The magnitude and causes of agglomeration economies*

Diego Puga§
IMDEA, Universidad Carlos III and *CEPR*

April 2009

Abstract: Firms and workers are substantially more productive in large and dense urban environments. There is substantial evidence of such agglomeration economies based on a clustering of production beyond what can be explained by chance or the heterogeneity of space, based on the spatial patterns of wages and rents, and based on the direct measurement of how productivity varies with the urban environment. There is, however, much more to be learned about the causes of agglomeration economies. We have good models of agglomeration through sharing and matching, but not a deep enough understanding of learning in cities. Despite recent progress, work distinguishing between alternative causes empirically is still in its infancy.

Key words: agglomeration economies.
JEL classification: R40

*This paper has been written for the *Journal of Regional Science*'s 50th Anniversary Symposium.

§Madrid Institute for Advanced Studies (IMDEA) Social Sciences, Antiguo pabellón central del Hospital de Cantoblanco, Carretera de Colmenar Viejo km. 14, 28049 Madrid, Spain (e-mail: diego.puga@imdea.org; website: http://diegopuga.org).

1. Introduction

The concentration of firms and people in cities creates congestion and bids up the price of land, and yet three quarters of the world population live in cities. What are the advantages of cities, that are able to offset the obvious costs and attract so many enterprises and workers to them? How large are these advantages and why do they arise? These are among the most fundamental questions in urban economics, since without answering them we cannot understand the very existence of cities. This paper reviews what we know about the magnitude and causes of the productive advantages of cities and also tries to identify the largest gaps in our knowledge of agglomeration economies.

Firms and workers are much more productive in large and dense urban environments than in other locations. It is also in large cities where the vast majority of substantial innovations emerge. The productivity advantages of cities and urban clusters with a high density of firms and workers have been perceived for a long time, and already received the attention of Adam Smith (1776) and Alfred Marshall (1890). Over the past thirty years, urban economists have been rather successful at documenting and quantifying these advantages. They have done so with three broad approaches. First, by showing that productive activities are much more clustered than would be expected if location was the result of a random outcome or merely reflected underlying differences across space leading to comparative advantage. Second, by studying spatial patterns in wages and land rents. If firms and workers are mobile and wages and land rents differ across space, higher wages and land rents in large and dense urban environments must reflect some productive advantage. Third, by looking at systematic variations in productivity across space. This last and most direct approach been particularly fruitful in recent years, thanks in part to the increasing availability of spatial data at the level of individual plants and workers.

Stories about the causes of agglomeration economies are as old as the realization that such advantages exist. The aforementioned classic works by Smith and Marshall contain frequently cited discussions of the advantages arising from the greater specialization made possible by larger markets, from sharing intermediate suppliers, from pooling in labour markets, or from the localized transmission of ideas. However, understanding the causes of agglomeration economies requires going much deeper than these classic stories or the many more than have followed. For instance, we may perceive some advantages from having a large local labour market, but what is the precise channel through which such advantages operate? Is it because a larger labour market improves matching between employers and employees? Or is it because large concentrations of employment iron out idiosyncratic shocks and improve establishments' ability to adapt their employment to good and bad times? Or perhaps because larger markets allow workers to specialize

in a narrower set of activities and improve their performance? And how important are these advantages relative through alternative sources of agglomeration economies not operating through the labour market? To answer such questions we need good models that formalize the microeconomic foundations of urban agglomeration economies, as well as detailed empirical work able to identify and quantify the precise mechanisms at work. This is an area where there has also been much recent progress. However, as we shall discuss in detail below, there are very substantive open questions that forthcoming research ought to address.

2. Evidence and magnitude of agglomeration economies

Excessive localization as a sign of agglomeration economies

One of the fundamental results in spatial economics is Starrett's (1978) spatial impossibility theorem. This states that, once we abstract from the heterogeneity of the underlying space, without indivisibilities or increasing returns, any competitive equilibrium in the presence of transport costs will feature only fully autarchic locations where every good will be produced at small scales (see Ottaviano and Thisse, 2004, for a detailed discussion). Thus, a substantial localization or spatial concentration of economic activity can be seen as a sign of agglomeration economies.[1]

Starrett's (1978) theorem already points to two reasons why, even in the absence of scale economies at the level of cities, one would not expect a uniform distribution of economic activity. First, because there is some inevitable lumpiness from small-scale indivisibilities in any production process. For instance, most technologies require production establishments within a certain size range. If we were to pin a map on the wall and, blindfolded, throw one dart at it for each establishment that exists in a sector, some areas would inevitably get several darts and others none. The 'dartboard approach' of Ellison and Glaeser (1997) yields measures of geographic concentration that correct for both the size distribution of plants and for differences in the size of the geographic areas. The second reason why, even without agglomeration economies, industries would not be uniformly distributed across space is that space itself is not uniform. Some areas are too arid or too rugged to be used. Furthermore, other activities are closely tied to geographic features such as natural resources and ports. However, Ellison and Glaeser (1999) find that natural advantages, even when very widely defined, can predict only about 20% of

[1]The terms localization and spatial concentration are used interchangeably. Note, however, that localization and specialization are related but distinct concepts. We measure localization for individual industries or activities and say that an industry is localized if it is concentrated in few places. We measure specialization for individual cities or geographical areas and say that an area is specialized if its employment or establishments are concentrated in few sectors.

industrial localization. Panel-data studies using area fixed-effects to capture any sort of localized advantage (e.g., Henderson, 1997) also find that such permanent advantages leave substantial agglomeration effects unexplained.

Duranton and Overman (2005) note three desirable properties of the spatial concentration index of Ellison and Glaeser (1997): it is comparable across industries, it controls for the concentration of overall economic activity when looking at individual sectors, and it controls for the distribution of plant sizes. They also suggest two additional desirable properties for a localization measure. One is to be able to accompany measures of localization with measures of statistical significance. Another desirable property is to avoid the bias created by aggregating data into large areas (boxes) instead of individual locations (points) — the so-called modifiable area unit problem. If a cluster of firms spreads across the border between two statistical units for which data is available, the close proximity of firms will be missed at least in part by localization indices based on area aggregates. We will see firms split across the two areas but not that firms are located within those areas so that they are close to each other — or even that the areas share a border, since areas are typically treated symmetrically. To get around this problem, they propose a distance-based approach that looks at the entire distribution of pairwise distances between plants and compares them to the distribution resulting from random allocations of plants. They find using data for the United Kingdom that about half of all 4-digit sectors are too localized at a 5% confidence level for their clustering to reflect merely a random outcome. In addition, localization mostly takes place within close distances — less than 50 kilometres.

Instead of looking at location in general, Rosenthal and Strange (2003) focus on the location decisions of new establishments, which are less constrained by past characteristics. Looking at the production environment in concentric rings of varying size around the zip code of new establishments, they also find that agglomeration effects decrease quite rapidly with distance.

Quantifying agglomeration economies through wages and rents

Comparing wages across areas provides more direct evidence of the existence and magnitude of agglomeration economies (see, e.g., Glaeser and Maré, 2001, Wheaton and Lewis, 2002, Combes, Duranton, and Gobillon, 2008, Combes, Duranton, Gobillon, and Roux, 2009*b*). This approach rests on the assumption that in competitive markets labour is paid the value of its marginal product. However, even if labour markets are not perfectly competitive, higher wages in large/dense urban areas can be seen as evidence of higher productivity. For workers, higher wages may be offset by larger commuting costs and housing costs. However, higher wages and land rents in large cities would lead firms to relocate elsewhere unless there were some significant productive advantages.

A key concern when interpreting the urban wage premium as evidence of agglomeration economies is that the ability of workers may also vary across cities. If more able workers sort into larger cities, then the urban wage premium may reflect this greater abilities instead of any intrinsic advantage to urban location. Glaeser and Maré (2001) discuss this issue at length and explore a number of solutions: controlling for observable skills, instrumenting for urban residence using the parental background, and finally exploiting the panel dimension of their data to include individual worker fixed-effects. They find that, even after all these corrections, there is significant wage premium associated with living and working in dense cities, although substantially smaller in magnitude than before taking unobserved ability into account. Glaeser and Maré (2001) argue that this drop in magnitude may be partly due to the fact that it takes time for workers to fully reap the benefits of locating in big cities. They find some support for this argument by looking at the wage path of rural-urban migrants. Working with French data, Combes *et al.* (2009*b*) suggest that accounting for differences in observed and unobserved ability using worker fixed-effects may reduce estimates of the magnitude of agglomeration economies by about one half. Their preferred estimate, when accounting both for omitted ability and the endogeneity of city sizes discussed below, yields an elasticity of wages with respect to urban density of 0.02.

An related approach is to study the spatial variation of rents. If firms are willing to pay higher rents in big cities it must be because there is some compensating productive advantage. A difficulty with this stretegy is finding the required data on comercial rents, so residential rents are sometimes used as a proxy (Dekle and Eaton, 1999).

Following (Roback, 1982), a particularly fruitful approach is to combine data on wages and rents. The beauty of her framework is that it helps disentangle the consumption amenities from the productive advantages of big cities. For workers, higher wages make them better off whereas higher rents make them worse off. Thus, greater consumption amenities in a city will make workers willing to accept lower wages and higher rents. For firms, both higher wages and higher rents mean increased costs. Thus, localized productive advantages will make firms willing to accept higher wages and higher rents. Consequently, both consumption amenities and productive advantages should be associated with higher rents. However, consumption amenities should be associated with lower wages whereas productive advantages should be associated with higher wages. Note, however, that this raises an additional concern when looking for agglomeration effects through wages. If big cities are associated with both better amenities and higher productivity, the net effect on wages may be ambiguous.

Productivity evidence of local increasing returns: more output out of the same inputs

By definition, local external scale economies imply that plants are able to produce more output with the same inputs in larger, denser, urban environments. Thus, perhaps the most natural way to directly measure the magnitude of agglomeration economies is to use data on outputs and inputs to estimate how productivity varies across space. The first influential modern study to do this, by Sveikauskas (1975), regressed log output per worker in a cross-section of city-industries on log city population and found that a doubling of population increases output per worker by about 6 percent.

Output per worker may not be the best measure of productivity in this context. As Moomaw (1983) points out, if capital is used more intensively in large cities, then estimating agglomeration economies using output per worker will lead to an upward bias in the estimated effects. For this reason the literature has moved towards using total factor productivity, calculated at the aggregate level for each area being considered or, more recently, at the plant level. A fundamental contribution to quantifying agglomeration economies through individual productivity estimates is Henderson (2003), who uses plant level data in high-tech and machinery sectors for the United States and is particularly careful in dealing with potential biases.

Another problem raised by early estimates of the magnitude of agglomeration economies, also discussed by Moomaw (1981), is that higher output per worker may be not so much a consequence of higher local employment or density but its cause. If a location has an underlying productive advantage, then it will tend to attract more firms and workers and become larger as a result. Following Ciccone and Hall (1996), the standard way to tackle this issue is to instrument for current size/density. Since there is substantial persistence in the spatial distribution of population but the drivers of high productivity today differ from those in the distant past, the usual instruments are historical data for size/density as well as characteristics that are thought to have affected the location of population in the past but that are mostly unrelated to productivity today. Ciccone and Hall (1996) find that reverse causality on this matter is only a minor issue. This conclusion has been confirmed by much of the subsequent literature. In a recent contribution, Combes *et al.* (2009*b*), use a wide range of historical and geological instruments — the latter justified because fertile soil was an important attraction for population in the past, cities with large historical populations tend to remain large today, but fertile soil no longer matters much for local productivity. They conclude that instrumenting only reduces the estimated magnitude of agglomeration economies by about one fifth. An alternative strategy to deal with a potential endogeneity bias is to exploit the time dimension of panel data. In particular, Henderson (2003) includes city-time fixed effects when estimating plant-level productivity, to capture any unobserved attributes that may

have attracted more entrepreneurs to a given city.

Finally, productivity estimates are also subject to the caveats discussed above in relation to unobserved differences in the quality of labour. However, it is in studies based on wages where these issues have so far been addressed.

After dealing with these potential concerns to different extents, the magnitudes that emerge from productivity studies suggest that a doubling of city size increases productivity by between 3 and 8 percent for a large range of city sizes.

Agglomeration or survival of the fittest?

While the productive advantages of large cities have usually been attributed to agglomeration economies (i.e., larger cities promote interactions that increase productivity), recently an alternative explanation has been offered based on 'firm selection' (i.e., larger cities toughen competition allowing only the most productive firms to survive). Melitz and Ottaviano (2008) model this argument by incorporating endogenous price-cost mark-ups in the framework of Melitz (2003). They show that the presence of more firms in larger markets makes competition tougher and this leads less productive firms to exit. Thus, higher average productivity in larger cities could result from a survival of the fittest rather than from a productivity boost based on agglomeration economies.

In a recent paper, Combes, Duranton, Gobillon, Puga, and Roux (2009*a*) develop a framework to distinguish between agglomeration and firm selection in explaining why average productivity is higher in larger cities. They nest a generalized version of the firm selection model of Melitz and Ottaviano (2008) (freed of distributional assumptions and extended to many cities) and a simple model of agglomeration in the spirit of Fujita and Ogawa (1982) and Lucas and Rossi-Hansberg (2002). This nested model enables them to parameterise the relative importance of agglomeration and selection.

The main prediction of the nested model is that, while selection and agglomeration effects both make average firm log productivity higher in larger cities, they have different predictions for how the shape of the log productivity distribution varies with city size. In particular, stronger selection effects in larger cities should lead to a greater left truncation of the distribution of firm log productivities in larger cities, as the least productive firms exit. Stronger agglomeration effects in larger cities should lead instead to a greater rightwards shift of the distribution of firm log productivities in larger cities, as agglomeration effects make all firms more productive.

While most models of agglomeration feature a representative firm, the model of Combes *et al.* (2009*a*) features heterogeneous firms and allows the benefits of agglomeration to vary systematically across firms within each city. In particular, through a simple technological complementarity between the productivity of firms and that of

workers, firms that are more productive per se are also better at reaping the benefits of agglomeration. The model predicts that if this additional effect holds in practice then agglomeration should lead not only to a rightwards shift but also to an increased dilation of the distribution of firm log productivities in larger cities.

To implement this model on exhaustive French establishment-level data, Combes *et al.* (2009*a*) develop a new quantile approach that allows estimating a relative change in left truncation, shift, and dilation between two distributions. Their main finding is that productivity differences across urban areas in France are mostly explained by agglomeration. The distribution of firms' log productivities in large French cities (population above 200,000) is remarkably well described by taking the distribution of firms' productivities in small French cities, dilating it, and shifting it to the right. This holds for the productivity distributions of firms across all sectors as well as most two-digit sectors when considered individually. For all manufacturing and business service sectors combined, relative to the rest of the country, the distribution of firms log productivities in large cities is shifted to the right by 0.09 and dilated by a factor of 1.22. Firms in large cities are thus on average about 9 percent more productive than in small cities. Because of dilation, this productivity advantage is only of 5 percent for firms at the bottom quartile and 14 percent for firms at the top quartile. On the other hand, they find no differences between small and large cities in terms of selection-driven left truncation of the log productivity distribution.

3. The causes of agglomeration economies

While there are solidly established conclusions regarding the existence of agglomeration economies, and a reasonably narrow range for their estimated magnitude, identifying the causes of agglomeration economies is proving more difficult.

There is a large theoretical literature that develops three broad classes of mechanisms to explain the existence of urban agglomeration economies (this classification follows Duranton and Puga, 2004, who cover the theoretical literature in detail). First, a larger market allows for a more efficient sharing of local infrastructure and facilities, a variety of intermediate input suppliers, or a pool of workers with similar skills. Second, a larger market also allows for a better matching between employers and employees, buyers and suppliers, or business partners. This better matching can take the form of improved chances of finding a suitable match, a higher quality of matches, or a combination of both. Finally, a larger market can also facilitate learning, for instance by promoting the development and widespread adoption of new technologies and business practices.

Going from modelling these mechanisms to identifying them empirically is not straightforward because they all share the prediction that productivity increases with the scale of an activity at a location. This 'Marshallian equivalence' (Duranton and

Puga, 2004) makes it very difficult to distinguish the main causes of the productivity advantages of cities. There are some clues in the aggregate estimates. For instance, the steep spatial decay of agglomeration economies (Rosenthal and Strange, 2003, Henderson, 2003, Desmet and Fafchamps, 2005) points to the importance of local interactions, and the rising wage profile following relocations to big cities (Glaeser and Maré, 2001) suggests an important dynamic component. However, to identify specific mechanisms, we must understand their microeconomic foundations and look for features that distinguish each particular mechanism. We now turn to reviewing the different mechanisms, providing a brief description of each theoretical argument and a review of the available empirical evidence supporting it.

Sharing facilities

Perhaps the simplest way to model localized increasing returns is to point to indivisibilities in the provision of certain goods or facilities, such as local infrastructure. Once the large fixed cost associated with a shared facility has been incurred, the larger the population that shares the facility the lower the cost per user. At the same time, congestion of the facility and crowding of the land surrounding it will place limits to growth of the user base (see Buchanan, 1965, for a pioneering discussion of the problems associated with the provision of shared facilities, and Scotchmer, 2002, for a detailed review of the literature).

While simple from a modelling perspective, this mechanism can be of practical importance in certain contexts. For instance, in their study of urban sprawl using remote-sending data, Burchfield, Overman, Puga, and Turner (2006) find that residences are closer to each other in cities where water provision relies on shared public facilities whereas urban development is more scattered in areas cities aquifers make individual household wells viable.

Sharing suppliers

A more sophisticated model of urban agglomeration economies through sharing is due to Abdel-Rahman and Fujita (1990). They capture the advantages for final producers of being able to share a larger common base of suppliers in larger and more specialized cities. In their model, perfectly competitive final goods firms use sector-specific intermediate inputs. These inputs are produced by a monopolistically competitive industry featuring Ethier's (1982) production-side version of the monopolistic competition framework of Dixit and Stiglitz (1977). Final goods are freely tradable while intermediates are only sold locally. Cities are subject to housing and commuting costs that increase with their population. In equilibrium, aggregate production at the city-sector level exhibits increasing

returns, despite constant returns to scale in perfectly-competitive final production. The reason is that an increase in final production, by virtue of expanding input sharing across a wider variety of suppliers, requires a less-than-proportional increase in primary factors.

Rosenthal and Strange (2001) study the empirical importance of sharing a common base of suppliers relative to other sources of agglomeration. They do so by regressing geographic concentration for each sector (measured by the index of Ellison and Glaeser, 1997) on proxies for different agglomeration motives computed also at the sector level (see also Audretsch and Feldman, 1996, for an earlier example of this approach using cross-sectional variation for identification). They measure the importance of input sharing for each sector using the purchase of manufactured and non-manufactured inputs as a share of value added. They find effects that are weak relative to other motives for agglomeration, with coefficients that are often statistically insignificant. Overman and Puga (2009), while studying the importance of the labour pooling mechanism discussed below, control for the importance of input sharing following the approach of Rosenthal and Strange (2001). When they simply use input purchases relative to value added as a proxy for the importance of input sharing, they find no support for this being an important determinant of spatial concentration. However, they then argue that, when an industry buys a lot from other industries, the importance of input sharing for its concentration will depend on whether those other industries are, in turn, spatially concentrated or dispersed. For instance, the meat processing industry is a large buyer of inputs from farms and from the plastic film industry. However, farms are very dispersed across the country and so is the plastic film industry, since it supplies many other sectors located in different places in addition to meat processing. Hence, the meat processing industry has no reason to concentrate spatially even if it makes large intermediate purchases: it can easily find its inputs everywhere. For a sector to cluster to share intermediate suppliers, it must be the case not only that the sector makes large purchases of intermediates but also that those intermediates are supplied by industries that are themselves very spatially concentrated. Following this line of reasoning, to better capture the importance of vertical linkages for a particular industry, Overman and Puga (2009) calculate sum of the Ellison and Glaeser (1997) index across all other industries weighted by the share each represents in the input shares of the industry being considered. This more sophisticated measure yields strong support for input sharing as a motive for agglomeration at the sector level.

In Abdel-Rahman and Fujita (1990), compared to a firm in a dense location, a relatively isolated firm produces using a narrower input mix purchased from outside suppliers. Alternatively, a relatively isolated firm might produce with a similar input mix but produce more of this inputs in-house. This possibility is discussed by Stigler (1951) and investigated empirically by Holmes (1999). He combines detailed plant level data for the United States with spatial data on input purchases. Using this, he first shows that the most

concentrated industries buy more inputs from outside suppliers in locations where they are clustered than in the rest of the country. Then Holmes (1999) regresses local purchased input intensity (differenced from the industry mean) on same-industry employment in the establishment's own county and in other counties within fifty miles (again differenced from the industry mean). The results indicate that the purchased-inputs intensity of a plant increases with the level of employment of neighboring plants in the same industry. This provides arguably the strongest empirical support to data for the importance of input sharing.

Sharing the gains from individual specialization

Adam Smith's (1776) famous pin factory example suggests that perhaps the presence of more workers in a given activity within a city increases output more than proportionately not just because extra workers can carry new tasks but because it allows existing workers to specialise on a narrower set of tasks. Several papers (Baumgardner, 1988, Becker and Murphy, 1992, Duranton, 1998, Becker and Henderson, 2000, Henderson and Becker, 2000) develop models exploring the idea that a larger market fosters specialization.

On the empirical side, the main evidence in favour of increasing specialization in larger markets comes from looking at professionals. For instance, Baumgardner (1988) shows that physicians perform a narrower range of activities in large markets. However, the work of Holmes (1999) discussed above can also be seen as supporting greater specialization in dense markets, although in this case it is greater specialization across firms as in-house input production is transferred to outside suppliers.

Sharing a labour pool

While there are various interpretations of labour market pooling as a source of agglomeration economies, some of them reviewed below, Marshall emphasized that "a localized industry gains a great advantage from the fact that it offers a constant market for skill" (Marshall, 1890, p. 271). Krugman (1991) formalizes this reasoning by considering a series of sectors where establishments experience idiosyncratic shocks. Individual profits are convex in the establishment-specific shock, since each establishment responds to the shock by adjusting its levels of both production and employment. However, changes in the establishment's employment affect local wages, and the effect is greater the more isolated the establishment is from other establishments in the same sector or using similar workers. If wages are higher when the establishment wants to expand production in response to a positive shock and lower when it wants to contract production in response to a negative shock, this limits the establishment's ability to adapt its employment level to good and bad times. Consequently, establishments that tend to experience substantial changes in

their employment relative to other establishments using workers with similar skills will find it advantageous to locate in places where there is a large number of workers with such skills. Here agglomeration arises because large concentrations of employment iron out idiosyncratic shocks and facilitate the transfer of labour from low to high productivity establishments.

Overman and Puga (2009) look at this mechanism empirically. They measure the likely importance of labour pooling by calculating the fluctuations in employment of individual establishments relative to their sector and averaging by sector. They finds that sectors whose establishments experience more idiosyncratic volatility are more spatially concentrated, even after controlling for a range of other industry characteristics. Overman and Puga (2009) study labour pooling looking at each sector separately. However, labour pooling could work across sectors if these use workers with similar skills. Dumais, Ellison, and Glaeser (1997) study the motives for agglomeration by considering which industries coagglomerate. They find that industries with similar labor mixes enjoy the largest benefit from proximity, which is indicative of labour pooling mattering also for bringing sectors together.

Better matching

Another advantage of of thick labour markets, in addition to the labour pooling argument discussed above, is that they lead to better matching between employers and employees — note that a similar argument can be made about matching between buyers and suppliers, or between business partners. Helsley and Strange (1990) formalize this argument by considering firms with heterogenous skill requirements represented by equally-spaced points on a circumference. Workers also have differentiated skills uniformly spread over the same circumference, and must incur in more costly retraining the greater the difference between their skill and the kill required by their employer. A larger city allows for the skill space to be more densely covered by firms, and thus reduces the average cost of mismatches.

Better matching can also take the form of improved chances of finding a suitable match. The frictional search literature often works with an aggregate matching function that yields the number of matches as a function of vacancies and job seekers. However, microeconomic foundations for such a matching function typically yield constant instead of increasing returns to scale. An important exception is the stock-flow matching framework of Coles (1994) and Coles and Smith (1998). Consider an unemployed worker who can simultaneously apply to all job vacancies that may suit her. In the first instance, the worker applies to the entire stock of available vacancies. If none of the applications are successful, the worker subsequently only applies to newly posted vacancies. Similarly, a

new vacancy receives applications from the entire stock of workers, but if none of these initial applications result in a suitable pairing, from then on the vacancy only receives applications from newly unemployed workers. The total number of matches is then the sum of matches between the flow of vacancies and the stock of unemployed workers and of matches between the flow of unemployed workers and the stock of vacancies. This yields naturally a matching function that exhibits increasing returns to scale. The intuition is simple: in a market with more job opportunities that can be explored simultaneously it is less likely that none of them work out.

Finally, there are interesting interactions between better matching in terms of improved chances of finding a suitable match and in terms of a higher quality of matches. A higher probability of matching in thicker markets allows firms and workers to be more "choosy", increasing the average quality of matches but somewhat reducing the higher probability of matching (Berliant, Reed, and Wang, 2000).

On the empirical side, Gan and Li (2004), after presenting a model where, as in Coles (1994) and Coles and Smith (1998), the probability of matching increases with the thickness of the market, test its predictions empirically. They compare fields of different size in the academic recruitment market for new PhDs in Economics, and find that a field of specialization with more job openings and more candidates offers a higher probability of matching. Another piece of evidence in favour of larger cities facilitating matching is the work of Costa and Kahn (2000). They show that couples in which both spouses have college degrees are increasingly likely to be located in the largest metropolitan areas, and not just because they meet there. One explanation is that college-educated couples are more likely to face a co-location problem and moving to big cities increases the chances that both find suitable matches. Gathering evidence on the other matching mechanism discussed, that thicker markets improve the quality of matches, is more complicated because of the intrinsic difficulties of measuring match quality. Hopefully, the increasing availability of matched employer-employee micro-data will encourage more work on agglomeration through matching.

Learning

Agglomeration mechanisms directly dealing with learning have received much less attention in the theoretical literature than the sharing and matching mechanisms discussed above. There are, however, some noteworthy exceptions. Glaeser (1999) develops a model in which young workers migrate to big cities because interactions with experienced workers helps them acquire valuable skills and experienced workers remain in cities to share the rents of this learning process. Besides this purposeful transmission of knowledge, the informal literature on learning in cities has also emphasized the casual and unintended

flows of information facilitated by big cities. However, despite being often motivated by examples that are specifically spatial in nature, the literature on social learning has so far not produced models capturing this microfoundation of urban agglomeration economies. Nevertheless, spatial interactions motivate the externality used in the richest models studying the spatial allocation of production and housing within a city (Fujita and Ogawa, 1982, Ota and Fujita, 1993, Lucas and Rossi-Hansberg, 2002).

In addition to facilitating the transmission of knowledge, cities are also seen as promoting the creation of new knowledge. The work of Jacobs (1969) is often associated with the idea that diversified urban environments facilitate search and experimentation in innovation. Duranton and Puga (2001) develop microeconomic foundations for such a role. In their model, a young firm needs a period of experimentation to realise its full potential — the entrepreneur may have a project, but may not know all the details of the product to be made, what components to use, where to source them, which workers to hire, and how to finance the venture. A diverse city provides many alternatives to try without having to relocate and this creates dynamic advantages to urban diversity. When combined with more standard static agglomeration economies and urban congestion costs, this justifies the coexistence of diversified and specialised cities and the agglomeration of firms at different stages of their life-cycle in cities of each type. Young firms locate in more diverse urban environments and, when their products mature, relocate to more specialized towns. Looking at establishment relocations across France, Duranton and Puga (2001) find evidence of the relocation pattern predicted by their model.

On the empirical side, there is strong support for the idea that cities facilitate innovation, the diffusion of knowledge, and the acquisition of skills. A priori, the diffusion of knowledge might seem hard to trace. However, Jaffe, Trajtenberg, and Henderson (1993) track knowledge flows by looking at patent citations. They show that inventors are much more likely to cite prior patents with inventors from the same city than a randomly drawn control sample of cited patents. Audretsch and Feldman (1996) show that innovative activity, as measured by significant new product introductions, tends to cluster geographically to a greater extent in industries where new economic knowledge plays a more important role. This greater spatial concentration of innovation holds even after controlling for the concentration of production.No (2003), using data on the adoption of advanced manufacturing technologies in Canada, finds that adoption is more likely in locations with more prior adopters, particularly if they use similar technologies but do not compete in the same detailed sector. Charlot and Duranton (2004) use survey data on communication between workers to show that workplace communication is more extensive in urban areas. Some results from studies quantifying agglomeration economies already discussed above Rosenthal and Strange (2003), Henderson (2003), Glaeser and Maré (2001) might also be interpreted as supporting the relevance of learning mechanisms.

4. Conclusions

Despite the broad agreement on the magnitude of agglomeration economies at the urban level, the literature has been far less successful at distinguishing between the possible sources. This requires models that work out the microfoundations to help identify distinguishing features and empirical work that carefully exploits these for identification. On the theoretical side, we have good models of agglomeration through sharing and matching, but not a deep enough theoretical understanding of learning in cities. On the empirical side, evidence of matching as a source of agglomeration is perhaps most needed. However, despite several notable existing contributions, there is room for much more work able to credibly claim identification of a particular driver of agglomeration.

References

Abdel-Rahman, Hesham M. and Masahisa Fujita. 1990. Product variety, Marshallian externalities, and city sizes. *Journal of Regional Science* 30(2):165–183.

Audretsch, David B. and Maryann P. Feldman. 1996. R&D spillovers and the geography of innovation and production. *American Economic Review* 86(3):630–640.

Baumgardner, James R. 1988. The division of labor, local markets, and worker organization. *Journal of Political Economy* 96(3):509–527.

Becker, Gary S. and Kevin M. Murphy. 1992. The division of labor, coordination costs, and knowledge. *Quarterly Journal of Economics* 107(4):1137–1160.

Becker, Randy and J. Vernon Henderson. 2000. Intra-industry specialization and urban development. In Jean-Marie Huriot and Jacques-François Thisse (eds.) *Economics of Cities: Theoretical Perspectives*. Cambridge: Cambridge University Press, 138–166.

Berliant, Marcus, Robert R. Reed, III, and Ping Wang. 2000. Knowledge exchange, matching, and agglomeration. Discussion Paper 135, Federal Reserve Bank of Minneapolis.

Buchanan, James M. 1965. An economic theory of clubs. *Economica* 32(125):1–14.

Burchfield, Marcy, Henry G. Overman, Diego Puga, and Matthew A. Turner. 2006. Causes of sprawl: A portrait from space. *Quarterly Journal of Economics* 121(2):587–633.

Charlot, Sylvie and Gilles Duranton. 2004. Communication externalities in cities. *Journal of Urban Economics* 56(3):581–613.

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

Coles, Melvyn G. 1994. Understanding the matching function: The role of newspapers and job agencies. Discussion Paper 939, Centre for Economic Policy Research.

Coles, Melvyn G. and Eric Smith. 1998. Marketplaces and matching. *International Economic Review* 39(1):239–255.

Combes, Pierre-Philippe, Gilles Duranton, and Laurent Gobillon. 2008. Spatial wage disparities: Sorting matters! *Journal of Urban Economics* 63(2):723–742.

Combes, Pierre-Philippe, Gilles Duranton, Laurent Gobillon, Diego Puga, and Sébastien Roux. 2009*a*. The productivity advantages of large cities: Distinguishing agglomeration from firm selection. Discussion Paper 7191, Centre for Economic Policy Research.

Combes, Pierre-Philippe, Gilles Duranton, Laurent Gobillon, and Sébastien Roux. 2009*b*. Estimating agglomeration effects with history, geology, and worker fixed-effects. In Edward L. Glaeser (ed.) *The Economics of Agglomeration*. Cambridge, MA: National Bureau of Economic Research, forthcoming.

Costa, Dora L. and Matthew E. Kahn. 2000. Power couples: Changes in the locational choice of the college educated, 1940-1990. *Quarterly Journal of Economics* 115(4):1287–1315.

Dekle, Robert and Jonathan Eaton. 1999. Agglomeration and land rents: Evidence from the prefectures. *Journal of Urban Economics* 46(2):200–214.

Desmet, Klaus and Marcel Fafchamps. 2005. Changes in the spatial concentration of employment across US counties: A sectoral analysis 1972–2000. *Journal of Economic Geography* 5(3):261–284.

Dixit, Avinash K. and Joseph E. Stiglitz. 1977. Monopolistic competition and optimum product diversity. *American Economic Review* 67(3):297–308.

Dumais, Guy, Glenn Ellison, and Edward L. Glaeser. 1997. Geographic concentration as a dynamic process. Working Paper 6270, National Bureau of Economic Research.

Duranton, Gilles. 1998. Labor specialization, transport costs, and city size. *Journal of Regional Science* 38(4):553–573.

Duranton, Gilles. 2008. California dreamin': The feeble case for cluster policies. Processed, University of Toronto.

Duranton, Gilles and Henry G. Overman. 2005. Testing for localization using micro-geographic data. *Review of Economic Studies* 72(4):1077–1106.

Duranton, Gilles and Diego Puga. 2001. Nursery cities: Urban diversity, process innovation, and the life cycle of products. *American Economic Review* 91(5):1454–1477.

Duranton, Gilles and Diego Puga. 2004. Micro-foundations of urban agglomeration economies. In Vernon Henderson and Jacques-François Thisse (eds.) *Handbook of Regional and Urban Economics*, volume 4. Amsterdam: North-Holland, 2063–2117.

Duranton, Gilles and Diego Puga. 2005. From sectoral to functional urban specialisation. *Journal of Urban Economics* 57(2):343–370.

Ellison, Glenn and Edward L. Glaeser. 1997. Geographic concentration in US manufacturing industries: A dartboard approach. *Journal of Political Economy* 105(5):889–927.

Ellison, Glenn and Edward L. Glaeser. 1999. The geographic concentration of industry: Does natural advantage explain agglomeration? *American Economic Review Papers and Proceedings* 89(2):311–316.

Ethier, Wilfred J. 1982. National and international returns to scale in the modern theory of international trade. *American Economic Review* 72(3):389–405.

Feldman, Maryann P. and David B. Audretsch. 1999. Innovation in cities: Science-based diversity, specialization and localized competition. *European Economic Review* 43(2):409–429.

Fujita, Masahisa and Hideaki Ogawa. 1982. Multiple equilibria and structural transition of non-monocentric urban configurations. *Regional Science and Urban Economics* 12(2):161–196.

Gan, Li and Qi Li. 2004. Efficiency of thin and thick markets. Working Paper 11813, National Bureau of Economic Research.

Glaeser, Edward L. 1999. Learning in cities. *Journal of Urban Economics* 46(2):254–277.

Glaeser, Edward L. and David C. Maré. 2001. Cities and skills. *Journal of Labor Economics* 19(2):316–342.

Harrison, Bennett, Maryellen R. Kelley, and Jon Gant. 1996. Specialization versus diversity in local economies: The implications for innovative private-sector behavior. *Cityscape* 2(2):61–93.

Helsley, Robert W. and William C. Strange. 1990. Matching and agglomeration economies in a system of cities. *Regional Science and Urban Economics* 20(2):189–212.

Henderson, J. Vernon. 1974. The sizes and types of cities. *American Economic Review* 64(4):640–656.

Henderson, J. Vernon. 1997. Externalities and industrial development. *Journal of Urban Economics* 42(3):449–470.

Henderson, J. Vernon. 2003. Marshall's scale economies. *Journal of Urban Economics* 53(1):1–28.

Henderson, J. Vernon and Randy Becker. 2000. Political economy of city sizes and formation. *Journal of Urban Economics* 48(3):453–484.

Holmes, Thomas J. 1999. Localization of industry and vertical disintegration. *Review of Economics and Statistics* 81(2):314–325.

Jacobs, Jane. 1969. *The Economy of Cities*. New York: Random House.

Jaffe, Adam B., Manuel Trajtenberg, and Rebecca Henderson. 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics* 108(3):577–598.

Kelley, Maryellen R. and Susan Helper. 1999. Firm size and capabilities, regional agglomeration, and the adoption of new technology. *Economics of Innovation and New Technology* 8(1–2):79–103.

Krugman, Paul R. 1991. *Geography and Trade*. Cambridge, MA: MIT Press.

Lucas, Robert E., Jr. and Esteban Rossi-Hansberg. 2002. On the internal structure of cities. *Econometrica* 70(4):1445–1476.

Marshall, Alfred. 1890. *Principles of Economics*. London: Macmillan.

Melitz, Marc and Gianmarco I. P. Ottaviano. 2008. Market size, trade and productivity. *Review of Economic Studies* 75(1):295–316.

Melitz, Marc J. 2003. The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica* 71(6):1695–1725.

Moomaw, Ronald L. 1981. Productivity and city size? a critique of the evidence. *Quarterly Journal of Economics* 96(4):675–688.

No, Joung Yeo Angela. 2003. Agglomeration effects in the diffusion of advanced manufacturing technologies. Processed, University of Toronto.

Ota, Mitsuru and Masahisa Fujita. 1993. Communication technologies and spatial organization of multi-unit firms in metropolitan areas. *Regional Science and Urban Economics* 23(6):695–729.

Ottaviano, Gianmarco I. P. and Jacques-François Thisse. 2004. Agglomeration and economic geography. In Vernon Henderson and Jacques-François Thisse (eds.) *Handbook of Regional and Urban Economics*, volume 4. Amsterdam: North-Holland, 2563–2608.

Overman, Henry G. and Diego Puga. 2008. Labour pooling as a source of agglomeration: An empirical investigation. Processed, IMDEA.

Overman, Henry G. and Diego Puga. 2009. Labour pooling as a source of agglomeration: An empirical investigation. In Edward L. Glaeser (ed.) *The Economics of Agglomeration*. Cambridge, MA: National Bureau of Economic Research, forthcoming.

Rauch, James E. 1993. Productivity gains from geographic concentration of human-capital - evidence from the cities. *Journal of Urban Economics* 34(3):380–400.

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

Rosenthal, Stuart S. and William Strange. 2004. Evidence on the nature and sources of agglomeration economies. In Vernon Henderson and Jacques-François Thisse (eds.) *Handbook of Regional and Urban Economics*, volume 4. Amsterdam: North-Holland, 2119–2171.

Rosenthal, Stuart S. and William C. Strange. 2001. The determinants of agglomeration. *Journal of Urban Economics* 50(2):191–229.

Rosenthal, Stuart S. and William C. Strange. 2003. Geography, industrial organization, and agglomeration. *Review of Economics and Statistics* 85(2):377–393.

Scotchmer, Suzanne. 2002. Local public goods and clubs. In Alan J. Auerbach and Martin Feldstein (eds.) *Handbook of Public Economics*, volume 4. Amsterdam: North-Holland, 1997–2042.

Smith, Adam. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. London: Printed for W. Strahan, and T. Cadell.

Starrett, David A. 1978. Market allocations of location choice in a model with free mobility. *Journal of Economic Theory* 17(1):21–37.

Stigler, George J. 1951. The division of labor is limited by the extent of the market. *Journal of Political Economy* 59(3):185–193.

Sveikauskas, Leo. 1975. Productivity of cities. *Quarterly Journal of Economics* 89(3):393–413.

Wheaton, William C. and Mark J. Lewis. 2002. Urban wages and labor market agglomeration. *Journal of Urban Economics* 51(3):542–562.