Chapter 49

EVIDENCE ON THE NATURE AND SOURCES OF AGGLOMERATION ECONOMIES

STUART S. ROSENTHAL

Department of Economics and Center for Policy Research, Syracuse University, Syracuse, NY 13244-1020, USA
e-mail: ssrosent@maxwell.syr.edu
url: http://www-cpr.maxwell.syr.edu/faculty/rosenthal/

WILLIAM C. STRANGE

Rotman School of Management, 105 St. George St., University of Toronto, Toronto, ON M5S 3E6, Canada
e-mail: wstrange@rotman.utoronto.ca
url: http://www.rotman.utoronto.ca/~wstrange/

Contents

Handbook of Regional and Urban Economics, Volume 4. Edited by J.V. Henderson and J.F. Thisse

DOI: 10.1016/S0169-7218(04)07049-2

Abstract

This paper considers the empirical literature on the nature and sources of urban increasing returns, also known as agglomeration economies. An important aspect of these externalities that has not been previously emphasized is that the effects of agglomeration extend over at least three different dimensions. These are the industrial, geographic, and temporal scope of economic agglomeration economies. In each case, the literature suggests that agglomeration economies attenuate with distance. Recently, the literature has also begun to provide evidence on the microfoundations of external economies of scale. The best known of these sources are those attributed to Marshall (1920): labor market pooling, input sharing, and knowledge spillovers. Evidence to date supports the presence of all three of these forces. In addition, there is also evidence that natural advantage, home market effects, consumption opportunities, and rent-seeking all contribute to agglomeration.

Keywords

agglomeration economies, productivity, external economies, microfoundations, urban growth

JEL classification: R0, O4, D2, C1

1. Introduction

The degree of concentration of economic activity is striking. Roughly 75% of Americans live in cities as defined by the Census Department, and yet cities occupy only 2% of the land area of the lower 48 states. A similar story could be told for any other developed county: labor and capital are both heavily concentrated in cities.

It is not just aggregate activity that is agglomerated. Individual industries are concentrated too. Figure 1, for instance, presents the density of employment in the furniture industry (SIC). Most of the country has almost no employment in the industry, as the map shows. The map also shows that the counties that do have employment are not randomly scattered across the U.S. They are disproportionately located in the western part of North Carolina and in other nearby locations. Clearly, furniture is an industry that makes use of particular raw materials, especially wood. Forestry is an important industry in North Carolina and elsewhere in the Southeast, so the location is sensible because of the access it offers to raw materials. But there are a lot of other equally sensible locations elsewhere in the county, from Maine to Oregon. Clearly, something beyond locating near raw materials sources is taking place.

The macro pattern of Figure 1 repeats itself in Figure 2, a map of the location of software producers (SIC 7371–7373 and 7375) in the vicinity of San Francisco. The map reports both the locations of existing establishments and the locations where new establishments are created (births). As can readily be seen, both are concentrated. In this case, there is no material input that is analogous to trees. Despite this, activity is highly concentrated in what is known as the Silicon Valley north of San Jose and in San Jose itself. Again, something is going on that is leading to this kind of geographic concentration.

This chapter will survey empirical work on the forces that lead to concentration, both of industries in clusters and of aggregate activity in cities. These forces are known variously as agglomeration economies or external economies of scale. In surveying the empirical work, the chapter will be concerned with two related questions: what is the nature and what are the sources of the increasing returns that produce agglomeration? In considering the nature of agglomeration economies we will be concerned with a number of smaller questions. Are they local, as seems to be the case in software, or do they operate at a regional scale, as seems to be the case for furniture? Are they restricted to individual industries like software and furniture, or are their effects comprehensive, extending across all activities? What is the dynamic nature of agglomeration economies? Are the effects of proximity felt immediately or does agglomeration have its positive effect on productivity only with a lag? Finally, are the effects dependent simply on the amount of activity that takes place somewhere, or is the nature of local interactions important to the process of agglomeration? All of these questions relate to what we will define as the scope of agglomeration economies. The empirical answers to these questions will be discussed together in Section 2.

The second broad question concerns the sources of agglomeration economies. Marshall (1920) suggests three. The first of these is the sharing of inputs whose production

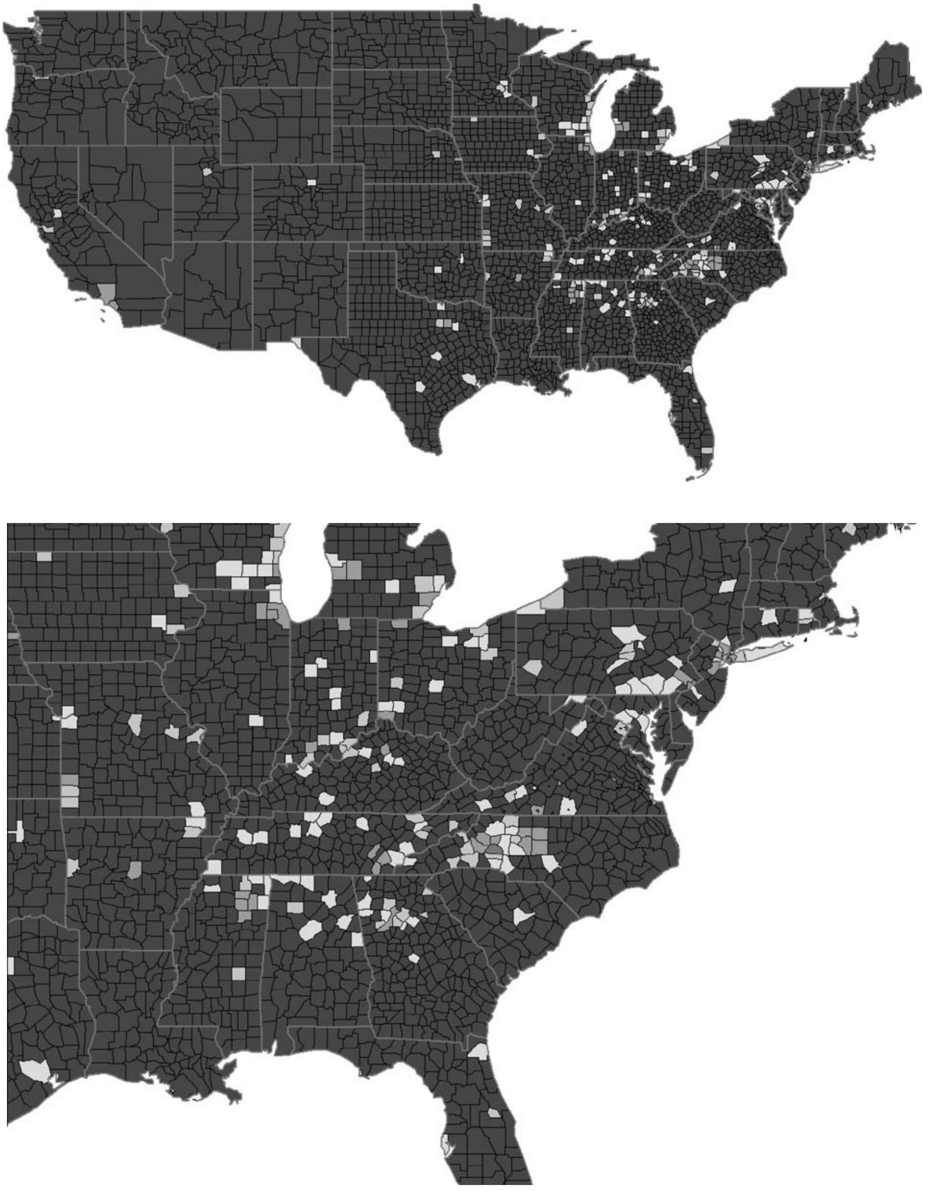

Figure 1. Furniture employment (SIC 25) per square mile. Fourth quarter 2002; source: Dun and Bradstreet. Red: greater than 10; orange: 4 to 10; dark yellow: 2 to 3; light yellow: 1 to 2; green: 0. For a colour reproduction of this figure see the colour figures section, page 3053.

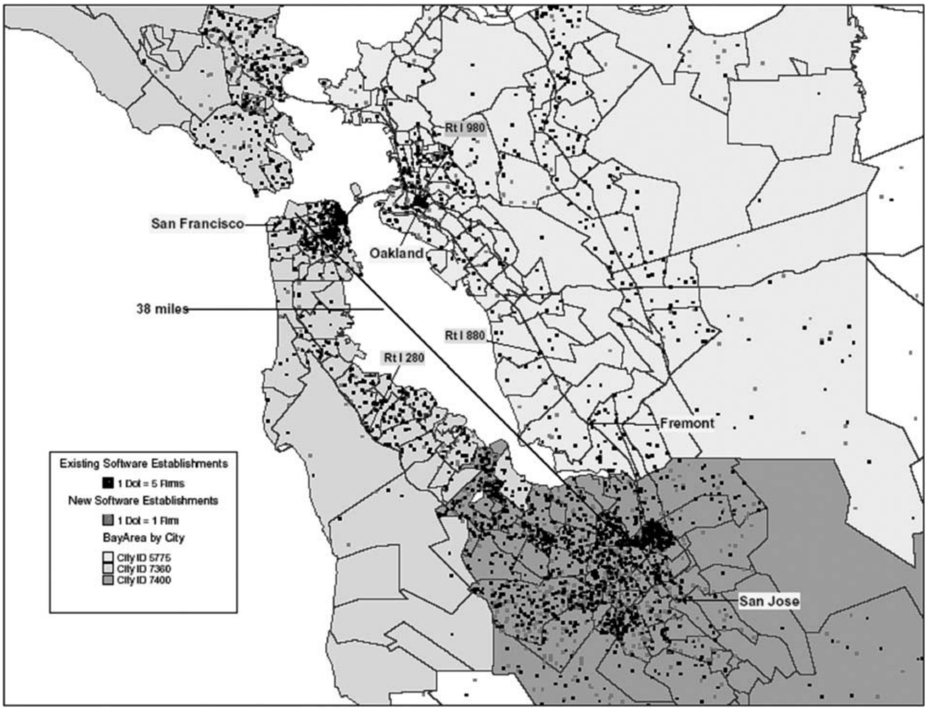

Figure 2. Existing and new software establishments in Silicon Valley. Fourth quarter 1997; source: Dun and Bradstreet. For a colour reproduction of this figure see the colour figures section, page 3054.

involves internal increasing returns to scale. The second is labor market pooling, where agglomeration allows a better match between an employer's needs and a worker's skills and reduces risk for both. The third source is spillovers in knowledge that take place when an industry is localized, allowing workers to learn from each other.[1] Other sources have been suggested more recently. These include home market effects, where the concentration of demand encourages agglomeration, and economies in consumption, where cities exist because people like the bright lights. On the negative side, it has also been suggested that agglomeration is related to rent-seeking, with inefficient mega-cities arising more frequently in undemocratic countries. This so-called urban primacy has many effects, with one being to redistribute the government's expropriated resources among the urban mob. Section 3 considers the empirical work that has addressed these issues.

Sections 2 and 3 review an econometric literature that is only about thirty years old. This literature has made substantial progress, especially in recent years as more refined

[1] In another chapter in this volume, Duranton and Puga (2004) propose a different taxonomy: matching, sharing, and learning.

data have become available. This has allowed researchers to ask questions that could not have been asked with more aggregate data. For example, evaluating the geographic extent of agglomeration economies is not possible without geographically refined data. Access to better data has also enabled researchers to answer old questions with greater precision, such as whether agglomeration economies are industry-specific or extend to the entire city. Despite the impressive record of progress of this program of formal econometric work, we believe there is much to be learned from less formal research. In Section 4, we consider some representative case studies. This is obviously a much older way to understand the facts that bear on agglomeration than through regression analysis. Even so, we believe it is an important part of the entire empirical story, both confirming and placing in context the formal empirical work and identifying important details in the big picture of agglomeration that the formal work misses.

We now turn to the scope of agglomeration economies.

2. The scope of urban increasing returns

2.1. Introduction

External economies exist when the scale of the urban environment adds to productivity. There are at least three dimensions over which these externalities may extend. We refer to the extent of the externality as its scope. The first and most familiar is the industrial scope. This is the degree to which agglomeration economies extend across industries, possibly even across all industries in a city, rather than being confined within industry boundaries. This distinction is well-known, with the economies of scale that arise from spatial concentration of activity within a given industry being known as localization economies. The externalities that arise from the concentration of all economic activity, or from city size itself, are known as urbanization economies. As will become apparent, empirical evidence in the literature suggests that as agents become closer in industrial space (i.e., their production processes become more similar), then there is greater potential for interaction.

The second kind of scope is geographic. Nearly every textbook in urban economics begins by explaining why cities exist. The answer is that proximity is advantageous. Thus, the discussion of agglomeration begins with the idea that geographic distance is crucial to understanding cites. The aspect of geographic distance that will matter most here is the attenuation of agglomeration economies with distance: if agents are physically closer, then there is more potential for interaction.

The third kind of scope is temporal. It is possible that one agent's interaction with another agent at a point in the past continues to have an effect on productivity in the present. For example, learning may take place only gradually, and awareness of a location's supply chain possibilities may take time to develop. Of course, such knowledge can decay over time. This means that in addition to the fairly well-known static agglomeration economies, there may also be dynamic agglomeration economies. That two agents who are separated temporally continue to affect each other is logically similar to the way that agents who are separated in physical or industrial space interact. The degree to which these time-separated interactions continue to be potent defines the temporal scope of agglomeration economies.

This section will examine recent empirical studies that shed light on each of these three aspects of the scope of external economies of scale. Table 1 provides a selective overview of the literature. We will begin by characterizing how one might proceed given a hypothetical "perfect" data set, free of measurement error, with no omitted variables, and including instruments that resolve all issues related to endogenous regressors. Against the backdrop of this ideal, we will discuss estimation strategies have been pursued in the presence of the imperfect data sets that actually are available. We then examine the evidence on the industrial, geographic, and temporal scope of agglomeration. Finally, we conclude the section by discussing empirical literature that sheds light on the manner in which the industrial organization and business "culture" of the local economy affects the generation and reception of external economies of scale.

Table 1
The scope of agglomeration economies

Issue	Paper	Key finding
Industrial scope	Moomaw (1981, 1983), Nakamura (1985), Calem and Carlino (1991)	Urbanization
	Nakamura (1985), Henderson (1986), Sveikauskas (1975), Henderson, Kuncoro and Turner (1995), Rosenthal and Strange (2001), Henderson (2003a)	Localization
	Glaeser et al. (1992)	Diversity
Temporal scope	Glaeser et al. (1992), Henderson, Kuncoro and Turner (1995), Henderson (1997)	Past agglomeration causes growth
	Glaeser and Maré (2001)	Lags in wage effects
Geographic scope	Rosenthal and Strange (2003), Ciccone (2002)	Attenuating effects
	Dekle and Eaton (1999)	National effects
	Ciccone and Hall (1996), Ciccone (2002)	Employment density
Organizational/ competitiveness	Glaeser et al. (1992), Henderson, Kuncoro and Turner (1995), Combes (2000b), Rosenthal and Strange (2003)	Competition
	Rosenthal and Strange (2003)	Small firms generate more
	Henderson (2003a)	Small firms receive more

2.2. Strategies for evaluating the scope of agglomeration economies

2.2.1. Context

External economies are by definition shifters of an establishment's production function. The first issue that must be confronted is whether the effect is Hicks neutral, or whether it augments labor or some other input in the production function. We will suppose the change to be neutral, consistent with empirical evidence from Henderson (1986). Given the Hicks neutrality assumption, an establishment's production function may be written as $g(A)f(x)$, where x is a vector of the usual inputs (land, labor, capital, and materials) and A characterizes the establishment's environment. The latter allows for the influence of agglomeration.

A general specification of agglomeration economies is that the aggregate urban external effect arises as the sum of a large number of individual externalities. We will treat the externalities as being between establishments, although they could instead be between individuals. Consider two establishments, j and k. The effect of establishment k on establishment j depends on the scale of activity at both establishments. In addition, the impact of k on j also depends on the distance between the two establishments, where distance is measured over three different dimensions. First, the influence of j on k depends on the geographic distance between the two establishments, d^{G}_{jk}. Second, it also depends on the type of industrial activity that takes place at the two establishments. It is natural to refer to this as the industrial distance between j and k, denoted here as d^{I}_{jk}. Two establishments carrying out the same kind of production would have $d^{\mathrm{I}}_{jk} = 0$, and d^{I}_{jk} would increase as the production processes become more dissimilar. Third, the impact of the interaction may extend temporally. At any point in time, establishment j may currently benefit from interaction with establishment k at some point in the past. This temporal dimension of distance is denoted d^{T}_{jk}. For example, for an interaction two years ago, d^{T}_{jk} would equal two.

An increase in any of these kinds of distance – spatial, industrial, or temporal – presumably leads to the attenuation of the agglomerative effect of establishment k on establishment j's production function. Formally, let the set of establishments with which establishment j might possibly benefit from interacting with be defined as K. Assume that all benefits to j from interaction with establishment $k \in K$ equal $q(x_j, x_k)a(d^{\mathrm{G}}_{jk}, d^{\mathrm{I}}_{jk}, d^{\mathrm{T}}_{jk})$. The first expression, $q(x_j, x_k)$, reflects benefits from interaction that depend on the scales of j's and k's activities, denoted by their input vectors x_j and x_k. For example, it is common to suppose that the strength of the interaction is captured by the size of establishment k's workforce, with other characteristics of establishment k having no effect. The second expression captures the attenuation of the interaction as establishments become more distant from each other. Specifically, holding the scale of the interaction constant, the benefit of an interaction with establishment $k \in K$ at geographic distance d^{G}_{jk}, industrial distance d^{I}_{jk}, and temporal distance d^{T}_{jk} is defined as $a(d^{\mathrm{G}}_{jk}, d^{\mathrm{I}}_{jk}, d^{\mathrm{T}}_{jk})$. The total benefit of agglomeration enjoyed by establishment j is then equal to the sum over interaction partners of the agglomerative effect as a function of geographic, industrial, and temporal distance:

$$A_j = \sum_{k \in K} q(x_j, x_k) a\left(d^{\mathrm{G}}_{jk}, d^{\mathrm{I}}_{jk}, d^{\mathrm{T}}_{jk}\right). \tag{2.1}$$

The construction of (2.1) immediately suggests some issues that bear on the estimation of agglomeration economies. The first is that A varies across establishments because each belongs to a given industry and is situated at a unique location over a particular period of time. The second issue is that each dimension of agglomeration economies could in principle be measured continuously. This would require some attempt to capture the attenuation of agglomeration economies as establishments move farther apart, both in the standard sense of physical space but also in the more novel sense of industrial and temporal space.

It is fair to say that relatively little of the empirical work on the scope of agglomeration economies has addressed the issues of establishment uniqueness and continuity. Instead, with regard to geography, most studies have grouped industries and plants into politically defined regions such as Metropolitan Statistical Areas (MSAs) or counties. Activity in neighboring regions is then typically assumed, usually implicitly, to have no effect on the region in question, and all activity within the specified region is treated as being situated at exactly the same spot. With regard to the type of industrial activity, most studies have collapsed industrial activity into just two broad categories: activity within an establishment's industry (i.e., SIC code) and activity outside of the establishment's industry. This, of course, does not capture the possibility that some industries belonging to different industry categories are close cousins, while others are hardly related at all.[2] With regard to temporal dimensions of agglomeration, several studies have considered the influence of time, but most have not.

Assuming that A_j could be fully specified and measured without error, the equation to be estimated is

$$y_j = g(A_j) f(x_j). \tag{2.2}$$

y_j is establishment j's output, x_j represents j's traditional inputs and A_j is given in (2.1). In principle, estimates of Equation (2.2) would provide measures of the productivity effects of the industrial, spatial, and temporal dimensions of agglomeration. In practice, attempts to estimate (2.2) face many challenges. We will now set out the challenges in detail.

2.2.2. Measuring the scope of agglomeration

In order to estimate an approximation to Equation (2.2), measures of A must first be constructed that correspond to the three dimensions of the scope of agglomeration economies. Thus, for a given geographic distance from establishment j, measures of

[2] Ellison and Glaeser (1997) examine exactly this issue when they construct measures of co-agglomeration.

A should ideally include the amount of economic activity present in a variety of different industries at different distances in industrial space from j. This would allow one to determine the industries that benefit from proximity. Including measures of physical distance would allow one to determine how close establishments need to be in order to benefit from their agglomeration. Finally, it would also be desirable to allow for dynamic externalities and consider the impacts of historic activity. Obtaining all these controls is a daunting challenge. Thus, most models of agglomeration bear on one or perhaps two of the key aspects of scope, but never all three.

2.2.3. *Estimating the production function: omitted variables and simultaneity*

The most natural way to understand agglomeration economies is to directly estimate the production function, (2.2). In carrying out this estimation, it is necessary to have measures of the various elements of x_j, including employment, land, capital, and materials. Labor inputs are perhaps the easiest to measure, since many data sets provide counts of workers, hours worked, and on occasion, proxies for skill level (e.g., education). Data on purchased materials are available in some data sets, but data on materials produced internally typically are not. See Ciccone and Hall (1996) and Henderson (2003a) for discussions of this issue. Few data sets make available measures of land use and information on the stock of capital, information essential to estimating (2.2). Thus, a fundamental challenge that must be faced in estimating a production function is in finding data on inputs.

The issue of measurement error has been central to the literature since the outset. Because this is an old issue and one that has already been surveyed with considerable care [Eberts and McMillen (1999)], our treatment will be relatively brief. First, it is clear that the absence of data on capital can affect the estimates. For instance, Sveikauskas (1975) lacks data on capital. As Moomaw (1983) points out, however, if capital is used more intensively in large cities, then the error terms will be positively correlated with the city size terms, leading to upward bias in coefficient estimates. In fact, Moomaw shows that this can inflate estimates by a factor of four.[3] Second, land is also an important input, and its contribution to production is also difficult to measure. Land will be used less intensively in large cities, so presumably this omission would lead to downward bias in the estimates.

A more recent effort to estimate (2.2) directly is Henderson (2003a). We believe that this paper is a model of a productivity-based study of agglomeration, coming closest to the ideal that we discussed at the beginning of the section. In this paper, Henderson constructs a panel of plant-level data from the Longitudinal Research Database (LRD) including measures of the capital stock, materials, and labor. Using the LRD's microdata, Henderson controls for industrial scope in the usual way by dividing activities

[3] A related literature considers the impact of public infrastructure (i.e., roads and bridges) on productivity. See Holtz-Eakin (1994). These studies also wrestle with measuring private capital.

into those that take place within a given industry and those that do not. Henderson also draws on the panel structure of the data to address issues related to the temporal scope of agglomeration. For the most part, Henderson considers county and MSA-level indicators, rather than using variables that directly reflect proximity. An exception to this is some analysis of neighboring counties. While Henderson's work is also noteworthy for the careful treatment of the data, the strength of the empirical work rests primarily with the use of plant-level information and detail on purchased factor inputs available from the confidential LRD files. While these data appear to offer some of the best opportunities for making contributions to the understanding of agglomeration, access to them is tightly guarded. This means that many researchers choose to work with other less ideal data.[4]

Even when plant-level data are available, direct estimation of equations such as (2.2) requires that the analyst address challenging endogeneity problems. Agglomeration economies enhance plant productivity, but successful entrepreneurs also seek out productive locations. If overachieving entrepreneurs were disproportionately found in agglomerated areas, this would cause one to overestimate the relationship between agglomeration and output. Henderson initially attempts to address this problem through two-stage least squares (2SLS) using local environment measures as instruments. The instrument list includes cross-sectional MSA attributes such as the market potential of the MSA, county air quality attainment status and other variables thought to be strictly exogenous. However, Henderson notes that these regressors make weak instruments, rendering the 2SLS approach ineffective.[5]

Next, Henderson (2003a) estimates a version of (2.2) drawing on the panel structure of his data and imposing constant slope coefficients over time. Time-differencing the data, he estimates this system by generalized method of moments (GMM) using predetermined industry environment variables as instruments (e.g., lagged levels of different types of local employment). Once more, however, Henderson finds that the instruments are weak, though not as weak as the cross-sectional instruments for the 2SLS model. In addition, by using predetermined data for instruments in conjunction to differencing the data over time, he is forced to dramatically reduce the sample over which the estimation is conducted.

After experimenting with both 2SLS and GMM, Henderson concludes that controlling for endogeneity through the use of fixed effects is superior. Specifically, he estimates his productivity equation including MSA-time specific fixed effects in addition to plant fixed effects. By adding the MSA-time fixed effects the hope is that this will capture the influence of unobserved attributes that might have drawn a given entrepreneur

[4] In order to gain access to the LRD data researchers must become sworn "employees" of the U.S. Census and conduct their research in a secure room at one of the Census research stations set up for such purposes. Census research stations are currently found in Washington, DC, Boston, Pittsburgh, Chicago, and San Francisco. In addition, access to the confidential Census files is costly and requires a level of funding typically only available from a major grant.

[5] See Hanson (2001) for more on the endogeneity issue.

to the area and that might otherwise be correlated with the error term in the estimating equation. Including MSA-time specific fixed effects is appealing and may well be one of the most effective ways to address the endogenous nature of the local industrial environment. Nevertheless, even this approach may still be exposed to endogeneity problems because the presence of a plant in a given MSA and time period represents the outcome of a profit-maximizing choice.

2.2.4. Indirect strategies for measuring the influence of agglomeration on productivity

Estimating the production function directly is not the only way to look for evidence of agglomeration economies. Because of the challenges associated with that approach, many recent studies have begun to favor one of four indirect approaches.

The first of these is to consider growth. Glaeser et al. (1992) and Henderson, Kuncoro and Turner (1995), for example, examine the impact of MSA-level agglomeration on employment growth. In the case of Glaeser et al. (1992), growth is measured using data from the County Business Patterns while Henderson, Kuncoro and Turner (1995) rely on the Census of Manufactures. The idea here is that agglomeration economies enhance productivity and productive regions (e.g., MSAs) grow more rapidly as a result.

Studying the growth of total employment presents different challenges than estimating productivity directly. Data on total employment are often readily available and the analysis lends itself to linear regressions. However, existing employers are constrained by prior choices, most importantly the level and kind of capital previously installed. Those fixed factors affect how the employer values the marginal worker, and consequently how it changes its employment level in response to a change in its environment. In principle, this difficulty can be overcome by looking at changes in total employment over a sufficiently long time frame so that there are no fixed factors and all establishments are effectively new. Even then, one still has to address endogeneity problems: not only is the growth of total employment in a given area sensitive to the composition of employment in the area (an agglomeration effect), but growth affects the level and composition of employment. Implementing this approach, therefore, ideally requires a long panel and effective instruments to control for endogenous variables. The primary approach used to address this problem in the Glaeser et al. (1992) and Henderson, Kuncoro and Turner (1995) papers is to use deeply lagged levels of past conditions of the MSAs as regressors.[6]

A different approach to studying the scope and effect of agglomeration on productivity has been to focus on births of new establishments and their employment. This approach was taken by Carlton (1983) and by Rosenthal and Strange (2003). The idea here is that entrepreneurs seek out profit-maximizing locations and are disproportionately drawn to the most productive regions. As with the other approaches, focusing on

[6] Glaeser et al. (1992) use 1956 employment levels to help explain growth over the 1956 to 1987 period. Henderson, Kuncoro and Turner (1995) use 1970 employment levels to help explain growth over the 1970 to 1987 period.

births has both advantages and disadvantages. On the positive side, data on purchased factor inputs (e.g., capital stock, labor, materials, and land) are not required, new establishments are largely unconstrained by previous decisions, and new establishments make their location and employment decisions taking the existing economic environment as exogenously given.

Studying plant births also presents difficulties. The principal drawback is that many locations do not receive any births in a given period, which can lead to technical challenges on the econometric side. In addition, births are more likely to occur in areas where there is already an existing concentration of industrial activity as spinoffs. Rosenthal and Strange (2003) control the zeros problem by using tobit models and comparing results to those from probit models that look for positive versus zero births. In addition, Rosenthal and Strange (2003) control for "churning" effects by studying zipcode level employment data and including MSA fixed effects as control variables. Even if an entrepreneur is tied to the local MSA because of past employment and other factors, the entrepreneur will still seek out the profit maximizing location within the MSA.

The third approach used to examine the scope and influence of agglomeration is to study wages. This approach rests on the assumption that in competitive markets labor is paid the value of its marginal product. Even without perfect competition, in more productive locations, wages will therefore be higher. Recent examples of this approach include Glaeser and Maré (2001) and Wheaton and Lewis (2002). An advantage of this approach is that wage data are readily available. Moreover, by focusing on wages this makes feasible the use of a variety of widely available datasets, such as the public access version of the Census, the Consumer Population Survey (CPS), and various panel studies including the Panel Study of Income Dynamics (PSID). Here too important issues associated with omitted variables and endogenous regressors arise. Glaeser and Maré (2001) provide a particularly careful discussion of these problems, a discussion to which we will return later in this section.

The final approach is to use rents. The idea here stems from the quality-of-life literature [Rosen (1979) and Roback (1982)]. If firms are paying higher rents in a particular location all else equal, then the location must have some compensating productivity differential.[7] Dekle and Eaton (1999) use this approach to measure agglomeration economies in Japan. One difficulty with implementing this approach is finding reasonably refined data on rents. For instance, Dekle and Eaton are forced to use housing rent data as a proxy for commercial and industrial rents.

An additional and potentially tricky issue in looking at agglomeration through wages or rents is that in theory, productivity differentials should be capitalized in both wages and rents. The degree they are captured in one or the other depends on elasticities in the markets for land and labor and also on the presence of other local attributes like natural amenities. Thus, although positive evidence of wage or rent capitalization is evidence

[7] See Blomquist, Berger and Hoehn (1988), Gyourko and Tracy (1991), and Gabriel and Rosenthal (2004) for empirical work in this area.

of the existence of agglomeration economies, the absence of evidence of capitalization into one of the two is not evidence of the absence of agglomeration economies. For instance, if households prefer big cities because of amenities associated with big city life, this will work to raise rents and reduce wages in big cities. If firms find big city workers to be more productive, this works to raise wages and rents. If the household amenity effect is sufficiently large, this will lead to lower wages and higher rents in big cities despite the existence of agglomeration economies. Of course, the empirical relevance of this point depends on the degree to which amenity and firm productivity effects are correlated. If firms and households care about different things, for example, if firms care about proximity to firms in the same industry and households care about having a baseball team, then this problem would not arise.

2.3. Industrial scope

2.3.1. Urbanization or localization economies

The oldest debate in the empirical literature on agglomeration economies concerns industrial scope. Specifically, the debate concerns whether agglomeration economies are related to the concentration of an industry or to the size of a city itself, regardless of competition. In other words, the debate concerns the relative importance of localization and urbanization economies.

The earliest precise discussion of the microfoundations of agglomeration stemming from localization is in Marshall's (1920). Over a hundred years later, the analysis remains fresh:

> When an industry has thus chosen a locality for itself, it is likely to stay there long: so great are the advantages which people following the same skilled trade get from neighborhood to one another. The mysteries of the trade become no mysteries; but are as it were in the air, and children learn many of them unconsciously ... Employers are apt to resort to any place where they are likely to find a good choice of workers with the special skill which they require ... The advantages of variety of employment are combined with those of localized industries in some of our manufacturing towns, and this is a chief cause of their continued economic growth. [Marshall (1920, 271)]

Famous modern examples of highly localized industries include computers (Silicon Valley, Route 128 in Boston) and carpets (Dalton, Georgia), but also less well-known concentrations such as that of furniture manufacturing in High Point North Carolina as discussed in the Introduction.

An equally influential discussion of microfoundations related to the industrial scope of spatial concentration is found in Jacobs (1969). In contrast to Marshall's treatment of urban specialization, Jacobs stresses the importance of urban diversity. Her argument is that diversity fosters cross-fertilization of ideas. For instance, she notes that New

York's brassiere industry evolved not from the lingerie industry but from dressmakers' innovations.

This emphasis is clearly different than Marshall's, and this had led to the discussion of localization vs. urbanization being characterized as a contest between Marshall and Jacobs. This is not completely fair to Marshall, who explicitly recognized the value of urban diversity, both as a way to achieve domestic complementarity and to reduce risk:

> On the other hand a localized industry has some disadvantages as a market for labour if the work done in it is chiefly of one kind, such for instance as can be done only by strong men. In those iron districts in which there are no textile or other factories to give employment to women and children, wages are high and the cost of labour dear to the employer, while the average money earnings of each family are low. But the remedy for this evil is obvious, and is found in the growth in the same neighbourhood of industries of a supplementary character. Thus textile industries are constantly found congregated in the neighbourhood of mining and engineering industries, in some cases having been attracted by almost imperceptible steps; in others, as for instance at Barrow, having been started deliberately on a large scale in order to give variety of employment in a place where previously there had been but little demand for the work of women and children ... A district which is dependent chiefly on one industry is liable to extreme depression, in case of a falling-off in the demand for its produce, or of a failure in the supply of the raw material which it uses. This evil again is in a great measure avoided by those large towns or large industrial districts in which several distinct industries are strongly developed. If one of them fails for a time, the others are likely to support it indirectly; and they enable local shopkeepers to continue their assistance to workpeople in it. [Marshall (1920, pp. 273–274)]

Various studies have attempted to identify the impact of urbanization economies. As above, because this older work is nicely reviewed elsewhere [Eberts and McMillen (1999)], our discussion of this issue will be selective. Shefer (1973) considered a cross-section of MSAs and a group of industries, concluding that doubling city size would increase productivity by between 14 and 27%. Sveikauskas (1975) found only an increase of 6–7%, which is more in line with later work in this area. Segal (1976) improved on the capital stock measures of the earlier studies, and found that productivity was roughly 8% higher in cities with populations of two million or more. Fogarty and Garofalo (1978) find an increase in productivity of about 10% when city population is doubled. Moomaw (1981) finds the increase to be 2.7%. Tabuchi (1986) finds that doubling population gives a 4.3% increase in productivity. In sum, doubling city size seems to increase productivity by an amount that ranges from roughly 3–8%.[8]

The relative impact on productivity of localization and urbanization together are examined by Nakamura (1985) and Henderson (1986). Nakamura considers Japan, while

[8] Combes, Duranton and Gobillon (2003) find slightly smaller estimates after controlling for worker skills in a study using French wage data. Cingano and Schivardi (2003) also find a smaller estimate.

Henderson considers the U.S. and Brazil.[9] Both estimate production functions separately for two-digit manufacturing industries. Urbanization is proxied by total employment in the city. Localization is proxied by employment in the industry. While there is evidence of urbanization economies in several industries, there is evidence of localization economies in more. Some industries exhibit no evidence of external economies at all. Nakamura summarizes his work as finding that a doubling of industry scale leads to a 4.5% increase in productivity, while a doubling of city population leads to a 3.4% increase. Henderson finds almost no evidence of urbanization economies and substantial evidence of localization. Taken together, Henderson and Nakamura are more favorable to the existence of localization economies than urbanization economies. In addition, the variation across industries strongly suggests that one ought to estimate agglomeration economies separately for different industries.

Other papers have considered both localization and urbanization together. Moomaw (1983) finds evidence of both. In a births model, Rosenthal and Strange (2003) consider the impacts of own-industry and total employment in an analysis that directly considers the geographic scope of agglomeration economies. They find stronger evidence of localization. Most recently, Henderson (2003a) also finds localization effects to be strongest.

2.3.2. *Specialization and diversity*

There are other ways to specify industrial scope. One is to consider the degree to which a city's employment is specialized. This is typically measured as the share of a city's employment in a particular industry [i.e., Glaeser et al. (1992) and Henderson, Kuncoro and Turner (1995)]. Glaeser et al. consider growth over the period 1956–1987 of industries that were among a city's top six in 1956. All industries are grouped together in the estimation. They find that specialization does not encourage growth. Henderson et al. also consider growth, in this case from 1970–1987. They estimate separately for eight industries, three that were rapidly evolving high-technology industries during the period covered by the data and five that were mature industries with stable technologies. For the high-technology industries, they also find specialization to lack a positive effect on growth. For the mature industries, in contrast, they find a positive effect of specialization. This result is parallel to Duranton and Puga (2001a), who use French data to show that while new industries evolve in diverse cities, they move to specialized ones after reaching maturity.

Combes (2000b) argues that this specialization analysis must be viewed with some caution. The issue is in the interpretation of the specialization variable. Henderson, Kuncoro and Turner (1995) note that an increase in specialization holding sectoral employment constant leads to more growth among the mature industries. Combes notes that

[9] It is interesting to note that Henderson's (1986) use of the refined Brazilian data was made possible by a mistake where data that should have been censored were not. This suggests that obtaining data that gets close to that required for ideal estimation has at least some element of luck to it.

the only way that specialization could be greater holding sectoral employment constant would be for the entire city to become smaller. Thus, the result that specialization encourages growth could be interpreted as saying that small cities grow more quickly. Controlling for total city employment instead of sectoral employment gives an entirely different picture, one where specialization fails to encourage growth.

Another issue that complicates interpretation of specialization variables concerns absolute versus relative effects. Explicit theories of the microfoundations of agglomeration economies have nearly always been based on the idea that an increase in the absolute scale of activity has a positive effect. For instance, more workers allows better matching in Helsley and Strange (1990). This sort of model suggests that an increase in the scale of an industry increases productivity. It does not make direct predictions regarding the impact of the industry's share of employment in a particular city or regarding the city's share in the industry relative to other cites. The city share of employment variable is typically introduced as a kind of net effect. The absolute scale of own-industry activity has a positive effect if localization economies are at work, but if the city's share of employment in the industry is small, that means that there is a lot of other activity in the city. Through congestion, this could have a negative effect. Presumably, the relative share variable matters in the context of a firm selecting the location that has the greatest net effect as described above.[10]

The flip side of specialization is diversity, a different way to conceive of an urbanization economy. Considering the diversity of employment rather than simply city size is very much in the spirit of Jacobs, whose many stories capture the idea that a city's diversity of activity can breed cross-fertilization of technology and so lead to innovation and growth. The idea of diversity being important was developed prior to Jacobs by Chinitz (1961), who argues that its diversity was one reason that New York performed better than did Pittsburgh after World War II. The issue of diversity has been addressed in various ways in the literature.

As noted above, Glaeser et al. (1992) consider the determinants of the growth of the largest six industries within a given MSA. To evaluate what they call "Jacobs externalities", they control for the fraction of MSA employment in the MSA's 7th through 12th largest industries. A large MSA employment share in the 7th through 12th largest industries in the MSA is indicative of a diverse industrial base. As an alternative, Henderson, Kuncoro and Turner (1995) use a Herfindahl–Hirschman index (HHI) of employment diversity at the MSA level. As usual, HHI measures the aggregate difference between a city's sectoral employment pattern and one that would arise if employment were uniformly distributed. Rosenthal and Strange (2003) also use a Herfindahl measure calculated in the same manner, except their measure is based on zipcode level employment and they also include MSA fixed effects in their models. The MSA fixed effects capture the influence of broader MSA-wide diversity effects, while the Herfindahl controls

[10] We discuss specialization further in the context of labor market pooling in Section 3.

for within-MSA variation in employment diversity. These measures all capture the absolute level of employment diversity in a given region, regardless of whether that region is defined based on MSA boundaries, zipcodes, or other geographic dimensions.

The results of these papers are fairly consistent. Glaeser et al. (1992) find that diversity encourages growth. Rosenthal and Strange (2003) find that diversity encourages births. Henderson, Kuncoro and Turner (1995) find diversity to encourage growth among high-technology firms. Given the different specifications and the variety of industries considered, these results together strongly suggest that diversity is helpful. This is consistent with Jacobs, but it is not inconsistent with Marshall. In any case, the importance of the diversity of the rest of the city's employment does not itself rule out a parallel effect associated with the concentration of employment in a particular industry.

2.3.3. *Other work on industrial scope*

The issue of industrial scope has been considered extensively. In addition to the growth studies mentioned above, Combes (2000a) considers the effects of industrial scope on growth in France over the period 1984–1993. As he persuasively points out, it is worthwhile to consider agglomeration economies separately across countries. In France, labor mobility is lower than in the U.S., as it is in the rest of Europe. Further, European unemployment rates are higher. Either of these could impact the urban development process, and so it makes sense to analyze the process of agglomeration separately in the European environment. Combes finds that there is a substantial difference between the results for service industries and manufacturing. For manufacturing, specialization and diversity both have negative impacts on growth in all but a few sectors. For services, specialization continues to have a negative effect, but the effect of diversity becomes positive. It is worth pointing out that Combes' data includes 341 employment areas that cover the whole territory of France. This means that there is less likely to be a selection problem than in papers that have confined their analysis to urban areas only. This point is also made in Rosenthal and Strange (2003).

A different approach is taken by Wheaton and Lewis (2002). They identify an urban wage premium associated with both an increase in the concentration of own-industry employment in the city and with an increase in the specialization of the city in the industry. Agglomeration economies have been studied by looking at rents by Dekle and Eaton (1999), who look only at the effects of aggregate activity. Hence, they provide evidence of urbanization.

2.3.4. *Continuity and industrial scope*

In conclusion, it should be emphasized once more that, relative to the ideal approach outlined earlier, most papers to date have adopted restrictive treatments of industrial scope. The standard approach has been to distinguish between activity in an establishment's own industry from activity outside of the own industry. Thus, the question of the rate at which agglomeration economies attenuate as nearby activity becomes increasingly dissimilar is virtually unexplored. This is not surprising. While the concept of industrial distance has some useful parallels to geographic distance, the analogy is not exact. That is because the way to measure the geographic distance between two locations is clear. The distance between two industries is not. The closest that anyone comes to defining industries in a way that captures industrial distance is in the Cluster Mapping Project described in Section 4. That project defines 41 clusters of "related" industries based on supply relationships, the similarity of production, and so forth. However, the exact algorithm by which clusters are defined is proprietary, which is obviously a serious obstacle to economists using this sort of procedure.

The paper in the economics literature that makes the most substantial progress in characterizing industrial distance is Ellison and Glaeser (1997). They characterize the extent of co-agglomeration among two digit industries, showing that there are many instances of industries apparently affecting each other. They also consider the forces that govern co-agglomeration, showing that when there are upstream–downstream linkages, co-agglomeration is greater. In sum, Ellison and Glaeser (1997) provide strong evidence suggesting further consideration of industrial distance to be warranted.

A recent paper by Duranton and Puga (2001b) also makes some progress in this direction. They show that increasingly cities are organized not so much around traditional industrial classifications, but instead along functional lines. Specifically, they show that cities emphasize managerial and information oriented activity of the type that benefits from face-to-face contacts. Thus, the era of the one-industry town may be waning.[11]

2.4. Geographic scope

Until very recently the standard approach has been to define geography based on political boundaries such as states, MSAs, and counties. Establishment and industrial activity is then grouped within these locations treating all entities within a given location as being located at precisely the same spot. In addition, researchers have typically assumed, usually implicitly, that activity outside of a location has no effect on activity within the location. In other words, spatial lags are ignored. Studies that fit this characterization include those of Glaeser et al. (1992) and Henderson, Kuncoro and Turner (1995).

An important departure from this tradition is Ciccone and Hall (1996). In explaining state-level labor productivity, they include measures of county-level employment density. Since the theory of agglomeration is almost entirely concerned with density,

[11] It is worth pointing out that most of the papers that we have studied in this review consider the agglomeration of manufacturing industries. This is problematic given the central role of the service sector in the economies of cities today. Kolko (1999) helps to remedy this shortcoming by considering the agglomeration of service industries. He finds that the need for the sort of labor found in cities is a continuing agglomerative force, with this access continuing to be valuable even with improvements in information technology that allow communication at a distance.

this approach is welcome. They find a positive effect of density, with a doubling associated with roughly a 5% increase in productivity. Of course, states are large, and this approach implicitly supposes that an establishment in Buffalo is closer to one in New York City than would be an establishment in northern New Jersey. In a follow-up paper based on European data, Ciccone (2002) finds effects that are only slightly smaller than in the U.S. He finds that the elasticity of labor productivity with respect to employment density is 4.5% in Europe, compared to 5% in the U.S.

Dekle and Eaton (1999) use rents to consider agglomeration economies using Japanese prefecture-level data. For both finance and manufacturing, they find evidence of agglomeration economies, although the magnitude is roughly one-quarter of the Ciccone and Hall (1996) estimate. They also find evidence that an increase in activity across all of Japan increases productivity in any prefecture, which suggests a large geographic scope.

More recently still, Rosenthal and Strange (2003) provide a micro-level analysis of the geographic scope of agglomeration economies. The environment of an establishment is measured by constructing rings around the centroid of the establishment's zipcode. Rings of 1, 5, 10, and 15 miles are included. For five of the six industries studied (computer software, apparel, food processing, machinery, and fabricated metals) new arrivals are more likely to be attracted to zipcodes as employment in the own industry within one mile increases. Employment in the own industry just five miles away, however, has a much smaller effect, as does employment further out in the ten and fifteen mile rings. It is worth pointing out, however, that for some industries the effect remains significant at fifteen miles. In sharp contrast, employment outside of the own industry has an inconsistent and frequently insignificant effect.

These results are important because they provide evidence that agglomeration economies may attenuate rapidly across geographic space. Indirect evidence consistent with these results is also provided in Henderson (2003a). He finds that employment activity in a plant's own county affects plant productivity. But employment activity in neighboring counties is not found to affect the plant's productivity.

Rosenthal and Strange (2001) also provide indirect evidence that agglomeration economies attenuate across geographic space. In this paper, the level of agglomeration for individual 4-digit SIC industries is regressed on industry characteristics in an attempt to identify the micro-determinants of the sources of agglomeration economies (e.g., labor pooling, knowledge spillovers, shared inputs). The study uses the Ellison–Glaeser (1997) index of agglomeration calculated at the zipcode, county, and state levels of geography. We will discuss the paper's conclusions regarding microfoundations in Section 3. At this point, we emphasize just those results that have a geographic dimension. The first of these is that reliance on factors sensitive to shipping costs (manufactured inputs, natural resource inputs, and perishability of products) influences agglomeration at the state level. In contrast, knowledge spillovers impact highly localized agglomeration, while labor impacts agglomeration at all levels of geography. This is consistent with attenuation of agglomeration economies, with the geographic scope differing for the various agglomerative forces. It seems sensible to us that knowledge spillovers would have a different geographic scope than would, for instance, input sharing.[12]

Finally, Ellison and Glaeser (1997) present a measure of an industry's concentration, and compare this index at different levels of geography. This leads them to conclude that much of the effects are localized, but that there are also spillovers. In a follow-up paper on measuring localization, Duranton and Overman (2002) find geographic localization to take place at a relatively small scale, less than fifty kilometers.

2.5. *Temporal scope*

The key issue regarding the temporal scope of agglomeration economies is whether agglomeration economies are static or dynamic. This issue is touched on in the growth models of Glaeser et al. (1992) and Henderson, Kuncoro and Turner (1995). Both of these papers show that the characteristics of a city can impact its growth over a period of twenty years or more. This does not necessarily mean that the economic environment twenty years or so earlier continues to have a direct impact on growth. Instead, the effect may be indirect, an accumulation of much shorter direct effects over the period. For example, externalities arising from shared inputs make a city more attractive and draw additional industry to the local area. As this occurs, the ability to share inputs further increases, contributing to further growth of the urban area.[13] Although this is clearly a dynamic process, it is not a direct dynamic effect with a twenty-year reach. A different estimating strategy would be required to estimate that kind of effect.

The direct dynamic effect is most naturally thought of as a knowledge spillover, although it could take other forms. The idea is that if knowledge were to take time to accumulate, having a lot of activity a few years ago could directly influence today's productivity. In a sense, urban areas can be thought of as schools in which managers and workers can continually add to their skills. As the time spent in a local environment increases, knowledge of local business contacts and Marshall's "secrets of the trade" would also increase. Two otherwise identical enterprises in the same city may therefore benefit differently from the local agglomeration depending on how long each has been present. Similarly, two otherwise identical cities would offer different sorts of increasing return depending on their histories. We refer to the historical component of agglomeration economies as their temporal scope. The parallel to distance is natural. Two establishments at a physical distance have less effect on each other. The same is true for two establishments at a greater industrial distance from each other. Likewise, the effect is likely smaller for two establishments who are farther apart in time.

Henderson (1997) is the paper that most directly addresses the temporal scope of agglomeration economies. He uses lags and differencing methods to identify the impact

[12] It should also be emphasized that the geographic scope of different sources of agglomeration economies has only occasionally been considered. This issue is addressed more fully in Section 4.

[13] See Rauch (1993b) or Helsley and Strange (2001) for models of this kind of dynamic growth.

on growth of the local environment at some point in the past. As usual, he estimates separate models for different industries. Again, as usual, the estimates vary by industry, with some having externalities with a long temporal scope and others having externalities that are closer to being static. The largest effects on productivity are typically from own-industry employment at two to five years in the past. Henderson speculates that these lagged effects may arise because it takes time to learn from neighbors. If the knowledge does not lead to portable increases in productivity, which would be the case with knowledge about local networks, then mobility across locations may be reduced.

A more recent paper by Glaeser and Maré (2001) takes a different approach to the temporal scope of agglomeration economies. They employ three different major datasets, drawing on the advantages of each to offset limitations of the others.[14] The estimation involves individual wage rates being regressed on a variety of local attributes and worker characteristics. The key result is that workers earn higher wages in large cities, with the urban wage premium equaling 33%. It is possible that this result could be explained by selection instead of by agglomeration economies. This possibility is addressed by looking at the effects of urbanization on recent migrants. The key idea is that if selection is at work, then recent migrants would receive higher wages since they would be, by hypothesis, the most able. The conclusion of this analysis is that there remains a substantial urban wage premium, perhaps 20%.

The part of the analysis that bears most closely on the temporal scope of agglomeration economies comes from considering the timing of the urban wage premium. Glaeser and Maré (2001) report evidence that long-time residents in bigger cities earn a premium over workers newly arrived to the same city. In addition, they find that when long-time urban workers leave their city, their wages in their new location are higher the larger the size of the previous city of residence. These results complement those of Henderson (1997) and provide further evidence that agglomeration economies have a dynamic component. However, neither Henderson (1997) nor Glaeser and Maré (2001) address the mechanism by which the dynamic spillover occurs. Both here as in many other part of this literature, therefore, further research is warranted.

2.6. *Industrial organization, "culture", and the transmission of agglomeration economies*

Thus far, we have dealt with agglomeration economies as technological phenomena. The results that we have reported address the degree to which external economies in production exist and their industrial, geographic, and temporal scope. At this point, we will discuss a rather different approach to agglomeration economies, one that stresses incentives and organizational considerations rather than technology. As will become apparent, these considerations influence the degree to which a given pattern of agglomeration creates external economies.

[14] The datasets studied include the PSID, the National Longitudinal Survey of Youth (NLSY), and the decennial Census.

2.6.1. Competition

The key idea here is found in Porter (1990). He argues based on case evidence (to be discussed in Section 4) that local competition encourages innovation by forcing firms to innovate or fail. In this view, for any given set of industrial clusters, competitive pressure enhances productivity. Since a firm's competitors are by definition within its own industry, this is a localization-based view of agglomeration. A different possibility, one that Glaeser et al. (1992) ascribe to Marshall (1920), Arrow (1962), and Romer (1986), is that the presence of local competition will decrease productivity because of incomplete appropriation of the returns from innovation.

Glaeser et al. (1992) test these ideas in the growth models discussed earlier. They include as a regressor the ratio of establishments per employee in a city for a given industry relative to establishments per employee for the entire U.S. As this ratio decreases the local environment in the given industry is thought to become more competitive. Measuring local competition in this manner, Glaeser et al. find that an increase in competition is positively associated with growth.[15]

In related work, Henderson (2003a) considers the influence of the average size of plants in the establishment's own industry and county on individual plant productivity. He also controls for the number of own-industry plants present in the county, along with other plant-specific attributes, industry-time fixed effects, and plant-location fixed effects. Henderson finds that average employment per plant does not positively affect productivity for the high-tech and machinery industries. In contrast, the number of plants in the own industry in the plant's county does positively affect productivity among high-tech industries (the effect is insignificant for machinery). He interprets these findings as providing evidence that localization economies arise from the presence of establishments *per se*, rather than size of the establishment.

Rosenthal and Strange (2003) employ similar competition variables when studying the number of births of new establishments and their employment. The specific controls are the number of establishments per worker in the establishment's own industry and the number of establishments per worker in other industries. Both of these variables are measured at the zipcode level. For all six industries studied, results indicate that average establishment size outside of the own industry has a significant negative influence on arrivals of new establishments. But for five of the six industries average establishment size within the own industry has a significant and positive influence on arrivals, while for the sixth industry this variable is insignificant. The own-industry finding echoes results from Glaeser et al. (1992). As in that paper, one possible interpretation is that the presence of smaller establishments implies a more competitive environment and that competition is good for growth. But, as will be discussed below, an alternative

[15] Glaeser et al. (1992) also note that this result could reflect the fact that smaller establishments grow more quickly. That interpretation is consistent with empirical evidence from the industrial organization literature. See, for example, Evans (1987).

interpretation is that establishment size may be associated with a different way of doing business. In particular, smaller establishments may be more flexible and open to nearby companies, and therefore might make good neighbors.

2.6.2. *Industrial organization and business culture*

The other key idea in this section is due to Saxenian (1994). In her comparison of the differences in performance between the Silicon Valley and Boston's Route 128, she argues that local technological capabilities are not the fundamental source. The primary cause is instead the differences in local industrial organization and culture. The key difference is that the Silicon Valley is in some sense more entrepreneurial than Route 128. Saxenian's analysis is discussed further in Section 4. We will at this point instead discuss attempts to understand the nature of agglomeration by estimating the kinds of effects that Saxenian identifies.

In a model of births, Rosenthal and Strange (2003) look in two ways for evidence of an impact of firm size. First, they partition the agglomeration variables according to the size of the establishment in which a neighboring employee works. This allows the estimation of different agglomerative effects for a worker in a small firm compared to a worker in a large firm. The idea is that the small firm is likely to be more open, with a greater external effect being the predicted result. The second organizational regressor is based on whether the employee works at a subsidiary establishment or at an independent establishment. The latter is presumably more open, while the former is presumably more closed.

The results of this estimation are at least partly consistent with Saxenian. Adding an additional employee at a small firm typically has a significant and positive effect on births and new firm employment. Adding the employee at a large firm typically has an insignificant effect. To the extent that small firms are more open, this result is consistent with Saxenian. The performance of the subsidiary/nonsubsidiary variable is unexpected, with an extra worker at a subsidiary establishment having a larger effect. This is not consistent with Saxenian, suggesting as it does that corporate establishments may have larger effects on the productivity of neighbors. One possible explanation for this result may be that in some sense the quality of the interactions with nearby employees of subsidiary plants is greater than those of nearby independent plants. However, the reason for such quality differentials remains to be explored.

Henderson (2003a) also considers establishment size, but in a different way. He finds that small firms enjoy a larger increment to their productivity as own-industry employment increases in the same city. This result on how firm size can impact the reception of agglomerative spillovers complements the Rosenthal–Strange (2003) result on how firm size can impact the transmission of agglomerative spillovers. Both are consistent with Saxenian.

Florida and Gates (2001) take a very different approach to the quantification of a city's environment. They find that cities that have many "bohemians", defined as artistic occupation categories, tend to innovate more than do less creative cities. Similarly, cities with large gay populations are also more innovative. The results are interpreted as suggesting that tolerant environments are more innovative. While one would not want to read too much into these sorts of correlations, they are consistent with the idea that culture matters.

2.6.3. The urban rat race

Another agglomeration effect can be shown to follow from the different incentives possessed by urban residents. Specifically, cities can either inspire or require hard work of their residents, a kind of urban rat race. Rosenthal and Strange (2002) consider this issue by looking at the connection between agglomeration and work behavior. The paper begins by looking at the facts: urbanization is shown to be positively related to work hours for full-time workers in professional occupations, even after controlling for individual worker attributes and for occupational fixed effects. However, that pattern largely disappears after controlling for the localization of the worker's occupation. In addition, the pattern is never present among non-professional workers.

To investigate the source of these effects, two simple models are specified: a selection model in which hard working individuals choose to locate in an active professional environment, and an urban rat race model in which proximity to workers of a similar type causes individuals to work longer hours. If the intrinsic taste for hard work persists over time, then the selection model implies that workers of all ages should work longer hours in agglomerated environments. In contrast, the rat race model is based on the idea that competition encourages individuals to work longer hours when it is important to be noticed. This effect is likely most pronounced among young professionals who have the most to gain from reputation building.

The paper employs differencing methods to test for the presence of these effects using 1990 Census data on full-time workers. For professional workers in their 30's (defined as "young") and 40's (defined as "middle-aged"), work hours are longer in locations where the density of employment in the worker's occupation is high. No such effect is present for non-professional workers of any age. Adding controls for the proximity to rivals with whom the worker is most likely to compete – defined as individuals who earn a similar wage in the national wage distribution for the worker's occupation – does not change this result. Findings based on this specification also indicate that both young and middle-aged professional workers work longer hours in areas with a high concentration of individuals in their professions. In addition, young professionals are found to work longer hours when both rivals are present and the rewards to advancement are high. Absent such potential rewards, the presence of rivals does not differentially affect the work habits of young versus middle-aged professionals. It should also be noted that these results are robust to controls for occupation fixed effects and also the concatenation of occupation and MSA fixed effects (over 6000 fixed effects in all). On the whole, this work confirms the long held belief that cities attract industrious workers. The research also seems to identify an overlooked aspect of the urbanization–productivity relationship, that cities encourage hard work.

3. The sources of urban increasing returns

As noted earlier, there are many potential sources of agglomeration economies. A complete understanding of urban development clearly requires that these sources be understood. Some of these microfoundations were suggested by Marshall (1920), including knowledge spillovers, labor market pooling, and input sharing. There are many other causes of agglomeration that were not discussed by Marshall, including home market effects, urban consumption opportunities, and rent-seeking. The literature on the theoretical microfoundations of agglomeration economies is surveyed in another chapter [Duranton and Puga (2004)]. This section will consider econometric evidence on microfoundations. Table 2 provides an overview of some of the studies to be discussed. Three

Table 2
The Marshallian microfoundations of agglomeration economies

Microfoundation	Paper	Key results
Natural advantage	Kim (1999), Ellison and Glaeser (1999)	Factor endowments matter
Input sharing	Holmes (1999)	More purchased inputs in clusters
Labor market pooling	Diamond and Simon (1990)	Labor market risk capitalized in wages
	Kahn and Costa (2001)	High-education married locate in large cities
Knowledge spillovers	Jaffe, Trajtenberg and Henderson (1993)	More citations of a patent in the same MSA, controlling for industry effects
	Rauch (1993a), Moretti (2000)	City average education raises wage
Home market effects	Davis and Weinstein (1999)	For some industries, regional development explained by market access
Consumption	Tabuchi and Yoshida (2000)	Real wages lower in cities (reflecting consumption possibilities)
	Glaeser, Kolko and Saiz (2001)	Various measures of consumption possibilities predict growth
	Waldfogel (2003)	Agglomeration enhances consumption possibilities in radio listening
Rent seeking	Ades and Glaeser (1995)	Dictatorship predicts mega-cities
	Henderson (2003b)	Dictatorship encourages urban primacy, which discourages growth
Multiple	Rosenthal and Strange (2001)	Evidence of labor market pooling at state, county, and zipcode levels of geography; of knowledge spillover and input sharing at zipcode and state levels
	Dumais, Ellison and Glaeser (1997)	Strongest evidence for labor market pooling, some evidence for knowledge spillovers and input sharing
	Audretsch and Feldman (1996)	Evidence of input sharing and knowledge spillovers at the state level

other chapters will also consider this sort of evidence [Audretsch and Feldman (2004), Moretti (2004), and Head and Mayer (2004)], so in some places, we will refer to them.

As in Section 2, it is helpful to begin by considering how an ideal empirical analysis would proceed. Clearly, the same sort of evidence that would have been required to estimate the contributions of urbanization and localization would be required to identify the microfoundations of agglomeration economies. In fact, even more detailed data would be needed, since it would no longer be sufficient to observe the contribution of own-industry employment to productivity. Instead, it would be necessary to observe the specific channels by which own-industry activity impacts productivity. For instance, in evaluating plant level productivity, information on the local knowledge that spilled-in would be required. Regarding inputs and workers, the data would have to include information on the quality of the match between worker and input and the employing firm. It is hard to imagine the ideal data ever being available, and so some sort of econometric compromise is necessary.

3.1. Increasing returns or natural advantage?

We will begin by considering an approach that looks at microfoundations in a negative way. Estimates of the impact of agglomeration on productivity reviewed in Section 2 should be interpreted as providing an upper bound on the magnitudes of the various sorts of agglomeration economies. This is because agglomeration arises both because of the benefits of locating in areas endowed with natural advantages and also because of the influence of agglomeration economies. The estimated impact of agglomeration economies, therefore, can be refined by determining the share of productivity that can be attributed to a location's natural advantage instead of to agglomeration economies.

The potential role of natural advantage is easy to see. The steel industry in North America, for example, was initially concentrated in the Great Lakes region largely because of the presence of iron ore and coal. Similarly, it is certainly true that at least part of California's growth can be attributed to its climate, which would allow employers to pay lower wages than where the weather was less pleasant. In order to analyze the sources of agglomeration economies, it is therefore necessary to begin by looking at the degree to which natural advantage explains location.

There is a long history of empirical research on industrial location that has considered the role of natural advantage. As always, Marshall (1920, p. 269) is seminal, noting that:

> Many various causes have led to the localization of industry, but the chief causes have been physical conditions; such as the character of the climate or the soil, the existence of mines and quarries in the neighborhood, or within easy access by land or water.

Fuchs (1962) is an exhaustive North American reference, documenting the importance of access to resources for manufacturing industries.

Natural advantage has also been considered more recently. Kim (1995, 1999) and Ellison and Glaeser (1999) have argued that natural advantages are very important in determining agglomeration. This means a smaller role for increasing returns in the process of agglomeration. Looking at agglomeration between 1860–1987, Kim (1995) regresses a location quotient, measuring the concentration of industry, against measures of plant size, natural resources, and industry and time dummies. The positive coefficient on the natural resources variable is interpreted as being consistent with an important role for natural advantage in determining agglomeration.[16] Over a slightly different period, 1880–1987, Kim (1999) estimates a Rybczynski equation, regressing state employment in a given industry on factor endowments. This approach assumes that all factors of production are immobile, including labor. It is maintained that the residuals in this estimation are upper bounds on the strength of agglomeration economies. In a similar way, Ellison and Glaeser (1999) employ predicted state level employment variables to account for the importance of natural advantage in agglomeration. Both Kim and Ellison and Glaeser conclude that natural advantage is important. Specifically, Ellison and Glaeser show that the percentage of agglomeration that is predicted by the natural advantage proxies is roughly 20%. Ellison and Glaeser argue that the proxies are imperfect, and so the fraction of agglomeration that could possibly be explained is larger than 20%. Exactly how much larger is unclear, which seems to leave a lot of agglomeration unexplained by the natural advantage proxies.

A more serious issue in interpreting these results is that they assume all factors of production to be immobile, including labor. It is, of course, precisely the mobility of labor that leads to agglomeration in the presence of external increasing returns in production. Or in reverse: external increasing returns lead to the agglomeration of labor. Attributing the productivity of agglomerated labor to natural advantage seems, therefore, to be questionable, at least for more recent periods (i.e., labor probably was fairly immobile in the early periods considered by Kim (1999)).

3.2. What do the productivity studies have to say about microfoundations?

Not much. Although claims are made regarding the nature of the externalities that are measured, it can be shown that agglomeration economies whose sources are knowledge spillovers, labor market pooling, or input sharing all manifest themselves in pretty much the same way. To see this, suppose that a firm's profits equal $g(A)f(x) - c(x) + \varepsilon$, where A is an index of agglomeration economies, x is a vector of the traditional inputs of production, and ε is an independent and identically distributed error term, distributed across establishments according to the cumulative distribution function $\Phi(\varepsilon)$. The first-order conditions for optimal input demands are of the form $g(A)\partial f/\partial x_i = \partial c/\partial x_i$

[16] He argues that there is little evidence of external economies impacting regional specialization. This conclusion is based on the fact that high-technology industries are not significantly more concentrated than are other industries and on the claim that theories of agglomeration imply that high-technology industries are more sensitive to external economies. The fact is interesting, but the claim, while tempting, is not an immediate consequence of any theories of agglomeration of which we are aware. In fact, Duranton and Puga (2001a) note that specialized cities are attractive for mature industries.

where x_i is an input (i.e., labor) and $\partial c/\partial x_i$ is the marginal cost of the input (i.e., the wage). An increase in $g(A)$ from any source will lead to higher productivity, higher employment of the input and, through the usual derived demand channel, to an increase in the price of the input. Thus, high productivity, employment, wages, and rents reflect the presence of agglomeration economies. They are not, however, evidence of any particular agglomeration economy.

The same observational equivalence holds for growth. This point is made precisely in Helsley and Strange (2001). The thrust of the argument is as follows. Growth requires profitability. Profitability requires productivity, which may be enhanced in a dynamic sense by agglomeration economies. Suppose that each firm, actual or potential, requires a specific heterogeneous input, indexed by y in the characteristic space, which is taken to be the unit circle. Prior activity makes these inputs available. The greater the amount of prior activity, the thicker will be the input market. This means that agglomeration can enhance growth by allowing future activity to be carried out at lower cost, since the costs of adjusting pre-existing inputs will decrease with the amount of activity. Thus, growth fosters agglomeration by making inputs available to entrepreneurs. However, the input can be anything: a supply channel developed by existing firms, a labor market pool that prior employment has created, or an idea that flows from existing activity. As long as current activity creates something that future entrepreneurs can draw from (physical inputs, a labor pool, or local knowledge), agglomeration will enhance growth.

This observational equivalence is somewhat unfortunate, since it means that the well-developed literature on the scope of agglomeration does not directly shed a lot of light on the microfoundations of agglomeration economies. There have occasionally been instances where results on a growth–localization or growth–diversity relationship were interpreted as knowledge spillovers. While it is true that knowledge spillovers are certainly one possible channel by which growth could be influenced by localization or by diversity, it is not the only one. These claims should, therefore, be taken with some caution.

In order that the productivity studies could be interpreted as bearing on the microfoundations of agglomeration economies, it is necessary to interpret the results in a more structural way. Two approaches of this kind have been suggested. One, proposed by Glaeser and Maré (2001), is to look to the dynamic structure of agglomeration economies for evidence of microfoundations.[17] As noted in the previous section, they find the positive effect of urbanization on wages to occur with a lag. One is tempted to interpret this result as reflecting knowledge spillovers among workers, with the slow increase in wage reflecting the accumulation of knowledge. The other approach, proposed by Henderson (2003a) is to look at the effect of the number of firms, rather than their employment levels, on the productivity of neighbors. He argues that this variable is more likely to capture knowledge spillovers than other microfoundations. For this to

[17] Glaeser and Maré (2001) are explicitly aware of the observational equivalence problem, noting that an increase in wages can reflect any sort of agglomeration economy.

be true, it must be the case, however, that the amount of knowledge that can spill out from a firm does not increase with its activity level. It must also be true that the spilling out of knowledge must be independent of the number of workers with the knowledge. An equivalent interpretation of a positive sign on a numbers variable would be that each firm needs its own input, and so the market thickness effects in the input market would depend on numbers and not levels. Of course, this interpretation has the same sort of qualifications that the knowledge spillover interpretation did. It does seem clear, nonetheless, that an effect of the number of firms rather than their employment levels is not consistent with the idea of labor market pooling.

In sum, the many excellent studies of productivity have told us about the existence of agglomeration economies and also about their scope across industries, locations, and time. They have not, however, had much to say about the sources of agglomeration economies. These approaches are the subject of the rest of this section.

3.3. *Individual microfoundations*

One way to analyze the sources of agglomeration economies without being able to make inferences from data on productivity, growth, or wages is to look at proxies for the microfoundation. For example, although it is difficult to link Marshallian input sharing directly to productivity, it may be possible to determine when input sharing is taking place. By relating this information to location patterns, one may be able to assess the theoretical claims on input sharing and agglomeration. In principle, this method can be applied to the other Marshallian microfoundations of labor market pooling and knowledge spillovers, as well as to other potential sources such as urban consumption opportunities.

3.3.1. *Input sharing*

Marshall's notion of input sharing depends crucially on the existence of scale economies in input production. If there were no scale economies, then a downstream firm could procure inputs at the same low price in isolation as it could in the midst of other similar firms. In the case where there are scale economies, however, the isolated firm will be at a disadvantage. Only downstream firms located where the industry is concentrated will be able to outsource their input demands to producers who are able to achieve an efficient scale of production.

Holmes (1999) considers the connection between the characteristics of a firm's location – concentrated or not – on input sharing. His insightful treatment of employment concentration is based on U.S. Census data on manufacturing from 1987 at the establishment level. This is matched with data on purchased inputs, also from the Census, that is available for locations. The locations by which this is defined vary by industry.[18]

[18] For instance, he notes that data on purchased inputs for the creamery butter industry is partitioned into only two locations: Wisconsin and the rest of the U.S.

Dividing purchased inputs by sales gives purchased input intensity, which is his measure of vertical disintegration and therefore of input sharing. These data are used in two ways. First, the differences between purchased input intensity between the most concentrated location and the rest of the U.S. are compared. This analysis shows that the most concentrated industries exhibit a relationship that is consistent with input sharing. For instance, the pantyhose industry is concentrated in North Carolina with 62% of national employment, and purchased input intensity of 53%, compared to 40% nationally. This pattern is repeated for other concentrated industries.

The second approach uses regression analysis. The dependent variable is the difference between purchased input intensity and industry mean purchased input intensity. This is regressed on the amount of same-industry employment in the establishment's own county and in other counties whose geographic centers are within fifty miles, again differenced from an industry mean. The results are again consistent with input sharing. Averaging across industries, moving from an unconcentrated location (499 or fewer neighboring employees in the same industry) to a concentrated location (10,000–24,999 neighboring employees) results in a 3% increase in purchased input intensity. Of course, as noted in Section 2, there is nothing in the theory to suggest that the effects would be the same across industries. To deal with this, Holmes considers the ten most concentrated industries. For these, he finds an effect that is roughly twice as large. In sum, there is consistent evidence of a positive relationship between vertical disintegration and industrial concentration that is strongly suggestive of Marshallian input sharing.

If there were input sharing, one would also expect there to be a greater fraction of input suppliers carrying out specialized functions. Thus, with highly refined data it might be possible to test for the existence of input sharing by looking for the presence of specialized input suppliers at locations where an industry is concentrated. Unfortunately, the nature of industry classification, specifically the practice of putting vertically linked stages of production into the same category, makes this test difficult to carry out. Holmes notes that the textile industry is defined in a way that overcomes this difficulty. He therefore uses the industry's specialized finishing plants (SIC 226) and the entire textile industry (SIC 22) to look for a relationship between the fraction of specialized plants and his measure of industry concentration. The relationship among the median specialized input fraction and concentration is exactly consistent with the existence of input sharing.[19]

In our view, the work by Holmes (1999) offers the most direct and compelling evidence to date of the importance of input sharing as a source of agglomeration economies. Even here, however, interpretation issues arise. In particular, plants with idiosyncratically strong tendencies to rely on shared inputs will naturally seek out locations where such opportunities are present, while agglomeration itself enhances opportunities for input sharing. From an econometric perspective, this implies that plant location is endogenous causing Holmes to overestimate the causal influence of agglomeration on opportunities for input sharing. Nevertheless, the positive relationship identified

[19] In a related paper, Ono (2002) finds there to be more outsourcing in large cities.

in Holmes' work between input sharing and agglomeration can only arise if agglomeration facilitates such opportunities. In that sense, findings by Holmes (1999) unambiguously support the presence of input sharing as a source of agglomeration economies.[20]

Several other papers also provide evidence supportive of the role of input sharing. The importance of input sharing at the macro level is touched on in Bartlesman, Caballero and Lyons (1994). They present a growth model that uses panel data methods to consider the possible presence of thick market effects on productivity. While this analysis is not focused on cities, it does seem to be relevant. The key result of the "within" estimates – drawing on temporal variation to identify effects – is the finding that customer thickness has a large effect on growth, but supplier thickness does not. The "between" results – estimates that draw on cross-sectional variation to identify effects – are exactly opposite. The cross-sectional results show a strong reduced form relationship between supplier thickness and productivity. Together, Bartlesman, Caballero and Lyons (1994) reconcile these findings by observing that over the short-term customers are crucial. Over the long term, it is the suppliers who matter most.

Holmes and Stevens (2002) is also relevant to input sharing, although it does not address the issue directly. They find that establishment size is larger where an industry is concentrated. This result is identified using a "size coefficient", equal to the ratio of the mean establishment size (measured in value of output) at a location to the mean size in the sector across the U.S. This result contrasts with Vernon's (1960) analysis of firm sizes in the New York region. It does not, of course, directly address the degree of vertical integration. In sum, the papers reviewed above provide fairly strong evidence that input sharing is important, both for cities and overall.[21]

3.3.2. Knowledge spillovers

Knowledge spillovers may be the most interesting of the microfoundations, speaking as they do to so many different areas of economics, including growth theory and the economics of human capital. There is no doubt, however, that knowledge spillovers are difficult to identify empirically. Unlike input sharing, for instance, knowledge is often exchanged without being bought and sold – the word "spillover" is important [see Helsley and Strange (2002)]. Even if there is an exchange, it is more likely to be a complicated joint venture between organizations, the kind of transaction for which data are routinely collected. Thus, the econometrician faces many challenges in measuring knowledge spillovers.

One way to deal with the challenges is to look for direct evidence of knowledge spilling over. This has the obvious advantage of its directness, but it does not directly

[20] The R^2 values in the input share regression in Holmes (1999) are also quite low. Although to some extent this is to be expected given that Holmes differences away the industry means from the data, the very low R^2 values are suggestive that other unspecified factors also contribute to input sharing.

[21] There are many papers in the tradition of the New Economic Geography that also consider input sharing as a possible source of agglomeration. See the chapter by Head and Mayer (2004) for a discussion.

tie the knowledge spillover to either agglomeration or productivity. One example of this kind of research is Jaffe, Trajtenberg and Henderson (1993). They provide probably the most compelling evidence to date for the idea that knowledge spillovers are important and that such spillovers attenuate with geographic distance. They identify a "paper trail" of knowledge spillovers in the location of patent citations. The key result is that patent citations are highly spatially concentrated, with citations 5 to 10 times as likely to come from the same SMSA as control patents. Other papers looking at knowledge spillovers in this way include Jaffe (1989) and Acs, Audretsch and Feldman (1992).[22] It is important to remember that even in this kind of research, the issue of industrial scope remains.

This kind of research has proceeded in parallel with work that has attempted to study the impact of industry characteristics on innovation, presumably the most important direct outcome of knowledge spillovers. Of course, innovative activity in the broadest sense is very difficult to measure, forcing researchers to use narrower proxies. One example of this kind of research is Audretsch and Feldman (1996). In addition to looking as well at the determinants of agglomeration (see below), they consider the impact of agglomeration on innovation. Their measure of innovation is based on a count of significant product introductions by the Small Business Administration in 1982. Those data are not commonly collected and are particularly valuable. Audretsch and Feldman (1996) regress the spatial concentration of innovative activity conditional on a number of local and industry specific attributes, including university research in the field, expenditures on research and development, and the availability of skilled labor. They also control for the industry-specific spatial concentration of manufacturing activity. Results confirm that knowledge oriented industries have more spatially concentrated innovative activity, consistent with the presence of knowledge spillovers.

This strand of the literature is surveyed fully in another chapter [Audretsch and Feldman (2004)], so we will not comment on it further. Instead, we will move on to consider a related strand of the urban literature that deals with human capital externalities. Since this literature is also the subject of another chapter [Moretti (2004)], our goal will be simply to characterize how the human capital literature relates to the microfoundations of agglomeration economies. We will consider only highlights from the literature, and not even all of them.

The key idea in these papers is that workers are the primary vehicle of knowledge spillovers. This seems to be what Marshall had in mind in his discussion of Sheffield cutlery workers taking advantage of the secrets of their trade that are available as local public goods. Rauch (1993a) considers this issue by looking at the impact of the average level of education on wages and rents. Using a hedonic model in the spirit of Roback (1982), he shows that the presence of public good effects should increase wages where

[22] Almeida and Kogut (1999) employ the Jaffe, Trajtenberg and Henderson (1993) methodology to look at the characteristics of regions impact knowledge spillovers. Their key result is that regions differ in the way that knowledge spills-over, a result reminiscent of Saxenian (1994).

average education levels are high, since workers will be more productive and employers will be willing to pay high wages in competing for them. In addition, his model shows that rents will be high too, since the productivity enhanced high wages will naturally be capitalized into housing prices.

Rauch employs 1990 Census data to test these predictions. The primary conclusion is that both wages and rents rise significantly with average education. The magnitudes are nontrivial, with a one year increase in average schooling leading to an increase of 3% in wages and 13% in rents. One potential difficulty with this result is that schooling, whether at the average or individual level, is endogenous. Acemoglu and Angrist (1999) use compulsory schooling laws as an instrument for the local level of education. They then estimate the effects of local education, finding a positive effect, but one that is small and insignificantly different from zero. Following up on Rauch, Moretti (2000) considers the impact of the presence of college graduates on a city's wages. He finds a positive effect of this kind of human capital. Together with Rauch and Acemoglu–Angrist, this might suggest that the human capital externalities depend on highly educated workers.

One objection that one can make to looking at the effects of local education levels as a way to understand knowledge spillovers is that the exact channel of interaction is unspecified. Charlot and Duranton (2002) address this issue using survey data. They find reported workplace communication to be more extensive in urban areas. They then show this communication to impact wages. This impact is small, however, amounting to roughly one-tenth of the urban education premium. This may mean that the education premium depends on something other than knowledge spillovers or that the surveys do not uncover all of the knowledge spillovers that take place. The issue of human capital in cities continues to receive considerable attention. As noted above, we believe that with only a few exceptions, the literature has provided little compelling evidence on the mechanism by which knowledge spillovers are transmitted.

3.3.3. *Labor market pooling*

There are two related interpretations of labor market pooling. One is that workers should be better matched in large cities (an urbanization effect) or in industrial concentrations (a localization effect). It is not clear how one might assess the quality of a worker's match. One possibility would be to look at termination rates, controlling for measures of the health of the local economy and the industry. To the extent that bad matches lead to termination, this would be one way to look for direct evidence of labor market pooling. However, in the absence of a good replacement, a firm would be less likely to fire a worker of any given match quality. Thus, the viability of this approach is unclear. Another possible approach to identifying labor market pooling would be to look at turnover. The implications of the labor-market pooling hypothesis are that workers can readily change jobs and that firms can just as readily change employees. This has the same difficulty as using termination, however, since it is not clear that turnover would actually occur in a thicker labor market.[23] Another way to look at match quality is to look at specialization. In this vein, Baumgardner (1988) shows that physicians perform a narrower range of activities in large markets. This confirmation of Smith on the division of labor is also consistent with Marshallian labor market pooling in that it shows that agglomeration can foster specialization.

The other interpretation of labor market pooling is that it is fundamentally about risk. Workers and firms confront two kinds of risk in choosing to work for a particular employer in a particular city. One is worker and firm-specific: for reasons particular to the match between worker and firm, the employment relationship may be terminated. Assuming that the termination did not result from exiting the labor market completely, the worker needs another job, and the firm needs another worker. If the worker's skills and the firm's labor requirements are specific to an industry, then these needs will be easier to meet in a location where the industry is concentrated. Thus, the worker and firm-specific risk will be reduced by localization.

The second source of risk is industry-specific. Suppose that industries are subject to shocks. In this case, an industry shock could result in a worker losing a job. The worker will not find getting a job to be easy if the rest of the employers in the area are in the same negatively shocked industry. Thus, locating in a specialized city exposes a worker to greater risk. This analysis implies that industry specific shock discourages localization, while the match-specific shocks described above encourage it. It is, therefore, an empirical question of how the degree of local specialization impacts the labor market.[24]

Simon (1988) considers the relationship between the unemployment rate and a city's specialization. The latter is measured using an industry-based Herfindahl index. He shows that unemployment is greater the more specialized is a city, which is consistent with the industry-specific shocks being important. To the extent that risk is greater in a specialized city, then workers will require higher wages as compensation. Diamond and Simon (1990) address this issue, showing that workers demand higher wages in more specialized cities. This effect is tied to risk by showing that the higher wages are related to the specific measures of the cyclical variability of an industry's employment. In sum, this work is a formal treatment of Hoover's (1948) idea that one of the benefits of urbanization is the diversification it provides against sector-specific shocks to the local economy.

[23] Recent evidence on turnover is provided by Fallick, Fleischman and Rebitzer (2003) and Di Addario and de Blasio (2003). Fallick et al. find that computer industry worker mobility rates are greater in the Silicon Valley than outside of California. This effect is not found for other industries. Di Addario and de Blasio also fail to find a general increase in job mobility in the Italian Industrial Districts, which are considered by many to be European cousins of the Silicon Valley.

[24] A related point is made by Krugman (1991), who shows that if demand shocks are uncorrelated across individual establishments within a given industry, then firms operating in a pooled environment face an elastic labor supply function, enjoy stable wages, and are able to expand and contract during good and bad times. Krugman shows the benefits firms gain from the pooled environment during good times more than offset costs that arise from their inability to offer lower wages during bad times. On average, firms earn higher profits in the pooled environment, even when firms are risk neutral.

The notion that risk is lower and matches are better in large cities is crucial in Costa and Kahn (2001). They consider "power couples", defined as married couples where both partners have at least a bachelor's degree. They document a substantial increase in the fraction of these couples located in large cities, from 32% in 1940 to 50% in 1990. This is an increase that is substantially greater than the increase in the population of the large cities over that period. One explanation for this is that these couples met and married in large cities. Another is that large cities offer a resolution to the dual career problem by increasing the probability that both partners will be able to find jobs that are closely matched to their abilities. They test between these hypotheses by looking at the differences between power couple location patterns and the location patterns of other types of couples, singles, and "potential" couples (referred to as "incidental" couples in the paper). Using the differencing strategy, Costa and Kahn (2001) conclude that 36% of the increase in the concentration of power couples in large cities is explained by the dual career hypothesis.[25] This story is very much in the spirit of the statistical explanations of agglomeration economies that have been offered since Marshall (1920) and Hoover (1948). It has important implications for the future of cities. If the productivity of the highly educated is crucial for economic performance [as in Moretti (2000)] and if these workers continue to marry each other, then large cities will have at least one large productivity advantage relative to small cities.

3.3.4. Home market effects

Marshall is properly regarded as the first word on agglomeration economies. He is not the last. Various other explanations for agglomeration have been proposed.

One of these is the home market effect. Suppose that increasing returns lead to the concentration of employment into a large factory. This in turn, creates a large market, which, in the presence of transportation costs induces other firms to choose the same location. The idea here is that the interaction between internal scale economies in production and transport costs lead to a "magnification", where home market size expands in a self-reinforcing process of agglomeration.

One of the first formal treatments of the home market effect is Krugman (1980). Davis and Weinstein (1996) look at the relative magnitude of home market effects based on increasing returns and the more traditional Heckscher–Ohlin effects in a sample of OECD countries. They find substantially stronger evidence for Heckscher–Ohlin. A followup paper, Davis and Weinstein (1999), looks at regional agglomeration, instead of international, by considering data on Japanese prefectures. Here, in contrast, there are substantial increasing returns effects in eight out of the nineteen manufacturing industries that they study. This leads them to conclude that the home market effect may be an important determinant of regional concentration, of which the agglomeration into cities is an important instance.

[25] The use of difference-in-difference estimation methods in Costa and Kahn (2001) is highly effective. We believe that this approach could prove useful in the study of other problems in the agglomeration literature.

In a number of careful papers looking at location patterns on the U.S.–Mexico border before and after NAFTA, Hanson (1998a, 1998b) also looks at home market effects, among other things. He finds that increases in the openness of the Mexican market led to a substantial shift in the economic geography of manufacturing, with a new concentration near the border arising and the traditional concentration around Mexico City declining. This is consistent with what one would expect as trade liberalization makes the home market less important.

Taken together, these papers provide evidence of the existence of home market effects. A more complete survey of this literature is provided in another chapter [Head and Mayer (2004)].

3.3.5. Consumption

There has recently been work on agglomeration that has emphasized the consumption possibilities of large cities as sources of agglomeration. This is in clear contrast to the traditional analysis of agglomeration that has focused on the ability of cities to enhance productivity. Glaeser, Kolko and Saiz (2001) argue that there are four fundamental ways that large cities enhance consumption. First, there may be goods and services available in large cities that are not available elsewhere (i.e., opera or restaurants). Second, large cities may offer various aesthetic charms (i.e., the Los Angeles climate or Paris architecture). Third, large cities may allow the provision of public goods that would not be possible in a smaller place (i.e., specialized schools). Fourth, the relatively dense settlement of a large city allows speed of interaction that would not be possible in a smaller city (i.e., social interactions).

Glaeser, Kolko and Saiz (2001) present several different kinds of evidence to indicate the importance of consumption for cities. One is the rise of reverse commuting, which grew by 2.79% between 1980 and 1990, and at a roughly similar pace in the two previous decades. This suggests that some individuals working at suburban sites are willing to incur higher central city house prices (quality adjusted) for the opportunity to enjoy central city consumption amenities. Another kind of evidence emerges from reduced-form regressions of urban growth on various measures of a city's consumption possibilities. A temperate climate is a strong predictor of growth, as are the presence of museums and theaters. Cities with a large number of bowling alleys did not grow. Of course, there may be problems of endogeneity in that growth helps to finance cultural amenities while cultural amenities and public infrastructure can promote growth. Finally, Glaeser et al. also present evidence of an increasing concentration of wealth in central locations. On balance, the patterns noted by Glaeser et al. are consistent with the idea that consumption possibilities are a source of agglomeration economies.

Waldfogel (2003) and George and Waldfogel (2003) also deal with the implications of agglomeration for consumption. The key idea is that a larger market may allow goods to be more closely tailored to individual consumers' tastes. Waldfogel (2003) examines radio listening patterns, and finds that the average fraction of the population listening to radio increases by 2 percent with a one million person increase in a city's population.

This result is analogous to urbanization effects in manufacturing productivity. Waldfogel also finds that the number of radio stations targeting African American and Hispanic groups increases with the size of those populations apart from city size per se. This result is analogous to localization effects in manufacturing productivity. These patterns are mirrored in George and Waldfogel (2003), who analyze the newspaper purchases at the zipcode level. They find consumption externalities, with black newspaper purchases increasing with the size of the black community and decreasing with the size of the white community. White newspaper purchases, in contrast, increase with the size of the white community but are unaffected by the size of the black community. These findings suggest that local newspapers cater to the tastes of their dominant customer group, within group preference externalities are positive, and across group externalities are negative.

Tabuchi and Yoshida (2000) look at nominal and real wages to arrive at estimates of the relative strengths of the production and consumption inducements to agglomeration. They find that the elasticity of nominal wage with respect to city size is 10%. Since firms would not pay higher wages in larger cities unless there were also a corresponding increase in productivity, this nominal wage effect is interpreted as a traditional agglomeration economy in production. The elasticity of real wage with respect to city size is between -7 and -12%, depending on specification. This is interpreted as an agglomeration economy in consumption: workers acceptance of lower real wages in cities implies a corresponding consumption benefit.

The idea that workers would be willing to give up real wage to enjoy a city's consumption amenities is a central feature of a related literature on urban quality of life [e.g., Blomquist, Berger and Hoehn (1988), Gyourko and Tracy (1991)]. In that literature, the value that workers place on the opportunity to live in one city over another is measured by the difference in real wage necessary for the worker to be indifferent between the two areas. Gabriel and Rosenthal (2004) extend that idea to also measure the quality of a city's business environment – equal to the sum of land rent and wage a firm would be willing to incur for the opportunity to locate a worker in a given city. They create and analyze a panel of such measures for 37 cities over the 1977 to 1995 period. Results support the idea that cities attractive to industry grow larger, which is consistent with standard arguments in the agglomeration literature.

3.3.6. Rent-seeking

Marshall's analysis of agglomeration is that cities exist and industries concentrate because this sort of agglomeration is useful. This need not be the case. Ades and Glaeser (1995) demonstrate a relationship between rent-seeking and the formation of mega-cities, the "urban giants" that are homes to disproportionate shares of their countries' populations. Using a cross-section of 85 countries, they find that economic factors are part of the explanation of this kind of urban primacy. Specifically, tariff barriers lead to a larger degree of urban primacy, and the development of an inter-city transport network leads to less. Political factors are even more important than economic ones. Ades and Glaeser (1995) show this by constructing an index of political instability based on the numbers of coups, revolutions, and strikes. This index is shown to be associated with greater urban concentration. Centralized political systems – measured by an index of political rights such as voting procedures – have the same effect. For instance, an unstable dictatorship has 37% of its urban population in the largest city, while a stable democracy has only 23%. Of course, political stability and political rights are potentially endogenous to the level of urbanization. In order to address the issue of causality, Ades and Glaeser instrument using predetermined political characteristics, regional political characteristics, and regional infrastructure. In addition, they use the timing of urban growth to address the causality issue. The results of both procedures lead Ades and Glaeser to conclude that politics contributes to spatial concentration.

In order to be more specific about the mechanisms by which this occurs, Ades and Glaeser consider a number of case studies of urban gigantism. These cases consider classical Rome, Tudor and Stuart London, Edo during the period of the Shoguns, Buenos Aires around 1900, and Mexico City today. All of these had at least some element of the political causation of urban primacy. In Rome, for example, citizens could avail themselves of the fruits of empire by locating in the capital city and accepting imperial bread in return for not challenging the current ruler. In Mexico's centralized government, albeit nominally a federal one, it is possible to obtain valuable patronage (i.e., land), by locating in Mexico City. The implicit threat of a riot is parallel to the Roman situation. In sum, the ability to engage in rent-seeking seems to be one force that leads to the concentration of population in mega-cities.

Henderson (2003b) takes this analysis a step further. This paper also considers the determinants of urban primacy. As in Ades and Glaeser (1995), centralization, specifically a non-federal system, is shown to be positively related to urban primacy. Having a poor national road network also has a strong effect. The additional step in Henderson is the consideration of the impact of primacy on growth. As in Williamson (1965), this effect is shown to be significantly negative. For instance, for Argentina to be one standard deviation above its best level of primacy (around 0.25) would cost 1.3% growth. This is a striking result. It implies that anything that leads to excessive urban concentration – politics or roads or something else – can have substantial negative effects on economic performance.

Holmes (1998) establishes the importance of public policy in general by looking at location patterns around borders. More precisely, he uses the presence of a right-to-work law as a discrete proxy for a state's attitudes towards business. He shows that there is a discontinuous and negative change moving across a border from into a state with a right-to-work law from one without such a law. Thus, the problem of rent-seeking can be seen as part of broader pattern of public policy impacting location. [26] In contrast, Carlton (1983) finds a more moderate effect of public policy instruments in his study of firm location.

[26] Note, however, that there is nothing in the Holmes (1998) paper that suggests that a shift of manufacturing activity a short distance across state borders is inherently inefficient. This is in contrast to the Ades and Glaeser (1995) and Henderson (2003b) papers where excessive urbanization does implicitly imply a less than efficient allocation of activity across locations.

3.4. The relative importance of Marshallian microfoundations

A number of factors have been identified by theory as possibly having led to agglomeration. We have thus far looked at tests for the importance of these factors that are separate in the sense that they usually consider only one or at most two of the possible microfoundations. It would obviously be helpful if it were possible to disentangle the relative importance of these.

In the absence of data that would allow direct tests of the relative impact of, for instance, the three Marshallian microfoundations on productivity, other approaches must be employed. A natural one is to use the variation among industries to identify the impact of microfoundations. This section reviews papers that have, in different ways, followed this broad approach.

Audretsch and Feldman (1996) use a state-level spatial statistic to measure geographic concentration, both of aggregate employment and of innovation. The statistic is defined as $G = \sum_i (x_i - s_i)^2$, where x_i is location i's share of total employment and s_i is the location's share of employment in a particular industry. It takes on a value of zero when an industry is allocated across space in exactly the same way as total employment. It takes on a value close to one (depending on the size of the industry itself) when the industry is completely concentrated in one location.

Audretsch and Feldman begin by considering the determinants of the concentration of innovation. Controlling for the overall level of industrial concentration, which would lead to the concentration of innovation even in the absence of knowledge spillovers, they consider the determinants of the concentration of innovation. They show that innovation is more concentrated in an industry with a high ratio of R&D to sales, a greater reliance on skilled labor, and where more university research is devoted to research relevant to that industry.

Having looked at the concentration of industry, it is natural to look also at the concentration of production. Audretsch and Feldman (1996) show that industries are more concentrated when the share of inputs purchased from mining and agriculture is greater, confirming the previously reported results on natural advantage. An industry is also more concentrated the greater are the R&D/sales ratio and the proportion of skilled labor. Both of these are consistent with knowledge spillovers playing a role in the process of agglomeration. The latter may also be interpreted as being consistent with the labor market pooling hypothesis, since it is likely that skilled labor is more specialized than is unskilled labor. This would mean that both firms and workers would suffer more if close matches were difficult to find, which is the heart of the pooling analysis. Audretsch and Feldman also find that a greater mean shipping distance of output, which they interpret as a proxy for transport costs, increases concentration. This may be problematic, since if an industry were to concentrate for other reasons, it would find itself shipping its output a greater distance.

From an econometric perspective, the principal challenge Audretsch and Feldman (1996) face in this exercise is to control for the endogenous relationship between spatial concentration of an industry's innovative activity and the spatial concentration of the industry's production. They address this by re-estimating their model by 3SLS using natural resources and transportation costs as instruments for the innovation variable and university research as the instrument for the production variable. Results from this final exercise are quite similar to those from the ordinary least squares regressions.

Rosenthal and Strange (2001) also consider the determinants of agglomeration. This paper employs the Ellison and Glaeser (1997) index of spatial concentration instead of using the spatial G as in Audretsch and Feldman (1996). The difficulty with this statistic is that $G > 0$ does not necessarily imply that an industry is over-concentrated. If the industry was made up of a small number of large plants, and there was no agglomerative force – either an externality or a natural advantage – then G would take on a large value simply because of the industrial organization of the industry.

To address this problem, Ellison and Glaeser propose the following index of concentration: $\gamma = [G - (1 - \sum_i x_i^2)H]/[(1 - \sum_i x_i^2)(1 - H)]$. $H = \sum_j z_j^2$ is a Herfindahl index of the J plants in the industry, with z_j representing the employment share of the jth plant. For a perfectly competitive industry with a large number of small plants, H approaches zero and γ approaches $G/(1 - \sum_i x_i^2)$.[27] In this case, G measures spatial concentration without any contamination associated with industrial organization. More generally, γ takes on a value of zero when an industry is as concentrated as one would expect from a random location process, while a positive value of γ indicates excess concentration. The index is used to measure the level of spatial concentration among manufacturing industries at the zipcode, county, and state levels, in the fourth quarter of 2000.

These measures are regressed on a large number of industry characteristics, including proxies for the three Marshallian microfoundations and controls for transport costs and natural advantage. The Marshallian regressors include a measure of prior innovation in the industry and the use of both manufactured and service inputs. The model also includes variables that proxy the importance of labor market pooling by characterizing the degree of labor specialization in the industry, including labor productivity, the number of manager's per production worker, and the educational characteristics of an industry's workforce. The regressions are carried out using 4-digit manufacturing industries as observations. In addition, all of the regressions are estimated separately for concentration at the zipcode, county, and state levels, since the causes of agglomeration could well differ at different levels of geographic aggregation.

The paper's results provide evidence of the importance of all three sources of localization economies. The evidence is strongest for labor market pooling, with proxies having a positive impact on agglomeration at all levels of geography. The proxies for knowledge spillovers impact agglomeration positively only at the zipcode level. Reliance on manufactured inputs or natural resources positively affects agglomeration at the state

[27] The $(1 - \sum_i x_i^2)$ term is included in order that the index have the property that $E(\gamma) = 0$ when neither agglomerative spillovers nor natural advantage are present [see Ellison and Glaeser (1997) for details]. For the state, county, and zipcode levels that Rosenthal and Strange (2001) consider, $(1 - \sum_i x_i^2)$ is close to one, taking on values of 0.9997, 0.9954, and 0.9578, respectively.

level but has little effect on agglomeration at lower levels of geography. The same is true for the perishability of output, a proxy for transportation costs. In contrast, reliance on service inputs reduces state-level agglomeration. Taking all of these results together, an interesting pattern emerges, with industry attributes sensitive to shipping costs (reliance on manufactured inputs, reliance on natural resource inputs, marketing of perishable products) influencing agglomeration at the state level, knowledge spillovers impacting highly localized agglomeration, and labor impacting agglomeration at all levels of geography. These findings are largely robust, holding in both ordinary least squares (OLS) and 2-digit Standard Industry Classification (SIC) fixed effect specifications, with alternative MSA-based measures of geography, and when industries are aggregated from the 4-digit to the 3-digit level.

In related work, Lovely, Rosenthal and Sharma (2004) examine the spatial concentration of headquarter activity of exporters. Exporting requires specialized knowledge of foreign markets and should, therefore, contribute to spatial concentration. Lovely et al. test that idea by applying differencing methods to 4-digit industry-level data for the fourth quarter of 2000. Using data comparable to that in Rosenthal and Strange (2001), they find that when export related information is difficult to obtain, exporter headquarter activity is more highly agglomerated relative to headquarter activity in the domestic-only sector of the same industry. These findings support the idea that the need to acquire information contributes to agglomeration. However, the results do not identify the mechanism by which the need to acquire information contributes to agglomeration.

Another approach is taken by Dumais, Ellison and Glaeser (1997).[28] They look at the microfoundations of agglomeration economies by considering which industries co-agglomerate. Specifically, they consider the patterns of growth in industries, and look at the degree to which industries grow more robustly in locations where other industries use similar labor mixes, are related as suppliers of inputs or demanders of outputs, or are linked technologically. They consider several patterns of growth, including new firm births, old firm births, and closures (negative growth). They also consider a reduced form levels model, regressing the level of employment as a function of the environment created by co-agglomerated industries.

The key to this paper is its clever treatment of the relationships among industries. Input and output relatedness are defined relative to state level industrial mixes. Thus, if an industry employed a lot of fabricated metals, one would expect that if input sharing were an important agglomerative force, then the industry would be more likely to be found in states with a lot of fabricated metal production. Similarly, the fabricated metal producers would be expected to locate near their downstream consumers for the same reason. The labor mix fit is defined as the difference between an industry's employment shares by occupation and the state shares outside of the industry. This measure

[28] This is a deliberate reference to the working paper, rather than to the published paper of the same title, Dumais, Ellison and Glaeser (2002), since the latter omits the material on the sources of agglomeration economies.

is then interacted with a plant closure rate to directly address the notion that it is the risk of job loss that prompts labor market pooling. The knowledge spillover variable is calculated relative to industry mix. Two weighting schemes for intellectually related industries are employed. The first is Scherer's (1984) technology flow matrix, which was calculated from R&D and patent data. This estimates the degree to which innovative activity in one industry is likely to benefit another industry. Thus, it captures one sort of knowledge spillover: public good aspects of R&D activity. It does not really capture Marshall's "secrets of the trade". The other weighting scheme is calculated based on the co-ownership of firms in different industries. It is designed to capture scope economies, which are presumably related to the potential for intellectual spillover between industries. Again, this is a partial proxy, but as Dumais et al. note, knowledge spillovers are very difficult to capture.

These variables are used in the estimation of a number of growth models along with proxies for state level employment and standard agglomeration economies that lead firms in an industry to locate together. The model also includes state and industry fixed effects. The kinds of growth considered include new firm births, old firm births (both instances of expansion) and closure. Four sets of five year changes are used over the 1972–1992 period. The preferred specification is nonlinear (the log of 1 plus the change in employment) to deal with outliers.

While there is evidence that is consistent with all three Marshallian microfoundations, the strongest evidence by far is for the labor mix variable. The evidence on input sharing is mixed, with the presence of likely customers encouraging only new firm births, while the presence of likely input source encourages only old firm births (new plants). Neither has an impact on closures. The knowledge spillover variables have somewhat stronger effects, with the scope-economies proxy having more impact than the technology-flows proxy. One potential difficulty with this approach – recognized by Dumais et al. – is that in spatial equilibrium, the marginal impact of various sources of agglomeration economies on growth is zero, complicating efforts to uncover evidence of the sources of external economies. One way to deal with this is to regress levels of employment on the proxies. All of them are significant in this model, except for the presence of input suppliers.

Two recent papers have also addressed multiple microfoundations. Holmes (2002) considers the sources of agglomeration economies for sales offices. He employs differences between of high- and low-demand products to identify the degree to which market concentration and knowledge spillovers contribute to concentration. He finds that the former accounts for approximately half of the concentration of sales offices. Rigby and Essletzbichler (2002) consider impacts of various proxies for microfoundations on productivity using micro data from the LRD. These proxies are close to those used by Dumais, Ellison and Glaeser (1997), including measures of similarity for inputs, labor, and technology. They find evidence of all three Marshallian microfoundations. This paper's methodology clearly goes far beyond the old localization vs. urbanization question by focusing on the impacts of specific Marshallian agglomeration economies on productivity. In that sense, this paper approaches the ideal discussed at the beginning of this section.

Identifying the sources of agglomeration economies by looking across industries has the problem of constraining different industries to be affected by agglomeration in the same way. One way around this would be to focus on a particular industry. Holmes (1999) treatment of the textile industry is one example of this. A more extensive industry study is Klepper (2001), who considers the car industry. Klepper documents a number of patterns. First, the industry was dominated for the better part of a century by four highly successful companies that were founded in Detroit. Their founders had experience in related industries, including carriages and engines. Second, there were unusually many spinoffs in Detroit in the early stages of the industry (prior to 1916). Third, the spinoffs usually located near their parent firms. For instance, 50 of the 61 spinoffs of Detroit area automakers chose to locate in Detroit. Fourth, the spinoffs of successful companies themselves experienced greater success, a heritage effect. In fact, in a model of survival, although a Detroit dummy is significant in a simple model, it is not in a model that controls for corporate heritage. In addition, the spinoffs of successful companies have lower hazards even when they do not locate in Detroit.

Klepper's interpretation of these patterns is not favorable to standard theories of agglomeration. He argues that Marshallian agglomeration economies should have benefited all firms and not just the spinoffs of the early successful firms in the industry. On the other hand, the success of the early giants of the industry and the high rate of spinoffs in Detroit can be seen as being consistent with Marshallian notions of agglomeration. In any case, Klepper's work in this area is a compelling example of an agglomeration economy that operates through the specific channel of spinoffs. This would seem more consistent with a particular kind of knowledge spillovers than one that emphasizes input sharing or labor market pooling.[29]

4. Case evidence

Thus far, this chapter has considered the econometric evidence on the scope and sources of agglomeration economies. One of the chapter's themes has been the impact of increasingly refined data on the kinds of research being carried out. For instance, as noted in Section 2, the ability to geocode plant level data has made possible research that considers the geographic scope of agglomeration economies. This kind of analysis allows the study of the particulars of agglomeration economies, of which there are many.

Of course, even the most refined data set and most sophisticated econometric techniques will not be able to address all of the idiosyncratic conditions that contribute to

[29] Klepper and Simons' (2000a, 2000b) are analyses of the tire and television industries. Although agglomeration is not their central focus, like Klepper (2001), these papers make compelling cases for the use of industry studies in understanding industry location.

agglomeration. Thus, there is much to be learned about the nature of agglomeration from case studies. There is far more evidence of this kind than we can review in one brief section. We will, therefore, limit our discussion to three kinds of case studies, each shedding considerable light on the microfoundations of agglomeration economies: Hoover and Vernon (1959) and others on external economies in New York, Porter (1998) and others on clusters and competitiveness, and Saxenian (1994) on cultural and organizational differences between the Silicon Valley and Boston's Route 128.

4.1. The New York Metropolitan Region Study

The New York Metropolitan Region Study took place around 1960. It produced nine books, and a number of other reports. The motivation for the project was practical. The Regional Plan Association wanted to understand and thus forecast the growth of Greater New York City. It is not hard to see that the relevance of this project goes far beyond New York City. Two particularly notable volumes produced in this study were Hoover and Vernon (1959) and Vernon (1960), the latter being a summary.

The key idea from this project that concerns us here is external economies. These are defined as "… 'economies' which establishments obtain through sharing the services of specialists external to themselves" [Vernon (1960, p. 9)]. It is argued that these external economies have been central throughout New York's history. Early, traders benefited from scheduled sailings, wholesalers, and ship brokers not available elsewhere. Later, traders benefited from investments in the harbor, canals, and railroads. These shared inputs in trade later became foundations of shared inputs in finance, with maritime insurance underwriting forming a basis for other kinds of investment. Thus, static external economies affecting one industry became dynamic, and came to affect the city's entire economy.

The heart of urban sharing is communication. If it were possible to make use of Vernon's specialists at a distance, then incurring the costs of urbanization would be uneconomical. The reason that it is not possible to escape from the city in this way is that interactions are, in Vernon's usage, "unstable". A clothing designer does not know the button that best suits a dress until the dress is already under design. Only a particular button will do, and time is of the essence. Because of this, the designer communicates closely with supporting suppliers. It is this kind of interaction that is at the heart of external economies. Vernon reports that in a sample of "communication-oriented industries", 77.5% were located in New York City proper, a testament to the value of proximity [Vernon (1960, p. 123)]. A corollary of this is that urban concentration allows small firms to survive and even thrive. Vernon (1960, p. 110) reports that although the Core of the region contains 49.6% of manufacturing employment, but 61.2% of the region's employment at small firms (60 employees or fewer) and only 7.7% of the employment at large firms (more than 240 employees).

The analysis of external economies in the New York Metropolitan Region Study provides insights that complement the findings of the econometric work that we have reviewed in this chapter. The treatment of the central role of the transportation sector is one example. It shows how one specific industry, shipping, fostered the development of another, mercantile trade, and ultimately led to the development of yet another industry, finance. These are the kinds of specific relationships between various economic activities that can be handled nicely in a case analysis, but may get lost in formal econometrics.

The New York Metropolitan Region Study has also been valuable as a source of inspiration for econometric work. Vernon's analysis of the industrial organization of New York's textile industry has the important conclusion that small firms and large cities naturally go together. As noted above, Holmes and Stevens (2002) look at this relationship across industries and across the county. Somewhat surprisingly, they find that plants in concentrated environments are on average larger rather than smaller. Their interpretation of this result is that it is the increased productivity associated with the concentration that allows the plants to become larger. Vernon's result is shown to apply to a specific industry in a specific city, with the lesson being that as valuable as case analysis may be, it is dangerous to draw universal conclusions from a single case.

4.2. Regional Clusters of Innovation Project

This research has a practical motivation that parallels the New York Metropolitan Region Study, with the goal being to achieve an understanding of the forces that govern competitiveness in the hope of fostering it. The basic approach is found in Porter (1990). The heart of the approach is that a business will become more productive when factor markets are favorable, when suppliers are available, when consumers are demanding, and when competitive pressures compel sustained innovation.

The reason that this analysis is relevant here is that these conditions are more likely to be present when an industry is spatially concentrated [Porter (1998)]. The concentration is referred to as a "cluster". The Cluster Mapping Project[30] offers the following definition:

> A cluster is a geographically proximate group of interconnected companies and associated institutions in a particular field, including product producers, service providers, suppliers, universities, and trade associations. Clusters arise out of the linkages or externalities that span across industries in a particular location.

The Cluster Mapping Project identifies 41 clusters in North America, including, for example, Aerospace engines, with 18.26% of national employment in Hartford, and another 18.0% in Cincinnati and Phoenix together.

Recently, the Regional Clusters of Innovation Project has looked at the competitiveness of five specific metropolitan areas: San Diego, Wichita, North Carolina's Research Triangle, Pittsburgh, and Atlanta. As of this writing, preliminary reports are available for the first three, as is a national report.[31] These reports provide detailed portraits of

[30] http://data.isc.hbs.edu/isc/cmp_data_glossary.jsp.
[31] http://www.compete.org.

Twenty-first Century cities in much the same way that the New York Metropolitan Region Project painted a picture of post-war New York.

This analysis adds to the econometric work on agglomeration in several ways. First, by focusing on clusters, which are explicitly defined to include related industries, some of the problems associated with a narrow treatment of industry scope (see Section 2) are mitigated. Second, by focusing on policy towards agglomeration, the projects show the importance of institutions for the process of agglomeration. For instance, it is argued that all levels of government have a potential to foster productive patterns of agglomeration. Local taxes and state funding for universities are two examples. Informal organizations can have an impact too. One example of this is the alumni organization UCSD CONNECT, which helps establish and maintain links that can be valuable in rapidly changing industries such as pharmaceuticals. The bottom line of this analysis is that institutions matter for agglomeration, a point that is only tangential to most of the econometric literature.

4.3. Regional advantage

Probably the most famous instance of clustering is the Silicon Valley. That the Santa Clara Valley was not the only possible center for the computer industry is the subject of a joke among programmers, who refer to the location of their jobs relative to the Valley's fictitious silicon mines. Saxenian (1994) deals with the forces that operated on the industry in the late-20th century. She points out that in the mid-1970s, both Boston (especially around Route 128) and the San Jose to Palo Alto corridor (the so-called Silicon Valley) were essentially equal in their positions as centers of electronics and high-technology. The next decade witnessed a movement offshore of semiconductor production, which hurt the Silicon Valley, and a shift away from minicomputers, which hurt Route 128. The Silicon Valley made its well-known and successful transition to software and other computer related industries. Route 128 had more difficulties. There are two explanations for this divergence. One is that either location could have become dominant in software based on its characteristics, but that the random hand of history selected the Silicon Valley as the industry core. The other explanation is that the locations did not have identical characteristics, and that the Silicon Valley offered a more productive environment.

Both locations had many of the characteristics that could be expected to attract high-technology employment including educated workforces and proximity to research universities. Saxenian argues that the key difference between the Silicon Valley and Route 128 is in their industrial systems. In her view [Saxenian (1994, p. 7)], a local industrial system has "three dimensions: local institutions and culture, industrial structure and corporate organization". Regarding the Silicon Valley, she maintains [Saxenian (1994, p. 37)] that

> The decentralized and fluid environment accelerated the diffusion of technological capabilities and know-how within the region… When engineers moved between

> companies, they took with them the knowledge, skills, and experience acquired at their previous jobs.

In contrast, Route 128 is presented by Saxenian as being relatively rigid and hierarchical, while the Silicon Valley is presented as being flexible and entrepreneurial. This certainly seems to be the view of the industry. Saxenian quotes Jeffrey Kalb, an entrepreneurial refugee from the Digital Electronics corporation:

> There's a fundamental difference in the nature of the industry between Route 128 and [the Silicon Valley]. Route 128 is organized into large companies that do their own thing... It's very difficult for a small company to survive in that environment... The Valley is very fast-moving and start-ups have to move fast. The whole culture of the Valley is one of change. We laugh about how often people change jobs. The joke is that you can change jobs and not change parking lots. There's a culture associated with that which says that moving is okay, that rapid change is the norm, that it's not considered negative on your resume... So you have this culture of rapid decisions, rapid changes, which is exactly the environment that you find yourself in as a startup.

Thus, the key difference between the two locations is neither a natural advantage (silicon mines) nor a traditional agglomeration economy. It is instead a difference in the organization of resources that drives the difference. The difficulty of quantifying these sorts of organizational differences [Rosenthal and Strange (2003) notwithstanding] are one of the reasons that a case approach like Saxenian's is valuable.

Given his importance in the study of agglomeration, it seems appropriate to give the last word here to Marshall, who in some sense appreciated the importance of intangible aspects of the economic environment as implied at least loosely by Marshall in 1920 (1920, p. 270):

> We have seen how physical nature acts on man's energies... but we have also seen how the use he makes of these advantages depends on his ideals of life, and how inextricably therefore the religious, political and economic threads of the world's history are interwoven.... .

5. Conclusion

This chapter has considered the empirical literature on agglomeration economies. Some of the questions in the literature are relatively old, such as the debate over whether localization or urbanization economies are of greater importance. Other questions are relatively new, such as the consideration of the geographic, temporal, and organizational dimensions of agglomeration economies. Likewise, attempts to understand the microfoundations of agglomeration economies are also relatively recent.

Whether the questions are old or new, the answers are potentially of considerable importance for policy. For instance, the increased integration and size of the European

Union has led to interest in the degree to which a core–periphery regional development pattern is likely to emerge. Since this depends on the nature of agglomeration economies, so too should regional development policy. Similar issues confront policy-makers in the rest of the world. For example, the Silicon Valley is perceived as a success that other locations would like to duplicate. Doing so obviously depends on the nature and sources of agglomeration economies.

As this review indicates, there is a lot that we do not yet know about agglomeration economies. We believe, however, that recent developments in the literature gives cause for optimism. Increasingly, researchers have made use of large micro data sets that have allowed for ever more refined studies of agglomeration. Such data sets allow for more reliable estimation, and also help resolve a variety of econometric problems. They also allow for the consideration of issues like the micro geographic scope of agglomeration that earlier data sets could not address. We are, therefore, confident that the impressive progress of recent years will continue, closing the gaps in knowledge that this review has identified.

Acknowledgements

We are grateful to Gilles Duranton, Vernon Henderson, Jacques Thisse and the participants of a presentation at the North American Regional Science Association Meetings in November, 2002. Any errors are ours alone. We are also grateful for the financial support of the National Institute of Aging, the Connaught Fund at the University of Toronto, and the Social Sciences and Humanities Research Council of Canada.

References

Acemoglu, D., Angrist, J. (1999). “How large are the social returns to education? Evidence from compulsory schooling laws”. NBER Working Paper No. 7444.

Acs, Z.J., Audretsch, D.B., Feldman, M.P. (1992). “Real effects of academic research: Comment”. American Economic Review 82 (1), 363–367.

Ades, A.F., Glaeser, E.L. (1995). “Trade and circuses: Explaining urban giants”. Quarterly Journal of Economics 110, 195–227.

Almeida, P., Kogut, B. (1999). “Localization of knowledge and the mobility of engineers in regional networks”. Management Science 45 (7), 905–917.

Arrow, K.J. (1962). “Economic welfare and the allocation of resources for invention”. In: Nelson, R.R. (Ed.), The Rate and Direction of Inventive Activity. Princeton Univ. Press, Princeton, NJ, pp. 609–626.

Audretsch, D.B., Feldman, M.P. (2004). “Knowledge spillovers and the geography of innovation”. In: Henderson, J.V., Thisse, J.-F. (Eds.), Handbook of Regional and Urban Economics, vol. 4. Elsevier, Amsterdam, pp. 2713–2739. This volume.

Audretsch, D.B., Feldman, M.P. (1996). “R&D spillovers and the geography of innovation and production”. American Economic Review 86 (3), 630–640.

Bartlesman, E.J., Caballero, R.J., Lyons, R.K. (1994). “Customer and supplier-driven externalities”. American Economic Review 84 (4), 1075–1084.

Baumgardner, J.R. (1988). "Physicians' services and the division of labor across local markets". Journal of Political Economy 96 (5), 948–982.

Blomquist, G., Berger, M., Hoehn, J. (1988). "New estimates of the quality of life in urban areas". American Economic Review 78, 89–107.

Calem, P.S., Carlino, G.A. (1991). "Urban agglomeration economies in the presence of technical change". Journal of Urban Economics 29, 82–95.

Carlton, D.W. (1983). "The location and employment choices of new firms: An econometric model with discrete and continuous endogenous variables". Review of Economics and Statistics 65, 440–449.

Charlot, S., Duranton, G. (2002). "Communication externalities in cities". London School of Economics Working Paper.

Chinitz, B.J. (1961). "Contrasts in agglomeration: New York and Pittsburgh". American Economic Review 51, 279–289.

Ciccone, A. (2002). "Agglomeration effects in Europe". European Economic Review 46, 213–227.

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

Cingano F., Schivardi F. (2003). "Identifying the sources of local productivity growth". Journal of the European Economic Association. In press.

Combes, P.-P. (2000a). "Economic structure and local growth, France, 1984–1993". Journal of Urban Economics 47, 329–355.

Combes, P.-P. (2000b). "Marshall–Arrow–Romer externalities and city growth". CERAS Working Paper No. 99-06.

Combes, P.-P., Duranton, G., Gobillon, L. (2003). "Wage differences across French local labor markets: Endowments, skills, and interactions". Working Paper.

Costa, D.L., Kahn, M.E. (2001). "Power couples". Quarterly Journal of Economics 116, 1287–1315.

Davis, D.R., Weinstein, D.E. (1996). "Does economic geography matter for international specialization?" NBER Working Paper No. 5706.

Davis, D.R., Weinstein, D.E. (1999). "Economic geography and regional production structure: An empirical investigation". European Economic Review 43, 379–407.

Dekle, R., Eaton, J. (1999). "Agglomeration and land rents, evidence from the prefectures". Journal of Urban Economics 46, 200–214.

Di Addario, S., de Blasio, G. (2003). "Labor market pooling: Evidence from the Italian industrial districts". Bank of Italy Working Paper.

Diamond, C.A., Simon, C.J. (1990). "Industrial specialization and the returns to labor". Journal of Labor Economics 8 (2), 175–201.

Dumais, G., Ellison, G., Glaeser, E. (1997). "Geographic concentration as a dynamic process". NBER Working Paper No. 6270.

Dumais, G., Ellison, G., Glaeser, E. (2002). "Geographic concentration as a dynamic process". Review of Economics and Statistics 84 (2), 193–204.

Duranton, G., Overman, H.G. (2002). "Testing for localisation using micro-geographic data". London School of Economics Working Paper.

Duranton, G., Puga, D. (2001a). "Nursery cities: Urban diversity, process innovation, and the life-cycle of products". American Economic Review 91 (5), 1454–1477.

Duranton, G., Puga, D. (2001b). "From sectoral to functional urban specialisation". University of Toronto Economics Department Working Paper.

Duranton, G., Puga, D. (2004). "Micro-foundations of urban agglomeration economies". In: Henderson, J.V., Thisse, J.-F. (Eds.), Handbook of Regional and Urban Economics, vol. 4. Elsevier, Amsterdam, pp. 2063–2117. This volume.

Eberts, R.W., McMillen, D.P. (1999). "Agglomeration economies and urban public infrastructure". In: Cheshire, P., Mills, E.S. (Eds.), Handbook of Regional and Urban Economics, vol. 3. North-Holland, New York, pp. 1455–1495.

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 an industry: Does natural advantage explain agglomeration?" American Economic Association Papers and Proceedings 89, 311–316.
Evans, D.S. (1987). "Tests of alternative theories of firm growth". Journal of Political Economy 85 (4), 657–674.
Fallick, B.C., Fleischman, C., Rebitzer, J. (2003). "Job hopping in the Silicon Valley: The micro-foundations of an industry cluster". Case Western Reserve University Working Paper.
Florida, R., Gates, G. (2001). Technology and Tolerance: The Importance of Diversity to High-Tech Growth. Brookings Institution, Washington, DC.
Fogarty, M.S., Garofalo, G.A. (1978). "Urban spatial structure and productivity growth in the manufacturing sector of cities". Journal of Urban Economics 23, 60–70.
Fuchs, V. (1962). Changes in the Location of Manufacturing in the U.S. Since 1929. Yale Univ. Press, New Haven.
Gabriel, S.A., Rosenthal, S.S. (2004). "Quality of the business environment versus quality of life: Do firms and households like the same cities?" The Review of Economics and Statistics 86 (1).
George, L., Waldfogel, J. (2003). "Who benefits whom in daily newspaper markets?" Journal of Political Economy. In press.
Glaeser, E.L., Maré, D.C. (2001). "Cities and skills". Journal of Labor Economics 19 (2), 316–342.
Glaeser, E., Kolko, J., Saiz, A. (2001). "Consumer city". Journal of Economic Geography 1, 27–50.
Glaeser, E.L., Kallal, H.D., Scheinkman, J.A., Shleifer, A. (1992). "Growth in cities". Journal of Political Economy 100, 1126–1152.
Gyourko, J., Tracy, J. (1991). "The structure of local public finance and the quality of life". Journal of Political Economy 99 (4), 774–806.
Hanson, G.H. (1998a). "Regional adjustment to trade liberalization". Regional Science and Urban Economics 28 (4), 419–444.
Hanson, G.H. (1998b). "North American economic integration and industry location". Oxford Review of Economic Policy 14 (2), 30–44.
Hanson, G.H. (2001). "Scale economies and the geographic concentration of industry". Journal of Economic Geography 1, 255–276.
Head, K., Mayer, T. (2004). "The empirics of agglomeration and trade". In: Henderson, J.V., Thisse, J.-F. (Eds.), Handbook of Regional and Urban Economics, vol. 4. Elsevier, Amsterdam, pp. 2609–2669. This volume.
Helsley, R.W., Strange, W.C. (1990). "Agglomeration economies and matching in a system of cities". Regional Science and Urban Economics 20, 189–212.
Helsley, R.W., Strange, W.C. (2001). "Innovation and input sharing". Journal of Urban Economics 51 (1), 25–45.
Helsley, R.W., Strange, W.C. (2002). "Knowledge barter". University of Toronto Faculty of Management Working Paper.
Henderson, J.V. (1986). "Efficiency of resource usage and city size". Journal of Urban Economics 19, 47–70.
Henderson, J.V. (1997). "Externalities and industrial development". Journal of Urban Economics 42, 449–470.
Henderson, J.V. (2003a). "Marshall's scale economies". Journal of Urban Economics 53, 1–28.
Henderson, J.V. (2003b). "The urbanization process and economic growth: The so-what question". Journal of Economic Growth 8, 47–71.
Henderson, J.V., Kuncoro, A., Turner, M. (1995). "Industrial development in cities". Journal of Political Economy 103, 1067–1085.
Holmes, T.J. (1998). "The effects of state policies on the location of manufacturing: evidence from state borders". Journal of Political Economy 106 (4), 667–705.
Holmes, T.J. (1999). "Localization of industry and vertical disintegration". Review of Economics and Statistics 81 (2), 314–325.

Holmes, T.J. (2002). "The role of cities: evidence from the placement of sales offices". Federal Reserve Bank of Minneapolis Research Department Staff Report 298.

Holmes, T.J., Stevens, J.J. (2002). "Geographic concentration and establishment scale". Review of Economics and Statistics 84 (4), 682–691.

Holtz-Eakin, D. (1994). "Public sector capital and the productivity puzzle". Review of Economics and Statistics 76 (1), 12–21.

Hoover, E.M. (1948). The Location of Economic Activity. McGraw–Hill, New York.

Hoover, E.M., Vernon, R. (1959). Anatomy of a Metropolis. Harvard Univ. Press, Cambridge, MA.

Jacobs, J. (1969). The Economy of Cities. Vintage, New York.

Jaffe, A.B. (1989). "Real effects of academic research". American Economic Review 79 (5), 957–970.

Jaffe, A.B., Trajtenberg, M., Henderson, R. (1993). "Geographic localization of knowledge spillovers as evidenced by patent citations". Quarterly Journal of Economics 108, 577–598.

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 (4), 881–908.

Kim, S. (1999). "Regions, resources and economics geography: Sources of U.S. regional comparative advantage, 1880–1987". Regional Science and Urban Economics 29, 1–32.

Klepper, S., (2001). "The evolution of the U.S. automobile industry and Detroit as its capital". Carnegie Mellon Working Paper.

Klepper, S., Simons, K.L. (2000a). "Dominance by birthright: Entry of prior radio producers and competitive ramifications in the U.S. television receiver industry". Strategic Management Journal 21, 997–1016.

Klepper, S., Simons, K.L. (2000b). "The making of an oligopoly: Firm survival and technological change in the evolution of the U.S. tire industry". Journal of Political Economy 108 (4), 728–760.

Kolko, J. (1999). " Can I get some service here? Information technology, service industries, and the future of cities". Harvard University Working Paper.

Krugman, P.R. (1980). "Scale economies, product differentiation, and the pattern of trade". American Economic Review 70, 950–959.

Krugman, P.R. (1991). Geography and Trade. MIT Press, Boston.

Lovely, M., Rosenthal, S., Sharma, S. (2004). "Information, agglomeration and the headquarters of U.S. exporters". Regional Science and Urban Economics. In press.

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

Moomaw, R.L. (1981). "Productivity and city size: a critique of the evidence". Quarterly Journal of Economics 96, 675–688.

Moomaw, R.L. (1983). "Is population scale a worthless surrogate for business agglomeration economies?" Regional Science and Urban Economics 13, 525–545.

Moretti, E. (2000). "Estimating the social return to higher education". NBER Working Paper No. 9018.

Moretti, E. (2004). "Human capital externalities in cities". In: Henderson, J.V., Thisse, J.-F. (Eds.), Handbook of Regional and Urban Economics, vol. 4. Elsevier, Amsterdam, pp. 2243–2291. This volume.

Nakamura, R. (1985). "Agglomeration economies in urban manufacturing industries: a case of Japanese cities". Journal of Urban Economics 17, 108–124.

Ono, Y. (2002). "Outsourcing business services and the scope of local markets". Federal Reserve Bank of Chicago Working Paper.

Porter, M. (1990). The Competitive Advantage of Nations. Free Press, New York.

Porter, M. (1998). "Clusters and the new economics of competition". Harvard Business Review (November–December).

Rauch, J. (1993a). "Productivity gains from geographic concentration of human capital: Evidence from the cities". Journal of Urban Economics 34, 380–400.

Rauch, J. (1993b). "Does history matter only when it matters little?" Quarterly Journal of Economics 108, 843–867.

Rigby, D.L., Essletzbichler, J. (2002). "Agglomeration economies and productivity differences in U.S. cities". Journal of Economic Geography 2, 407–432.

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

Romer, P. (1986). “Increasing returns and long run growth”. Journal of Political Economy 94, 1002–1037.
Rosen, S. (1979). “Wage-based indexes of urban quality of life”. In: Mieszkowski, P., Straszheim, M. (Eds.), Current Issues in Urban Economics. Johns Hopkins Univ. Press, Baltimore, pp. 391–429.
Rosenthal, S.S., Strange, W.C. (2001). “The determinants of agglomeration”. Journal of Urban Economics 50, 191–229.
Rosenthal, S.S., Strange, W.C. (2002). “The urban rat race”. Syracuse University Working Paper.
Rosenthal, S.S., Strange, W.C. (2003). “Geography, industrial organization, and agglomeration”. Review of Economics and Statistics 85 (2), 377–393.
Saxenian, A. (1994). Regional Advantage: Culture and Competition in Silicon Valley and Route 128. Harvard University, Cambridge, MA.
Scherer, F. (1984). “Using linked patent data and R&D data to measure technology flows”. In: Griliches, Z. (Ed.), R&D, Patents, and Productivity. Univ. of Chicago Press, Chicago.
Segal, D. (1976). “Are there returns to scale in city size?” Review of Economics and Statistics 58, 339–350.
Shefer, D. (1973). “Localization economies in SMSAs: A production function analysis”. Journal of Regional Science 13, 55–64.
Simon, C.J. (1988). “Frictional unemployment and the role of industrial diversity”. Quarterly Journal of Economics 103, 715–728.
Sveikauskas, L.A. (1975). “The productivity of cities”. Quarterly Journal of Economics 89, 393–413.
Tabuchi, T. (1986). “Urban agglomeration, capital augmenting technology, and labor market equilibrium”. Journal of Urban Economics 20, 211–228.
Tabuchi, T., Yoshida, A. (2000). “Separating agglomeration economies in consumption and production”. Journal of Urban Economics 48, 70–84.
Vernon, R. (1960). Metropolis 1985. Harvard Univ. Press, Cambridge, MA.
Waldfogel, J. (2003). “Preference externalities: An empirical study of who benefits whom in differentiated product markets”. RAND Journal of Economics. In press.
Wheaton, W.C., Lewis, M.J. (2002). “Urban wages and labor market agglomeration”. Journal of Urban Economics 51, 542–562.
Williamson, J. (1965). “Regional inequality and the process of national development”. Economic Development and Cultural Change 13, 3–45.