Agglomeration and Productivity: New Estimates and Macroeconomic Implications*

Morris A. Davis
University of Wisconsin
mdavis@bus.wisc.edu

Jonas D.M. Fisher
Federal Reserve Bank of Chicago
jfisher@frbchi.org

Toni M. Whited
University of Wisconsin
twhited@bus.wisc.edu

November 19, 2008

Preliminary and Incomplete

Abstract

We construct a dynamic general equilibrium model of cities and aggregate growth and use it to identify the effects of local agglomeration on aggregate per capita consumption. There is a fixed amount of land in each city; workers live and work in the same city and value consumption and land for housing; and identical firms in each city produce output using a Cobb-Douglas combination of land, labor and capital. The level of firm output in each city is also affected by Ciccone and Hall (1996)'s density externality. Our model predicts a relationship between wages, output prices, land prices, and labor input which we use to estimate the structural parameters of the model with panel data. We find a statistically significant and economically meaningful role for density in accounting for changes in output: Our estimates imply that along a balanced growth path, agglomeration accounts for 5 percent of growth in per-capita consumption.

Journal of Economic Literature Classification Numbers: E0, E1, E2, R1, R3
Keywords: Density of economic activity, growth, human capital externalities

*The views expressed herein are those of the authors and do not necessarily represent those of the Federal Reserve Bank of Chicago or the Federal Reserve System.

1 Introduction

There is widespread agreement that cities emerge because of *agglomeration* effects: workers become more productive when they are near other workers. Cities emerge because of agglomeration, and most aggregate growth actually occurs in cities. So, how important is local agglomeration for aggregate growth? To answer this question, we build and estimate a dynamic general equilibrium model of cities and aggregate growth in which city-based agglomeration affects per capita consumption growth. We estimate the structural parameters of our model using panel data and use these estimates to quantify the impact of local agglomeration on per capita consumption growth.

Our model embeds into the neoclassical growth model a version of Roback (1982)'s model of cities, augmented with agglomeration effects in the way proposed by Ciccone and Hall (1996). There is a representative household consisting of a unit measure of members. The large household structure allows for full risk sharing within each household, a standard device for studying complete markets allocations. After observing total factor productivity in each city, individual household members are randomly allocated to cities, where they work and consume goods and housing. Identical competitive firms in each city produce intermediate goods with perfectly mobile capital and local land and labor. The intermediate goods are used in the production of final consumption and investment. Land at each location is used in housing and production. We study the stationary competitive equilibrium of our model and show that along the balanced growth path per capita consumption growth depends on exogenous total factor productivity change and the average household or employment growth. The household-employment growth rate effect is determined by the net impact of agglomeration. We use our model and data on wages and prices for a sample of US cities to estimate the net impact on local productivity of agglomeration, thereby yielding the contribution to per capita consumption growth.

Our model predicts a specific relationship between wages, output and land prices, plus labor input, and exogenous total factor productivity. We exploit this relationship to estimate the key parameter underlying local agglomeration and its effect on aggregate growth. Our estimation uses three sets of data. Our first data set combines wage data from the CPS, output-price data from the Bureau of Economic Analysis

(BEA), and land-price data from a study by Davis and Palumbo (2008). This annual panel data set covers 42 MSAs over the 1985-2004 period. The second annual data set uses the same annual wage and output-price data, but uses data on house prices (rather than land prices) for 149 MSAs over the 1985-2006 period. We create the house price data by merging available house price indexes with information in the 1990 Decennial Census. Our third panel data set merges the BEA data on output prices with data on wage and house price exclusively from the Decennial Census for the years 1970, 1980, 1990, and 2000.

The twenty year panel sample of the first two data sets is particularly amenable to applying recent advances in dynamic panel data analysis. We use the Arellano and Bover (1995) estimation strategy on the annual data. Our key finding from these data is that local agglomeration effects are statistically significant and economically important, regardless of the data source we use. However, the exact estimates are sensitive to our choice of data. When we use the land data, we estimate that a doubling of output density increases output by 1.2 percent. Our estimates from the house-price data are higher; they suggest that a doubling of output density increases output by 7.5 percent. On the census data we explore several instrumental variables strategies and find [**TBD** xxx].

Depending on whether we use land or house prices we find between 1.6% and 9% of per capita consumption growth is due to local agglomeration. According to our model, Ciccone and Hall (1996)'s estimates of the agglomeration coefficient translate to a 6% contribution to per capita consumption growth. We prefer our focus on cities since they seem a more natural economic entity compared to the counties at the base of Ciccone and Hall (1996). There are two other key differences: we include an alternative use for land at a location and we allow for heterogeneity in output across locations.

Our work is related to [**TBD** .... old literature, ciccone-hall, ciccone-peri and related cites, island models of cities/other island models, production function, sources of growth - neutral, investment-specific, agglomeration...., Lucas - only externality can imagine is the one that creates cities].

The rest of the paper is organized as follows. The next section describes our model economy, including the equations underlying our estimation. After that we describe our econometric strategy and then the data. Section five discusses our empirical results and quantifies the impact of agglomeration on per capita consumption growth. Section six concludes.

2 Model Economy

We embed a version of Roback (1982)'s model of cities, modified to include agglomeration effects in the way proposed by Ciccone and Hall (1996), into the neoclassical growth model. We use our model to derive the equations underlying our estimates of agglomeration's effect on local productivity and to show how to quantify the impact of local agglomeration on aggregate growth. We also characterize the cross-sectional pattern of output, employment and prices in the model.

2.1 Economic Environment

We consider a discrete time infinite horizon model with complete markets and no aggregate uncertainty. There are M locations called cities where distinct intermediate goods are produced by competitive firms. Competitive firms whose location is unspecified combine the intermediate goods into a final good. We normalize each city's endowment of land to unity.[1]

There is a representative household consisting of a unit measure of members. The large household structure allows for full risk sharing within each household, a standard device for studying complete markets allocations. Each household member supplies one unit of labor inelastically and enjoys utility from consumption, c and housing, h. For simplicity housing is just land; there are no structures in the model. The household derives income from labor and capital by allocating its members and its stock of capital, K across the cities, and by collecting rent on its land. Members are costlessly and randomly allocated across cities each period after the household observes the current level of total factor productivity in each city. Each member

[1] This is a normalization since the problem we study can always be rewritten in terms of shares of total land used in residential and non-residential activities as long as the local production function for intermediate goods is rescaled appropriately.

must consume and enjoy housing in the same city that they work. Capital is perfectly mobile and along with land is also allocated after the household observes productivity.

The representative intermediate goods producer in each city uses a Cobb-Douglas technology with inputs of land, l labor, n and capital, k. Total factor productivity (TFP) in each city, s is taken by households and firms to be given. It is specified as

$$s = z^{(1-\alpha)\phi} \left[\frac{y}{l}\right]^{\frac{\lambda-1}{\lambda}}$$

where z is an exogenous city-specific stationary stochastic term and y is total city output.[2] The ratio of output to land is called the *density of economic activity*. If $\lambda = 1$ density has no impact on productivity and if $\lambda > 1$ firms' productivity is increasing in density. Ciccone and Hall (1996) show how this production technology can be derived as the reduced form of a micro founded model. Final goods producers combine the output of each city into consumption and investment goods using a constant elasticity of substitution production function.

Land and labor are traded in local markets while the intermediate and final goods as well as capital services are traded in economy-wide markets. For simplicity, we initially assume the aggregate level of technology and the number of households are constant and focus on a stationary equilibrium.

2.2 Stationary Competitive Equilibrium

The stationary competitive equilibrium for the case without density effects, $\lambda = 1$, consists of the following. In each period, the household chooses how much to consume of the final good and housing, and how to allocate its members and stock of capital to each city, in order to maximize the equally weighted utility of all the household's members subject to the technological constraints on allocating members across cities, taking as given the stochastic process for TFP in each city and all prices. Each intermediate and final goods producer maximizes profits by choosing inputs taking their prices as given. The aggregate intermediate and final good and capital service markets and the local labor and land markets all clear.

[2] We could incorporate non-stationary technology along the lines pursued by Alvarez and Shimer (2008) In their model idiosyncratic technology follows a random walk. They achieve a stationary aggregate distribution by assuming that locations die with constant probability replaced instantaneously by the same number of new locations drawn from the stationary distribution of technology.

In cases with a positive effect of density on productivity, $\lambda > 1$, an additional *productivity consistency* condition must be satisfied. Recall that we assume that the distribution of total factor productivity is taken as given by households and firms. When $\lambda > 1$ the decisions of households and firms affect the distribution of productivity, through the resulting distribution of density. It follows that in a competitive equilibrium, the solutions of the household and firm problems must generate the underlying distribution of productivity that was taken as given when calculating those solutions.

2.3 Planning Problem

We initially focus on the case without any impact of density on productivity, $\lambda = 1$. The competitive equilibrium for this case solves a planning problem and the two welfare theorems apply. Without aggregate uncertainty we seek stationary decision rules and so we drop the time subscript where appropriate. Instead of indexing cities according to their current productivity level, it is convenient to index them according to their productivity history, s^t. Without loss of generality suppose s^t can take on a finite set of values. Let $\pi(s^t)$ be the distribution of cities across productivity histories.

Since the household perfectly insures itself against consumption risk, we write the planning problem as

$$\max \sum_{t=0}^{\infty} \beta^t \sum \pi(s^t) n(s^t) \left[\ln c_t + \psi \ln h(s^t) \right]$$

subject to

$$\begin{aligned}
\sum \pi(s^t) n(s^t) c_t + K_{t+1} - (1-\kappa) K_t &\leq \left[\sum \left(\pi(s^t) y(s^t) \right)^{\eta} \right]^{\frac{1}{\eta}} && (1) \\
y(s^t) &\leq s_t l(s^t)^{1-\phi} k(s^t)^{\alpha\phi} n(s^t)^{(1-\alpha)\phi} \\
\sum \pi(s^t) k(s^t) &\leq K_t \\
n(s^t) h(s^t) + l(s^t) &\leq 1, \ \forall s^t && (2) \\
\sum \pi(s^t) n(s^t) &\leq 1 \\
& K_0 \text{ given}
\end{aligned}$$

where the unindexed summations are over histories s^t. The right-hand side of (1) is the function mapping intermediate into final goods. In this mapping $\eta \leq 1$ and in addition we assume $\eta \neq 0$. Intermediate goods are perfect substitutes if $\eta = 1$ and as $\eta \to -\infty$ the final good is a Leontief function of intermediate goods. The variables c_t and $h(s^t)$ are in per capita terms. Finaly $\psi > 0$ and $0 \leq \kappa \leq 1$ is capital's rate of depreciation.

Let the Lagrange multipliers for the above restrictions be $\beta^t \nu_t$, $\beta^t \nu_t \pi(s^t) q(s^t)$, $\beta^t \nu_t r_t$, $\beta^t \nu_t \pi(s^t) p(s^t)$ and $\beta^t \nu_t \theta_t$. Then the first order conditions for c_t, $y(s^t)$, K_{t+1}, $k(s^t)$, $l(s^t)$, $h(s^t)$ and $n(s^t)$ are

$$\frac{1}{c_t} = \nu_t \tag{3}$$

$$q(s^t) = Y_t^{1-\eta} \left(\pi(s^t) y(s^t)\right)^{\eta-1} \tag{4}$$

$$\nu_t = \beta \nu_{t+1} \left[r_{t+1} + 1 - \kappa\right] \tag{5}$$

$$r_t = \alpha \phi q(s^t) s_t l(s^t)^{1-\phi} k(s^t)^{\alpha\phi-1} n(s^t)^{(1-\alpha)\phi} \tag{6}$$

$$p(s^t) = (1-\phi)\, q(s^t) s_t l(s^t)^{-\phi} k(s^t)^{\alpha\phi} n(s^t)^{(1-\alpha)\phi} \tag{7}$$

$$\frac{\psi}{h(s^t)} = \nu_t p(s^t) \tag{8}$$

$$\ln c_t + \psi \ln h(s^t) = \nu_t \theta_t + \nu_t \left[c_t + p(s^t) h(s^t) - w(s^t)\right] \tag{9}$$

where

$$w(s^t) = (1-\alpha)\phi q(s^t) s_t l(s^t)^{1-\phi} k(s^t)^{\alpha\phi} n(s^t)^{(1-\alpha)\phi-1} \tag{10}$$

In the competitive equilibrium we use the final good as the numeraire. Then, $w(s^t)$, $p(s^t)$ and $q(s^t)$ are wages, land rents and intermediate good prices at cities with history s^t, and r_t is the economy's rental rate on capital.

Most of the first order conditions are familiar and need no discussion. There are two key differences with the traditional growth model. First, wages and land prices are idiosyncratic to location. That wages depend on the local level of productivity follows directly from the inclusion of housing in preferences. Without a preference for housing, wages would be equalized across cities. This can be seen by inspecting (9). The only way for this condition to hold without housing is for wages to be equalized across cities.

Second, the *location indifference* condition (9) does not appear in the growth model. This condition contrasts with the static Roback (1982) model where utility is equated across cities. In our dynamic model with complete markets, utility net of the consumption price of the transfer is equated. Since transfers sum to zero, *ex ante expected* utility is equalized.

Without density effects the competitive equilibrium exists and is unique, by standard arguments. With density effects, the competitive equilibrium is found by finding the planning problem's solution that satisfies the productivity consistency condition. We have not found conditions under which a solution to this fixed point problem exists or is unique. However, using the argument in Kehoe, Levine and Romer (1992) we know that for λ sufficiently close to unity, there exists a unique competitive equilibrium, because the model without density effects has a unique equilibrium. Equilibrium existence and uniqueness under the size of density effects we estimate remain open questions since we do not need to find a solution of the model to identify the impact of density on local productivity or quantify the density effect on aggregate growth, and so do not do so in this paper.

2.4 Equation Underlying Our Estimation

We now derive a key equation underlying our estimation. Consider any two cities with different productivity histories, indexed i and j, in some arbitrary period. Output in a city is

$$y = z^{(1-\alpha)\phi} \left[\frac{y}{l}\right]^{\frac{\lambda-1}{\lambda}} l^{1-\phi} k^{\alpha\phi} n^{(1-\alpha)\phi.} \tag{11}$$

Let $\delta = \lambda\phi$. This parameter measures the net effect on productivity of diminishing returns to land, otherwise known as congestion, and density. Solving for output and dividing output in city i by that in city j yields

$$\frac{y_i}{y_j} = \left[\frac{z_i}{z_j}\right]^{\delta(1-\alpha)} \left[\frac{l_i}{l_j}\right]^{1-\delta} \left[\frac{k_i}{k_j}\right]^{\delta\alpha} \left[\frac{n_i}{n_j}\right]^{\delta(1-\alpha)}.$$

Using (6) to eliminate the capital terms in this last equation we have

$$\frac{y_i/l_i}{y_j/l_j} = \left[\frac{z_i}{z_j}\right]^{\frac{\delta(1-\alpha)}{1-\delta\alpha}} \left[\frac{n_i/l_i}{n_j/l_j}\right]^{\frac{\delta(1-\alpha)}{1-\delta\alpha}} \left[\frac{q_i}{q_j}\right]^{\frac{\delta\alpha}{1-\delta\alpha}} \tag{12}$$

Equation (12) is closely related to equation 19 in Ciccone and Hall (1996), which is the basis for their estimation equation. It differs in two key respects. First, because Ciccone and Hall (1996) assume intermediate goods are perfect substitutes in producing final goods, $\eta = 1$, their equation does not contain any output prices. In this sense their equation may be subject to an omitted variable bias. Second, they assume all land in a given location is used in production while we have competing uses for land. Since the allocation of land between residential and non-residential uses in U.S. cities is unavailable to us, we cannot use (12) as a basis for estimation. Further complicating our analysis, annual data on real city output that is produced by the BEA is available for too short a time period (2001-2006) to be useful.

We eliminate land and output from (12) using (7) and (10) to arrive at

$$\frac{w_i}{w_j} = \frac{z_i}{z_j} \left[\frac{p_i}{p_j}\right]^{\frac{\delta-1}{\delta(1-\alpha)}} \left[\frac{q_i}{q_j}\right]^{\frac{1}{\delta(1-\alpha)}}. \tag{13}$$

This equation underlies our estimation procedure described below. It is important to note that (13) follows directly from profit maximization by firms and the functional form of the production function. In particular, it does not depend on the assumption of perfect risk sharing or any assumptions about preferences. If we were to add other local factor inputs to the model with competing uses, such as energy, their prices would also appear in (13) in the same way that land prices do.

2.5 Characterization of the Cross-Section of Cities

It is useful to characterize the cross-sectional relationship among the endogenous variables in the model. Our characterization strategy is to consider small perturbations to the level of technology of a given city, say city i, holding fixed all aggregate variables and local variables in other cities.

Combining (3), (8) and (9) and differentiating it is straightforward to show that

$$\frac{d\ln p_i}{d\ln z_i} = \frac{w_i n_i}{p_i n_i h_i} \frac{d\ln w_i}{d\ln z_i}. \tag{14}$$

It follows immediately that the model predicts land prices and wages are positively correlated in a cross-section of cities.

The cross-sectional relationship between employment and wages is derived as follows. From (2), (3) and (8)

$$\frac{d\ln l_i}{d\ln z_i} = \frac{p_i n_i h_i}{p_i l_i} \left[\frac{d\ln p_i}{d\ln z_i} - \frac{d\ln n_i}{d\ln z_i}\right]. \tag{15}$$

Taking logs of (7) and (10) and subtracting one resulting expression from the other leaves an equation in employment, wages, land and its price. Substituting for land in this last equation using (15) and the price of land using (14) and solving for the employment elasticity yields

$$\frac{d\ln n_i}{d\ln z_i} = \frac{\frac{p_i l_i}{p_i n_i h_i} w_i n_i + w_i n_i - p_i l_i}{p_i n_i h_i + p_i l_i} \frac{d\ln w_i}{d\ln z_i} \tag{16}$$

Any empirically plausible calibration of the model will have labor's share exceeding land's share in production, *i.e.* $w_i n_i > p_i l_i$. From now on we assume that this condition is satisfied. It follows that employment is positively correlated with wages.

The relationships of output and its price with wages depend on the degree of substitutability of intermediate goods in the production of final goods. To see this, eliminate the local output price from (10) using (4),

$$\frac{d\ln y_i}{d\ln z_i} = \frac{1}{\eta}\left[\frac{d\ln n_i}{d\ln z_i} + \frac{d\ln w_i}{d\ln z_i}\right].$$

Use this last expression to substitute for output in (4),

$$\frac{d\ln q_i}{d\ln z_i} = \frac{\eta-1}{\eta}\left[\frac{d\ln n_i}{d\ln z_i} + \frac{d\ln w_i}{d\ln z_i}\right]$$

Given our previous result for the employment elasticity, it is clear that if $\eta > 0$ then output is positively correlated with wages and the local output price is negatively correlated with wages. The opposite relationships hold if $\eta < 0$.

Land use in production is positively related to wages. This follows from substituting for prices and employment in (15) using (14) and (16):

$$\frac{d\ln l_i}{d\ln z_i} = \frac{n_i h_i}{n_i h_i + l_i}\frac{d\ln w_i}{d\ln z_i}.$$

This last equation is interesting in light of our result for output. The two results imply that the density of employment activity can be positively or negatively related to wages in the cross section, depending on the degree of substitutability of intermediate goods in the production of final goods, η. In particular, even without any effects of density on total factor productivity, we could observe a positive relationship between wages, or labor productivity since this is proportional to wages in our model, and density in a cross-section of cities.

2.6 Amenities

Cities clearly vary by amenities in addition to productivity. Therefore it is natural to ask whether our results in the previous two subsections are affected by including amenities in the model. The short answer to this question is that if amenities appear in individuals' preferences in a logarithmically separable way, then none of our results thus far are affected. We now elaborate briefly on this answer.

Suppose the new planning objective is given by

$$\sum_{t=0}^{\infty} \beta^t \sum \pi(s^t) n(s^t) \left[\ln c_t + \psi \ln h(s^t) + \varsigma \ln a(s^t) \right],$$

where we redefine $\pi(s^t)$ to be the distribution of cities across productivity *and* amenity histories, $a(s^t)$ denotes the quantity of amenities at a city with productivity-amenity history s^t, and $\varsigma > 0$. The only first order condition affected from the problem without amenities is the location indifference condition. This becomes

$$\ln c_t + \psi \ln h(s^t) + \varsigma \ln a(s^t) = \pi_t \theta_t + \pi_t \left[(1+\psi)c - w(s^t) \right].$$

Amenities matter because they effect the planner's allocation of labor, which then effects relative prices across cities. We pointed out earlier that introducing housing was important for generating a non-trivial wage distribution in the model. This is only true if we exclude amenities. Inspection of the new indifference condition shows that wages can differ by city if amenities do even without housing.

Since the other first order conditions are unaffected by introducing amenities, the equation underlying our estimation, (13), is also unchanged. It is also the case that the relationship between the endogenous variables with idiosyncratic variation in technology remains qualitatively unchanged with idiosyncratic variation in amenities. This can be verified by following the same proof strategy as before. Assuming that amenities are orthogonal to productivity, it follows that the general model including both sources of cross-sectional variability will share the same relationship between endogenous variables as with cross-sectional variation in technology only.

2.7 Heterogeneous Workers

So far we have assumed that workers are homogeneous. This is a questionable assumption from an empirical perspective. Locations with a high density of economic activity may have high wages because of a concentration of high human capital workers, not because of a density effect on productivity. Without accounting for cross-sectional variation in the distribution of human capital we could overstate density effects. Ciccone and Hall (1996) address this concern by accounting for the average level of educational attainment in a location. [**TBD** xxx rewrite this next sentence]. Ciccone and Peri (2006) have raised doubts about this approach. Their argument is focused on the issue of identifying human capital externalities, but is relevant in our context as well. They argue that the kind of approach adopted by Ciccone and Hall (1996) confounds positive externalities with wage changes due to a downward sloping demand curve for human capital. Consistent with Ciccone and Peri (2006), we address the issue by allowing high skill and low skill workers to be imperfect substitutes in producing labor services. We now derive a version of the equation underlying our estimation which incorporates heterogeneous workers.

Suppose that effective labor input in (11) is a constant elasticity of substitution aggregate of unskilled and skilled labor:

$$n = \left[\sigma u^{\xi} + (1-\sigma)e^{\xi}\right]^{1/\xi}.$$

We can exploit previous derivations from above by using the notion of the demand for composite labor n. Composite labor satisfies (10) as before. Efficient use of unskilled and skilled labor implies

$$\begin{aligned} w_{ui} &= \sigma w_i {n_i}^{1-\xi} {u_i}^{\xi-1} \\ w_{ei} &= (1-\sigma) w_i {n_i}^{1-\xi} {e_i}^{\xi-1} \end{aligned}$$

where w_{ei} and w_{ui} are wages paid to skilled and unskilled workers respectively in city i. Notice that

$$\frac{w_{ui}u_i + w_{ei}e_i}{w_{ui}u_i} = 1 + \frac{1-\sigma}{\sigma} m_i^{\xi} \equiv \chi_i,$$

so that

$$w_{ei} = (1-\sigma)\sigma^{1/\xi-1}(1-\alpha)\phi w_i \chi_i^{1/\xi-1} m_i^{\xi-1}$$

where w_i is interpreted as the implicit wage for the composite labor input, n_i, and $m_i = e_i/u_i$. Substituting for composite wages using (13), we have the version of our key estimation equation incorporating skilled and unskilled labor:

$$\frac{w_{ei}}{w_{ej}} = \frac{z_i}{z_j}\left[\frac{p_i}{p_j}\right]^{\frac{\delta-1}{\delta(1-\alpha)}}\left[\frac{q_i}{q_j}\right]^{\frac{1}{\delta(1-\alpha)}}\left[\frac{\chi_i}{\chi_j}\right]^{1/\xi-1}\left[\frac{m_i}{m_j}\right]^{\xi-1}. \tag{17}$$

Equation (17) reduces to equation (13) if $\xi = 1$, that is if unskilled and skilled labor are perfect substitutes.

2.8 Balanced Growth

Our choices of preferences and technology guarantee that the model admits balanced growth. We now derive the balanced growth path and show how to use this to quantify the impact of local agglomeration effects on aggregate growth. Assume the neutral technology and number of households grow at gross rates $\gamma \geq 1$ and $\mu \geq 1$ respectively, and that the model's stationary representation has a unique solution. Assume the neutral technology evolves as

$$z_t = \gamma^t \tilde{z}_t$$

where $\tilde{z}_t$ is a stationary random variable.

The land constraints (2) implies that per capita residential land, $h(s^t)$ must shrink at the rate of population growth and land used for production, $l(s^t)$, cannot grow. The aggregate resource constraint (1) says that per capita consumption and capital grow at the same rate. The labor demand condition says that wages grow as the same rate as per capita output, and therefore per capita consumption and per capita capital. Denote this common rate γ_c.

Output in any given city satisfies (11). Dividing this equation through by n yields an expression involving per capita variables:

$$y/n = z^{(1-\alpha)\phi}\left[\frac{y/n}{l/n}\right]^{\frac{\lambda-1}{\lambda}}[l/n]^{1-\phi}[k/n]^{\alpha\phi}$$

Collecting terms in this expression and using the balanced growth requirement that all cities must grow at the same rate we arrive at

$$\gamma_c = \gamma^{\frac{(1-\alpha)\delta}{1-\alpha\delta}}\mu^{\frac{\delta-1}{1-\alpha\delta}}. \tag{18}$$

Equation (18) shows that per capita consumption growth depends on the rate of technology growth and, if $\delta > 1$, the rate of population growth as well. This dependence on population growth derives from the fact that the the density of economic activity must grow along a balanced growth path if land used in production does not grow. This provides an additional impetus to total factor productivity growth. A key implication of (18) is that it is the net effect of diminishing returns to land and the density of economic activity which determines the ultimate effect of agglomeration on per capita consumption growth. Finally, notice that without any productivity effect of density, $\lambda = 1$ (18) implies

$$\gamma_c = \gamma^{\frac{(1-\alpha)\phi}{1-\alpha\phi}} \mu^{\frac{\phi-1}{1-\alpha\phi}}.$$

In this case per capita consumption growth is reduced because of decreasing returns to land.

3 Econometric Strategy

This section describes how we estimate the parameters of our model, in particular the agglomeration parameter, δ. To begin, take logs of equation (17) and rearrange terms:

$$\begin{aligned} &\ln(w_{ei}) - \ln(w_{ej}) \\ &= \tfrac{\delta-1}{\delta(1-\alpha)}\left[\ln(p_i) - \ln(p_j)\right] \;+\; \tfrac{1}{\delta(1-\alpha)}\left[\ln(q_i) - \ln(q_j)\right] \\ &\quad + \tfrac{1-\xi}{\xi}\left[\ln(\chi_i) - \ln(\chi_j)\right] \;+\; (\xi-1)\left[\ln(m_i) - \ln(m_j)\right] \\ &\quad + \ln(z_i) - \ln(z_j) \end{aligned}$$

At all points in time, this equation holds for any two MSAs i and j. Therefore, it must also hold for MSA i and the average of the $j = 1, \ldots, N$ MSAs. Denote a hatted variable as the difference of a logged variable of MSA i and the average of the same logged variable for all j MSAs in a sample, i.e.

$$\hat{w}_{ei} \;\equiv\; \ln(w_{ei}) \;-\; \frac{1}{N}\sum_{j=1}^{N} \ln(w_{ej}) \;.$$

It follows that

$$\hat{w}_{ei,t} = \tfrac{\delta-1}{\delta(1-\alpha)}\hat{p}_{i,t} \;+\; \tfrac{1}{\delta(1-\alpha)}\hat{q}_{i,t} \;+\; \tfrac{1-\xi}{\xi}\hat{\chi}_{i,t} \;+\; (\xi-1)\,\hat{m}_{i,t} \;+\; \hat{z}_{i,t} \;. \tag{19}$$

Equation (19) holds at all dates. The time subscript t has been included in all variables.

We do not observe $\hat{z}_{i,t}$. The model tells us that $z_{i,t}$ and the other right hand side variables are correlated, implying we have to estimate the structural parameters of the model, δ, α, and ξ using an instrumental variables approach. If we were to only have one wave of data available in estimation, we would argue for an instrument that is correlated with the observable variables and uncorrelated with the $\ln(\hat{z}_{i,t})$ term. This is the econometric approach adopted by Ciccone and Hall (1996), for example. Since we have access to panel data, we adopt a different estimation strategy which exploits the panel nature of our data.

We assume that for every MSA i, $\ln(z_{i,t})$ has the following stochastic process:

$$\begin{aligned} \ln(z_{i,t}) &= \ln(z_{i,0}) + gt + \ln(\tilde{z}_{i,t}) \\ \ln(\tilde{z}_{i,t}) &= \rho \ln(\tilde{z}_{i,t-1}) + v_{i,t} \ . \end{aligned}$$

g is the deterministic rate of growth and ρ is the AR(1) coefficient on the cyclical component of $\ln z$. Both g and ρ are assumed common to all MSAs. $z_{i,0}$ is an initial level of multi-factor productivity specific to MSA i and $v_{i,t}$ is an i.i.d. shock (again, specific to i) to the level of multi-factory productivity that is assumed orthogonal to all variables dated $t-1$ and earlier.

We now use our assumption that $\ln(\tilde{z}_{i,t})$ follows an AR(1) process to rewrite equation (19) into an equation with an unobserved variable that is uncorrelated with all variables from previous years. This will form the basis of our instrumental variables estimation. Given the "hat" variable definition (log deviation from average) and the assumption that ρ and g are common to all MSAs, note that

$$\hat{z}_{i,t} - \rho \hat{z}_{i,t-1} = (1-\rho)\,\hat{z}_{i,0} + [v_{i,t} - \bar{v}_t] \ , \tag{20}$$

where $\bar{v}_t$ is the average value of $v_{j,t}$ for all j MSAs in the sample in period t. For simplicity going forward, we will redefine $\epsilon_{i,t} = v_{i,t} - \bar{v}_t$. Now from each variable in equation (19) subtract ρ times its once-lagged value. This is a valid operation because equation (19) holds at all dates. Then, using the results from (20) we derive

$$\begin{aligned} &\hat{w}_{ei,t} \\ &= \rho\, \hat{w}_{ei,t-1} + \tfrac{\delta-1}{\delta(1-\alpha)} [\hat{p}_{i,t} - \rho\, \hat{p}_{i,t-1}] + \tfrac{1}{\delta(1-\alpha)} [\hat{q}_{i,t} - \rho\, \hat{q}_{i,t-1}] \\ &+ \tfrac{1-\xi}{\xi} [\hat{\chi}_{i,t} - \rho\, \hat{\chi}_{i,t-1}] + (\xi - 1)[\hat{m}_{i,t} - \rho\, \hat{m}_{i,t}] + \hat{z}_{i,0} + \epsilon_{i,t} \ . \end{aligned} \tag{21}$$

To make further progress on estimating ρ, δ, α, and ξ, we need to identify a set of instruments that are correlated with the wage, price, and skill variables and are uncorrelated with $\epsilon_{i,t}$. In addition, we need to eliminate the unobserved variable $\hat{z}_{i,0}$. This variable is a "fixed effect" in the sense that it varies across MSAs, but is fixed over time in each MSA.

Finding valid instruments for equation (21) is straightforward because $\epsilon_{i,t}$ is an i.i.d. shock; any variable dated $t-1$ or earlier is potentially a valid instrument. However, because we use all same-MSA variables as instruments $\{\hat{w}_{ei}, \hat{p}_i, \hat{q}_i, \hat{\chi}_i$, and $\frac{\hat{e_i}}{u_i}\}$, we follow standard practice and omit the $t-1$ observation of these variables as instruments and only use these variables dated $t-2$ and earlier for the instruments. This allows that the variables that appear on the right-hand side of equation (21), some dated at $t-1$, are potentially measured with classical measurement error.

Several strategies have been proposed to handle the MSA-level fixed effect. The most common is the first difference method. Letting Δ denote the first-difference operator, equation (21) can be written

$$\begin{aligned} &\Delta\hat{w}_{ei,t} \\ &= \rho\,\Delta\hat{w}_{ei,t-1} + \tfrac{\delta-1}{\delta(1-\alpha)}\left[\Delta\hat{p}_{i,t} - \rho\,\Delta\hat{p}_{i,t-1}\right] + \tfrac{1}{\delta(1-\alpha)}\left[\Delta\hat{q}_{i,t} - \rho\,\Delta\hat{q}_{i,t-1}\right] \\ &+ \tfrac{1-\xi}{\xi}\left[\Delta\hat{\chi}_{i,t} - \rho\,\Delta\hat{\chi}_{i,t-1}\right] + (\xi-1)\left[\Delta\hat{m}_{i,t} - \rho\,\Delta\hat{m}_{i,t}\right] + \Delta\epsilon_{i,t}\,. \end{aligned} \tag{22}$$

In practice, researchers use the level of all same-MSA model variables, $\hat{w}_{ei,t-s}$, $\hat{p}_{i,t-s}$, $\hat{q}_{i,t-s}$, $\hat{\chi}_{i,t-s}$, and $\widehat{\frac{e_{i,t-s}}{u_{i,t-s}}}$, for $s \geq 3$ as instruments when estimating model parameters based on this equation. Thus, four waves of data are required for consistent estimation.

A second strategy that has been proposed is to simply assume that the level of the fixed effect, $\hat{z}_{i,0}$, is uncorrelated with the first-differences of all model variables. In this application, equation (21) is used directly in estimation, but the instruments used are $\Delta\hat{w}_{ei}$, $\Delta\hat{p}_i$, $\Delta\hat{q}_i$, $\Delta\hat{\chi}_i$, and $\Delta\frac{\hat{e_i}}{u_i}$, all dated $t-2$ and earlier. As before, four waves of data are required. A widely used procedure, first advocated by Blundell and Bond (2000), is to combine these two strategies to estimate model parameters. We are uncomfortable assuming that the fixed effect is uncorrelated with lagged changes of model variables, since we have no theory suggesting this to be the case, and so do not use the Blundell and Bond (2000) estimation procedure. Instead we focus on procedures that directly remove the MSA-level fixed effects. When five or more waves of data are available, first-differencing as in equation (22) to eliminate the MSA-specific intercept does not use all the available information. For this reason we use the strategy proposed by Arellano and Bover (1995) which takes advantage of the additional information on the fixed effects which comes with long panel data.

In this approach, each time-t variable in equation (21) is expressed as a deviation from the average of all future observations for MSA i in the sample. For any variable x_t with observations $t = 1, \ldots, T$, define the Arellano-Bover difference operator as of date t, Δ_t^*, as follows:

$$\Delta_t^* x_t \;=\; x_t \;-\; \frac{1}{T-t} \sum_{s>t} x_s \; .$$

Since equation (21) holds at all dates t, the Arellano-Bover difference operator can in principal be directly applied to all terms. However, the variance of the unobserved ϵ term once this difference is applied

$$\Delta_t^* \epsilon_t \;=\; \epsilon_t \;-\; \frac{1}{T-t} \sum_{s>t} \epsilon_s$$

changes at each date t. In order to keep the variance of this error term constant, each variable must be multiplied by a weight equal to

$$\omega_t \;=\; \left(\frac{T-t}{T-t+1} \right)^{1/2} \; .$$

When applied to our problem, the Arellano and Bover (1995) method to difference out the fixed effect has the form

$$\begin{aligned} \omega_t \Delta_t^* \hat{w}_{ei,t} = \; & \rho \, \omega_t \Delta_t^* \hat{w}_{ei,t-1} + \tfrac{\delta-1}{\delta(1-\alpha)} \left[\omega_t \Delta_t^* \hat{p}_{i,t} \;-\; \rho \, \omega_t \Delta_t^* \hat{p}_{i,t-1} \right] \\ & + \tfrac{1}{\delta(1-\alpha)} \left[\omega_t \Delta_t^* \hat{q}_{i,t} \;-\; \rho \, \omega_t \Delta_t^* \hat{q}_{i,t-1} \right] + \tfrac{1-\xi}{\xi} \left[\omega_t \Delta_t^* \hat{\chi}_{i,t} \;-\; \rho \, \omega_t \Delta_t^* \hat{\chi}_{i,t-1} \right] \\ & + \; (\xi - 1) \left[\omega_t \Delta_t^* \tfrac{\widehat{e_{i,t}}}{u_{i,t}} \;-\; \rho \, \omega_t \Delta_t^* \tfrac{\widehat{e_{i,t-1}}}{u_{i,t-1}} \right] + \; \omega_t \Delta_t^* \epsilon_{i,t} \; . \end{aligned} \tag{23}$$

Note that all variables, including the time $t-1$ variables, have the same Arellano-Bover difference operator Δ_t^* and weight ω_t. We use the levels of $\hat{w}_{ei,t-s}$, $\hat{p}_{i,t-s}$, $\hat{q}_{i,t-s}$, $\hat{\chi}_{i,t-s}$, and $\frac{\widehat{e_{i,t-s}}}{u_{i,t-s}}$ for $s = \{3, 4\}$ as instruments in estimating equation (23). We do not use $s \geq 5$ for instruments because typically they do not add much information to help identify parameters.

4 Data

We estimate the parameters of the model using three data sources. The exact procedures we use to construct and merge the data are detailed in the data appendix. A brief summary follows here.

Our first data set is a balanced annual panel covering 42 MSAs over the 1985-2004 period. We construct our wage and employment variables – $\hat{w}_{ei,t}$, $\hat{\chi}_{i,t}$, and $\widehat{\frac{e_{i,t}}{u_{i,t}}}$ – from the March Current Population Survey (CPS). All workers with at least four years of college are considered as high-skill workers and all other workers are labeled as low skill. We use data from the Bureau of Economic Analysis (BEA) to construct output prices $\hat{q}_{i,t}$ by MSA. Specifically, we create a price index for output that varies by MSA by merging annual data on income earned by industry by MSA from the Regional Economic Accounts with price indexes for industry output from the Annual Industry Accounts. Because the mix and intensity of industries varies by MSA, and price indexes vary by industry, the price index for output produced in an MSA will vary across MSAs. Finally, we use data on the price of land in residential use as estimated by Davis and Palumbo (2008) to construct $\hat{p}_{i,t}$. Davis and Palumbo (2008) construct time-series estimates of the price of land for 46 MSAs. Four of the MSAs in their study had to be discarded because of lack of corresponding CPS and BEA data, explaining the dimensions of the panel data in this first data set.

The data on output prices is available from 1969-2006. We normalize the price index for MSA-level output to equal 1.0 in every MSA in 1969. This arbitrary normalization introduces another MSA-level fixed effect that is differenced-out by the Arellano and Bover (1995) procedure. The first year of data for both the CPS and the Davis and Palumbo (2008) data is 1985.[3] The CPS data are available through 2006, but the last year of Davis and Palumbo (2008) data is 2004, explaining the range of the sample. The CPS and BEA data are annual. The Davis and Palumbo (2008) data are quarterly; we set the annual estimate equal to the average of the reported quarterly values.

Our second data set is a balanced annual panel covering 149 MSAs over the 1985-2006 period. As with the first data set, we use annual CPS data to construct $\hat{w}_{ei,t}$, $\hat{\chi}_{i,t}$,

[3]The CPS identifies only 15 metropolitan areas in prior years.

and $\frac{\hat{e_{i,t}}}{u_{i,t}}$ variables and data from the BEA to construct output prices $\hat{q}_{i,t}$. We combine micro-level data from the 1990 Decennial Census of Housing (DCH) with price indexes for existing homes from the Office of Federal Housing Enterprise Oversight (OFHEO) to compute estimates of the price of housing, by MSA, over the 1985-2006 period. We use the data from the 1990 DCH to estimate the level of house prices by MSA in 1990, and then use the OFHEO price indexes to extrapolate the price level backwards to 1985 and forwards to 2006. After dropping MSAs with incomplete or missing wage, employment, output price or house price data, we are left with 149 MSAs, including all of the 42 MSAs covered in the first data set.

Our third and final data set is balanced annual panel covering 83 MSAs over 4 years. As before, the data on output prices is from the BEA. All other variables on wages, employment, and house prices are taken from the 1970, 1980, 1990, and 2000 DCH.

Why three data sets? The first data set is exactly consistent with the model. We use the second data set (house prices rather than land prices) for two reasons: By using house-price data, we extend the sample by two years and more than triple the sample of MSAs from 42 to 149. Related, because we have many more MSAs in the sample, this data set includes the experiences of MSAs with relatively small populations that are largely excluded from the Davis and Palumbo (2008) data. We use the third data set (the DCH data) because it includes the experiences of MSAs prior to 1985. Also, since the DCH data are used to construct the wage and employment variables, sample sizes are much larger than in the first two data sets that use the CPS data to construct these data.

Table 1 summarizes the key variables in each of our three data sets. Given the definition of the hatted variables as deviations from the average, the sample average of all variables is zero in every year. Therefore, in table 1 we report standard deviations. Since the variables are in logs, the standard deviations we report have the interpretation of cross-sectional percent standard deviations of the levels. In every data set, output prices are least dispersed, land/house prices are most dispersed, and high-skill wage rates are less dispersed than the inverse of the share of labor income accruing to unskilled workers which are less dispersed then high-low skill ratios. Table 2 shows more details, specifically the standard deviations of all variables by year. This table enables a full comparison of the data over time (within data sets) and across data sets at a point in time. In every data set, wages paid to high-skill workers have become more dispersed over time and the price of land/housing has also become more dispersed over time.[4] Wages are more dispersed and house prices are less dispersed in our second data set than in our first data set.

5 Results and Analysis

Table 3 reports our parameter estimates and standard errors for equation (23) from our first two data sets. Standard errors are computed using the "delta" method. Column (1) reports the estimates for the first data set – the data set with land prices – when all parameters are unrestricted in the estimation procedure. When our estimation algorithm runs unrestricted, we estimate that capital's share of output is negative, which is infeasible. In column (2), we report estimates for the parameters of the model after forcing α to exactly equal 0.30. Columns (3) and (4) are identical to columns (1) and (2), except that in these columns we report the estimates for the second data set (house prices). Columns (5) and (6) report estimates for the second data set, and for α set to 0.30, but with the sample restricted: In column (5), we only consider the 42 MSAs used in the first data set, but over the entire 1985-2006 sample period. We use the same sample in column (6) as column (5), but with the time period restricted to be 1985-2004, the identical time period as the first data set.

6 Conclusions

TBD

References

Alvarez, F. and R. Shimer (2008). Search and rest unemployment. Unpublished University of Chicago manuscript.

[4] Output prices have become more dispersed over time, but this may be due to our normalization of the price of output to equal 1.0 in every MSA in 1969.

Arellano, M. and O. Bover (1995). Another look at the instrumental variables estimation of error-components models. *Journal of Econometrics 68*(1), 29–51.

Blundell, R. and S. Bond (2000). Gmm estimation with persistent panel data: An application to production functions. *Econometric Reviews 19*(3), 321–340.

Ciccone, A. and R. E. Hall (1996). Productivity and the density of economic activity. *American Economic Review 86*(1), 54–70.

Ciccone, A. and G. Peri (2006). Identifying human-capital externalities: Theory with applications. *Review of Economic Studies 73*, 381–412.

Davis, M. and J. Heathcote (2007). The price and quantity of residential land in the united states. *Journal of Monetary Economics 54*, 2595–2620.

Davis, M. and M. Palumbo (2008). The price of residential land in large u.s. cities. *Journal of Urban Economics 63*, 352–384.

King, M., S. Ruggles, T. Alexander, D. Leicach, and M. Sobek (2004). Integrated public use microdata series, current population survey: Version 2.0. [machine-readable database]. Technical report, Minneapolis, MN: Minnesota Population Center [producer and distributor].

Roback, J. (1982). Wages, rents, and the quality of life. *Journal of Political Economy 90*(6), 1257–1278.

A Data Appendix

In this appendix, we document how we construct key variables from our various data sources (section A.1) and then document how we merge our different data sources together to create our three data sets for estimation (section A.2).

A.1 Data Sources

A.1.1 CPS Data (Wages and Hours Worked by Skill)

The March CPS data are available for download at *http://cps.ipums.org/cps/* as part of the Integrated Public Use Microdata Series (IPUMS-CPS) project at the University of Minnesota Population Center: See King et al. (2004).

We download the March CPS data from 1986 through 2007. We chose 1986 as our starting year because the CPS identifies only 15 metropolitan areas in prior years. The CPS wage and employment questions refer to the "previous calendar year." Therefore, data for any given year's CPS is treated as data appropriate for the previous calendar year. For example, variables generated from the March 2005 CPS would be treated as data for the year 2004.

In each year of our data, we use the following criteria to restrict the sample (with IPUMS-CPS variables in italics)

- Respondent lives in a household, not in group quarters or vacant units ($gq = 1$)
- Is aged 20 to 65 ($age \geq 20$ and $age \leq 65$)
- Wage and salary income in the previous calendar year is identified and is nonzero ($incwage > 0$ and $incwage < 999998$)
- Educational attainment is recorded ($educrec \geq 1$ and $educrec \leq 9$)
- Has an identified metro area of residence (*metarea* non missing)[5]

For each MSA, we create the following three variables:

1. Ratio of labor input of high skill to labor input of low skill, e_i/u_i
2. Ratio of total wages paid to total wages paid to low skill workers, χ_i
3. Average weekly wage of high skill workers, $w_{e,i}$.

[5] According to notes from the IPUMS-CPS, the metro area of residence was not collected from respondents, but added by the Census Bureau. The metro areas of residence is based on FIPS codes used in the 1990 census.

We use the *educrec* categorical variable to label respondents as either "low" or "high" skill workers. High skill workers are assumed to have completed 1+ years of college. Everyone else in the sample is assumed to be a low skill worker.

e_i is created as the total of weeks worked the previous calendar year ($wkswork1$) multiplied by the number of hours per week the respondent usually worked ($uhrswork$) for high skill workers. u_i is created as the same product, but for low skill workers. For each respondent, we weigh the product of $wkswork1$ and $uhrswork$ using the IPUMS-CPS sampling person weights, $perwt$.

χ_i is computed as

$$\frac{w_{e,i}e_i + w_{u,i}u_i}{w_{u,i}u_i} = \frac{\sum_{j\epsilon MSA_i} perwt_j \cdot wages_j}{\sum_{j\epsilon MSA_i} perwt_j \cdot wages_j \cdot 1\{unskilled_j\}}$$

for respondent j in MSA i, i.e. as the sum of all workers' pre-tax wage and salary income for the previous calendar year ($incwage$) divided by the sum of all low skill workers' pre-tax wage and salary income for the previous calendar year. We weigh pre-tax wage and salary income for all persons using the IPUMS-CPS sampling person weights.

$w_{e,i}$ is created as the sum of all high skill workers' pre-tax wage and salary income for the previous calendar year (created as an input into χ_i) divided e_i.

A.1.2 BEA Data (Output Prices)

We use two data sources from within the BEA web site: The Annual Industry Accounts, *http://www.bea.gov/industry/index.htm#annual*, and the Regional Economic Accounts data on Local Area Personal Income, *http://www.bea.gov/regional/reis/*.

Chain-type price indexes for industry output are available over the 1947-2007 period in the Annual Industry Accounts. Many industry price indexes are missing in 2007, so we do not use data from that year. To construct a price index for output produced by MSA, we merge this information with MSA-level data on earnings by industry that is available in Tables CA05 and CA05N of the Regional Economic Accounts. Earnings is inclusive of wage and salary disbursements, supplements to wages and salaries, and proprietors' income.

Thus, we assume that the price of output varies across MSAs because that industry composition varies across MSAs, and the price index for industry output varies across industries.

Denote $g_{t,j}$ as the growth rate of the price of industry output j from periods t to $t+1$ and g_t^i as the growth rate of the price of all output produced in MSA i between years t and $t+1$. Assuming output from $j = 1, \ldots, N$ industries is produced in MSA i in year t, we set the growth rate of the price of output produced in MSA i between years t and $t+1$ as

$$g_t^i \;=\; \sum_{j=1}^{N} \omega_{t,j}^i \; g_{t,j} \; . \tag{24}$$

The weight on each industry, $\omega_{t,j}^i$, is the share of total MSA earnings attributable to earnings of industry j in MSA i in year t:

$$\omega_{t,j}^i \;=\; \frac{\epsilon_{t,j}^i}{\sum_{k=1}^{N} \epsilon_{t,k}^i} \; , \tag{25}$$

where $\epsilon_{t,j}^i$ stands for total earnings of employees in industry j in MSA i during year t. In these computations, we only consider earnings from non-farm private industries. For each MSA, we construct a price index for output, normalized to 1.0 in the year 1969, that is consistent with the sequence of time-series estimates of g_t^i.

Ideally, we would compute the growth rate of the price of output produced in city i between years t and $t+1$ as

$$\sum_{j=1}^{N} \phi_{t,j}^i \; g_{t,j} \; , \tag{26}$$

with $\phi_{t,j}^i$ equal to the fraction of the nominal value of output in year t in MSA i that is accounted for by industry j. In an environment in which (a) output in each industry is produced by a set of identical firms all using a Cobb-Douglas combination of capital, labor, and land and (b) the labor-share of output is identical in each industry, assumptions that hold in our model, then industry j's share of nominal GDP in MSA i in year t, $\phi_{t,j}^i$, is equal to its earnings share $\omega_{t,j}^i$, and equations (24) and (26) are equivalent. In these calculations, we assume that proprietors' income are payments to labor.[6]

A few details complicate these calculations.

First, on a somewhat infrequent basis, Tables CA05 and CA05N do not report estimates of earnings for a given industry in an MSA in a given year. In these cases, we set earnings for this industry-MSA-year cell to zero.[7] Also, some of the industry-MSA-year employment estimates are marked with code E. According to the BEA web site, these estimates "constitute the major portion of the true estimate." In these cases, we assume that the reported estimate is equal to the actual estimate.

Second, the definition of industries in the Regional Accounts is not consistent across years. Table CA05 reports employment based on SIC-industry classifications over the 1969-2000 period and CA05N reports employment based on NAICS industry classifications spanning the years 2001-2006.

We map SIC and NAICS industry employment from Tables CA05 and CA05N to prices from the Annual Industry Accounts according to the tables below. The tables below list all the categories of nonfarm private employment. The sum of the earnings estimates in each of these categories is considered as total nonfarm private earnings, and is used to compute the denominator of equation (25).

[6]In the event that proprietors' income includes some payments to capital, equations (24) and (26) are equivalent as long as capital's share of proprietors' income and the fraction of earnings attributable to proprietors' income are both constant across industries.

[7]The three reasons that are listed for omission are (a) avoid disclosure of confidential information (code D), (b) earnings are less than $50,000 (code L), or (c) data not available for this year (code N). These omissions occur in approximately six percent of industry-MSA-year cells from 1969 to the mid-1990s and about thirteen percent of cells from the mid-1990s through 2006.

Data for Earnings Weights, $w^i_{t,j}$ Regional Accounts Table CA05, 1969-2000		Data for Growth in Prices, $g^p_{t,j}$ Industry Accounts, 1969-2001	
Line	Label	Line	Label
100	Agricultural services, forestry fishing and other	3	Agriculture, forestry, fishing and hunting
200	Mining	6	Mining
300	Construction	11	Construction
400	Manufacturing	12	Manufacturing
500*	Transportation and public utilities less electric, gas, and sanitary services	36	Transportation and warehousing
570	Electric, gas, and sanitary services	10	Utilities
610	Wholesale trade	34	Wholesale trade
620	Retail trade	35	Retail trade
700	Finance, insurance and real estate	50	Finance, insurance, real estate, rental and leasing
800	Services	59	Professional and business services

* See text for details.

Data for Earnings Weights, $w^i_{t,j}$ Regional Accounts Table CA05N, 2001-2005		Data for Growth in Prices, $g^p_{t,j}$ Industry Accounts, 2001-2006	
Line	Label	Line	Label
100	Forestry, fishing, related activities and other	5	Forestry, fishing and related activities
200	Mining	6	Mining
300	Utilities	10	Utilities
400	Construction	11	Construction
500	Manufacturing	12	Manufacturing
600	Wholesale trade	34	Wholesale trade
700	Retail trade	35	Retail trade
800	Transportation and warehousing	36	Transportation and warehousing
900	Information	45	Information
1000	Finance and insurance	51	Finance and insurance
1100	Real estate and rental and leasing	56	Real estate and rental and leasing
1200	Professional, scientific and technical services	60	Professional, scientific and technical services
1300	Management of companies and enterprises	64	Management of companies and enterprises
1400	Administrative and waste services	65	Administrative and waste management services
1500	Educational services	69	Educational services
1600	Health care and social assistance	70	Health care and social assistance
1700	Arts, entertainment and recreation	75	Arts, entertainment and recreation
1800	Accommodation and food services	78	Accommodation and food services
1900	Other services except public administration	81	Other services except government

In all cases except one, there is an exact correspondence of earnings estimates from Tables CA05 and CA05N to prices from the Annual Industry Accounts. For the SIC category of "Transportation and public utilities," line 500 of Table CA05, there is no clean analogous price index in the Annual Industry Accounts. Instead, the Annual Industry Accounts includes separate price indexes for "Transportation and warehousing" and "Utilities." In Table CA05, we therefore separate earnings of the single Transportation and public utilities into earnings in two categories: Earnings from utilities ("electric, gas, and sanitary services", line 570) and earnings from transportation and public utilities less earnings from utilities (i.e. line 500 less line

570).

A.1.3 Davis and Palumbo Data (Land Prices)

Davis and Palumbo (2008) report the quarterly value of land on a typical owner-occupied single-family lot for 46 MSAs over the 1984:4 - 2004:4 period. We directly use their data, but define the annual estimate as the average of the quarterly readings. Davis and Palumbo (2008) estimate the value of land as the market value of housing units less an estimate of the replacement cost of the structure (after accounting for depreciation) – see their paper for details.

A.1.4 OFHEO Data and 1990 Decennial Census of Housing (House Prices)

We create annual estimates over the 1985-2006 period of the average price of single-family owner-occupied housing units, by MSA, using a two-step procedure.

In the first step, we estimate the average price of single-family owner-occupied housing units in 1990 using household-level data from the 1990 Decennial Census of Housing (DCH). These data are available for download at *http://usa.ipums.org/usa/* as part of the Integrated Public Use Microdata Series (IPUMS-USA) project at the University of Minnesota Population Center. We use data from the 1990 DCH, rather than the 2000 Census of Housing, because more 1990 DCH reports more areas.

With IPUMS-USA variables in italics, we restrict the 1990 DCH sample to owner-occupied non-farm households living in single-unit residences in an identifiable MSA ($metarea > 0$, $ownershg = 1$, $farm \neq 1$, and $unitsstr \in \{3, 4\}$) who live in households and do not live in group quarters ($gq \in \{3, 4, 6\}$) and where the reported value of the house is not missing and is nonzero ($valueh > 0$ and $valueh < 999999$). The reported value of housing units, *valueh*, is top-coded at \$400,000 in 1990; we replace this top-coded value with \$662,000, which is the value of housing units worth more than \$400,000 in 1990.[8] Conditional on these restrictions, we compute the weighted average value of units by MSA using the sampling weight variable *hhwt*. These calculations yield estimates of the average price of housing for 272 metro areas as identified

[8]This estimate was first reported by Davis and Heathcote (2007) and is based on proprietary data in the 1989 Survey of Consumer Finances.

in the 1990 DCH.

In the second step, we extrapolate the price of housing in each metro area forward from 1990 to 2006 and backwards from 1990 to 1985 using MSA-specific constant-quality price indexes for owner-occupied single-family existing homes. These price indexes are published by the Office of Federal Housing Enterprise Oversight (OFHEO), until recently the regulator of the Fannie Mae and Freddie Mac organizations, and are available at *http://www.ofheo.gov/hpi_download.aspx*. The reported price indexes are quarterly; we convert to annual using a simple average of the quarterly readings.

For most metropolitan areas, OFHEO reports the price index for the MSA that are given by the list in the November, 2007 report of the Office of Management and Budget,[9] and so the mapping of MSAs from the IPUMS *metarea* variable to the OFHEO *cbsa* variable is one-to-one. For a few of the larger MSAs – Boston, Chicago, Dallas, Detroit, Los Angeles, Miami, New York, Philadelphia, San Francisco, Seattle, and Washington, DC – OFHEO reports the price indexes for metropolitan divisions within the MSA. In these cases, we set the growth rate of house prices of the entire MSA as the simple average of the growth rate of the underlying metropolitan divisions in the MSA.

In the OFHEO data, quite a few MSAs do not have a price index that extend back to 1985. After merging the 1990 DCH with the OFHEO price indexes, and eliminating the MSAs for which price indexes are not available back to 1985, we are left with annual estimates of house prices over the 1985-2006 period for 178 MSAs.

A.1.5 Decennial Data on Wages, Employment, House Prices, 1970-2000

We use decennial IPUMS data from 1970 to 2000 available at *http://usa.ipums.org/usa/*. Specifically, we use the 1% Metro sample for 1970[10], 1980, 1990, and 2000. In each year, we use the same respondent criteria as with the March CPS data described in section A.1.1. Note also the added requirement that the respondent is an employed, non-military civilian. The resulting sample restrictions are:

[9] For a complete list of the counties comprising each MSA, go to *http://www.census.gov/population/www/metroareas/metrodef.html*.

[10] The Form 2 version.

- Respondent is aged 20 to 65 ($age \geq 20$ and $age \leq 65$)
- Lives in a household, not in group quarters or vacant units ($gq = 1$)
- Has an identified metro area of residence ($metarea$ non-zero, non-missing)
- Educational attainment is recorded ($educrec \geq 1$ and $educrec \leq 9$)
- Wage and salary income in the previous calendar year is identified and is nonzero ($incwage > 0$ and $incewage < 999998$)
- Is an employed, non-military civilian ($empstatd = 10$ or $empstatd = 12$)

We create the following three variables for each MSA:

1. Ratio of labor input of high skill to labor input of low skill, e_i/u_i
2. Ratio of total wages paid to total wages paid to low skill workers, χ_i
3. Average hourly wage of high skill workers, $w_{e,i}$

We use the *educrec* categorical variable to label respondents as either "low" or "high" skill workers. High skill workers are assumed to have completed 1+ years of college. Everyone else in the sample is assumed to be a low skill worker.

e_i and u_i are the total number of hours worked by skilled and unskilled workers, respectively, in the previous calendar year for MSA i. It is calculated as the sum of hours worked last year (IPUMS variables $uhrswork_j \times wkswork1_j$, i.e. respondent j's usual hours worked last year multiplied by weeks worked last year) weighted by each respondent's sample weight, $perwt_j$. Since the 1970 sample only contains categorical codes for weeks and hours worked (IPUMS variables $wkswork2$ and $hrswork2$), we impute the 1970 values for $uhrswork$ and $wkswork1$ by using the national weighted average of $uhrswork$ and $wkswork1$ within each $hrswork2$ and $wkswork2$ category in 1980. For example, if a person belonged to the "50-52 weeks" category in 1970, then that person was assigned the national average among respondents in the "50-52 weeks" category in 1980, which was 51.7945.

The variables χ_i and $w_{e,i}$ are computed using the same proceudre as with the CPS data, described in section A.1.1. The average house price variable was created from *valueh*, whose top codes have been set to according to the scheme in Davis and Heathcote (2007):

- 2000: replace \$999,998 with \$1,860,000
- 1990: replace \$400,000 with \$662,000
- 1980: replace \$200,000 with \$350,000
- 1970: replace \$50,000 with \$87,500

In addition, for consistency in the *valueh* variable across years, we include only single-family housing units ($unitsstr = 3$ or $unitsstr = 4$).

A.2 Merging the Data

A.2.1 Annual Data Set with Land Prices, 1985-2004

Our first data set we use in estimation contains annual data with land prices over the 1985-2004 period. To create this data set, we merge the CPS data on wages and employment (section A.1.1) with the BEA data on output prices (A.1.2) and the Davis and Palumbo (2008) data on land prices (section A.1.3), which yields annual data on 42 MSAs.[11] In every MSA and date, the minimum number of respondents from the CPS is never less than 100. The median number of respondents is about 420 until about 2000, at which point the median jumps to about 650. The maximum number of respondents is typically over 4,000.

We merge these data by MSA. Note that the MSA definitions may not be completely consistent. In the BEA data, MSAs definitions are given by the list in the November, 2007 report of the Office of Management and Budget.[12] The MSA definitions in the CPS data are consistent with the definitions as of the 1990 Census. The Davis and Palumbo (2008) data are created by merging data from various sources, and

[11] We discard information on four areas reported by Davis and Palumbo (2008) because information on these areas is not available in the CPS.

[12] For a complete list of the counties comprising each MSA, go to *http://www.census.gov/population/www/metroareas/metrodef.html*.

the MSA definition of one of the key sources (data from the R.S. Means corporation) is not known.

A.2.2 Annual Data Set with House Prices, 1985-2006

Our second data set we use in estimation contains annual data with house prices over the 1985-2006 period. To create this data set, we merge the CPS data on wages and employment (section A.1.1) with the BEA data on output prices (A.1.2) and the annual data we construct on house prices by merging the OFHEO price indexes with information on house prices in the 1990 Decennial Census of Housing (section A.1.4). After all data are merged, we are left with a balanced panel of 146 MSAs.

Since we include many more relatively small MSAs in our sample than the first data set described in the previous section (A.2.1), the typical number of observations we use to compute our wage and employment variables is lower than the first data set. Across our 146 MSAs, the median number of observations in the CPS data is about 130 until about the year 2001, at which point the median number of observations jumps to approximately 170.[13] Until 2001, about 12 percent of our MSAs (around 18 MSAs) have less than 50 observations per year in the CPS data; after 2001, only 4 percent of our MSAs (6 MSAs) have less than 50 observations per year.

All data files are merged by MSA, and the caveat from the previous section – that the MSA definitions may not be completely consistent across data sets – applies.

A.2.3 Decennial Data Set with House Prices, 1970-2000

Our third data set in estimation contains data from the 1970, 1980, 1990, and 2000 Decennial Census of housing on wages, employment, and house prices. To create this data set, we merge the Decennial Census data on wages, employment, and house prices (section A.1.5) with the BEA data on output prices (A.1.2). After all data are merged, we are left with a balanced panel of 83 MSAs. As noted earlier, the MSA definitions from the Decennial Census files may not be completely consistent with the

[13] About 35 percent of our MSAs have more than 200 CPS observations in any year up through 2001; after 2001, the fraction of MSAs with more at least 200 CPS observations in any year rises to 45 percent.

MSA definitions from the BEA data.

Table 1: Summary of Data

	Standard Deviations:		
Variable	Data Set 1	Data Set 2	Data Set 3
$\hat{w}_{ei,t}$	0.105	0.164	0.096
$\hat{p}_{i,t}$	0.961	0.406	0.319
$\hat{q}_{i,t}$	0.053	0.072	0.051
$\hat{\chi}_{i,t}$	0.195	0.288	0.135
$\widehat{\frac{e_{i,t}}{u_{i,t}}}$	0.255	0.399	0.281

Data set 1:
Annual data, 1985-2004, 42 MSAs: $\hat{w}_{ei,t}$, $\hat{\chi}_{i,t}$, $\widehat{\frac{e_{i,t}}{u_{i,t}}}$ from CPS, $\hat{p}_{i,t}$ from Davis and Palumbo (2008), $\hat{q}_{i,t}$ from BEA.

Data set 2:
Annual data, 1985-2006, 149 MSAs: $\hat{w}_{ei,t}$, $\hat{\chi}_{i,t}$, $\widehat{\frac{e_{i,t}}{u_{i,t}}}$ from CPS, $\hat{p}_{i,t}$ constructed from 1990 DCH and OFHEO house price indexes, $\hat{q}_{i,t}$ from BEA.

Data set 3:
Decennial data, 1970-2000, 83 MSAs: $\hat{w}_{ei,t}$, $\hat{\chi}_{i,t}$, $\widehat{\frac{e_{i,t}}{u_{i,t}}}$, $\hat{p}_{i,t}$ from 1970-2000 DCH, $\hat{q}_{i,t}$ from BEA.

Table 2: Standard Deviations of the Data by Year

	Data Set 1:					Data Set 2:					Data Set 3:				
year	$\hat{w}_{ei,t}$	$\hat{p}_{i,t}$	$\hat{q}_{i,t}$	$\hat{\chi}_{i,t}$	$\frac{\widehat{e_{i,t}}}{u_{i,t}}$	$\hat{w}_{ei,t}$	$\hat{p}_{i,t}$	$\hat{q}_{i,t}$	$\hat{\chi}_{i,t}$	$\frac{\widehat{e_{i,t}}}{u_{i,t}}$	$\hat{w}_{ei,t}$	$\hat{p}_{i,t}$	$\hat{q}_{i,t}$	$\hat{\chi}_{i,t}$	$\frac{\widehat{e_{i,t}}}{u_{i,t}}$
1970											0.087	0.214	0.003	0.194	0.322
1980											0.080	0.288	0.041	0.139	0.265
1985	0.080	1.078	0.043	0.171	0.294	0.144	0.327	0.056	0.231	0.407					
1986	0.097	1.063	0.039	0.194	0.292	0.152	0.350	0.050	0.232	0.399					
1987	0.085	1.074	0.043	0.180	0.281	0.161	0.380	0.055	0.254	0.415					
1988	0.105	1.155	0.042	0.184	0.286	0.151	0.407	0.055	0.258	0.418					
1989	0.101	1.201	0.042	0.188	0.301	0.158	0.428	0.054	0.249	0.416					
1990	0.098	1.210	0.042	0.162	0.240	0.144	0.428	0.055	0.240	0.392	0.101	0.416	0.052	0.099	0.273
1991	0.103	1.162	0.042	0.198	0.280	0.140	0.413	0.055	0.280	0.430					
1992	0.103	1.069	0.043	0.177	0.243	0.137	0.398	0.056	0.272	0.409					
1993	0.087	0.979	0.045	0.188	0.250	0.147	0.381	0.059	0.267	0.372					
1994	0.092	0.906	0.046	0.193	0.250	0.142	0.362	0.061	0.296	0.403					
1995	0.116	0.866	0.048	0.212	0.256	0.191	0.352	0.063	0.295	0.394					
1996	0.114	0.841	0.051	0.189	0.227	0.188	0.343	0.067	0.295	0.392					
1997	0.120	0.829	0.053	0.202	0.253	0.170	0.342	0.070	0.321	0.442					
1998	0.120	0.809	0.055	0.191	0.236	0.169	0.348	0.072	0.343	0.435					
1999	0.109	0.813	0.059	0.216	0.255	0.168	0.362	0.077	0.299	0.407					
2000	0.134	0.830	0.065	0.242	0.252	0.193	0.392	0.084	0.337	0.442	0.115	0.330	0.078	0.081	0.265
2001	0.116	0.823	0.067	0.202	0.204	0.190	0.411	0.087	0.307	0.386					
2002	0.090	0.830	0.070	0.192	0.220	0.154	0.428	0.090	0.268	0.352					
2003	0.129	0.834	0.073	0.232	0.263	0.184	0.446	0.094	0.285	0.360					
2004	0.112	0.860	0.075	0.215	0.244	0.172	0.485	0.097	0.300	0.367					
2005						0.163	0.531	0.100	0.314	0.363					
2006						0.185	0.548	0.102	0.361	0.382					

Table 3: Parameter Estimates (Standard Errors in Parentheses)

	Data Set 1:		Data Set 2:			
					42 MSAs	42 MSAs
	42 MSAs, 1985-2004		149 MSAs, 1985-2006		1985-2006	1985-2004
Parameter	Unrestricted	$\alpha = 0.30$	Unrestricted	$\alpha = 0.30$	$\alpha = 0.30$	$\alpha = 0.30$
	(1)	(2)	(3)	(4)	(5)	(6)
ρ	0.491	0.568	0.376	0.543	0.591	0.514
	(0.023)	(0.021)	(0.032)	(0.027)	(0.017)	(0.022)
α	-0.066		-0.492			
	(0.120)		(0.244)			
δ	1.026	1.012	1.106	1.075	1.037	1.072
	(0.007)	(0.004)	(0.045)	(0.022)	(0.011)	(0.015)
ξ	0.590	0.550	0.655	0.683	0.590	0.578
	(0.005)	(0.005)	(0.014)	(0.016)	(0.005)	(0.005)
J-test	57.892	58.190	115.609	116.117	61.225	58.914
p-value	1.000	1.000	0.999	0.999	1.000	1.000
d.o.f.	146	147	166	167	167	147