CHAPTER 22

The Geography of Development Within Countries

Klaus Desmet*, **J. Vernon Henderson**†
*Department of Economics, Southern Methodist University, Dallas, TX, USA
†Department of Geography, London School of Economics, London, UK

Contents

Abstract

This chapter describes how the spatial distribution of economic activity changes as economies develop and grow. We start with the relation between development and rural–urban migration. Moving beyond the coarse rural–urban distinction, we then focus on the continuum of locations in an economy and describe how the patterns of convergence and divergence change with development. As we discuss,

Handbook of Regional and Urban Economics, Volume 5B
ISSN 1574-0080, http://dx.doi.org/10.1016/B978-0-444-59531-7.00022-3

these spatial dynamics often mask important differences across sectors. We then turn our attention to the right tail of the distribution, the urban sector. We analyze how the urban hierarchy has changed over time in developed countries and more recently in developing countries. The chapter reviews both the empirical evidence and the theoretical models that can account for what we observe in the data. When discussing the stylized facts on geography and development, we draw on empirical evidence from both the historical evolution of today's developed economies and comparisons between today's developed and developing economies.

Keywords

Geography, Development, Space, Growth, City-size distribution, Spatial distribution of economic activity, Developed countries, Developing countries, Urban hierarchy, Industrialization and urbanization

JEL Classification Codes

R1, R11, R12, O18

22.1. INTRODUCTION

As economies grow and develop, the spatial distribution of the population, employment, and production changes. Probably the most prominent feature of this spatial transformation is increased urbanization. Between 1950 and 2009, the world's urban population more than quadrupled from 732 million to 3.4 billion, as the world moved from being under 30% urbanized to over 50% urbanized. Understanding the patterns of this rapid transformation is of paramount importance to policy makers. More than 80% of governments are concerned about the geographic distribution of people, and nearly 70% of them have implemented policies to reduce internal migration (United Nations, 2010). The goal of this chapter is to review what we know about the spatial distribution of economic activity and development. An important point we will make is that this spatial transformation can be viewed at different spatial scales and through different lenses. Which one is more useful will largely depend on the issue of interest.

One traditional divide is to contrast rural and urban areas, but that fails to capture the full richness of a country's spatial transformation. Rather than splitting up locations into two types (urban or rural), one often finds it useful to think of locations as a continuum, going from more rural (smaller and/or less dense) to more urban (larger and/or denser). The distribution of the population and economic activity along that continuum changes radically with development, and these changes mark how we view the overall geography of a country. What happens with aggregate employment and production often masks interesting differences across sectors. Manufacturing and services have exhibited very different spatial growth patterns over time.

Once a country becomes more urbanized, these changes and the spatial distribution are often viewed through a narrower lens that focuses on the urban sector. Within the urban sector there is enormous heterogeneity across the hierarchy of cities, and the

transformation of activities differs across that hierarchy. Finally, we note that while much of what we see is driven by market forces, the role of government in economies has grown. As a result, in today's developing countries, economic policies can have a strong effect on both the location and the concentration of economic activities.

This chapter reviews the models and evidence that characterize these processes. Section 22.2 starts by looking at the urban–rural divide and then focuses on the continuum. It analyzes population and income convergence versus divergence and the reshaping of the location patterns of people and economic activity, especially in today's richer countries as they developed through the nineteenth century into the twentieth century. Another issue of interest that we discuss is the link between an economy's overall spatial structure and its aggregate growth. Section 22.3 also focuses on the continuum, but takes a sectoral approach by looking at the structural transformation of economic activities as a country develops and matures. The distribution of economic activity differs across sectors, and these differences change over time as countries develop. Section 22.4 looks at the urban sector, with particular attention on the urban hierarchy. It explores aspects of the transformation of the urban sector over the last 100 years in more developed countries and the more recent, rapid changes in developing countries. Section 22.4 also discusses the key issue of how government policies in today's developing countries affect the transformation and the concentration of economic activities.

When discussing how the spatial distribution of economic activity changes with development, we draw on evidence both from comparing today's developed and developing economies and from analyzing the long-run evolution of today's developed countries. Although using historical evidence from today's developed countries to explain the spatial patterns of present-day developing countries is useful, this should be done with care. For example, because of trade and comparative advantage, the role of the structural transformation from agriculture to manufacturing in explaining urbanization in today's developing countries may be different from its role in nineteenth century Europe.

22.2. DEVELOPMENT AND THE AGGREGATE SPATIAL DISTRIBUTION

We start by discussing models of rural–urban migration. This coarse-grained look at the shift from the rural to the urban sector that occurs with development is the typical approach used by development economists. We cover recent developments to this paradigm that originally dates back to Lewis (1954). Then we turn to the perspective of a continuum which covers the national geography at a finer spatial scale.

22.2.1 Development: Urban versus rural

The link between urbanization and development has been emphasized both in the context of the transition from Malthusian to modern growth and in the work on rural–urban migration in developing countries. Much of the literature has emphasized the link between

development, industrialization, and urbanization. However, in light of the recent experience of Africa and the Middle East, urbanization and industrialization may not always go hand in hand, especially for countries whose incomes are heavily resource dependent.

22.2.1.1 Industrialization and urbanization

While the literature on the transition to modern growth is extensive, most of the competing models aim to capture the gradual transition from an agricultural-based rural economy to an industrial-based urban economy. In a context where incomes are growing, most articles generate this result by assuming an income elasticity of less than 1 for food items, leading to an increasing share of expenditure on urban goods. A simple way of modeling this is to introduce a subsistence constraint into standard Cobb–Douglas preferences, which yields a Stone–Geary utility function:

$$U(c_a, c_m) = (c_a - \bar{c}_a)^{1-\alpha} c_m^{\alpha}, \tag{22.1}$$

where c_a is agricultural consumption, c_m is manufacturing consumption, and $\bar{c}_a$ is the agricultural subsistence constraint. These preferences have been used in many models of industrialization (see, e.g., Caselli and Coleman, 2001; Desmet and Parente, 2012). Such a setup creates a direct link between income *per capita*, industrialization, and urbanization, in as far as the industrial sector is less land intensive and more urbanized than the agricultural sector.

Another way of generating industrialization is by having an elasticity of substitution between agriculture and industry of less than 1:

$$U(c_a, c_m) = \left(\alpha_a c_a^{\frac{\sigma-1}{\sigma}} + \alpha_m c_m^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}}, \tag{22.2}$$

where $\sigma < 1$. This approach to the structural transformation, taken by Ngai and Pissarides (2007) and Desmet and Rossi-Hansberg (2014a), implies that employment will shift out of agriculture into industry if agricultural productivity growth is higher.

Independently of whether we assume (22.1) or (22.2), an "agricultural revolution" must have preceded the industrial revolution. This idea is emphasized in the work by Nurkse (1953), Rostow (1960), Schultz (1968), and Diamond (1997) who argue that high agricultural productivity was a precondition for industrial takeoff. Consistent with this, Allen (2004) finds that output per worker in English agriculture doubled between 1600 and 1750, ahead of the industrial revolution. Greater agricultural efficiency allowed the economy to overcome the "food problem" and created a surplus of workers who could then engage in other activities, such as manufacturing. In modern developing countries, such as India, the Green Revolution has played a similar role. Work by Gollin et al. (2007) shows in a quantitative model that differences in agricultural total factor productivity (TFP) are key in explaining the differential timing of takeoff across countries. Note, however, that this positive link between agricultural productivity

and industrialization may be reversed when we allow for trade. As shown by Matsuyama (1992), in an open economy higher agricultural productivity may lock in a comparative advantage in that sector, thus delaying industrialization.

In most models of the industrial revolution and the transition to modern growth, the link to space and urban–rural migration is indirect. It is only in as far as we equate agriculture with rural and industry with urban that we get clear implications for the changing spatial distribution of economic activity. In some models, the transition from agriculture to manufacturing is implicit (Galor and Weil, 2000; Lucas, 2004), whereas in others it is explicit (Hansen and Prescott, 2002; Tamura, 2002; Doepke, 2004; Galor et al., 2009; Desmet and Parente, 2012). But in most of this literature, the focus is not on rural–urban migration *per se*. There are some exceptions though, such as Lucas (2004) and Henderson and Wang (2005), which we discuss in the next paragraphs.

22.2.1.2 Rural–urban migration and the transition to modern growth

Lucas (2004) proposes a model of infinitely lived dynasties to analyze the link between the structural transformation, urban–rural migration, and the shift from a traditional technology (with no growth) to a modern technology (with unbounded growth). In the rural sector, human capital is useless, whereas in the urban sector it increases productivity. Human capital accumulation depends on the time invested and on the human capital frontier.

The Lucas (2004) model captures some of the stylized facts of rural–urban migration. First, as the economy develops, people move gradually from the rural sector to the urban sector. Over time, as the human capital frontier moves out, it becomes less costly to accumulate human capital, making cities more attractive. The human capital externality—the fact that cities are good places to accumulate human capital—is key for this result. Second, recent arrivals do not work and instead spend their time improving their human capital. This is akin to the Harris and Todaro (1970) model where many of the recent arrivals are unemployed. The difference here is that unemployment is voluntary. Third, because the representative agent is a family, when migrants first arrive in the city, they are subsidized by the ones that stayed behind, and they later reimburse the rural part of the family through remittances.

In contrast to Lucas (2004), where there is only one consumption good, in Henderson and Wang (2005), the urban and rural sectors produce different goods. There is accumulation of human capital fueling growth in both sectors, although human capital externalities in the urban sector are posited to be greater. Demand for food is completely income inelastic. As human capital accumulates, and people become richer, there is a shift of the population and production out of the farm/rural sector into the urban/industrial sector and development. This leads to increased urbanization, as existing cities grow and new cities arise.

The fact that urbanization and industrialization often tend to go hand in hand (in many countries and models) raises the question of which one drives which. There are

good reasons to believe that causality runs both ways. In fact, the model of Lucas (2004) and many other models of the structural transformation feature such circular causality. Empirically, while urbanization certainly helps industrialization, it is not necessarily the case that industrialization started in the large urban centers. In the historical case of the US Northeast, for example, Williamson (1965) shows that the degree of urbanization exhibited convergence during the period of industrialization, suggesting that industrialization was particularly strong in the least urbanized places. The same was not true though for the United States as a whole, where there was divergence in urbanization during the nineteenth century. As in the US Northeast, also in England, many of the "hot spots" of the industrial revolution were initially relatively small towns that grew into large cities. Liverpool's population, for example, multiplied by more than 60 times over the course of a century and a half, from 6000 in 1700 to 376,000 in 1850 (Bairoch et al., 1988). For developing counties, the issue is in part related to the adaptation of technologies: their agricultural and industrial revolutions often involve the simultaneous importing of world technologies in both sectors. In many developing countries, the rural sector has had fairly high levels of nonfarm activity and an important presence of traditional industries for decades.

22.2.1.3 Rural–urban migration without industrialization

Although urbanization and industrialization often go together, this is not always the case. Gyourko et al. (2013) document that urbanization in Africa (and the Middle East) has proceeded at about the same pace as in Asia, in spite of Africa having a much lower level of industrialization. Asia follows the standard development process: higher income, industrialization, and urbanization all proceed simultaneously. This gives rise to what they refer to as "production cities." In contrast, in Africa, surplus income from the exports of natural resources leads to greater demand for nontradable goods which are produced in urban centers. This gives rise to what they refer to as "consumption cities." This disconnect between industrialization and urbanization has also been noted by Fay and Opal (2000) and Henderson et al. (2013).

We now describe the Gollin et al. model in some more detail. They propose a small open economy model with four sectors (food, tradable goods, nontradable goods, and natural resources). By assumption, food production is a rural activity, whereas tradable and nontradable goods are produced in cities. Natural resources are sold internationally and have no domestic market. In this model, a positive shock to natural resources (an increased stock or an increased price) will lead to urbanization without industrialization. Through a standard Rybczynski effect, more labor will be employed in natural resources and less in food and tradable goods. In contrast, because of the positive income effect, the demand for nontradable goods will increase. As a result, the food and tradable good sector will shrink, and the nontradable good sector will expand. If the expansion of the nontradable good sector outweighs the contraction of the tradable good sector, urbanization will increase in the absence of industrialization.

In addition to focusing on the role of natural resources, the work by Gollin et al. also illustrates that comparative advantage and trade can change a country's standard development path. Because of trade, not all countries may need to go through a structural transformation from agriculture to manufacturing as they develop. In that sense, international trade may make the relation between development and the spatial concentration of economic activity more heterogeneous. A broader implication is that using the historical experience of developed countries to "predict" what will happen in developing countries, though useful, should be done with caution. Comparative advantage in early developers may very well be different from comparative advantage in late developers, thus changing the relation between development, industrialization, and urbanization. This connects back to the work of Matsuyama (1992) which we discussed earlier.

That urbanization has proceeded without industrialization does not necessarily imply that urbanization has proceeded without growth. In fact, in Gollin et al., the growth of the urban nontradable good sector is a direct consequence of the positive income shock coming from natural resources. Not everyone agrees though. Fay and Opal (2000), for example, claim that Africa has urbanized in the absence of economic growth. However, given the severe measurement problems that plague income *per capita* in Africa, Henderson et al. (2013) are skeptical of that claim. In fact, when using human capital accumulation, as measured by average number of years in school, they find that the relation with urbanization is not different in Africa compared with the rest of the world.

22.2.2 Development: Continuum of locations

Although increased urbanization is a basic fact of development, limiting the focus to the urban–rural distinction may ignore some of the richer growth dynamics. After all, there are denser and less dense rural areas, and there are bigger and smaller cities. In this section, we take a comprehensive approach. Rather than focusing on cities of different sizes, we focus on all locations. This is important for at least four reasons. First, cities are not islands, and they form part of the overall spatial distribution of the population and economic activity. Second, when going back in time, or when focusing on developing countries, we find the percentage of the population living in rural areas is not trivial. Third, some of the stylized facts that hold for cities may no longer hold when all locations are included. Fourth, when we limit our focus to cities, we introduce a selection bias that we need to be aware of, since by definition cities are locations that benefited from high growth at some point in the past.

In what follows we start by analyzing some of the stylized facts related to growth across locations of different sizes and densities, and then briefly discuss some models that are able to capture the observed dynamics. We also review recent work that focuses on the link between the economy's overall spatial structure and its aggregate growth. It emphasizes the need to develop models that reconcile the main macro growth facts with the observed spatial heterogeneity of economic activity.

22.2.2.1 Facts

22.2.2.1.1 Population growth dynamics and Gibrat's law

Several articles have looked at whether growth is orthogonal to size when considering the entire distribution of locations. Holmes and Lee (2010) divide the United States into a grid of 6 mile by 6 mile squares (93.2 km^2) and find an inverted-U relation between size and the growth of the population between 1990 and 2000. Squares with a population of less than 1000 have an average growth rate of 0.054; growth peaks at 0.149 for squares with a population between 10,000 and 50,000; and growth then declines to 0.06 for squares with more than 500,000 people. This translates into growth peaking in locations with a population density between 100 and 500 per square kilometer. This suggests that Gibrat's law can be rejected when looking at the entire distribution of locations. Using the same time period but focusing on census places, Eeckhout (2004) finds that growth satisfies Gibrat's law. Part of the difference from the findings of Holmes and Lee (2010) is that the census places in Eeckhout (2004) cover only 74% of the US population, leaving out some of the areas with very low population densities.

If there is some doubt about the orthogonality of growth to size in recent times, there is even more doubt when going back in time. Gibrat's law is, if anything, a fairly recent phenomenon. Michaels et al. (2012) use data on US subcounty divisions (in particular, minor civil divisions) to analyze the relation between population density and population growth over the period 1880–2000. As shown in Figure 22.1, the data show a U-shaped relation which becomes flat for high-density locations. Low-density locations, with

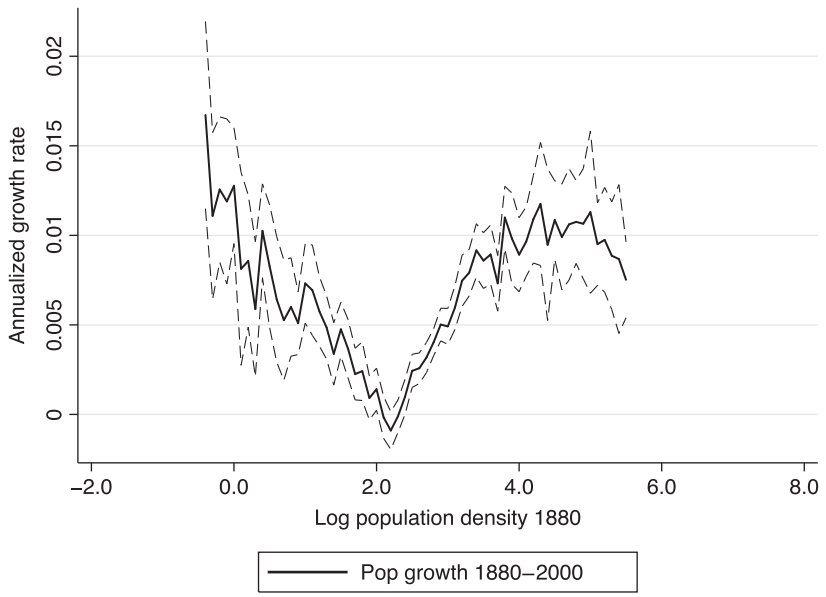


Figure 22.1 Population growth from 1880 to 2000 for US minor civil divisions. *Source: Michaels et al. (2012).*

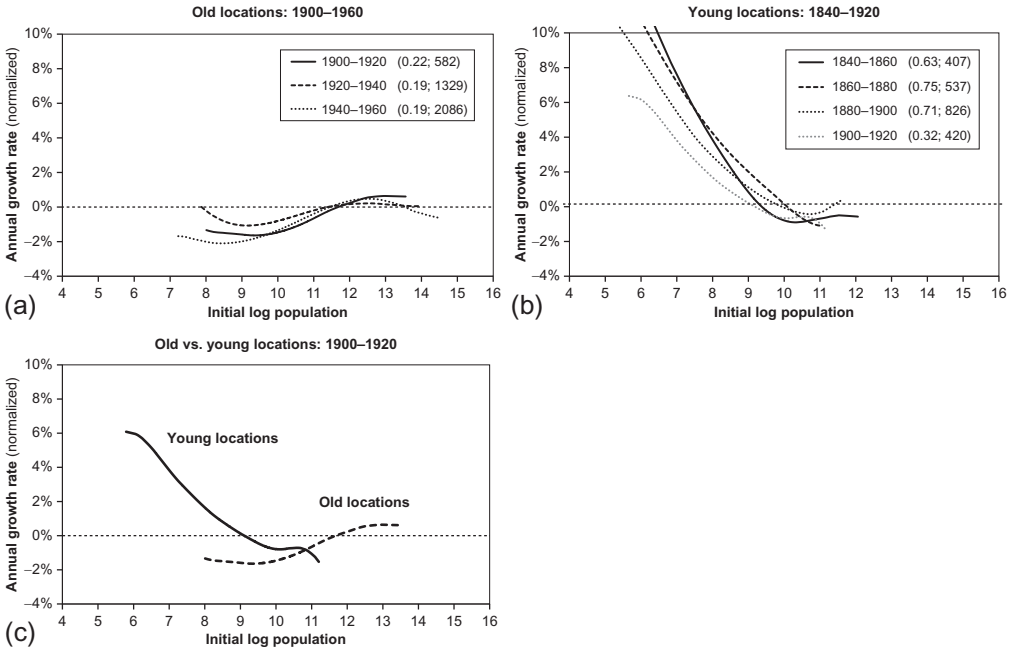
populations of less than 7 per square kilometer in 1880 (less than log 2), exhibit a negative relation between the initial density in 1880 and growth over the period 1880–2000. For medium-density locations, with populations between 7 and 55 per square kilometer (between log 2 and log 4), this relation is positive. It is only for the highest-density locations, with populations above 55 per square kilometer in 1880 (above log 4), that the relation becomes orthogonal. That is, if Gibrat's law holds, it holds only for high-density cities and not for rural areas. This finding illustrates that exclusively focusing on cities is misleading in terms of changes to the whole geography, especially taking into account that about half of the US population in 1880 lived in the intermediate range of locations that experienced divergent growth over the following century. Michaels et al. relate this finding to the structural transformation: divergent growth is most prominent in locations that are transitioning from being agricultural based to being manufacturing based, which reshapes the entire national economic geography.

In a related study, Desmet and Rappaport (2013) use data on US counties from the decennial censuses, starting in 1800, and analyze the relation between size and growth over ten 20-year periods until 2000. They strongly reject orthogonal growth until very recently. Until the 1940s, smaller counties exhibited dispersion (convergence), medium-sized counties exhibited concentration (divergence), and large counties exhibited dispersion (convergence). In more recent time periods, the dispersion at the lower end has disappeared, although the medium-sized counties continue to show some tendency toward further concentration. They show that the nonorthogonality at the lower tail of the distribution is intimately related to the age of a location. Figure 22.2 a and b shows how newly settled (young) locations tend to grow faster than long-settled (old) locations. Young locations exhibit strong convergent growth, whereas old locations exhibit slight divergent growth, except for the largest ones. Although most young locations are also small, not all old locations are large. As can be seen in Figure 22.2 c, the distinction between young and old is therefore not just picking up a size effect.

When the westward settlement of the United States came to an end, convergent growth among smaller locations weakened and disappeared. The importance of settlement for understanding US growth dynamics was emphasized in earlier work by Beeson and DeJong (2002). As for the divergent growth of medium-sized locations, Desmet and Rappaport (2013) relate it to evidence regarding either the declining share of land in production (as in Michaels et al., 2012) or increasing agglomeration economies owing to the introduction of new technologies (as in Desmet and Rossi-Hansberg, 2009).

22.2.2.1.2 Population distribution

Another important finding is that for the last 200 years, the spatial distribution of the population (and population density) has been close to lognormal. This is true, for example, when focusing on the distribution of population levels across US counties as early as 1790, as can be seen in Figure 22.3. The distribution of population densities across minor



A location is "young" if no more than 40 years have passed since the state or territory in which it is located first had two or more counties with positive population. A location is old if more than 60 years have passed since it experienced its final significant geographic change.

Figure 22.2 Population growth from 1800 to 2000 for US counties. *Source: Desmet and Rappaport (2013).*

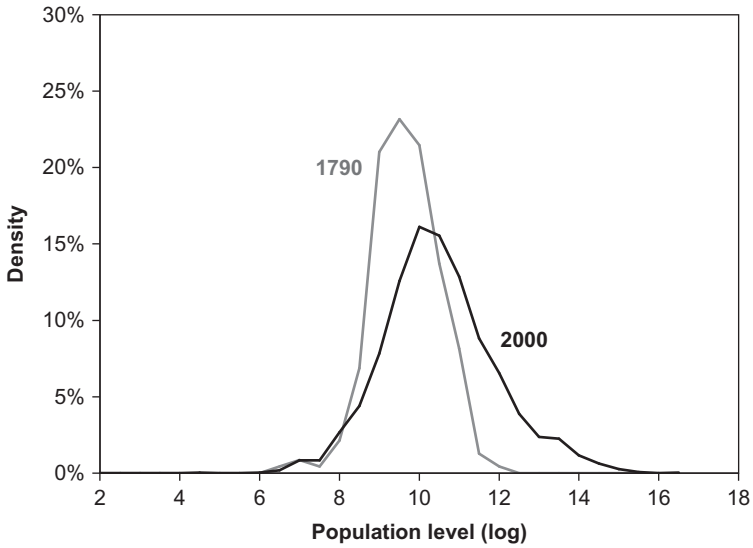


Figure 22.3 Logarithmic population distribution from 1790 to 2000 for US counties. *Source: Desmet and Rappaport (2013).*

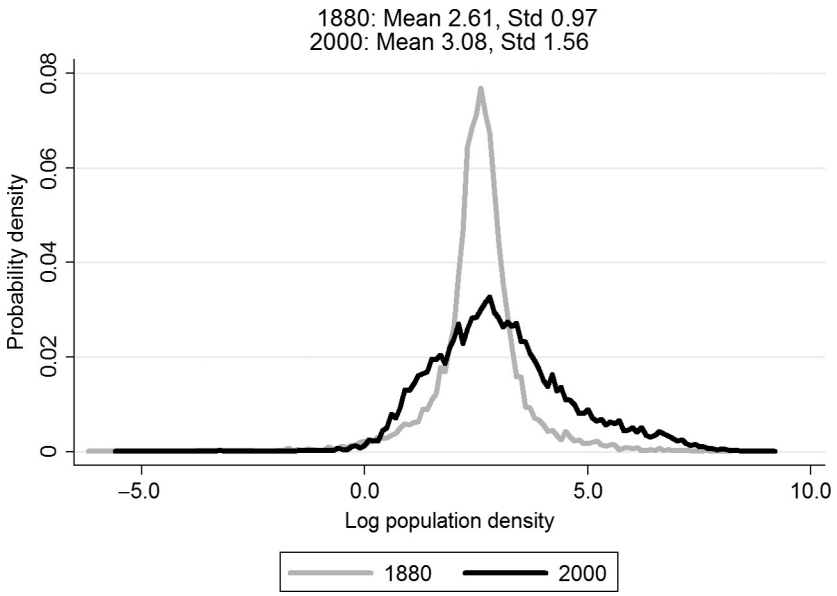


Figure 22.4 Logarithmic population densities from 1880 to 2000 for US minor civil divisions. *Source:* *Michaels et al. (2012)*.

civil divisions in 1880 in Figure 22.4 shows a similar picture. Although the population distribution has essentially remained lognormal (in both levels and densities), the dispersion has increased, mainly because the larger locations today are much larger than they were in the nineteenth century, whereas the smaller locations are not.

22.2.2.1.3 Income growth dynamics

In addition to the focus on population dynamics, some articles have emphasized income *per capita* dynamics. While Michaels et al. (2012) show how the structural transformation can help us understand differential population growth across locations, Caselli and Coleman (2001) argue that the structural transformation can account for the observed income *per capita* convergence across US regions over the last century. Between 1880 and 1980, the South/North relative wage increased from 0.4 to 0.9. In 1880, there was a strong negative correlation between income per worker and the share of employment in agriculture across US states. Over the following century, the states which had most agriculture initially were also the ones where agriculture declined the most and where income *per capita* growth was strongest. These findings taken together, Caselli and Coleman (2001) show that this can explain regional convergence in income *per capita*. This is broadly consistent with evidence provided by Kim and Margo (2004), who show that US income *per capita* diverged across regions during the nineteenth century and early twentieth century, and then started converging dramatically. As in Caselli and Coleman

(2001), they relate this to changes in industrial structure across regions. During the industrialization of the Northeast and the formation of the manufacturing belt, regional differences in specialization increased, and with it regional differences in income *per capita*. At the beginning of the twentieth century, this trend reversed, and regional specialization started to decline (Kim, 1998).

This pattern of rising regional divergence followed by a process of regional convergence is common across countries. The relation between income *per capita* and regional dispersion in income *per capita* often exhibits an inverted-U-shaped pattern, a phenomenon Kim (2009) refers to as a "spatial Kuznets curve." In agrarian economies, regional differences are limited. Early industrialization leads to clusters of manufacturing activity emerging in particular locations, leading to an increase in regional income dispersion. As industrialization spreads and agriculture loses importance across the economy, those income differences decline. This pattern has been documented in the *2009 World Development Report* (World Bank, 2009) for both developing countries and developed countries.

22.2.2.1.4 Relation between spatial agglomeration and growth

The discussion above focused on the relation between development and the convergence (or divergence) of income *per capita* across space. Another, not less important, question is how the overall spatial structure of the economy affects *aggregate*, rather than *local*, growth. Since policy makers often try to affect the spatial distribution of economic activity—as mentioned in Section 22.1, nearly 70% of governments implement policies that slow down urbanization—having a convincing answer to this question would seem to be of much interest. Unfortunately, empirical studies are scarce. One of the few examples is the study of Brülhart and Sbergami (2009), who use measures of the spatial concentration of employment for a panel of European countries, and find that greater spatial concentration promotes growth up to a GDP *per capita* threshold of around $12,000 (in 2006 prices). Using urbanization as a proxy for spatial concentration, they find similar results for a large panel of countries across the globe.

22.2.2.2 Theory

22.2.2.2.1 Population growth dynamics and Gibrat's law

There are different possible explanations for why the spatial distribution of economic activity or the population changes with a country's development. Michaels et al. (2012) propose a model that explains how the structural transformation from agriculture to nonagriculture affects the relation between population (or employment) density and growth. Since the timing of the structural transformation is related to an economy's level of development, their theory provides valuable predictions for how spatial growth patterns change along an economy's development path. Their theory also has implications for the evolution of the dispersion of the population over time.

The model consists of a continuum of locations that can produce agricultural or nonagricultural goods using land and labor. Time is discrete and is indexed by t. Workers are perfectly mobile across locations. Workers' preferences between the consumption of agricultural goods, c_A, and the consumption of nonagricultural goods, c_N, are of the constant elasticity of substitution type,

$$\left(ac_A^{\rho} + (1-a)c_N^{\rho}\right)^{\frac{1}{\rho}},$$

with an elasticity of substitution between both types of goods of less than 1, that $1/(1-\rho) < 1$. The production function is of the Cobb–Douglas type for land and labor. Output in sector j and sector i in period t is given by

$$Y_{jit} = L_{jit}^{\eta_j} \Gamma_{jt} \theta_{jit} L_{jit}^{\mu_j} H_{jit}^{1-\mu_j},$$

where L_{jit} and H_{jit} denote land and labor used, and where TFP depends on external economies of scale, $L_{jit}^{\eta_j}$, a sectoral productivity component common across locations, Γ_{jt}, and a location-specific sectoral productivity component, θ_{jit}. Agriculture is assumed to be more land intensive than nonagriculture, so $\mu_A < \mu_N$, and agriculture benefits less from agglomeration economies than nonagriculture, so $\eta_A < \eta_N$. The location-specific sectoral productivity component, θ_{jit}, is hit by idiosyncratic shocks ϕ_{jit}:

$$\theta_{jit} = \phi_{jit} \theta_{jit-1}^{\nu_j},$$

where the parameter ν_j is key, as it is inversely related to the mean reversion in location-specific productivity. In particular, if $\nu_j = 0$, there is no persistence in productivity, so we get high mean reversion; if $\nu_j = 1$, there is high persistence in productivity, so we get low mean reversion. It is assumed that mean reversion in agriculture is greater than in nonagriculture, so $\nu_A < \nu_N$.

Since workers can costlessly relocate, utility equalizes across locations, and the dynamic equilibrium collapses to a sequence of static equilibria. As long as agglomeration economies are not too strong compared with dispersion forces coming from land, the equilibrium of the economy is stable and unique. The theory generates the following results. First, population density is lower in locations specialized in agriculture than in those specialized in nonagriculture. This reflects the relatively higher land intensity in agriculture ($\mu_A < \mu_N$) and the relatively weaker agglomeration forces in that same sector ($\eta_A < \eta_N$). Second, the dispersion in population density is greater across nonagricultural locations than across agricultural locations. This is a consequence of the weaker mean reversion in nonagriculture, which implies the variance and the maximum value of productivity is greater in nonagriculture than in agriculture. Third, the structural transformation displaces the population from agricultural locations to nonagricultural locations, and also makes some locations switch from agriculture to nonagriculture. Relative increases in agricultural productivity, together with an elasticity of substitution of less than 1 between agricultural and nonagricultural goods, underlie this phenomenon.

These predictions are able to explain two of the more relevant features of the data. One is the increased dispersion in population density over time, as shown in Figure 22.4. The greater relative dispersion in nonagricultural employment density implies that the structural transformation away from agriculture increases the overall dispersion in population density. Another is the nonlinear relation between the initial population density and growth: as shown in Figure 22.1, for low-density locations the relation is negative, for medium-density locations the relation is positive, and for high-density locations the relation is orthogonal. On the one hand, for agricultural locations (which tend to be low-density places), strong mean reversion in productivity implies a negative relation between density and growth. Among those locations, the higher-density ones are those which had the highest productivity shocks in the past; in the presence of mean reversion, their relative productivity (and density) is therefore likely to go down. On the other hand, for nonagricultural locations (which tend to be high-density places), there is no relation between density and growth. The absence of mean reversion as $\nu_N \rightarrow 1$ implies that growth is essentially orthogonal to density, so Gibrat's law holds for those locations. In between those two extremes, we have the nonspecialized medium-density locations where the share of agriculture, on average, decreases with the initial density. The structural transformation leads to greater population growth in those locations with a higher proportion of nonagriculture, thus implying a positive relation between the initial population density and growth.

An alternative explanation by Desmet and Rappaport (2013) focuses on transition dynamics and entry. In their one-sector model, locations gradually enter over time. Upon entry, they draw a productivity from a distribution. Frictions on positive population growth slow the upward transition to each location's steady state and so cause population growth from low levels to be characterized by convergence. The congestion arising from the fixed supply of land in each location gradually diminishes over time. This is consistent with either a decrease in land's share of factor income (as in Michaels et al., 2012) or an increase in the effect of agglomeration on productivity (as in Desmet and Rossi-Hansberg, 2009). As this allows steady-state population levels to become more sensitive to underlying differences in exogenous productivity, it introduces a force toward divergence. Once entry is complete and the degree of net congestion has stabilized, the assumed orthogonality of productivity growth causes population growth to be orthogonal as well.

22.2.2.2.2 Income growth dynamics

The models mentioned above remain silent on income *per capita* differences across space, essentially because there is only one type of labor and all workers are perfectly mobile across locations. Caselli and Coleman (2001) introduce different skill types. Although workers are geographically mobile, regional differences in skill composition will lead to income *per capita* differences.

To be more precise, they propose a North–South model of the structural transformation with three basic assumptions. First, TFP growth is higher in agriculture than in manufacturing. The production technologies in food and manufacturing in region i and time t use land (T), labor (L), and capital (K) and are of Cobb–Douglas type of the form

$$F_t^i = A_{\mathrm{f}t}^i (T_{\mathrm{f}t}^i)^{\alpha_T} (L_{\mathrm{f}t}^i)^{\alpha_L} (K_{\mathrm{f}t}^i)^{1-\alpha_T-\alpha_L}$$

and

$$M_t^i = A_{\mathrm{m}t}^i (T_{\mathrm{m}t}^i)^{\beta_T} (L_{\mathrm{m}t}^i)^{\beta_L} (K_{\mathrm{m}t}^i)^{1-\beta_T-\beta_L},$$

where the South has a comparative advantage in agriculture and the North has a comparative advantage in manufacturing. As mentioned before, it is assumed that (exogenous) TFP growth in agriculture, g_{f}, outpaces that in manufacturing, g_{m}.

Second, there is a cost of acquiring nonfarm skills, and this cost drops over time. The demographic structure is that of a dynasty, with a constant population and a probability of death in each period. In each period each person is endowed with one unit of time. When born, a person decides whether to immediately start working on the farm, or to first spend $\xi_t \zeta^i$ units of time getting trained to work in manufacturing, where ξ_t captures the economy's overall efficiency in providing training and ζ^i is distributed among the people of a generation according to a time-invariant density function $\mu(\zeta^i)$. Assuming that ξ_t drops over time implies that training becomes cheaper over time. As a result, the cutoff ζ^i below which individuals invest in skill acquisition rises over time, implying more people become skilled.

Third, the income elasticity of demand for agricultural goods is less than 1. In particular, the period utility derived from consuming food, c_{f}, and manufactured goods, c_{m}, is

$$u(c_{\mathrm{f}t}^i, c_{\mathrm{m}t}^i) = \frac{\left((c_{\mathrm{f}t}^i - \gamma)^\tau (c_{\mathrm{m}t}^i)^{1-\tau}\right)^{1-\sigma}}{1-\sigma},$$

where $\gamma > 0$ is the subsistence constraint on food consumption, implying the less than unit income elasticity of demand for food.

Because of the initially high cost of acquiring nonfarm skills, the relative supply of manufacturing workers is low, implying a substantially higher manufacturing wage. Given that the South has a comparative advantage in agriculture, this implies a wage gap in favor of the North, in spite of labor being mobile across regions. As the overall economy becomes richer because of general productivity growth, the demand for manufacturing goods increases, shifting labor from agriculture to manufacturing. This process is further reinforced by the faster TFP growth in agriculture compared with manufacturing. With a declining weight of agriculture in the economy, average wage differences across regions drop. The falling cost of acquiring nonfarm skills enhances this convergence across regions and has the additional advantage of leading to a reduction in

wage differences not just across regions, but also within regions between farm and non-farm workers. It is this latter feature which the model would not be able to capture if it did not assume a falling cost of acquiring manufacturing skills.

Whereas this model predicts that the structural transformation leads to income convergence across regions, it is likely that in the early stages of industrialization the opposite happened. In the model, part of the convergence between North and South happens because average wages converge as a result of the sectoral composition becoming more similar across regions. During the early stages of industrialization, when the North shifted increasingly into manufacturing, the opposite should have happened. As mentioned before, this would be consistent with the evidence in Kim and Margo (2004), who describe a process of income divergence during the nineteenth century, followed by convergence, which is particularly strong during the second half of the twentieth century.

An assumption in most of these models is that labor is freely mobile across regions. This does not necessarily contradict the evidence of nominal and real wages being substantially higher in the West than in the rest of the country during the nineteenth century (Easterlin, 1960; Rosenbloom, 1990; Mitchener and McLean, 1999). As in Caselli and Coleman (2001), this gap might be due to differences in skills. This does not seem to be the entire story though, since these differences also existed within occupations. Focusing on 23 occupations, Rosenbloom (1990) documents within-occupation average real wage differences of more than 50% between the West and the South in 1870; by the end of the nineteenth century, this difference continued to exist, although it had been cut in half. This suggests that labor markets were not completely integrated, and that moving costs were driving a wedge between wages in the West and the rest of the country.

22.2.2.2.3 Gibrat's law and Zipf's law

An interesting related question is how Gibrat's law is connected to Zipf's law. Theory says that proportionate (or random) growth should give rise to a lognormal distribution (Gibrat, 1931). That is, Gibrat's law implies a lognormal distribution. Consistent with this, Eeckhout (2004), using data on census places, shows that growth between 1990 and 2000 satisfies Gibrat's law and that the size distribution of places is lognormal. Since the lognormal distribution and the Pareto distribution are very different, Gibrat's law seems to be inconsistent with the observation that the city-size distribution conforms to Zipf's law.[1] The puzzle is partly resolved when it is realized that cities make up the upper tail of the size distribution of all locations, and at that upper tail the lognormal distribution is actually very similar to the Pareto distribution. So although Gibrat's law does not imply a Pareto distribution overall, in the upper tail they are similar (see Ioannides and Skouras, 2013 for a further discussion). Note that there are restrictions on the stochastic process which can lead Gibrat's law to imply Zipf's law. For example, Gabaix (1999) shows that if cities cannot fall below a minimum size, then Gibrat's law implies a city

[1] Section 22.4.2.1 has a longer discussion on city-size distributions.

size distribution that converges to Zipf's law. The intuition is simple: we get the density function peaking at the minimum city size and at the same time the lower bound on size pushes more cities to become large, implying the fatter upper tail, characteristic of Zipf's law (see Duranton and Puga, 2014 for a review of this literature).

Returning to the observed lognormality in the size distribution of all places, an open question is whether the lognormality is due to past proportionate (or random) growth or whether it is due to some underlying lognormal distribution of locational characteristics. The finding in both Michaels et al. (2012) and Desmet and Rappaport (2013) that the orthogonality of growth across locations in the United States was categorically violated until recently sheds doubt on whether random growth can have caused the present-day lognormal distribution of the population. This doubt is further enhanced once we observe that the distribution of the population in 1790 was already lognormal, as shown in Figure 22.3. More consistent with the observed growth rates is that the combined underlying determinants of the steady-state population are distributed lognormally (Krugman, 1996; Rappaport and Sachs, 2003). This does not require any one characteristic of a location to be distributed lognormally. As shown by Lee and Li (2013), as long as there enough factors, the population distribution will be lognormal, even if none of the factors individually is lognormally distributed.

22.2.2.2.4 Spatial agglomeration and aggregate growth

So far we have analyzed spatial growth patterns, but we have not focused on the relation between space and aggregate growth. That is, how does the overall spatial structure of an economy affect its aggregate growth rate? There exist some dynamic extensions of two-region *new economic geography* models which were reviewed in the previous edition of this handbook (Baldwin and Martin, 2004). Although these extensions analyze the relation between geography and growth, their focus on a small number of locations limits their ability to capture the overall spatial distribution of the economy.

Incorporating a richer spatial structure into dynamic growth models is complex because it increases the dimensionality of the problem. As discussed in Desmet and Rossi-Hansberg (2010), models become quickly intractable and unsolvable when agents' decisions depend on the distribution of economic activity over both time and space. In recent years, some attempts have been made to incorporate forward-looking agents into models with a continuum of locations (Brock and Xepapadeas, 2008; Boucekkine et al., 2009; Brock and Xepapadeas, 2010). Unfortunately, to keep these spatial dynamic models solvable, they do not take into account many relevant spatial interactions, such as transportation costs and factor mobility.

To get around this problem, Desmet and Rossi-Hansberg (2014a) impose enough structure so that future allocation paths do not affect today's decisions. This result is obtained by assuming that workers are freely mobile and that innovation by firms diffuses across space. The model strikes a balance between being tractable and having a rich spatial structure that allows it to connect with the data. They use their theoretical framework to

study the spatial and aggregate evolution of the US economy over the last half century. To highlight some of the main features of the model, we present here a simplified one-sector version. Land and agents are located on the unit interval $[0,1]$, time is discrete, and the total population is $\bar{L}$. Agents solve

$$\max_{\{c(\ell,t)\}_0^\infty} E\sum_{t=0}^{\infty}\beta^t U(c(\ell,t))$$
$$\text{subject to}$$
$$w(\ell,t)+\frac{\bar{R}(t)}{\bar{L}}=p(\ell,t)c(\ell,t), \quad \text{for all } t \text{ and } \ell,$$

where $c(\ell,t)$ is consumption at location ℓ and time t, $p(\ell,t)$ is the price of the consumption good, $w(\ell,t)$ denotes the wage, and $\bar{R}(t)$ are total land rents, so $\bar{R}(t)/\bar{L}$ is the dividend from land ownership, assuming that agents hold a diversified portfolio of land. Free mobility implies that utility equalizes across locations.

Firms use land and labor to produce. Production per unit of land at location ℓ at time t is $Z(\ell,t)L(\ell,t)^\mu$, where $\mu<1$,$Z(\ell,t)$ denotes TFP, and $L(\ell,t)$ is the amount of labor per unit of land used. A firm's TFP depends both on technology diffusion and on innovation decisions. Technology diffuses between time periods. Before the innovation decision at time t, a firm at location ℓ has access to

$$Z^-(\ell,t)=\max_{r\in[0,1]} \mathrm{e}^{-\delta|\ell-r|}Z^+(r,t-1), \tag{22.3}$$

where the "minus" superscript in Z^- refers to the technology a location has access to before innovation, whereas the "plus" superscript in Z^+ refers to the technology a location ends up using after the innovation decision. In addition to the technology it gets access to through diffusion, a firm can decide to buy a probability $\phi\leq 1$ of innovating at cost $\psi(\phi)$. A firm that obtains the chance to innovate draws a technology multiplier z from a Pareto distribution with shape parameter a and lower bound 1, so the expected technology for a given ϕ is

$$E(Z^+(\ell,t)|Z^-)=\left(\frac{\phi+a-1}{a-1}\right)Z^-.$$

The innovation draws are independent and identically distributed across time, but not across space. Hence, innovation draws are spatially correlated, and firms that are located arbitrarily close to each other obtain exactly the same innovations. The timing of the problem is as follows. During the night, between periods $t-1$ and t, technology diffuses locally. This leads to a level of technology $Z^-(\ell,t)$ in the morning. Each firm then decides on how many workers it wants to hire, how much it wants to bid for land, and how much to invest in innovation. Only the firm that offers the highest bid for land in a given location gets to rent the land. Investment in innovation, if it occurs, then leads to a new technology, $Z^+(\ell,t)$. Production happens at the end of the period.

We now turn to the firm's problem. The objective function of a firm in a given location ℓ at time t_0 is

$$\max_{\{\phi(\ell,t),L_i(\ell,t)\}_{t_0}^{\infty}} E_{t_0}\left[\sum_{t=t_0}^{\infty}\beta^{t-t_0}\begin{pmatrix} p(\ell,t)\left(\frac{\phi(\ell,t)}{a-1}+1\right)Z^{-}(\ell,t)L(\ell,t)^{\mu} \\ -w(\ell,t)L(\ell,t)-R(\ell,t)-\psi(\phi(\ell,t))\end{pmatrix}\right],$$

where β is the discount factor and $R(\ell,t)$ is the firm's bid rent, which is chosen to maximize the probability of winning the auction to rent land. As discussed in Desmet and Rossi-Hansberg (2012), in this setup firms invest in innovation, in spite of operating in a perfectly competitive market, because it allows them to bid a higher price for land. Returning to the above maximization problem, we recall that labor is freely mobile and that firms compete for land and labor every period with potential entrants that, because of diffusion, have access to the same technology. The decision on how many workers to hire and how much to bid for land are therefore static problems. The only problem that is in principle dynamic is the innovation decision, but here as well the dynamic problem simplifies to a static one. The continuity in the diffusion process and the spatial correlation in innovation realizations guarantee that a firm's decisions do not affect the expected technology it wakes up with tomorrow. Hence, future allocation paths do not affect a firm's decision today. This key result is what makes the dynamic spatial model solvable and computable.

The importance of this framework is that it not only has implications for the interaction between density and growth at the local level, but it also analyzes the interaction between the spatial distribution of economic activity and aggregate growth. When applying their framework to the evolution of the US economy in the last 50 years, Desmet and Rossi-Hansberg (2014a) can account for the main spatial patterns, such as the evolution in the dispersion of land prices, as well as the main macroeconomic stylized facts, such as the evolution of aggregate growth and wages. More broadly, the aim is to develop a unified framework to study the interaction between space and the macroeconomy. In other work, the same authors use a similar setup to quantitatively analyze the impact of global warming on both the spatial distribution of economic activity and global welfare (Desmet and Rossi-Hansberg, 2014b).

22.3. DEVELOPMENT, SPACE, AND INDUSTRIES

Although we touched upon the structural transformation from agriculture to manufacturing, our main focus in the previous section was on aggregate population growth across different locations. In this section, we delve deeper into the incentives of different industries to concentrate or disperse, and analyze the geography of sectoral employment growth. In particular, we are interested in the differences between manufacturing and services.

This is related to the broader question of how the spatial distribution of economic activity changes with development, for at least two reasons. First, if spatial growth patterns differ across sectors, then a country's overall spatial organization will change as it develops and the relative importance of different sectors changes. Second, for a given sector, spatial growth patterns may also change over time, as sectors transition from being young to being maturer. In what follows, we discuss some of the recent empirical findings, as well as theories that can account for them.

22.3.1 Manufacturing versus services

In recent decades, US manufacturing has become spatially more dispersed and services have become spatially more concentrated. On the basis of US county employment data between 1970 and 2000, Table 22.1 shows that the difference in the logarithm of employment between the 70th percentile and the 30th percentile decreased in manufacturing and increased in services. This implies manufacturing became more equally spread across US counties, whereas the opposite happened to services. When the standard deviation of the logarithm of employment is used as an alternative measure of the degree of concentration, the result is similar. Since services started off being less concentrated than manufacturing, this implies services becoming more like manufacturing in their degree of spatial concentration.

Does this mean that manufacturing is dispersing and services are concentrating across all locations? To get a more precise idea, Desmet and Fafchamps (2006) and Desmet and Rossi-Hansberg (2009) run nonlinear kernel regressions of the form

$$L^i_{t+s} = \phi(L^i_t) + e^i_t,$$

where L^i_t is the logarithm of employment in year t and county i. Figure 22.5 shows that the tendency toward greater geographic dispersion in manufacturing is happening across the entire distribution. Counties with less manufacturing have been experiencing faster manufacturing employment growth than those with more manufacturing. In the case of services, the picture is more complex: the relation between size and growth is S shaped.

Table 22.1 Spatial concentration of employment (as a logarithm)

	1970	2000
Difference between 70th percentile and 30th percentile		
Manufacturing	1.81	1.74
Services	1.29	1.52
Standard deviation		
Manufacturing	2.05	1.89
Services	1.40	1.52

Source: REIS, Bureau of Economic Analysis.

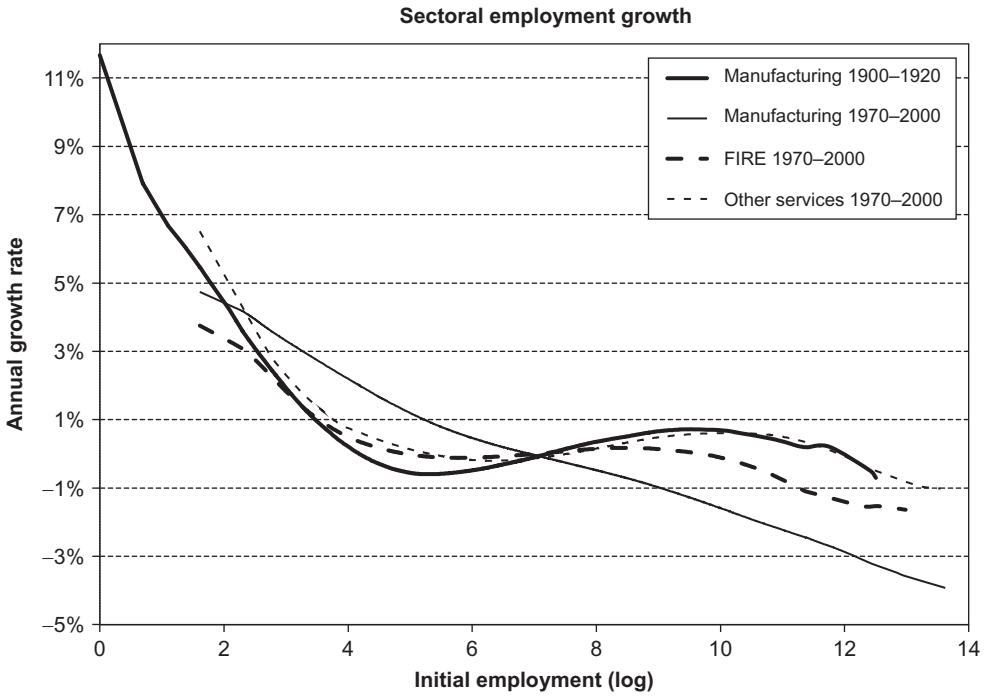


Figure 22.5 Sectoral employment growth in US counties. *Source:* *Desmet and Rossi-Hansberg (2009)*.

The tendency toward a greater geographic concentration of services employment has mainly affected mid-sized service employment clusters. European regions look similar to US counties: deconcentration in manufacturing and greater concentration in services.

22.3.2 Life cycle of industries and spatial distribution

One possible explanation for this trend is the greater land intensity of services. As services compete for the same land as manufacturing in high-density urban environments, services are increasingly competing manufacturing out of cities. Another possible explanation has to do with the life cycle of an industry. Younger industries stand more to gain from knowledge spillovers, which are enhanced by the geographic concentration of economic activity. This could help us understand the recent tendency toward greater concentration in services. As shown by Hobijn and Jovanovic (2001), the impact of information technology (IT) is greater in services than in manufacturing. They compute IT intensity—the share of IT equipment in the total stock of equipment—in different sectors in 1996, and find a figure of 42.4% in services and a much lower 17.9% in manufacturing. Using alternative definitions of the importance of IT, Triplett and Bosworth (2002) and Basu and Fernald (2007) find similar differences between manufacturing and services.

To operationalize the idea of the age of an industry, Desmet and Rossi-Hansberg (2009) propose using the time elapsed since the introduction of a general-purpose technology (GPT). David and Wright (2003) and Jovanovic and Rousseau (2005) argue that the two major GPTs of the twentieth century were electricity and IT. As for their timing, Jovanovic and Rousseau (2005) propose identifying the starting date of a GPT by taking the year in which it reaches a 1% diffusion. In electricity, this corresponds to 1894, the year of the first hydroelectric facility at Niagara Falls, and in IT this corresponds to 1971, the year of the Intel 4004 microprocessor. As the ending date of a GPT, they take the year when the diffusion curve becomes more or less flat. In the case of electricity, this corresponds to 1929, whereas in IT that point has not been reached yet. Roughly speaking, this makes the period between 1900 and 1920 for electricity comparable to the period between 1970 and 2000 for IT.

While IT is viewed as mainly affecting services, electricity's impact was mostly felt in the manufacturing sector (David and Wright, 2003). If age plays an important role in the spatial growth patterns of an industry, we would expect the spatial growth pattern of manufacturing at the beginning of the twentieth century to look similar to that of services at the end of the twentieth century. As seen in Figure 22.5, this is indeed the case. The spatial growth pattern of manufacturing at the beginning of the twentieth century looks very different from that of manufacturing at the end of the twentieth century, but very similar to that of services at the end of the twentieth century. This finding implies that when analyzing the relation between space and growth, not only the sectoral composition of the economy matters but also the age of the different sectors plays a role. There is nothing inherent about manufacturing exhibiting a tendency toward greater dispersion; indeed, when the sector was young, it became increasingly concentrated.

Motivated by this evidence, Desmet and Rossi-Hansberg (2009) provide a theory for how an industry's spatial growth is related to its life cycle. The model has three forces. First, local knowledge spillovers constitute an agglomeration force that incentivizes the geographic concentration of economic activity. Second, crowding costs coming from land constitute a dispersion force. Third, technology diffuses over space. This constitutes an additional dispersion force. The relative importance of these three forces will depend both on a location's size and on an industry's age. Together, they will be able to capture how the scale dependence of an industry's growth evolves over its life cycle.

To see this, it will be convenient to distinguish between three types of locations in terms of their employment size: small locations, medium-sized locations, and large locations. In small locations, knowledge spillovers have little bite, so TFP is determined by the technology a location has access to through diffusion. Hence, among smaller places, we should see increasing divergence. In medium-sized locations, knowledge spillovers become the dominating force. With congestion forces still being weak, we see increasing concentration among medium-sized places. In large locations, local congestion forces start to dominate knowledge spillovers from neighboring locations. Among those large places, we should thus see increasing dispersion.

The above description suggests an S-shaped relation between size and growth: dispersion among both small and large locations, and concentration among medium-sized locations. Going back to Figure 22.5, we see this description is consistent with the growth patterns of services in recent decades and manufacturing at the beginning of the twentieth century. We referred to those as "young" industries at the beginning of their life cycle. In contrast, "maturer" industries, such as manufacturing in recent decades, exhibit increased dispersion across all sizes. The absence of increased concentration in medium-sized locations reflects knowledge spillovers being less likely to outdo the productivity a location has access to through diffusion. Medium-sized locations that could benefit from knowledge spillover from neighboring locations have already done so, and no longer benefit from further increases in their productivity. As this happens, the upward-sloping part of the S-shaped relation between size and growth, present in younger industries, disappears as industries reach the later stages of their life cycle.

Of course, services and manufacturing are broad sectors; not all subsectors of services benefit from IT in the same way. With a focus on two-digit Standard Industrial Classification (SIC) subsectors of services, evidence obtained by Chun et al. (2005), Caselli and Paternò (2001), and McGuckin and Stiroh (2002) suggests that the most IT-intensive sector is legal services and the least IT-intensive sector is auto repair. Using employment at the two-digit SIC level from the County Business Patterns dataset spanning the period 1977–1997, Figure 22.6 shows employment growth in legal services and auto repair. As expected, legal services exhibit the S-shaped spatial growth pattern. In contrast, auto repair looks like a mature sector, with convergence across the entire distribution. In the same way that not all service activity is concentrating, not all manufacturing is dispersing. We would expect manufacturing activities that most strongly benefit from knowledge spillovers to have less of an incentive to disperse. This explains the findings of Fallah and Partridge (2012), who show that high-tech manufacturing pays a relatively higher price for remoteness. In particular, a 1 km increase in the distance from the nearest metropolitan area decreases employment growth by 0.2% in high-tech manufacturing, compared with a 0.1% decrease in manufacturing overall. We would therefore expect high-tech manufacturing sectors to remain more clustered than the rest of the industry.

The more general link between an industry's life cycle and its spatial distribution has also been analyzed by Henderson (2010), who provides evidence of standardized manufacturing dispersing and high-tech manufacturing concentrating. In the specific case of the Internet, Forman et al. (2005) show that its use diffused rapidly across the United States, but its more complex applications, such as e-commerce, predominantly located and developed in cities, where there was more easy access to complementary inventions and activities.

The pattern of spatial concentration followed by spatial dispersion as industries mature has been noted in other countries. For example, in Section 22.4, we will discuss data which indicate that Seoul transformed from being a manufacturing center to a service center from 1970 on. Similarly to what happened in the United States and Europe,

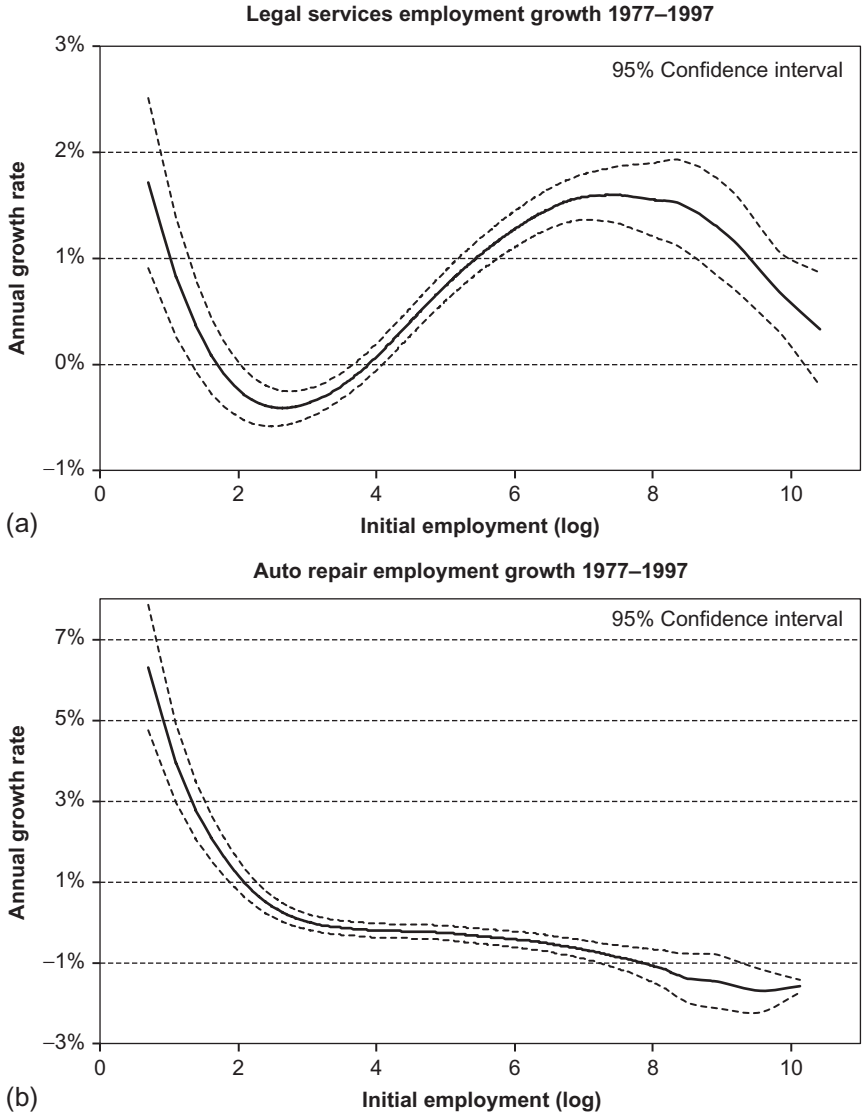


Figure 22.6 Sectoral employment growth in US counties: (a) from Desmet and Rossi-Hansberg (2009) and (b) from Desmet and Rossi-Hansberg (2009).

the loss of manufacturing employment in Seoul benefited the rural areas and the small towns, which experienced an industrial transformation after 1980. Similarly, in the 1990s, the correlation between the manufacturing-service ratio and the size of a city in China was −0.20, implying that larger cities were relatively more service oriented (Au and Henderson, 2006a). Consistent with this, China's 2008 economic census indicates that telecommunications, software, information, and broadcasting services are

highly concentrated at the upper end of the size distribution of counties. It is, of course, important to note that the timing of these transformations may differ across countries. For example, whereas in the United States manufacturing had become a mature industry by, say, the 1960s, in South Korea this same stage was reached only in, say, the 1980s. This underscores a point we made before: to understand the relation between development and space, it is important to know not just the relative sizes of different sectors but also their ages.

The appearance of clusters during the early stages of an industry's life cycle is not a recent phenomenon. Trew (2014), for example, documents the emergence of industrial hot spots in nineteenth century England. In 1750, two counties in England, Lancashire and the West Riding, had between 65% and 70% of all employment in the country's secondary sector. These were not necessarily the densest areas initially, but they experienced tremendous population growth as the industrial revolution took off. In the nineteenth century, Sheffield, for example, grew from a town of 60,000 inhabitants to a large city of 450,000 inhabitants. London, the country's biggest city, was also a major manufacturing center, as were some of the other large cities, such as Manchester and Birmingham (Shaw-Taylor and Wrigley, 2008).

22.3.3 Ruralization versus suburbanization

Although manufacturing clusters are spreading out, they often do not move far away. If so, manufacturing growth should be lower in the clusters themselves but higher in areas close to the clusters. Using data on US counties for the last three decades of the twentieth century, Desmet and Fafchamps (2005) find exactly this. In particular, having 1% more manufacturing employment locally lowered manufacturing employment growth by around 2% annually, whereas having 1% more manufacturing employment 40–50 km away increased manufacturing employment growth by 0.1–0.2% annually. These figures refer to manufacturing clusters, rather than to aggregate clusters.

When we look at total employment, the tendency of manufacturing is to suburbanize rather than to ruralize. If so, manufacturing growth should be relatively low in locations with high aggregate employment and relatively high in locations close to aggregate clusters. Again, this is what Desmet and Fafchamps (2005) find. Having 1% more total employment locally lowered manufacturing employment growth by around 0.2% annually, whereas having 1% more manufacturing employment 40–50 km away increased manufacturing employment growth by a little less than 0.01% annually. Though small, the effects are statistically significant, and amount to something much larger once we take into account that we are looking at average annual growth over a period of three decades.

22.3.4 The cost of remoteness

The general tendency toward greater dispersion is mitigated in several ways. First, as already mentioned, high-tech manufacturing tends to remain clustered in high-density

areas to take advantage of knowledge spillovers. Second, the cost of remoteness differs not only across sectors (high-tech vs. low-tech sectors) but also across functions within sectors. With the fragmentation of the value chain, we are witnessing firms locating headquarters and business services in larger cities and production facilities in smaller cities. The evidence for this is reviewed in the next section when models of functional (as opposed to product) specialization by cities are discussed. In general, since 1950, larger cities have moved toward management activities as opposed to production activities, while smaller cities have moved in the opposite direction (Duranton and Puga, 2005). Although the fragmentation of the value chain and the spatial division of labor respond to standard forces of comparative advantage, there are limits to their scope. For example, Tecu (2013) finds that an average US chemical firm is 1.8% more productive in R&D (in terms of patents) if it increases the number of production workers by 10% in the same metropolitan statistical area. In the average metropolitan statistical area, having an average-sized production facility increases the productivity of R&D by 2.5 times in the chemical industry. Doubling the number of production workers has nearly as large an effect on a firm's R&D productivity as doubling the total number of patents generated in the metropolitan statistical area.

The trade-off between moving to cheaper locations and benefiting from proximity may explain the tendency of the different units of multiestablishment firms to locate not too far from each other. In the UK manufacturing industry, for example, establishments that belong to the same firm tend to cluster no more than 50 km from each other, whereas there is no evidence of such clustering by establishments that are part of different firms (Duranton and Overman, 2008).

22.4. THE URBAN SECTOR

Sections 22.2 and 22.3 started with the urban–rural divide and then turned to an analysis of the evolution of economic activity across the continuum of space in a country, moving from the least to the most densely populated locations. For the continuum, the focus was on the spatial transformation in uses of these spaces: how the distribution of the population and the distribution of industrial and service activities change across the continuum with economic growth and technological change. This section has a narrower focus, which is the subject of a large body of literature. We look at the right tail of the continuum in the denser locations that are typically labeled as urban. Because of the sheer volume of the population living at high densities in this tail, it is often the focus of public policy and institutional reform initiatives, as well as people's images of other countries as defined by their largest cities.

This right tail, the urban sector, consists of a hierarchy of cities of very different sizes and functions that transform over time, as suggested before by the results for the continuum. Within the urban sector, cities specialize relatively, and to some degree absolutely,

in particular export activities, giving cities different sizes and different compositions of production activities, occupations, and functions. There are strong patterns in the variation of compositional specialization across the urban hierarchy by city size at a point in time, as well as variation over time within the urban hierarchy depending on the level of economic development. This urban literature has traditionally focused both on the reasons for and the extent of extreme agglomeration and on analyzing why production activities and occupations vary across the hierarchy and over time. There are a number of chapters in prior handbooks which detail work in the literature up to the early years of the twenty-first century (e.g., Abdel-Rahman and Anas, 2004; Duranton and Puga, 2004; Fujita et al., 2004; Gabaix and Ioannides, 2004; Holmes and Stevens, 2004; Henderson, 2005). We focus on developments since then.

In Section 22.4.1, we start by reviewing some basic facts on specialization within urban hierarchies in different countries today, and then turn to a discussion of models that capture key aspects of the industry, occupation, and functional specialization we see across parts of the urban hierarchy. In Section 22.4.2, we take a more dynamic look, building on the analysis of the structural transformation in Section 22.3. We look at how the products and functions of bigger versus smaller cities have altered dramatically over the last 25 years in particular countries, both developed and developing, with aspects of that transformation depending on the stage of economic development. In Section 22.4.3, we turn to an examination of some policies which have strong effects on aspects of a country's urban hierarchy and thus may affect the relative efficiency of the spatial organization of production.

22.4.1 Production patterns in the urban hierarchy

22.4.1.1 Facts

Older work characterized product specialization in two ways. One way was by using cluster analysis to classify cities as steel cities, auto cities, wood product cities, and the like. The second way was to see how the elasticities of sectoral employment with respect to city size differ across sectors (Henderson, 1997; Kolko, 1999; Black and Henderson, 2003). For the United States some facts emerge. Small and medium-sized cities were historically relatively specialized in manufacturing, but that degree of specialization has declined as the country has deindustrialized. Specialization in standardized services by smaller and medium-sized cities has increased. Bigger cities have a much more diverse industrial base, and they are much more engaged in the provision of more sophisticated business and financial services.

Here we evaluate more recent work. For developed countries, recent contributions characterize specialization not just by products but also by functions and occupations, with the idea that bigger cities are specialized more in more highly skilled occupations and functions. Although specialization may still be characterized by elasticities of sectoral employment with respect to city size to show what is produced more in different parts of the urban hierarchy, the literature now also uses spatial Gini or Krugman indices to

characterize the degree to which different cities are specialized (without reference to what they are specialized in *per se*). Another body of literature, which we do not review here, focuses on identifying which specific industries are more clustered in space versus more diffusely spread out, compared with a random allocation across space (Ellison and Glaeser, 1997; Duranton and Overman, 2005).

For developing countries, there are a few recent articles looking at specialization that offer a somewhat different perspective. One more innovative article focuses on a different dimension: the division of labor between and within cities, as it varies across a less developed hierarchy. In Section 22.4.2, we also look at some recent patterns concerning urban specialization in China.

22.4.1.1.1 What big cities do and their skill composition

Figure 22.7 shows different manufacturing industries and their elasticities of local employment with respect to the metropolitan area population. The figure tells us two things. First, traditional industries producing standardized products such as wood products, furniture, and paper products have low elasticities, consistent with findings in the earlier work noted above. Higher technology industries such as the computer and

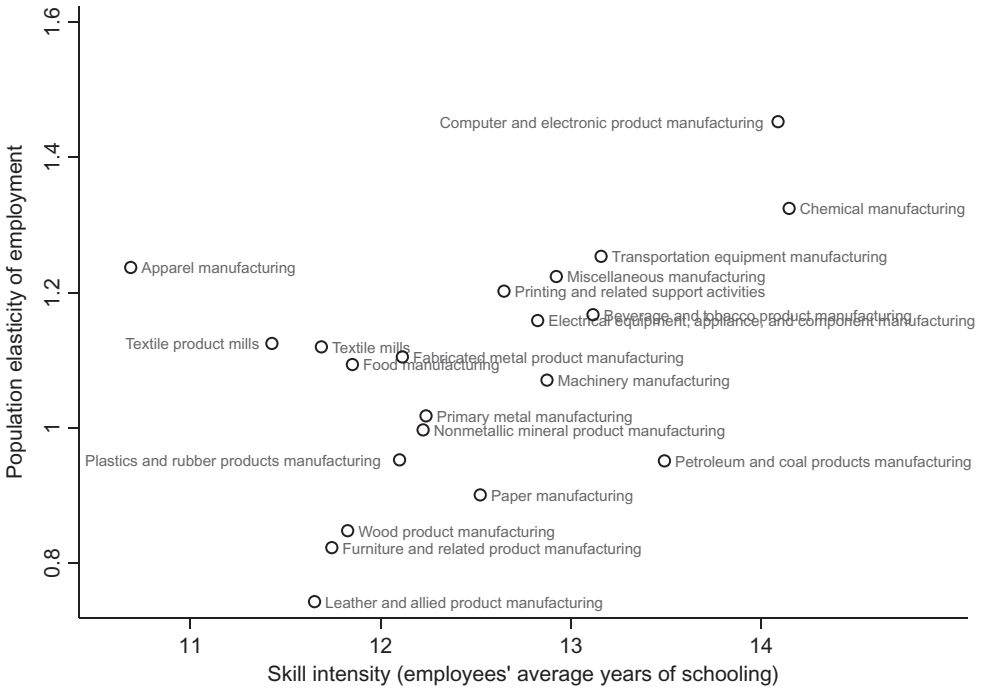


Figure 22.7 Relative industry composition position in the urban hierarchy and relationship to industry skill intensity. *From Davis and Dingel (2013, Figure 14).*

electronics industries have higher elasticities, consistent with the idea that high-tech production benefits from the diverse environments of large cities. Second, in general the skill intensity of industries is correlated with these elasticities, suggesting skill intensity also rises with city size (as Davis and Dingel, 2013 show separately). Skill intensity is measured by the average number of years of schooling of people working in an industry nationally. The only strong outlier is apparel which is a low skill industry and which has a high elasticity. This may reflect the recent surge in the immigrant proportion of the workforce in the apparel industry, where migrants' first landing points are disproportionately cities such New York and Los Angeles. Figure 22.7 covers only manufacturing.

Figure 22.8 looks at the role of big cities for the universe of workers, focusing on occupational mix. Traditional occupations such as farming, food preparation, and health care support have again lower city size elasticities and low education, while computing and mathematical, architectural, and engineering occupations have higher elasticities and high skill levels. Taken together, the findings show that today the United States has lower-skilled workers in standardized manufacturing and services in smaller cities, with higher-skilled workers in often more innovative and creative industries and occupations in larger cities.

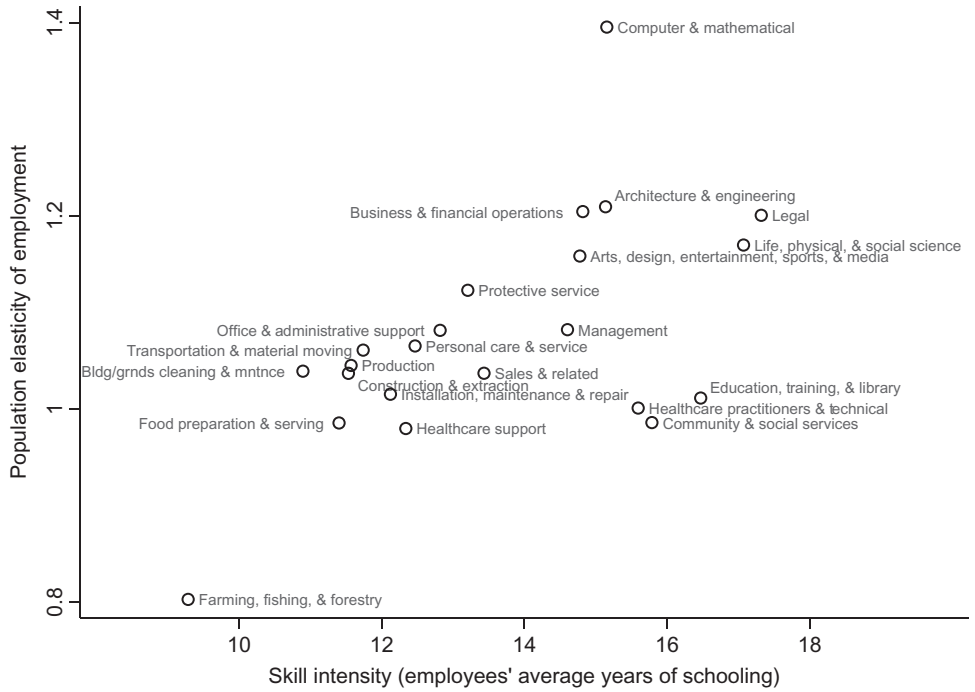


Figure 22.8 Relative industry occupational position in the urban hierarchy and relationship to industry skill intensity.

22.4.1.1.2 Specialization in the urban hierarchy

The next feature concerns the degree to which cities are specialized. For individual cities, the standard measure of the degree to which a city is specialized is the "Gini" specialization index taken from Krugman (1992):

$$K_j = \frac{1}{2}\sum_{i}^{n} |s_{ij} - s_i|,$$

where s_{ij} is subindustry i's share of city j's total industry employment, and s_i is industry i's share of national total industry employment. The higher the index, the more specialized (less diverse) the locality is. The range is from 0, where the city's shares of different subindustries perfectly mimic the nation's share of different industries, to values approaching 1 for a city that is completely specialized in a minor product nationally. An alternative index takes the squares of the deviations, thus giving more weight to bigger deviations. However, that index is mechanically affected by the count of industries in the SIC, which changes over time (the index falls mechanically as the number of industries rises). The Krugman Gini coefficient is free of that mechanical effect.

The second and third columns in Table 22.2 show the Gini coefficient for different size classes of cities for 1977 and 1997 based on County Business Patterns data. Two things are apparent for the United States. First, going down the urban hierarchy by city size class, we find specialization increases sharply. Second, the specialization index has declined for size categories over time, consistent with the manufacturing diffusion analysis in Section 22.3.

The next columns in Table 22.2 deal with a different aspect of activity composition across the urban hierarchy: how firms organize their internal functions by size class, leading to functional specialization by firms across the urban hierarchy. Duranton and Puga (2005) calculate the average number of executives and managers relative to production workers in manufacturing in all cities for 1950 and for 1990. Then they calculate that number within each size class and show the percent deviation of the size class in that year

Table 22.2 Specialization in manufacturing and function specialization across size classes of cities

Population (millions)	Sectoral specialization (Gini coefficient)		Functional specialization (management vs. production)	
	1977	1997	1950	1990
> 5	0.377	0.374	+10.2%	+39.0%
1.5–5	0.366	0.362	+0.3%	+25.7%
0.5–1.5	0.397	0.382	−10.9%	−2.1%
0.25–0.5	0.409	0.376	−9.2%	−14.2%
0.075–0.25	0.467	0.410	−2.1%	−20.7%
< 0.075	0.693	0.641	−4.0%	−49.5%

Source: Duranton and Puga (2005).

from the national average. Bigger cities have relatively more managers and executives in both years, but the degree of differentiation of managers and executive shares between small and large cities has increased enormously over time. Underlying this is a large increase over time in functional specialization by firms nationally (Kim, 1999), with production-oriented versus management-oriented activity increasingly in different locations.

Several articles explore functional specialization by firms across the urban hierarchy in recent years using micro data. Ono (2003) shows that in bigger cities, production plants found there rely more on the headquarters to buy business services for them. More generally, the headquarters are located in larger cities and enjoy a high degree of local scale externalities (Davis and Henderson, 2008). The headquarters outsource most services functions and are located in larger service-oriented cities (Aarland et al., 2007). Smaller cities house plants of firms in certain sectors of manufacturing and have relatively less business and financial services.

There are other dimensions to urban specialization and the hierarchy. A notable article by Fafchamps and Shilpi (2005) documents a pattern in specialization that may be typical in developing countries, using data from the Nepal Labour Force Survey. Note that countries at Nepal's stage of development have yet to develop a sophisticated manufacturing structure, let alone a corporate structure to produce. The data show how individuals allocate their hours to wage work, nonfarm self-employment, agriculture, construction, food processing, handicrafts, and other work. It also lists the main occupation of an individual for 56 International SIC occupation codes. Looking at patterns across 719 wards covering most of Nepal, the study authors have two key findings. The first concerns specialization in the allocation of time by individuals, which reveals a new result for the literature. Such specialization declines as people in a ward are less exposed to nearby urban populations, or live in less urbanized areas. An increase in the urban population nearer to a person induces more specialization in individual tasks—more Adam Smith specialization. The second finding concerns specialization at the ward level by the local population, where, as people in a ward are less exposed to nearby urban populations, ward specialization increases. This pattern suggests that wards nearer to cities can support a more diverse set of individual occupations while remote places are more specialized, paralleling at a different spatial scale what we saw in the second and third columns in Table 22.2. In contrast, Adam Smith specialization for individuals increases with greater exposure to urban markets.

22.4.1.2 Modeling the urban hierarchy

22.4.1.2.1 Initial attempts at a hierarchy: A benchmark model of city sizes and hierarchies

The initial systems of cities literature (Henderson, 1974) envisioned an equilibrium system with differing types and corresponding sizes of cities, where each type was specialized in the production of one traded good. The idea was that if scale effects were ones of localization (internal to the own industry), specialization accentuated the efficiency gains from

agglomeration relative to the congestion costs of increasing city size. The model has workers with identical skills and capital. The equilibria are free-mobility ones, meaning that workers are free to move across cities and in equilibrium earn equal utility everywhere. Henderson considered two regimes. In the first regime, there are agents who operate in national land markets to set up cities, such as developers or city governments. City sizes under such a regime are determined by developers or local governments which set sizes to optimize local net rents or per worker utility. Equilibrium sizes represent a trade-off between the marginal benefits of enhanced scale effects in production versus the marginal costs of increased commuting or generalized congestion from increasing city size. In an urban hierarchy, bigger types of cities are specialized in industries with greater marginal benefits of enhancing scale. In the second regime, there are no organizing agents operating in national land markets, and cities form through what is now called self-organization, a term introduced by Krugman (1996). With self-organization workers move across cities to equalize private marginal products but individually cannot act to internalize externalities. As we will see, in this regime, there are a continuum of potential equilibria where cities are generally too large.

Initial extensions of the basic model focused on modeling the microfoundations of local scale economies, which in Henderson are presented as traditional black-box scale externalities. Fujita and Ogawa (1982) model information spillovers as decaying with distance, which provides an incentive for people to cluster in agglomerations. Helsley and Strange (1990) model how the efficiency of search and matching in labor markets improves in thicker markets. Building on earlier work developing monopolistic competition models in urban economics,[2] Abdel-Rahman and Fujita (1993) model diversity of local nontraded intermediates which increases with urban scale, and thus provides greater choice and efficiency for final good producers in the city. Duranton and Puga (2004) present a detailed analysis of how to add other, more behavioral elements when thinking about microfoundations of scale externalities.

Another set of initial extensions focused on two aspects of urban hierarchies. First, rather than cities being specialized in one export good, in reality most cities export more than one good. Second, as we have seen, diversity of traded good production (i.e., manufacturers) increases as we move up the urban hierarchy. An early attempt to add such elements was by Abdel-Rahman and Fujita (1993), who looked at a world with two cities and two final traded goods (which can be produced with diversified intermediate nontraded inputs). Each final good requires fixed inputs, where the requirement is lower for one good than for the other. However, if the two industries colocate, these fixed costs can be reduced for firms in each sector. In their equilibrium, there is a city specialized in the good with the lower fixed costs, and the other, potentially larger city is diversified. Tabuchi and Thisse (2011) present a similar model and outcome, but now the two goods

[2] See, for example, the special issue of *Regional Science and Urban Economics* in 1988 edited by Fujita and Rivera-Batiz.

differ in the unit costs of intercity trade. In this case, the specialized city is the one with the lower unit trade costs. While these models do give specialized versus diversified cities, the environment is not rich. The number of cities is exogenously set at two and the distinction between goods has limited intuition.

In the recent literature, much more sophisticated modeling of production and labor force structure across the urban hierarchy has emerged. We turn to these in the next section, but as a reference point, we review key aspects of the basic model reviewed in detail in both Duranton and Puga (2004) and Abdel-Rahman and Anas (2004).

A benchmark model. For our benchmark, we use as microfoundations for scale externalities the diversity of intermediate inputs framework. It is straightforward to reformulate the model to allow the scale benefits to be other microfoundations, such as greater scale promoting greater specialization of workers in their tasks (Becker and Henderson, 2000; Duranton and Puga, 2004). A city has production functions for final and intermediate producers, respectively, of

$$y = \left(\int_0^m x(h)^{1/(1+\varepsilon)} \mathrm{d}h \right)^{1+\varepsilon}$$

and

$$X(h) = \beta l(h) - \alpha,$$

where $l(h)$ is labor input for firm h, and $x(h)$ and $X(h)$ are, respectively, inputs of type h for a final good firm and output of the intermediate good producer of type h. For other notation m is the endogenous number of intermediate good producers, L is the effective city labor force, Y is total final good output, the price of the final good is the numéraire, and the price of intermediate inputs is q. Using key results from standard cost minimization of final producers and from profit maximization and competition among intermediate producers,[3] we find the reduced form expressions for final good output per worker in the city and wages, respectively, are

$$Y/L = CL^{\varepsilon}$$

and

$$w = (\beta/(1+\varepsilon))m^{\varepsilon} = (\varepsilon/\alpha)^{\varepsilon}(\beta/(1+\varepsilon))^{1+\varepsilon}L^{\varepsilon}.$$

Both output per worker and wages increase with city scale, as ultimately measured by total effective employment. Note the reduced form specification looks like black-box externalities.

[3] For cost minimization, we have that the direct elasticity of derived demand is approximated by $-(1+\varepsilon)/\varepsilon$ and that final price $1 = \left(\int_0^m q(h)^{-1/\varepsilon}\mathrm{d}h\right)^{-\varepsilon} = qm^{-\varepsilon}$, where the last term emerges in the symmetric equilibrium. Profit maximization conditions by intermediate producers and free entry allow us to solve for the wage level $w = \beta/(1+\varepsilon)q$, firm output $X = \alpha/\varepsilon$, and the number of such producers in the city $m = \beta\varepsilon/((1+\varepsilon)\alpha)L$.

Given these positive benefits of increasing scale, what economic forces serve to limit city sizes and serve as a counterbalance to scale benefits from agglomeration? To answer this requires the introduction of sources of urban diseconomies. Such diseconomies are typically modeled as coming from increases in urban commuting costs. The standard approach assumes a monocentric city with fixed lot sizes where all production occurs at a point in the city center. Following the specifics in Duranton and Puga (2004) for a linear city, each worker is endowed with one unit of time and the working time is $1 - 4\tau u$, where u is the distance from the city center and 4τ is the unit commuting costs. It is then easy to derive expressions for the effective labor force L, for total rents in the city, and for the net wage after rents and commuting costs, all as functions of the city population N.[4] For use below we have

$$L = N(1-\tau N);\ \ \text{net wage income} = w(1-2\tau N);\ \ \text{total rents} = w\tau N^2.$$

The final step is to introduce the mechanism to determine city sizes. The standard one following the first regime in Henderson (1974) assumes the existence of "large agents" operating in national land markets who serve to coordinate agglomeration. These could be developers who own city land and set city sizes and any subsidies to workers or firms to maximize their profits, or alternatively (and equivalently) they could be city governments, who can tax away land rent income from landowners and set city sizes to maximize real income per worker. As an example, developers seek to maximize

$$\text{Profits} = \text{total rents} - \text{worker subsidies} = w\tau N^2 - sN = \left(\frac{\varepsilon}{\alpha}\right)^{\varepsilon}\left(\frac{\beta}{1+\varepsilon}\right)^{1+\varepsilon}\tau N^{2+\varepsilon}(1-\tau N)^{\varepsilon} - sN$$

$$\text{subject to } \bar{y} = \left(\frac{\varepsilon}{\alpha}\right)^{\varepsilon}\left(\frac{\beta}{1+\varepsilon}\right)^{1+\varepsilon}N^{\varepsilon}(1-\tau N)^{\varepsilon}(1-2\tau N) + s,$$

where s is any subsidy developers pay workers to join their city and $\bar{y}$ is the going real income available for workers in national labor markets, as perceived by any city. These subsidies could also go to firms, but in this simple example this is irrelevant.[5] Assuming that developers maximize profits with respect to s and N and that, with competition, cities earn zero profits, solving the problem gives the equilibrium (and efficient) city size[6]:

[4] The population comes from integrating over the two halves of the city, each of length $N/2$. The rent gradient is derived by equating rent plus commuting costs for a person at u with that of a person at the city edge, where rents are 0. Total rents come from integrating over the rent gradient.

[5] There is no misallocation of resources here, despite fixed costs of production and monopolistic competition, because diversified inputs are the only factor of production and enter symmetrically.

[6] There is also the Henry George theorem where all rents in the city are paid out to workers in subsidies to cover the marginal externalities they generate (more workers bring more varieties and greater efficiency of final good producers). In particular, $dY/dN = (1+\varepsilon)[(\varepsilon/\alpha)^{\varepsilon}(\beta/(1+\varepsilon))^{1+\varepsilon}N^{\varepsilon}(1-\tau N)^{\varepsilon}(1-2\tau N)]$, where the term in the square brackets is the private benefit of adding a worker (his/her net wage) and ε times the expression in square brackets is the externality, which also equals s in equilibrium.

$$N^* = \frac{\varepsilon}{\tau(1+2\varepsilon)}; \;\; \partial N^*/\partial \tau < 0, \, \partial N/\partial \varepsilon > 0.$$

As constructed, this is also the size that maximizes net income per worker, y, including the subsidy set equal to average land rents.[7] This implies that y is an inverted-U-shaped function of N with equilibrium and optimum city size at this maximum. That equilibrium and optimal size coincide in this context depends on the use of subsidies to residents to effectively internalize scale externalities, as financed by land rents. If, for example, land rents go to absentee owners, as reviewed in Abdel-Rahman and Anas (2004), cities will be too small.

There are some loose ends before proceeding to recent developments. What happens under the self-organization regime? The requirement for a Nash equilibrium in worker location choices is that no worker wants to change cities in equilibrium. Given that income, y, is an inverted-U-shaped function of city size, this has two implications. The first is that the equilibrium size is at the peak or to the right of the peak where $dy/dN < 0$. That is, if a worker moves to another city (by increasing its size), he/she would earn less than what he/she earned in the city he/she left (where real income would rise as he/she left). Thus, it is also the case that cities to the left of the peak where $dy/dN > 0$ cannot be Nash equilibria. The second implication is that all cities be of the same size so as to equalize real incomes. There is then a continuum of equilibria in city sizes between the peak and a size to the right of the peak, N_{max}, where $y(N_{max}) = y(N;N = 1)$. Beyond N_{max} workers would deviate to form a city of size 1, which would then induce migration flows and self-reorganization until there was a new equilibrium where all cities again had a common size between N^* and N_{max}.

Thus, in general, city sizes under self-organization are oversized, potentially enormously so. However, there are models where under self-organization there are unique and more reasonable city size solutions. In the absence of optimizing city land developers, Henderson and Venables (2009) show that in a world with durable housing capital as a commitment device equilibrium city sizes are unique and that, while cities are oversized, they are only modestly so. Behrens et al. (2014) have another, reasonable self-organization equilibrium for the special case they focus on with a continuum of heterogeneous workers. However, most of the literature avoids the self-organization paradigm by assuming either that the number of cities is fixed so city formation is not an issue or that, with an endogenous number of cities, there are land developers who act as optimizing agents to achieve potentially efficient and unique outcomes.

Finally, as alluded to above, to get a hierarchy we would specify that there is a variety of final consumer products, or sectors, each produced with different degrees of scale economies (ε) in their use of local nontraded intermediate inputs. Having different ε is generally enough to guarantee specialization and a hierarchy, but that is fully assured

[7] That is, N^* maximizes $[(\varepsilon/\alpha)^{\varepsilon}(\beta/(1+\varepsilon))^{1+\varepsilon}N^{\varepsilon}(1-\tau N)^{\varepsilon}(1-\tau N)]$ such that $dy/dN = 0$ and $d^2y/dN^2 < 0$.

if we also assume that the production of inputs is specific to each final good sector. As noted earlier, this assumption of "localization" economies means that there are no benefits to industries from colocating. With costs on the commuting consumption side (higher commuting distances and rents), specialized cities are more efficient than diversified ones, as they more fully exploit localization economies. With a fixed set of final goods, in a developer regime, we will have different types of cities, each specialized in one type of product as in Henderson (1974). The sizes of a city by type increase as ε increases across types. While here specialization involves final goods that are uniform in quality, in many recent applications, they could be diversified products within sectors (or types of cities) sold under monopolistic competition. We could also have each city specialized in one particular variety of a traded product Y under monopolistic competition with differ values of ε, as in Au and Henderson (2006a), who estimate a simple structural model applied to China.

22.4.1.2.2 The second generation of hierarchy models

With this simple benchmark in mind, we now turn to the second generation of models developed in the early years of the twenty-first century by Duranton and Puga.

Nursery cities and the product cycle. The second generation of hierarchy models starts with Duranton and Puga (2001), who have an endogenous number of cities, introduce at least one type of diverse city, and develop models that relate to the larger economics literature. In their 2001 article, they build upon the product cycle hypothesis from international trade. That model seeks to explain why product innovations are carried out in major centers (in our case big cities), but, once standardized, production moves to lower-cost sites (in our case smaller cities). In Duranton and Puga (2001), there are m types of final goods, each produced by firms using varieties of type-specific diversified intermediate nontraded inputs. Diversified nontraded inputs of type j must be produced by workers with the same labeled aptitude, where there are thus m types of workers. Final good firms are subject to a probability δ of dying in a period, so there is firm turnover, with new firms appearing in each period. Most critically, each new firm draws an ideal type of intermediate input it must use, but it does not know what that type is. It experiments with different intermediate inputs of type j produced by workers with aptitude j, producing prototypes at a higher cost until it finds its ideal type. Once it chances upon its ideal type, its costs of production fall (thus signaling that the producer has found its ideal type).

How does this fit into urban structure? Using the developer framework for how cities are established, in equilibrium there are specialized cities, where for type j there are only workers and intermediate producers of type j in the city and all final good producers in the city have discovered their ideal type is j. For those specialized cities, scale benefits arise only from having more type j intermediate producers. Thus, as in the previous

subsection, specialization comes from maximizing scale benefits relative to commuting costs, by having only type *j* producers, given an absence of any cross-industry scale effects.

The second type of city is novel: a diversified nursery city. In such a city, all sectors are represented and there are equal numbers of each type of worker and of each type of intermediate good producer. Final good producers produce prototypes as they seek to learn their ideal technology. Why does this experimentation occur in diversified cities, rather than in specialized ones? Duranton and Puga (2001) assume that to move from city to city is costly; a final producer loses a period of production. Thus, to experiment by visiting different specialized cities can become quite costly, whereas to shift input types to experiment within the same city is costless.

Note two key aspects of the nursery city equilibrium just portrayed. The cost of moving across cities (loss of production for a period) must be sufficiently high relative to the scale economies from being in a specialized city, so new firms do not experiment in only specialized cities. But it cannot be so high that once firms know their ideal type they do not want to move to a specialized city (with its lower production costs) for the horizon of their life. Note that this tension also places limits on how relatively high the probability of dying may be. The triumph of the model is not just having a new type of diversified city, but in also formalizing an urban version of the product cycle model. Recently, empirical work and some theoretical work have focused more directly on the role of innovation in cities; this work is reviewed in the chapter by Carlino and Kerr (2015) in this handbook.

Functional specialization and diversity. Duranton and Puga (2005) explore a different type of hierarchy where rather than distinguishing only between product types, they also distinguish between functions. Production units of a firm use intermediate physical inputs and service inputs provided by their headquarters. The headquarters produces these services with intermediate service inputs and labor. Both services and physical intermediate inputs are produced with labor and are not tradable across cities. As in the nursery city model, there are workers belonging to specific occupations (aptitudes) and thus firms in different sectors. Production units use sector-specific intermediate inputs. In contrast, the headquarters of different firms in any city use a common set of business service inputs. So all types of headquarters use lawyers and accountants, but only apparel firms use textile inputs. Firms may spatially integrate so the headquarters and production are located in the same city or they may be multilocation firms, with their headquarters and production units in different cities. Most critically to get their results, multilocation production raises the cost of a production unit to acquire its headquarter services by a factor $\rho > 1$, relative to it being in the same location. However, having the headquarters in separate specialized business service cities allows a greater diversity of intermediate business services of benefit to all types of firms and their headquarters.

Given these implicit trade-offs, the equilibrium has a multilocation pattern for firms, and there are two sets of cities. One set comprises cities specialized in headquarter and business service production. The other set comprises cities specialized in the production

of one type of final good and their corresponding intermediate inputs. Duranton and Puga (2005) call this functional specialization by cities, where now the diversified city is one where the headquarters of different production sectors enjoy a diversity of common business service inputs that are not traded across cities. We note this functional specialization equilibrium will not exist if the cost of having production units acquire headquarter services from other cities is high enough.[8]

22.4.1.2.3 The third-generation models

In the last few years, several articles have introduced more sophisticated considerations into modeling urban hierarchies. Prior work, even in the second-generation models, took a simple approach to looking at urban specialization and diversity. Very recent work has introduced several innovations. First and foremost is allowing for labor heterogeneity, not just different labor types (horizontal differentiation) but also different labor talents or skills (vertical differentiation). This introduces the possibility of labor sorting by talent across the urban hierarchy. Second, in bigger cities, competition among firms may be "tougher" and different qualities of firms may survive. Third, there may exist more complex sorting by industries across the urban hierarchy, based on more complex specifications of interindustry interactions and scale externalities.

Such sorting is critical to the evaluation of urban productivity. In developing and even developed countries, some policy makers evaluate that bigger cities are inherently more productive. That has in certain instances become a basis for advocating that these cities should be effectively subsidized at the expense of smaller cities, an issue we will return to in Section 22.4.3.2. However, small cities persist in developed market economies, suggesting that they are competitive and thus productive. The issue is that in the data we typically observe higher measured output per worker in bigger cities, which could be a basis for the evaluation of policy makers. But this does not mean that bigger cities are more productive. The puzzle can be explained by the types of sorting just noted.

First, we know from Figures 22.7 and 22.8 that more educated and higher-skilled workers sort into bigger cities. So if we observe higher output per worker in a bigger city, the question is to what extent is that because of pure productivity effects versus because of higher quality labor. Models that tackle sorting across cities help us to understand that issue better. Second, if competition in bigger cities is tougher so that only higher productivity firms survive there, that also lowers the component of higher output per worker in bigger cities owing to pure productivity effects. Third, there is industry sorting across cities, where only certain types of industries are found in bigger cities. In the early part of this section, we discussed the idea that industries with greater

[8] If ρ exceeds a critical value, then the equilibrium has only integrated production. Then each city type specializes in production of one type of final output and hosts just the headquarters of the firms in that city and their corresponding intermediate physical and business service suppliers.

localization economies of scale should be found in larger types of cities, with also higher costs of living, while those with lower localization economies may be found in smaller types of cities. Equilibrium in national labor markets with equalized real wages will also require higher output per worker and wages in bigger cities to offset higher costs of living in those cities. The key is that different sizes of cities house different industries; or smaller cities are competitive in what they produce. However, recent work suggests that the issue is more complicated when there are cross-industry externalities. Maybe an industry with high localization economies in a bigger city would really benefit from having an industry with lower localization economies colocate there, but that may not be realized in a market equilibrium and makes local policy enactment and evaluation of productivity more complicated.

There are several articles that tackle theoretical models of sorting of workers across cities, apart from empirical modeling, which we do not cover here (e.g., Baum-Snow and Pavin, 2012). The first article we look at examines sorting across cities, with the distinction that the model links such sorting across cities to residential sorting within cities (Davis and Dingel, 2013). To achieve this neat link, sorting in that article always goes in the direction of having more skilled workers sort into bigger cities. Another article, which focuses only on sorting across cities, questions the presumption that there is monotonic sorting (Eeckhout et al., 2014). The third article we discuss combines sorting of workers across cities with the idea that competition may be tougher in bigger cities (Behrens et al., 2014). This article has a number of nice innovations, one being the endogenous formation of firms within cities. Finally, we analyze the article by Helsley and Strange (2014) on sorting of industries across cities in the face of cross-industry scale externalities.

Sorting within and across cities. Davis and Dingel (2013) develop a model of sorting across and within cities, albeit in a context where the number of cities is set exogenously. Cities have internal space, which is required if workers are going to sort with regard to where to live within the city. Similar to the benchmark model, final output is produced just with intermediate inputs, but now from a fixed set of intermediate input sectors. In Davis and Dingel (2013), intermediate inputs are sold competitively, traded costlessly within and across cities, and produced by workers with different skills, where there is perfect substitutability among skills in production in any sector σ. The higher σ, the more "advanced" the sector is, as defined below. A worker living in city c at location δ in the city with skill ω chooses which sector σ to work in so as to maximize wages net of rent, or

$$\max_{\sigma} p(\sigma)A(c)D(\delta)H(\omega,\sigma) - r(\delta,c).$$

The worker takes the price $p(\sigma)$ of output in the sector as given. Locations in a city are ordered by values of δ, with the most desirable at $\delta = 0$, and $D' < 0$. While the interpretation can be quite general, to fix ideas and to meet a regularity condition that better locations be "relatively scarcer" in a smaller city, we adopt the Davis and Dingel example

where all cities are circular with fixed lot sizes, δ is the distance from the city center, and $D(\delta)$ is linear. Note furthermore that $r(\delta,c)$ is the rent at location δ in city c, and $A(c)$ is an urbanization productivity level in the city where, for $Lf(\omega,c)$ being the quantity of ω skilled people in city c, $A(c) = J\left(L\int_{\omega\in\Omega} j(\omega)f(\omega,c)\mathrm{d}\omega\right), J', j' > 0$. Heterogeneous individuals have density function $f(\omega)$ on support $\Omega \in [\underline{\omega},\bar{\omega}]$. An equilibrium will have relatively more high-skilled people in bigger cities and thus A is higher in bigger cities because of both scale and skill composition. Finally, worker technology, $H(\omega,\sigma)$, increases with ω and is supermodular (Costinot, 1999), so H has a larger value for the same skill ω in a more advanced sector.

To solve for the within-city and across-city sorting, Davis and Dingel (2013) utilize the perfect substitutability of skill in production of intermediate inputs. Then, in equilibrium the marginal returns to ω in sector σ are independent of the assignment of ω's to the sector. With perfect substitutability, the worker's choice of σ simplifies to $M(\omega) = \max_\sigma p(\sigma)H(\omega,\sigma)$ and defines $G(\omega) \equiv H(\omega,M(\omega))p(M(\omega))$, $G' > 0$, where then a worker's choice of the sector in a city is independent of his/her location choice δ. This in turn yields a simplified location problem within the city of $\max_\delta A(c)D(\delta)G(\omega) - r(\delta,c)$. Within a city, higher-skilled people outbid lower-skilled people for better locations, because they have a higher willingness to pay for better locations, or $\frac{\partial^2}{\partial\delta\partial\omega}A(c)D(\delta)G(\omega) < 0$.

We can now turn to some properties of an illustrative equilibrium with two cities, where city c will be larger than city c' in equilibrium. If we think of $A(c)D(\delta)$ as measuring the attractiveness of a location in city c, then $A(c) > A(c')$ and $L(c) > L(c')$. Why? With rents standardized to zero at each city edge, those least desirable locations in each city in equilibrium must have the same general attractiveness as they will house the same type of worker, the lowest-skilled ($\underline{\omega}$) people. Across cities, in the larger city, the highest-skilled people will live nearest to the city center in locations that are more desirable than any in the smaller city. Only the very highest skilled people in $(\omega,\bar{\omega}]$ are found in the larger city living between $\delta(c) = 0$ and $\delta(c) = \widetilde{\delta}$. At $\delta(c) = \widetilde{\delta}$ and $\delta(c') = 0$ across the respective cities, workers have the same skills $\widetilde{\omega}$ at those respective locations and pay the same rents. After that there are people of all lower skills in both cities. For a person of skill $\omega < \widetilde{\omega}$ found in both cities, they must be equally well off, pay the same rents, and face the same $A(c)D(\delta)$. Thus, an equally skilled person in the bigger city will have a higher δ and a lower $D(\delta)$.

Given the assumption noted earlier of "relative scarcity" of better locations in smaller cities, Davis and Dingel (2013) are able to show that the bigger city not only houses an entire segment of the highest-skilled people but for skill groups in common it has relatively more higher-skilled people than lower-skilled people. Under that condition, employing results in Costinot (2009), they show that higher-skilled people are more likely to work in higher-quality sectors and that such sectors are disproportionately found

in the bigger city. These results become the basis for the empirical work reviewed earlier in this chapter on the greater intensity of higher-skilled types in larger cities. In summary, the model neatly shows the intertwining of skill and scale externalities in defining productivity differences between bigger and smaller cities, even allowing for skill differentials to contribute to scale externalities. And interestingly, it links sorting across cities to within-city sorting.

Eeckhout et al. (2014) present a related model which focuses just on sorting across cities. However, in their work, sorting need not be monotonic with ever-increasing skills in larger and larger cities. In their model, there are two cities with exogenous city-specific productivity differences and three skill types. Their innovation is to explore the effect of differing degrees of complementarity between skill types. They suggest that there is extreme skill complementarity, where the larger (higher inherent productivity) city has a greater representation of both the highest-skilled and lowest-skilled types. Intuitively, high-skilled workers consuming low-skill services enjoy benefits from mutual location; or office workers in Manhattan enjoy food carts.

Sorting and selection. In another ambitious article, Behrens et al. (2014) consider both sorting of skill types across cities and selection within cities as to which people become entrepreneurs versus stay as workers. Selection raises the issue of tougher competition in bigger cities and whether firms in bigger cities are inherently more productive. The output setup is as in the benchmark model with local intermediate inputs going into production of one final numéraire good sold competitively. Intermediate inputs are produced by an entrepreneur using labor, l, with output increasing with the productivity of the entrepreneur. Individuals are given a draw of talent, t, which they know upon the draw. Then when they choose (irreversibly) the city they will live in, they get a draw of luck, s. The distribution they draw s from is the same for each city. A person's potential productivity as an entrepreneur is $\psi = st$, with firm output being ψl. Intermediate goods are sold under monopolistic competition to final producers, so intermediate entrepreneurs earn profits, which is their compensation and which increase with their ψ.

A person decides whether to be a worker or an entrepreneur on the basis of ψ, where within a city there is a critical value of ψ. People with values below the critical value choose to be workers (with the same wage and productivity—workers get no benefits from higher talent or luck), and those with a value above the critical value are entrepreneurs. The selection issue is whether bigger cities then typically have better firms. What about sorting across cities? It turns out that the sorting and selection issues are interlinked. Although the details in the article are involved, the basic idea is that there are the scale effects as in the benchmark and the sorting benefits as in Davis and Dingel (2013). Bigger cities are more productive because they can house more varieties of intermediate inputs and these scale effects interact with talent. In a version of log-supermodularity, the

expected return to talent rises with city scale and cities with more talented people have higher marginal benefits of agglomeration for a given size to be traded off against urban diseconomies. While there is sorting, in equilibrium selection is unimportant: each city has the same proportion of workers to entrepreneurs. So although bigger cities have more talented firms, the relative cut is the same in each city regardless of the size.

In Behrens et al., free migration can lead to two types of equilibrium. There can be a symmetric equilibrium where all cities are identical if the variation in talent across people is limited. But the interesting equilibria are asymmetric ones. Behrens et al. do not have city developers, but rely on properties of their special case for an equilibrium to define city sizes that emerge through self-organization. In their special case, cities are talent homogeneous, rather than the usual case of each city having a segment of the talent distribution. Each city has only one talent t and city sizes increase as the talent level rises. Cities under self-organization are too large but not catastrophically so. Intuitively, if cities get too big, people with the intended talent will want to move to less talented cities. The uniqueness and existence of such equilibria are not guaranteed.

In summary, Behrens et al. (2014), similar to Davis and Dingel (2013), have talent or skill sorting across cities, with bigger cities being in part more productive simply because they have more talented workers. But they also deal with selection and the choice of people to become entrepreneurs, where the relative proportion of workers to firms is the same across cities of different sizes. In evaluating the sources of higher productivity in bigger cities, the modeling suggests that sorting by talent is a key source of higher output per worker, while selection is not. This finding parallels the empirical results in Baum-Snow and Pavin (2012).

Cross-fertilization in externalities and sorting of industries. The benchmark model assumes that industries are characterized by simple specifications of externalities, which potentially lead to urban specialization. In an innovative article, Helsley and Strange (2014) explore cross-fertilization among industries in a generalized fashion. They explore equilibria and optima where cities may be of mixed employment or nonspecialized, and there can be a hierarchy of the nonspecialized. The article focuses on how under self-organization specialized cities are not just oversized but may also have poor compositions of output. The article is summarized in the chapter by Behrens and Robert-Nicoud (2015) in this handbook. Some key results are as follows. Pareto efficient size cities where different city types have a plurality of own type population but an optimal mix of other worker types can never be stable equilibrium city compositions and sizes. In an equilibrium, own type workers will want to agglomerate further with own type workers in some types of cities, compared with the optimum. If it is efficient for all cities to specialize when complementarities are weak, we have the usual problem of equilibrium oversized specialized cities under self-organization. However, we can also have an equilibrium with specialized cities when the optimum would dictate mixing. And when we do have

mixing, compositions and sizes are inefficient. For policy makers, the issue of how to enact optimal or second-best policies and institutions is not covered in the article, but the challenge is there.[9]

22.4.2 Dynamics in the urban hierarchy

So far the models we have presented are basically static, or, at best, stationary-state models. To examine the rapid changes in urban systems which occur with economic development, we need models of the dynamics of the urban hierarchy. Such models are limited in number, and many were reviewed in earlier handbook chapters (Gabaix and Ioannides, 2004; Henderson, 2005). Given that, we take a different expositional approach in the first part of this subsection. We start by interweaving a discussion of empirical relationships about the dynamics of the urban hierarchy with a brief review of some key models which sought to explain certain patterns in the data. These have to do with growth of city sizes and numbers over time and the stability of the size distribution of cities over time. We then turn to a more recent focus on why production patterns of individual cities and cities across the urban hierarchy change with time. There we return to the format of presenting empirical relationships and then models that relate to them.

22.4.2.1 Facts and concepts concerning the size distribution of cities

22.4.2.1.1 Growth in city sizes

In general, over recent centuries cities have been growing in population size. To see this for a more recent time period for which there are good data, Henderson and Wang (2007) take all metropolitan areas in the world from 1960 to 2000, divide them by mean size in any time period to get a relative size distribution, and define a relative cutoff point so that the minimum size to mean size is stable over time. They call this the relative size distribution. Within this distribution, the absolute mean (and median) size doubled from 1960 to 2000 worldwide. There are two explanations for why city sizes have increased with time, given size is determined by a trade-off between scale economies and diseconomies.

The first is that scale economies have been increasing relative to diseconomies. Black and Henderson (1999a) present an endogenous growth model where economic growth is

[9] We note two comments on the model. The first is based on the specification of cross-industry externalities. There are I types of workers where any type is employed in the corresponding industry of that type. Output per worker in sector i in city j is a function of the vector of the number of workers in different sectors in the city, $\mathbf{n_j}$, such that output per worker in sector i in city j is $g_i(\mathbf{n_j})$. The comment is that the study authors assume not just that there are localization and urbanization economies where $\partial g_i/\partial n_{ij} > 0$ and $\partial g_i/\partial n_{kj} > 0$, but there are complementarities where $\partial^2 g_i/\partial n_{kj}\partial n_{ij} > 0$, which is a special assumption without empirical validation. Second, they only fully solve for self-organization equilibria. How equilibria would look with optimizing developers who can cross-subsidize worker types within different types of cities is less clear, as is what institutions or policies would be required to achieve optima.

fuelled by human capital (knowledge) accumulation. In that model, human capital externalities interact with scale economies at the local level to enhance overall economies of agglomeration. They correlate differential growth rates of US cities with differential growth rates of local human capital. Rossi-Hansberg and Wright (2007) performed similar modeling on growth of city sizes. Recent work by Desmet and Rossi-Hansberg (2009) has a more nuanced approach, involving the endogenous evolution of scale externalities as part of their work focusing on transition dynamics.

An alternative to increasing scale externalities as the explanation for increasing city sizes is that diseconomies have dissipated with technological progress. The Alonso–Muth model emphasizes the decline in commuting costs as a driver of city spread. But it could also be the basis of increasing city sizes, with an eye to the technological revolutions of the last 120 years—the development of transit systems, the invention of the automobile, and the construction of multilane high-speed highway systems with rays and rings for cities. Empirical work by Duranton and Turner (2012) supports this idea. Finally, if we think outside the traditional models, growth in human capital per person may be associated with better technologies and management techniques in planning of cities and in managing urban diseconomies.

Although we have two reasons for city sizes to be increasing, that does not mean there will be necessarily a shrinking number of cities. In Black and Henderson (1999a) and Rossi-Hansberg and Wright (2007), city numbers may also increase with national population growth, as long as the rate of national population growth exceeds the growth rate of individual city sizes. These articles assume a fully urbanized world. Growth in city numbers in developing countries is also driven by urbanization, or the move out of agriculture as discussed in Section 22.2.1.

22.4.2.1.2 Stability of the relative city size distribution and size ranking of larger cities

City size distributions for countries are remarkably stable over time, and some argue that they are either globally (Gabaix, 1999) or locally (Eeckhout, 2004; Duranton, 2007) approximated by a Pareto distribution and thus obey Zipf's law. Henderson and Wang (2007) illustrate this stability for the world size distribution from 1960 to 2000. Black and Henderson (2003) and Harris-Dobkins and Ioannides (2001) show this similarly for the United States over many decades. To be clear, these exercises look at just cities, not at the spatial transformation of the universe of space as described in Section 22.3. Theoretical modeling pioneered by Gabaix (1999) and developed more fully by Rossi-Hansberg and Wright (2007) and Duranton (2007) argues that stochastic processes in particular contexts such as ones that obey Gibrat's law generate a stable size distribution of cities over time approximated by Zipf's law. A potential problem is that these models also have all cities transiting continuously through the size distribution of cities, in partial contrast to the next fact.

Evidence suggests that the biggest cities historically tend to remain the relatively biggest cities in a country over long periods of time. There is little move downward from the

top rung of cities in a country (Eaton and Eckstein, 1997; Black and Henderson, 1999b, 2003). Eaton and Eckstein (1997) show that the ranking of cities by size has been remarkably stable in France and Japan over the prior 100 years or more. In a Markov process based on 10 decades of data, Black and Henderson (1999b, 2003) show that mean first passage times for a US city in the top 5% of population size to transit to the bottom 35% is many centuries (which is a time horizon way out of sample). The question is why are big cities so slow to move down the size ranking? Glaeser and Gyourko (2005) and Henderson and Venables (2009) claim that city durable capital is an explanation for why big cities retain populations in the face of bad shocks and competition. Arthur (1990) and Rauch (1993) stress information externalities are embedded in place, where bigger cities have a large accumulated stock of knowledge that is not readily transferable.

22.4.2.2 Churning and movement of industries across the urban hierarchy

In this part, we start with some facts about the movement of industries across cities, which is also related to the movement within cities (from the core to the periphery). We then turn to a discussion of two recent relevant models.

22.4.2.2.1 Facts about industry movement

Churning is the process whereby cities over time lose their existing export industry or industries, to be replaced by different export industries. So an automobile city of today may become an electronics city in the next decade. Churning can be defined directly with a churning index based on the work of Davis and Haltiwanger (1998) and used in Duranton (2007), or can be based on mean first passage times in a Markov process (Black and Henderson, 1999b, 2003; Duranton, 2007). The mean first passage time for a top city industry to transit from the top 5% to the bottom cell of five cells is a small fraction of the mean first passage time for a city in the top 5% of the population to transit to the bottom cell.

While Duranton observes generalized churning in US and French data, there are other sets of empirical findings, more specific to the economic development process. A first set of findings concerns the degree of specialization of cities in the urban hierarchy. As noted in Table 22.2, the degree of specialization of cities in the United States has declined over the last 30 years. In contrast, South Korea, at a different stage of development, showed increasing specialization of cities from 1983 to 1993 for most industries, while diversity increased at the more aggregate regional level. So regional economies diversified but industry concentration at the city level increased (Henderson et al., 2001). Another example is China, where average specialization at all spatial scales increased from 1995 to 2008. Table 22.3 shows that specialization for both urban counties and rural counties increased from 1995 to 2008, also at the larger spatial scales of the prefecture and metropolitan area (city proper). Note that Chinese cities in general as a group are less specialized than the individual urban districts making up those cities, consistent with an idea that there is neighborhood clustering of like activities within cities.

The second set of empirical findings concerns the phenomenon in developing countries of industrial decentralization from the cores of the largest metropolitan areas. At the early stages of national economic development, modern manufacturing in a country may be largely confined to the core city of the largest metropolitan area(s) for reasons discussed below. This concentration is followed by two stages of decentralization: first out of the core to peripheries of metropolitan areas and then from metropolitan areas to hinterlands.

The idea is illustrated using data for Korea and China. For Korea, Table 22.4 looks at the evolution of manufacturing shares within the national capital region of Seoul, Kyonggi province. While Seoul metropolitan area has retained a fairly constant share of the population in its local region, its share of manufacturing employment declined dramatically during the 1970s and the 1980s, starting at 76% in 1970 and declining to 30% by 1993. This is movement of industry out of Seoul to nearby satellite cities and ex-urban areas. Table 22.4 also compares the evolution in just the 10 years from 1983 to 1993 of shares of national manufacturing employment held by the three main metropolitan areas in Korea, their satellite cities, and then the rest of the country. This is a second stage of decentralization where the three core metropolitan areas continue to lose share. The losses are no longer to the satellite cities, but are beyond, to the hinterlands. The

Table 22.3 Changing specialization in China (three-digit industry breakdown)

	1995 Krugman Gini coefficient (manufacturing)		2008 Krugman Gini coefficient (manufacturing)	
	Mean	**Median**	**Mean**	**Median**
Prefecture	0.4033	0.3978	0.4694	0.4741
City proper (urban districts, 2010)	0.3059	0.2863	0.3525	0.3460
County (rural units, 2010)	0.4218	0.4185	0.4612	0.4574
County (urban units, 2010)	0.4359	0.4294	0.4825	0.4749

Source: Authors' own calculations, based on about 150 three-digit industries in each year which show positive employment.

Table 22.4 Stages of decentralization in Korea

Share of Seoul in Kyonggi province (National Capital Region)				
	1970	1980	1983	1993
Population	62%	63%	67%	61%
Manufacturing Employment	76%	61%	45%	30%
Share of national manufacturing employment				
	1983		1993	
Seoul, Pusan, and Taegu metro areas	44%		28%	
Satellite cities of Seoul, Pusan, and Taegu metro areas	30%		30%	
Other cities, rural areas	26%		42%	

Source: Henderson et al. (2001) and related calculations.

hinterlands' share rose from 26% to 42% in 10 years, at a time when their population share declined modestly. This shift to the hinterlands is correlated with the extensive investment in highways and telecommunications Korea undertook in the early 1980s to service hinterland areas. The overall dispersion of manufacturing is also consistent with manufacturing becoming a mature industry, as discussed in Section 22.3.

For China, Table 22.5 shows the decline in the shares of areas defined as core urban counties of metropolitan areas in 1990 in national manufacturing employment from 1995 to 2008. New urban counties are on the periphery of these 1990 urban cores, or are the new suburbs. Their employment shares more than double. But hinterland towns labeled as county towns also see a modest rise in their shares. Note the high concentration of services in the original urban cores in 2008 (we do not know the 1995 numbers for services), far in excess of their shares of either population or manufacturing employment.

Desmet et al. (2015) show a corresponding trend in India, looking at the growth of manufacturing versus service employment in districts where they are initially concentrated versus in districts where they are not. The time period is short, 2000–2005, but still the patterns are striking. As illustrated in Figure 22.9, they fit locally a trend with error bands, and, as the trend moves to higher-density districts with fewer observations, the error bands widen. For manufacturing, there is strong mean reversion whereby districts with high densities in 2000 grow much more slowly than districts with low densities in 2000. The pattern for services is quite different. High-density districts on average have higher growth rates than at least the middle-density districts. At the upper end, growth rises with density. Overall, this suggests decentralization of manufacturing as in the Korean and Chinese cases, while services are concentrating even more in the high-concentration districts found in the biggest cities.

22.4.2.2.2 Modeling industry movement across cities

The facts presented in the previous section concern churning and the general turnover of industries in cities, changes in the degree of specialization in the urban hierarchy, and

Table 22.5 Stages of decentralization in China

Share of "nation"	Industry		Services	Population
	Share of national industry employment		Share of national services employment	Share of national population
	1995	2008	2008	2000
Urban counties in both 1990 and 2000	52%	41%	58%	28%
New urban counties	6.2%	13%	6.3%	5.5%
County towns	18%	22%	12%	18%
Other rural counties in 2010	24%	24%	24%	48%

Source: Authors' own calculations.

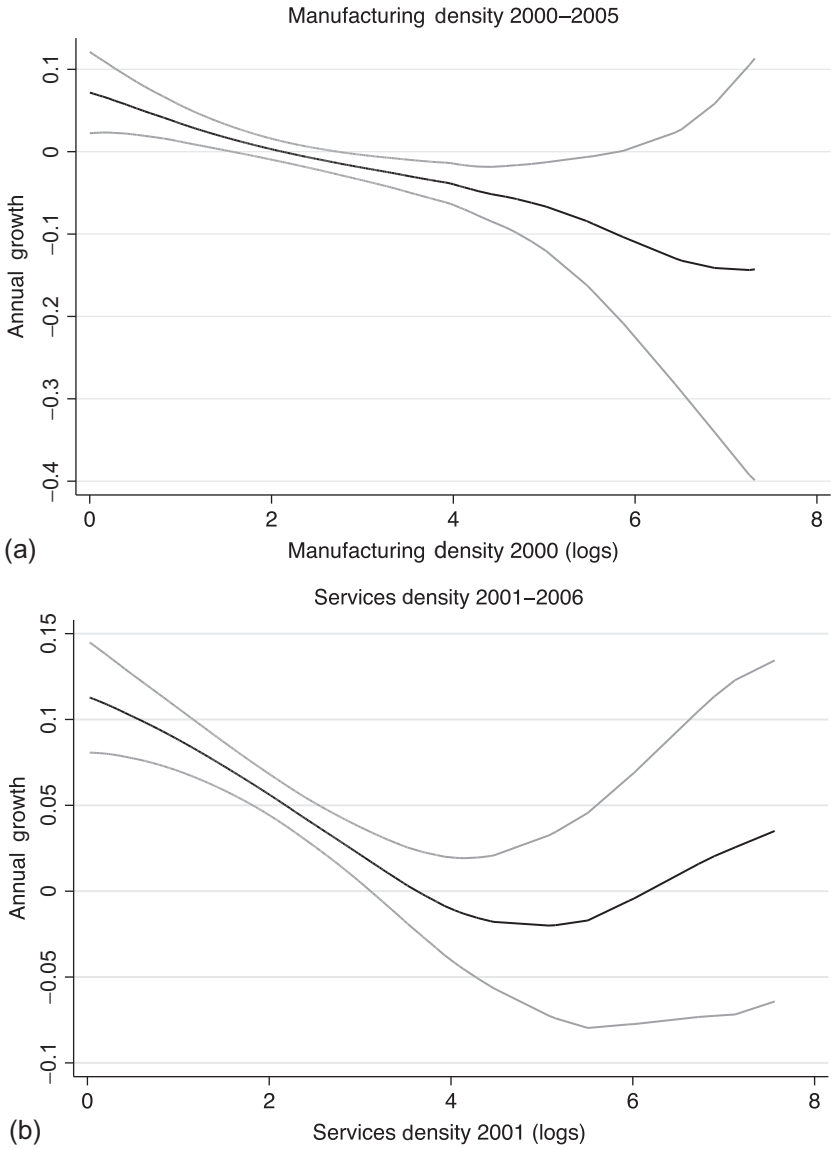


Figure 22.9 Decentralization of manufacturing and centralization of services in India: (a) manufacturing and (b) services. *Source: (a, b)* Desmet et al. (2015).

patterns of industry movements across the urban hierarchy. We outline a model that deals with churning *per se* and then review a variety of relevant models that deal with industry movements across the hierarchy.

Churning. Duranton (2007) adapts the Grossman–Helpman quality ladder model to an urban setting, with the aim of presenting a model of the following facts. Cities are slow to

change their relative sizes; the overall size distribution of cities is remarkably stable; and industries move quickly across cities, with cities changing production patterns. In the Grossman–Helpman quality ladder model, there are a fixed set of consumer goods, but they can be produced with ever-increasing quality. Quality is a ladder process where there is one current best quality $\bar{j}(z,t)$ for sector z at time t. Research by research firms is ongoing to improve that quality. Eventually that research leads in sector z to an advance discovered by one research firm. That firm then gets the (nontransferable) patent to produce that product and focuses on production activities, enjoying monopoly rents in production in industry z until there is the next move up the ladder. Only the research firm that discovers the latest quality level produces the product, pricing so as to exclude potential lower-quality producers.

Duranton adds an urban component. He assumes a fixed given number of cities, each specialized in the production of a different first-nature good, assumptions which anchor cities so that none can disappear, or become unpopulated. The action lies in second-nature goods which involve innovation and are completely footloose. Duranton makes two key assumptions. Production must occur in the place where a winning research firm makes a discovery. Production requires information from the research firm which can be transmitted only locally, such as through hiring the former research workers of the winning firm to be involved in production. Second, in order to be productive, all research firms focused on innovating in z must be located in the current city where z is produced. If all innovations, as in Grossman–Helpman quality ladder model, are within the own industry, then production would never move. Duranton introduces cross-industry innovation. The cumulated expenditure $\lambda^k(z)$ by research firm k focused on innovation z has a probability $\beta\lambda^k(z)$ of inducing a winning move up the quality ladder in industry z, but also a probability $\gamma\lambda^k(z)$, $\gamma < \beta$, of inducing a winning innovation in industry z'. The probability of an innovation in industry z is $\beta\lambda(z) + \gamma\sum_{z\neq z'}\lambda(z')$, where $\lambda(z)$ is the cumulated expenditures of all research firms focused on innovation in z. If a research firm working on z happens to make a winning innovation in z', the production of z' moves to the city where this firm is located, generating churning.

In the steady state, there are several key results. First, there is industry churning: the location of production for second-nature products with footloose production will change over time driven by cross-industry innovation.[10] Second, the innovation process leads to a stable size distribution of cities that locally approximates Zipf's law, so the size distribution of cities remains time invariant. Third, however, there is motion for individual cities. Bigger cities which (by accident) have accumulated innovations and production will lose and gain sectors over time, but the net result will be mean reversion, with

[10] An older heuristic version of this was that traditional producers and their nearby research firms become "complacent" and the innovation occurs in new locations.

the biggest cities growing more slowly (or losing employment) relative to smaller cities. So there is a transition process where small cities move up and bigger cities move down (slowly).

Explaining patterns of movement across the urban hierarchy as economic development proceeds. We have two sets of shifts across the urban hierarchy. The first is the changing degree of specialization of cities and regions. The second is the movement of industry out of the core of the largest metropolitan area(s) to peripheries and then to hinterlands. How might we explain these shifts? In Section 22.4.3, we will argue that public policy, transport investments, and innovations may play a role in explaining both of them. Here we focus on modeling that involves changes in production technology. In Desmet and Rossi-Hansberg (2009), as reviewed in Section 22.3, two waves of GPT—electricity and IT—induced first the concentration of initially more dispersed manufacturing into high-density locations in the United States, and then several decades later the deconcentration of manufacturing from the most densely populated areas, to be replaced by services. Both also involve changing specialization at a more aggregate level. Correspondingly, we might think of developing countries experiencing technology transfers and adaptation. Learning with adaptation of foreign technologies is initially efficiently concentrated in the densest locations; but, as manufacturing technologies are adapted and standardized, scale externalities may diminish and disperse. Manufacturing moves out of the most densely populated locations, to be replaced by services.

The ideas in Desmet and Rossi-Hansberg (2009) also relate to within-metropolitan-area deconcentration historically in the United States and in developing countries today. In the United States, there was the shift of manufacturing with electrification to continuous-process production in the early twentieth century, where continuous-process production requires single-story buildings and hence a lot of land. Land being much cheaper at the city fringe than in the center provides an incentive for manufacturing to relocate out of core cities as it did in the early and middle twentieth century in the United States. Related to this, if the service sector within a city starts to enjoy greater marginal local agglomeration benefits than manufacturing, that makes the service sector better able to outcompete manufacturing for high-priced land in city centers, a point further developed in Desmet and Rossi-Hansberg (2014a).[11]

22.4.3 Policies affecting the spatial allocation of resources

Government policies and institutions strongly influence the structure of the urban hierarchy. There are a whole range of policies, such as those governing trade, minimum

[11] Related to this, in the Fujita and Ogawa (1982) model, a decline in manufacturing externalities (the value of information spillovers within the city as technology standardizes) leads to the formation of more urban centers away from the core, fostering the development of subcenters to which workers can commute more cheaply.

wages, capital markets, and fiscal decentralization, which in older work (Renaud, 1981; Henderson, 1988) as well as more recent work are recognized as affecting the allocation of resources across the urban hierarchy. For example, policies which affect the national composition of products then affect the sizes and numbers of cities producing products favored by trade policies. As such, these policies will differentially affect cities through the urban hierarchy. So if trade policies favor steel at the expense of textiles, the national composition of cities will change so that the relative number of cities engaged in steel production or inter-related products will increase. These may be bigger types of cities than those engaged in textile and related production such as apparel production. Minimum-wage policies which fix nominal wages may bite only in big cities with higher nominal wages but not higher real wages.

It is beyond the scope of this chapter to review all these policies. While many have been covered in older research, on some there has been a lot of policy work but little recent hard-core research. Of particular concern is financing by local governments and the institutions that allow cities to tax for current expenditures and borrow for capital projects such as infrastructure investments. As an example, in developing countries with weak institutions, metropolitan governments generally are not able to finance capital projects by borrowing either on bond markets or from international banks (given public infrastructure cannot be used as collateral). Borrowing is essential to efficient allocations given both limited current tax capacity and the fact that the benefits extend far into the future, so ideally financing is spread over time (Henderson and Venables, 2009). National governments can offer financing or guarantee loans, but then there is a problem of default by local governments on any loans granted to them. Of course, the national government can use grants to selectively finance local projects, but selection may be based on political considerations and less on local economic conditions. And national governments may be restricted in their revenue sources and ability to borrow as well. In short, it may be that many cities cannot access sufficient money and have deficient infrastructure investments (and some targeted cities may have excessive investments). We know of no hard core research on what the impact of underfunding (or overfunding) is on urban quality of life and growth of city populations or productivity. What are the productivity losses for a city such as Dar es Salaam with horrendous congestion, with little public transport, and with poor underfunded road networks? What will be the impact of the development of bus rapid transit now being constructed? We simply do not have findings from research which deals with such questions at a city or national scale.

In this section, we focus on two types of policies for which there is recent research and are fundamentally spatial in nature. The first policy concerns the causal effects of transport infrastructure investments linking cities and regions, as well as locations within cities, on urban form and city growth. The second policy concerns urban, or what we will label as big-city, bias in the allocation of public resources and the operation of markets.

22.4.3.1 Transport investments and technological change

Modeling suggests that transport infrastructure investments are responsible for changing patterns of specialization and growth of towns and regions observed in the data. This is the subject of a chapter in this handbook by Redding and Turner (2015), and our coverage is brief. An old debate concerns the effect on hinterland towns of improved linkages to the national centers of economic activity: linkages offer better access to markets but remove protection from outside competition for local producers. The work by Donaldson (2014) on historical India, which is based on the model of Eaton and Kortum (2002), suggests that transport investments lowering costs of trade between locations benefit essentially all cities or regions by allowing them to specialize in the production of goods for which they have more of a comparative advantage and to shed production of others and import them as transport costs fall.[12] For our purposes, the key is increased specialization on a wide-scale basis, consistent with the data on China and Korea we reviewed above. In the new economic geography models pioneered by Krugman (1991), transport improvements on a cruder scale lead first to centralization and specialization of the "core" region in manufacturing, consistent with the above analysis. But further improvements (at a later stage of development) can lead to decentralization of manufacturing to periphery regions if core regions become congested (Puga, 1999), as suggested by recent US data. In the simple new economic geography models, specialization and concentration are intertwined.

Transport investments also have strong effects on within-city decentralization of industry. In the United States, historically, goods moved across cities by rail, being shipped from terminals in or near city centers to other cities. Transport within the city to rail terminals by, for example, horse-drawn wagons was very expensive, so firms tended to cluster around the rail terminal in the city center. With the development of trucking and then the highway system, Meyer et al. (1965) argue that the construction of ring roads in cities in the 1950s and 1960s permitted various types of manufacturing to decentralize from urban cores to suburban areas with cheaper land and then to ship goods to rail sidings and suburban terminals by ring roads. For China, in a corresponding phase during the 1995–2008 period, Baum-Snow et al. (2013) show that rail and ring roads causally led to decentralization of manufacturing within Chinese cities.

22.4.3.2 Urban and political city bias

There is a development literature based on the two-sector model (e.g., see Ray, 1998 for a synopsis) which talks about biases and/or policy distortions in labor or capital markets nationally that favor the urban sector and may draw in excessive numbers of migrants to

[12] Empirical work on China is less conclusive. Faber (2014) and Banerjee et al. (2012) reach opposite conclusions on the effect of transport improvements on the fortunes of hinterland areas that are "treated" with transport investments versus those that are not.

cities. Alternatively, there may be migration restrictions such as China's *hukou* system that restrain the extent of rural–urban migration. Here we turn to a related version of biases, where one city, or more generally, larger and politically connected cities are favored relative to other cities and the rural sector. As we will see, favoring a certain city may make that city either larger or smaller than it otherwise would be.

As reviewed by Henderson (1988) and Duranton (2008), the standard modeling of the effects of political bias assumes that favored cities are larger than they would be in the absence of favoritism. There is a system of cities in an economy of different types and equilibrium sizes. Under a developer regime, cities tend to operate near the peak of their inverted-U-shaped real income curves, at sizes where real incomes across different types of cities for a representative worker are equalized. In national labor markets, any one city faces a horizontal labor supply curve at that going real income. A city that is favored in capital markets or with special public services has an inverted-U shaped real income curve is shifted up—it can pay higher real income/utility at any size. If the city is subsidized for capital costs, that raises the marginal product of labor a competitive firm can pay. With unrestricted immigration, the size of the favored city expands beyond the peak of its inverted U. The equilibrium is the point down the right-hand side from the peak where city real income equals the going real income in national labor markets in other nonfavored cities. The implication is that, in a free-migration equilibrium, the benefits of favoritism are dissipated through increased commuting costs, or increased city disamenities more generally, as the city size expands beyond the peak potential real income point.

Empirically Ades and Glaeser (1995) and Davis and Henderson (2003) examine indirect evidence. As hinted at by the title of the article by Ades and Glaeser (trade and circuses), there seems to be a clear bias toward capital cities in many countries, especially before democratization. Relative to their economic position, they are much larger than other cities, indicating a bias toward investing in capital cities.

As a more specification example, for China there is indirect evidence given in Au and Henderson (2006b), who infer differential rates of return to the urban sector versus the rural sector and for different types of firms within the urban sector. Direct evidence is harder to find. While articles generate fiscal numbers showing higher *per capita* public expenditures in different classes of cities, it is hard to distinguish if that is bias, or if it is simply that it is efficient for public sectors to be larger in bigger cities, with their greater congestion and environmental issues. Capital markets where we expect an efficient allocation equalizes rates of return across cities can provide readier documentation. Jefferson and Singh (1999) estimate higher rates of return to rural-based firms compared with urban-based firms in the early 1990s in China. Cai and Henderson (2013) show that the rates of return to capital in China differ not only by firm type (lower for state-owned firms) but also by city type. All types of firms in political cities such as Beijing on average are favored (earn lower rates of return) than firms in ordinary-prefecture-level cities in

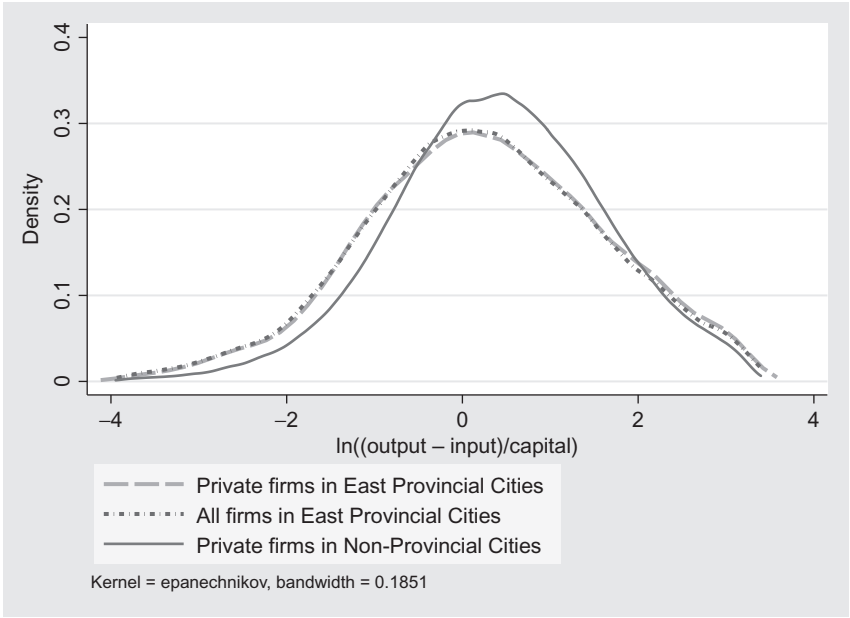


Figure 22.10 2007 Distribution of after-tax value added divided by net asset value (as proportional to the returns on capital).

China. Figure 22.10 shows the distribution of returns for private sector firms in ordinary prefecture cities versus the three main provincial-level and heavily politically favored cities in eastern China, Beijing, Shanghai, and Tianjin.

Favoritism raises another critical issue. From the discussion of the inverted-U curve of real income against city size, it is apparent that cities would want to resist inward migration beyond the peak. If cities could price discriminate and city populations could be fixed, either "original" incumbent residents or a developer could restrict inward migration to the favored city and charge fees to marginal migrants (Henderson, 1988; Behrens and Robert-Nicoud, 2015). According to the specific framework, city size is set to some real income maximizing size (for either the developer in a developer-controlled city or incumbent residents who control a city) between the peak and the free migration equilibrium. However, entry fees and price discrimination are not the direct institutions for cities in countries. Rather, it is through land markets and regulation that residents or city governments attempt to restrict city size.

In developed countries the tool to restrict size is exclusionary zoning. In the super-star cities article by Gyourko et al. (2013), favored cities, in their case cities favored with natural amenities, attempt to restrict inward migration through exclusionary zoning. Such zoning can effectively fix the number of dwelling units permitted in a locality. With that restriction, the key to entry is getting one of the fixed number of lots in the city. Lot prices rise so as to lower utility from entering the city to the outside option for the marginal

entrant. In their model, higher-skilled, higher-income people have a greater willingness to pay for the amenities of super-star cities. Thus, as the national population and real incomes rise, super-star cities both have higher price increases and a shift in the population composition toward higher-income people, who outbid others for the amenities of these cities. In these frameworks, the key assumption is that all dwelling units are provided in a formal sector governed by zoning laws.

In developing countries, the restrictions are different. Until recently in China, there have been explicit migration restrictions, directly limiting mobility. Desmet and Rossi-Hansberg (2013) find that the dispersion of amenities is greater across Chinese cities than across US cities, and show how this can be interpreted as evidence of migratory restrictions to some of the country's favored cities. In the absence of such restrictions, they find that some of China's largest cities would become even larger, and that overall welfare would increase significantly. Their article is also an example of how quantitative models in urban economics can be used to estimate the welfare impact of different policies.

Most countries do not have direct migration restrictions, and in China these are now disappearing. Restrictions take a different form. They involve land markets and the public sector, but not zoning, which restricts entry directly. In developing countries, there are informal housing sectors, which violate whatever regulations potentially govern the formal sector. As discussed in the chapter in this handbook by Brueckner and Lall (2015), governments in developing countries either do not have the power or political will to stop the development of informal sectors or permit them to develop in a second-best framework. Informal sectors may involve "squatting" (Jimenez, 1984), which means collective illegal seizure of land or illegal or quasi-legal development of land that is legally owned. One example of the latter is *loteamentos* in Brazil, which are developments in violation of national zoning laws but built on legally held land. Another example is the development of urban villages in cities in China. Urban villages are on land within the city that is still owned by a rural collective. Typically these were the former living areas of farm villages, where the city annexed the farm land but not the living area. These living areas are then intensively developed into high-density "slum" housing for migrants. This escape valve would then allow a free-migration equilibrium to emerge, but with one catch.

The catch involves the provision of local public services and becomes the basis for restricting inward migration. As Cai (2006) discusses for China, urban villages do not receive services from the city (central water or sewerage, garbage collection), and their children are generally excluded from state schools. This forces high-cost and/or low-quality provision of such services for migrants in these settlements, making migration much more costly for them. As nicely illustrated in Duranton (2008), effectively incumbent residents face one inverted U, while at the margin inward migrants face a different one that is shifted down. This reduces the population at which that national supply curve

of the population to the city intersects the effective real income curve for migrants. In China, this policy has been called "lifting the door sill" (Cai, 2006). For Brazil, Feler and Henderson (2011) attempt to estimate the causal effects on population growth especially of low-skilled people of denial of centralized water provision to likely migrant housing areas in Brazil. In Brazil, localities were not required to service areas which were not in the formal sector in the 1980s. This analysis suggests that the emergence of slum areas in cities in developing countries in some contexts reflects in part a strategic decision of localities to try to restrict inward migration, especially into favored larger cities such as national capitals (e.g., Beijing) or the seats of political elites (e.g., Shanghai and São Paulo).

22.5. CONCLUDING REMARKS

In this chapter, we have described recent theory and evidence of how the spatial distribution of economic activity changes as a country grows and develops. In doing so, we focused on different geographic units, starting with the coarse urban–rural distinction, then going to the entire distribution, and finishing with its upper tail—the cities. When addressing the question of how an economy's spatial organization changes with development, the literature has often analyzed the long-term patterns of today's developed countries, notably the United States. Undoubtedly, the past spatial development of the United States holds valuable lessons for today's developing countries, so this strategy is often both useful and appropriate. At the same time, today's world is different from the one faced by the United States and other developed countries in the nineteenth and twentieth centuries. For example, the increasing impact of trade may imply that some countries can urbanize without industrializing. Traditionally the paucity of geographically disaggregated data has limited the extent of empirical analysis on developing countries. However, the rapidly increasing availability of data, together with geographic information system tools, is changing this. In fact, as this chapter has made clear, the last decade has seen a growing number of empirical studies using data from developing countries. We believe that there is a need for more such studies in order to elicit the stylized facts which should form the basis for further theoretical work on the link between geography and development. More work is also needed in modeling and understanding the relationship between space and development. The spatial distribution of economic activity affects growth, and vice versa. An economy's degree of urbanization is not only a consequence of its development, it is also a determinant of its growth. To understand better these links, there is a need for more micro studies and for more quantitative work with an emphasis on counterfactual policy experiments. It is clear that a country's spatial organization is not independent of its macroeconomic performance. Regional and urban economists should therefore continue their efforts to develop the tools needed to inform policy makers of how regional and spatial policies affect welfare and growth.

REFERENCES

Aarland, K.J., Davis, J.C., Henderson, J.V., Ono, Y., 2007. Spatial organization of firms. Rand J. Econ. 38, 480–494.

Abdel-Rahman, H., Anas, A., 2004. Theories of systems of cities. In: Henderson, J.V., Thisse, J.F. (Eds.), Handbook of Regional and Urban Economics, vol 4. Elsevier, Amsterdam, pp. 2293–2339.

Abdel-Rahman, H., Fujita, M., 1993. Specialization and diversification in a system of cities. J. Urban Econ. 33, 189–222.

Ades, A.F., Glaeser, E.L., 1995. Trade and circuses: explaining urban giants. Q. J. Econ. 110, 195–227.

Allen, R.C., 2004. Agriculture during the industrial revolution, 1700–1850. In: Floud, R., Johnson, P. (Eds.), The Cambridge Economic History of Modern Britain, vol. 1. Cambridge University Press, Cambridge, United Kingdom. Industrialisation 1700–1860 (Chapter 1).

Arthur, B., 1990. Silicon valley locational clusters: when do increasing returns to scale imply monopoly. Math. Soc. Sci. 19, 235–251.

Au, C.C., Henderson, J.V., 2006. Are Chinese cities too small? Rev. Econ. Stud. 73, 549–576.

Au, C.C., Henderson, J.V., 2006. How migration restrictions limit agglomeration and productivity in China. J. Econ. Dev. 80, 350–388.

Bairoch, P., Batou, J., Chèvre, P., 1988. La population des villes européennes de 800 à 1850. Centre d'Histoire Economique Internationale de l'Université de Genève, Librairie Droz.

Baldwin, R.E., Martin, P., 2004. Agglomeration and regional growth. In: Henderson, J.V., Thisse, J.F. (Eds.), Handbook of Regional and Urban Economics. first ed., vol. 4. Elsevier, Amsterdam, pp. 2671–2711 (Chapter 60).

Banerjee, A., Duflo, E., Qian, N., 2012. On the road: transportation infrastructure and economic development. NBER Working paper 17897.

Basu, S., Fernald, J., 2007. Information and communications technology as a general purpose technology: evidence from US industry data. Ger. Econ. Rev. 8 (2), 146–173.

Baum-Snow, N., Pavin, R., 2012. Understanding the city size wage gap. Rev. Econ. Stud. 79, 88–127.

Baum-Snow, N., Brandt, L., Henderson, J.V., Turner, M., Zhang, Q., 2013. Roads, railways and decentralization of Chinese cities. Brown University, processed.

Becker, R., Henderson, J.V., 2000. Intra-industry specialization and urban development. In: Huriot, J.M., Thisse, J. (Eds.), The Economics of Cities: Theoretical Perspectives. Cambridge University Press, Cambridge, UK, pp. 138–166.

Beeson, P.E., DeJong, D.N., 2002. Divergence. Contrib. Macroecon. 2 (1), Article 6, B.E. Press.

Behrens, K., Robert-Nicoud, F., 2015. Agglomeration theory with heterogeneous agents. In: Duranton, G., Henderson, J.V., Strange, W. (Eds.), Handbook of Regional and Urban Economics, vol. 5. Elsevier, Amsterdam.

Behrens, K., Duranton, G., Robert-Nicoud, F., 2014. Productive cities: sorting, selection, and agglomeration. J. Polit. Econ. 122, 507–553.

Black, D., Henderson, J.V., 1999. A theory of urban growth. J. Polit. Econ. 107 (2), 252–284.

Black, D., Henderson, J.V., 1999. Spatial evolution of population and industry in the USA. Am. Econ. Rev. Pap. Proc. 89 (2), 321–327.

Black, D., Henderson, J.V., 2003. Urban evolution in the USA. J. Econ. Geogr. 3, 343–372.

Boucekkine, R., Camacho, C., Zou, B., 2009. Bridging the gap between growth theory and the new economic geography: the spatial Ramsey model. Macroecon. Dyn. 13, 20–45.

Brock, W., Xepapadeas, A., 2008. Diffusion-induced instability and pattern formation in infinite horizon recursive optimal control. J. Econ. Dyn. Control. 32, 2745–2787.

Brock, W., Xepapadeas, A., 2010. Pattern formation, spatial externalities and regulation in coupled economic-ecological systems. J. Environ. Econ. Manag. 59, 149–164.

Brueckner, J., Lall, S., 2015. Cities in developing countries: fueled by rural-urban migration, lacking in tenure security, and short of affordable housing. In: Duranton, G., Henderson, J.V., Strange, W. (Eds.), Handbook of Regional and Urban Economics, vol. 5. Elsevier, Amsterdam.

Brülhart, M., Sbergami, F., 2009. Agglomeration and growth: cross-country evidence. J. Urban Econ. 65 (1), 48–63.

Cai, F., 2006. Floating populations: urbanization with Chinese characteristics. CASS mimeo.

Cai, W., Henderson, J.V., 2013. The Bias towards political cities and state owned firms in China's capital markets. LSE, processed.
Carlino, G., Kerr, W., 2015. Agglomeration and innovation. In: Duranton, G., Henderson, J.V., Strange, W. (Eds.), Handbook of Regional and Urban Economics, vol. 5. Elsevier, Amsterdam.
Caselli, F., Coleman II., W.J., 2001. The U.S. structural transformation and regional convergence: a reinterpretation. J. Polit. Econ. 109, 584–616.
Caselli, P., Paternò, F., 2001. ICT accumulation and productivity growth in the United States: an analysis based on industry data. Temi di Discussione 419, Banco d'Italia.
Chun, H., Kim, J.W., Lee, J., Morck, R., 2005. Information technology, creative destruction, and firm-specific volatility. Unpublished manuscript.
Costinot, A., 1999. An elementary theory of comparative advantage. Econometrica 77, 1165–1192.
David, P.A., Wright, G., 2003. General purpose technologies and surges in productivity: historical reflections on the future of the ICT revolution. In: David, P.A., Thomas, M. (Eds.), The Economic Future in Historical Perspective. Oxford University Press, Oxford, UK.
Davis, D., Dingel, J., 2013. The comparative advantage of cities. Columbia University, processed.
Davis, S., Haltiwanger, J., 1998. Measuring gross worker and job flows. In: Haltiwanger, J.C., Manser, M.E., Topele, R.H. (Eds.), Labor Statistics Measurement Issues. University of Chicago Press, Chicago.
Davis, J., Henderson, J.V., 2003. Evidence on the political economy of the urbanization process. J. Urban Econ. 53, 98–125.
Davis, J., Henderson, J.V., 2008. Agglomeration of headquarters. Reg. Sci. Urban Econ. 63, 431–450.
Desmet, K., Fafchamps, M., 2005. Changes in the spatial concentration of employment across U.S. counties: a sectoral analysis 1972–2000. J. Econ. Geogr. 5, 261–284.
Desmet, K., Fafchamps, M., 2006. Employment concentration across U.S. counties. Regional Sci. Urban Econ. 36, 482–509.
Desmet, K., Parente, S.L., 2012. The evolution of markets and the revolution of industry: a unified theory of growth. J. Econ. Growth 17, 205–234.
Desmet, K., Rappaport, J., 2013. The settlement of the United States, 1800–2000: the long transition to Gibrat's law. CEPR Discussion Paper #9353.
Desmet, K., Rossi-Hansberg, E., 2009. Spatial growth and industry age. J. Econ. Theory 144, 2477–2502.
Desmet, K., Rossi-Hansberg, E., 2010. On spatial dynamics. J. Reg. Sci. 50, 43–63.
Desmet, K., Rossi-Hansberg, E., 2012. Innovation in space. Am. Econ. Rev. Pap. Proc. 102, 447–452.
Desmet, K., Rossi-Hansberg, E., 2013. Urban accounting and welfare. Am. Econ. Rev. 103, 2296–2327.
Desmet, K., Rossi-Hansberg, E., 2014a. Spatial development. Am. Econ. Rev. 104, 1211–1243.
Desmet, K., Rossi-Hansberg, E., 2014b. On the spatial economic impact of global warming. Working paper.
Desmet, K., Ghani, E., O'Connell, S., Rossi-Hansberg, E., 2015. The spatial development of India. J. Reg. Sci. 55, 10–30.
Diamond, J., 1997. Guns, Germs, and Steel: The Fates of Human Societies. W.W. Norton, New York.
Doepke, M., 2004. Accounting for fertility decline during the transition to growth. J. Econ. Growth 9, 347–383.
Donaldson, D., 2014. Railroads of the raj: estimating the impact of transportation infrastructure. Am. Econ. Rev., forthcoming.
Duranton, G., 2007. Urban evolutions: the fast, the slow, and the still. Am. Econ. Rev. 97, 197–221.
Duranton, G., 2008. Viewpoint: from cities to productivity and growth in developing countries. Can. J. Econ. 41, 689–736.
Duranton, G., Overman, H.G., 2005. Testing for localization using micro-geographic data. Rev. Econ. Stud. 72, 1077–1106.
Duranton, G., Overman, H.G., 2008. Exploring the detailed location patterns of U.K. manufacturing industries using microgeographic data. J. Reg. Sci. 48, 213–243.
Duranton, G., Puga, D., 2001. Nursery cities. Am. Econ. Rev. 91, 1454–1477.
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.
Duranton, G., Puga, D., 2005. From sectoral to functional urban specialisation. J. Urban Econ. 57, 343–370.
Duranton, G., Puga, D., 2014. The growth of cities. In: Durlauf, S.N., Aghion, P. (Eds.), Handbook of Economic Growth, vol. 2. Elsevier, Amsterdam.

Duranton, G., Turner, M., 2012. Urban growth and transportation. Rev. Econ. Stud. 79, 1407–1440.
Easterlin, R.A., 1960. Interregional difference in per capita income, population, and total income, 1840–1950. In: Parker, W. (Ed.), Trends in the American Economy in the Nineteenth Century, Studies in Income and Wealth. Princeton University Press, vol. 24. Princeton, NJ, pp. 73–140.
Eaton, J., Eckstein, Z., 1997. Cities and growth: evidence from France and Japan. Reg. Sci. Urban Econ. 27, 443–474.
Eaton, J., Kortum, S., 2002. Technology, geography, and trade. Econometrica 70, 1741–1779.
Eeckhout, J., 2004. Gibrat's law for (all) cities. Am. Econ. Rev. 94, 1429–1451.
Eeckhout, J., Pinheiro, R., Schmidheiny, K., 2014. Spatial sorting. J. Polit. Econ. 122, 554–620.
Ellison, G., Glaeser, E.L., 1997. Geographic concentration in U.S. manufacturing industries: a dartboard approach. J. Polit. Econ. 105, 889–927.
Faber, B., 2014. Trade integration, market size, and industrialization: evidence from China's National Trunk Highway System. Rev. Econ. Stud. forthcoming.
Fafchamps, M., Shilpi, F., 2005. Cities and specialization: evidence from South Asia. Econ. J. 115, 477–504.
Fallah, B., Partridge, M., 2012. Geography and high-tech employment growth in U.S. counties. MPRA Paper 38294.
Fay, M., Opal, C., 2000. Urbanization without growth: a not-so-uncommon phenomenon. World Bank Policy Research Working paper Series 2412.
Feler, L., Henderson, J.V., 2011. Exclusionary policies in urban development. J. Urban Econ. 69, 253–272.
Forman, C., Goldfarb, A., Greenstein, S., 2005. Geographic location and the diffusion of internet technology. Electron. Commer. Res. Appl. 4, 1–13.
Fujita, M., Ogawa, H., 1982. Multiple equilibria and structural transition of non-monocentric configurations. Reg. Sci. Urban Econ. 12, 161–196.
Fujita, M., Henderson, J.V., Kanemoto, Y., Mori, T., 2004. The spatial distribution of economic activities in Japan and China. In: Henderson, J.V., Thisse, J.F. (Eds.), Handbook of Regional and Urban Economics, vol. 4. Elsevier, Amsterdam.
Gabaix, X., 1999. Zipf's law for cities: an explanation. Q. J. Econ. 114, 739–767.
Gabaix, X., Ioannides, Y., 2004. The evolution of city size distributions. In: Henderson, J.V., Thisse, J.F. (Eds.), Handbook of Regional and Urban Economics, vol. 4. Elsevier, Amsterdam.
Galor, O., Weil, D., 2000. Population, technology, and growth: from the Malthusian regime to the demographic transition and beyond. Am. Econ. Rev. 90, 806–828.
Galor, O., Moav, O., Vollrath, D., 2009. Inequality in landownership, human capital promoting institutions and the great divergence. Rev. Econ. Stud. 76 (1), 143–179.
Gibrat, R., 1931. Les inégalités économiques: applications aux inégalités de richesses, à la concentration des entreprises, aux populations des villes, aux statistiques des familles, etc., d'une loi nouvelle, la loi de l'effet proportionnel. Librairie du Recueil Sirey, Paris.
Glaeser, E., Gyourko, J., 2005. Urban decline and durable housing. J. Polit. Econ. 113, 345–375.
Gollin, D., Parente, S.L., Rogerson, R., 2007. The food problem and the evolution of international income levels. J. Monet. Econ. 54, 1230–1255.
Gyourko, J., Mayer, C., Sinai, T., 2013. Superstar cities. Am. Econ. J. Econ. Policy 5, 167–199.
Hansen, G., Prescott, E.C., 2002. Malthus to Solow. Am. Econ. Rev. 92, 1205–1217.
Harris, J.R., Todaro, M.P., 1970. Migration, unemployment and development: a two-sector analysis. Am. Econ. Rev. 60, 126–142.
Harris-Dobkins, L., Ioannides, Y.M., 2001. Spatial interactions among U.S. cities: 1900–1990. Reg. Sci. Urban Econ 31, 701–731.
Helsley, R.W., Strange, W.C., 1990. Agglomeration economies and matching in a system of cities. Reg. Sci. Urban Econ. 20, 189–212.
Helsley, R.W., Strange, W.C., 2014. Coagglomeration, clusters and the scale and composition of cities. J. Polit. Econ. 122, 1064–1093.
Henderson, J.V., 1974. The sizes and types of cities. Am. Econ. Rev. 64 (4), 640–656.
Henderson, J.V., 1988. Urban Development: Theory, Fact, and Illusion. Oxford University Press, New York.
Henderson, J.V., 1997. Medium size cities. Reg. Sci. Urban Econ. 27, 583–612.

Henderson, J.V., 2005. Urbanization and growth. In: Aghion, P., Durlauf, S. (Eds.), Handbook of Economic Growth. Elsevier, Amsterdam.
Henderson, J.V., 2010. Cities and development. J. Reg. Sci. 50, 515–540.
Henderson, J.V., Venables, A., 2009. The dynamics of city formation. Rev. Econ. Dyn. 12, 233–254.
Henderson, J.V., Wang, H.G., 2005. Aspects of the rural-urban transformation of countries. J. Econ. Geogr. 5, 23–42.
Henderson, J.V., Wang, H.G., 2007. Urbanization and city growth: the role of institutions. Reg. Sci. Urban Econ. 37, 283–313.
Henderson, J.V., Lee, T., Lee, Y.J., 2001. Scale externalities in a developing country. J. Urban Econ. 49, 479–504.
Henderson, J.V., Roberts, M., Storeygard, A., 2013. Is urbanization in sub-Saharan Africa different? Policy Research Working paper Series 6481. World Bank.
Hobijn, B., Jovanovic, B., 2001. The information-technology revolution and the stock market: evidence. Am. Econ. Rev. 91, 1203–1220.
Holmes, T.J., Lee, S., 2010. Cities as six-by-six-mile squares: Zipf's law? NBER. In: Glaeser, E. (Ed.), Agglomeration Economics. National Bureau of Economic Research, pp. 105–131 (Chapter).
Holmes, T., Stevens, J.J., 2004. Spatial distribution of economic activities in North America. In: Henderson, J.V., Thisse, J.F. (Eds.), Handbook of Regional and Urban Economics, vol. 4. Elsevier, Amsterdam.
Ioannides, Y., Skouras, S., 2013. US city size distribution: robustly Pareto, but only in the tail. J. Urban Econ. 73 (1), 18–29.
Jefferson, G.H., Singh, I., 1999. Enterprise Reform in China: Ownership, Transition, and Performance. World Bank, Washington, DC and Oxford University Press, New York.
Jimenez, E., 1984. Tenure security and urban squatting. Rev. Econ. Stat. 66, 556–567.
Jovanovic, B., Rousseau, P.L., 2005. General purpose technologies. In: Aghion, P., Durlauf, S. (Eds.), Handbook of Economic Growth. Elsevier, Amsterdam (Chapter 18).
Kim, S., 1998. Economic integration and convergence: U.S. regions, 1840–1987. J. Econ. Hist. 58, 659–683.
Kim, S., 1999. Regions, resources, and economic geography: the sources of USA comparative advantage, 1880–1987. Reg. Sci. Urban Econ. 29, 1–32.
Kim, S., 2009. Spatial inequality and development: theories, facts and policies. In: Buckley, R., Annez, P., Spence, M. (Eds.), Urbanization and Growth. The International Bank for Reconstruction and Development and World Bank, Washington, DC.
Kim, S., Margo, R.A., 2004. Historical perspectives on U.S. economic geography. In: Henderson, J.V., Thisse, J.F. (Eds.), Handbook of Regional and Urban Economics. first ed., vol. 4. Elsevier, Amsterdam, pp. 2981–3019 (Chapter 66).
Kolko, J., 1999. Can I get some service here: information technology, service industries, and the future of cities. Harvard University, Mimeo.
Krugman, P., 1991. Increasing returns and economic geography. J. Polit. Econ. 99 (3), 483–499.
Krugman, P., 1992. Geography and Trade. MIT Press, Gaston Eyskens Lecture Series, Cambridge, MA.
Krugman, P., 1996. Confronting the mystery of urban hierarchy. J. Jpn. Int. Econ. 10, 399–418.
Lee, S., Li, Q., 2013. Uneven landscapes and city size distributions. J. Urban Econ. 78, 19–29.
Lewis, W.A., 1954. Economic development with unlimited supplies of labour. Manch. Sch. 22 (2), 139–191.
Lucas, R.E., 2004. Life earnings and rural-urban migration. J. Polit. Econ. 112 (S1), S29–S59.
Matsuyama, K., 1992. Agricultural productivity, comparative advantage, and economic growth. J. Econ. Theory 58, 317–334.
McGuckin, R.H., Stiroh, K.J., 2002. Computers and productivity: are aggregation effects important? Econ. Inq. 40, 42–59.
Meyer, J.R., Kain, J.F., Wohl, M., 1965. The Urban Transportation Problem. Harvard University Press, Cambridge.
Michaels, G., Rauch, F., Redding, S., 2012. Urbanization and structural transformation. Q. J. Econ. 127, 535–586.

Mitchener, K.J., McLean, I.W., 1999. U.S. regional growth and convergence, 1880–1980. J. Econ. Hist. 59, 1016–1042.
Ngai, R.L., Pissarides, C.A., 2007. Structural change in a multisector model of growth. Am. Econ. Rev. 97 (1), 429–443.
Nurkse, R., 1953. Problems of Capital Formation in Underdeveloped Countries. Oxford University Press, New York.
Ono, Y., 2003. Outsourcing business services and the role of central administrative offices. J. Urban Econ. 53 (3), 377–395.
Puga, D., 1999. The rise and fall of regional inequalities. Eur. Econ. Rev. 43, 303–334.
Rappaport, J., Sachs, J.D., 2003. The United States as a coastal nation. J. Econ. Growth 8, 5–46.
Rauch, J.E., 1993. Does history matter only when it matters a little? The case of city-industry location. Q. J. Econ. 108, 843–867.
Ray, D., 1998. Development Economics. Princeton University Press, Princeton (Chapter 3).
Redding, S., Turner, M., 2015. Transportation costs and the spatial organization of economic activity. In: Duranton, G.J., Henderson, V., Strange, W. (Eds.), Handbook of Regional and Urban Economics, vol. 5. Elsevier, Amsterdam.
Renaud, B., 1981. National Urbanization Policy in Developing Countries. Oxford University Press, Oxford.
Rosenbloom, J.L., 1990. One market or many? Labor market integration in the late nineteenth-century United States. J. Econ. Hist. 50, 85–107.
Rossi-Hansberg, E., Wright, M., 2007. Urban structure and growth. Rev. Econ. Stud. 74 (2), 597–624.
Rostow, W.W., 1960. The Stages of Economic Growth: A Non-Communist Manifesto. Cambridge University Press, Cambridge, UK.
Schultz, T.W., 1968. Economic Growth and Agriculture. McGraw-Hill, New York.
Shaw-Taylor, L., Wrigley, E.A., 2008. The Occupational Structure of England c.1750 to c.1871, http://www.geog.cam.ac.uk/research/projects/occupations/.
Tabuchi, T., Thisse, J.F., 2011. A new economic geography model of central places. J. Urban Econ. 69, 240–252.
Tamura, R., 2002. Human capital and the switch from agriculture to industry. J. Econ. Dyn. Control. 27, 207–242.
Tecu, I., 2013. The location of industrial innovation: does manufacturing matter? US Census Bureau Center for Economic Studies Paper No. CES-WP-13-09.
Trew, A., 2014. Spatial takeoff in the first industrial revolution. Rev. Econ. Dyn 17, 707–725.
Triplett, J.E., Bosworth, B.P., 2002. 'Baumol's Disease' has been cured: IT and multifactor productivity in U.S. services industries. In: Jansen, D.W. (Ed.), The New Economy and Beyond: Past, Present, and Future. Edward Elgar Publishing, Cheltenham, UK and Northampton, MA, pp. 34–71.
United Nations, 2010. World Population Policies 2009. United Nations, New York.
Williamson, J., 1965. Antebellum urbanization in the American Northeast. J. Econ. Hist. 25, 592–608.
World Bank, 2009. World Development Report 2009: Reshaping Economic Geography. The World Bank, Washington, DC.