Journal of Economic Geography **12** (2012) pp. 435–453
Advance Access Published on 4 May 2011
doi:10.1093/jeg/lbr006

Specialized knowledge and the geographic concentration of occupations

*Todd M. Gabe**,† *and Jaison R. Abel***

*School of Economics, University of Maine, 5782 Winslow Hall, Orono, ME 04469, USA.
** Research and Statistics Group, Federal Reserve Bank of New York, 237 Main Street, Suite 1200, Buffalo, NY 14203, USA.
†Corresponding author: Todd M. Gabe, School of Economics, University of Maine, 5782 Winslow Hall, Orono, ME 04469, USA. *email* <todd.gabe@umit.maine.edu>

Abstract

This article examines the effects of specialized knowledge on the geographic concentration of occupations across US metropolitan areas. Controlling for a wide range of other attributes, empirical results reveal that occupations with a unique knowledge base exhibit higher levels of concentration than those with generic knowledge requirements. This result is robust to the use of several model specifications and instrumental variables estimation that relies on an instrument set representing the means by which people acquire knowledge. Thus, the study suggests that the benefits of labor market pooling are particularly apparent in cases where workers require a specialized knowledge base.

Keywords: Agglomeration, geographic concentration, occupations, labor market pooling, knowledge
JEL classification: R12, J24, R20
Date submitted: 22 January 2010 **Date accepted:** 14 March 2011

1. Introduction

One of the most striking features of the US economy is the high geographic concentration of people and firms.[1] Indeed, the mere existence of cities is testament to the idea that people—and the businesses that employ them—benefit from the close proximity of others (Lucas, 1988). Textile manufacturing, which generates a large share of its output at a very few locations, is one of the most-cited examples of an industry with a high geographic concentration of activity (Krugman, 1991; Ellison and Glaeser, 1997; Duranton and Overman, 2005). A frequently mentioned example of an industry cluster is the strong presence of computer programming and other high-technology firms located in and around Silicon Valley (Saxenian, 1994; Fallick et al., 2006).

These locational patterns are advantageous to workers and firms: people can move among employers without retooling and businesses have access to a deep pool of labor with the skills they need. Indeed, in their study of industry co-agglomeration, Ellison et al. (2010) found that sectors employing the same types of workers tend to

1 This statement applies to many other nations (Maurel and Sedillot, 1999; Devereux et al., 2004; Andersson et al., 2005; Barrios et al., 2005; McCann and Simonen, 2005).

co-agglomerate. Moreover, several studies have uncovered empirical evidence of increased labor mobility (i.e. movement among jobs) that is facilitated by industry agglomeration (Fallick et al., 2006; Freedman, 2008). A common theme in the existing literature is that the benefit of agglomeration resulting from labor market pooling is particularly helpful when workers require a specialized skill set, whether it is the ability to turn fibers into textiles or write computer programs.

This article examines the geographic concentration of occupations across the US metropolitan areas. We focus on the importance of labor market pooling as measured by the extent to which occupations require a specialized knowledge base related to a wide variety of topics. Our analysis of occupations provides a new way to look at the forces that influence the geographic concentration of economic activity. Industry-centric studies focus on where similar types of goods and services are made, since sectors are assigned based on a firm's primary output. In contrast, recent occupational-based approaches to urban and regional analysis emphasize what people do in their jobs (Feser, 2003; Markusen, 2004; Florida et al., 2008; Gabe, 2009; Scott, 2009; Bacolod et al., 2009a, 2009b). Here, we use occupations to understand the knowledge required to perform a job.

For at least two of Marshall's (1920) ideas about the benefits of agglomeration, the effects seem to be more pronounced for occupations (i.e. tasks and activities people perform in their jobs) than industries (i.e. goods and services provided). Agglomeration facilitates knowledge spillovers because it allows individuals to share ideas and tacit knowledge (Kloosterman, 2008; Ibrahim et al., 2009). A computer programmer, for example, presumably benefits more from proximity to others involved in similar day-to-day activities (e.g. interacting with computers, using technology) than he or she gains from working next to others in the same industry (e.g. a software company's receptionist, human resources specialist, or chief executive).

Likewise, the basic idea behind labor market pooling—that agglomeration provides a thick labor market for those who possess or require a particular skill set—seems to apply more readily to occupations than industries. In an analysis of industry agglomeration, Rosenthal and Strange (2001, 205) suggest that labor market pooling is the most problematic of the Marshallian micro-foundations to measure because 'it is difficult to identify industry characteristics that are related to the specialization of the industry's labor force'.[2] This is not the case with occupations. Some jobs require a very specific knowledge and skill set that is specialized to the task at hand, while other occupations call for a more generic set of knowledge and skills.

Knowledge is measured in the study using data from the US Department of Labor's Occupational Information Network (O*NET), which covers 33 subject areas ranging from physics to fine arts. We incorporate information on all 33 topics into a single measure of the extent to which an occupation's overall knowledge profile differs from the average US job. Locational Gini coefficients, based on micro-level data from the 2008 American Community Survey on 1.3 million individuals in 284 US metropolitan

2 Rosenthal and Strange (2001) use three measures (e.g. net productivity, an indicator of 'brains to brawn' and the percentage of workers with advanced degrees) as proxies for the importance of labor market pooling. Overman and Puga (2010) point out the limitations of these indicators and, instead, focus on the effects of idiosyncratic firm-level employment shocks on industry agglomeration. Their results suggest establishments that expand while the overall industry declines (or vice versa) benefit more from agglomeration than plants in sectors with homogeneous employment shocks.

areas, are used to represent the geographic concentration of occupations. The regression models estimated in the article examine the relationship between the locational Gini coefficients and our measure of specialized knowledge, while controlling for other occupational-level attributes expected to influence the geographic concentration of economic activity.

Explanations about the forces influencing the locational patterns of economic activity have evolved from Marshall's (1920) discussion of the sources of industry agglomeration (i.e. pooled labor force, availability of non-traded inputs and knowledge spillovers) to Duranton and Puga's (2004) more formal exposition of these micro-foundations based on spatial externalities arising from sharing, matching and learning. In our analysis, the preferred explanation for the agglomeration effects of labor market pooling is that it makes workers and businesses better off when firms face idiosyncratic demand shocks (Krugman, 1991; Overman and Puga, 2010).

Duranton and Puga (2004, 2085) characterize this explanation as one of 'sharing risk', in which there are 'efficiency gains from sharing resources (e.g. skilled workforce) among firms that do not know *ex ante* how many of these resources they will need'. An important feature of the existing theories of agglomeration is that they predict benefits of labor market pooling [or, using the terminology of Duranton and Puga (2004), 'risk sharing'] in cases where firms require a specific type of labor. For example, in Krugman's (1991, 38) labor market pooling model, he assumes that firms 'use the same distinctive kind of skilled labor'. Likewise, Overman and Puga (2010) assume that workers have skills that are specific to a given sector.

Consistent with this idea, our regression results show that occupations with knowledge profiles that differ from the average US job have higher levels of concentration than those with more generic knowledge requirements. This finding is robust to the use of several model specifications, including fixed-effects models that control for an occupation's major occupational category and instrumental variables estimation that relies on an instrument set representing the means by which people acquire knowledge. Thus, as commonly assumed in models of industry concentration, the main findings from this research suggest that the benefits of labor market pooling are particularly apparent in cases where workers require a specialized skill set.

2. The geographic concentration of the US occupations

Following Krugman (1991), Audretsch and Feldman (1996) and Jensen and Kletzer (2006), we use locational Gini coefficients to measure the geographic concentration of occupations across US metropolitan areas.[3] The locational Gini coefficient (LGINI) for the US Census occupations, indexed by k, is calculated as (Kim et al., 2000):

$$\mathrm{LGINI}_k = \Delta/4\,u, \tag{1}$$

3 Although Rosenthal and Strange (2001) found that the determinants of agglomeration differ somewhat at the zip code, county and state levels, metropolitan areas are an appropriate geographic unit of observation to examine the concentration of occupations because they represent the labor market areas in which workers interact and move among jobs. Moreover, data limitations prevent us from examining occupational concentration at the zip code and county levels. Our reliance on occupations as the unit of observation necessarily excludes any within-occupation variation that may exist.

where $\Delta = \{1/[n(n-1)]\} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|$

$i, j =$ US metropolitan areas ($i \neq j$),

$u =$ mean of x_i,

$x_{i(j)} =$ [metro area *i*'s (*j*'s) share of employment in k/ metro area *i*'s (*j*'s) share of total employment],

and, $n = 284$, the number of US metropolitan areas included in the analysis.

Locational Gini values close to zero suggest that employment in the occupation is widely dispersed across US metropolitan areas and spread out in a manner similar to the distribution of overall employment. Values close to 0.5 suggest that workers in the occupation are geographically concentrated in a single metropolitan area, or located in very few places.

We used individual-level data from the 1% Public Use Microdata Sample (PUMS) of the 2008 American Community Survey to construct locational Gini coefficients (Ruggles et al., 2010). This dataset contains over 3 million observations, but our sample size is smaller because we focus on individuals for whom we know their occupation and metropolitan area location. After removing people without an identified occupation (e.g. retirees, students, etc.) and those who live outside of metropolitan areas, we constructed the locational Gini coefficients using data on 1,321,889 individuals. The 2008 American Community Survey includes information on 284 metropolitan areas, as defined by the US Census, and 470 occupations.[4]

To assess whether the geographic concentration of occupations reveals anything beyond what is already known through previous studies on the geographic patterns of industries, we estimated a second set of locational Gini coefficients using occupational employment figures that were estimated based on the industries present in a given metropolitan area. To do so, we used information on the 1.3 million individuals from the 2008 American Community Survey to construct an occupational-industry matrix. Then, we used this matrix along with industry employment data for each metropolitan area to estimate occupational employment. If the occupational Gini coefficients were dictated solely by the location patterns of the industries they support, we would expect the predicted values to be reasonably close approximations of the actual geographic concentration of occupations. However, the correlation between these two measures is only 0.60, which—although positive—suggests that the forces influencing concentration differ between industries and occupations.

Table 1 presents information on the actual geographic concentration of the US occupations, summarized by major standard occupational classification (SOC) category. We find that the most geographically concentrated jobs are in the broad categories of farming, fishing and forestry occupations; production occupations; and life, physical and social science occupations. The major occupational categories that exhibit the lowest average values of the locational Gini coefficients include food preparation and serving related occupations; building and grounds cleaning and maintenance occupations; and community and social services occupations. Workers in these latter broad job categories tend to be geographically dispersed across the group of 284 US metropolitan areas.

4 The sample size is less than 470 occupations in our regression analysis because we removed public administration- and military-related categories, as well as occupations with missing information for one or more of the explanatory variables.

Table 1. Geographic concentration of major occupational categories, 2008

SOC category	Description	Average LGINI
11-0000	Management occupations	0.2204
13-0000	Business and financial operations occupations	0.2765
15-0000	Computer and mathematical occupations	0.2766
17-0000	Architecture and engineering occupations	0.3210
19-0000	Life, physical and social science occupations	0.3662
21-0000	Community and social services occupations	0.1769
23-0000	Legal occupations	0.1829
25-0000	Education, training and library occupations	0.1909
27-0000	Arts, design, entertainment, sports and media occupations	0.2797
29-0000	Healthcare practitioners and technical occupations	0.2900
31-0000	Healthcare support occupations	0.2252
33-0000	Protective service occupations	0.2901
35-0000	Food preparation and serving related occupations	0.1514
37-0000	Building and grounds cleaning and maintenance occupations	0.1751
39-0000	Personal care and service occupations	0.2819
41-0000	Sales and related occupations	0.1888
43-0000	Office and administrative support occupations	0.2417
45-0000	Farming, fishing and forestry occupations	0.4137
47-0000	Construction and extraction occupations	0.3143
49-0000	Installation, maintenance and repair occupations	0.3057
51-0000	Production occupations	0.3675
53-0000	Transportation and material moving occupations	0.3188

Notes: Locational Gini values are based on 1,321,889 people included in the 2008 American Community Survey 1% Public Use Microdata Sample (PUMS). Individuals are located in one of 284 US metropolitan areas.

Tables 2 and 3 show the 20 most and least geographically concentrated US occupations. Jobs that involve aspects of textile manufacturing (e.g. textile knitting and weaving machine setters; textile winding, twisting and drawing out machine setters) exhibit high levels of geographic concentration, similar to the ranking of textiles among the most concentrated manufacturing industries reported by Krugman (1991), Ellison and Glaeser (1997) and Duranton and Overman (2005). As expected, southern metropolitan areas such as Danville, Virginia and Chattanooga, Tennessee–Georgia, have the highest relative employment shares in these textile-related occupations. Other occupations that are highly concentrated include gaming workers (e.g. gaming cage workers), shoemakers (e.g. shoe machine operators and tenders) and specialized engineers (e.g. marine engineers and naval architects; petroleum, mining and geological engineers). On the other end of the spectrum, the least geographically concentrated occupations include secretaries and administrative assistants; retail salespersons; elementary and middle school teachers and registered nurses. These jobs tend to be evenly dispersed in proportions similar to overall employment across the US metropolitan areas.

This descriptive analysis provides insight into some of the factors that might influence the concentration of occupations in the United States. For example, natural advantage appears to be important for the locational patterns of several of the most concentrated occupations (e.g. petroleum, mining and geological engineers; riggers). However, it is

Table 2. Twenty most geographically concentrated occupations, 2008

Occupation	Locational Gini	Metro w/ highest relative employment share
Shoe machine operators and tenders	0.4898	Danville, VA
Textile knitting and weaving machine setters, operators and tenders	0.4785	Danville, VA
Textile winding, twisting, and drawing out machine setters, operators, and tenders	0.4748	Chattanooga, TN–GA
Tire builders	0.4732	Gadsden, AL
Marine engineers and naval architects	0.4732	Bremerton, WA
Forging machine setters, operators and tenders	0.4723	Rockford, IL
Gaming cage workers	0.4695	Las Vegas, NV
Rail-track laying and maintenance equipment operators	0.4690	Redding, CA
Subway, streetcar, and other rail transportation workers	0.4687	Redding, CA
Brokerage clerks	0.4686	Florence, AL
Drilling and boring machine tool setters, operators and tenders	0.4682	Decatur, AL
Model makers and patternmakers	0.4682	Benton Harbor, MI
Petroleum, mining and geological engineers, including mining safety engineers	0.4675	Houma-Thibodoux, LA
Purchasing agents and buyers, farm products	0.4668	Sioux City, IA–NE
Railroad brake, signal and switch operators	0.4664	St. Joseph, MO
Heat treating equipment setters, operators and tenders	0.4652	Jackson, TN
Cleaning, washing and metal pickling equipment operators and tenders	0.4647	Danville, VA
Riggers	0.4631	Bremerton, WA
Tool grinders, filers and sharpeners	0.4624	Sioux City, IA–NE
Explosives workers, ordnance handling experts and blasters	0.4615	Jacksonville, NC

Notes: Locational Gini values are based on 1,321,889 people included in the 2008 American Community Survey 1% Public Use Microdata Sample (PUMS). Individuals are located in one of 284 US metropolitan areas. A metropolitan area's relative employment share is defined as its share of employment in a particular occupation divided by its share of overall employment.

Table 3. Twenty least geographically concentrated occupations, 2008

Occupation	Locational Gini
Secretaries and administrative assistants	0.0518
First-line supervisors/managers of retail sales workers	0.0579
Retail salespersons	0.0592
Elementary and middle school teachers	0.0614
Cashiers	0.0627
First-line supervisors/managers of office and administrative support workers	0.0761
Janitors and building cleaners	0.0766
Bookkeeping, accounting and auditing clerks	0.0771
Registered nurses	0.0775
Receptionists and information clerks	0.0818
Stock clerks and order fillers	0.0823
Cooks	0.0836
Customer service representatives	0.0849
Miscellaneous managers, including postmasters and mail superintendents	0.0857
Waiters and waitresses	0.0885
Office clerks	0.0895
Child care workers	0.0900
Driver/sales workers and truck drivers	0.0911
Construction laborers	0.0930
Food service managers	0.0943

Notes: Locational Gini values are based on 1,321,889 people included in the 2008 American Community Survey 1% Public Use Microdata Sample (PUMS). Individuals are located in one of 284 US metropolitan areas.

also clear that many of the most concentrated occupations require highly specialized skills (e.g. textile knitting and weaving machine setters, operators, and tenders; marine engineers and naval architects). Moreover, the least geographically concentrated occupations tend to have generic knowledge requirements that are easily transferable across a wide range of jobs (e.g. secretaries and receptionists; salespersons and cashiers). With this in mind, we now turn to a more formal analysis of the determinants of occupational concentration.

3. Determinants of occupational concentration

To investigate whether specialized knowledge influences the geographic concentration of occupations in the United States, we estimate a regression model that examines the relationship between the locational Gini coefficients described above and a new measure of the distinctiveness of the cognitive skills required to perform a job. The regression model controls for a wide range of factors that may also influence occupational concentration, such as other agglomerative forces (e.g. knowledge spillovers), natural advantage and transport costs, to isolate the effect of specialized knowledge. Specifically, using occupations as observations, we estimate the following model:

$$\mathrm{LGINI}_k = \beta_1 \text{ specialized knowledge}_k + \beta \boldsymbol{X}_k + \varepsilon_k, \tag{2}$$

where $k \equiv$ occupation, $\boldsymbol{X} \equiv$ vector of controls and $\varepsilon \equiv$ error term. This is the same general approach used by Rosenthal and Strange (2001) to examine the concentration of manufacturing industries. With this specification, the key coefficient we estimate is identified by the cross-sectional variation in the knowledge requirements across occupations. Summary statistics for the variables used in the regression analysis are presented in Table 4, while a detailed description of these variables is provided below.

3.1. Specialized knowledge

The explanatory variable of key interest, used to measure the distinctiveness of the cognitive skills required in a job, is the extent to which an occupation's knowledge profile differs from the US norm. This variable, *specialized knowledge*, is constructed using information from the US Department of Labor's Occupational Information Network (O*NET, version 12.0) on the importance and level of knowledge required in 33 subjects (see Table 5).[5] The O*NET, which is based on employee surveys and input from professional occupational analysts, asks respondents to rate on a scale of 1–5 the importance of these knowledge areas to a person's job. For topics that are rated as at least 'somewhat important' (i.e. a score of 2 or higher), the respondent is asked to rate on a scale of 1–7 the level of knowledge required.

Using data from 33 different questions (i.e. one for each of the subject areas) included on the O*NET survey, we multiplied the importance (i.e. scale of 1–5) and level (i.e. scale of 1–7) ratings to construct occupational-level knowledge indices (Feser, 2003). With this information and estimates from the 2008, American Community Survey on total US employment by occupation, we calculated the (weighted) average US knowledge requirements in each of the 33 topics. To measure the extent to which an individual occupation's knowledge profile differs from the norm, we constructed the *specialized knowledge* variable as:

$$\text{Specialized knowledge}_k = \sum_{z=1}^{33} (\mathrm{KI}_{k,z} - \mathrm{KI}_{ave,z})^2, \tag{3}$$

where the subscript z indicates the knowledge area, KI is the knowledge index, the subscript k indicates the occupation, and the subscript *ave* indicates the average US occupation. Low values of this variable indicate that the occupation's knowledge profile is similar to the average US job, while high values suggest that the occupation requires specialized knowledge.

3.2. Other determinants of occupational concentration

Of Marshall's three explanations of the benefits of agglomeration, knowledge spillovers have received the most attention in the literature (Jaffe et al., 1993; Audretsch and Feldman, 1996; Kloosterman, 2008; Ibrahim et al., 2009). The idea here is that agglomeration (and, in our case, a high geographic concentration of activity) allows workers to learn job-specific tasks and stay with current new developments as if they were 'in the air'. The variable *update knowledge* is used to measure the extent to which

5 See Peterson et al. (2001) for a detailed discussion of O*NET.

Table 4. Descriptive statistics ($n = 433$)

Variable	Description	Mean	Standard deviation
LGINI	Locational Gini coefficient calculated across 284 US metropolitan area	0.285	0.116
Specialized knowledge	Variable measuring the difference between an occupation's knowledge profile and the knowledge profile of the average US occupation	695.0	405.4
Update knowledge	Index value that measures the importance (scale of 1–5) and level (scale of 1–7) of occupational-related activity titled 'updating and using relevant knowledge'	14.98	5.76
Machinery	Index value that measures the importance (scale of 1–5) and level (scale of 1–7) of occupational-related activity titled 'controlling machines and processes'	8.76	6.96
Interaction with public	Index value that measures the importance (scale of 1–5) and level (scale of 1–7) of occupational-related activity titled 'performing for or working directly with the public'	10.33	7.17
Average establishment size	Average size of businesses that employ workers in the occupation	57.19	75.50
Agriculture	Share of people in occupation who work in agricultural-related industry	0.023	0.118
Mining	Share of people in occupation who work in mining-related industry	0.017	0.089
Industry distribution	Entropy measure of an occupation's distribution across 19 major NAICS industrial categories	0.541	0.286
Occupation size	Number of people in occupation included in the 2008 American Community Survey 1% Public Use Microdata Sample (PUMS)	2,824	5,096
Difference in education	Variable measuring the squared difference between the share of workers in an occupation with at least a BA/BS degree and the average value across all occupations	0.081	0.104
Difference in experience	Variable measuring the squared difference between the months of experience required for an occupation and the experience requirements of the average US occupation	309.4	507.2
Difference in training	Variable measuring the squared difference between the months of job training required for an occupation and the training requirements of the average US occupation	69.76	176.3

Table 5. Knowledge areas

Administration and Management	Building and Construction	Education and Training
Clerical	Mechanical	English Language
Economics and Accounting	Mathematics	Foreign Language
Sales and Marketing	Physics	Fine Arts
Customer and Personal Service	Chemistry	History and Archeology
Personnel and Human Resources	Biology	Philosophy and Theology
Production and Processing	Psychology	Public Safety and Security
Food Production	Sociology and Anthropology	Law and Government
Computers and Electronics	Geography	Telecommunications
Engineering and Technology	Medicine and Dentistry	Communications and Media
Design	Therapy and Counseling	Transportation

an occupation requires workers to keep up with new information relevant to the job. Presumably, a high geographic concentration of employment would be beneficial to workers in occupations that place a high premium on keeping current with new developments. *Update knowledge* is constructed as an index (i.e. 'importance' multiplied by the 'level' of this activity required in an occupation) using O*NET data on an occupational activity titled 'updating and using relevant information'. Unlike the *specialized knowledge* variable that is constructed using information from 33 questions included on the O*NET survey, the *update knowledge* variable uses information from a single O*NET question.

Along with ideas about labor market pooling and knowledge spillovers, Marshall (1920) suggested that agglomeration (and, in our case, the geographic concentration of economic activity) facilitates the sharing of intermediate inputs. Specialized machinery and equipment, especially items that exhibit increasing returns to scale in their use, are examples of inputs that workers and firms may locate around. Marshall (1920, 225) notes 'the economic use of expensive machinery can sometimes be attained in a very high degree in a district in which there is a large aggregate production of the same kind, even though no individual capital employed in the trade be very large'. We use the variable, labeled as *machinery*, to represent the use of machinery in a person's job. It is constructed as an index, similar to the *update knowledge* variable, using information on the importance and level of the O*NET occupational activity titled 'controlling machines and processes'.

The explanatory variable *interaction with public* represents a sort of transport cost that is expected to affect agglomeration. In an analysis of the agglomeration of service industries, Kolko (2010) suggests that transport costs dictate that low-value services delivered through face-to-face contact should be geographically dispersed. Moreover, jobs characterized by heavy interaction with the general public typically require face-to-face contact, which limits an occupation's tendency to agglomerate (Storper and Venables, 2004). We constructed the *interaction with public* variable as an index using information from O*NET on the importance and level of an occupational activity titled 'performing for or working directly with the public'.

To account for the importance of establishment-level economies of scale, the regression model includes the variable *average establishment size*. It is constructed by matching occupations to industries using individual-level data from the 2008 American Community Survey. After determining the sectors that correspond to each of the occupations, we calculated an average employment size using industry-level data from County Business Patterns. The explanatory variables labeled as *agriculture* and *mining* measure the percentage of occupational employment in agricultural- or mining-related industries. These variables, which were constructed in a similar fashion using information from the 2008 American Community Survey, account for the importance of natural advantage and raw material use in geographic locational patterns (Kim, 1995; Ellison and Glaeser, 1999).

The final two explanatory variables used in our analysis of the geographic concentration of occupations are *industry distribution* and *occupation size*. The variable *industry distribution* uses a simple entropy measure to capture the distribution of workers across 19 major NAICS industrial categories. Low values indicate that workers in an occupation are spread across industries, while high values of *industry distribution* suggest that occupational employment is concentrated in a few sectors. This variable, constructed using individual-level data from the 2008 American Community Survey, controls for instances in which the locational patterns of occupations are intimately connected to the industries that they support. Finally, a weakness of the locational Gini when studying industries is that it could suggest high levels of concentration in cases where sectors comprised of a few large companies locate in a dispersed, random pattern (Ellison and Glaeser, 1997).[6] To address this limitation of the locational Gini coefficient, the regression models estimated in the paper control for the number of workers employed in each occupation. This variable, *occupation size*, accounts for cases in which a high geographic concentration might be explained by the location decisions of a relatively small group of individuals.

4. Regression results

4.1. Baseline results and robustness checks

Table 6 presents OLS regression results on the effects of specialized knowledge on the geographic concentration of occupations. Findings from the baseline regression analysis, reported in the first column of results, indicate that our measure of specialized knowledge has a positive and significant effect on the geographic concentration of occupations. Since the dependent variable and *specialized knowledge* both enter into the regression as natural logs, the estimated coefficient can be interpreted as an elasticity. Specifically, our baseline estimate suggests that a doubling of the *specialized knowledge* variable, greater than a one and one-half standard deviation increase, is associated with a 4.1% increase in the locational Gini coefficient.

Other results from the baseline regression model are generally consistent with expectations based on Marshall's ideas and other studies that have examined the geographic concentration of economic activity. The positive relationship between the locational Gini coefficient and the *update knowledge* variable suggests occupations that require workers to keep current with new information are associated with high levels of geographic concentration. Results from the baseline regression model also show that,

6 The Ellison–Glaeser concentration measure addresses this concern by incorporating information about the size distribution of firms in the industry (i.e. the Herfindahl index). In the case of occupations, information needed to calculate Herfindahl indices—namely, firm-level employment data—is not readily available.

Table 6. OLS regression results: effects of specialized knowledge on the geographic concentration of occupations

Variable	Estimated coefficients (robust standard errors in parentheses)				
Constant	0.485***	0.509***	0.344***	0.211	0.498***
	(0.112)	(0.138)	(0.116)	(0.143)	(0.115)
Specialized knowledge	0.041***	0.036**	0.046***	0.043**	0.040**
	(0.015)	(0.017)	(0.016)	(0.019)	(0.016)
Update knowledge	0.051**	0.032	0.025	0.062**	0.053**
	(0.020)	(0.022)	(0.023)	(0.027)	(0.020)
Machinery	−0.023***	−0.029***	−0.016*	−0.019**	−0.022***
	(0.007)	(0.011)	(0.009)	(0.009)	(0.008)
Interaction with public	−0.029***	−0.008	−0.020*	−0.020**	−0.025***
	(0.008)	(0.009)	(0.011)	(0.010)	(0.008)
Average establishment size	−0.017*	0.003	0.012	0.005	−0.016*
	(0.009)	(0.012)	(0.012)	(0.017)	(0.008)
Agriculture	0.172*	0.153	NA	0.170*	NA
	(0.099)	(0.106)		(0.091)	
Mining	−0.037	−0.010	NA	0.005	NA
	(0.050)	(0.057)		(0.039)	
Industry distribution	0.002	0.019	0.021	0.052***	−0.003
	(0.013)	(0.015)	(0.016)	(0.017)	(0.013)
Occupation size	−0.304***	−0.298***	−0.297***	−0.275***	−0.307***
	(0.007)	(0.008)	(0.007)	(0.009)	(0.007)
Dummy variables indicating major SOC category	No	Yes	No	No	No
Variables indicating the share of occupational employment by major NAICS category	No	No	Yes	No	No
Adjusted r^2	0.898	0.904	0.905	0.884	0.901
Sample	$n=433$	$n=433$	$n=433$	Excludes occupations with more than 50% employment in non-tradable industries ($n=287$)	Excludes occupations with more than 50% employment in agriculture and mining ($n=420$)

Notes: All variables except agriculture, mining and occupation (i.e. SOC dummies) and industry (i.e. NAICS employment shares) controls are measured in logs.
***, **, * denote significance at the 0.01, 0.05 and 0.10 levels, respectively.

surprisingly, jobs requiring the use of machinery exhibit low levels of geographic concentration, other things being equal. In addition, as expected, jobs that involve substantial interaction with the public tend to be geographically dispersed.

With respect to the variable *average establishment size*, the baseline regression results suggest that internal economies of scale do not appear to influence the geographic concentration of occupations. To explain this somewhat counterintuitive finding, we note that many of the jobs characterized by the largest average employment size fall in the major SOC categories of Healthcare Practitioners and Technical Occupations (SOC 29-0000) and Healthcare Support Occupations (SOC 31-0000). Hospitals—which employ a large proportion of workers in these occupations—tend to be large in size and geographically dispersed across metropolitan areas. Focusing on the remaining control variables used in the baseline regression model, we find that occupations steeped in agricultural-related industries, as well as those with small numbers of people, tend to be geographically concentrated.

The other results shown in Table 6 provide robustness checks of the estimates from our baseline analysis. The second column of results is from a regression model that includes a set of 21 dummy variables that indicate the occupation's major SOC category. Rosenthal and Strange (2001, 207) used a similar approach in their analysis of the geographic concentration of industries; however, they noted that 'the fixed effects potentially soak up much of the meaningful variation in the data'. The third column of results is from a model that includes a set of 18 variables that indicate the share of occupational employment by major industrial (e.g. NAICS) category. Whereas the fixed effects model controls for omitted factors that vary by broad occupational category, this version of the model accounts for missing variables related to the industries associated with the occupations.

The final two columns of Table 6 show results from regression models that selectively exclude occupations from the sample used in the baseline analysis. First, we removed occupations that have more than 50% of employment in 'non-tradable' industries. We used information from Jensen and Kletzer (2006), reported by major NAICS category, to identify industries categorized as non-tradable. Jensen and Kletzer (2006, 77) examined the geographic concentration of industries and occupations to determine 'activities that are potentially exposed to international trade'. Their approach is based on the logic that geographically concentrated activities are traded domestically and, thus, are 'potentially tradable internationally'. By focusing our analysis on occupations with 50% or less employment in non-tradable industries, we have omitted occupations that have geographically dispersed locational patterns in order to serve local residents.

It is also likely that occupations with a high percentage of employment in extractive industries have locational patterns dictated by the presence of natural resources (Ellison and Glaeser, 1999). Thus, in the final column of Table 6, we remove occupations with more than 50% of employment in agricultural- and mining-related industries. Whereas the fourth column of results focuses on occupations that are not dependent on the settlement patterns of people, the final set of results shown in the table examine the geographic concentration of occupations that are not dependent on the location of natural resources.

Regression results from the models used as robustness checks suggest, as was the case in the baseline analysis, that occupations with distinctive cognitive skills are more geographically concentrated than jobs with knowledge profiles that are similar to the norm. Estimated coefficients corresponding to the *specialized knowledge* variable are in the range of 0.036–0.046, which is similar to the estimate from the baseline model (i.e. 0.041). In the two right-hand columns of Table 6, the estimated coefficients corresponding to the other explanatory variables are generally similar to the baseline results. However, in the fixed effects model and the specification that accounts for the share of employment by major industrial category, several of the control variables (e.g. *update knowledge*, *interaction with public*, *agriculture*) have lower *t*-statistics than in the baseline regression analysis. This suggests, as anticipated by Rosenthal and Strange (2001), that the occupational fixed effects and industry controls remove meaningful variation in some of the factors expected to influence geographic concentration.

4.2. Instrumental variables estimation

The final set of regression results are from an IV model used to account for potential endogeneity between the locational Gini coefficient and the *specialized knowledge* variable. Focusing on industries, Ellison et al. (2010, 1196) employed a similar approach to mitigate the concern that 'industrial relationships may be the result of co-location instead of the cause of co-location'. In our analysis of occupations, it is possible that patterns of geographic concentration might influence the types of knowledge that are required in a job. Instrumental variables estimation provides a strategy to address this identification problem. However, implementing IV estimation requires that we identify instruments that are correlated with the extent to which an occupation's knowledge requirements differ from the norm (i.e. relevant) but unrelated to its level of geographic concentration (i.e. exogenous).

To construct our instrument set, we focus on the means by which people acquire knowledge to explain differences in the knowledge profiles among occupations. We use information on an occupation's experience, training and educational requirements—specifically, how these ways of obtaining knowledge differ from the average occupation—as instruments for the *specialized knowledge* variable. The experience and training variables measure the amount of work experience and job training required by someone in an occupation. For these job attributes, the O*NET survey includes a set of response categories, such as 'up to and including 1 month' and 'over 8 years, up to and including 10 years'. Using this information and the percentage of O*NET survey respondents who selected each category, we calculated the average amounts of experience and training required by workers. The education variable, calculated using data from the 2008 American Community Survey, is the percentage of individuals in an occupation with at least a 4-year college degree. For all three of these ways that people acquire knowledge, the variables enter into the first-stage regression model as the squared difference between the value for a given occupation and the average US job. The logic of our instrument set is that distinct knowledge profiles likely arise from distinct education, experience, or training; while the extent to which an occupation's education, experience and training differs from the norm is not directly related to the geographic concentration of occupations.

First-stage regression results presented in Table 7 indicate that, as expected, the *difference in education*, *difference in experience* and *difference in training* variables have a positive and statistically significant effect on the distinctiveness of an occupation's knowledge profile. To assess the strength of these instruments, we used the Stock and Yogo (2005) weak instrument test that compares the first-stage *F*-statistic to a critical

Table 7. IV Regression results: effects of specialized knowledge on the geographic concentration of occupations

Variable	Estimated coefficients (robust standard errors in parentheses)	
	First stage	Second stage
Constant	5.290***	−0.021
	(0.249)	(0.288)
Difference in education	0.073***	NA
	(0.016)	
Difference in experience	0.024**	NA
	(0.012)	
Difference in training	0.029**	NA
	(0.011)	
Specialized knowledge	NA	0.138***
		(0.053)
Update knowledge	0.546***	−0.0004
	(0.051)	(0.033)
Machinery	−0.022	−0.021***
	(0.025)	(0.008)
Interaction with public	−0.066***	−0.021**
	(0.024)	(0.009)
Average establishment size	0.042	−0.022**
	(0.033)	(0.009)
Agriculture	0.448**	0.136
	(0.186)	(0.101)
Mining	0.237	−0.063
	(0.183)	(0.053)
Industry distribution	0.019	−0.002
	(0.037)	(0.014)
Occupation size	−0.037**	−0.301***
	(0.016)	(0.007)
F-statistic for weak instrument test	13.303+	NA
p-value for over-identification test	NA	0.406
p-value for endogeneity test	NA	0.042

Notes: All variables except agriculture and mining are measured in logs. ***, ** denote significance at the 0.01 and 0.05 levels, respectively. IV estimates obtained using limited information maximum likelihood (LIML) estimator. The symbol '+' denotes we can reject the null hypothesis of weak instruments based on the Stock and Yogo test ($\alpha = 0.05$) using the 10% maximal LIML size threshold (i.e. critical value of 6.46 in a model with three instruments).

value that depends on the number of endogenous variables, the size of the instrument set, and the tolerance for the 'size distortion' of a test ($\alpha = 0.05$) of the null hypothesis that the instruments are weak.[7] We can reject the null hypothesis of weak instruments based on the Stock and Yogo (2005) test using a 10% maximal size threshold.

7 The size distortion tolerance (e.g. 10%) accounts for the idea that using the weakest combination of instruments might lead to a conclusion of biased second-stage estimates (from a Wald test), whereas using the entire group of instruments does not.

With the relevance criterion satisfied, we now consider the exogeneity of the instrument set. Our key identifying assumption here is that any relationship between the geographic concentration of occupations and the means by which people acquire knowledge occurs though the specialized knowledge variable. That is, firms and workers locate around a specific skill set and do not typically consider how the skills were acquired (Marshall, 1920; Krugman, 1991). While some recent theoretical papers point out some exceptions to this rule, either when firms bear the costs of training or when human capital investments are general to an industry (Matouschek and Robert-Nicoud, 2005; Almazan et al., 2007), the education, experience and training variables used in our analysis do not consider these more nuanced dimensions of skill acquisition. Additionally, since the variables we employ are measured as squared differences relative to the average US job rather than their actual levels, occupations that require exceptionally low or high amounts of human capital investments exhibit similar values for these variables. This means that, even if a relationship exists between the amount of education, training or experience required in an occupation and its geographic concentration, the construction of our instrumental variables breaks this link. Thus, it is plausible that our instrument set is exogenous. Indeed, Sargan over-identification test results, with a p-value of 0.406, indicate that the instruments are uncorrelated with the error term.[8] As our instrument set satisfies the relevance and exogeneity conditions, we conclude that they are valid.

Second-stage regression results presented in Table 7 suggest that the *specialized knowledge* variable has a positive and significant effect on the locational Gini coefficient when potential endogeneity is taken into account.[9] The estimated coefficient corresponding to the *specialized knowledge* variable is 0.138 in the IV model, which is considerably higher than our OLS estimates. The difference between the IV and OLS estimates suggests that there may be some measurement error in the *specialized knowledge* variable. However, as is typically the case, it is important to note that the precision of our estimates is reduced considerably in the IV model. The baseline OLS estimate of the effect of specialized knowledge on the geographic concentration of occupations is contained within the 95% confidence interval of the IV estimate. Nonetheless, a Wu–Hausman test indicates that the results from the IV model differ from those obtained using OLS. Overall, though, findings from the IV model diminish concerns that our main results are being driven by endogeneity between patterns of geographic concentration and the distinctiveness of an occupation's knowledge profile.

5. Conclusions

Researchers have long been interested in understanding the factors that influence the concentration of economic activity. Alfred Marshall's (1920, 225) ideas about labor market pooling, which suggest employers locate around 'workers with the skills which they require' and workers seek out places 'where there are many employers who need

8 This test of over-identifying restrictions is computed as $N \times R^2$, where N is the number of observations and the r-squared value is from a regression of the residuals from the second-stage regression on all of the exogenous variables and instruments. The test statistic is distributed χ^2 with degrees of freedom equivalent to the number of over-identifying restrictions.

9 We employ LIML for our instrumental variables regression analysis as Stock and Yogo (2005) demonstrate that it is superior to 2SLS in the presence of weak instruments.

such skill as theirs' emphasize the strong connection between geographic concentration and the specialization of work-related tasks. Focusing on the knowledge requirements of a wide variety of jobs, this paper presents new evidence on the importance of labor market pooling as a factor influencing the locational patterns of occupations. Specifically, our findings suggest jobs that draw from a specialized knowledge base are geographically concentrated, whereas occupations with generic knowledge requirements are more dispersed across US metropolitan areas.

This article extends the existing literature in several ways. First, our analysis considers the entire spectrum of the economy—ranging from service providers (e.g. waiters and waitresses) and medical professionals (e.g. registered nurses) to jobs that are closely associated with manufactured goods (e.g. shoe machine operators, tire builders). In contrast, previous studies on the geographic concentration of industries have generally focused on sectors within a major industrial sector, such as manufacturing (Ellison and Glaeser, 1997; Rosenthal and Strange, 2001; Ellison et al., 2010) or services (Kolko 2010). The broader approach taken in this study provides a more complete understanding of the factors that influence the geographic concentration of economic activity.

Second, this article focuses on the distinctiveness of a worker's knowledge profile as a key factor influencing the geographic concentration of occupations. Conceptual models proposed by Krugman (1991) and Overman and Puga (2010) predict benefits of labor market pooling in cases where firms use specific types of labor. Our analysis of occupations, unlike industries for which it is difficult to identify 'characteristics that are related to the specialization of the industry's labor force' (Rosenthal and Strange 2001, 205), allows us to develop a new measure of specialization. This gets at the heart of Marshall's argument about the benefits of labor market pooling. Such behavior is advantageous if firms need and workers possess a specialized knowledge base, whereas a high geographic concentration of activity is less important in occupations with generic knowledge requirements where suitable workers and jobs are easy to find.

Third, following Ellison et al. (2010), this article presents one of the first attempts to account for endogeneity between geographic concentration and, in our case, the knowledge required in a job. Our IV regressions utilize a set of valid instruments that represent the means by which people obtain knowledge (i.e. education, experience and training). The IV results, along with additional regressions that use different sub-samples of the data and that include 'tight controls' for occupational and industrial categories, confirm the robustness of our baseline regression results demonstrating a positive relationship exists between the locational Gini coefficients and the *specialized knowledge* variable.

Along with this key result, we found that several of the control variables provide evidence (although not consistent across all of the models estimated in the article) on the importance of occupation size, interaction with the public, establishment size and the process of updating knowledge as determinants of the geographic concentration of occupations. This represents what we believe to be one of the first attempts to examine these types of locational patterns. Numerous studies have looked at both the causes and consequences of the geographic concentration of industries. Thus, we have a developed a good understanding about why similar goods and services are produced within a close geographic proximity, and what these types of locational patterns mean for regional economic growth. What has been missing is an empirical analysis of the locational patterns of workers involved in similar job-related tasks and activities. Our work on this topic has helped to illuminate the importance of specialized knowledge as a key determinant of geographic concentration, which has been an illusive task in many past studies focusing on industries.

Acknowledgements

The authors would like to thank Henry Overman and three anonymous referees for their insightful comments. The views and opinions expressed here are solely those of the authors and do not necessarily reflect those of the Federal Reserve Bank of New York, the Federal Reserve System, or the University of Maine. Gabe's involvement in the project was supported, in part, by the Maine Agricultural and Forest Experiment Station. MAFES External Publication 3184.

References

Almazan, A., DeMotta, A., Titman, S. (2007) Firm location and the creation and utilization of human capital. *Review of Economic Studies*, 74: 1305–1327.
Andersson, R., Quigley, J., Wilhelmsson, M. (2005) Innovation, knowledge spillovers and local labour markets. *Papers in Regional Science*, 84: 446–464.
Audretsch, D., Feldman, M. (1996) R&D spillovers and the geography of innovation and production. *American Economic Review*, 86: 630–640.
Bacolod, M., Blum, B., Strange, W. (2009a) Skills in the city. *Journal of Urban Economics*, 65: 136–153.
Bacolod, M., Blum, B., Strange, W. (2009b) Urban interactions: soft skills vs. specialization. *Journal of Economic Geography*, 9: 227–262.
Barrios, S., Bertinelli, L., Strobl, E., Teixeira, A.-C. (2005) The dynamics of agglomeration: evidence from Ireland and Portugal. *Journal of Urban Economics*, 57: 170–188.
Devereux, M., Griffith, R., Simpson, H. (2004) The geographic distribution of production activity in the UK. *Regional Science and Urban Economics*, 34: 533–564.
Duranton, G., Overman, H. (2005) Testing for localization using micro-geographic data. *Review of Economic Studies*, 72: 1077–1106.
Duranton, G., Puga, D. (2004) Micro-foundations of urban agglomeration economies. In J. V. Henderson, J. F. Thisse (eds) *Handbook of Regional and Urban Economics*, vol. 4, Chapter 48, pp. 2063–2117. Amsterdam: Elsevier/North-Holland.
Ellison, G., Glaeser, E. (1997) Geographic concentration in U.S. manufacturing industries: a dartboard approach. *Journal of Political Economy*, 105: 889–927.
Ellison, G., Glaeser, E. (1999) The geographic concentration of industry: does natural advantage explain agglomeration. *American Economic Review*, 89: 311–316.
Ellison, G., Glaeser, E., Kerr, W. (2010) What causes industry agglomeration? Evidence from coagglomeration patterns. *American Economic Review*, 100: 1195–1213.
Fallick, B., Fleischman, C., Rebitzer, J. (2006) Job-hopping in Silicon Valley: some evidence concerning the microfoundations of a high-technology cluster. *Review of Economics and Statistics*, 88: 472–481.
Feser, E. (2003) What regions do rather than make: a proposed set of knowledge-based occupation clusters. *Urban Studies*, 40: 1937–1958.
Florida, R., Mellander, C., Stolarick, K. (2008) Inside the black box of regional development – human capital, the creative class and tolerance. *Journal of Economic Geography*, 8: 615–649.
Freedman, M. (2008) Job hopping, earnings dynamics, and industrial agglomeration in the software publishing industry. *Journal of Urban Economics*, 64: 590–600.
Gabe, T. (2009) Knowledge and earnings. *Journal of Regional Science*, 49: 439–457.
Ibrahim, S., Fallah, M. H., Reilly, R. (2009) Localized sources of knowledge and the effect of knowledge spillovers: an empirical study of inventors in the telecommunications industry. *Journal of Economic Geography*, 9: 405–431.
Jaffe, A., Trajtenberg, M., Henderson, R. (1993) Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics*, 108: 577–598.

Jensen, J. B., Kletzer, L. (2006) Tradable services: understanding the scope and impact of services offshoring. In L. Brainard, S. Collins (eds) *Brookings Trade Forum 2005: Offshoring White-Collar Work*, pp. 75–133. Washington, DC: Brookings Institution Press.

Kim, S. (1995) Expansion of markets and the geographic distribution of economic activities: the trends in U.S. regional manufacturing structure, 1860-1987. *Quarterly Journal of Economics*, 110: 881–908.

Kim, Y., Barkley, D., Henry, M. (2000) Industry characteristics linked to establishment concentrations in nonmetropolitan areas. *Journal of Regional Science*, 40: 231–259.

Kloosterman, R. (2008) Walls and bridges: knowledge spillover between 'superdutch' architectural firms. *Journal of Economic Geography*, 8: 545–563.

Kolko, J. (2010) Urbanization, agglomeration, and co-agglomeration of service industries. In E. L. Glaeser (ed.) *Agglomeration Economics*, Chapter 5, pp. 151–180. Chicago: University of Chicago Press for the NBER.

Krugman, P. (1991) *Geography and Trade*. Cambridge, MA: The MIT Press.

Lucas, R. (1988) On the mechanics of economic development. *Journal of Monetary Economics*, 22: 3–42.

Markusen, A. (2004) Targeting occupations in regional and community economic development. *Journal of the American Planning Association*, 70: 253–268.

Marshall, A. (1920) *Principles of Economics*. London: Macmillan.

Matouschek, N., Robert-Nicoud, F. (2005) The role of human capital investments in the location decisions of firms. *Regional Science and Urban Economics*, 35: 570–583.

Maurel, F., Sedillot, B. (1999) A measure of the geographic concentration in French manufacturing industries. *Regional Science and Urban Economics*, 29: 575–604.

McCann, P., Simonen, J. (2005) Innovation, knowledge spillovers and local labour markets. *Papers in Regional Science*, 84: 465–485.

Overman, H., Puga, D. (2010) Labour pooling as a source of agglomeration: an empirical investigation. In E. L. Glaeser (ed.) *Agglomeratin Economics*, Chapter 4, pp. 133–150. Chicago: University of Chicago Press for the NBER.

Peterson, N., Mumford, M., Borman, W., Jeanneret, R., Fleishman, E., Levin, K., Campion, M., Mayfield, M., Morgeson, F., Pearlman, K., Gowing, M., Lancaster, A., Silver, M., Dye, D. (2001) Understanding work using the occupational information network (O*NET): implications for practice and research. *Personnel Psychology*, 54: 451–492.

Rosenthal, S., Strange, W. (2001) The determinants of agglomeration. *Journal of Urban Economics*, 50: 191–229.

Ruggles, S., Alexander, J. T., Genadek, K., Goeken, R., Schroeder, M. D., Sobek, M. (2010) *Integrated Public Use Microdata Series: Version 5.0* [Machine-readable database]. Minneapolis: University of Minnesota.

Saxenian, A. (1994) *Regional Advantage: Culture and Competition in Silicon Valley and Route 128*. Cambridge, MA: Harvard University Press.

Scott, A. (2009) Human capital resources and requirements across the metropolitan hierarchy of the USA. *Journal of Economic Geography*, 9: 207–226.

Stock, J. H., Yogo, M. (2005) Testing for weak instruments in linear IV regression. In D. W. K. Andrews, J. H. Stock (eds) *Identification and Inference for Economic Models: A Festrschrift in Honor of Thomas Rothenberg*, pp. 80–108. Cambridge: Cambridge University Press.

Storper, M., Venables, A. (2004) Buzz: face-to-face contact and the urban economy. *Journal of Economic Geography*, 4: 351–370.