Agglomeration and Sorting

Yujiang Chen and Coen Teulings[*]

June 2018

[Preliminary]

Abstract

Recent papers suggest a strong interaction between agglomeration externalities and human capital. We analyse a Mincerian wage equation with regional fixed effects and variation in the return to human capital, using data on 47 states and 34 metropolitan areas for the US. Agglomeration externalities are strongly related to the occupational structure. We show that regional difference in house prices offset these externalities.We develop a multi-region model with regional heterogeneity in workers and jobs, tradable versus non-tradable (land) commodities, consumption amenities, regional house prices, non-homothetic utility, and interregional labour mobility. The model allows for two forms of spatial organization: cities and rural areas. The model fits the regional data on the fixed wage effects, the return and mean level of human capital, land prices, and the city-rural area distinction well. The contribution of agglomeration externalities to GDP capitalizes in land values. We use these land values to calculate the value of agglomeration.

JEL classification: J24, J31, I26, R12, R13

Keywords: Wage Differential, Occupation, Human Capital, Geographic Sorting, Spillover Effect, Spatial Equilibrium

[*] Yujiang Chen, University of Cambridge, yc349@cam.ac.uk; Coen Teulings, Utrecht University and University of Cambridge, c.n.teulings@outlook.com. We thank Daron Acemoglu, Gabriel Ahlfeldt, David Albouy, Michael Amior, David Autor, Nicholas Bloom, Jan Eeckhout, Pieter Gautier, Edward Glaeser, Esteban Rossi-Hansberg, William Janeway, Koen Jochmans, Alan Manning and Guy Michaels, for helpful comments. We also benefited from very useful feedback from seminar audiences at Cambridge, CREST, CSIC, IFS, MIT, and Utrecht. All mistakes remain our own.

1 Introduction

Around 1920, Frits Philips was pondering where to set up the new factory of electric light bulbs. He considered several villages in the South Eastern part of the Netherlands, like Helmond, Veghel, and Veldhoven. He ended up in building his factory in the village of Eindhoven. Subsequently, that little village went through several decades of exceptional growth. By 1950, Eindhoven was the 7th city of the Netherlands, and 20 years later it had climbed to the 5th rank, a position Eindhoven still holds. Philips Electronics built extensive laboratory facilities, which were renowned in the industry. The city started its own technical university. From 1970 onwards, Philips Electronics went through a difficult episode. It had difficulty marketing its excellent technological innovations and went almost bankrupt. The renowned laboratories were closed down. Eventually, Philips decided to move its headquarters to Amsterdam, seeking a more open labour market and a better connection to the outside world. Eindhoven experienced a deep trough. But in the end, the backbone of former researchers of Phillips' laboratories, well trained engineers, many of them receiving their education at Eindhoven's technical university, saved the city. There were many new startups, often supported by Philips. Nowadays, the city is striving again, hosting ASML, the world leader in new production technologies for ICs. The current market capitalization of ASML exceeds that of its parent Philips.

This story is just one of many. Glaeser (2005) compared the different development of Boston and Detroit. Why did Boston recover from the closing down of its harbour activities, while Detroit is still struggling after the demise of the automobile industry? The answer is in the permanent inflow of new excellently trained workers, which were able - in Glaeser's words - to reinvent the city. Lucas (1988, 2001, 2009) used standard economic models to explain the importance of entrepreneurship, human capital externalities and the city agglomeration impact. In a recent paper, Gennaioli, La Porta, Lopez-de-Silanes and Shleifer (2013) use Lucas's model as framework for the analysis of the regional distribution of human capital and economic activity in more than 100 countries across the world. They find very strong agglomeration forces. Human capital tends to cluster within a small number of regions within each country. GDP per capita and wages in these regions are much higher than the nationwide average. These wage differentials are much higher than can be explained from the standard private return to human capital. For example, a simple regression of the regional GDP per capita on the regional mean years of education yields returns to a year of education far above any reasonable estimate of the private return to education, e.g. 54% for Brazil, 31% for India, 23% for Colombia, and 55% for Russia.

In this paper, we discuss the relationship among education, population, agglomeration, using a spatial equilibrium framework and discuss the importance of occupation structure. Our first conclusion is that knowledge spillover effects are big and significant. We present an empirical analysis on regional average wages, education levels and occupation structure. We use the

Current Population Survey in United States from 1979 to 2015, combining other regional level variables. US data provides a wide regional disparity and large data availability. In line with Gennaioli et al.(2013), year of education matters in term of the local wage determination. Knowledge spillover leads to a general higher social return to workers. Higher educated workers benefit other workers in the same area by increasing the overall wage. It is the quality of workers, rather than the quantity that matters. We use a simple comparative advantage human capital and occupation model to evaluate the results. By comparing and discussing the single index model and double index model, we provide evidence that the observed social return cannot be fully explained by unobservable worker characteristics. The observable features which describe the human capital level can nicely predict the unobservable part. Taking this into consideration, the estimated knowledge spillover can be a little smaller but still economically and statistically significant. Using only a few variables, this framework gives almost complete descriptions of the interregional average wage differential.

Our second conclusion is that shocks to the occupational structure can explain a large fraction of spillover and agglomeration effects. By empirical example, we show that occupation structures also have significant impacts. Areas with more complex occupations in general have higher income. Spill-overs come from certain occupational structures. We decompose the impact by occupation and find out that computer science and financial services generate most of the occupation spillover. Interestingly, entrepreneurs have no impact in our sample. The full model gives both a static and dynamic explanation on regional wage determination and occupation structure shocks. We integrated the cities and states sample, by setting up both city and non-city regions in a spatial equilibrium and agglomeration model, similar to Lucas and Rossi-Hansberg (2002). Three agglomeration forces are considered. In city regions, workers commute to the center and knowledge spillover is fully transmitted into total productivity. However, due to the commuting cost, the productivity is partly consumed by the crowding effect. In non-city areas, there is no commuting cost, but the knowledge spillover will diminish along the distance, at a given decay rate. The last distance concept is the transportation cost. It gives a larger market for regions with better connections with their neighboring regions, due to the larger market demand. Wage and housing price or the land price are the outcome of local market and play the role to guide the inflow and outflow of labour force, to further make the economy reach the equilibrium states. In production technology, we use the search and matching friction model by Gautier and Teulings (2006), who focus on search frictions. Here, search frictions are ignored.

Last, we estimate the welfare gain from agglomeration and the knowledge spillover. Using the theoretical model, we estimate the equation system using instrumental variables, labour demand Bartik instruments, January temperature, and the roughness of land area. Estimations indicate that occupation structure shocks can explain the correlations among the key variables. We further calculate the macro valuation of externalities, i.e. the welfare gain from agglomerations, using housing wealth and its counterfactual value.

2 Some empirical evidence

2.1 A simple statistical model

The strawman of our paper is a simple statistical model following ideas of the hedonic "kissing curves" models by Rosen (1974), Sattinger (1975), and Teulings (1995) that serves as a first description of our data. The model describes the wage rate for workers i which differ by their level of human capital and who can take jobs with different occupational complexity in an economy with multiple regions r. Both worker's human capital h_i and the occupational complexity o_i can be captured by a single index. Each region has a separate labour market. Relative wages may differ between region. However, within each region, wages are increasing in workers' human capital and there is perfect positive sorting of job complexity by human capital. Since wages are an increasing function of human capital within each region, we can use a wage equation to obtain a metric for the worker's human capital. We do this by means of a simple log linear wage equation[1]

$$\begin{aligned} w_{ir} &= \omega_0 + \omega_r + \rho_r h_i + e_{ir}, \qquad (1) \\ h_i &\equiv \widehat{h}_i + \underline{h}_i, \\ \widehat{h}_i &\equiv \omega' x_i, \end{aligned}$$

where w_{ir} is the observed log hourly wage for worker i working in region r, where $\widehat{h}_i$ and $\underline{h}_i$ are the observed and unobserved part of human capital respectively, where e_{ir} reflects measurement error in the observed wage, and where x_i is a vector of standard personal characteristics like age, years of education, gender, marital status, and race. Without loss of generality, we assume the components $\widehat{h}_i$ and $\underline{h}_i$ to be orthogonal over the full sample of all regions. For the sake of convenience, all elements of x_i are demeaned over the full sample; hence, $\mathrm{E}\left[\widehat{h}_i\right] = 0$. Similarly, we assume that $\underline{h}_i$ has mean zero over the full sample; however, within a region, its mean might be different from zero due to selective interregional migration. The parameter ω_0 is an overall intercept, ω_r is a dummy for each region r, ρ_r is the return to h_i in each region r, and ω is a vector of parameters of the same dimension as the vector x_i which is common to all regions; the parameter vector aggregates the components x_i into the single human capital index h_i. For future reference, it is useful to define

$$\widehat{\omega}_{0r} = \omega_r + \rho_r \mathrm{E}\left[\underline{h}_i | r\right]. \qquad (2)$$

[1] The linearity of this equation is not an important restriction on the generality of the analysis, see Gautier and Teulings (2008). Suppose wages are an increasing but non-linear function of some human capital index h^*: $w = w(h^*) = w(\omega' x)$, with $w'(h^*) > 0$. By defining a transformed human capital index $h = w(h^*)$ the linearity can be imposed without loss of generality. The non-linearity in the relation with the vector x can then be addressed by applying a polynomial in x.

Estimation of equation (1) by standard techniques, see the discussion below, yields an estimate of $\widehat{\omega}_{0r}$, but not of ω_r, because $\underline{h}_i$ and hence $\mathrm{E}[\underline{h}_i|r]$ is unobserved. Referring h_i as the human capital of a worker glosses over all kind of hairy issues like whether the effect of gender or race might attributed to differences in human capital or that these variables might be proxies for all kind of other processes, like interrupted careers of women or discrimination against women and blacks. Since our aim is just to agglomerate all observable variables in a single index that reflects the earning capacity of a worker, we sidestep these issues.

Equation (1) imposes a multiplicative restriction compared to a more general model

$$w_{ir} = \omega_0 + \omega_r + \omega_{xr}' x_i + e_{ir}'.$$

In equation (1), interregional variation in ρ_r affects the returns to the each components of x_i proportionally, while in the general model, the returns to each component of x_i can vary independently between regions.[2] Substitution of the definition of $\widehat{h}_i$ in the wage equation yields

$$w_{ir} = \omega_0 + \omega_r + \rho_r \left(\omega' x_i \right) + e_{ir}^{(h)}, \tag{3}$$

where $e_{ir}^{(h)} = \rho_r \underline{h}_i + e_{ir}$. The full model in equation (1) is unidentified: multiplying the vector ω by a constant and dividing all ρ_r by that same constant changes does not change anything. The same applies to ω_0 and ω_r. Hence, we add convenient normalizing assumptions for the mean of ρ_r and ω_r across regions.

$$\mathrm{E}_r \left[\omega_r \right] = 0, \qquad \mathrm{E}_r \left[\rho_r \right] = 1. \tag{4}$$

For future reference, we refer to R_h^2 as the share of observed human capital in the total variance of human capital

$$R_h^2 \equiv \frac{\mathrm{Var}\left[\widehat{h}_i\right]}{\mathrm{Var}\left[\widehat{h}_i\right] + \mathrm{Var}\left[\underline{h}_i\right]}. \tag{5}$$

Equation (3) can be estimated by NLLS (Non-Linear Least Squares). The non-linearity stems from the multiplicative restriction on the coefficients $\rho_r \omega$. We can apply a simple iterative scheme for the estimation of the model. First, estimate equation (1) setting $\rho_r = 1$ for all regions to obtain a first estimate for $\omega_0, \widehat{\omega}_{0r}$, and ω. Next, calculate $\widehat{h}_{ir} = \omega' x_i$ and estimate equation (3) to obtain a second estimate for ω_0 and $\widehat{\omega}_{0r}$ and a first estimate for ρ_r. Then, calculate $x_{ir}^{(2)} = \rho_r x_{ir}$ and use these data to reestimate equation (1) from the first step to obtain new parameters for $\omega_0, \widehat{\omega}_{0r}$, and ω, etc.. In practice, the estimates from this second step deviate hardly from the first step. We can test the non-linear restriction $\omega_r = \rho_r \omega$ by means of a likelihood ratio test. Due to the large number of data, this test will surely reject this non-linear restriction. Alternatively, we can ask what share of

[2] The saturated model has $R \times (K+1)$ parameters, where R is the number of regions and K is the dimension of the vector x_{ir}, while equation (1) has only $2R + K + 1$ coefficients.

the variance of a model with a separate parameter vector ω_r for each region r relative to variation spanned by a restricted model with a single common vector ω for all regions is covered by the restriction $\omega_r = \rho_r \omega$. For future reference, we define the mean level of observed human capital in region r as

$$\widehat{H}_r \equiv \mathrm{E}\left[\widehat{h}_{ir}|r\right];$$

H_r is defined similarly; the means of $\widehat{H}_r$ and H_r over the full sample are zero by construction (since x_i is demeaned over de full sample and $\mathrm{E}[\underline{h}_i] = 0$).

These equations describe the supply side. We use a similar model for the demand side. Since wages are increasing in human capital and since there is perfect positive sorting of job complexity by human capital, wages are increasing in the job complexity. Hence, analogous to equation (1), we obtain a measure of occupational complexity using the following model

$$\begin{aligned} w_{ir} &= \omega_0 + \chi_{0r} + \chi_{or} o_i + e_{ir}, \\ o_i &\equiv \widehat{o}_i + \underline{o}_i, \\ \widehat{o}_i &\equiv \chi' z_i, \\ \mathrm{E}_r\left[\chi_{0r}\right] = 0, &\qquad \mathrm{E}_r\left[\chi_{hr}\right] = 1, \\ \widehat{O}_r &\equiv \mathrm{E}\left[\widehat{o}_i|r\right]. \end{aligned} \tag{6}$$

where z_i is a vector of occupational dummies; like x_i, z_i is demeaned over the full sample.[3] We shall refer to $\widehat{o}_i$ as the level of observed occupational complexity. Again, the means of $\widehat{O}_r$ and O_r over the full sample are zero. R_o^2 is defined analogously to R_h^2.

2.2 Data

We draw data from five different sources. Individual level data are taken from the Current Population Survey, Merged Outgoing Rotation Groups (CPS-MORG) from 1979 till 2015. We use the hourly wage, years of education, occupation, industry and other demography information as gender, age, marital status, and race. Our sample includes all workers age 16 to 64.

For our classifications of regions, we select 34 MSAs for which both individual level and regional level data set are available. We then take the remaining part of each state as one non-city region. The definition of MSAs changes overtime. To make the samples consistent, we match different ID of these areas over time. From 1979 to 1985, we use 1970 Census ranking to identify MSAs. From 1986 to 1988, we use CMSA and PMSA identifier. From 1989 to 2003, we use MSAFIPS and for the rest of samples we use CBSAFIPS. Out of the total sample, 2,027,727 obser-

[3] By construction, the intercept of this regression must be the same as in equation (1), since all explanatory variables are demeaned in both regressions. The element of χ corresponding to a particular occupation is the mean log relative wage in that occupation in the full sample.

vations, 36.7%, are living in MSAs and the proportion is stable over time. We have 47 Non-MSA state regions. As is common practice, we exclude Hawaii and Alaska. Furthermore, New Jersey is excluded since all of its area is part of NY-NJ MSA, leaving us with 34 MSAs and 47 rural areas, 81 regions in total. The list of MSAs is in appendix, see Table A1.

The definition of occupation and industry varies over time. We use the occupation definition proposed by Autor and Dorn (2013), which uses the similarity of tasks to form a consistent definition of occupations over time. We consider 330 different 3-digit occupations. We use the industry definition suggested by Autor, Levy, and Murnane (2003) and the crosswalk constructed by IPUMS.

The regional population and employment data are taken from US Census Bureau. The Housing Price Index (HPI) is taken from the Federal Housing Finance Agency All-Transaction Indexes, both for states and for Metropolitan Statistical Areas (MSAs). To make the housing price comparable across regions, we also calculated the housing value index, using the additional information from Zillow Research, Zillow Home Value Index. We use the estimated median home value for all homes within a region. State level data for two states, Maine and Louisiana, are missing from this data set. Instead, we use the available average home value at the county level. Proximity data is from US Department of transportation. We collect information on the bordering regions and whether a given region is close to sea coast or one of the main navigable rivers. Average January temperature data is from US National Oceanic and Atmospheric Administration, 1981-2010 US climate normals, following Glaeser (2004). All regional data covers the research period 1979 to 2015 at annual frequency.

2.3 Summary statistics

We use these data to estimate equation (1) and (6). We add time dummies to account for nation wide nominal wage growth. Estimation results for the parameter vector ω are presented in the Appendix, estimation results for χ are available from the authors on request. Three examples of the index $\widehat{h}_i$ characterize the distribution. The 10th percentile of the distribution of $\widehat{h}_i$ is -0.426, a typical worker in this group is a black married female with 10 years of education and 26 years of experience. The median value of $\widehat{h}_i$ is -0.012, corresponding to a white married male with 12 years of education and 8 years of experience. The 90 percentile is 0.429, which corresponds to a white married female with 18 years of education and 21 year of experience. On the demand side, the 10th percentile of the distribution of $\widehat{o}_i$ is -0.465, which are occupations like laundry and dry cleaning. The median is -0.019: health technicians. The 90th percentile is 0.396, which are financial service sales. Clearly, the median level of $\widehat{h}_i$ and $\widehat{o}_i$ should be close to zero by construction, since we demeaned x_i and z_i across the full sample; the only reason for the medians not being exactly zero is that the median is not equal to the mean.

Table 1: Summary Statistics

Individual level			Var Decomposition %					
Variable	Mean	S.D.	Inter-region			Intra-region		
w_{ir}	2.387	0.5691	4.15			95.85		
$\hat{h}_i$	0.000	0.3432	0.92			99.08		
$\hat{o}_i$	0.000	0.3341	1.07			98.93		
$\text{cor}(\hat{h}_i, \hat{o}_i)$	0.5655							
Regional level			Correlation Matrix					
Variable	Mean	S.D.	H_r	O_r	$\widehat{\omega}_{0r}$	ρ_r	χ_{0r}	χ_{or}
$\widehat{H}_r$	0.0010	0.0323	1					
$\widehat{O}_r$	0.0010	0.0320	0.7854	1				
$\widehat{\omega}_{0r}$	-0.0637	0.0811	0.6363	0.7913	1			
ρ_r	0.9860	0.0480	0.1180	0.4462	0.5493	1		
χ_{0r}	-0.0397	0.0785	0.7270	0.7265	0.9695	0.4622	1	
χ_{or}	0.9861	0.0456	0.2019	0.5356	0.5311	0.8603	0.4376	1

Note: The summary statistics for key individual and regional level variables. Log wage from data. Human capital index and occupation index are calculated using equations in section 2. Mean and standard deviation are presented. Variance decomposition represents the contribution of inter- and intra- region.
Data Source: Current population survey, the US Census Bureau,the Federal Housing Finance Agency, Zillow Research, the US Department of transportation, and the US National Oceanic and Atmospheric Administration.

The estimation results for equation (1) are summarized in Table A2. The variance of in the human capital index $\widehat{h}_i$ is $(0.343/0.569)^2 = 36\%$ of the variance of log wages (the numbers for $\widehat{o}_i$ are very similar). Most of the variance in observed human capital is within regions; only 1% of the variance is between regions. The standard deviations of the mean observed levels of human capital and occupational complexity, $\widehat{H}_r$ and $\widehat{O}_r$, are 3%, while the standard deviation of inter-regional wage differentials is $\sqrt{0.0415} \times 0.569 = 11\%$. Hence, since the average return to human capital is normalized to unity, observed human capital explains only only $(0.03/0.11)^2 = 7\%$ of the interregional variation in wages. The estimated region fixed effect $\widehat{\omega}_{0r}$ accounts for a large part of the remaining variation; its standard deviation is 8%. The mean observed levels of human capital and occupational complexity, $\widehat{H}_r$ and $\widehat{O}_r$, are positively correlated across regions: a region with a well-educated workforce tends to have many jobs in complex occupations. Finally, we observe that the returns to human capital and to occupational complexity, ρ_r and χ_{or}, vary substantially between regions (std.dev. 5%), but are strongly correlated across regions (86%). Since their means and standard deviations are also almost equal, this justifies a simplifying assumption.

The Equal Return Assumption

$$\rho_r = \chi_{or}, \qquad \forall r.$$

Though this assumption does not hold exactly, it provides a good first order approximation of the data. It simplifies the subsequent analysis. Elimination of w_{ir} from equation (1) and equation (6) gives

$$o_i = \frac{\omega_r - \chi_{0r}}{\chi_{or}} + \frac{\rho_r}{\chi_{or}} h_i.$$

The assignment of human capital type h_i to occupation type o_i is an increasing linear function with the interregional variation in its coefficients. This equation can be simplified by imposing the Equal Return Assumption

$$o_i = \frac{\omega_r - \chi_{0r}}{\chi_{or}} + h_i. \tag{7}$$

Taking expectations in equation (7) for region r shows that the same relation between h_i and o_i that holds at the individual level, also holds at the aggregate level between H_r and O_r. Solving for $\frac{\omega_r - \chi_{0r}}{\chi_{or}}$ and substitution in equation (7) yields

$$O_r = \frac{\omega_r - \chi_{0r}}{\chi_{or}} + H_r, \tag{8}$$
$$h = o - O_r + H_r.$$

Under the Equal Return Assumption, the interregional variation in the assignment of human capital to occupational complexity is a simple mean shift. There is no a priori reason for this assumption to hold. It just turns out to be a good description of the data.

2.4 The Proportionality Assumption

Our interpretation of the occupational structure as capturing the supply side of the labour market is not undisputed. For example, Autor and Dorn (2013) have argued that our observed measures of human capital are imperfect and that occupation is just an alternative measure for the worker's human capital. This section addresses this issue: what information can we extract from data on the worker's occupation?

We refer to the model set out in Section 2.1 as the *Single Index Model*. This model claims that a worker's human capital can be meaningfully summarized in a single index h (we drop the index i in what follows, since it is by now clear which variables are defined on the individual level). Its decomposition in an observed and an unobserved component is irrelevant from an economic point of view; both components are perfect substitutes. The model implies that high human capital workers sort into complex occupations; see equation (7). Since human capital and occupational complexity are only partially observed, the model predicts that the observed part of occupational complexity will be correlated to the unobserved part of human capital. Hence, in a regression with both human capital variables and occupational dummies, the effect of occupational dummies proxies unobserved human capital, and the other way around. However, this view implies that when aggregating over all individuals in region r, the means of observed human capital and occupational complexity, $\widehat{H}_r$ and $\widehat{O}_r$, are an imperfect measure of supply and demand for human capital in region r due to the selectivity in the unobserved components.

The alternative view is dubbed the *Double Index Model*. Where the Single Index Model assumes that observed and unobserved characteristics can be aggregated into a single index h, the Double Index Model assumes that $\widehat{h}$ and $\underline{h}$ measure economically different aspects of workers' human capital, which are required for different occupations and for which the return might therefore vary independently between regional markets. For example, let us presume for the sake of the argument that formal education is conducive to a student's analytical skills, but not to her emotional skills. In that case, the index h alone is a sufficient statistic for IQ, but not for EQ. Hence, the occupational assignment of a worker cannot be predicted based on the index h alone; one needs information on its decomposition in $\widehat{h}$ and $\underline{h}$. In a wage regression with both standard human capital variables and occupational dummies, the occupational dummies pick up the effect of EQ on the occupational assignment.

A meaningful comparison of both interpretations of the data requires us to raise the hurdle for the Single Index Model by making a further assumption. This assumption increases the empirical content of the model, which makes it therefore more easy to reject the model (hence, we refer to it as the Extended Single Index Model). Equation (7) specifies the relation between observed indexes h and o implied by the model. However, it does not specify the relation between the unobserved and the unobserved components of both indexes. The assumption below fills this gap.

The Proportionality Assumption

When a worker with human capital $h = \widehat{h} + \underline{h}$ is assigned to an occupation $o = \widehat{o} + \underline{o}$ in region r, then the index h is a sufficient statistic for the expectation of the observed and unobserved component of the occupational complexity indexes $\widehat{o}$ and $\underline{o}$ respectively. The components of h, $\widehat{h}$ and $\underline{h}$, do not provide further information about these expectations. In particular, the following relations apply

$$\begin{aligned} \mathrm{E}\left[\widehat{o}|h,r\right] &= R_o^2 \mathrm{E}\left[o|h,r\right], \\ \mathrm{E}\left[\underline{o}|h,r\right] &= \left(1-R_o^2\right)\mathrm{E}\left[o|h,r\right], \end{aligned} \tag{9}$$

where $\mathrm{E}[o|h,r]$ is given by equation (7). Similarly, o is a sufficient statistic for the expectation of the observed and the unobserved component of human capital, $\widehat{h}$ and $\underline{h}$, and mutatis mutandis the same relations as in equation (9) apply.

This assumption is a natural extension of the idea that observed and unobserved human capital are perfect substitutes and that the decomposition of h in these components is therefore irrelevant

Table 2: Intraregional Variance Decomposition

No.	Variance	Data	Formula	Calculated
1.	$\text{Cov}\left[\widehat{h}, \widehat{o}\right] / \text{Var}[w\|r]$	21%	$(1-E)\, R_h^2 R_o^2$	20%
2.	$\text{Cov}\left[\widehat{h}, \underline{o}\right] / \text{Var}[w\|r]$	16%	$(1-E)\, R_h^2 \left(1-R_o^2\right)$	17%
3.	$\text{Cov}[\underline{h}, \widehat{o}] / \text{Var}[w\|r]$	16%	$(1-E)\left(1-R_h^2\right) R_o^2$	17%
4.	$\text{Cov}[\underline{h}, \underline{o}] / \text{Var}[w\|r]$	17%	$(1-E)\left(1-R_h^2\right)\left(1-R_o^2\right)$	15%
5.	$\text{Var}[e] / \text{Var}[w\|r]$	30%	E	30%

Note: The decomposition of intraregional variance of wages. Human capital index and occupation index are calculated using equations in section 2. *Data Source*: Current population survey.

for the decomposition of the occupational complexity of the worker's job into its observed and unobserved component.

The Proportionality Assumption implies that the intraregional variance in log wages can be decomposed into five orthogonal components. Let E be the share of the measurement error in the variance of the data on wages. Then,

The percentage for the first three components can be calculated directly from the data.[4] The sum of the final two components can then be calculated as a residual item. However, its decomposition into both components cannot be derived from the data. We apply an independent estimate of the measurement error in wages by Angrist and Krueger (1999). Then, the covariance between the unobserved components follows as a residual item.

As a first test of the Extended Single Index Model, we check whether this distribution into five components is consistent with the Proportionality Assumption. We have the last column of Table 2. A similar calculation as for R_h^2 yields the same value R_o^2 (which is accidental; it follows from the equality of Component 2 and 3). We use these numbers to calculate the predicted share of Component 1 and 4 in the total variance; the result is in line which the outcome that is predicted by the Proportionality Assumption. This is a first confirmation of the Extended Single Index Model.

The second test directly compares the Single to the Double Index Model. If $\widehat{h}$ and $\underline{h}$ measure different aspects of human capital, as in the Double Index Model, they span a two-dimensional space. Different combinations of $\widehat{h}$ and $\underline{h}$ make a worker apt for different occupations even when their sum h is the same. Since our occupational classifications has more than 300 entries, this classification can be expected to span this two dimensional space. The linear combination that correlates best to the observed component $\widehat{h}$ should therefore be different from the linear combi-

[4] For the second component, we use

$$\text{Cov}\left[\widehat{h}, \underline{o}\right] = \text{Var}\left[\widehat{h}\right] - \text{Cov}\left[\widehat{h}, \widehat{o}\right]$$

which holds since $\widehat{o} + \underline{o} = o = h$. A similar relation applies for $\text{Cov}[\underline{h}, \widehat{o}]$.

nation that correlates best to the unobserved component $\underline{h}$. The estimation results on equation (3) provide an estimate of the observed component $\widehat{h} = \omega' x$. We also have a measure of the unobserved component: $\rho_r^{-1}\left(w - \omega_0 - \rho_r \widehat{h}\right)$. Hence, we run two regressions

$$\begin{aligned} \widehat{h} &= \chi^{(1)\prime} z + e^{(1)}, \\ \frac{w_i - \omega_0 - \omega_r}{\rho_r} - \widehat{h} &= \underline{h} + \rho_r^{-1} e = \chi^{(2)\prime} z + e^{(2)}, \end{aligned}$$

where $e^{(1)}$ and $e^{(2)}$ are individual specific error terms. If the Single Index Model holds, the correlation between $\chi^{(1)\prime} z$ and $\chi^{(2)\prime} z$ should be close to unity. Instead, if the Double Index Model holds, this correlation should be substantially lower. Moreover, the R^2 of the first regression should be $R_o^2 = 0.53$ (the share of observed occupational complexity in the total variance), while the R^2 of the second regression should be $\frac{1-R_h^2}{1-R_h^2+E} R_o^2 = 0.32$; the R^2 of this second regression is lower since the measure of unobserved human capital is diluted by the measurement error in wages. The actual correlation between $\chi^{(1)\prime} z$ and $\chi^{(2)\prime} z$ is 0.84, while the actual R^2s are 0.36 and 0.11 respectively. This provides strong evidence in favour of the Extended Single Index Model.

The Extended Single Index Model yields an estimate of the actual means H_r and O_r of human capital and occupational complexity based on the means $\widehat{H}_r$ and $\widehat{O}_r$ of their observed components. Taking expectations for region r in equation (9) yields

$$\begin{aligned} H_r &= R_h^{-2} \widehat{H}_r = 1.89 \widehat{H}_r, \\ O_r &= R_o^{-2} \widehat{O}_r = 1.89 \widehat{O}_r, \end{aligned} \tag{10}$$

where we use $R_h^2 = R_o^2 = 0.53$ in the last step. These relations extend the selection process for a single occupation o to the regional level. Though the relation between o and h is deterministic, see equation (7), the deconvolution of h in its observed and unobserved component is random. The workforce in region r is a selective sample from the nation's total workforce. This selection follows the same rules as the selection process for an individual occupation. The selection affects the observed and the unobserved component proportional to their share in the variance of h for the total population.

Under the Proportionality Assumption, the mean of observed human capital, $\widehat{H}_r$, is therefore an underestimation of the mean of total human capital, H_r, since it ignores the selectivity in the unobserved component. A similar argument applies to the mean of occupational complexity. One might object that this assumption about the selectivity of unobserved components is speculative. However, this approach follows immediately from the Proportionality Assumption for the selection process at the level of individual occupations, for which we provided strong empirical evidence. It is hard to conceive a model where this process would apply at the level of individual occupations, while it would not apply when aggregating over all occupations in a region. Equation (10) implies

$$\mathrm{E}\left[\underline{h}_i|r\right] = \frac{1-R_h^2}{R_h^2}\widehat{H}_r,$$

Hence, by equation (2), ω_r satisfies

$$\omega_r = \widehat{\omega}_r - \rho_r\frac{1-R_h^2}{R_h^2}\widehat{H}_r = \widehat{\omega}_{0r} - \rho_r 0.89\widehat{H}_r, \tag{11}$$

where we use again $R_h^2 = 0.53$ in the last step. Using equation (10) for calculation of the total interregional variance in human capital, human capital explains $1.89^2 \times (0.03/0.11)^2 = 27\%$ rather than 9% of the interregional variation in log wages; see Section 2.3.

One can take the argument in favour of the Double Index Model one step further, by considering a *Multiple Index Model*. Where the Double Index Model assumes that $\widehat{h}$ and $\underline{h}$ are different components of human capital, the Multiple Index Model claims that the observed human capital $\widehat{h}$ measures general human capital, but that occupations require specific skills that are badly measured by years of education but are picked up by the occupational classification. In this view, the effect of occupations in a wage regression measures the market value of this occupation specific human capital. Taking it to the extreme, a secretary is called a secretary, not because she has a different job than her boss, but because she has a typing certificate, unlike her boss. The arguments in favour of the Single Index Model in comparison to the Double Index Model apply a fortiori to the Multiple Index Model.

2.5 What explains the interregional variation?

Table 1 has documented the substantial interregional variation in the intercept $\widehat{\omega}_{0r}$ and the return to human capital ρ_r in regression for log wages. What explains this variation? This section provides a first answer to this question. We regress the regional dummies ω_r (derived from $\widehat{\omega}_{0r}$ using the correction in equation (11)) on a number of explanatory variables, see Table 3. Column (1) includes only the regional mean observed level of human capital H_r (using the correction in equation (10)). It comes in significantly with a coefficient of 0.38, the R^2 being 0.12. Our methodology yields an easy interpretation of the magnitude of the coefficient.[5] A coefficient of unity would imply that when we raise the observed human capital of all workers by 0.01, which increases their wage via the term $\rho_r\widehat{h}$ by 1% due to our normalization, then there is an indirect effect via the intercept ω_{0r} of 0.38%. It is tempting to interprete this as evidence of agglomeration externalities from a better skilled workforce.

[5] The reader might who is sceptical about the proposed correction for selectivity might wonder what would be the size of the spillover when using the uncorrected estimated $\widehat{\omega}_{0r}$ and $\widehat{H}_r$. In that case, the coefficient would be much larger, 1.58.

Table 3: Interregional variation of the intercept

	(1)	(2)	(3)	(4)	(5)
VARIABLES	Adj. Avg. Wage ω_r				
Human Capital Index H_r	0.376	0.250	0.00768	-0.330	-0.531
	(3.44)	(3.11)	(0.09)	(-2.67)	(-2.97)
Occupation Index O_r				0.575	0.736
				(2.95)	(3.11)
City Dummy			-0.308	-0.299	-0.284
			(-2.93)	(-5.36)	(-4.27)
City x ln Population			0.0262	0.0237	0.0225
			(3.51)	(6.07)	(4.77)
Spatial Lag		0.865	0.747	0.685	0.658
		(7.75)	(7.50)	(8.20)	(7.72)
South Dummy					-0.0229
					(-2.24)
Constant	-0.0653	0.000480	-0.0358	-0.0301	-0.0239
	(-9.36)	(0.05)	(-3.10)	(-3.29)	(-2.75)
R-squared	0.118	0.480	0.688	0.760	0.773
R-MSE	0.0630	0.0486	0.0382	0.0337	0.0330

Note: Columns (1)-(5) present the estimated social returns using OLS regression. Dependent variable is the average wage controlling for individual human capital w_r. Human Capital index measures the average human capital in a region. Occupation index measures the occupation complexity of local labour market. Metro Dummy equals one if an observation is city area, zero if it is non-city area. Log MSA Population is the reported population in each region. Spatial local wage is the average of all the neighbouring region wages. Detailed definitions in section 2. Robust t-statistics in parentheses.

In column (2) we measure the potential interactions between regions by adding a spatial lag.[6] Agglomeration benefits in neighboring regions might spill over to the own region. The spatial lag is highly significant and the R^2 increases to 0.48. The coefficient on H_r drops slightly. In column (3) we add city variables: a city dummy and the log population of a city; the combination of both terms implies that any city with more than 100,000 inhabitants pays higher wages than the countryside. This applies to all cities in our sample. The R^2 increases to 0.69. The coefficient on H_r is no longer significant. Clearly, the city variables and the mean level of education level are highly correlated. The spatial lag term remains largely unaffected.

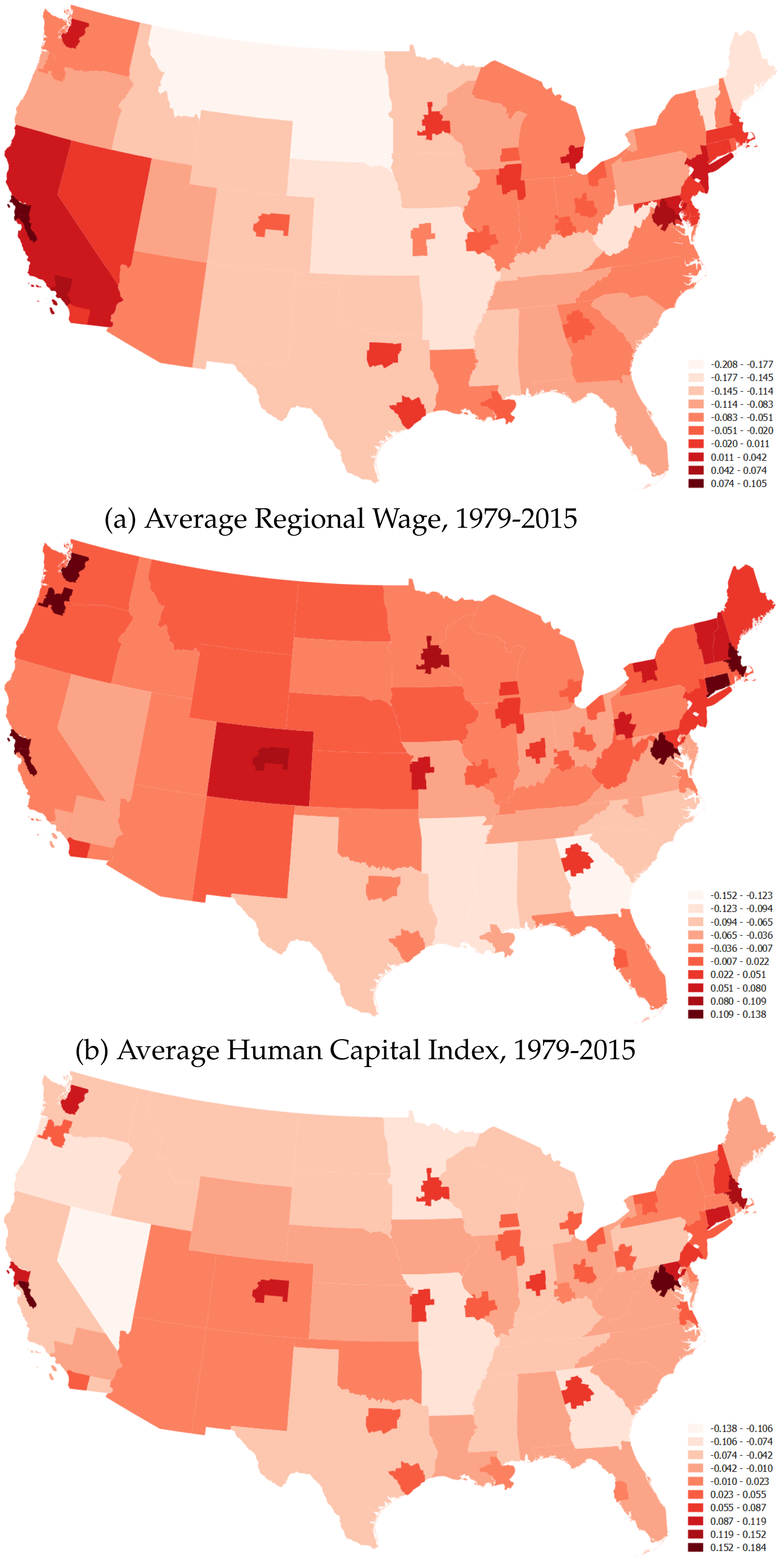
The surprise comes in column (4), where the mean occupation index O_r is added as a regressor. The R^2 jumps again, to 0.76 in this case. The coefficient on H_r flips signs, while the coefficient on O_r takes over and is clearly significant. Just five variables explain the main part of the variation of regional fixed effects in a log linear wage regression. Apperently, spillovers do not stem from better educated workers, but from the more complex occupations that these workers have. As documented in Table 1, both variables are highly correlated (79%). Nevertheless, we are able to separately estimate their effect with reasonable precision.

Finally, column (5) repeats the same regression with a dummy for the South, as is a standard practice in many empirical studies. Its coefficient is significant but small. The lower wages in the South are largely explained by the five variables in column (4).

Figure 1 shows the region fixed ω_r for each region. Three observations catch the eye. First, there exist large interregional wage differentials. Second, cities have higher fixed effects than the surrounding regions. Most cities are concentrated in small number of regions, usually close to navigable water: the Northeast and the West coast, the Mississippi River basin, and the Lake region. We experimented a bit by adding a dummy for the proximity to navigable waters, but after controlling for our four main variables, this did not have a significant effect on the region fixed effects ω_r. Third, the spillover effects between neighboring regions are substantial. The rural area of states with one or more cities tend to benefit from their presence.

A feature that is deeply wired into the theoretical structure of the hedonic "kissing curves" models of Rosen (1974), Sattinger (1975), and Teulings (1995) is that the market return to the human capital index h is a function of the assignment of workers to jobs: the higher the complexity of the occupation o to which a worker with human capital h is assigned in a particular region, the higher the return to human capital in that region. This feature is derived from the first order condition for optimal assignment: the return to human capital for a particular level of human capital must be equal to its marginal productivity in the occupation to which a worker with that human capital is assigned in that region. Hence, if the assignment of h-type workers to o-type occupations

[6] Let $\mathbf{P}$ be an $R \times R$ matrix, where R is the number of regions. If region r and s are neighbors, then the corresponding elements in matrix $\mathbf{P}$ is equal to 1, otherwise it is zero. Then, the matrix is normalized by dividing each row by its row-total.



(a) Average Regional Wage, 1979-2015

(b) Average Human Capital Index, 1979-2015

(c) Average Occupation Index, 1979-2015

Figure 1: Average Regional Differentials in Key Variables I

is the same in two regions, so must be the return to human capital.

In the next section, we present a version of this model where the return $\mathrm{d}w/\mathrm{d}h$ in region r is a decreasing linear function of the difference between a worker's human capital index h and complexity index o of the job to which she is assigned in the market equilibrium in this region

$$\frac{\mathrm{d}w}{\mathrm{d}h} = 1 - \gamma\left(h - o\right), \tag{12}$$

where γ is a positive parameter. The sign of this parameter is the second order condition for optimal assignment. Taking expectations over all individuals in region r yields

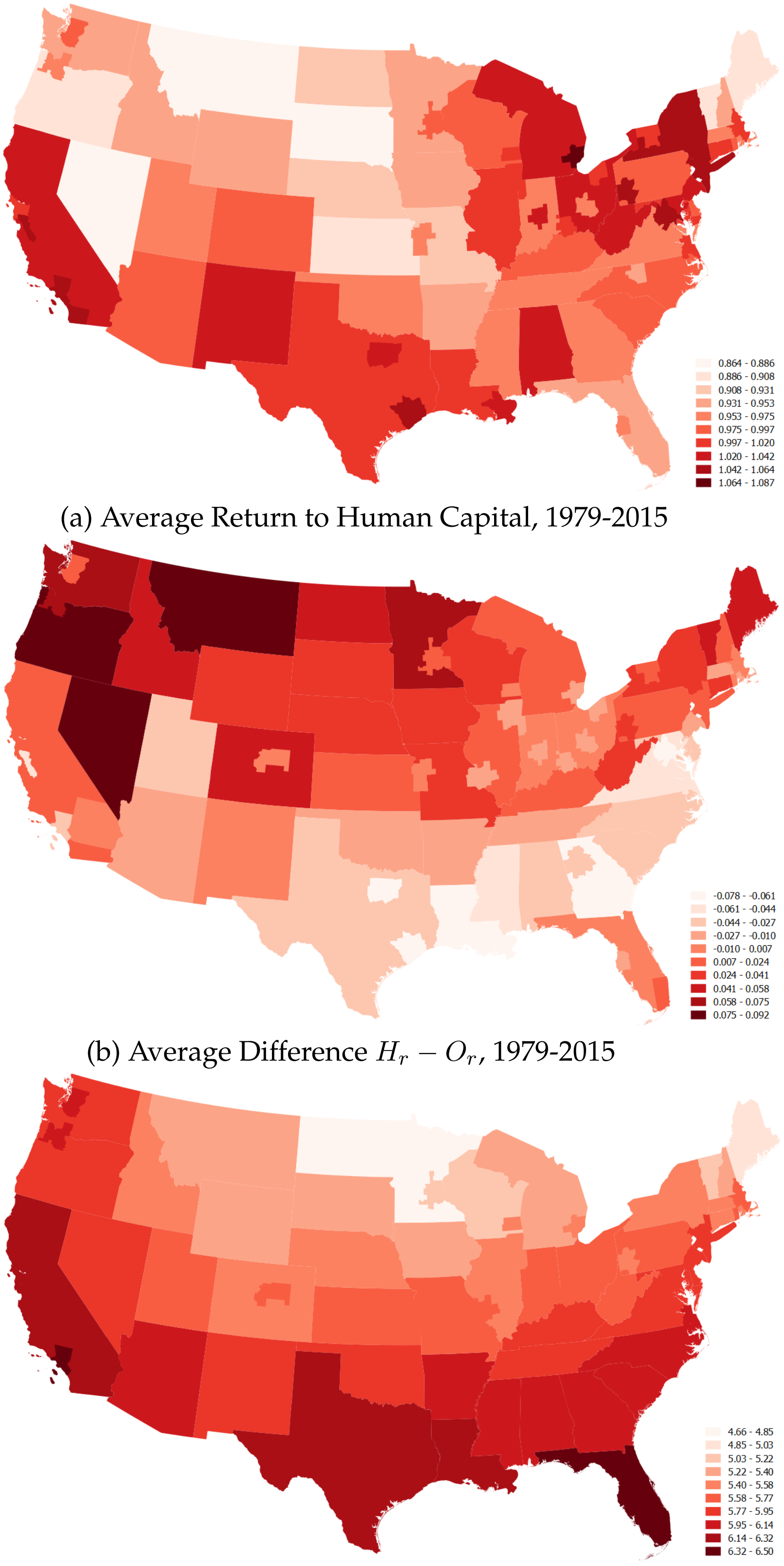
$$\mathrm{E}\left[\frac{\mathrm{d}w}{\mathrm{d}h}|r\right] = 1 - \gamma\left(H_r - O_r\right). \tag{13}$$

Equation (8) reveals that for the assignment of h to o to be constant between two regions, both regions must have the same value for $H_r - O_r$; hence the coefficient for both variables should have the same magnitude, but opposite sign. Note that equation (13) is consistent with the normalizations that we have applied: $\mathrm{E}[\rho_r] = 1$ and $\mathrm{E}[H_r]$ =$\mathrm{E}[O_r] = 0$, see equation (4) and (6).

Equation (13) specifies a negative interregional relation between the return to human capital ρ_r and the net supply of human capital $H_r - O_r$. Table 4 presents estimation results for this equation. In column (1), we enter H_r and O_r separately. Both variable have the right sign. Though the coefficients for O_r and H_r are similar in magnitude, the equality restriction on both coefficients is strongly rejected; see column (2).[7] The estimation results imply $\gamma = 0.60$, which is lower than the inverse elasticity of substitution between high and low skilled workers estimated by Katz and Murphy (1992) and Teulings and Van Rens (2008), see the latter for a discussion. Keeping the mean occupational complexity constant, an increase of the average human capital by 1% reduces the return to human capital by 0.6%.

Equation (13) implies that full interregional equalization of relative wages would require the regional supply and demand for human capital to move in tandem, $H_r = O_r$ for all regions. As documented in Table 2 and Figure 2, supply and demand are strongly correlated, but not perfectly. More importantly, we have document that there is sufficient interregional variation in $H_r - O_r$ to separately identify the effect of both variables in the regressions for both ω_r and ρ_r. This paper addresses the question what explains these interregional differences in $H_r - O_r$ and, by implication, what explains the interregional variation in the return to human capital ρ_r. Column (3) and (4) of Table 3 provide some first provisional answers. We regress $H_r - O_r$ and ρ_r on a city dummy, log city size, and the mean occupational complexity in the region. As the theory predicts, the coefficients in column (3) and (4) have opposite sign, see equation (13): a factor that raises $H_r - O_r$ reduces ρ_r. However, the coefficients are weakly significant at best. Cities and

[7] The $F\left(2, 79\right)$ test statistic is 29.3.



(a) Average Return to Human Capital, 1979-2015

(b) Average Difference $H_r - O_r$, 1979-2015

(c) Average ln January Temperature, 1979-2015

Figure 2: Average Regional Differentials in Key Variables II

in particular large cities have a higher return to human capital than other regions. The same applies to regions with an high mean occupational complexity. This implies that an increase in the demand for human capital O_r is not entirely offset by a corresponding increase in its supply H_r. Obviously, the regression results in column (3) and (4) should be interpreted with caution, as O_r is endogenous. For $H_r - O_r$, the negative sign can just be an artifact of regression to the mean.

Table 4: Return to Education

	(1)	(2)	(3)	(4)
VARIABLES	ρ_r	ρ_r	ρ_r	$H_r - O_r$
H_r	-0.478			
	(-4.47)			
O_r	0.733		0.191	-0.150
	(6.78)		(1.87)	(-1.57)
$H_r - O_r$		-0.601		
		(-5.00)		
City Dummy			-0.216	0.0655
			(-2.24)	(0.48)
City x ln Population			0.0171	-0.00530
			(2.57)	(-0.56)
Constant	0.986	0.986	0.974	0.00470
	(223.89)	(211.85)	(135.13)	(0.79)
R-squared	0.340	0.248	0.267	0.113
R-MSE	0.0395	0.0419	0.0419	0.0382

Note: Columns (1)-(4) present the determinates of the return to human capital using OLS regression. Dependent variable is the average returns to human capital w'_r. Human Capital index measures the average human capital in a region. Occupation index measures the occupation complexity of local labour market. Detailed definitions in section 2. Robust t-statistics in parentheses.

These remarks on the regression results in Table 3 can be extended to all results presented in Table 2. They have to be taken with a grain of salt, since both H_r and O_r are endogenous. They only provide a point of reference for the model to be developed in next section. Four conclusions are relevant from this perspective. First, our results suggest that there are positive agglomeration externalities in the upper tier of the labour market, for workers with either high human capital or complex jobs. Second, these agglomeration benefits go hand in hand with the location of high end activities in cities. Third, the evidence suggests that these externalities are more likely to be related to the regional mean of occupational complexity than of human capital. Finally, we have documented that there is substantial interregional variation in the return to human capital.

2.6 Agglomeration and house prices

In a world where interregional labour mobility is free, house prices differentials are the prime reason why wage differentials can persist. Table 5 presents the estimation results of two regressions. Column (1) reports the results when regressing the log housing value regional characteristics. The higher temperature leads to a higher housing value. A similar effect shows in human capital but is comparatively stronger. City have higher housing price which is positively correlated with the size of city. Larger cities have even higher price. Column (2) reports a regression without regional temperature and the city effect is persistent. The occupational structure has a positive but less significant effect.

Table 5: Housing Value

	(1)	(2)
VARIABLES	avg. ln Housing Value	
ln Jan Temp	0.304	
	(2.96)	
ln City Population	0.221	0.245
	(2.77)	(2.89)
Metro Dummy	-3.287	-3.550
	(-2.97)	(-3.03)
Occ Index	0.965	1.966
	(0.86)	(1.66)
HC Index	2.813	1.269
	(3.22)	(1.74)
Constant	9.882	11.60
	(17.22)	(283.39)
Observations	81	81
R-squared	0.456	0.377
R-MSE	0.254	0.270

Note: Columns (1)-(2) present the determinates of the log housing value using OLS regression. Dependent variable is the average returns to human capital w_r'. Human Capital index measures the average human capital in a region. Occupation index measures the occupation complexity of local labour market. Detailed definitions in section 2. Robust t-statistics in parentheses.

The next section will provide a formal model that brings together the various pieces that have been discussed till sofar, based on the assumption that agglomeration externalities are driven by mean occupational complexity. The structure of the model will suggest a number of instruments, which allows us to estimate the model and to test this assumption.

3 Spatial Equilibrium Model

3.1 General structure

We consider an economy consisting of regions indexed r. For simplicity, there is no physical capital in this economy. All land rents are earned by a class of absentee landlords. Each region has three exogenous characteristics: its mean occupational rank O_r, an exogenous consumption amenity, the January temperature T_r, and whether the region is organized as an urban or a rural area. Workers are endowed with a level of human capital h. Each worker supplies one unit of labour. Her wage is her only source of income. In equilibrium, the log wage $w_r(h)$ of a worker with human capital h living in region r is a linear function of her human capital,

$$w_r(h) = \omega_r + \rho_r h. \tag{14}$$

The coefficients ω_r and ρ_r differ between regions and are endogenously determined. We assume that there is perfect competition on all labour, product, and land markets. The model consist of four building blocks:

1. workers' utility function: costless interregional labour mobility sets the utility of a worker with human capital h equal to some exogenous benchmark for that level of human capital in all regions.

2. regional labour markets: workers with human capital h are assigned to jobs in occupation o; regional relative wages adjust to make this assignment a profit maximizing choice of firms, depending on the regional supply and demand of human capital, H_r and O_r. Free entry of firms drives their profits down to zero.

3. agglomeration externalities: depending on the regional mean level of occupational complexity, on the average lot size chosen by the workers in the region, and on spatial form of a region (urban versus rural), a region benefits from agglomeration externalities.

4. regional housing markets: workers choose the lot size of their house as to maximize their utility. Regional log land prices p_r adjust to clear the land market. The sum of the log land price and the log average lot size yields the log average regional house price v_r. Competition between regions drives land prices down to the point where workers are indifferent between regions.

Each of these four blocks will be discussed in the next subsections. The equilibrium to this economy can be described by relations that are largely linear in h and in the exogenous aggregate variables O_r and T_r and in the endogenous aggregate variables H_r, ω_r, ρ_r, and p_r. At some points, we apply a first order Taylor expansion of the equilibrium; details are discussed in the Appendix. Like in Section 2, the nation wide means of these aggregate variables are normalized to zero

$$\mathrm{E}\left[O_r\right] = \mathrm{E}\left[H_r\right] = \mathrm{E}\left[\omega_r\right] = \mathrm{E}\left[\rho_r - 1\right] = \mathrm{E}\left[p_r\right] = \mathrm{E}\left[v_r\right] = \mathrm{E}\left[T_r\right] = 0. \tag{15}$$

Since their variances are small, this justifies the application of a first order approximation; see the Appendix for details.

3.2 Workers' utility

Workers choose in which region r to live and work at the beginning of their career. They do so as to maximize their utility. Regional mobility at the beginning of the career is costless. Worker mobility will therefore equalize the utility of each h-type worker across all regions. In equilibrium, this utility is equal to some nation-wide exogenous benchmark $u\left(h\right)$. This benchmark utility depends on the human capital of a worker since her human capital enhances her earning capacity. Without loss of generality, this outside benchmark is normalized to the human capital index h,

$$u\left(h\right) = h. \tag{16}$$

Workers derive utility from the private consumption of tradeables and non-tradeables and from the availability of amenities/public goods. Tradeables are traded across regions at a constant nation-wide log price p, which is normalized to zero without loss of generality, $p = 0$. The non-tradeable consumption good is land, either directly, land that is used for residential purposes, or indirectly, land that is used e.g. for shopping malls, where the price of the merchandise reflects cost differentials due to variation in the price of land, or the land that is used by workers providing non-tradeable services and who get compensated for the higher land prices by higher wages.

The private benefits of amenities cannot be priced. We consider two type of amenities: the January temperature T_r and the regional fixed effect in wages ω_r. People prefer to live in regions where January temperature is more agreeable, see e.g. Glaeser (2009). Ahlfeldt, Redding, Sturm and Wolf (2015) show in their study of Berlin that there are strong and highly localized agglomeration externalities in residential areas. A high local density allows a dense network of services like shops, restaurants, and cultural performances to be sustained. For the sake of convenience, we apply a reduce form model for this agglomeration externality on the consumption side by assuming that this externality is proportional to the applomeration externality on the production side ω_r.

The equilibrium condition for market for interregional migration can be derived most easily in the form of a cost function. We allow the utility function of workers to be non-homothetic, so that land can be a normal good. We use the expenditure function proposed by Hanoch (1975) and Comin et al. (2015). As we show in the appendix, a first order Taylor expansion of the log of this expenditure function reads

$$\underbrace{\omega_r + \rho_r h}_{\text{income = cost}} = \underbrace{h}_{\text{benchmark u}} + \underbrace{(1 - \varepsilon h)\,\lambda p_r}_{\text{price index}} - \underbrace{(\alpha - h\xi)'\,x_r}_{\text{public goods}}, \tag{17}$$

where $x_r \equiv [T_r, \omega_r]'$ is the vector of public goods, where p_r is the price of land in region r, and where λ is the average land share in expenditure. Equation (17) holds exactly when utility function is homothetic and takes the Cobb Douglas form ($\varepsilon = 0, \eta = 1$), where the land share λ does not depend on either the price of land p_r or the benchmark level of utility h; when the elasticity of substitution η between land and other consumption is different from unity, deviations of the land share from the nation wide mean have only a second order effect on the cost function. Hence, this effect drops out in this first order approximation of the cost function.

The left hand side is log income (= cost from the perspective of a cost function) of obtaining a utility level h. The first term on the right hand side is the benchmark utility level. The second term is the price index in region r. The log price of tradeables is normalized to zero. Hence, it drops out. The log price of land enters the equation pre-multiplied by the average land share in the economy, λ. We apply a non-homothetic utility function, where the income elasticity of the demand for land is less then one; ε is one minus the income elasticity of land. If the utility function were homothetic, $\varepsilon = 0$, the land share is independent of the benchmark utility level h; the term $\lambda\varepsilon h p_r$ drops out in that case. When $\varepsilon > 0$, land is a normal good in consumption; the price index is less sensitive to price of land for high human capital workers in that case. Albouy, Ehrlich and Liu (2016) provide evidence that land is indeed a normal good.

The final term measures the compensating differentials for regional amenities. Other things equal, regions with high amenities will have lower cost. Like the term for the price of land, this term is comprised of two sub-terms. The parameter vector $\alpha \equiv [\alpha_T, \alpha_\omega]'$ measures the compensating differential for one unit of the amenity as a share of disposable income. The parameter vector $\xi \equiv [\xi_T, \xi_\omega]'$ measures the income effect. If the utility function were homothetic in the benefits of amenities, the compensating differentials would be independent of the benchmark utility level h; hence, $\xi = 0$. For $\xi > 0$, the amenities are normal goods.

Equation (17) must hold identically for all h. This yields two conditions

$$\begin{aligned} \omega_r &= \lambda p_r - \alpha' x_r, \\ \rho_r &= 1 - \varepsilon \lambda p_r + \xi' x_r. \end{aligned} \tag{18}$$

The first condition states the level of log wages in region r must compensate for the level of log land prices multiplied by the value share of housing and for the availability of amenities. When land and amenities are normal goods ($\varepsilon > 0$, $\xi > 0$), the second condition states that the return to human capital is lower in regions with high land prices, since workers with a high benchmark level of utility need less compensation for high land prices, since land has a smaller share in their consumption basket, while the return to human capital is higher in regions with a large endowment of public goods, since the compensating differential rises less than proportional to the benchmark utility level. Note that for $p_r = x_r = 0$, ρ_r is equal to unity, which is the slope of the benchmark utility with respect to the worker's human capital, see equation (16).

3.3 Regional labour markets

The production structure is similar to that in Rosen (1974), Sattinger (1975), Teulings (1995), and Teulings and Van Rens (2008). A region produces a tradeable commodity that is sold on the nation wide market for tradeable commodities. This commodity is characterized by the mean level of occupational complexity of the inputs of occupational effort that is required to produce this commodity. For example, the region San Jose produces new IT applications. The mean level occupational complexity needed to produce these applications is very high. For another example, Oregon is mainly engaged in forestry. Lumberjacks is a relatively simple occupation. Regions need some mixture of occupations to produce their commodity, but the mean of the distribution of occupational complexity differs; the mean is assumed to be a sufficient statistic for the characterization of the interregional differences. The consumption of tradeables in each region consist of an appropriate mix of the commodities produced in all regions. This mix is the same across regions.

Firms producing the occupational input for the production of this regional composite commodity operate a constant returns to scale technology which is common to all regions. Let $y(h,o)$ be the log output of a worker with human capital h in occupation o. We make two assumptions:

1. $y_h(h,o) > 0$: high human capital workers are more productive in any occupation; hence wages will be increasing in human capital in any market equilibrium.

2. $y_{ho}(h,o) > 0$: high human capital workers have a Ricardian comparative advantage in complex jobs (log supermodularity); as in equation (7), within each region r, high human capital workers are assigned to more complex jobs.

These assumptions are the same as those for the strawman model in Section 2.1. For $\gamma(h-o) < 1$, the following specification of $y(h,o)$ satisfies these assumptions

$$y(h,o) = y_0 + h - o - \frac{1}{2}\gamma(h-o)^2. \tag{19}$$

Firms in region r offering jobs of occupational complexity o hire workers of human capital type h in order to minimize their cost of production subject to the regional wage function $w_r(h)$. Hence, the optimal level of human capital $h_r(o)$ satisfies

$$h_r(o) = \arg\min_h \left[w_r(h) - y(h,o)\right].$$

Substitution of equation (14) for $w_r(h)$ and equation (19) for $y(h,o)$ and solving the problem yields

$$w_r'(h) = \rho_r = 1 - \gamma\left[h_r(o) - o\right]. \tag{20}$$

The linearity of $w_r(h)$, and hence the return to human capital being the same for all levels of human capital in region r implies that $h_r(o) - o$ must be a constant. Taking expectations within region r yields

$$\rho_r = 1 - \gamma\left(H_r - O_r\right). \tag{21}$$

which is identical to equation (13): an excess supply of human capital reduces its return. The parameter γ measures the curvature of the output function $y(h,o)$; its sign is the second order condition of the cost minimization problem. We estimated $\gamma = 0.60$, see Table 3. Though the market assignment $h_r(o)$ is increasing within region r, this relation is not necessarily increasing across regions. When $H_r - O_r$ is higher than average in a region, workers in that region are assigned to less complex jobs. Hence, the return to their human capital will be lower. Since $h_r(o) - o$ is a constant, the optimal assignment must satisfy

$$h_r(o) = o - O_r + H_r, \tag{22}$$

which is identical to equation (8). As we documented in Section 2.4, this relation is based on the Equal Return Assumption, which is roughly consistent with the data. It is interesting to consider what would happen if this assumption were not satisfied. Then, $h_r(o) - o$ would vary by o by equation (7). Hence, $w_r'(h)$ would not be constant within a region and $w_r(h)$ would therefore be non-linear. Consider the case that $h_r(o) - o$ is decreasing in r. Then, the variance is larger for occupational complexity than for human capital: there are many complex and simple jobs, but few intermediates.[8] Hence, $w_r'(h)$ is increasing in h; see equation (20). This is the polarization phenomenon discussed by Autor and Dorn (2013). Polarization is therefore consistent with the Extended Single Index Model. Polarization is ruled out only when we introduce the Equal Return Assumption. Autor and Dorn (2013) provide a more complicated model of polarization, which involves three rather than one human capital index. Occam's razor suggests that one should use the more parsimonious model. Our empirical evidence shows that polarization does not play an

[8] Equation (7) implies

$$\text{Var}\left[o_i|r\right] = \left(\frac{\rho_r}{\chi_{or}}\right)^2 \text{Var}\left[h_i|r\right].$$

Hence, $\rho_r > \chi_{or}$ implies Var$[o_i|r]$ >Var$[h_i|r]$.

important role in our cross-section of regions. This is likely to be different when evaluating the evolution of $w_r(h)$ over time, as Autor and Dorn have shown.

Substitution of equation (21) in equation (18) yields

$$1 - \rho_r = \gamma(H_r - O_r) = \varepsilon\omega_r + (\varepsilon\alpha - \xi)' x_r. \tag{23}$$

Since the return to human capital and its net supply move in opposite direction, they are driven by the same variables, which should enter both equations in the same proportions and with opposite sign. These implications can be tested.

The return to human capital is decreasing in the fixed effect, ω_r. A higher fixed effect leads to higher land prices, which are less harmful for high human capital workers since land is a normal good. The same reasoning applies for compensating differentials for regional amenities, $\alpha' x_r$; again, amenities increase land prices, which is less detrimental for high human capital workers. The term $\xi' x_r$ measures the income effect of amenities. When amenities are normal goods, $\xi > 0$, they are relatively less attractive for higher income types. Their presence requires therefore an increase in the return to education.

As we have documented in Section 2.4, the return to human capital is higher in urban than in rural regions. Cities have higher house prices, which *other things equal* would lead to the opposite result: a lower return to human capital. The only force that yields a higher return to human capital in cities is the income effect of amenities, $\xi' x_r$. Hence, amenities in cities must be normal goods, which runs counter to the evidence by Diamond (2016).

3.4 Regional land markets

A first order Taylor expansion of the log demand for land $l_r(h)$ of an individual with benchmark utility h in region r yields

$$l_r(h) = \ln\lambda + \overline{\varepsilon} h - \eta\overline{\lambda} p_r - \alpha' x_r, \tag{24}$$

see the Appendix, where η is the price-elasticity of land, and where a bar on top of a parameter denotes its complement with respect to one, so: $\overline{\varepsilon} = 1 - \varepsilon$. For $p_r = x_r = h = 0$, the log demand for land is equal to the land share in total expenditure $\ln\lambda$ (since $p_r = 0$ implies that the price of land is equal to unity). The demand for land depends negatively on regional land prices (due to the substitution effect) and on amenities (due to the negative income effect of the compensating differential for these amenities). Since land is a normal good, the demand for land increases less than proportional to the benchmark utility.

For the evaluation of our data on house prices, we use the log price of a plot of land for an average worker type H_r in region r

$$v_r \equiv l_r\left(H_r\right) + p_r = \ln\lambda + \overline{\varepsilon} H_r + \left(1 - \overline{\lambda}\eta\right) p_r - \alpha' x_r.$$

Regional amenities push house prices up. For the Cobb Douglas case ($\varepsilon = 0, \eta = 1$), the value of a house is proportional to the log income of the worker $H_r + \lambda p_r$ (the term λp_r is price compensation in wages for house prices, see equation (18))

$$v_r = \ln\lambda + H_r + \lambda p_r - \alpha' x_r.$$

3.5 Agglomeration externalities

Intraregional agglomeration externalities or knowledge spillovers are modelled similar to Gennaioli, La Porta, Silanes and Shleifer (2013). Where Gennaioli et.al (2013) assume that knowledge spillovers depend on supply of human capital, we assume these externalities depend on its demand. Hence, rather than the mean of human capital, H_r, we enter on the mean of occupational complexity, O_r. Hence, our specification reads

$$\omega_r = \psi\left(n_r + \theta O_r\right), \tag{25}$$

where n_r is the number of workers that contribute to the externalities, and where ψ and θ are weakly positive parameters. For $\psi = 0$, there are no knowledge spillovers. For $\theta = 0$, knowledge spillovers depend only on the number of workers, not on the occupational structure. Due to our normalization of the average return to human capital to unity, $\theta = 1$ would imply that knowledge spillovers are proportional to the total wage bill. Gennaioli et.al. (2013) reported evidence that knowledge spillovers increase more than proportional to the average level of human capital of the regional workforce. This is the case if $\theta > 1$.

Our modeling of the intraregional spatial structure combines ideas from Lucas and Rossi-Hansberg (2002), Rossi-Hansberg and Wright (2007), and Ahlfeldt, Redding, Sturm and Wolf (2015). In the cities considered in Lucas and Rossi-Hansberg and Ahlfeldt et.al. workers have to commute between their home and job location and ideas have to travel between the locations of different jobs. We consider two archetypical spatial structures: rural areas and cities. In a rural area, people work at their home location. Hence, workers do not commute and ideas have to travel. In a city, it is exactly the opposite: jobs are concentrated in a Central Business District (CBD). Hence, ideas don't travel, but then workers have to commute. We discuss both archetypes below.

3.5.1 Rural areas

In rural areas, workers work at the same location as they live and all h-type workers are spread homogeneously across space. Like Lucas and Rossi Hansberg (2002) and Ahlfeldt et.al. (2015), the travel of knowledge spills across space comes at a cost: at distance s, only a fraction $1-\delta s$ of the spillover survives. The maximum distance knowledge spillovers can travel is therefore δ^{-1}. Only workers working within a distance δ^{-1} contribute to the knowledge spillover for a particular worker. Hence, the knowledge spillover ω_r^r in region r (the superfix r denotes rural areas) reads

$$\begin{aligned} \omega_r^r &= \psi \ln\left(\int_0^{\delta^{-1}} 2\pi s\left(1-\delta s\right) e^{\theta O_r - l_r} \mathrm{d}s\right) \\ &= \psi\left(\ln\frac{\pi}{3} - 2\ln\delta + \theta O_r - l_r\right), \end{aligned} \tag{26}$$

see equation (25). In the first line, $2\pi s$ is the circumference of the circle at distance s of the own location, $1-\delta s$ is the fraction of the spillovers that survives at that distance, and e^{-l_r} is the population density. Substitution of equation (24) for l_r, see Appendix, yields

$$\begin{aligned} \omega_r^r &= \Psi^r\left[\psi_0 + \left(\theta - \overline{\varepsilon}\right) O_r + \psi_T T_r\right], \\ \Psi^r &\equiv \frac{\psi}{1-\psi\cdot\psi_\omega}, \end{aligned} \tag{27}$$

The parameter Ψ^r has to be positive for a bounded solution to exist. Hence

$$1-\psi\cdot\psi_\omega < 1 \Rightarrow \Psi^r > \psi. \tag{28}$$

We assume this condition to hold. Hence, the elasticity of the knowledge spillover ω_r^r with respect to occupational complexity is $\Psi^r\left(\theta-\overline{\varepsilon}\right)$: a higher complexity yields more spillovers, which raises house prices and therefore reduces land use, which allows a further increase in spillovers by a higher population density. Our composite parameter $\Psi^r\left(\theta-\overline{\varepsilon}\right)$ corresponds to the parameter θ reported by Gennaioli et.al. (2013). Gennaioli et.al. (2013) report a value of $\theta = 6$. January temperature increases knowledge spillovers, since it raises house prices and therefore increases the population density.

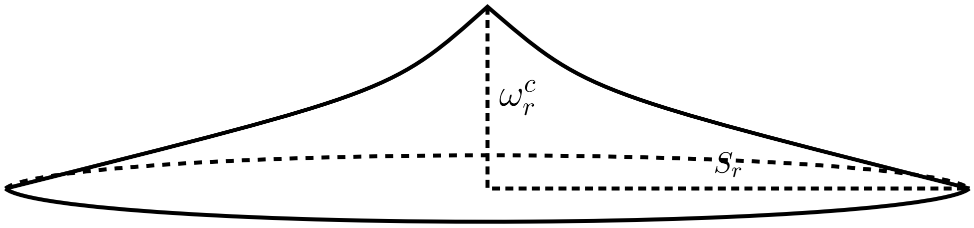
For $\varepsilon = 0$ (homothetic utility) and $\alpha = \xi = 0$ (no amenities), ψ_ω is equal to $\frac{\overline{\lambda}}{\lambda}\eta$. Empirical estimates of ψ vary between 0.05 and 0.20 (for the latter, see Ahlfeldt et.al. 2015), depending on the precise context.[9] With a land share λ of about $0.15-0.20$ and a Cobb Douglas utility function ($\eta = 1$), condition (28) is critical. If land and other consumption were more easily substitutable, $\eta > 1$, agglomeration benefits would be unbounded since workers would reduce their land use

[9] When ψ is analyzed at the level of the city, its value is much lower than in studies like Ahlfeldt et.al., who evaluate the spill-over a one point in space and where contributions of workers at some distance are discounted at a rate δ.

in response to the high price for land, which enables them to co-locate in ever higher densities. Teulings, Ossokina, and Groot (2018) and Albouy, Ehrlich and Liu (2016) report evidence that the elasticity of substitution is less than unity.

3.5.2 Cities

Like Lucas and Rossi-Hansberg (2002) and Rossi-Hansberg and Wright (2007), cities are assumed to have a circular shape. Like Rossi-Hansberg and Wright (2007), all employment is localized in the CBD at the city-center, which does not use any land at all; all jobs are therefore concentrated at a single point in space and hence, there is no loss in the transmission of ideas. Workers live in the area around the CBD. The cost of commuting to the CBD is a share κ per unit of distance. Hence, somebody living at a distance s from the CBD works only a fraction $1 - \kappa s$ of her time. The worker's output and the knowledge spillovers she creates are similarly affected. Let S_r denote the edge of the city (the maximum distance from the CBD). By construction, $S_r < \kappa^{-1}$: commuting beyond that distance is useless as it leaves no time for working.



In Lucas and Rossi-Hansberg (2002) and Ahlfeldt et.al. (2015), land use varies within a city, depending on the price of land at a each location within the city. This is a more difficult structure than we can handle in our setting with heterogeneous workers. We therefore introduce a city council that equalizes the private cost of commuting as a share of the wage rate $w_r(h)$ of the inhabitants living within city boundary S_r by imposing a balance budget system of taxes and subsidies. Inhabitants living close to the CBD pay a tax, while those living at the edge of the city receive a subsidy. These taxes and subsidies balance all commuting cost differentials within the city. The merit of this assumption is that it makes each location equally attractive, irrespective of the worker's human capital. Hence, land prices and the consumption of land are flat within the city.

The city council sets the boundary S_r such that a worker living just outside the city is indifferent between commuting to the CBD to benefit from agglomeration externalities and working at his home location. Hence, ω_r^c (the superfix c denotes cities) satisfies

$$\omega_r^c = -\ln(1 - \kappa S_r) \tag{29}$$

Where people commute in cities, so that ideas don't have to travel and hence, δ is irrelevant, ideas travel in rural areas, so that people don't have to commute and hence, κ is irrelevant. Define Δ to be

$$\Delta \equiv \ln \frac{\delta}{\kappa} > 0. \tag{30}$$

Ahlfeldt et.al. (2015) report $\delta > \kappa$ and hence $\Delta > 0$, see their Table V.

Cities can only be an efficient spatial organization if the cost of commuting is smaller than the spatial decay of knowledge spillovers. If not, it is cheaper to let ideas rather than workers travel, since commuting reduces the time available for knowledge sharing *and* production, while the travel of idea only reduces the former. Hence, if $\kappa > \delta$ (or: $\Delta < 0$), it is always more efficient to let ideas travel and not people.[10] In rural areas, the radius of the area around a location that contributes to the agglomeration benefits is fixed at δ^{-1}. In an urban area, that radius is determined endogenously, by equation (29). Hence, an increase in e.g. O_r raises not only the spillovers generated within a fixed area. It also increases the area involved.

Log city size is about twice as sensitive to a change in O_r than the agglomeration effect. The reason is that an increase in the agglomeration effect raises the radius of the city, but that increase translates quadratically into the size of the population, since land is a two-dimensional space.

This paper takes the spatial form and the mean occupational complexity of a region as exogenous, as a tribute to our lack of understanding of their causal relationship. Both are highly persistent. However, our model suggests that highly complex activities can be expected to locate in cities in the long run, either because a region transforms into a city when it hosts complex activities, or because complex activities relocate to cities. Table A3 provides some evidence on this. Regions are ranked by their mean occupational complexity. Urban regions rank high, rural regions rank low. The separation is almost perfect. Whatever the direction of causation, the long run outcome fits our prediction. The spill-over effect in cities ω_r^c and their log population size n_r satisfy

$$\begin{aligned} \omega_r^c &= G^{-1}(2\Psi^r \Delta + \omega_r^r) \simeq \Psi^c \left[\psi_0 + 2\Delta + (\theta - \overline{\varepsilon})\, O_r + \psi_T T_r\right] \\ n_r &= H\left(\omega_r\right), \end{aligned} \tag{31}$$

where $G(\cdot)$ is a function with $G'(\cdot) > 1$. The derivation of these relation is relegated to the Appendix. The spill-over effect is more sensitive to O_r and T_r in cities than in rural areas. The intuition is that the area that contributes to knowledge spill-overs is fixed at δ^{-1} in rural areas, while S_r varies with ω_r^c in cities; see equation (29).

[10] This can be see immediately by realizing that equation (29) implies that $S_r < \kappa^{-1}$. If $\kappa > \delta$, then the radius of area for which a rural region extracts knowledge spill-overs exceeds that for which a city can do.

3.6 Overview of the equilibrium

An equilibrium to this multi-region economy is a set of supplies of human capital H_r, fixed effects in the wage function ω_r, returns to capital ρ_r, log house prices v_r (for all regions), and the log population size n_r (for cities), conditional on their occupational structure O_r, January temperature T_r, and their spatial organization (urban versus rural). These variables must satisfy equations (18), (23), (35), (38) and (31). These equations can solved for the endogenous variables.

$$\begin{aligned}
1-\rho_r &= \varepsilon_\omega \omega_r + \varepsilon_T T_r, \\
H_r - O_r &= \gamma^{-1}\left(\varepsilon_\omega \omega_r + \varepsilon_T T_r\right), \\
v_r &= \ln\lambda + \lambda_\omega \omega_r + \lambda_T T_r + \overline{\varepsilon} O_r, \\
\omega_r^r &= \Psi^r\left[\psi_0 + \left(\theta - \overline{\varepsilon}\right) O_r - \psi_T T_r\right], \\
\omega_r^c &= \Psi^c\left[\psi_0 + 2\Delta + \left(\theta - \overline{\varepsilon}\right) O_r - \psi_T T_r\right],
\end{aligned}$$

where

$$\begin{aligned}
\varepsilon_\omega &\equiv \varepsilon + \varepsilon\alpha_\omega - \xi_\omega, \\
\varepsilon_T &\equiv \varepsilon\alpha_T - \xi_T, \\
\lambda_\omega &\equiv 1 + \frac{\overline{\lambda}}{\lambda}\overline{\eta}\left(1+\alpha_\omega\right) + \frac{\overline{\varepsilon}}{\gamma}\varepsilon_\omega, \\
\lambda_T &\equiv \frac{\overline{\lambda}}{\lambda}\overline{\eta}\alpha_T + \frac{\overline{\varepsilon}}{\gamma}\varepsilon_T, \\
\psi_\omega &\equiv \frac{1+\alpha_\omega}{\lambda} - \lambda_\omega = \frac{1-\overline{\lambda}\overline{\eta}}{\lambda}\left(1+\alpha_\omega\right) - \frac{\overline{\varepsilon}}{\gamma}\varepsilon_\omega - 1, \\
\psi_T &\equiv \frac{\alpha_T}{\lambda} - \lambda_T = \frac{1-\overline{\lambda}\overline{\eta}}{\lambda}\alpha_T - \frac{\overline{\varepsilon}}{\gamma}\varepsilon_T.
\end{aligned}$$

These equation are not fully reduced form, since ω_r appears on the right hand side of the first three equations, to facilitate the interpretation of the results.

3.7 Identification and testing

Table 6 provides an overview of the (composite) parameters that can be established from the various equations.

The model yields a number of testable implications. First, the coefficients in the equations for the return to human capital, $1-\rho_r$, and the net supply of human capital, $H_r - O_r$, should have

Table 6: Overview of the (composite) Parameters

(composite) parameters	value	derived parameters	equation/source
η	0.75	$-$	Albouy et al. (2016) Table 1, Teulings et.al.
ε	0.50	$-$	Albouy et al. (2016) Table 1
λ	0.30	$-$	Glaeser et al (2008)
γ	1.00	$-$	
ε_ω	-0.6	ζ_O, ω_O	
ε_T	-0.01	$\omega_T, \varepsilon_\omega, \zeta_T$	
λ_ω	4.8	$v_O, \varepsilon, w_\omega$	
α_ω	6.1	$\lambda, \eta, \lambda_\omega, \varepsilon, \gamma, \varepsilon_\omega$	
λ_T	-0.04	$v_T, \omega_T, \lambda_\omega$	
α_T	-0.06	$\lambda, \eta, \lambda_T, \varepsilon, \gamma, \varepsilon_T$	
ψ_ω	18.8	$\lambda_\omega, \alpha_\omega, \lambda$	
ψ_T	-0.16	$\lambda_T, \alpha_T, \lambda$	
Ψ_c	0.24	ω_T^c, ψ_T	
Ψ_r	0.13	$\omega_O^c, \Psi^c, \omega_O^r$	
θ	3.33	$\omega_O^c, \Psi^c, \varepsilon$	
ψ	0.04	ψ_ω, Ψ^r	

equal sign and should vary proportionally. Second, the impact of O_r and λ_T on ω_r should vary proportionally between rural and urban regions and the impact should be larger for cities. Next, the impact of $\widetilde{\omega}_r$ should be the same for rural and urban regions. The testable implication from the equation on n_r^c will be weak.

4 Empirical analysis

We calculate the value of agglomeration in an open system, which keeps the outside option of workers constant. The regional mean occupational complexity and January temperature are held constant. Since the outside utility level is constant, worker will migrate between the regions and land prices will adjust, such that workers' utility in each region is equal to outside option.

We then calculate the value of land for each region and for the nation as a whole. For this exercise, we have to treat rural and urban regions differently. For rural regions, we keep the total land surface constant. The total land surface is set equal to the number of workers in our sample times the calculated land use in the initial equilibrium. For urban regions, we take a different approach. We calculate the new optimal city size S_r. From there, it is a standard exercise to calculate the total land value of the city. This approach implies that we allow cities to vary in size without accounting for what happens to the land if a city shrinks or where the land comes from when the city grows. The only issue that matters is how far away a plot of land is from the city center. We shall evaluate ex post how much distortion this approach yield by pricing changes in land use of the city according to the rural region in the corresponding state.

Table 7: Bartik IV Regression Results with 34 MSAs

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	O_r	$1-\rho_r$	$H_r - O_r$	v_r	w_r	n_r
Panel A: City and Non-City Sample (81 Observations)						
ln Jan Temp	-0.00585	-0.0284	-0.0539	0.161	0.0424	
	(-1.03)	(-2.36)	(-5.59)	(1.78)	(2.93)	
Bartik IV	1.524					
	(19.41)					
Bartik IV sq.	3.698					
	(3.40)					
Occ IV		-0.390	-0.290	3.110	0.542	
		(-3.64)	(-3.37)	(3.86)	(4.19)	
Metro Dummy	0.0246	-0.00620	0.0128	-0.0943	0.0269	
	(4.63)	(-0.48)	(1.24)	(-0.98)	(1.73)	
Constant	0.0217	0.181	0.305	10.70	-0.321	
	(0.67)	(2.66)	(5.59)	(20.96)	(-3.91)	
R-squared	0.911	0.356	0.394	0.251	0.513	
R-MSE	0.0185	0.0393	0.0315	0.295	0.0474	
Panel B: City Sample (34 Observations)						
ln Jan Temp		0.00400	-0.0343	0.476	0.0392	0.397
		(0.23)	(-2.04)	(3.12)	(1.84)	(1.19)
Occ IV		-0.185	-0.195	5.016	0.684	2.311
		(-1.43)	(-1.60)	(4.54)	(4.41)	(0.96)
Constant		-0.0261	0.198	8.659	-0.282	11.87
		(-0.25)	(1.98)	(9.55)	(-2.22)	(5.99)
R-squared		0.068	0.160	0.463	0.403	0.062
R-MSE		0.0349	0.0330	0.300	0.0420	0.655
Panel C: Non-city Sample (47 Observations)						
ln Jan Temp		-0.0452	-0.0644	-0.00426	0.0456	
		(-2.95)	(-5.60)	(-0.04)	(2.33)	
Occ IV		-0.608	-0.386	1.107	0.360	
		(-3.73)	(-3.15)	(1.05)	(1.72)	
Constant		0.269	0.361	11.57	-0.345	
		(3.09)	(5.53)	(20.63)	(-3.10)	
R-squared		0.335	0.480	0.025	0.158	
R-MSE		0.0399	0.0300	0.257	0.0510	

Note: Columns (1)-(6) present the Bartik IV regression results with 34 Cities and 47 non-city areas. Dependent variables are the average occupation index, local average log wage, housing value normalised by land share and the local wage level, returns to human capital, difference between average human capital index and the occupation index, and the log local population. Log Jan Temp is the average log local January temperature. Bartik IV and Bartik IV sq are defined in section 5. Occ IV is the estimated average occupation index using IV correction in first stage. Metro Dummy equals one if an observation is city area, zero if it is non-city area. Detailed definitions of equations are in section 2. Robust t-statistics in parentheses.

The equilibrium without agglomeration can be conveniently summarized as follows:

$$\begin{aligned}
1-\rho_r &= \left(\varepsilon\alpha_T-\xi_T\right)T_r,\\
H_r-O_r &= \gamma^{-1}\left(\varepsilon\alpha_T-\xi_T\right)T_r,\\
p_r &= \lambda^{-1}\alpha_T T_r,\\
v_r &= \ln\lambda+\frac{\overline{\lambda}}{\lambda}\eta\alpha_T T_r,\\
l_r &= \ln\lambda+\frac{1}{2}\sigma^2\overline{\varepsilon}^2+\overline{\varepsilon}O_r-\lambda_T T_r\\
\omega_r &= 0,\\
n_r^c &= \psi_0+2\Delta-\overline{\varepsilon}O_r+\lambda_T T_r.
\end{aligned}$$

The equilibrium is inefficient, since workers do not take into account the benefit to other workers of them moving to the city. We can calculate that welfare improvement from the equilibrium allocation to the optimal one.

5 Conclusion

This paper estimates the knowledge spillover effect, using within country variation with US CPS data. The first founding is human capital spillover effect is strong, and cannot fully addressed by unobservable abilities. Agglomeration effect is significant, with faster development in cities and more population flow into cities. We add two important element into the standard model, first is the regional occupation structures, reflecting the complexity of tasks of different regions, and the second is the different knowledge spillover channel in city and non-city regions. Hence, we considered three distance components, the decay of knowledge spillover in distance, the commuting cost and the transportation cost. All these distance concepts play key roles in considering the spatial equilibrium.

Our conclusion are education spillover effect is strong but also the occupation structure. Take different transmission channel into consideration, we estimate the loss of welfare, measured by housing wealth, is at least 15 percent is not higher. Using the spatial equilibrium model, we also point out that the loss of welfare may partly compensated by the less crowed effect. Hence, the amenity increase can be a reason that we see less welfare loss.

The shortcoming of the paper is that the empirical part relies heavily on the city dummy, which make the explanatory power less strong. Definition of regions also matters. In some of our sample, non-city and city region have very small difference, in term of land area and population.

Appendix

Price index and the demand for land

Let $E\left(P, P_r, U\right)$ be expenditure as a function of the prices P and P_r and the utility level U. Following Hanoch (1975) and Comin et al. (2015), the expenditure function can be defined as

$$E\left(P, P_r, U\right) = \left(\overline{\lambda} P^{\overline{\eta}} + \lambda U^{-\varepsilon/\overline{\lambda}} P_r^{\overline{\eta}}\right)^{\overline{\eta}^{-1}} U^{1+\lambda\varepsilon/\left(\overline{\lambda}\overline{\eta}\right)},$$

where a bar on top of a variable denotes its complement with respect to 1, hence: $\overline{\lambda} = 1 - \lambda$. Define $e\left(p, p_r, h\right) \equiv \ln E\left(e^p, e^{p_r}, e^h\right), z \equiv \overline{\eta} p_r - \varepsilon h/\overline{\lambda}$, and $U \equiv e^h$. Taking logs and using $\ln\left(\overline{\lambda} + \lambda e^z\right) = \lambda z + \frac{1}{2}\lambda\overline{\lambda} z^2 + O\left(z^3\right)$ and $P = 1$ yields

$$\begin{aligned} e\left(0, p_r, h\right) &= \overline{\eta}^{-1} \ln\left(\overline{\lambda} + \lambda e^z\right) + \left[1 + \lambda\varepsilon/\left(\overline{\lambda}\overline{\eta}\right)\right] h \\ &= \overline{\eta}^{-1}\left(\lambda z + \frac{1}{2}\lambda\overline{\lambda} z^2\right) + \left[1 + \lambda\varepsilon/\left(\overline{\lambda}\overline{\eta}\right)\right] h + O\left(z^2\right) \\ &= \left(1 - \varepsilon h\right)\lambda p_r + h + O\left(h^2\right) + O\left(p_r^2\right), \end{aligned} \tag{32}$$

where $e\left(p, p_r, h\right) \equiv \ln E\left(e^p, e^{p_r}, e^h\right)$, which proves equation (17). Differentiating the first line in equation (32) with respect to p_r yields

$$\begin{aligned} \ln e_2\left(0, p_r, h\right) &= \ln\lambda + z - \ln\left(\overline{\lambda} + \lambda e^z\right) = \ln\lambda + \overline{\lambda} z + O\left(z^2\right) \\ &= \ln\lambda - \varepsilon h + \overline{\lambda}\overline{\eta} p_r + O\left(z^2\right). \end{aligned} \tag{33}$$

By Shephard's Lemma, the partial derivative with respect to P_r is the demand land L_r

$$E_2\left(P, P_r, U\right) = L_r.$$

By its definition $e_2\left(p, p_r, h\right)$ satisfies

$$\begin{aligned} e_2\left(0, p_r, h\right) &= \frac{E_2\left(1, e^{p_r}, e^h\right)}{E\left(1, e^{p_r}, e^h\right)} e^{p_r} = \frac{L_r P_r}{E}, \\ \ln e_2\left(0, p_r, h\right) &= l_r\left(h\right) + p_r - w_r\left(h\right), \end{aligned} \tag{34}$$

where we use $e\left(p, p_r, h\right) = w_r\left(h\right)$ in the final line. Substitution of this expression in equation (33) yields

$$l_r\left(h\right) = \ln\lambda + \overline{\varepsilon} h - \left(1 - \overline{\lambda}\overline{\eta}\right) p_r + w_r\left(h\right) + O\left(z^2\right)$$

using equation (17) for $w_r(h)$ and rearranging terms yields equation (24). Substitution of equation

(18) and (23) yields the average log value of a house in region r[11]

$$
\begin{aligned}
v_r &= l_r\left(H_r\right) + p_r = \ln\lambda + \overline{\varepsilon} H_r + (1-\eta\overline{\lambda})p_r - \alpha' x_r \\
&= \ln\lambda + \lambda_\omega \omega_r + \lambda_T T_r + \overline{\varepsilon} O_r, \\
\lambda_\omega &\equiv 1 + \frac{\overline{\lambda}}{\lambda}\overline{\eta}\left(1+\alpha_\omega\right) + \frac{\overline{\varepsilon}}{\gamma}\varepsilon_\omega, \\
\lambda_T &\equiv \frac{\overline{\lambda}}{\lambda}\overline{\eta}\alpha_T + \frac{\overline{\varepsilon}}{\gamma}\varepsilon_T, \\
\varepsilon_\omega &\equiv \varepsilon + \varepsilon\alpha_\omega - \xi_\omega, \\
\varepsilon_T &\equiv \varepsilon\alpha_T - \xi_T
\end{aligned} \tag{35}
$$

(36)

Regional labour markets

A region produces a composite commodity, characterized by the mean occupational complexity O_r. This composite commodity is produced by a CRS Leontieff technology with requires as inputs the output of each occupation o in fixed proportions. The required input of occupation o for one unit of the composite commodity is characterized by the normal density function with mean O_r and standard deviation σ

$$
\text{input of occupation } o = \phi\left(\frac{o-O_r}{\sigma}\right), \tag{37}
$$

where $\phi(\cdot)$ is the density function of the standard normal distribution. Let $f_r(h)$ be the density function of h among labour supply in region r. Then, the market equilibrium for the output of occupation o in region r is characterized by

$$
\underbrace{y_r^* + \ln\phi\left(\frac{o-O_r}{\sigma}\right)}_{\text{Leontieff for type } o} = \underbrace{\ln f_r\left[h_r\left(o\right)\right]}_{\text{labour supply type } h_r(o)} + \underbrace{\ln\frac{dh_r\left(o\right)}{do}}_{\text{Jacobian}} + \underbrace{y\left[h_r\left(o\right),o\right]}_{\text{productivity}},
$$

where y_r^* is log output per worker of the composite commodity of region r. The left hand side is log demand for the output of occupation o, the right hand side is log supply. Demand is equal to the log output of the composite commodity plus the log input of occupation o per unit of output. Supply is equal to log supply of human capital type $h_r(o)$ (which is equal to the log density plus the log Jacobian) times log productivity of $h_r(o)$ in occupation o. Equation (19) specifies a relation for $y(h,o)$. It is convenient to define

$$
y_0 = \omega_r + O_r.
$$

[11] We use the average log value and the log average value of a house interchangebly. The difference between both is $\frac{1}{2}\sigma^2\rho_r^{-2}$. We treat these terms as a normalizing, thereby ignoring variation in ρ_r.

The term ω_r is the agglomeration spill over in region r, the term O_r reflects the log price level for commodities of complexity level O_r. As we have shown in Section 2.2, the specification of $y\,(h, o)$ and the linearity of the wage function (14) yields a linear expression for the optimal assignment $h_r\,(o)$, see equation (22),

$$h_r\,(o) - H_r = o - O_r.$$

This expression implies that $\mathrm{d}h_r\,(o)\,/\mathrm{d}o = 1$ and

$$y\,[h_r\,(o)\,, o] = \omega_r + H_r - \frac{1}{2}\gamma\,(H_r - O_r)\,.$$

Hence,

$$\begin{aligned} y_r^* &= \omega_r + H_r, \\ \ln f_r\,(h) &= \ln \phi\left(\frac{h - H_r}{\sigma}\right) \Rightarrow h|r \sim N\left(H_r, \sigma^2\right). \end{aligned}$$

We drop the second order terms in H_r and O_r in the first line.

Agglomeration in rural areas

Substitution in equation (26) of equation (24) for l_r yields

$$\begin{aligned} \omega_r^r &= \psi\left(\ln\frac{\pi}{3} - 2\ln\delta + \theta O_r - l_r\right) = \psi\left(\ln\frac{\pi}{3} - 2\ln\delta + \theta O_r - v_r + p_r\right) \qquad (38) \\ &= \psi\left[\psi_0 + (\theta - \overline{\varepsilon})\,O_r + \psi_\omega\omega_r + \psi_T T_r\right] \\ &= \Psi^r\left[\psi_0 + (\theta - \overline{\varepsilon})\,O_r + \psi_T T_r\right], \\ \Psi^r &\equiv \frac{\psi}{1 - \psi\cdot\psi_\omega}, \\ \psi_\omega &\equiv \frac{1+\alpha_\omega}{\lambda} - \lambda_\omega = \frac{1-\overline{\lambda\eta}}{\lambda}\,(1+\alpha_\omega) - \frac{\overline{\varepsilon}}{\gamma}\varepsilon_\omega - 1, \\ \psi_T &\equiv \frac{\alpha_T}{\lambda} - \lambda_T = \frac{1-\overline{\lambda\eta}}{\lambda}\alpha_T - \frac{\overline{\varepsilon}}{\gamma}\varepsilon_T, \\ \psi_0 &\equiv \ln\frac{\pi}{3} - 2\ln\delta - \ln\lambda. \end{aligned}$$

Agglomeration in urban areas

The population of a city satisfies

$$\begin{aligned} n_r &= \ln\left[\int_0^{S_r} 2\pi s e^{-l_r} \mathrm{d}s\right] \\ &= \ln\pi - 2\ln\kappa + 2\ln\left(\kappa S_r\right) - v_r + p_r \\ &= \psi_0 + \ln 3 + 2\Delta + 2\ln\left(\kappa S_r\right) + \psi_\omega \omega_r + \psi_T T_r - \bar{\varepsilon} O_r, \end{aligned} \tag{39}$$

where we substitute equation (29) for κS_r, equation (30) for $\ln\frac{\delta}{\kappa}$ and equation (24) for l_r in the second line. Average commuting cost as a share of labour supply satisfy

$$\begin{aligned} f_r &= \ln\left[\int_0^{S_r} 2\pi s\left(1-\kappa s\right)\mathrm{d}s\right] - \ln\left[\int_0^{S_r} 2\pi s \mathrm{d}s\right] \\ &= \ln\left(1 - \frac{2}{3}\kappa S_r\right) \end{aligned}$$

Hence,

$$\begin{aligned} \omega_r^c &= \psi\left(n_r + f_r + \theta O_r\right) \\ &= \psi\left(\psi_0 + 2\Delta + (\theta - \bar{\varepsilon})O_r + 2\ln\left(\kappa S_r\right) + \ln\left(3 - 2\kappa S_r\right) + \psi_\omega \omega_r + \psi_T T_r\right) \\ &= \Psi^r\left(\psi_0 + 2\Delta + (\theta - \bar{\varepsilon})O_r + 2\ln\left(1 - \exp(-\omega_r)\right) + \ln\left(1 + 2\exp(-\omega_r)\right) - \psi_T T_r\right) \\ G(\omega_r^c) &= \omega_r^c - 2\Psi^r \ln\left(1 - \exp(-\omega_r)\right) - \Psi^r \ln\left(1 + 2\exp(-\omega_r)\right) \\ &= 2\Psi^r\Delta + \omega_r^r \\ \omega_r^c &= G^{-1}(2\Psi^r\Delta + \omega_r^r) \end{aligned}$$

Identification

Define Q_x^z to be the estimated coefficients

$Q \in \left\{\zeta_r, v_r, \omega_r\right\}, \zeta_r \equiv H_r - O_r, \gamma\left(1 - \rho_r\right)$, we drop the subscript r in the following

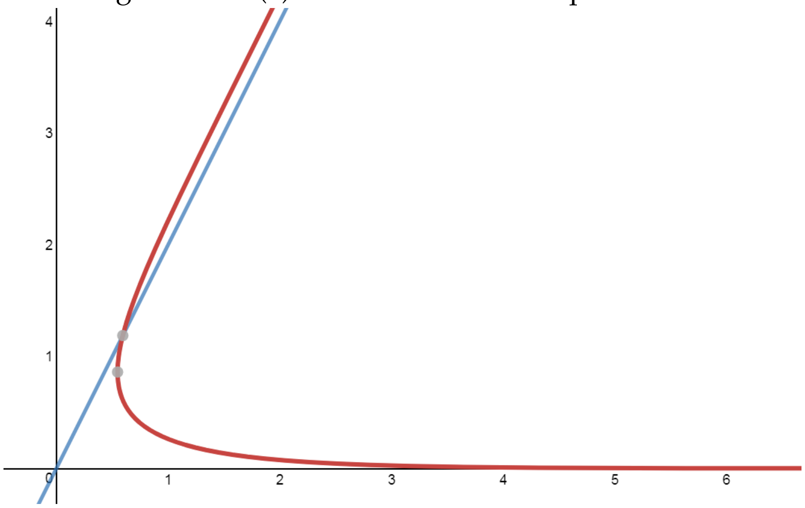
$z \in \{r, c, \text{blank}\}$: rural, cities, all

$x \in \{O, T, \omega\}$

$\gamma = 1$

Then we have the following relations:

Figure A1: G(a) function with model parameters



Note: The plot shows $\omega_r^c = G^{-1}(2\Psi^r\Delta + \omega_r^r)$ and $\omega_r^r = \Psi^r\left[\psi_0 + (\theta - \bar{\varepsilon})\, O_r + \psi_T T_r\right]$, with model parameter estimates in Table **??**. The red curve is the city agglomeration effect ω_r^c of area r. The blue line ω_r^r is the non-city agglomeration effect of area r.

$$
\begin{aligned}
\varepsilon_\omega &= \frac{\zeta_O}{\omega_O} = -\frac{0.34}{0.54} = -0.6 \\
\varepsilon_T &= \zeta_T - \varepsilon_\omega \omega_T = -0.04 + 0.63 \times 0.04 = -0.01 \\
\lambda_\omega &= \frac{v_O - \bar{\varepsilon}}{\omega_O} = \frac{3.11 - 0.50}{0.54} = 4.8 \\
\alpha_\omega &= \frac{\lambda}{\bar{\lambda}\bar{\eta}}\left(\lambda_\omega - 1 - \frac{\bar{\varepsilon}}{\gamma}\varepsilon_\omega\right) - 1 = \frac{0.30}{0.70 \times 0.25}\left(4.82 - 1 + 0.50 \times 0.63\right) - 1 = 6.1 \\
\lambda_T &= v_T - \lambda_\omega \omega_T = 0.16 - 4.82 \times 0.04 = -0.04 \\
\alpha_T &= \frac{\lambda}{\bar{\lambda}\bar{\eta}}\left(\lambda_T - \frac{\bar{\varepsilon}}{\gamma}\varepsilon_T\right) = \frac{0.30}{0.70 \times 0.25}\left(-0.042 + 0.5 \times 0.015\right) = -0.06 \\
\psi_\omega &= \frac{1 + \alpha_\omega}{\lambda} - \lambda_\omega = \frac{1 + 6.08}{0.30} - 4.82 = 18.78 \\
\psi_T &= \frac{\alpha_T}{\lambda} - \lambda_T = \frac{-0.06}{0.3} + 0.042 = -0.16 \\
\Psi^c &= -\frac{\omega_T^c}{\psi_T} = \frac{0.039}{0.16} = 0.24 \\
\theta &= \frac{\omega_O^c}{\Psi^c} + \bar{\varepsilon} = \frac{0.68}{0.24} + 0.5 = 3.33 \\
\Psi^r &= \frac{\omega_O^r}{\omega_O^c}\Psi^c = \frac{0.36}{0.684} \times 0.24 = 0.13 \\
\psi &= \frac{\Psi^r}{1 + \psi_\omega \Psi^r} = \frac{0.13}{1 + 18.78 \times 0.13} = 0.04
\end{aligned}
$$

Table A1: CBSA Observations Distribution Among States

CBSA	State I	State II	State III	State IV	Pct SI	Pct SII	Pct SIII	Pct SIV	NAME
31100	CA				100.00%				Los Angeles-Long Beach-Anaheim, CA
40140	CA				100.00%				Riverside-San Bernardino-Ontario, CA
41740	CA				100.00%				San Diego-Carlsbad, CA
41860	CA				100.00%				San Francisco-Oakland-Hayward, CA
41940	CA				100.00%				San Jose-Sunnyvale-Santa Clara, CA
19740	CO				100.00%				Denver-Aurora-Lakewood, CO
47900	DC	VA	MD		45.91%	25.90%	28.19%		Washington-Arlington-Alexandria, DC-VA-MD-WV
33100	FL				100.00%				Miami-Fort Lauderdale-West Palm Beach, FL
45300	FL				100.00%				Tampa-St. Petersburg-Clearwater, FL
12060	GA				100.00%				Atlanta-Sandy Springs-Roswell, GA
16980	IL	IN	WI		98.23%	1.77%	0.00%		Chicago-Naperville-Elgin, IL-IN-WI
26900	IN				100.00%				Indianapolis-Carmel-Anderson, IN
35380	LA				100.00%				New Orleans-Metairie, LA
14460	MA	NH			86.75%	13.25%			Boston-Cambridge-Newton, MA-NH
12580	MD				100.00%				Baltimore-Columbia-Towson, MD
19820	MI				100.00%				Detroit-Warren-Dearborn, MI
33460	MN	WI			99.99%	0.01%			Minneapolis-St. Paul-Bloomington, MN-WI
28140	MO	KS			45.36%	54.64%			Kansas City, MO-KS
41180	MO	IL			80.98%	19.02%			St. Louis, MO-IL
24660	NC				100.00%				Greensboro-High Point, NC
15380	NY				100.00%				Buffalo-Cheektowaga-Niagara Falls, NY
35620	NY	NJ			69.24%	30.76%			New York-Newark-Jersey City, NY-NJ
40380	NY				100.00%				Rochester, NY
17140	OH	KY			77.70%				Cincinnati, OH-KY-IN
17460	OH				100.00%				Cleveland-Elyria, OH
18140	OH				100.00%				Columbus, OH
38900	OR	WA			91.57%	8.43%			Portland-Vancouver-Hillsboro, OR-WA
37980	PA	NJ	DE	MD	62.06%	23.32%	14.62%	0.00%	Philadelphia-Camden-Wilmington, PA-NJ-DE-MD
38300	PA				100.00%				Pittsburgh, PA
19100	TX				100.00%				Dallas-Fort Worth-Arlington, TX
26420	TX				100.00%				Houston-The Woodlands-Sugar Land, TX
47260	VA				100.00%				Virginia Beach-Norfolk-Newport News, VA-NC
42660	WA				100.00%				Seattle-Tacoma-Bellevue, WA
33340	WI				100.00%				Milwaukee-Waukesha-West Allis, WI

Note: Information for 34 city areas: CBSA code in 2013, city belong to which state(s) and the percentage of sample observations in the CPS 1979-2015, name of cities. *Data sources*: the Current Population Survey MORG and the US Census Bureau.

Table A2: Individual Mincerian Wage Regression

Variables	Coefficient	t-stat	Variables	Coefficient	t-stat
Male	0.306	(639.4)	Edu = 0	-0.535	(-104.7)
Single	0.0200	(20.95)	Edu = 1	-0.480	(-38.08)
Male × Single	-0.202	(-253.7)	Edu = 2	-0.503	(-81.41)
Single × Time Trend	-0.000648	(-16.77)	Edu = 3	-0.491	(-91.94)
Divorced	-0.00696	(-5.693)	Edu = 4	-0.422	(-76.86)
Male × Divorced	-0.0792	(-66.45)	Edu = 5	-0.456	(-121.1)
Divorced × Time Trend	-1.09e-05	(-0.196)	Edu = 6	-0.404	(-136.1)
Black	-0.0969	(-100.7)	Edu = 7	-0.330	(-119.7)
Black × South	-0.0377	(-29.65)	Edu = 8	-0.243	(-131.4)
Other Race	-0.0808	(-71.86)	Edu = 9	-0.235	(-187.4)
Other × South	-0.00518	(-2.189)	Edu = 10	-0.171	(-191.3)
Year of Experience	0.0261	(52.50)	Edu = 11	-0.140	(-170.4)
Exp × Edu	0.00177	(45.78)	Edu = 13	0.0750	(100.9)
Exp^2 / 100	-0.0248	(-10.77)	Edu = 14	0.153	(206.9)
Exp^2 / 100 × Edu	-0.0100	(-55.07)	Edu = 15	0.184	(150.8)
Exp^3 / 100000	-0.0820	(-2.738)	Edu = 16	0.382	(380.4)
Exp^3 / 100000 × Edu	0.125	(50.98)	Edu = 17	0.436	(250.6)
Edu in y9297	0.00704	(36.87)	Edu = 18	0.527	(343.1)
Constant	1.194	(516.9)			
Observations	5,316,676				
R-squared	0.597				
R-MSE	0.414				

Note: Table presents the estimated β using OLS regression. Dependent variable is the log hourly wage. Mincer wage regression includes individual characteristics x, gender, year of education, year of experience, race, marital status, and the interaction of these factors. All the regressions include time x region dummies. Robust t-statistics in parentheses.

Table A3: Ranking of Regions in Occupational Structure

Region	Occ Index	Type	Region	Occ Index	Type
Washington, DC	0.093	City	Virginia	-0.004	Non-City
San Jose, CA	0.093	City	Greensboro, NC	-0.007	City
Boston, MA	0.070	City	Los Angeles, CA	-0.008	City
San Francisco, CA	0.056	City	Wyoming	-0.009	Non-City
Seattle, WA	0.056	City	Kansas	-0.012	Non-City
Denver, CO	0.053	City	North Carolina	-0.012	Non-City
Baltimore, MD	0.050	City	Florida	-0.013	Non-City
Connecticut	0.047	Non-City	Alabama	-0.014	Non-City
Minneapolis, MN	0.041	City	Louisiana	-0.014	Non-City
Atlanta, GA	0.039	City	West Virginia	-0.015	Non-City
Philadelphia, PA	0.037	City	Maine	-0.016	Non-City
Kansas City, MO	0.032	City	Buffalo, NY	-0.016	City
Indianapolis, IN	0.029	City	Ohio	-0.017	Non-City
New Hampshire	0.029	Non-City	Nebraska	-0.017	Non-City
Dallas, TX	0.029	City	South Carolina	-0.017	Non-City
Chicago, IL	0.027	City	Michigan	-0.018	Non-City
Portland, OR	0.026	City	Illinois	-0.018	Non-City
Houston, TX	0.024	City	Riverside, CA	-0.019	City
Milwaukee, WI	0.024	City	Maryland	-0.019	Non-City
Pittsburgh, PA	0.020	City	Iowa	-0.020	Non-City
Detroit, MI	0.020	City	Miami, FL	-0.020	City
Rochester, NY	0.019	City	Tennessee	-0.021	Non-City
Cleveland, OH	0.019	City	Kentucky	-0.023	Non-City
New York, NY	0.018	City	Pennsylvania	-0.025	Non-City
St Louis, MO	0.017	City	North Dakota	-0.025	Non-City
Columbus, OH	0.017	City	Wisconsin	-0.027	Non-City
San Diego, CA	0.016	City	Mississippi	-0.029	Non-City
Virginia Beach, VA	0.013	City	Washington	-0.029	Non-City
Cincinnati, OH	0.012	City	Texas	-0.030	Non-City
Massachusetts	0.011	Non-City	Montana	-0.031	Non-City
New Orleans, LA	0.011	City	California	-0.031	Non-City
Utah	0.008	Non-City	Indiana	-0.031	Non-City
Colorado	0.008	Non-City	Idaho	-0.033	Non-City
Rhode Island	0.007	Non-City	South Dakota	-0.037	Non-City
Tampa, FL	0.006	City	Georgia	-0.037	Non-City
Vermont	0.004	Non-City	Arkansas	-0.038	Non-City
Arizona	0.002	Non-City	Missouri	-0.040	Non-City
New Mexico	0.001	Non-City	Oregon	-0.044	Non-City
New York	0.000	Non-City	Minnesota	-0.049	Non-City
Oklahoma	-0.003	Non-City	Nevada	-0.074	Non-City
Delaware	-0.003	Non-City			

Note: Average local occupation index and region type. 34 cities are denoted by the name of largest city with the abbreviation of the state. 47 non-city areas are denoted by the name of the states. Detailed definitions of occupation index in section 2. *Data sources*: Current Population Survey MORG and author's own calculations.

Table A4: Bartik IV Regression Results with 34 MSAs Over-time

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
VARIABLES	O_r	H_r	w'_r	p_r	w_r	n_r	O_r	H_r	w'_r	p_r	w_r	n_r	O_r	H_r	w'_r	p_r	w_r	n_r
			1979-1986						*1993-2000*						*2007-2015*			
Panel A: City and Non-City Sample (81 Observations)																		
log Jan Temp	0.00631	-0.0188	-0.0134	0.189	0.0332		0.00197	-0.0251	0.0550	0.127	0.0325		-0.00527	-0.0284	0.0519	0.210	0.0263	
	(1.70)	(-2.53)	(-0.89)	(2.95)	(1.76)		(0.52)	(-4.03)	(3.48)	(1.52)	(1.67)		(-0.96)	(-4.02)	(3.83)	(1.84)	(1.68)	
Bartik IV	1.830	1.211	1.167	4.926	3.555		1.816	1.283	1.336	6.563	3.272		1.618	1.291	2.112	7.623	2.419	
	(19.66)	(6.46)	(3.08)	(3.06)	(7.52)		(20.66)	(8.87)	(3.64)	(3.38)	(7.24)		(13.62)	(8.45)	(7.20)	(3.08)	(7.12)	
R-squared	0.848	0.371	0.117	0.220	0.460		0.857	0.557	0.326	0.213	0.476		0.719	0.537	0.532	0.216	0.475	
R-MSE	0.0124	0.0250	0.0505	0.215	0.0630		0.0127	0.0209	0.0529	0.280	0.0652		0.0183	0.0236	0.0453	0.381	0.0524	
Panel B: City Sample Only (34 Observations)																		
log Jan Temp	0.00711	-0.0152	-0.0734	0.423	0.0177	0.180	-0.0104	-0.0272	0.0239	0.361	0.00386	0.325	-0.00257	-0.0246	0.0121	0.703	0.0527	0.448
	(1.28)	(-1.27)	(-3.48)	(3.88)	(0.65)	(0.50)	(-2.00)	(-2.50)	(0.84)	(2.43)	(0.16)	(0.92)	(-0.26)	(-1.90)	(0.57)	(3.64)	(2.46)	(1.20)
Bartik IV	1.745	1.296	-0.697	12.10	3.578	11.41	1.856	1.291	-0.523	13.97	3.271	9.843	1.813	1.487	1.988	18.18	3.073	9.140
	(10.79)	(3.71)	(-1.13)	(3.80)	(4.53)	(1.08)	(13.79)	(4.57)	(-0.71)	(3.64)	(5.24)	(1.07)	(7.78)	(4.89)	(4.01)	(4.01)	(6.11)	(1.05)
R-squared	0.808	0.391	0.323	0.488	0.442	0.096	0.886	0.567	0.146	0.422	0.559	0.078	0.698	0.527	0.426	0.513	0.645	0.080
R-MSE	0.0108	0.0233	0.0412	0.213	0.0528	0.706	0.0101	0.0212	0.0552	0.289	0.0469	0.690	0.0194	0.0253	0.0413	0.377	0.0419	0.726
Panel C: Non-City Sample Only (47 Observations)																		
log Jan Temp	0.00141	-0.0255	0.0102	0.0765	0.0202		0.00508	-0.0292	0.0583	0.00434	0.0249		-0.00941	-0.0345	0.0640	-0.0344	0.000386	
	(0.31)	(-2.64)	(0.52)	(1.07)	(0.92)		(1.20)	(-4.06)	(3.19)	(0.05)	(1.06)		(-1.53)	(-4.39)	(3.79)	(-0.28)	(0.02)	
Bartik IV	1.486	0.651	1.694	0.226	1.147		1.242	0.645	2.035	-1.361	0.383		1.088	0.616	0.967	-0.827	0.497	
	(10.21)	(2.13)	(2.74)	(0.10)	(1.66)		(9.30)	(2.85)	(3.53)	(-0.47)	(0.51)		(6.02)	(2.67)	(1.95)	(-0.23)	(0.89)	
R-squared	0.760	0.217	0.160	0.027	0.083		0.741	0.436	0.420	0.077	0.238		0.540	0.481	0.418	0.090	0.178	
R-MSE	0.0119	0.0251	0.0506	0.185	0.0568		0.0110	0.0186	0.0472	0.239	0.0610		0.0159	0.0203	0.0436	0.316	0.0490	

Note: Dependent variables are the average occupation index O_r, the average human capital index H_r, average return to human capital w'_r, average housing value p_r, the average wage w_r. Columns (1)-(6), (7)-(12) and (13)-(18) present results using samples from 1979-1986, 1993-200 and 2017-2015. Panel A shows results with mix city and non-city samples. Panel B shows city sample only. Panel C shows non-city sample only. Log Jan Temp is the average log local January temperature. Bartik IV defined in section 5. Detailed definitions of equations are in section 2. Robust t-statistics in parentheses.

Table A5: Bartik IV Regression Results with 222 MSAs

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES	O_r	O_r	$1 - w_{hr}$	$H_r - O_r$	v_r	w_{0r}	n_r
Panel A: City and Non-City Sample (267 Observations)							
ln Jan Temp	-0.0135		-0.0202	-0.0727	0.0691	0.0439	
	(-1.70)		(-1.51)	(-8.36)	(0.55)	(4.05)	
Bartik IV	1.661	1.739					
	(22.42)	(23.88)					
Bartik IV sq.	3.993	4.412					
	(4.75)	(5.26)					
Occ IV			-0.255	-0.160	0.617	-0.00168	
			(-3.72)	(-3.57)	(0.95)	(-0.03)	
Metro Dummy	0.0295		-0.0498	0.00716	0.0703	0.0111	
	(3.62)		(-3.64)	(0.80)	(0.54)	(1.00)	
Constant	0.0372	-0.0165	0.179	0.410	11.43	-0.243	
	(0.83)	(-4.84)	(2.36)	(8.31)	(15.97)	(-3.95)	
R-squared	0.719	0.704	0.131	0.229	0.007	0.071	
R-MSE	0.0471	0.0482	0.0798	0.0520	0.753	0.0647	
Panel B: City Sample (222 Observations)							
ln Jan Temp			-0.0165	-0.0680	0.142	0.0424	0.649
			(-1.07)	(-6.81)	(1.21)	(3.41)	(3.59)
Occ IV			-0.255	-0.144	0.864	-0.0207	3.371
			(-3.49)	(-3.04)	(1.56)	(-0.35)	(3.93)
Constant			0.107	0.390	11.08	-0.224	9.537
			(1.20)	(6.72)	(16.29)	(-3.09)	(9.06)
R-squared			0.054	0.189	0.015	0.053	0.102
R-MSE			0.0833	0.0538	0.631	0.0672	0.976
Panel C: Non-City Sample (45 Observations)							
ln Jan Temp			-0.0360	-0.0983	-0.324	0.0562	
			(-1.51)	(-6.31)	(-0.69)	(2.93)	
Occ IV			-0.229	-0.511	-4.779	0.438	
			(-0.89)	(-3.05)	(-0.94)	(2.13)	
Constant			0.270	0.530	13.28	-0.282	
			(2.04)	(6.14)	(5.09)	(-2.65)	
R-squared			0.059	0.506	0.027	0.208	
R-MSE			0.0604	0.0395	1.193	0.0486	

t-statistics in parentheses

Note: Columns (1)-(7) present the Bartik IV regression results with 222 Cities and 45 non-city areas. Dependent variables are the average occupation index, local average log wage, housing value normalised by land share and the local wage level, returns to human capital, difference between average human capital index and the occupation index, and the log local population. Log Jan Temp is the average log local January temperature. Bartik IV and Bartik IV sq are defined in section 5. Occ IV is the average occupation index after IV correction. Metro Dummy equals one if an observation is city area, zero if it is non-city area. Detailed definitions of equations are in section 2. Robust t-statistics in parentheses.