0.044)	0.508** (0.092)	2.013** (0.072)	1.386** (0.091)
$\Delta \bar{d}_{jlt}$	-0.681** (0.059)	-0.012 (0.112)	-1.342** (0.102)	-0.566** (0.146)	-0.187** (0.046)	0.034 (0.092)	0.469** (0.065)	0.967** (0.110)
$\Delta \ln \widehat{N}_{jlt} \cdot \bar{d}_{jlt}$			0.051** (0.007)	0.063** (0.010)			-0.111** (0.029)	-0.133** (0.008)
N	53532	53532	53532	53532	57679	57679	57679	57679
F	292.05		194.28		461.72		357.81	
Year	1980	1980	1980	1980	1990	1990	1990	1990
MSA Effects		X		X		X		X
Occupation Effect		X		X		X		X

Regressions are at the MSA-occupation level. They are weighted by MSA-occupation employment. Standard errors clustered by MSA. Stars indicate significance level. * : 5%, ** : 1%.

Table 5 shows the regression of $\Delta \ln w_{j\mu}$ on the instruments. Again, this shows that the Bartik demand shock instrument is highly correlated with changes in log wages. Columns (1)-(2) and (5)-(6) provide some evidence that the predicted change in AOD instrument reduces wages, especially from 1990-2000.

Table 5: Main estimates, reduced-form regressions of change in wages on instruments.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	$\Delta \ln w_{jit}$	$\Delta \ln w_{jit}$	$\Delta \ln w_{jit}$	$\Delta \ln w_{jit}$	$\Delta \ln w_{jit}$	$\Delta \ln w_{jit}$	$\Delta \ln w_{jit}$	$\Delta \ln w_{jit}$
$\Delta \ln N_{jit}$	0.161** (0.022)	0.236** (0.038)	0.412** (0.037)	0.274*** (0.040)	0.183** (0.019)	0.223** (0.040)	0.342** (0.022)	0.323** (0.038)
$\Delta \bar{d}_{jit}$	-0.194** (0.015)	0.071 (0.059)	0.314** (0.048)	0.151* (0.072)	-0.087** (0.022)	-0.174** (0.060)	0.056* (0.026)	-0.067 (0.059)
$\Delta \ln N_{jit} \cdot \bar{d}_{jit}$			1.168* (0.578)	-1.229* (0.603)			-0.024** (0.002)	-0.015** (0.003)
N	53532	53532	53532	53532	57679	57679	57679	57679
F-stat	114.95		88.29		49.91		83.46	
Year	1980	1980	1980	1980	1990	1990	1990	1990
MSA Effects		X		X		X		X
Occupation Effect		X		X		X		X

Regressions are at the MSA-occupation level. They are weighted by MSA-occupation employment. Standard errors clustered by MSA. Stars indicate significance level. * : 5%, ** : 1%.

Table 6 presents the results from the regression of $\Delta \bar{d}_{jlt}$ on the instruments. $\Delta \hat{d}_{jlt}$ appears to be a strong instrument, although it is far stronger in the 1980-1990 period than in the 1990-2000 period.

Table 6: Main estimates, reduced-form regressions of change in AOD on instruments.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	$\Delta \bar{d}_{jit}$	$\Delta \bar{d}_{jit}$	$\Delta \bar{d}_{jit}$	$\Delta \bar{d}_{jit}$	$\Delta \bar{d}_{jit}$	$\Delta \bar{d}_{jit}$	$\Delta \bar{d}_{jit}$	$\Delta \bar{d}_{jit}$
$\Delta \ln \hat{N}_{jit}$	-0.149** (0.039)	0.195** (0.042)	-0.064 (0.044)	0.131* (0.051)	-0.047 (0.028)	-0.611** (0.074)	-0.050 (0.039)	-0.576** (0.065)
$\Delta \bar{d}_{jit}$	0.982** (0.040)	0.387** (0.068)	1.155** (0.065)	0.253** (0.097)	0.441** (0.055)	-0.708** (0.093)	0.438** (0.064)	-0.671** (0.095)
$\Delta \ln \widehat{N_{jit} \cdot d_{jit}}$			-0.013** (0.005)	0.015* (0.015)			0.000 (0.004)	-0.005 (0.005)
N	53532	53532	53532	53532	57679	57679	57679	57679
F	524.94		372.20		35.64		24.90	
Year	1980	1980	1980	1980	1990	1990	1990	1990
MSA Effects		X		X		X		X
Occupation Effect		X		X		X		X

Regressions are at the MSA-occupation level. They are weighted by MSA-occupation employment. Standard errors clustered by MSA. Stars indicate significance level. * : 5%, ** : 1%.

The final reduced-form table, Table 7, gives the results of a regression of the change in the interaction between employment and AOD on the instruments. This shows that the instrument for the interaction is strong, although the fact that the coefficient on the instrument is negative in the 1990-2000 period is troubling, because it should be positively correlated.

Table 7: Main estimates, reduced-form regressions of change in interaction on instruments.

	(1)	(2)	(3)	(4)
	$\Delta(\ln N_{jlt} \cdot \bar{d}_{jlt})$	$\Delta(\ln N_{jlt} \cdot \bar{d}_{jlt})$	$\Delta(\ln N_{jlt} \cdot \bar{d}_{jlt})$	$\Delta(\ln N_{jlt} \cdot \bar{d}_{jlt})$
$\Delta \ln \hat{N}_{jlt}$	3.468** (0.321)	1.304** (0.418)	12.300** (0.421)	4.874** (0.704)
$\Delta \bar{d}_{jlt}$	1.739* (0.858)	-3.248** (0.864)	7.536** (0.560)	1.757* (0.736)
$\Delta \ln \widehat{N_{jlt} \cdot \bar{d}_{jlt}}$	0.184** (0.061)	0.576** (0.062)	-0.699** (0.047)	-0.868** (0.056)
N	53532	53532	57679	57679
F	169.63		286.62	
Year	1980	1980	1990	1990
MSA Effects		X		X
Occupation Effect		X		X

Regressions are at the MSA-occupation level. They are weighted by MSA-occupation employment. Standard errors clustered by MSA. Stars indicate significance level. * : 5%, ** : 1%.

In Table 8, we can directly evaluate the predictions of the model. As I showed in section 3, the predictions of the model can be stated as two predictions in terms of the inverse labor supply curve.

First, an increase in AOD leads to a decrease in employment and an increase in wages. In this framework, that corresponds to a shift upward of the inverse labor supply curve, so that for a given level of employment, wages are higher. This corresponds to a positive effect of $\Delta \bar{d}_{jlt}$ on Δw_{jlt} in the inverse labor supply curve. We can see estimates of this effect in the second row of Table 8. Consider the estimates in columns (2) and (6) where no interaction is included. For the 1980-1990 period, I find no evidence of an effect. However, for the 1990-2000 period I do find evidence of a positive effect, as predicted by the model.

Second, an increase in AOD leads to an increase in the slope of the inverse supply curve; supply becomes less elastic. A shift in the demand curve should lead to a greater change in wages and a smaller change in employment. Estimates of this effect are in the third row of Table 8. The preferred estimates are in columns (4) and (8). In 1990-2000 there is no evidence of an effect, but in 1980-1990 there is evidence of a negative effect, contrary to the predictions of the theory.

Table 8: Main estimates, 2SLS estimates.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Δw_{jit}	Δw_{jit}	Δw_{jit}	Δw_{jit}	Δw_{jit}	Δw_{jit}	Δw_{jit}	Δw_{jit}
$\Delta \ln N_{jit}$	0.138** (0.017)	0.422** (0.079)	25.887 (27.207)	2.217** (0.636)	0.138** (0.016)	0.780** (0.230)	7.276 (12.537)	16.415 (44.223)
$\Delta \bar{d}_{jit}$	-0.102** (0.025)	0.197 (0.123)	37.375 (40.187)	1.648** (0.523)	-0.138* (0.054)	0.283* (0.141)	11.530 (20.486)	16.939 (46.363)
$\Delta (\ln N_{jit} \cdot \bar{d}_{jit})$			-4.673 (4.940)	-0.304** (0.098)			-1.116 (1.964)	-2.599 (7.065)
N	53532	53532	53532	53532	57679	57679	57679	57679
Year	1980	1980	1980	1980	1990	1990	1990	1990
MSA Effects		X		X		X		X
Occupation Effect		X		X		X		X

Regressions are at the MSA-occupation level. They are weighted by MSA-occupation employment. Standard errors clustered by MSA. Stars indicate significance level. * : 5%, ** : 1%.

6 Conclusion

In this paper, I present a supply-side mechanism for the clustering of workers in similar occupations. Workers choose to locate in larger labor markets for their skills because of the combination of the risk of job separation and moving costs. I construct a measure of average occupational distance to determine how well a worker's current occupation matches to other occupations employed in that area. I then construct Bartik-style instruments to isolate variation in labor demand and AOD unrelated to supply factors. I use these instruments to estimate the inverse labor supply curve and determine the effect of AOD on the position and slope of this curve. I do not find substantial evidence for the predictions of the theory.

One reason I may not find substantial evidence for the predictions is that the method may not be ideal for this theory.

The methods used here are very broad, taking into account every occupation and location in the country. It may be that this effect operates in a few occupations or locations and are washed out on average. Future work could study specific industries, in particular those industries where there is a high risk of separation, like technology start-ups or restaurants. This effect should be stronger in industries with a lot of risk because outside options in the labor market will be far more important.

There are other concerns with the econometrics used in this paper. I use only 13 industry shares as underlying instruments to construct 3 instruments for change in log employment, AOD, and their interaction. This may not provide enough variation to estimate the effects I am interested in. Another explanation could be that the fixed effects took out much of the variation in AOD. I noted above that most variation in AOD was across occupation rather than across location within occupation. By including occupation fixed effects, I am removing this variation.

References

1. Acemoglu, Daron and David H. Autor, "Skills, Tasks, and Technologies: Implications for Employment and Earnings" in *Handbook of Labor Economics*, v. 4, Orley Ashenfelter and David E. Card (eds.), Amsterdam: Elsevier, 2010.
2. Autor, David H. and David Dorn, "The Growth of Low Skill Service Jobs and the Polarization of the U.S. Labor Market," *American Economic Review*, 103(5), 1553-1597, 2013.
3. Autor, David H., David Dorn, and Gordon H. Hanson, "The China Syndrome: Local Labor Market Effects of Import Competition in the United States," working paper, 2011.
4. Autor, David H. and Michael Handel, "Putting Tasks to the Test: Human Capital, Job Tasks and Wages," MIT Working Paper, 2009.
5. Autor, David H., Frank Levy, and Richard J. Murnane, "The Skill Content of Recent Technological Change: An Empirical Exploration," *Quarterly Journal of Economics*, 118(4), 1279-1333, 2003.
6. Bartik, Timothy J., *Who Benefits From State and Local Economic Development Policies*, W.E. Upjohn Institute for Employment Research: Kalamazoo, Mich, 1991.
7. Blanchard, Olivier and Lawrence Katz, "Regional Evolutions", *Brookings Papers on Economic Activity*, 1-75, 1992.
8. Bleakley, Hoyt and Jeffrey Lin, "Thick-Market Effects and Churning in the Labor Market: Evidence from U.S. Cities," working paper, 2007.
9. Dorn, David, "Essays on Inequality, Spatial Interaction, and the Demand for Skills," University of St. Gallen Dissertation no. 3613, 2009.

10. Gathmann, Christina and Uta Schönberg, "How General is Human Capital? A Task-Based Approach," *Journal of Labor Economics*, 28(1), 1-49, 2010.
11. Goldsmith-Pinkham, Paul, Brian Kovak, Isaac Sorkin, and Henry Swift, "Is General Equilibrium Important for Studying Migration? Theory and Implications", unpublished draft, 2014.
12. Goldsmith-Pinkham, Paul, Isaac Sorkin, and Henry Swift, "The Econometrics of Shift-Share Instruments", unpublished draft, 2014.
13. Jacobson, Louis S, Robert J. LaLonde and Daniel G. Sullivan, "Earnings Losses of Displaced Workers", *The American Economic Review*, 83(4), 685-709, 1993.
14. Jayachandran, Seema, "Selling Labor Low: Wage Responses to Productivity Shocks in Developing Countries", *Journal of Political Economy*, 114(3), 538-575, 2006.
15. Marshall, Alfred, *Principles of Economics*, London: MacMillan, 1920.
16. Notowidigdo, Matthew J., "The Incidence of Local Labor Demand Shocks," working paper, 2011.
17. Papageorgiou, Theodore, "Worker Sorting and Agglomeration Economies", mimeograph, 2011.
18. Polataev, Maxim and Chris Robinson, "Human Capital Specificity: Evidence from the Dictionary of Occupational Titles and Displaced Worker Surveys, 1984-2000," *Journal of Labor Economics*, 26(3), 387-420, 2008.
19. Ruggles, Steven, J. Trent Alexander, Katie Genadek, Ronald Goeken, Matthew B. Schroeder, and Matthew Sobek. *Integrated Public Use Microdata Series: Version 5.0* [Machine-readable database]. Minneapolis: University of Minnesota, 2010.

20. Saez, Emmanuel and Thomas Piketty, “Income Inequality in the United States, 1913-2002” in *Top incomes over the twentieth century: A contrast between Continental European and English-Speaking countries*, ed. A.B. Atkinson and T. Piketty, Oxford: Oxford University Press, 2007.

7 Data Appendix

7.1 IPUMS

Data on wages and employment come from the IPUMS USA Census microdata extract (Ruggles, et al (2010)). I use data from the 1980, 1990, and 2000 Censuses.

I use the following sample selection procedures. I keep only MSAs that are identified in at least two decades in the 1980, 1990, and 2000 IPUMS data. I use only observations with non-military industries that are not “not applicable” or “experienced unemployed not classified by industry”. Also, I drop any observations with no occupation reported or if the occupation is reported as military or unemployed. I drop any non-prime-age individuals (17 and younger or 65 and older). I drop individuals who have non-zero business or farm income in order to make sure the hourly wage calculation is correct. I drop individuals living in group quarters or with no hours worked last year.

I use the IPUMS person-level weights in order to generate MSA-level totals and averages. For some MSAs, the MSA code is only partially available. I re-weight these observations by the inverse of the proportion of the MSA for which the code is available. Since the observations that are coded to the MSA are not a random sample of the MSA as a whole (Ruggles, et al (2010)), as a robustness check I also present results using only MSAs in which the codes are fully identified.

I use industry coded to 1990 categories by IPUMS in order to compare industries across years. I also generate coarser industry categories based on those identified in the IPUMS documentation.

I use occupation coded to 1990 categories by IPUMS as modified by Dorn

Year	Top Code (nominal \$)	P95 (2000 \$)	Mean P95-P100 (2000 \$)	Implied Ratio	Adjusted Top Code (nominal \$)
1980	\$75,000	\$87,131	\$131,771	1.51	\$113,425
1990	\$140,000	\$96,665	\$168,060	1.74	\$243,401
2000	\$175,000	\$110,859	\$228.277	2.06	\$360,354

(2009).

I define hourly wages as total wage income divided by total hours worked last year. Hours worked last year is defined to be weeks worked last year multiplied by usual hours worked per week last year.

I adjust top-codes to reflect the mean in the tail of a Pareto wage income distribution using data from table B3 of Piketty and Saez (2007):

Following Notowidigdo (2011), I top-code hourly wages at the top-coded wage income amount divided by $50 * 35$, the number of full-time hours worked in a year, but replace this amount by the same amount multiplied by the top-code adjustment, as above. I also construct a full-time employment indicator to be at least 30 hours usually worked per week and at least 35 weeks worked last year.

7.2 O*NET

Data on the tasks that are associated with occupations comes from the Occupational Information Network (O*NET) data. O*NET was created by the Employment and Training Administration and the Department of Labor as an informational resource for workers and job seekers.² I use version 16.0 of the O*NET database, released in July 2011.³

The O*NET data was collected by sampling firms “expected to employ workers in the targeted occupations”⁴. After a firm was sampled, a random sample of workers within the firm in the targeted occupations was surveyed. The original version of O*NET included ratings from “occupational analysts”

²More information about O*NET can be found at <http://www.onetcenter.org/>. More information about the data collection process can be found at <https://onet.rti.org/>.

³The data can be accessed at <http://www.onetcenter.org/database.html?p=2>.

⁴Source: <http://www.onetcenter.org/dataCollection.html>

who produced ratings based on a combination of materials. Subsequent releases have included more and more data from actual workers and “occupational experts”, who are taken from professional association membership lists.

O*NET provides occupational ratings in many different categories, including “Knowledge”, “Work Context”, and so forth. There is a rating category for “Tasks”, however these tasks are extremely specific and are not common across occupations. Therefore, I will use O*NET’s “Generalized Work Activities” as a measure of tasks. Table 9 gives a list of the 41 work activities that are rated.

I use only ratings from workers and occupational experts who answered a questionnaire about work activities.⁵ For each work activity, respondents were asked to answer two questions:

1. “How important is WORK ACTIVITY to the performance of your current job?”
2. “What level of WORK ACTIVITY is needed to perform your current job?”

The importance question is rated on a scale from 1 to 5, labeled with adjectives ranging from “Not Important” to “Extremely Important”. The level question is rated on a scale from 1 to 7, but each work activity is labeled with specific examples. For example, for the activity “Operating Vehicles, Mechanized Devices, or Equipment” the examples given are:

- 2: “Drive a car.”
- 4: “Drive an 18-wheel tractor trailer.”
- 6: “Hover a helicopter in strong wind.”

In my measure of occupational similarity I prefer using the ratings from the level question for two reasons. First, because the level question provided specific examples that anchor the respondent’s answers, it is potentially more

⁵Sample questionnaire: <https://onet.rti.org/pdf/GeneralizedWorkActivities.pdf>

Table 9: Work Activities

Getting Information
Monitor Processes, Materials, or Surroundings
Identifying Objects, Actions, and Events
Inspecting Equipment, Structures, or Material
Estimating the Quantifiable Characteristics of Products, Events, or Information
Judging the Qualities of Things, Services, or People
Processing Information
Evaluating Information to Determine Compliance with Standards
Analyzing Data or Information
Making Decisions and Solving Problems
Thinking Creatively
Updating and Using Relevant Knowledge
Developing Objectives and Strategies
Scheduling Work and Activities
Organizing, Planning, and Prioritizing Work
Performing General Physical Activities
Handling and Moving Objects
Controlling Machines and Processes
Operating Vehicles, Mechanized Devices, or Equipment
Interacting With Computers
Drafting, Laying Out, and Specifying Technical Devices, Parts, and Equipment
Repairing and Maintaining Mechanical Equipment
Repairing and Maintaining Electronic Equipment
Documenting/Recording Information
Interpreting the Meaning of Information for Others
Communicating with Supervisors, Peers, or Subordinates
Communicating with Persons Outside Organization
Establishing and Maintaining Interpersonal Relationships
Assisting and Caring for Others
Selling or Influencing Others
Resolving Conflicts and Negotiating with Others
Performing for or Working Directly with the Public
Coordinating the Work and Activities of Others
Developing and Building Teams
Training and Teaching Others
Guiding, Directing, and Motivating Subordinates
Coaching and Developing Others
Provide Consultation and Advice to Others
Performing Administrative Activities
Staffing Organizational Units
Monitoring and Controlling Resources

informative than the importance question, which is somewhat vague. Second, the level question is more closely related to the skill or human capital actually required to perform the task. Taxi drivers and airplane pilots may rate the importance of vehicle operation a 5, but they would presumably rate the level 2 and 6, respectively. So while the importance question would not differentiate them, the level question would.

7.3 Merging the IPUMS and O*NET Data

Linking O*NET with the IPUMS is a somewhat complicated process. The main challenge is that O*NET and IPUMS use two different occupational coding schemes. O*NET uses the O*NET-SOC 2010 coding scheme. IPUMS has implemented a coding scheme in the OCC1990 variable that is consistent over time. To connect these two coding schemes requires several steps.

First, I convert from O*NET-SOC 2010 to SOC 2010 simply by aggregating up to the SOC level. O*NET-SOC 2010 nests the SOC, so this simply requires taking averages of variables within the SOC occupation codes. As weights, I use the O*NET sample sizes. This is not ideal, since these sample sizes do not necessarily reflect shares of the occupation within the population. Second, I convert from SOC 2010 to SOC 2000. This conversion is somewhat more complicated. If a single SOC 2000 code maps to multiple SOC 2010 codes, I take a weighted average of variables (using O*NET sample sizes as weights) within the SOC 2000 code. If the SOC 2000 code maps to a single SOC 2010 code (that may itself have multiple SOC 2000 codes mapping to it), I simply use the value of the variable from that SOC 2010 code.

The third step in this process moves from SOC 2000 to the IPUMS Census 2000's OCCSOC variable. Unfortunately, this variable's coding scheme does not cleanly match up with SOC 2000, as the name would imply. However, SOC 2000 does nest OCCSOC, so again I take averages using O*NET sample size weights. Finally, I need to convert from OCCSOC to (David Dorn) OCC1990. In order to do this, I merge all of the task rating data into the IPUMS using OCCSOC. I then collapse the data to the (David Dorn) OCC1990 level, taking

averages of the task ratings. Essentially, this uses the IPUMS weights to convert OCCSOC averages into OCC1990 averages.

The key problem here is that each step adds potential measurement error to the data. Miscoded workers will tend to bias my results toward zero.