The growth of cities

Gilles Duranton[*‡]
University of Pennsylvania and *CEPR*

Diego Puga[*§]
CEMFI and *CEPR*

May 2013

Abstract: Why do cities grow in population, surface area, and income per person? Which cities grow faster and why? To these questions, the urban growth literature has offered a variety of answers. Within an integrated framework, this chapter reviews key theories with implications for urban growth. It then relates these theories to empirical evidence on the main drivers of city growth, drawn primarily from the United States and other developed countries. Consistent with the monocentric city model, fewer roads and restrictions on housing supply hinder urban growth. The fact that housing is durable also has important effects on the evolution of cities. In recent decades, cities with better amenities have grown faster. Agglomeration economies and human capital are also important drivers of city growth. Although more human capital, smaller firms, and a greater diversity in production foster urban growth, the exact channels through which those effects percolate are not clearly identified. Finally, shocks also determine the fate of cities. Structural changes affecting the broader economy have left a big footprint on the urban landscape. Small city-specific shocks also appear to matter, consistent with the recent wave of random growth models.

Key words: urban growth, agglomeration economies, land use, transportation, amenities

JEL classification: C52, R12, D24

[*]This is a draft of a chapter written for eventual publication in the *Handbook of Economic Growth*, Volume 2, edited by Steven N. Durlauf and Philippe Aghion, to be published by Elsevier. We thank conference and seminar participants at the Paris School of Economics and the 2011 North American Meeting of the Regional Science Association for comments and discussions. This research was developed in part when Duranton was visiting the Paris School of Economics, whose financial support is gratefully acknowledged. Puga gratefully acknowledges funding from the European Communities's Seventh Framework Programme under ERC Advanced Grant agreement 269869.

[‡]Wharton School, University of Pennsylvania, 3620 Locust Walk, Philadelphia, PA 19104, USA (email: duranton@wharton.upenn.edu; website: https://real-estate.wharton.upenn.edu/profile/21470/).

[§]Centro de Estudios Monetarios y Financieros (CEMFI), Casado del Alisal 5, 28014 Madrid, Spain (e-mail: diego.puga@cemfi; website: http://diegopuga.org).

1. Introduction

In 2010, the mean population size of the 366 US metropolitan areas was 707,000, with a range from 18.3 million to just over 50,000. Between 2000 and 2010, these cities grew on average by 10.7%. The first decade of the 21st century was not exceptional for US urban growth. US metropolitan areas grew on average by 17.9% per decade since 1920, the earliest year for which consistent data is available.[1] This figure of 17.9% exceeds aggregate population growth by 5.3 percentage points even though a growing population could be accommodated in more cities instead of larger cities. When it comes to urban growth, the United States is not an exceptional country. In Spain, urban areas grew on average by 17.5% between 2000 and 2010, and by 18.1% per decade on average between 1920 and 2010, exceeding aggregate population growth in Spain by 9.2 percentage points. In France, metropolitan areas grew on average by 4% between 1999 and 2007, and by 7.7% per decade on average between 1936 and 2007, exceeding aggregate population growth in France by 2 percentage points.

Although cities tend to grow over time, they do not grow uniformly at the same rate. The standard deviation of the growth rate of US metropolitan areas between 2000 and 2010 is slightly larger than its corresponding mean. Observing individual city growth rates over a decade with means and standard deviations of about the same magnitude is typical. This is the case for the 1920–2010 period in the United States, in Spain, and in France. These figures about the mean and standard deviation of the growth rates of cities naturally lead to asking why cities keep growing even after countries are already highly urbanized, and why some cities grow faster than others.

Being able to answer these questions is important for at least three reasons. The first is that the population growth of cities is economically important in itself. Extremely large investments in building new housing and infrastructure must be made to accommodate the demographic growth of cities. For instance, American households spend about a third of their income on housing, according to the Consumer Expenditure Survey. For their part, various levels of the US government spend more than 200 billion dollar every year to maintain and expand the road infrastructure. Given that most of these investments are extremely durable, it is important to plan them properly and, for this, we need to understand why and how cities grow.

Second, urban economics has proposed a number of theories to explain the population size of cities. Following Alonso (1964), Mills (1967), and Muth (1969), a large literature has focused on the importance of location within the city and its impact on commuting costs as a key determinant of land use and housing development in cities. In turn, the ease of commuting, the availability of housing, and earnings determine the population size of cities. Following Rosen (1979) and Roback (1982), urban economists have also paid great attention to the role of amenities in attracting people to cities. Recognizing that earnings and productivity are themselves systematically related to the

[1] The computations for the United States are based on the 2009 definition of metropolitan areas. Using the earliest definition of metropolitan areas that can be applied to county population data, the 1950 Standard Metropolitan Statistical Areas, we observe a mean growth of 7.3% between 2000 and 2010 and 15.8% by decade on average between 1920 and 2010. These lower figures probably understate the true population growth of US cities which, to some extent, grew through the expansion of their suburban areas that were not taken into account by the 1950 definition. On the other hand, the figures based on 2009 definitions probably overstate the true growth of US cities since they partly reflect the selection of the fastest growing cities that became the largest and form the existing set of metropolitan areas.

population size of cities, much work has been devoted to modelling the productive advantages of cities or agglomeration economies explicitly (e.g., Fujita, 1988, Helsley and Strange, 1990, Glaeser, 1999, Duranton and Puga, 2001). The tradeoff between agglomeration economies and urban costs, at the core of systems of cities models building on Henderson (1974), is widely accepted as the key explanation behind the existence of cities and provides some important implications for their population growth. Finally, the existence of some regularities in the size distribution of cities and in the patterns of urban growth has motivated alternative approaches which emphasize the importance of random shocks in urban growth (e.g., Gabaix, 1999*a*).

These theories offer useful guidance to conduct empirical work on urban growth by providing us with specifications and by highlighting a number of identification pitfalls. Conversely, an evaluation of the key drivers of urban growth is also an evaluation of the predictions of the core approaches to the economics of cities.

A third reason to study urban growth is that cities offer an interesting window through which to study the process of economic growth. How cities grow and why may hold important lessons for how and why economies grow. Existing theories of economic growth emphasize the importance of direct interactions. Such interactions often involve direct physical proximity between individuals and are thus naturally studied within cities. Taking the advice of Lucas (1988) seriously, it may be in cities that economic growth is best studied

We also note that the population growth of cities may be easier and simpler to study than the process of growth of entire countries. The large cross-country growth literature which builds on Barro's (1991) work is afflicted by fundamental data and country heterogeneity problems that are much less important in the context of cities within a country. Furthermore, cross-country growth regressions are plagued by endogeneity problems that are often extremely hard to deal with in a cross-country setting (Durlauf, Johnson, and Temple, 2005). As we show in this review, looking at cross-sections of cities within countries offers more hope of finding solutions to these identification problems.

To finish this introduction, we would like to delineate more precisely what this chapter does and what it does not do. First, we focus mostly on cities in developed economies. Most of the empirical evidence we discuss below originates from there, the United States in particular. Consistent with this, the theories we discuss consider implicitly 'mature' cities between which workers move. Rural-urban migrations, urbanization, and the role of cities in developing countries are not examined here. We refer instead the reader to Henderson (2005) in a previous volume of this *Handbook*. Second, we discuss and attempt to unify work that has taken place within one discipline, economics. We are aware that other social scientists in geography, planning, or sociology, have taken an interest in urban growth. We leave the bigger task of integrating cross-disciplinary perspectives to others (see Storper and Scott, 2009, for references and one such attempt).

2. Land use and transportation

Urban scholars have long recognized that transportation costs are a fundamental determinant of both the population size of cities and their patterns of land use.[2] To understand more precisely the articulation between transportation, land use, and city population, we start with a simple monocentric urban model in the spirit of Alonso (1964), Mills (1967), and Muth (1969).[3] We then use the predictions of this model to structure our examination of the empirical literature on cities and transportation. In subsequent sections, we also enrich this model to account for other features such as amenities and agglomeration economies.

2.1 The monocentric city model

Consider a linear monocentric city. Land covered by the city is endogenously determined and can be represented by a segment on the positive real line. Production and consumption of a numéraire good take place at a single point $x = 0$, the Central Business District (CBD). Preferences can be represented by a utility function $U(A, u(h, z))$ written in terms of the common amenity level enjoyed by everyone in the city, A, and a sub-utility $u(h, z)$ derived from individual consumption of housing, h, and of the numéraire, z. Commuting costs increase linearly with distance to the CBD, so that a worker living at distance x incurs a commuting cost τx. This leaves $w - \tau x$ for expenditure on housing and the numéraire.[4] Denoting by $P(x)$ the rental price of housing at a distance x from the CBD, we can use a dual representation of the sub-utility derived from housing and the numéraire, and represent preferences with

$$U(A, v(P(x), w - \tau x)) \,, \tag{1}$$

where $\frac{\partial U}{\partial A} > 0$, $\frac{\partial U}{\partial v} > 0$, $\frac{\partial v}{\partial P(x)} < 0$, and $\frac{\partial v}{\partial (w - \tau x)} > 0$.

All residents in the city are identical in income and preferences, enjoy a common amenity level, and are freely mobile within the city. At the residential equilibrium, residents must derive the same sub-utility from housing consumption and the numéraire:

$$v(P(x), w - \tau x) = \overline{v} \,. \tag{2}$$

Totally differentiating equation (2) with respect to x yields

$$\frac{\partial v(P(x), w - \tau x)}{\partial P(x)} \frac{\mathrm{d}P(x)}{\mathrm{d}x} - \tau \frac{\partial v(P(x), w - \tau x)}{\partial (w - \tau x)} = 0 \,, \tag{3}$$

[2] For early cities, urban historians insist on the difficulty of supplying their residents with food. See for instance Duby (1981–1983), de Vries (1984), or Bairoch (1988). For modern cities, the same scholars point at the cost of moving residents within cities as the being key impediment on urban growth. On that they agree with observers of contemporary cities such as Glaeser and Kahn (2004) who often mention the automobile as the single most important driver of urban change. LeRoy and Sonstelie (1983) and Glaeser, Kahn, and Rappaport (2008), among others, argue that the transportation technologies and their relative costs are also a major driver of where rich and poor residents live within cities. We do not address this last set of issues here.

[3] These models derive from a common ancestor, Thünen's (1826) *Isolated State*, who applied a similar logic to understand the spatial organization of crops in large farms. A detailed presentation of the monocentric model can be found in Fujita (1989).

[4] We generalize this specification below in several ways, including allowing commuting costs to be non-linear and endogenizing wages.

which implies

$$\frac{\mathrm{d}P(x)}{\mathrm{d}x} = -\frac{\tau}{-\frac{\partial v(P(x),w-\tau x)}{\partial P(x)} / \frac{\partial v(P(x),w-\tau x)}{\partial (w-\tau x)}} = -\frac{\tau}{h(x)} < 0\,, \tag{4}$$

where the simplification follows from Roy's identity. Equation (4) is often referred to as the Alonso-Muth condition. It states that, at the residential equilibrium, if a resident moves marginally away from the CBD, the cost of her current housing consumption falls just as much as her commuting costs increase. Thus, the price of housing decreases with distance to the CBD. Then, residents react to this lower price by consuming more housing (larger residences) the farther they live from the CBD. To see this, simply differentiate the Hicksian demand for housing with respect to x:

$$\frac{\partial h(P(x),\overline{v})}{\partial x} = \frac{\partial h(P(x),\overline{v})}{\partial P(x)} \frac{\mathrm{d}P(x)}{\mathrm{d}x} \geqslant 0\,. \tag{5}$$

Note this is a pure substitution effect, since utility is being held constant at $\overline{v}$. This also implies that the price of housing is convex in distance to the CBD: house prices do not need to fall as fast as commuting costs increase with distance to the CBD to keep city residents indifferent, since they enjoy having a larger house.

To supply housing, a perfectly competitive construction industry uses land and capital under constant returns to scale to produce an amount $f(x)$ of housing floorspace per unit of land at a distance x from the CBD. The rental price of land, denoted $R(x)$, varies across the city. The rental price of capital is constant and exogenously given, so we omit it as an argument of the unit cost function in construction $c(R(x))$. The zero-profit condition for the construction sector can then be written as

$$P(x) = c(R(x))\,. \tag{6}$$

Totally differentiating equation (6) with respect to x yields

$$\frac{\mathrm{d}P(x)}{\mathrm{d}x} = \frac{\partial c(R(x))}{\partial R(x)} \frac{\mathrm{d}R(x)}{\mathrm{d}x}\,, \tag{7}$$

which implies

$$\frac{\mathrm{d}R(x)}{\mathrm{d}x} = \frac{\mathrm{d}P(x)}{\mathrm{d}x} \frac{1}{\frac{\partial c(R(x))}{\partial R(x)}} = \frac{\mathrm{d}P(x)}{\mathrm{d}x} f(x) < 0\,, \tag{8}$$

where the simplification follows from the envelope theorem. Thus, the reduction in the price of housing as one moves away from the CBD gets reflected in a reduction in the price of land. The construction industry then reacts to lower land prices by building with a lower capital to land ratio (fewer stories and larger gardens) further away from the CBD.

Land is built if the rent $R(x)$ it can fetch in residential use is at least as high as the rent $\underline{R}$ it can fetch in the best alternative use (e.g., agriculture). The edge of the city is thus located at a distance $\overline{x}$ from the CBD such that $R(\overline{x}) = \underline{R}$. The physical extent of the city must also be sufficient to hold its population N:

$$N = \int_0^{\overline{x}} d(x)\mathrm{d}x\,, \tag{9}$$

where $d(x)$ denotes population density at a distance x from the CBD. Using equations (4) and (8), we can express population density as

$$d(x) = \frac{f(x)}{h(x)} = \frac{\frac{\mathrm{d}R(x)}{\mathrm{d}x} / \frac{\mathrm{d}P(x)}{\mathrm{d}x}}{-\tau / \frac{\mathrm{d}P(x)}{\mathrm{d}x}} = -\frac{1}{\tau}\frac{\mathrm{d}R(x)}{\mathrm{d}x} \,. \tag{10}$$

Substituting this expression for $d(x)$ into equation (9), solving the integral, and using $R(\overline{x}) = \underline{R}$ yields $N = \frac{R(0)-\underline{R}}{\tau}$. This implies a very simple expression for land rent at the CBD ($x = 0$):

$$R(0) = \underline{R} + \tau N \,. \tag{11}$$

Valuing equation (6) at $x = 0$ and using (11), we can write the price of housing at the CBD as $P(0) = c(\underline{R} + \tau N)$. Equation (2) holds for any location in the city, so valuing it at an arbitrary x and at $x = 0$, and using the previous expression for $P(0)$ yields

$$\begin{aligned} v(P(x), w - \tau x) = \overline{v} &= v(P(0), w) \\ &= v(c(\underline{R} + \tau N), w) \,. \end{aligned} \tag{12}$$

This can be inverted to solve for house prices $P(x)$ as a function of x, N, w, τ, and $\underline{R}$. That is the 'closed city' version of the monocentric city model, which treats population N as a parameter. The 'open city' version allows N to be endogenously determined by migration across cities to attain a common utility level $\overline{U}$. If the amenity level A is common to all cities, we only need to consider the sub-utility derived from housing and the numéraire and can write the condition of utility equalization across cities as:

$$v(c(\underline{R} + \tau N), w) = \overline{v} \,. \tag{13}$$

This spatial equilibrium condition can be inverted to solve for N as a function of $\overline{v}$, w, τ, and $\underline{R}$.[5]

Before going any further, it is worth asking whether the monocentric city model provides reasonable guidance for empirical work. The main issue with the monocentric city model is that it imposes a particular geography for employment.[6] Observation suggests that the geography of most cities is far less extreme than postulated by the monocentric city model, where all employment is concentrated in a single location. In 1996 only about 25% employees in US metropolitan areas worked within five kilometres of their CBD (Glaeser and Kahn, 2001). There is a tendency for employment to diffuse away from centres and for metropolitan areas to develop secondary centres (Anas, Arnott, and Small, 1998, McMillen, 2001). Despite these clear limitations, the monocentric model remains useful for a number of reasons. First, there is strong empirical support for the existence of declining gradients of land and housing prices, population density, and intensity of

[5]These models also deliver a number of proportionality results between urban aggregates such as total differential land rent and total commuting costs. We do not develop them here given our focus on population size. The interested reader can refer to Arnott and Stiglitz (1981) and Fujita (1989).

[6]Following Fujita and Ogawa (1982) and, more recently, Lucas and Rossi-Hansberg (2002), economists have attempted to endogenize the location of employment in cities. These models deliver very useful insights and, in some cases, plausible narratives about observed changes in urban forms. However, these models are too complex for their comparative statics results to be easily tested except in some specific dimensions like the number of subcentres (McMillen and Smith, 2003).

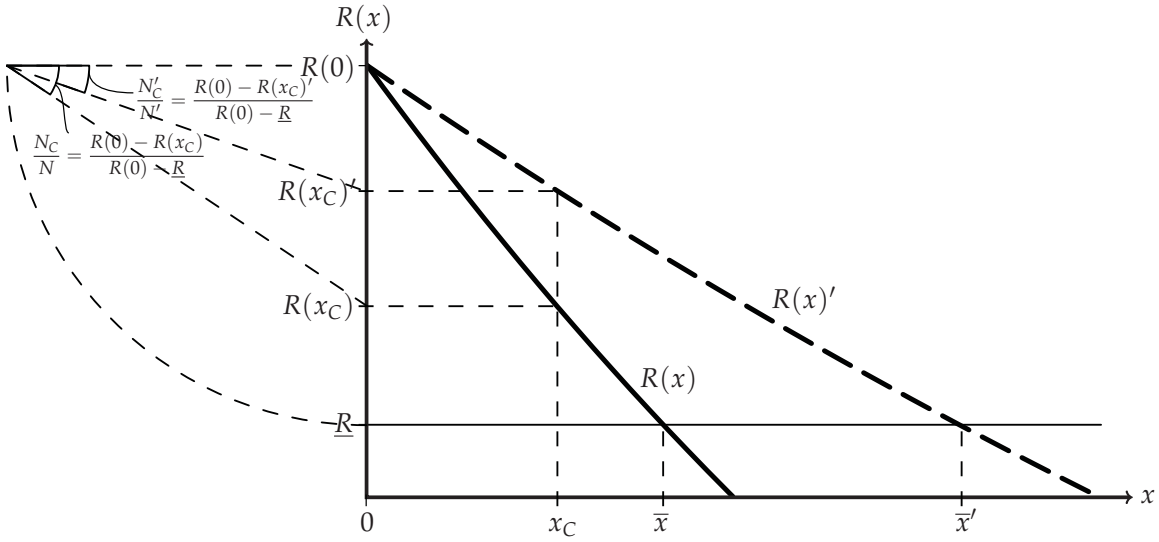


Figure 1: Residential and agricultural land rent against distance to the CBD

construction as predicted by the monocentric city model (see McMillen, 2006, for an introduction to the voluminous literature on this topic). In addition, the monocentric city model has comparative static properties relevant for urban growth that carry through to models with a richer spatial structure, including polycentric cities, and also to models without an explicit modeling of space within the city. Simple models that capture essential elements of reality, such as the monocentric city model, are useful because they provide a solid base to specify regressions, help us with identification, and facilitate the interpretation of results. However, we must also be careful not to give a narrow structural interpretation to parameters estimated using the monocentric model as motivation.

2.2 *Commuting infrastructure and population growth*

Local transportation improvements are often justified on the basis that they promote city growth. The monocentric city model sustains this claim. Consider a local improvement in transportation that lowers τ in one particular city within a large urban system. It follows immediately from equation (13) that a reduction in commuting costs increases this city's population with unit elasticity.

The intuition for this result is straightforward and is illustrated in figure 1. The figure plots land rent as a function of distance to the CBD before (solid downward sloping curve) and after (dashed downward sloping curve) a fall in τ. (Ignore for now the dashed lines to the left of the vertical axis, which will be used in section 2.3.) Since any change in this particular city is too small to affect the large urban system, the level of utility of every resident in the city must remain unchanged to satisfy the spatial equilibrium condition (13). Someone living at the CBD does not commute to work and is thus not directly affected by the fall in τ. The spatial equilibrium condition then implies that residential land rent at the CBD, $R(0) = \underline{R} + \tau N$, must remain unchanged, which requires population N to increase in the same proportion as τ falls. Everywhere beyond the CBD, residents benefit from the reduction in τ, but land and house prices increase as a result of immigration,

offsetting the utility gain from lower commuting costs. The shift in land rents pushes outwards the edge of the city, given by the intersection of $R(x)$ with $\underline{R}$, from $\bar{x}$ to $\bar{x}'$. The larger population is housed through a combination of this increase in the spatial size of the city and rising densities everywhere (as people react to rising house prices by reducing their housing consumption, and the construction industry reacts to rising land prices by building more floorspace per unit of land).

This prediction of a unit elasticity of city population with respect to commuting costs maps directly into the following regression:

$$\Delta_{t+1,t} \log N_i = \beta_0 - \beta_1 \Delta_{t+1,t} \log \tau_i + \epsilon_{it} \ . \tag{14}$$

where i indexes cities, $\Delta_{t+1,t}$ is a time-differencing operator between period t and period $t+1$, β_1 is the elasticity of interest (predicted to be unity), and ϵ_{it} is an error term which, for the time being, we can interpret as a random disturbance.

To begin, we note that testing whether the coefficient β_1 estimated in regression (14) differs from unity would be more than a test of the core mechanism of the monocentric city model. It would be a joint test of several assumptions in that model, including the linearity of commuting costs and free labour mobility. Because we do not expect all the conditions leading a unit population elasticity to hold, being able to reject that the estimated value of β_1 is exactly one is of secondary importance. Instead, we are primarily interested in knowing whether commuting costs affect the population of cities and how important this factor might be both in absolute terms and relatively to other drivers of urban growth. The advice of Leamer and Levinsohn (1995), "estimate, don't test", is particularly relevant here.

Equation (14) belongs to a much broader class of regressions where the growth of cities is regressed on a number of explanatory variables. Hence, the estimation issues raised by this regression also occur in most urban growth regressions. We discuss these general issues at length here and avoid repeating them when discussing similar regressions below.

A first key issue is the speed of adjustment. Rather than assume free labour mobility in a static sense, one could think of the equilibrium population N_i^* that satisfies equation (13) as a steady-state towards which the city converges. New housing takes time to build and we cannot expect an immediate adjustment of city population after a change in commuting costs. We might instead posit the following myopic adjustment process where $N_{it+1} = N_i^{*\lambda} N_{it}^{1-\lambda}$. The parameter λ can be interpreted as a rate of convergence. We have $\lambda = 0$ if residents cannot change city and $\lambda = 1$ if they fully adjust between any two periods.[7] Taking logs of this adjustment process equation implies $\Delta_{t+1,t} \log N_i = \lambda(\log N_i^* - \log N_{it})$. The spatial equilibrium condition (13) implies that τN_i^* should be constant in steady state, i.e. $\log N_i^* = \beta_0 - \beta_1 \log \tau_{it}$, with β_1 predicted to be unity. Combining these two equations leads to the following regression:

$$\Delta_{t+1,t} \log N_i = \lambda \beta_0 - \lambda \log N_{it} - \lambda \, \beta_1 \log \tau_{it} + \epsilon_{it} \ . \tag{15}$$

The choice between a 'changes on changes' regression like (14) and a 'changes on levels' regression like (15) matters because these two regressions use very different sources of variation in the data

[7] Baldwin (2001) shows how this ad-hoc migration specification can be consistent with forward-looking behaviour when migration across cities generates congestion frictions.

and, as a result, suffer from different identification problems. Ideally, this choice of specification should be driven by informed priors about how population adjusts. An advantage of equation (15) is that the speed of convergence λ is estimated together with the parameter of interest, β_1.

The recognition that transport improvements may take time to affect city growth and that other factors will influence this process creates additional identification concerns. As a first step, we should control for other factors that may simultaneously affect city growth. However, it is not possible to control for all such factors. If there are omitted variables that drive urban growth and are correlated with transportation costs, then ordinary least square (OLS) estimates of the effects of transport costs on city growth will be biased. A second, related, concern is possible reverse causation, where transport infrastructure is assigned on the basis of expected growth. Even in the absence of forward-looking infrastructure assignment, transport costs and future growth can be correlated. This is because we expect any measure of local transport costs to be serially correlated and, as discussed below, there is also persistence in urban growth.

As made clear by the rest of this chapter, these concerns of correlated omitted variables and reverse causation or endogeneity plague all city growth regressions. This is unsurprising. Regressions like (15) strongly resemble cross-country growth regressions in the line of Barro (1991).[8] It is well-known that these regressions are afflicted by serious problems of correlated omitted variables and endogeneity (Durlauf, Johnson, and Temple, 2005). In a cross-country setting, these problems are extremely hard to deal with and solutions are very few. Looking at cities within countries offers more hope regarding identification.[9]

One might be tempted to tackle the problem of correlated omitted variables by building a panel of cities and estimating a regression based on equations (14) or (15) with city fixed-effects. However, if transport infrastructure is allocated on the basis of the economic fortunes of cities the correlation between changes in the transport infrastructure and the growth residual (i.e., the error term) in the regression will be much stronger than the correlation between the level of infrastructure and the error. That is, fixed-effect and first-difference estimations can suffer from worse biases than simple cross-sectional estimations.

Duranton and Turner (2012) tackle correlated omitted variables and reverse causation using instrumental variables to estimate a regression that is very close to equation (15).[10] They use as dependent variable the change in log employment between 1983 and 2003 for US metropolitan areas. As a proxy for τ they use lane kilometres of interstate highways within metropolitan areas in 1983 (although interstate highways represent only a small proportion of the roadway, they carry a disproportionate amount of traffic). Duranton and Turner (2012) instrument interstate highways using three historical measures of roads: the 1947 highway plan that was the template for the

[8] In the neoclassical models of growth that underpin cross-country growth regressions, the 'changes on levels' specification for the regressions arises from the slow adjustment of capital in the process of convergence towards steady-state. In our case, this is driven by the slow adjustment of labour.

[9] Cross-country growth regressions are also afflicted by fundamental problems of data and cross-country heterogeneity which are much less important in the context of metropolitan areas within countries.

[10] Duranton and Turner (2012) derive their specification from a model where, unlike in the monocentric model, relative locations within cities do not matter even though residents have a demand for transportation within the city. As noted above, it is reassuring that better transportation is predicted to be a driver of urban growth in a class of models broader than the monocentric city model.

modern US interstate highway system, a map of 1898 railroads, and a map of old exploration routes of the continent dating back to 1528.

As usual with this type of strategy, it relies on the instruments being relevant, i.e., on their ability to predict roads conditional on the control variables being used. Denoting the instruments Z_i:

$$\mathbb{C}\text{ov}(\log \tau_i, Z_i|.) \neq 0\,. \tag{16}$$

This condition can be formally assessed (see Angrist and Pischke, 2008, for details). In the case of the three instruments used by Duranton and Turner (2012) the relevance condition is satisfied even when using a demanding set of control variables. This is because the 1947 highway map was by and large implemented, old railroads were turned into roads or highways were built alongside them, and many pathways discovered a long time ago through exploration are still pathways today.

The validity of the instruments also relies on them being exogenous, i.e., on them being correlated with population growth only through the roadway so that they are orthogonal with the error term:

$$\mathbb{C}\text{ov}(\epsilon_i, Z_i|.) = 0, \tag{17}$$

Establishing exogeneity is much harder than establishing relevance. The first step for the defense of any set of instruments is to show that they are not directly linked to the dependent variable. In the case at hand, the 1947 highway planners were interested in linking US cities together but were not concerned with future commuting patterns. Railroad builders in 1898 were interested in shipping grain, cattle, lumber, and passengers across the continent. Early explorers were interested in finding a wide variety of things, from the fountain of youth to pathways to the Pacific. This first step is necessary but not sufficient. The exogeneity condition (17) fails when an instrument is correlated with a missing variable that also affects the dependent variable. For instance, cities in more densely populated parts of the country in 1947 received more kilometres of planned highways. Those cities might also have grown less between 1980 and 2000. The second step is thus to use further controls, and in particular population controls, in the instrumental variable (IV) estimation to preclude such correlations with missing variables as much as possible.

Over-identification tests are the next element of any IV strategy. They can be conducted when there are more instruments than (endogenous) parameters to estimate. However, we expect very similar instruments to lead to very similar estimates and thus pass over-identification tests. This should not be taken as a proof of instruments validity. Over-identification tests are more meaningful when the instruments rely on very different sources of variation in the data.

Finally, a difference between OLS and IV estimates can be indicative of an OLS bias. However with invalid instruments, the bias on the IV estimate could be even worse. Thus, whenever there are significant differences between OLS and IV estimates it is important to provide out-of-sample evidence for the channels through which the OLS bias percolates.

In conclusion, any reasonable IV strategy needs to (i) establish the strength of its instruments, (ii) provide a plausible argument that the instruments are independent from the dependent variable, (iii) preclude alternative indirect channels of correlation between the instruments and the dependent variable, (iv) show that different instruments provide the same answer, and (v) provide out

of sample evidence explaining differences between OLS and IV estimates. This said, no IV strategy can be entirely full-proof since instrument validity relies on the absence of a correlation with an unobserved term as shown by equation (17). Despite their limitations, IV strategies are likely to remain an important part of the toolkit for the analysts of the growth of cities. Natural experiments and discontinuities are scarce and the context in which they take place is often very specific.[11]

Turning to the results of Duranton and Turner (2012), they find that a 10% increase in a city's stock of interstate highways in 1983 causes the city's employment to increase by about 1.5% over the course of the following 20 years when using IV, compared with about 0.6% when using OLS. The higher coefficient on the roadway with IV is consistent with the institutional context in which interstate highways are built in the United States. There is a funding formula that equalizes funding per capita and thus gives fewer roads to denser and fast-growing places where land is more expensive. In addition, this formula is not universally applied and many road projects are make-work subsidies for poorly performing places. Duranton and Turner (2012) provide evidence to that effect.

Note also that the estimated 0.15 elasticity of city employment with respect to the roadway is not directly comparable to the unit elasticity of city population with respect to transportation costs. This is because there is no proportional relationship between highways and commuting costs. The chief reason is that more roads beget more traffic, as shown by Duranton and Turner (2011) in a companion paper. As a result, the speed of travel declines only little when more roads are provided. In turn, this suggests that the proper estimation of (15) requires knowing more about the relationship between roads, traffic and speed of travel. This also calls for a more detailed modeling of the commuting technology. In addition, Duranton and Turner (2012) estimate that the adjustment of population to increased road provision is slow at the metropolitan level.

A more meaningful comparison is with other drivers of city growth. The elasticity of city growth with respect to roads estimated by Duranton and Turner (2012) implies that a one standard deviation of 1983 interstate highways translates into two-thirds of a standard deviation in city growth. This is comparable to the effect of one standard deviation in January temperatures found by Rappaport (2007) in his analysis of population displacement in the United States towards nicer weather. It is also slightly larger than the effect of one standard deviation in the initial stock of university graduates found by Glaeser and Saiz (2004). We discuss the role of amenities and human capital at greater length below.

2.3 Commuting infrastructure and land use

We have just seen that, following a decline in unit commuting costs, cities should experience an influx of population. To accommodate this larger population, cities physically expand outwards and also experience rising densities. Of these two channels, outwards expansion is more important. To see this, consider any arbitrary point x_C, and think of the segment of the city between the CBD and x_C as the historical central city, and the segment between x_C and the city edge $\overline{x}$ as the suburbs. Let

[11] Greenstone, Hornbeck, and Moretti (2010) and Holmes (1998) are key examples of the use of, respectively, quasi-experimental evidence and discontinuities in this area of research, although neither focuses on the effect of transport improvements that we discuss in this section.

$N_C = \int_0^{x_C} d(x)\mathrm{d}x$ denote the (endogenous) population of the central city. Then, using equations (10) and (11), we can calculate the share of population in the central city as

$$\frac{N_C}{N} = \frac{R(0) - R(x_C)}{R(0) - \underline{R}} \ . \tag{18}$$

A reduction in τ increases land rent at any given point beyond the CBD including x_C, but it does not affect land rent $R(0)$ at the CBD (where there is no need to commute and migration keeps utility unchanged) nor land rent at the city edge, which is fixed at $\underline{R}$. Then equation (18) implies that the share of population in the central city falls when commuting costs are reduced. This reduction in $\frac{N_C}{N}$ is shown graphically in figure 1 to the left of the vertical axis, based on equation (18). This has important implications for the analysis of suburbanization, since it implies that improvements in local transportation foster the suburbanization of population.

The positive relationship between roads and suburbanization implied by the monocentric city model is explored in Baum-Snow's (2007) pioneering work on US cities.[12] His main specification is of the following form:

$$\Delta \log N_{C(i)} = \beta_0 - \beta_1 \Delta\tau_i + \beta_2 \, x_{C(i)} + \beta_3 \Delta \log N_i + X_i \beta_4^{'} + \epsilon_i \ , \tag{19}$$

where the dependent variable is the change in log central city population between 1950 and 1990. His measure of commuting, $\Delta\tau_i$, is the change in the number of rays of interstate highways that converge towards the central city. The specification controls for the change in log population for the entire metropolitan area $\Delta \log N_i$ and the radius of the central city $x_{C(i)}$.

The key identification challenge is that rays of interstate highways going to the central city may not have caused suburbanization but instead accompanied it. Baum-Snow's (2007) innovative identification strategy relies on using the 1947 map of planned interstate highways. Planned rays of interstate highways are a strong predictor of rays that were actually built. As already argued, the 1947 highway plan was not developed with suburbanization in mind but aimed instead at linking cities between them. Finally, Baum-Snow (2007) also controls for a number variables such as changes in log income or changes in the distribution of income which could drive suburbanization and be associated with the assignment of interstate highways.

The main finding of Baum-Snow (2007) is that an extra ray of interstate highways leads to a decline in central city population of about 9 percent. This IV estimate is larger than its OLS counterpart, perhaps because more highways were built in cities that suburbanized less. This finding is confirmed when estimating the effect of highways using a panel of shorter first differences and city fixed effects.

More puzzling in light of the monocentric model is the fact that central cities experienced not only a relative decline but also an absolute decline in their population. Over 1950–1990, the population of central cities fell by an average 17 percent while total metropolitan area population rose by 72 percent. This evolution could be explained by a concomitant increase in incomes in

[12] Baum-Snow (2007) motivates his specification verbally with a closed-city (i.e., constant population) version of the monocentric city model. With constant population in the city, when a fall in commuting costs flattens the land and house price gradients, each resident consumes more housing and land. This expands the city boundary outwards and also (unlike in the open-city version of the model with endogenous population) reduces density. Suburbanization then follows from the relocation of some former central city residents to the suburbs.

the United States leading residents to consume more housing. In the monocentric city model, it follows from equation (12) that an increase in the wage w that affects all cities equally leaves their populations unchanged. By equation (11), land rent at the CBD is also unchanged. The land rent at the city edge must still equal the rent in the best alternative use, $\underline{R}$. If housing is a normal good, the economy-wide increase in w then simply makes the house-price gradient flatter. Differentiating equation (4) with respect to w, yields

$$\frac{\partial^2 P(x)}{\partial w \partial x} = \frac{\tau}{(h(x))^2} \frac{\partial h(x)}{\partial w} > 0 \tag{20}$$

Residents each consume more housing and this leads to a reduction in central city population (population in $x \in [0, x_c]$).

Other explanations for the decline of central cities in the United States have focused on a variety of social and material ills that have afflicted central cities such as crime (Cullen and Levitt, 1999), the degradation of the housing stock (Brueckner and Rosenthal, 2009), racial preferences (Boustan, 2010), and related changes in the school system (Baum-Snow and Lutz, 2011).[13]

The suburbanization of population is one of several phenomena that has been associated with urban sprawl. Another key dimension of sprawl is the scatteredness of development, i.e., how much undeveloped land is left between buildings. Burchfield, Overman, Puga, and Turner (2006) merge data based on high-altitude photographs from around 1976 with data based on satellite images from 1992 to track development on a grid of 8.7 billion 30×30 meter cells covering the United States. For each metropolitan area, they compute an index of sprawl measuring the percentage of undeveloped land in the square kilometer surrounding the average residential development. Burchfield, Overman, Puga, and Turner (2006) show that US metropolitan areas differ widely in terms of how scattered development is in each one of them, but for most individual areas the scatterredness of development has been very persistent over time. Among various factors that could potentially affect sprawl, they look at transportation. They find that a denser road network in the suburbs is not associated with more scattered development. At the same time, the car-friendliness of the city centre does matter. Cities that were originally built around public transportation (proxied by streetcar passengers per capita circa 1900) tend to be substantially more compact, even in terms of their recent development, than cities built from the start around the automobile. Other factors that lead to more scattered urban development include ground water availability, temperate climate, rugged terrain, specialization in spatially decentralized sectors, a high-variance over time in decade-to-decade local population growth, having large parts of the suburbs not incorporated into municipalities, and financing a lower fraction of local public services through local taxes.

[13] Existing evidence points at black in-migration followed by white flight and crime as being the two main factors. The race explanation is specific to the United States, and this may explain why it has experienced greater central city decline than other developed countries. These factors are, of course, in addition to the uniquely important role played by the car in the United States.

3. Housing

Our modeling of housing so far misses two key features that matter enormously in reality: the supply of housing is only imperfectly elastic and housing is durable. In themselves, these two characteristics do not determine whether a city will grow or decline. They will however determine how cities will react to positive and negative shocks.

To model the effects of imperfectly elastic housing supply and housing durability, we thus first need to enrich our model by incorporating an elastic demand for labour that helps determine the wage. We can then study how imperfectly elastic housing supply affects a city's population, wages, and house prices following a labour demand shock. Note that this extension to our model is of independent interest since it also allows us to study the effects of changes in labour demand on city growth. We return to this later in this chapter.

3.1 Housing supply restrictions

Suppose labour demand in each city depends negatively on wages w and positively on a local productivity shifter B_i, with i used to index cities. With a constant unit labour supply per worker, local labour supply is simply given by the local labour force N_i. We can then characterize the labour market equilibrium by a wage function:

$$w_i = w(B_i, N_i) \ , \tag{21}$$

with $\frac{\partial w_i}{\partial B_i} > 0$ and $\frac{\partial w_i}{\partial N_i} < 0$. Consider a positive shock to local productivity in a city, i.e., an exogenous increase in B_i in some city i. In the short run (where we take city workforce, and hence labour supply, to be fixed), such an increase in the demand for labor leads to higher wages since $\frac{\partial w_i}{\partial B_i} > 0$. The long-run consequences, however, depend on land and housing supply, which help determine the evolution of N_i.

In the standard monocentric city model, as developed above, the construction industry can develop as much land as necessary at the price of land $\underline{R}$ determined by its best alternative use. The construction industry can also redevelop already developed areas by increasing or decreasing the density of development.

Then, when a positive productivity shock increases the wage, this makes the city relatively more attractive and causes its population to grow. Substituting equation (21) into the spatial equilibrium condition of equation (13) and applying the implicit function theorem directly implies:

$$\frac{\mathrm{d}N_i}{\mathrm{d}B_i} = -\frac{\frac{\partial v}{\partial w_i}\frac{\partial w_i}{\partial B_i}}{\frac{\partial v}{\partial w_i}\frac{\partial w_i}{\partial N_i} + \frac{\partial v}{\partial P(x)}\frac{\partial c(R(x))}{\partial R(x)}\tau} > 0 \ . \tag{22}$$

To sign this derivative, recall from the statement of equation (1) that $\frac{\partial v}{\partial P(x)} < 0$ and $\frac{\partial v}{\partial w_i} > 0$; recall also that the envelope theorem, as used to simplify equation (8), implies that $\frac{\partial c}{\partial R(x)} = \frac{1}{f(x)} > 0$; and we have just stated that $\frac{\partial w_i}{\partial B_i} > 0$ and $\frac{\partial w_i}{\partial N_i} < 0$. This implies $\frac{\mathrm{d}N_i}{\mathrm{d}B_i} > 0$.

To house this larger population, new dwellings must be built, which requires an increase in land and house prices everywhere to make it worthwhile for the construction industry to outbid

alternative uses such as agriculture at the expanded urban fringe:

$$\begin{aligned}\frac{\mathrm{d}P(x)}{\mathrm{d}B_i} &= -\frac{\frac{\partial v}{\partial w_i}\left(\frac{\partial w_i}{\partial B_i} + \frac{\partial w_i}{\partial N_i}\frac{\mathrm{d}N_i}{\mathrm{d}B_i}\right)}{\frac{\partial v}{\partial P(x)}} \\ &= \tau\frac{\partial c(R(x))}{\partial R(x)}\frac{\mathrm{d}N_i}{\mathrm{d}B_i} > 0\,.\end{aligned} \tag{23}$$

The first line of equation (23) follows from substituting equation (21) into equation (2) and applying the implicit function theorem. Note that the short-run wage rise resulting from a positive productivity shock ($\frac{\partial w_i}{\partial B_i}$) is mitigated by the population growth that it triggers ($\frac{\partial w_i}{\partial N_i}$). This also dampens the increase in house prices. However, since population in the city grows following the positive productivity shock, the overall effect on house prices must still be an increase. This is the implication of the second line of equation (23), which is obtained by substituting equation (21) into (13), totally differentiating with respect to B_i, and using the resulting equation to substitute the right-hand side expression on the first line of (23). Local inhabitants react to higher house prices by choosing to live in smaller dwellings at any given distance from the CBD.[14]

In reality, the supply of land is not completely elastic, as assumed so far. It is limited both by geographical constraints and by land-use regulations.[15] This has important implications for the growth of cities. As a benchmark, consider the case where the supply of land is completely inelastic. For instance, a city could reach a green belt at its edge and the edge of the city would become fixed at $\overline{x}$. The spatial equilibrium condition of equation (13) is then replaced by

$$v(c(R(\overline{x}) + \tau N), w) = \overline{v}\,, \tag{24}$$

where land rent at the city edge is now strictly greater than the agricultural land rent: $R(\overline{x}) > \underline{R}$. A positive productivity shock still increases the wage and makes the city relatively more attractive. However, with an inelastic land supply the only way to house a larger population is through an increase in density. Compared with the case of an elastic supply of land, the green belt causes land rents to be higher everywhere in the city, from the fixed edge $\overline{x}$ to the CBD, which makes population grow by less. This comparison shows that land-use regulations affect the extent to which a positive shock that makes a city more attractive translates into higher house prices or more population.

The above comparative statics, by showing that the effect of B_i on city growth is mediated by restrictions on the supply of developable land, provide useful guidance for empirical work. They highlight that measures of the stringency and restrictiveness of land-use regulations cannot be used directly as explanatory variables in a city growth regression. Instead, the stringency of land-use regulations should be interacted with predictors of city growth. In cities that are predicted to grow, we expect strong population growth when land-use regulations are lax, and strong wage and housing price growth when they are stringent. In their analysis of US metropolitan areas

[14] A cross-sectional implication of these comparative statics is that high-productivity cities will tend to be larger in population, have higher density, pay higher wages, and have more expensive houses.

[15] In theory, cities could use land-use regulations to increase the supply of housing, for instance through densification schemes. In practice, land-use regulations in developed countries and many developing countries are, in most cases, geared towards restricting housing supply through, among others, minimum lot size regulations, maximum building height, green belts, and lengthy and cumbersome approval processes.

between 1980 and 2000, Glaeser, Gyourko, and Saks (2006) use two robust predictors of city growth to demonstrate this process. Since human capital is strongly correlated with city growth during this period, they use the initial share of the local population with a bachelor's degree as their first predictor. The second predictor is an index that exploits the idea that the sectoral composition of cities in an important determinant of the evolution of their labour demand, and sectors expand and contract differently as a result of largely national factors. As first suggested by Bartik (1991), cities with a high share of employment in sectors with high growth nationally are thus expected to grow faster in population. For a given predicted employment growth, Glaeser, Gyourko, and Saks (2006) show that highly regulated cities experienced lower population growth rate, higher income growth, and higher growth in housing prices.

This type of analysis raises again an inference problem due to the potential endogeneity of land-use regulation.[16] When trying to explain population growth in cities, we expect land-use regulations to be more stringent in what would otherwise be fast-growing cities because current land owners may lobby hard for stricter regulations when they expect housing prices to appreciate.[17] When trying to explain wage growth in cities, we also expect wages to rise faster in more regulated cities since only households with high wages may be able to afford more expensive houses in these cities. All this casts doubts on the direction of causality in the findings reported above.

While a complete disentangling of the stringency of land-use regulations, population growth, and wage growth in cities has escaped the literature so far, Saiz (2010) offers interesting findings about the exogeneity of land-use regulations.[18] Most importantly, land-use regulations are more stringent in cities where there is less 'usable' land. Usable land is defined as all land not covered by water or with a steep slope. By that measure, cities like San Francisco and Miami have very little usable land whereas cities like Atlanta and Columbus are largely unconstrained in their development. Saiz (2010) shows a strong link between these natural constraints and stringent land-use regulations. This suggests that, ultimately, the limits on city growth imposed by land-use regulations are geographical limits magnified by human interventions.

3.2 Housing durability

The durability of housing has important implications for city growth since people can move out of a city whereas houses cannot.

[16]There is also an important issue of how to measure regulation. There are three main approaches. The first is to estimate a wedge between property prices and construction costs as, among many others, Glaeser, Gyourko, and Saks (2005). This method has the obvious drawback of doing no more than putting a name on a residual. The second is to precisely document some key regulations at a high level of geographical resolution like Glaeser and Ward (2009). The difficulty of this exercise makes it difficult to go much beyond one metropolitan area which the analyst knows extremely well. The third approach tracks land-use regulations across a wide range of jurisdictions and aggregates them into one aggregate index such as the the Wharton Residential Land Use Regulatory Index (Gyourko, Saiz, and Summers, 2008). The drawbacks here are the possible heterogeneity in the data (e.g., a 'sanctuarized' greenbelt in one city may not be the same thing as a slow moving greenbelt in another) and aggregation biases.

[17]On the other hand, employers may lobby for laxer regulations when they expect their activities to expand.

[18]Given the difficulty of the exercise, a better identification of zoning issues is likely to use good restrictions coming from plausible theories of housing regulations. See Fischel (2000), Ortalo-Magné and Prat (2010), or Hilber and Robert-Nicoud (2013) for recent contributions.

When a city experiences a positive shock, as we saw in section 3.1, more workers are attracted to it and additional housing is built. On the other hand, when a city experiences a negative shock and some workers leave, existing housing is not destroyed. More specifically, if housing is durable, its supply will be kinked — with a steep slope below its current equilibrium level and a flatter slope above this level. This suggests an interesting asymmetry between city growth and city decline. When cities grow, they experience moderate house price increases and large population changes. When cities contract, they experience large house price drops and small population changes.

Glaeser and Gyourko (2005) document this asymmetry between urban growth and urban decline using data for 321 US cities for each decade between 1970 and 2000. This asymmetry also holds for the more recent past. In the United States, according to the US census, 17 metropolitan areas, including Las Vegas (NV) and Raleigh (NC), enjoyed a population growth more than 20 percentage points above the mean of 10.7% between 2000 and 2010. On the other hand, New Orleans was the only city which declined by more than 10% during the same period.[19] Even Youngstown (OH) and Johnstown (PA) —- which come just before New Orleans at the bottom of the growth ranking for 2000-2010 — did not decline by more than 6 percent over a decade.

Housing is durable but not permanent. It depreciates slowly over time. This suggests another step to the argument above. After a negative shock, some households leave, and housing prices decline, which induces many to stay. Then, over time, the housing stock depreciates and housing supply declines. Since house prices, that is, the market values of properties, may be well below their construction costs, houses that depreciate are not likely to be refurbished. Households will thus slowly leave the city as the housing stock slowly depreciates. Put differently, housing decline is expected to be persistent. Indeed, urban decline one decade is a strong predictor of urban decline the following decade whereas city growth one decade is a less strong predictor of city growth for the following decade (Glaeser and Gyourko, 2005).

Glaeser and Gyourko (2005) also argue that those who stay in declining cities because of low housing prices are likely to be those with the lowest labor market opportunities in case of out-migration. They provide evidence that declines in population are associated with declines in human capital in their sample of US cities.

4. Urban amenities

Following the work of Rosen (1979) and Roback (1982), urban economists have paid great attention to the role of amenities in attracting people to cities. If cities differ in terms of their amenity level, the spatial equilibrium condition (13) must be taken up one level. Substituting equation (12) into the initial utility function (1), and indexing cities by i, we can write this more general version of the spatial equilibrium condition as:

$$U(A_i, v(c(\underline{R} + \tau N_i), w)) = \overline{U} \, . \tag{25}$$

Recall from the presentation of equation (1) that $\frac{\partial U}{\partial A_i} > 0$ and $\frac{\partial U}{\partial v} > 0$. Recall also from the derivation of equation (22) that $\frac{\partial v}{\partial N_i} = \frac{\partial v}{\partial P(x)} \frac{\partial c(R(x))}{\partial R(x)} \tau < 0$. Applying the implicit function theorem

[19] Adding to this, the decline of New Orleans was due an extremely rare weather event that caused the sudden and massive destruction of its housing stock.

to (25) directly implies:

$$\frac{\mathrm{d}N_i}{\mathrm{d}A_i} = -\frac{\frac{\partial U}{\partial A_i}}{\frac{\partial U}{\partial v}\frac{\partial v}{\partial N_i}} > 0\,. \tag{26}$$

This suggests a first obvious channel through which amenities can affect urban growth: cities where amenities improve become relatively more attractive and grow in population, while cities where amenities deteriorate lose population. These changes in the supply of amenities are sometimes the result of local improvements. One can think of the cleaning and rejuvenation of old historical downtowns, particularly in Europe. Other instances of local changes in the supply of amenities are the result of some economy-wide shock that affects cities heterogeneously. For example, the invention of air-conditioning has reduced the disamenity of extremely hot summer weather in cities in the Southern United States.

There are two other possibilities which are less well understood but, possibly, at least as relevant empirically. First, some demographic changes might be at play. Cities with nice downtowns may be particularly appealing to childless educated workers in their twenties or early thirties whereas cities with mild winters may be particularly attractive to pensioners. These two groups have grown substantially in size and so did these two types of cities.

Second, aggregate economic growth increases wages. If amenities complement other goods, higher wages lead to an increased appeal of high-amenity cities. In this context, migration to high amenity cities is a consequence of economic growth raising the demand for amenities, not of changes in the supply of amenities.

To understand this argument in greater depth, let us return to the spatial equilibrium condition described by equation (25). Since we are now considering an economy-wide increase in the wage, we can not treat the common utility level $\overline{U}$ as a constant; instead, it will change equally in all cities. Totally differentiating (25) with respect to w yields

$$\frac{\partial U}{\partial v}\left(\frac{\partial v}{\partial N_i}\frac{\mathrm{d}N_i}{\mathrm{d}w} + \frac{\partial v}{\partial w}\right) = \frac{\mathrm{d}\overline{U}}{\mathrm{d}w}\,. \tag{27}$$

Totally differentiating (27) with respect to A_i results in

$$\frac{\partial^2 U}{\partial A_i \partial v}\left(\frac{\partial v}{\partial N_i}\frac{\mathrm{d}N_i}{\mathrm{d}w} + \frac{\partial v}{\partial w}\right) + \frac{\partial U}{\partial v}\frac{\partial v}{\partial N_i}\frac{\partial^2 N_i}{\partial A_i \partial w} = 0\,. \tag{28}$$

Rearranging implies

$$\frac{\partial^2 N_i}{\partial A_i \partial w} = -\frac{\partial^2 U}{\partial A_i \partial v}\frac{\frac{\partial v}{\partial N_i}\frac{\mathrm{d}N_i}{\mathrm{d}w} + \frac{\partial v}{\partial w_i}}{\frac{\partial U}{\partial v}\frac{\partial v}{\partial N_i}}\,. \tag{29}$$

To sign this expression, note that the numerator of the fraction on the right, $\frac{\partial v}{\partial N_i}\frac{\mathrm{d}N_i}{\mathrm{d}w} + \frac{\partial v}{\partial w} = \frac{\mathrm{d}v}{\mathrm{d}w}$, must be positive. This is because when rising wages cause a movement of population across cities, some cities must lose population while others gain population. In cities that lose population, $\frac{\mathrm{d}v}{\mathrm{d}w} > 0$, and by equation (28) the same must be true in every city to maintain a spatial equilibrium. Stated differently, an economy-wide increase in wages must cause utility to rise everywhere. Since $\frac{\partial U}{\partial v} > 0$ and $\frac{\partial v}{\partial N_i} < 0$, it follows that $\frac{\partial^2 N_i}{\partial A_i \partial w}$ has the same sign as $\frac{\partial^2 U}{\partial A_i \partial v}$. Hence, if utility is supermodular in the level of amenities and the sub-utility derived from housing and other goods ($\frac{\partial^2 U}{\partial A_i \partial v} > 0$), an

economy-wide increase in income makes cities with greater amenities grow in population relative to other cities. Intuitively, if the value that consumers place on additional amenities increases as they are able to afford a better bundle of housing and other goods, aggregate economic growth makes high-amenity cities relatively more attractive.

Both supply and demand channels suggest a link between amenities and urban growth. Taken literally, the supply explanation described by equation (26) suggests regressing changes in population on changes in amenities:

$$\Delta_{t+1,t} \log N_i = \beta_0 + \beta_1 \Delta_{t+1,t} A_i + \epsilon_{it} \, . \tag{30}$$

This regression mirrors regression (14), with the only difference that now the level of amenities replaces (log) commuting costs as the driver of city growth. By the same argument that was applied to regression (14), if the adjustment of population is sluggish after a change in amenities, one is naturally led to estimate instead

$$\Delta_{t+1,t} \log N_i = \beta_0 - \lambda \log N_{it} + \beta_1 A_{it} + \epsilon_{it} \, . \tag{31}$$

This is the counterpart to the transportation regression (15) and, equivalently to that case, the coefficient of interest, β_1, measures the effect of amenities on population in cities and λ the speed of population adjustment.

Turning to the demand-for-amenities explanation, the comparative statics of equation (29) suggest regressing local population changes on local amenities interacted with national wage growth. In practice, interacting national wage growth and amenities is likely to be problematic for several reasons. First, sluggish population adjustment is likely to make extremely difficult the identification of faster population growth for cities with higher amenities during periods when national wage growth is higher. For instance, cyclical downturns, which imply both lower wages and less mobility, are likely to act as a confounding factor. Second, cyclical behaviour and sluggish adjustment also suggest measuring population growth over periods of five or ten years, which limits the potential length of a panel of city growth. To avoid these problems, one may prefer to rely on cross-sectional variation rather than longitudinal variation and check whether, against the background of rising incomes nationwide, high-amenity cities have attracted more people. Note that this leads to regression (31) again.

Since both demand and supply explanations can be used as motivations for the specification of equation (31), estimating such a regression will help identify the overall effect of amenities on urban growth but will not assist us much in disentangling demand and supply explanations. While we return to this issue later, at this stage the most pressing issue is how to measure amenities.

From the spatial equilibrium we can define the shadow price of amenities as

$$Q = h(0) \frac{\mathrm{d}P(0)}{\mathrm{d}A_i} \, , \tag{32}$$

which is the extra housing cost that a resident is willing to pay to live in a city with higher amenities. Note that this is valued at the CBD ($x = 0$) so that commuting costs do not have to be considered separately (utility equalization within the city implies that the same shadow price

applies to other locations within the city with higher commuting costs and lower housing prices). In equation (32), the shadow price of amenities only depends on housing variables (h and P) and the level of amenities (A). In a more general setting where amenities and land enter production as well as consumption, expression (32) also contains a wage term as in Roback (1982) and subsequent literature:

$$Q = h(0)\frac{\mathrm{d}P(0)}{\mathrm{d}A_i} - \frac{\mathrm{d}w}{\mathrm{d}A_i}\,. \tag{33}$$

This wage term reflects that amenities can affect not just land prices but also wages.[20]

There is a long tradition of empirical research motivated by equation (33) that attempts to value amenities by separately regressing housing expenditures and wages in cities on a set of broadly defined amenities (e.g., Blomquist, Berger, and Hoehn, 1988). The amenities considered range from the availability of good restaurants to nice architecture to low crime or richly endowed public libraries. The coefficients on each of these amenities in the housing regression and in the wage regression are used in place of respectively $h(0)\frac{\mathrm{d}P(0)}{\mathrm{d}A_i}$ and $\frac{\mathrm{d}w}{\mathrm{d}A_i}$ in equation (33) to compute a shadow price of each individual amenity. The overall value of amenities in each location can then be assessed by aggregating its bundle of amenities valued at its estimated shadow price. Using this approach to estimate the impact of amenities on city growth is problematic. Most of the amenities that are usually considered are likely to be endogenous to city growth. For instance whether good restaurants cause city growth or result from it is unclear. Then, aggregating an arbitrary number of poorly identified coefficients is unlikely to be informative about the effects of amenities on city growth. In fact, this approach often provides quality-of-life rankings that seem hard to reconcile with commonly accepted notions of attractiveness.[21]

The main advantage of the standard approach building on Roback (1982) is that it allows to see whether the main effect of amenities is to raise the utility of residents or to provide a productivity advantage to firms. If amenities mainly raise the utility of residents, these will be willing to accept higher rents or lower wages in order to enjoy them ($h(0)\frac{\mathrm{d}P(0)}{\mathrm{d}A_i} > 0$ and $\frac{\mathrm{d}w}{\mathrm{d}A_i} < 0$ in equation 33). If amenities mainly provide a productivity advantage, firms will be willing to incur higher building costs or higher wage costs to locate where they can benefit from them ($h(0)\frac{\mathrm{d}P(0)}{\mathrm{d}A_i} > 0$ and $\frac{\mathrm{d}w}{\mathrm{d}A_i} > 0$).

This approach is useful to study not just amenities but also other city characteristics. For instance, Ottaviano and Peri (2006) study the extent to which a greater diversity of countries of origin among residents of US cities is associated with productivity advantages or consumption amenity advantages. Following Altonji and Card (1991) and Card (2001), they instrument the diversity of each city by combining historical stocks of immigrants by origin at the local level with immigration flows by origin at the national level (under the assumption that recent immigrant flows sort across cities proportionately to historical stocks of the same origin). They find that cities with a greater diversity of countries of origin have both higher wages and higher land rents and

[20] There are two different channels. First, amenities may impact productivity directly. For instance a coastal location may lower trade costs while being enjoyed as a consumption amenity. Second, consumption amenities can affect wages indirectly when land is a factor of production, since higher amenities imply higher land prices. Then firms substitute away from land in production, which lowers the marginal product of labour. In addition, Moretti (2011) shows that with imperfectly mobile workers, the expression that values amenities should also contain a term to reflect imperfect mobility. The estimation of this mobility term is an open challenge.

[21] For instance, Blomquist, Berger, and Hoehn (1988) rank Pueblo (CO), a county in Macon (GA), and one in Binghamton (NY) as three of the most desirable places to live in the United States whereas New York City is close to the bottom.

conclude that it is the higher productivity effect of diversity that dominates. Given the discussion above, this approach is most useful when applied to study a single well-defined amenity or city characteristic that is either exogenous or appropriately instrumented.

To solve the problems of mixing heterogenous amenities many of which are endogenous, much of recent research on urban amenities has focused on the weather. That weather variables should be valued highly by consumers is needed for them to play an important role in location decisions and more specifically in the growth of cities. Reassuringly, the literature that values amenities usually estimates high shadow prices for climate-related variables.[22] The weather is also often deemed to be 'exogenous'. Although most manifestations of the weather are not a consequence of city growth, some caution here is nonetheless needed since most measures of weather are likely to be correlated to other determinants of urban growth. This suggests enriching regression (31) with a number of control variables and assessing the robustness of the estimated weather coefficients against the inclusion of these controls.

As argued by Glaeser, Kolko, and Saiz (2001), weather — as measured by January and July temperatures — is one of the most reliable predictors of city growth in recent US history. Warmer temperatures in January and cooler temperatures in July are both strongly associated with city growth. These findings are confirmed and greatly extended in Rappaport's (2007) comprehensive study. His main conclusion is that nice weather (in the form of mild winters and summers that are not too hot) is a major engine of population growth for US counties between 1970 and 2000. More specifically, he shows that a standard deviation in January temperature is associated with a 0.6 standard deviation in population growth. For July temperature, one standard deviation is associated with a 0.2 standard deviation in population growth. For European countries, Cheshire and Magrini (2006) also reach similar conclusions.

This said, in the United States the correlation between summer and winter temperatures and city growth reflects to a large extent the rise of Southern cities. As pointed by Glaeser and Tobio (2008), Southern cities which offer milder winters and warmer summers also differ from other US cities in the evolution of their wages and housing costs. These are potentially two important missing variables in regression (31), since both housing costs and wages affect the spatial equilibrium condition (25). Importantly, housing in Southern cities appears to have become relatively cheaper. This obviously raises some doubts about the importance of the weather as a key driver of city growth.[23]

In this respect, Rappaport (2007) makes important observation that the effects of nice weather in the United States are also observed outside of the South for areas with mild summers. In addition, for these areas the development of air-conditioning made little difference, if any. This result is more immediately consistent with explanations that rely on a rising demand for amenities than

[22] This is true of Blomquist, Berger, and Hoehn (1988), Albouy (2008), and many others.

[23] Pushing the logic of the Roback (1982) model, Glaeser and Tobio (2008) find that the relative decline in the costs of housing and rising productivity in the US South imply no increase in the willingness to pay for Southern amenities (actually these two features imply a decline).

those that highlight supply changes.[24]

Another direction taken by recent research is to look for a 'summary variable' that would proxy for the entire bundle of amenities in a city. The first possibility, suggested by Glaeser, Kolko, and Saiz (2001), is to estimate the aggregate value of amenities in a city relative to another city or to the average city as the sum of the difference in housing costs minus the difference in wages. This builds again on the spatial equilibrium condition for workers, which implies that differences in real wages (i.e., nominal wages corrected of housing costs) should be offset by differences in amenities. Albouy (2008) implements this strategy empirically, while also correcting for differences in non-labour income, in federal taxes, and in the price of goods other than housing. He obtains an aggregate amenity value for each city that better corresponds to perceived notions of attractiveness. He then regresses this aggregate value on a number of individual amenity variables to study the relative importance of each. This approach seems promising. As Carlino and Saiz (2008) note, however, it still is subject to the concern that current property prices also partly reflect expectations about future population growth. Then studying the effects of amenities on urban growth by regressing population growth on an 'amenity index' that contains expectations of population growth is potentially problematic.

As another summary variables capturing a large set of amenities, Carlino and Saiz (2008) propose using the number of leisure visits to each city. They first show that leisure visits, as collected by a consultancy in the tourism industry, correlate well with alternative measures of amenities and quality of life, including Albouy's (2008). Second, they regress population growth between 1990 and 2000 for US metropolitan areas on leisure visits and find that the elasticity of population with respect to leisure visits is about 2 percent over this ten-year period. This coefficient is robust to the inclusion of many other control variables. This said, we can again imagine a number of ways leisure visits might be correlated with city growth without having a causal effect on the latter. Tourism is itself a strong growth industry. However the correlations are robust to the exclusion of the likes of Las Vegas and Orlando. In addition, fast-growing cities receive a greater inflow of newcomers who, in turn, may receive more visits from family and friends. We can again use an instrumental variable approach to circumvent this simultaneity problem. Carlino and Saiz (2008) use two exogenous determinants of leisure visits: the number of historic places and the coastal share within a ten kilometre radius of the central city. This instrumental variable approach leads to an even higher elasticity of city population with respect to amenities of 4 percent.

5. Agglomeration economies

The monocentric city model, by focusing on the trade-off between commuting costs and house prices within a single city, highlights the costs of bigger cities. To study meaningfully multiple cities within an urban system, we need to consider also the productive benefits of bigger cities.

[24] Matters are actually even more complicated than this because amenities and land-use regulations appear to interact in some interesting fashion. Gyourko, Mayer, and Sinai (2013) show that some cities with good amenities such as San Francisco, Santa Cruz, or Boston have imposed ever more restrictive zoning regulations. As a result, population growth has been limited but property price appreciation has been extremely strong. This has also led to the sorting of high-income workers in these "superstar cities."

For simplicity, let us abstract from amenity differences and refer back to the spatial equilibrium condition of equation (13). If we treat the wage w_i as a parameter independent of a city's population, then $\frac{\mathrm{d}v}{\mathrm{d}N_i} < 0$, so that any individual prefers to live alone than to live in a city of any size. Even if we endogenise the wage but have $\frac{\mathrm{d}w}{\mathrm{d}N_i} < 0$, as in section 3.1, it remains that $\frac{\mathrm{d}v}{\mathrm{d}N_i} < 0$. Stated differently, if new potential sites for cities are available, then agglomeration economies are essential to understand why cities exist at all.

A simple way to incorporate agglomeration economies into the monocentric city model is to recognize that the wage in each city i depends positively on its population: $w_i = w(N_i)$ with $\frac{\mathrm{d}w_i}{\mathrm{d}N_i} > 0$. Many urban models have this feature and there is broad empirical evidence supporting it, as discussed below. Now, as a city's population increases there is both the negative effect on residents' utility of rising urban costs ($\frac{\mathrm{d}v}{\mathrm{d}P(0)}\frac{\mathrm{d}P(0)}{\mathrm{d}N_i} < 0$) and the positive effect of stronger agglomeration economies ($\frac{\mathrm{d}v}{\mathrm{d}w_i}\frac{\mathrm{d}w_i}{\mathrm{d}N_i} > 0$). The fact that in reality there is no one extremely large city but instead multiple cities of finite population size suggests that v is a concave and non-monotonic function of N_i. Initially, agglomeration economies dominate and utility increases with city size. Eventually, higher costs of housing and commuting dominate and utility decreases with city size.

5.1 City formation and urban systems

We now develop a simple model of a system of cities in the tradition of Henderson (1974). Before turning to city creation, we begin by deriving an expression for $w(N_i)$ built on explicit microfoundations, following Abdel-Rahman and Fujita (1990). Suppose there are multiple perfectly competitive final sectors, identified by superindex j, each of which produces a homogenous final good that is freely tradable across cities. Final production technology features a constant elasticity of substitution across intermediate inputs that are sector-specific and non-tradable across cities, so that output in sector j in city i is

$$Y_i^j = B^j \left\{ \int_0^{m_i^j} \left[y_i^j(h) \right]^{\frac{1}{1+\sigma^j}} \mathrm{d}h \right\}^{1+\sigma^j}, \tag{34}$$

where h indexes intermediate varieties, $y_i^j(h)$ denotes intermediate input quantities, m_i^j denotes the endogenous mass of intermediates available in sector j in city i, B^j is a measure of technological development that will be useful for comparative statics, and $0 < \sigma_j < 1$. As in Ethier (1982), intermediates are produced by monopolistically competitive firms à la Dixit and Stiglitz (1977) with technology

$$y_i^j(h) = \beta^j l_i^j(h) - \alpha^j, \tag{35}$$

where $l_i^j(h)$ is firm-level employment. Workers are freely mobile across sectors as well as across locations.

In equilibrium, all firms in any given city and sector set the same profit-maximising price $q_i^j = w_i^j(1+\sigma^j)/\beta^j$. Free entry drives intermediate profits to zero: $q_i^j y_i^j - w_i^j l_i^j = 0$. Using equation (35) and the pricing rule to expand this expression and solving for y_i^j shows there is a fixed level of intermediate output consistent with zero profits in each sector:

$$y_i^j = \frac{\alpha^j}{\sigma^j}. \tag{36}$$

Equating (35) and (36) allows solving for the constant workforce of each intermediate supplier: $l_i^j = \alpha^j(1+\sigma^j)/(\beta^j\sigma^j)$. Hence, the equilibrium mass of intermediate producers in sector j of city i is

$$m_i^j = \frac{N_i^j}{l_i^j} = \frac{\beta^j\sigma^j}{\alpha^j(1+\sigma^j)}\, N_i^j \,. \tag{37}$$

By choice of units for intermediate output, we can set $\beta^j = (1+\sigma^j)(\alpha^j/\sigma^j)^{\sigma^j/(1+\sigma^j)}$. Substituting equations (36) and (37) into (34) yields aggregate production in sector j of city i as

$$Y_i^j = B^j \left[\left(y_i^j\right)^{\frac{1}{1+\sigma^j}} m_i^j\right]^{1+\sigma^j} = B^j \left(N_i^j\right)^{1+\sigma^j} \,. \tag{38}$$

Zero profits in final production imply that $w_i^j N_i^j = P^j Y_i^j$. Thus, wages are given by

$$w_i^j = P^j\, B^j \left(N_i^j\right)^{\sigma^j} \,. \tag{39}$$

Note that aggregate production is subject to increasing returns at the sector and city level. As the size of a sector in a city increases, it supports a wider range of shared intermediate suppliers. Gains from variety in final production then imply that it is possible to increase output more than proportionately relative to the increase in employment producing intermediates for this sector. The literature has explored many alternative agglomeration mechanisms with similar implications, both theoretically and empirically (see Duranton and Puga, 2004, Rosenthal and Strange, 2004, and Puga, 2010, for reviews).

In equilibrium, all cities are specialized in a single sector. To see this, note that any equilibrium must be such that wages are equalized across sectors in a city. Consider now a small perturbation in the distribution of employment across sectors within a city keeping its total population constant. Since $\partial w_i^j/\partial N_i^j > 0$, sectors that see employment rise begin paying higher wages and attract more workers, whereas sectors that see employment fall begin paying lower wages and lose more workers. The only equilibrium that is stable with respect to perturbations in the distribution of local employment across sectors has all local workers employed by the same sector. Two assumptions drive full urban specialization in this simple model. First, since intermediates are sector-specific, agglomeration economies arise within sector only. Thus, mixing multiple sectors in a single city would increase house prices and commuting costs without bringing any benefits relative to having sectors operate in separate cities. Second, final goods are assumed freely tradable, which eliminates the proximity-concentration trade-off that would otherwise arise. Cross-sector externalities and trade costs would provide static motives for a diversity of sectors in cities. See Duranton and Puga (2000) for a review of such static extensions. Below we also discuss a dynamic alternative proposed by Duranton and Puga (2001).

Next we model the internal structure of cities using a version of the monocentric city model that both generalizes and simplifies the version presented in section 2. In particular, let us generalize the specification for commuting costs so that they are not necessarily linear but instead have an elasticity γ with respect to distance. Further, we have so far had commuting costs incurred in units of the single consumption good in the economy. Since we are now considering multiple consumption goods, and each city is specialized in the production of just one such good, let us

have commuting costs in each city incurred in terms of the locally produced good j. We can then write commuting costs for a resident living at distance x from the CBD as $P^j \frac{1+\gamma}{\gamma} \tau x^\gamma$, where the normalization constant $\frac{1+\gamma}{\gamma}$ is just meant to simplify notation below.

At the same time, to obtain closed-form solutions without the need to specify a functional form for utility, let us make the simplifying assumption that all residences have the same size and are built with a constant capital to land ratio. Thus, every individual consumes one unit of floorspace built on one unit of land with a fixed amount of capital. Relative to the version of the monocentric city model seen before, this implies the restriction $h(x) = f(x) = 1$. Hence the physical extent of the city is the same as its population: $\overline{x} = N_i$.

Then, totally differentiating the spatial equilibrium condition with respect to x yields the land and house price gradients: $\frac{\mathrm{d}R(x)}{\mathrm{d}x} = \frac{\mathrm{d}P(x)}{\mathrm{d}x} = -P^j(1+\gamma)\tau x^{\gamma-1}$. Without further loss of generality, let us set the cost of capital per residence equal to zero (so that house and land prices are equal instead of differing by a constant) and the rental price of land when not in urban also equal to zero (otherwise, land prices would be higher everywhere by the value of that rent). Integrating the land price gradient $\frac{\mathrm{d}R(x)}{\mathrm{d}x}$ and using $R(\overline{x}) = \underline{R} = 0$ and $\overline{x} = N_i$ to obtain the integration constant, we can express house and land prices as

$$P(x) = R(x) = P^j \frac{1+\gamma}{\gamma} \tau \left(N_i^\gamma - x^\gamma\right) . \tag{40}$$

Note that with fixed housing consumption, utility equalization within the city implies that the sum of commuting costs and housing expenditure is the same for every resident, and equal to the house price at the CBD: $P^j \frac{1+\gamma}{\gamma} \tau x^\gamma + P(x) = P(0) = P^j \frac{1+\gamma}{\gamma} \tau N_i^\gamma$. Integrating $R(x)$, as given by equation (40), over the physical extent of the city yields total land rents:

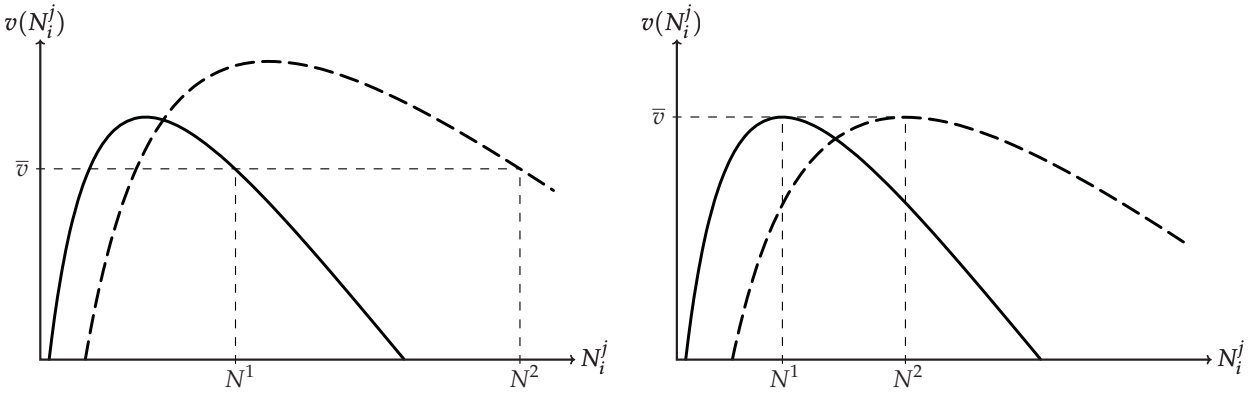
$$R_i = \int_0^{N_i} R(x)\mathrm{d}x = P^j \tau N_i^{1+\gamma} . \tag{41}$$

We now consider endogenous city formation in this framework. Following Becker and Henderson (2000) we study three alternative mechanisms: self-organization, land developers, and active local governments.

Let us begin with self-organization, so that cities arise as the result of the uncoordinated decisions of individual agents. Since free trade in final goods equalizes their prices across locations and since housing consumption is fixed, utility only depends on income net of housing costs and commuting costs. To compute income, we need to consider what happens to land rents. Under self-organization the simplest assumption is that land rents are shared by all residents in the city, each getting $\frac{R_i}{N_i}$. Recall that the sum of housing costs and commuting costs for every resident is equal to the price of housing at the CBD, $P(0)$. Using c_i to denote per-capita income net of housing costs and commuting costs, we can write this as $c_i = w(N_i) + \frac{R_i}{N_i} - P(0)$. Substituting equations (39), (40), and (41), this becomes

$$c_i = P^j \left(B^j N_i^{\sigma^j} - \frac{\tau}{\gamma} N_i^\gamma \right) . \tag{42}$$

We consider a continuum of cities. Under self-organization, the number (mass) of cities in each sector is given, since no agent is large enough to create a new city on their own. An equilibrium distribution of population across cities simply requires utility equalization and stability



Panel (a)
Self-organization

Panel (b)
Land developers

Figure 2: City sizes and utility

with respect to small perturbations. Panel (a) of figure 2 plots utility as a function of population size, as given by equation (42), for two cities specializing in different sectors.[25] Provided that $\gamma > \sigma^j$, for each city type, utility is a concave function of population: initially agglomeration economies dominate and utility increases with city size but eventually higher costs of housing and commuting dominate and utility decreases with city size. The difference in specialization affects the relationship between population and net income through differences in agglomeration economies (σ^j), in the productivity shifter (B^j), and in final goods prices (P^j).

Consider first the solid curve. For any given level of utility, there are at most two population sizes that provide this utility level, one above and one below the efficient size that maximizes utility in a city of that specialization. However, cities below the efficient size cannot survive small perturbations in the distribution of workers. This is because those that gain population get closer to the efficient size and attract even more workers while those that lose population get further away from the efficient size and lose even more workers. If cities specializing in sector 1 (with utility plotted as a solid curve) have the population marked by N^1 in panel (a) of figure 2, then cities specializing in sector 2 (with utility plotted as a dashed curve) have the population marked by N^2 to ensure workers have no incentive to migrate across cities. Thus, under self-organization, all cities of the same specialization have the same population size and this size cannot be smaller than the efficient city size for each sector.

The reason why cities tend to be too large under self-organization is a lack of coordination: each worker would prefer more cities of a smaller size each in their sector but a worker alone cannot create a new city. Following Henderson (1974), suppose all land at each potential its for a city is controlled by a land developer who collects land rents. There is free entry and perfect competition amongst land developers. Each of them announces a population size and specialization for their city as well as a level of transfers T_i^j that they are willing to provide to workers locating in their city. When active, each land developer seeks to maximize land rents $R_i = P^j \tau N_i^{1+\gamma}$, net of any

[25] The figure is plotted for $\gamma = \tau = 0.045$, $\sigma^1 = 0.038$, $\sigma^2 = 0.040$, $B^1 = B^2 = 1$, $P^1 = 1$, and $P^2 = 1.43$.

transfers:

$$\max_{\{T_i,N_i\}} \Pi_i = P^j \tau N_i^{1+\gamma} - T_i N_i \ , \tag{43}$$

subject to the participation constraint for workers. This constraint results from incorporating transfers into workers' income and ensuring that they achieve the same per-capita income net of housing and commuting costs $\overline{c}$ as in the best alternative location:

$$P^j B^j N_i^{\sigma^j} + T_i - P^j \frac{1+\gamma}{\gamma} \tau N_i^{\gamma} = \overline{c} \ , \tag{44}$$

Further, population must be positive: $N_i \geqslant 0$.

Solving for T_i^j in equation (44) and substituting this into (43) yields an equivalent programme:

$$\max_{\{N_i\}} \Pi_i = P^j B^j N_i^{1+\sigma^j} - P^j \frac{\tau}{\gamma} N_i^{1+\gamma} - \overline{c} N_i \ . \tag{45}$$

The equivalence between the programmes (43) and (45) shows that, in maximizing land rents net of transfers, developers behave as if running a factory-town in which they hired workers at their going net compensation levels ($\overline{v}$) and sold in national markets all output produced in the city ($P^j Y_i = P^j B^j N_i^{1+\sigma^j}$) net of commuting cost expenditure ($P^j \frac{\tau}{\gamma} N_i^{1+\gamma}$), keeping any residual as a profit.

The first order condition for (45) is

$$\overline{v} = P^j \left((1+\sigma^j) B^j N_i^{\sigma^j} - \frac{1+\gamma}{\gamma} \tau N_i^{\gamma} \right) \ . \tag{46}$$

Substituting this first-order condition into equation (45) yields maximized profits for the developer:

$$\Pi_i = P^j \tau N_i^{1+\gamma} - \sigma^j P^j B^j N_i^{1+\sigma^j} \ . \tag{47}$$

Equating this expression for Π_i with the expression in the original developer's program of equation (43) shows that, in maximizing their profits, developers offer each worker the following transfer:

$$T_i = \sigma^j P^j B^j N_i^{\sigma^j} \ . \tag{48}$$

This transfer covers the gap between the market wage $w_i^j = P^j B^j N_i^{\sigma^j}$ and the city-level marginal product of labour $(1+\sigma^j) P^j B^j N_i^{\sigma^j}$. Thus, in maximizing their profits, developers internalize the city-level externality created by agglomeration economies.

Free entry and perfect competition among land developers exhausts their profits. Using $\Pi_i = 0$ in equation (47) and solving for N_i, we obtain equilibrium city size in the presence of land developers:[26]

$$N_i = \left(B^j \frac{\sigma^j}{\tau} \right)^{\frac{1}{\gamma - \sigma^j}} \ . \tag{49}$$

This is the optimal city size. To see this, consider a situation where land at each site, instead of being owned by a developer, is shared by residents who elect a local government that runs the city so as to maximize their welfare. The utility of each resident in the city is given by equation (42). This utility is maximized for N_i given by equation (49).

[26] Note that the second-order condition for profit maximization requires $\gamma > \sigma^j$.

Panel (b) of figure 2 plots utility as a function of population size for two cities specializing in different sectors in the presence of competitive land developers.[27] There are two differences with respect to the equilibrium under self-organization of panel (a). First, taking final goods prices as given, land developers are induced through competition to create cities of the efficient size for their sector. If this was not the case, then another developer could enter and make a profit by offering a more efficient city and capturing as profit by means of lower transfers the difference in utility relative to the best alternative city. Optimal city size is obtained because, in equilibrium, developers must make transfers that cover the gap between private and social returns opened by agglomeration economies. With zero profits for developers, total land rents equal total transfers, and thus are just enough to cover that gap.[28] The second difference is that if, as shown by the dashed curve for sector 2 in panel (a), cities in a certain sector offer higher utility at the peak, this will attract more developers to this sector. Entry increases economy-wide output in that sector and lowers its general equilibrium final good price until all cities offer the same utility at the efficient size for their sector. This is what shifts downwards the utility for sector 2 in panel (b) relative to panel (a).

Equilibrium city sizes, as given by equation (49), are the result of what Fujita and Thisse (2002) call the "fundamental trade-off" of urban economics: between agglomeration economies, which make wages and productivity increase with city size, and crowding diseconomies, which makes commuting and housing costs increase with city size.[29] The lower the magnitude of commuting costs, as captured by τ, the larger is city size. This confirms that the implication of the monocentric city model that improvements in commuting infrastructure foster urban growth is robust to the introduction of agglomeration economies and endogenous city creation. A decrease in the elasticity of commuting costs with respect to distance, γ, similarly leads to larger cities. Turning to the other side of the trade-off, stronger agglomeration economies, as measured by σ^j, make cities larger. Since crowding costs are unlikely to be very different for workers engaged in different activities but agglomeration economies will be stronger in some sectors than in others, there is an important link between sectoral specialization and city size. In particular, cities specializing in sectors with higher agglomeration economies (high σ^j) will be larger in size. Black and Henderson (2003) show that US cities can be classified into groups with similar specialization and size.

In the model in this section, sectors are defined as groups of firms using similar bundles of

[27] The curves are plotted with developer profits driven to zero, in which case utility is still given by equation (42), like under self-organization. We use the same parameters as in panel (a), except P^2, which now adjusts through free entry by land developers until utility is equalized across cities.

[28] Using $\Pi_i = 0$ in equation (43) implies $T_i N_i = P^j \tau N_i^{(1+\gamma)} = R_i$, while multiplying both sides of equation (48) by N_i implies $T_i N_i = \sigma^j P^j B^j N_i^{(1+\sigma^j)}$. This is a classic result in urban economics known as the Henry George Theorem (Serck-Hanssen, 1969, Starrett, 1974, Vickrey, 1977). Its best-known version is associated with local public goods (Flatters, Henderson, and Mieszkowski, 1974, Stiglitz, 1977, Arnott and Stiglitz, 1979).

[29] Equation (49) reflects city sizes with developers. Under self-organization, we must be in the downward-sloping portion of the size-utility relationship depicted in figure 2 instead of at the optimum. For given goods prices, a fall in τ or γ or an increase in σ^j or B^j push the size-utility curve outwards and make city size increase, resulting in the same qualitative comparative statics as with developers. However, while with efficient city sizes the envelope theorem implies no additional effects operating through goods prices, without developers we must consider such general equilibrium price effects which, if sufficiently strong, could in principle offset the standard comparative statics. In section 6.4 we discuss further the importance of considering the elasticity of demand and price effects when looking at the impact of productivity changes on cities.

inputs. In a more general setting, the combinations of activities present in cities of different sizes depend on more complicated links, some of which form through agglomeration economies extending across sectors, and others through links between various parts of the production process within a firm. Duranton and Puga (2005) model such a multi-stage production process within each firm in a general equilibrium model of an urban system. They show that as technological developments and transport improvements facilitate the spatial fragmentation of activities within the firm, management and business service provision will tend to concentrate in larger cities whereas actual production will concentrate in smaller cities. They also who that such a process has occurred in the United States since the 1950s. In effect, this implies an increasing specialization by functions and occupations instead of traditional sectoral divisions.

The technological parameter in the model, B^j, allows us to study the effects on city growth of sectoral shocks and of aggregate growth. To study sectoral shocks, consider a continuum of sectors, so that a shock to just one of these sectors does not affect the entire economy. Then, by equation (49), a positive shock to B^j makes cities specializing in this sector grow in size. Higher supply lowers output prices in this sector, which induces some developers to move away from it until utility equalization, which implies that the value of output per worker net of commuting costs must stay constant, is restored (see equation 42). At the new equilibrium, there will be fewer cities specializing in the sector experiencing the positive technology shock and each of them will be larger in population.

Turning to aggregate technical change, consider a situation where $B^j = B$ is common to all sectors and experiences an increase. Given that this will affect every city, we can no longer treat utility as constant. Instead, it will change equally everywhere. By equation (42) and the envelope theorem,

$$\frac{B}{\overline{c}}\frac{\mathrm{d}\overline{c}}{\mathrm{d}B} = \frac{BN_i^{\sigma^j}}{BN_i^{\sigma^j} - \frac{\tau}{\gamma}N_i^{\gamma}} > 1\,. \tag{50}$$

Thus, cities amplify the growth effects of technical progress: per-capita income net of commuting costs increases more than proportionately with aggregate technical change.

By equation (49) aggregate technical also makes cities grow in population, with initially larger cities tending to grow more. Henderson and Wang (2007) suggest that over the last few decades this tendency of aggregate technical change to increase the relative size of then largest cities has been offset by the tendency of increased democratization around the world to lower urban concentration, thus helping keep city size distributions roughly stable. This effect of democratization can also be linked to the systems of cities model presented in this section. As we have seen, city developers or local governments can help cities get closer to their efficient size, while without them cities tend to be too few and too large. For this to be the case, Henderson and Wang (2007) argue that local governments need to be able to set up new cities, to finance new infrastructure so that existing towns can expand into cities, and to enable land development in well-functioning land markets where regulation is transparent and land ownership is clearly defined. Arguably, all these are characteristics that are closely related to more democratic regimes.

The operation of city developers and local governments in the model of this section is purely static. As a result, changes such as the sectoral shocks we have examined or aggregate population

growth cause swings in population sizes. Henderson and Venables (2009) develop a dynamic model of city formation where housing and urban infrastructure are durable. Then population changes smoothly and it is instead the price of housing that is subject to swings. In this dynamic version of the model, cities are created sequentially and city developers borrow to finance development. The subsidies paid by developers are no longer as in equation (48) and such that the total value of the subsidy equals the total value of the externality created by agglomeration economies. Instead the subsidy to the marginal migrant covers the marginal externality she creates. Cuberes (2011) provides empirical evidence of such sequential city growth: in many countries the largest cities grow more initially, but over time their growth tends to settle and smaller cities start growing faster.

Following Bartik (1991), the link between specialization and city size has often been used to predict city growth in multiple contexts. Applications include studying the interactions between land-use regulations and urban growth as described in section 3.1 above. For each city, the "Bartik predictor" takes employment growth by industry at the national level (excluding the city at hand) and averages it across industries using initial local employment shares as weights. This measure is sometimes alleged to provide a measure of city growth that is clean from city-specific shocks.[30]

While the Bartik predictor may be plausibly used as a measure of local labour demand shocks affecting a city, like in Glaeser and Gyourko (2005), using it as an instrument may be more problematic. To see this, consider regressing changes in local output or in local wages on changes in local employment to estimate the agglomeration elasticity σ as suggested by equations (38) or (39). If there are sectoral shocks with some unobserved component, for instance affecting B^j, these will become part of the error term in the regression. By construction, those sectoral shocks will also be part of the Bartik predictor. Thus, the Bartik predictor will violate the exogeneity requirement to be used as an instrument for changes in local employment.

Finally, note that the model above assumes free trade between cities. As noted above, cities fully specialize in equilibrium because of the combination of free trade in final goods and within-sector agglomeration economies. Introducing trade costs brings in the proximity-concentration trade-off that is familiar from international trade (Brainard, 1997). A diversity of sectors in a city reduces the strength of agglomeration economies but saves transport costs when supplying a mixed bundle of goods to local consumers. The prediction that lower transportation costs between cities should lead to greater urban specialisation has received mixed empirical support. With the secular decline in transport costs, one would expect urban specialization to increase. Instead, sectoral specialisation in US cities has declined since at least the 1970s while functional specialisation has increased during the same period (Duranton and Puga, 2005). Allowing for transportation costs to respond differently to changes in infrastructure generates a richer set of predictions. In particular, Duranton, Morrow, and Turner (2013) develop a framework where more highways to enter or

[30]It is worth noting that endogenous changes in the number and specialization of cities can alter the link between changes in national sectoral employment and changes in city sizes. For instance, as seen above when discussing the comparative statics on the productivity shifter B^j, sector-specific shocks that increase equilibrium city sizes typically lead to a consolidation of the sector's employment in fewer cities of larger size, so that cities with similar sectoral composition may experience very different changes during the adjustment. Alternatively, a positive demand shock may lead to a sector being present in more cities without significant changes in the size of cities that initially hosted the sector.

exit a city make it cheaper to export heavier goods which are more sensitive to the provision of roadway. For US cities they find that cities with more highways tend to be more specialised in the production of heavier goods and export them more.

Redding and Sturm (2008) consider a model similar to the one developed above but with only one sector for which differentiated varieties directly enter the utility function, as in Helpman (1998). There are still both agglomeration economies and crowding costs related to city size. In addition, the introduction of transport costs creates additional concentration and dispersion forces related to the relative location of cities in space. The interaction between transportation costs and increasing returns in production creates a home market effect where firms want to concentrate their production in cities with good access to large markets (Krugman, 1980). Counteracting this is the fact that firms close to large markets face a larger number of competitors. An important prediction of this framework is that cities with a better market access should be larger. Redding and Sturm (2008) successfully test this prediction using the division of Germany after the Second World War as a natural experiment. They show that West German cities located close to the Iron Curtain lost significant market access and declined in population relatively to other West German cities.

5.2 *Empirical magnitude of urban benefits and costs*

We have seen that agglomeration economies are essential to understand why cities exist at all, and their magnitude fundamentally affects city sizes and patterns of firm and worker location. Thus, quantifying agglomeration economies has been a key aim of the empirical literature in urban economics, especially in recent years.

Agglomeration economies imply that firms located in larger cities are able to produce more output with the same inputs. Thus, perhaps the most natural and direct way to quantify agglomeration economies is to estimates the elasticity of some measure of average productivity with respect to some measure of local scale, such as employment density or total population. This elasticity corresponds to parameter σ in the model just presented. In early work, Sveikauskas (1975) regressed log output per worker in a cross-section of city-industries on log city population and found an elasticity of about 0.06. More recent studies have obtained estimates of around 0.02 to 0.05, after dealing with three key potential problems in the original approach.

The first problem is that measuring productivity with output per worker will tend to provide upwardly biased estimates of σ, since capital is likely to be used more intensively in large cities. To address this concern, recent contributions focus on total factor productivity, calculated at the aggregate level for each area being considered or, more recently, at the plant level. A particularly influential contribution using this approach is that of Henderson (2003), who estimates total factor productivity using plant-level data in high-tech and machinery sectors for the United States.

A second concern when estimating agglomeration economies is that productivity and city size are simultaneously determined. If a location has an underlying productive advantage, then it will tend to attract more firms and workers and become larger as a result. Following Ciccone and Hall (1996), the standard way to tackle this issue is to instrument for the current size or density of an area. The usual instruments are historical population data for cities and characteristics that

are thought to have affected the location of population in the past but that are mostly unrelated to productivity today. The logic behind these instruments is that there is substantial persistence in the spatial distribution of population (which provides relevance), but the drivers of high productivity today greatly differ from those in the distant past (which helps satisfy the exclusion restriction). Most studies find that reverse causality is only a minor issue in this context and that estimates of σ are not substantially affected by instrumenting (Ciccone and Hall, 1996, Combes, Duranton, Gobillon, and Roux, 2010). An alternative strategy to deal with a potential endogeneity bias is to use panel data and include city-time fixed effects when estimating plant-level productivity, to capture any unobserved attributes that may have attracted more entrepreneurs to a given city (Henderson, 2003).[31] Finally, Greenstone, Hornbeck, and Moretti (2010) follow an ingenious quasi-experimental approach. They identify US counties that attracted large new plants involving investments above one million dollars as well as runner-up counties that were being considered as an alternative location by the firm. They find that, after the new plant opening, incumbent plants in chosen counties experience a sharp increase in total factor productivity relative to incumbent plants in runner-up counties.

A third concern with productivity-based estimates is that agglomeration economies are not the only reason why average productivity may be higher in larger cities. As in Melitz and Ottaviano (2008) or Syverson (2004), the large number of firms in larger cities may make competition tougher, reducing markups and inducing less productive firms to exit. In this case, higher average productivity in larger cities could result from firm selection eliminating the least productive firms rather than from agglomeration economies boosting the productivity of all firms. Combes, Duranton, Gobillon, Puga, and Roux (2012*b*) develop a framework to distinguish between agglomeration and firm selection. They nest a generalized version of the firm selection model of Melitz and Ottaviano (2008) and a simple model of agglomeration in the spirit of Fujita and Ogawa (1982) and Lucas and Rossi-Hansberg (2002). This nested model enables them to parameterise the relative importance of agglomeration and selection. The main prediction of their model is that, while selection and agglomeration effects both make average firm log productivity higher in larger cities, they have different predictions for how the shape of the log productivity distribution varies with city size. More specifically, stronger selection effects in larger cities, by excluding the least productive firms, should lead to a greater left truncation of the distribution of firm log productivities in larger cities. Stronger agglomeration effects, by making all firms more productive, should lead instead to a greater rightwards shift of the distribution of firm log productivities in larger cities. If firms that are more productive are also better at reaping the benefits of agglomeration, then agglomeration should lead not only to a rightwards shift but also to an increased dilation of the distribution of firm log productivities in larger cities.

Using a quantile approach that allows estimating a relative change in left truncation, shift, and dilation between two distributions and establishment-level data for France, Combes, Duranton, Gobillon, Puga, and Roux (2012*b*) conclude that productivity differences across urban areas in

[31] There are some clear limitations to this strategy. Changes in sectoral productivity are potentially determined simultaneously with changes in employment in the same sector. One may perhaps argue that employment adjusts only slowly after productivity shocks. This then calls for using high-frequency data but serial correlation is likely to be a major issue in this case.

France are mostly explained by agglomeration. They compare locations with above median employment density against those with below-median density (results are almost identical when comparing cities with population above or below 200,000). The distribution of firm log productivity in areas with above-median density is shifted to the right and dilated relative to areas below median density. On the other hand, they find no difference between denser and less dense areas in terms of left truncation of the log productivity distribution, indicating that firm selection is of similar importance in cities of different sizes. Their results show that firms in denser areas are thus on average about 9.7 percent more productive than in less dense areas. Put in terms of σ, this implies an elasticity of 0.032. However, the productivity boost of larger cities is greater for more productive firms, so the productivity gain is 14.4 percent for firms at the top quartile and only of 4.8 percent for firms at the bottom quartile.

For estimating the empirical magnitude of σ, an alternative to comparing establishments' productivity across cities is to compare workers' wages instead. As shown in equation (42), from the point of view of workers, higher wages in larger cities are offset by higher house prices. Looking at the spatial equilibrium from the point of view of firms, equation (39) shows that for firms to be willing to pay higher wages to produce in larger cities, there must be productive advantages that offset the higher costs. Thus, comparing wages across cities of different sizes also allows us to quantify the magnitude of agglomeration economies. This approach is used by Glaeser and Maré (2001), Combes, Duranton, and Gobillon (2008), Combes, Duranton, Gobillon, and Roux (2010) and De la Roca and Puga (2012), amongst others. A key concern when interpreting the existence of an earnings premium for workers with similar observable characteristics in larger cities is that there may be unobserved differences in worker ability across cities. Following Glaeser and Maré (2001), a standard way to tackle this concern is to use panel data for individual workers and introduce worker fixed-effects. Compared with a simple pooled OLS regression, a fixed-effects regression reduces the estimate of σ by about one-half (Combes, Duranton, Gobillon, and Roux, 2010). This drop in the estimated elasticity when worker fixed-effects are introduced is sometimes interpreted as evidence of more productive workers sorting into bigger cities. However, De la Roca and Puga (2012) argue that the drop is mostly due to the existence of important learning advantages of larger cities. A pooled OLS regression mixes the static advantages from locating in a larger city, with the learning effects that build up over time as workers in larger cities are able to accumulate more valuable experience, with any possible sorting. Introducing worker fixed-effects makes the estimation of agglomeration economies be based exclusively on migrants, and captures the change in earnings they experience when they change location. This implies that an earnings regression with worker fixed effects likely is expected to provide an accurate estimate of σ, capturing the static productive advantages of larger cities. Recent studies find the estimated value of σ thus estimated to be around 0.025 (Combes, Duranton, Gobillon, and Roux, 2010, De la Roca and Puga, 2012). At the same time, to more fully capture the benefits of larger cities, we should also study learning effects. We return to these below.

As we have seen, equilibrium and efficient city sizes are the result of a trade-off between agglomeration economies, as measured by σ, and urban crowding costs, as measured by γ. While there is now a large literature estimating the value of σ, the elasticity of urban productivity

advantages with respect to city size, much less is known about γ, the elasticity of crowding costs with respect to city size. Combes, Duranton, and Gobillon (2012*a*) develop a methodology to estimate this and apply it to French data. As highlighted by the monocentric city model studied in section 2, house prices within each city vary with distance to the city centre offsetting commuting costs. House prices at the city centre capture the combined cost of housing and commuting in each city, so they are a relevant summary of urban costs. Combes, Duranton, and Gobillon (2012*a*) use information about the location of parcels in each city and other parcel characteristics from recorded transactions of land parcels to estimate unit land prices at the centre of each city. They then regress these estimated (log) prices at the centre of each city on log city population to obtain an estimate of the elasticity of unit land prices at the centre of each city with respect to city population: 0.72. Multiplying this by the share of land in housing (0.25) and then by the share of housing in expenditure (0.23), yields an elasticity of urban crowding costs with respect to population of 0.041

Hence existing empirical estimates suggest that the difference between the crowding costs elasticity γ and the agglomeration elasticity σ is small, perhaps 0.02 or less.[32] This has some interesting implications. On the one hand, optimal city sizes as given by equation (49) should be highly sensitive to changes in agglomeration economies and productivity. On the other hand, mild deviations from optimal city sizes as described by equation (49) should have only a small economic cost. This in turn means that it may be important to better account for migration costs when studying cities: with free mobility small productivity shocks may have large consequences for city sizes, whereas if mobility costs are important migration may only weakly respond to shocks, since the net effect from changes in agglomeration benefits and crowding costs achieved by moving may be small.

6. Human capital and entrepreneurship

The models of cities considered so far are static. We have used comparative static results from those models to provide predictions about the effects of some manifestations of economic growth, such as better transportation or higher incomes, on the population and structure of cities. This unidirectional approach is valid if aggregate growth is not affected by the drivers of urban growth, as is arguably the case for urban amenities. However, the lack of feedback from cities to aggregate growth is questionable for the drivers of urban growth that we examine in this section: human capital and entrepreneurship. As discussed below, a good case can be made that human capital and entrepreneurship affect the growth of cities. Human capital and entrepreneurship are also arguably at the heart of the process of aggregate growth (Lucas, 1988, Aghion and Howitt, 1992). To explore two-way interactions between urban population growth and aggregate economic growth, dynamic models are needed.

[32]Unfortunately, the empirical literature only provides estimates for an average agglomeration elasticity for all cities not for city-specific agglomeration elasticities. There are sector-specific agglomeration elasticities available from the literature (e.g., Henderson, 2003) but they are subject to more serious identification concerns than agglomeration elasticities estimated at the city level since there is no good instrument for sectoral employment in cities. It is also unclear how elasticities for sectoral employment map into city-specific agglomeration elasticities given that most cities are far from being fully specialised.

As stressed in section 5, a complete modelling of cities must include some form of agglomeration benefits. It is possible that agglomeration economies are static (i.e., take place in production) and affect the dynamics of aggregate growth only indirectly. It also possible that agglomeration benefits are dynamic (i.e., take place in the accumulation of factors) and affect the dynamics of growth directly. In this section, we first explore a model in which agglomeration benefits are static but have dynamic implications before turning to dynamic benefits from agglomeration.

6.1 Human capital and urban growth: static externalities

The model that follows draws from Duranton and Puga (2013) and captures key elements from Black and Henderson (1999). There are N_{it} workers in city i at time t. The output of each of these workers is

$$y_{it} = B\, H_{it}^{\sigma}\, h_{it}^{\alpha}\, l_{it}^{1-\alpha}\ . \tag{51}$$

This production process offers constant returns at the individual level in the worker's human capital, h_{it}, and labour, l_{it}, but it is subject to a city-level externality in aggregate human capital, H_{it}^{σ}. Duranton and Puga (2013) develop micro-foundations for this production function in which the human capital externality arises by fostering entrepreneurship. Aggregate human capital in each city is the sum of the individual capital of its workers: $H_{it} = h_{it} N_{it}$. Each worker devotes a share δ of the unit of time that she has every period to accumulating human capital and a share $1-\delta$ to working.[33] As a result of this investment, human capital evolves according to the following accumulation equation:

$$h_{it} - h_{it-1} = b\,\delta\, h_{it-1}\ . \tag{52}$$

The parameter b measures the marginal return to the time devoted to human capital accumulation: $d(h_t/h_{t-1})/d\delta = b$. Note that human capital at time t needs to be a linear function of human capital at time $t-1$, as in equation (52), for self-sustained but non-explosive growth to be possible.

We identify the accumulation factor h_{it} with human capital and model its accumulation accordingly in equation (52) through a time investment made by individuals. Like Romer (1986), we could have labelled the accumulation factor physical capital instead. This would have made no difference to our modelling of production in equation (51) but would have required a different accumulation process to replace equation (52), since investment in physical capital is more appropriately modelled as foregone consumption measured in output rather than foregone time spent learning. We prefer to focus on human capital given the rich literature providing evidence about human capital externalities in cities.[34]

Another possibility would be to identify the accumulation factor with knowledge, following Romer (1990). The accumulation equation (52) would then be more appropriately modeled by

[33] In Black and Henderson (1999), the share of time devoted to human capital accumulation is endogenous. As in much of the endogenous growth literature, it ends up being constant in steady-state following intertemporal utility maximization by consumers with log-linear intertemporal preferences.

[34] With physical capital instead of human capital and a standard investment function where capital in t is equal to capital in $t-1$ minus depreciation plus foregone consumption, the production externality needs to be such that $\sigma = \alpha$ for self-sustained growth to be possible (Romer, 1986, Duranton and Puga, 2004). On the other hand, the accumulation equation no longer requires the linearity assumed in equation (52). In any case, the results obtained from both sets of assumptions are qualitatively the same.

describing firms conducting research and development. Successful innovators are rewarded with patents, while their innovation also increases a common stock of knowledge available to all, which in turn facilitates further innovations. While knowledge arguably plays an important role in long-run aggregate growth, using knowledge as accumulation factor in an urban context would force us to model its diffusion across cities to get non-trivial interactions between cities and aggregate growth. We return to this issue in the next section.

We model cities as in section 5.1. This implies that the consumption of a worker living in city i is $c_{it} = y_{it} - \frac{\tau}{\gamma} N_i^{\gamma}$. Substituting equation (51), $H_{it} = h_{it} N_{it}$ and $l_{it} = (1 - \delta)$ into this expression, we can write per-capita consumption as

$$c_{it} = B(1-\delta)^{1-\alpha} h_{it}^{\alpha+\sigma} N_{it}^{\sigma} - \frac{\tau}{\gamma} N_i^{\gamma} \,. \tag{53}$$

Since returns to human capital investments are the same everywhere, with perfect mobility across cities workers choose their city of residence at each period to maximize their present consumption. With profit-maximizing land developers, as in section 5.1, equilibrium city sizes are optimal and are given by:

$$N_{it} = \left(B(1-\delta)^{1-\alpha} h_{it}^{\alpha+\sigma} \frac{\sigma}{\tau} \right)^{\frac{1}{\gamma-\sigma}} \,. \tag{54}$$

Note this expression, which maximizes c_{it} in equation (53), is the same as equation (49) from section 5.1, with the productivity shifter B replaced by $B(1-\delta)^{1-\alpha} h_{it}^{\alpha+\sigma}$. In section 5.1, we treated the productivity shifter B as an exogenous parameter to see how aggregate or sectoral shocks would affect cities. The term $B(1-\delta)^{1-\alpha} h_{it}^{\alpha+\sigma}$ is instead endogenous and driven by human capital accumulation. As workers become more productive through their accumulation of human capital, they find it worthwhile to agglomerate in larger cities. Hence, when economic growth takes the form of human capital accumulation, it leads to growing city sizes ($\frac{\mathrm{d}N_{it}}{\mathrm{d}h_{it}} > 0$).

The relationship between human capital and growth does not stop here. The growth of cities, through agglomeration economies, amplifies the effects of human capital accumulation for aggregate growth. Following Duranton and Puga (2013), we can write the evolution of output per worker as:

$$\begin{aligned} \frac{y_{it}}{y_{it-1}} &= \left(\frac{h_{it}}{h_{it-1}} \right)^{\alpha+\sigma} \left(\frac{N_{it}}{N_{it-1}} \right)^{\sigma} \\ &= (1+b\delta)^{(\alpha+\sigma)\left(1+\frac{\sigma}{\gamma-\sigma}\right)} \\ &\approx 1 + b\delta \frac{\gamma(\alpha+\sigma)}{\gamma-\sigma} \,, \end{aligned} \tag{55}$$

where the first line of equation (55) is obtained from equation (51), the second line makes use of equations (52) and (54), and third provides a simple linear approximation when $b\delta$ is small. The last line of equation (55) shows that in absence of agglomeration economies ($\sigma = 0$) the growth rate of output is $b\delta\alpha$. With positive agglomeration economies ($\sigma > 0$), the growth rate of output per person is higher at $b\delta \frac{\gamma(\alpha+\sigma)}{\gamma-\sigma}$. We can compute the contribution of urban agglomeration to economic growth as

$$\frac{b\delta \frac{\gamma(\alpha+\sigma)}{\gamma-\sigma} - b\delta\alpha}{b\delta \frac{\gamma(\alpha+\sigma)}{\gamma-\sigma}} = \frac{(\alpha+\gamma)\sigma}{(\alpha+\sigma)\gamma} \,. \tag{56}$$

This expression represents the increase in the growth rate as the result of urban agglomeration economies ($\sigma > 0$) relative to the total growth rate that appears in equation (55). Empirically, recall from the discussion in section 5.2 that estimates in the literature of σ, the agglomeration coefficient, and γ, the urban costs coefficient, are small. If we use our preferred estimates of $\sigma = 0.025$ and $\gamma = 0.04$, equation (56) implies that cities account for 64% of aggregate growth.[35]

While this is a large number, we should keep in mind that we only consider growth from human capital accumulation and ignore other sources of growth such as physical capital accumulation and knowledge accumulation.[36] This nonetheless suggests that Lucas (1988) made an important point when he suggested looking at cities to understand the effects of human capital externalities. The large contribution of agglomeration to aggregate growth is also consistent with results from the human capital literature, which typically finds that external returns to human capital in cities are of about the same magnitude as private returns (e.g., Moretti, 2004*a*).

6.2 Human capital and urban growth: dynamic externalities

We now turn to the modeling of dynamic agglomeration effects. As suggested by Alfred Marshall long ago: "The mysteries of trade become no mysteries; but they are as it were in the air, children learn many of them unconsciously. Good work is rightly appreciated, inventions and improvements in machinery, in process and the general organization of the business have their merits promptly discussed: if one man starts a new idea, it is taken up by others and combined with suggestions of their own; and thus becomes the source of further new ideas" (Marshall 1890: iv.x.3). Several approaches have been developed to model these ideas. In an approach related to Black and Henderson (1999), Eaton and Eckstein (1997) adapt Lucas' (1988) model of human capital and growth to an urban context. To discuss their framework, let us start with a simple production function with no agglomeration effect. The output of a worker in city i is

$$y_{it} = B\, h_{it}^{\alpha}\, l_{it}^{1-\alpha}\,, \tag{57}$$

where, again, each worker devotes a share δ of her time to human capital accumulation, h_{it} is individual human capital and $l_{it} = 1 - \delta$ is individual labour. In contrast to equation (51), equation (57) has no externality in production. This externality now appears in the accumulation equation. Thus, instead of an accumulation equation like (52), where each worker builds on her own human capital, Eaton and Eckstein (1997) assume that all residents of city i learn from the same aggregate knowledge base H_{it}:

$$h_{it} - h_{it-1} = b\, H_{it}\, \delta\,. \tag{58}$$

It may seem natural, as before, to think of the city's knowledge base as the sum of the human capital of all residents $H_{it} = h_{it} N_{it}$. However, having dynamic scale effects in equation (58) would

[35]The computation also requires assigning a value to α. The 64% figure is obtained from $\alpha = 0.5$, following the finding by Mankiw and Weil (1992) of equal shares for labour and human capital in production. However, our results are not at all sensitive to this choice. With $\alpha = 0.7$ the contribution of urban agglomeration to aggregate growth is still 64%, with $\alpha = 0.3$ it is 65%. To a first approximation, the contribution of urban agglomeration to growth is σ/γ.

[36]Davis, Fisher, and Whited (2011) conduct a similar exercise within a neoclassical model of growth with physical capital and no human capital. They find a much smaller contribution of agglomeration to aggregate growth of about 10%.

imply that cities of different population size experience different growth rates. Ultimately the output of the entire economy would be dominated by that of the largest city, where output per worker would grow increasingly faster than in other cities.

An alternative way to think about the city's knowledge base H_{it} would be to equate it with the average human capital in the city: $H_{it} = \overline{h}_{it}$. This raises three problems. The first is that city size no longer matters since production now only depends on the individual's human capital and this accumulates at a rate that does not depend on city size. If urban costs increase with a city's size, efficiency then calls for the smallest possible cities. So instead of having one city of exploding size, we have all cities disappear. The second issue is that the process of economic growth can take place in each city separately and independently. This is arguably counterfactual. A third problem arises when we introduce some heterogeneity in individual human capital levels. Because equation (58) now implies that an individual's human capital increases more rapidly in cities with higher average human capital, this heterogeneity provides a strong incentive for sorting and leads again to faster growth in some cities

To avoid these three problems, Eaton and Eckstein (1997) propose a more complicated production function with static agglomeration economies as in equation (51). Although assuming agglomeration economies in production 'solves' the problem created by the lack of scale effects, it means this is no longer a model with dynamic agglomeration economies. Agglomeration effects essentially remain static. In response to the second issue of each city being a separate economy able to generate self-sustaining growth alone, Eaton and Eckstein (1997) equate the city knowledge base with the weighted sum of the average human capital of other cities: $H_{it} = \sum_j \phi_{ij} \overline{h}_{jt}$ where the weights ϕ_{ij} may depend on the distance between cities. While this still allows cities to be isolated growing economies, this process of diffusion is intuitively appealing. Finally, the third problem of sorting is 'solved' by considering *ex ante* identical workers and a steady state with symmetric growth in all cities so that workers remain identical.

The literature has followed two alternative strategies to reintroduce dynamic agglomeration economies without having one city dominate the entire urban system. The first is to limit how much can be learnt by, for instance, imposing a finite lifetime as in Glaeser (1999). The second strategy is to model the diffusion of innovations as Duranton and Puga (2001). Let us summarize these two approaches.

In a model of skill transmission inspired by Jovanovic and Rob (1989) and Jovanovic and Nyarko (1995), Glaeser (1999) formalizes the notion that the proximity to individuals with greater skills facilitates the acquisition of skills.[37] Glaeser (1999) considers overlapping generations of risk-neutral individuals who live for two periods (young and then old). Workers can be skilled or unskilled, and this affects their productivity: the output of an unskilled worker is lower than that of a skilled worker.

Each worker is born unskilled and chooses whether to spend her youth in the hinterland or in the city. In the hinterland, the cost of living is low but a worker remains unskilled. In the city, the cost of living is higher but a worker may become skilled after successfully meeting with an (old) skilled worker. The probability of a successful meeting increases with the number of skilled

[37] This model is generalized and exposed more formally in Duranton and Puga (2004).

workers in the city. The surplus created by this successful acquisition of skills is split between the young apprentice and her old master. When old, workers chose whether to relocate. Old unskilled workers can no longer become skilled so, given the higher cost of living in the city, they always live in the hinterland. Old skilled workers, however, may offset the higher cost of living in the city with their share of the surplus created by teaching young apprentices.

Provided the benefits from becoming skilled are sufficiently large and provided the probability of meeting a skilled worker in the city is sufficiently high, there is a steady state in which young workers move to the city. Those that become skilled then stay in the city while those that do not become skilled go to the hinterland in their second period.

In a different model of learning, Duranton and Puga (2001) propose a diffusion mechanism where the benefits from learning in one city can be exploited in another.[38] In this model, an entrepreneur can introduce a new product by paying a fixed cost of entry. At first, entrepreneurs need a period of experimentation to realize their full potential — they may have a project, but may not know all the details of the product to be made, what components to use, or what kind of workers to hire. There are many possible ways to implement this project, but one is better than all others.

More specifically, entrepreneurs can choose between many production processes, each associated with a different set of inputs. The ideal production process, which differs across entrepreneurs, is initially unknown. An entrepreneur can try to discover her ideal production process by sampling at most one production process each period and using it for prototype production. As soon as an entrepreneur samples her ideal production process, she knows this is it and can start mass-production. A proportion of firms randomly exit every period to ensure that new firms keep entering and learning is never exhausted.

The use of a particular production process, either for prototype production or mass-production, requires physical proximity with the corresponding input producers. As in the model described in section 5.1, input producers benefit from static agglomeration economies. The cost of using a given production process diminishes as more local firms use the same type of process because they can share intermediate suppliers. At the same time, relocating production across cities is costly, so entrepreneurs who have not yet discovered their ideal production process benefit from locating in a very diversified local economy to facilitate their learning. They would also like to face many suppliers for each set of inputs to enjoy lower costs. However, urban crowding places a limit on city size and consequently on how many processes can be widely used in a city.

Provided learning is important and moving costs are neither too high nor too low, an interesting equilibrium where both diversified and specialized cities arise endogenously can be sustained. It reconciles the needs for diversity and specialization along the life-cycle of firms. Entrepreneurs develop new products in cities with a diversified production structure. It allows them to sample easily and discover their ideal set of inputs. After discovering this ideal set of inputs, entrepreneurs are no longer interested in urban diversity. Because input producers in different sectors do not

[38] In Duranton and Puga (2001), the diffusion of innovations relies explicitly on factor mobility. This differs from the literature in international trade that models diffusion mechanisms occurring through the trade of goods or, directly, through diffusion spillovers (e.g., Grossman and Helpman, 1991*a*).

benefit from each other directly, industrial diversity makes cities more costly. As a result, entrepreneurs who have discovered their ideal set of inputs move away from a diversified city to a specialized cities so that they can benefit from agglomeration effects in the production of those inputs. Moving costs cannot be too high for relocation to occur after learning, nor so low that an entrepreneur can easily learn by constantly relocating. Further, the gains from learning need to be high enough to justify the foregone static agglomeration economies in the early phases. In this sense, we can think of diversified cities as 'nursery cities' where learning takes place and specialized cities as the places where the production of mature goods occurs.

The nursery cities model of Duranton and Puga (2001) proposes a theory of how innovation takes place and diffuses in space, while also matching observed patterns of firm relocations and a number of other facts about cities such as the coexistence of specialized and diversified cities (Duranton and Puga, 2000). It can also be used to explain why, even if innovation and learning concentrate in a few large and diverse cities, this does not imply that smaller and more specialized cities will disappear. Instead, the diffusion of innovations to exploit them in small specialized cities frees up large and diverse cities to concentrate in continuously feeding the growth process with new ideas.

6.3 Human capital

Empirically, the strong association between city human capital and city population growth has been noted for some time. Glaeser, Scheinkman, and Shleifer (1995), Simon and Nardinelli (1996), and Simon (1998) estimate regressions of the following form:

$$\Delta_{t+1,t} \log N_i = \beta_0 + \beta_1 \log N_{it} + \beta_2 h_{it} + X_{it}\beta_3^{'} + \epsilon_{it} \ , \tag{59}$$

where the dependent variable is the change in log population or log employment between t and $t+1$ in city i. The explanatory variable of interest h_{it} is a measure of human capital at time t. Finally, X_{it} is a set of controls for other engines of growth, which often includes region dummies, and initial population is also controlled for. To measure human capital, early work used a range of education variables (e.g., Glaeser, Scheinkman, and Shleifer, 1995) or rough proxies (such as the number of business professionals in Simon and Nardinelli, 1996, for 19th century England). More recent work (e.g., Simon and Nardinelli, 2002, Glaeser and Saiz, 2004) prefers the share of university graduates since this more discriminant measure of human capital is usually associated with stronger effects.

Note that our growth model from section 6.1 can be used to motivate this specification. Dividing equation (54) valued at time $t+1$ from the same equation valued at time t, and taking logs, we obtain

$$\Delta_{t+1,t} \log N_i = \frac{\alpha + \sigma}{\gamma - \sigma} \Delta_{t+1,t} h_{it} \ . \tag{60}$$

The main difference, leaving aside the controls and the error term, is that the theoretical equation (60) relates changes in population to changes in human capital whereas the empirical specification (59) relates changes in population to initial levels of human capital. However, if we assume, as in the regression relating city growth to roads in section 2, that population adjusts slowly to any

changes in human capital, we end up with a regression of changes on initial levels instead with initial population as an additional control (see equations 14 and 15).

In a thorough investigation of the relationship between human capital and city growth across US metropolitan areas between 1970 and 2000, Glaeser and Saiz (2004) conclude that one standard deviation in the share of university graduates in a city's workforce is associated with a quarter of a standard deviation of population growth during the following decade. Put differently, for an average city, a one percentage point higher share of university graduates is associated with around 0.5 percent population growth over the subsequent decade. This finding is representative of the findings in the rest of the literature.[39]

The strong association between human capital and city growth might be spurious for a number of reasons. For instance, more educated workers may be more mobile (or equivalently have stronger incentives to move) and, as a result, end up being over-represented in fast-growing cities. Alternatively, the effect may be stronger than estimated. This would occur, for example, if cities with more stringent zoning restrictions, which experience slower population growth, also retain a more educated workforce.

To investigate these concerns and to show that the effect of human capital on city growth is most likely causal, Glaeser and Saiz (2004) perform a number of robustness checks. First, they show that education levels affect city growth even after controlling for a wide array of city characteristics. Second, they show that the relationship between education levels and city growth holds when looking only at variations within cities over time. That is, a given city tends to grow faster during periods when its population is more educated. This indicates that the relationship between human capital and city growth is not driven by unobserved permanent characteristics that make cities grow faster and also attract more educated workers. Finally, to account for the possibility of a common determinant of both city growth and human capital, they use instrumental variables. To obtain an exogenous determinant of human capital in cities, they follow Moretti (2004*a*) and use the foundation of land grant colleges as an instrumental variable. Starting in 1862, land grant colleges were created in each state to foster agricultural and engineering education. They were usually placed in cities that were conveniently located (typically a central location in a state). Shapiro (2006) shows that these cities were not more educated before 1900 but gradually became more educated as the grant colleges developed, often turning into major universities. Glaeser and Saiz (2004), like Shapiro (2006), find that instrumenting city human capital by the presence of land grant colleges strongly suggests that the effect of education on city growth is causal and, if anything, leads to higher coefficients than indicated by the simple association in the data.

The literature has also provided less direct evidence about the role of human capital in city growth by investigating the channels through which it percolates. The model in section 6.1 proposes some direct benefits in production occurring through human capital externalities in cities (see equation 51). The notion that smart, educated people benefit from being surrounded by other smart, educated people has received support in the literature. Following Rauch (1993), Moretti

[39]The main exception is Glaeser, Ponzetto, and Tobio (2011). They fail to find a positive association between human capital and subsequent county population growth in the Eastern and Central United States for a few decades in the last two hundred years.

(2004*a*, 2004*b*) finds robust evidence of large external effects of university education on city wages and productivity.

The human capital externalities of the model in section 6.1 are micro-founded in Duranton and Puga (2013) through a link between human capital and entrepreneurship. Entrepreneurs may be over-represented among more educated workers. If this is the case, a more educated city is also a more entrepreneurial city, where more new firms are created and existing firms grow faster. Stronger population growth then naturally follows. We explore empirical evidence of this channel in greater depth below. For now, we note that when attempting to disentangle between different channels through which human capital affects city growth, Glaeser and Saiz (2004) and Shapiro (2006) provide evidence that most of the effects of human capital percolate through a productivity channel, either learning and human capital externalities or entrepreneurship and firm growth. De la Roca and Puga (2012) explicitly study learning effects, using rich administrative data for Spain that tracks workers' full employment histories. They find that, by working in bigger cities, workers not only obtain an immediate static earnings premium, as in the model of section 5.1, but are also able to accumulate more valuable experience, which increases their earnings faster. The additional value of experience accumulated in bigger cities persists even after workers move away and is even stronger for those with higher initial ability. This is evidence of the importance of learning in cities, providing support for the idea that cities foster the accumulation of human capital.

Higher productivity is not the only possible channel through which human capital can affect city growth. It could also be the case that more educated cities develop better amenities. These amenities are attractive to workers from other cities, particularly educated workers. Although Glaeser and Saiz (2004) and Shapiro (2006) only find modest support regarding the importance of amenities created by the presence of a skilled workforce, Diamond (2013) stresses this channel to explain the divergence in the skill composition of US cities in the last 30 years. The tension between these divergent findings will hopefully be resolved by future research.

A difficulty with human capital externalities and most forms of knowledge spillovers is that they are hard to track directly since they do not leave a paper trail. There is however one outcome of interactions that leaves some paper trail behind: innovations, when they are patented, contain citations to other patents. In their pioneering work, Jaffe, Trajtenberg, and Henderson (1993) show a local bias in citation patterns. A patent is more likely to be cited by a subsequent patent for which the inventor lives in the same US metropolitan areas than by a 'similar' patent for which the inventor lives in a different area. While this initial finding has been shown to be sensitive to what one means by 'similar' and how one defines the control group for citing patents (Thompson and Fox-Kean, 2005), more recent work has established it on firmer grounds (Murata, Nakajima, Okamoto, and Tamura, 2013) and evidenced a host of other phenomena associated with knowledge spillovers in innovative activity. For instance, Agrawal, Cockburn, and McHale (2006) show that citations for a given patent are also disproportionately often more likely to occur in locations where the cited inventor was living prior to obtaining this patent. In other research, Kerr (2010) shows that, for a given technology, patenting growth in a city is stronger after a breakthrough innovation and that this growth differential is higher for technologies that depend more heavily on immigrant innovators, who are arguably more mobile. It is beyond the scope of this chapter to review this

broad literature. We refer instead the reader to the survey of Carlino and Kerr (2013).

6.4 Entrepreneurship

To investigate the effect of 'agglomeration' on city growth Glaeser, Kallal, Scheinkman, and Schleifer (1992) propose the following regression:

$$\Delta_{t+1,t} \log N_i^j = \beta_0 + \beta_1 Spec_{it}^j + \beta_2 Div_{it}^j + \beta_3 EstSize_{it}^j + X_{it}^j \beta_4' + \epsilon_{it}^j \,, \tag{61}$$

where the dependent variable is the change in log employment between t and $t+1$ in city i and sector j. The use of log employment as the dependent variable is motivated by a positive link from productivity growth to employment growth. The explanatory variables are a measure of initial specialization, $Spec_{it}^j$, a measure of sectoral diversity faced by sector k in city i, Div_{it}^j, a measure of establishment size, $EstSize_{it}^j$, and a set of other controls X_{it}^j, such as wages, the national growth of sector j during the same period, and initial employment in the city and sector.

The main results of Glaeser, Kallal, Scheinkman, and Schleifer (1992) are: a negative coefficient on initial specialization, a negative coefficient on establishment size, and a strongly positive coefficient on diversity. The effects are quantitatively large. A standard deviation in specialization or diversity is associated with about 10% of a standard deviation in employment growth. A standard deviation in establishment size is associated with nearly a quarter of a standard deviation in employment growth. These results have been subsequently replicated in many countries and generally confirmed. See for instance Combes (2000) for France or Cingano and Schivardi (2004) for Italy. An important qualification of these findings by Henderson, Kuncoro, and Turner (1995) is that diversity appears to be particularly important for high-tech industries whereas specialization seems to play a positive role for mature industries. These results are consistent with those of the model of Duranton and Puga (2001) described above. In another important paper, Feldman and Audretsch (1999) use a measure of innovation instead of employment growth as dependent variable. They find a positive association between innovation and sectoral diversity (provided this diversity is relevant to the sector) and a negative association between innovation and specialization.

The regression described by equation (61) does not directly tie into the model described in section 6.1 nor into any of the frameworks described in section 6.2. In their work, Glaeser, Kallal, Scheinkman, and Schleifer (1992) interpret the coefficients on specialization, diversity, and establishment size as dynamic externalities affecting local employment growth in sectors. In particular, the coefficient on average establishment size is interpreted as a 'competition effect' (or even a 'Porter effect' after Porter, 1990). This interpretation is far-fetched since there is no obvious mapping of establishment size into the toughness of competition. In many reasonable models of industrial organization, tougher competition actually leads to larger firms (Sutton, 1991). It may be more reasonable to think of $EstSize_{it}^j$ as a broad measure of entrepreneurship, since higher entrepreneurship will lead to more start-ups, which will generally be smaller in size than more

mature firms.[40] This, in turn, is consistent with a suggestion initially made by Chinitz (1961) in his classic comparison of New York and Pittsburgh about the importance of small firms and entrepreneurship as a key determinant of the prosperity of cities. This would also be consistent with interpreting entrepreneurship as a form of human capital that would be particularly important in explaining the evolution of cities.

The regression described by equation (61) suffers from another interpretation issue. It is hard to separate mean-reversion in employment caused by measurement error from the true effect of initial specialization since initial employment in the city and sector must be used to compute initial specialization.

A third problem of interpretation, noted by Combes, Magnac, and Robin (2004) and Cingano and Schivardi (2004), is that the link between employment growth and productivity growth need not be positive. In a sector with constant markups, if the price elasticity of demand is larger than one, an increase in productivity implies a higher revenue and an increase in employment. However in sectors where demand is less elastic, the opposite holds. At the level of entire industries fast productivity growth will often lead to declining employment (as illustrated by many traditional manufacturing industries where the ability to produce goods has risen much faster than demand). This could also occur for sectors within cities when goods are differentiated across cities. This does not mean that regression (61) cannot uncover the 'agglomeration' determinants of urban growth. It simply suggests some caution when interpreting any positive effect of diversity, specialization, or establishment size. It need not be the case that diversity fosters productivity which in turn fosters employment growth. To explore this issue more depth, Cingano and Schivardi (2004) suggest running the following regression,

$$\Delta_{t+1,t} \log \text{TFP}_i^j = \beta_0 + \beta_1 Spec_{it}^j + \beta_2 Div_{it}^j + \beta_3 EstSize_{it}^j + X_{it}^j \beta_4' + \epsilon_{it}^j \,, \tag{62}$$

which mirrors equation (61) but uses growth in average firm-level total factor productivity in a city and industry instead of employment growth as dependent variable.

Interestingly, while the estimation of equation (61) by Cingano and Schivardi (2004) generally confirms the findings of Glaeser, Kallal, Scheinkman, and Schleifer (1992), their estimation of equation (62) yields a positive coefficient on specialization, an insignificant coefficient on diversity, and weak results regarding establishment size.[41] The difference in the sign of the coefficient on specialization is consistent with the intriguing possibility raised above: specialization may have strong effects on productivity and, because of inelastic demand, negative effects on employment.

A fourth issue is whether any effect of specialization, diversity, or establishment size can be interpreted as evidence of dynamic externalities. Dynamic externalities imply that the level of, say,

[40] This then begs the question of whether establishment size is a good measure of entrepreneurship and more generally raises the legitimate question of how best to measure entrepreneurship. In the case of a regression like (61), Glaeser and Kerr (2009) show that the results are the same with alternative measures of entrepreneurship such as the number of start-ups.

[41] Glaeser, Kallal, Scheinkman, and Schleifer (1992) also run a regression akin to (62) but use the change in log wage in cities and sectors as dependent variable instead of total factor productivity growth. They find tiny effect associated with their specialization variable and strong positive coefficients on initial employment in the city and sector. They also find a small positive coefficient on diversity and a negative coefficient on the number of establishments. To the extent that wages growth reflects productivity growth, these results are roughly consistent with those of Cingano and Schivardi (2004) .

establishment size, has an effect on the growth of employment. Static externalities, on the other hand, imply that establishment size measured in level has an effect on the level of employment. Put differently, with static externalities it is the first difference in establishment size which affects the growth rate of employment. To distinguish between static and dynamic effects, it would then seem natural to run the following regression

$$\Delta_{t+1,t} \log N_i^j = \beta_0 + \beta_1 \Delta_{t+1,t} EstSize_i^j + \beta_2 EstSize_{it}^j + X_{it}^j \beta_4' + \epsilon_{it}^j \,. \tag{63}$$

A positive coefficient on establishment size would be consistent with dynamic externalities whereas a positive coefficient on the change in establishment size would be consistent with static externalities. This interpretation is problematic because a gradual adjustment of employment following a change in $EstSize_{it}^j$ implies that even with only static externalities we could estimate a positive value for β_2. This is the same argument as with the gradual adjustment of population which follows on transportation improvements discussed above in equations (14) and (15).

To improve on regression (63), a possibility is to estimate models that examine the dynamics of both the number of establishment and their size with perhaps a rich lag structure to assess how much and how fast past values of both variables affect their contemporaneous values. Combes, Magnac, and Robin (2004), estimate the following type of auto-regressive system

$$\left\{ \begin{array}{lcl} \Delta_{t+1,t} \log m_i^j & = & \beta_0 + \beta_1 m_{it}^j + \beta_2 EstSize_{it}^j + X_{it}^j \beta_3' + \epsilon_{it}^j \,, \\ \Delta_{t+1,t} \log EstSize_i^j & = & \beta_4 + \beta_5 m_{it}^j + \beta_6 EstSize_{it}^j + X_{it}^j \beta_7' + \varepsilon_{it}^j \,, \end{array} \right. \tag{64}$$

where m_i^j is the number of establishments in sector j and city i and t measures years. Relative to equation (63), the system estimated in (64) decomposes the growth of employment in a city and industry into the growth in the number of establishments and the growth in their employment size. Combes, Magnac, and Robin (2004) also estimate systems with longer and richer lag structures. They find that a shorter lag structure like the one in equation (64) performs well. In turn, this suggests that the explanatory variables affect employment and establishment size 'fast'. This is consistent with local externalities being 'static' and not 'dynamic'. They also find that the number of establishments is more sensitive to the local structure of economic activity than establishment size. This last result is consistent with the more recent finding of Glaeser and Kerr (2009) that much of local entrepreneurship can be explained by the presence of many small suppliers. Rosenthal and Strange (2010) also highlight the importance of small establishments and suggest that their benefits arise from the greater diversity of specialized suppliers that they provide to local firms.

While interesting and insightful, the work discussed so far does not solve the endogeneity of the key explanatory variables in these regressions. This problem has been neglected by the literature. This is perhaps because regressions like (61) use growth over a period as dependent variable and establishment size at the beginning of the period as explanatory variable. However, using a predetermined variable as explanatory variable in a regression does not guarantee its exogeneity. Local entrepreneurs could enter in large numbers in a city and sector if they foresee strong future demand. That expectations of future growth should trigger entry today is only natural. This is the nature of business.

Glaeser, Kerr, and Ponzetto (2010) examine whether the presence of many small firms in a city and sector is driven by the demand for entrepreneurship or by its supply. To the extent that they

can be captured by higher sales per worker, demand factors do not appear to matter. Their findings point instead at the importance of the supply of entrepreneurship. This indirect approach, however does not entirely solve the causality issue. To tackle it head on, Glaeser, Pekkala Kerr, and Kerr (2012) take an instrumental variable approach. Returning to Chinitz's (1961) initial comparison of Pittsburgh and New York, they use the idea that cities closer to mines have been influenced by large mining firms. In turn, large firms are expected to reduce entrepreneurship by providing attractive employment opportunities for highly skilled workers. Large firms may also breed a local culture of 'company men' which also reduces entrepreneurship. Indeed, proximity to historical mines is associated with larger establishments today even in completely unrelated sectors. Using this instrument, they estimate an even larger effect of entrepreneurship on city growth than the one measured directly from the data. Because a mining past can be associated with a general decline in manufacturing, Glaeser, Pekkala Kerr, and Kerr (2012) replicate their main findings for cities outside the rust-belt. These findings also hold when, instead of focusing on overall employment, they only look at service sectors only remotely tied to mining. Overall, these results are supportive of the notion that entrepreneurship is an important engine of city growth.

7. Random urban growth

In our exposition of random urban growth models, we do not proceed as above with first a theoretical model followed by a discussion of the empirics. Instead, it is convenient to start with a discussion of a key fact about the size distribution of cities before presenting statistical processes that can account for this fact. We then discuss recent attempts at grounding these statistical processes into economic models before returning to a discussion of empirical issues.

7.1 The empirics of Zipf's law

Since Auerbach (1913), the distribution of city sizes has often been approximated with a Pareto distribution. To do this, a popular way is to rank cities in a country from the largest to the smallest and regress the rank on city population N_i in the following manner:

$$\log \text{Rank}_i = \beta_0 - \xi \log N_i + \epsilon_i \,. \tag{65}$$

The estimated coefficient ξ is the exponent, or shape parameter, of the Pareto distribution.[42] Zipf's law (after Zipf, 1949) corresponds to the statement that $\xi = 1$. This implies that the expected size of the second largest city is half the size of that of the largest, that of the third largest is a third of that of the largest, etc.[43]

The empirical validity of Zipf's law is hotly debated. The classic cross-country assessment of Rosen and Resnick (1980) is ambiguous because their average Pareto exponent of 1.14 for 44

[42] Regression (65) is not a standard regression. First, because the dependent variable is computed directly from the explanatory variable, measurement error on the 'true' size also affects the rank and thus leads to a downward bias for the standard errors with OLS. In addition, when $\xi = 1$ the ratio of the largest to the second largest city is equal to two in expectations but its 95% confidence interval is one to 20. Put differently, the largest city is on average more than twice as large as the second largest city. This biases the OLS estimate of ξ with small samples. See Gabaix and Ibragimov (2011) for a simple and elegant solution to this problem. See also the excellent survey of Gabaix and Ioannides (2004).

[43] The deterministic reformulation of Zipf's law is usually referred to as the rank-size rule.

countries has been interpreted as evidence both for and against Zipf's law. Follow-up work by Soo (2005) broadly confirms the results of Rosen and Resnick (1980).

A lot of the debate has centred around the validity of Zipf's law for US cities. Using less than 200 US cities, Krugman (1996) and Gabaix (1999*a*) conclude at a near perfect fit. On the other hand, Black and Henderson (2003) and Eeckhout (2004) dismiss Zipf's law. Black and Henderson (2003) use data for metropolitan areas for the entire 20th century. They argue that the Pareto exponent is 'far' from one at around 0.8 and that the linearity of the relationship between log size and log rank is questionable. Eeckhout (2004) uses data for US places and argues that their size distribution is better described by a log normal than by a Pareto distribution. Rozenfeld, Rybski, Gabaix, and Makse (2011) use high-resolution data for the United States and aggregate settlements that are close to each other into cities. When defined from the 'bottom up', they find that Zipf's law holds very well for cities with population above 10,000. Giesen, Zimmermann, and Suedekum (2010) argue that, for a number of countries, a distribution that is Pareto for both tails and log normal for its body (double Pareto lognormal) provides a better fit to the data. In the same spirit, Ioannides and Skouras (2013) estimate a variety of log-normal and Pareto shapes allowing for some switching between them or a mixture of both. They highlight the importance of the excellent fit of the Pareto in the upper tail where most of the population lives and some fragility in the lower tail where the results depend on the definition of cities being used.

Stepping back from these seemingly contradictory claims, the empirical debate is mainly about three issues. The first is what constitutes a proper definition for cities. Ideally, this definition should be given by the model at hand. As made clear above, many urban models have commuting patterns at their core. Practically, this argues in favour of defining cities from commuting patterns. However the notion of spatial continuity used by Rozenfeld, Rybski, Gabaix, and Makse (2011) is also legitimate since urban models also imply that cities should be constituted of contiguous commercial and residential areas with agriculture beyond the urban fringe.[44]

The second issue in this debate is about whether we observe a Pareto distribution. When distributions have the same number of parameters to be estimated, like with Pareto and log normal, they can be compared directly in terms of goodness of fit. This is nonetheless problematic because the goodness of fit may be different in different parts of the distribution. The Pareto distribution may offer a better fit in the upper tail whereas the lognormal may fit the body of the distribution better. In addition, distributions often have different numbers of parameters. Distributions with more parameters are expected to provide mechanically a better fit. For instance, a mixture of Pareto and lognormal is bound to do better than a simple Pareto or a simple log normal. The standard approach is then to rely on specification tests that weight the fit of a distribution relative

[44]In practice both types of definitions run into a number of problems. With commuting-based definitions, (sub-metropolitan) jurisdictions are aggregated to a given core when they send a minimum fraction of their workers to this core. The procedure is repeated until no jurisdiction remains to be aggregated to the resulting metropolitan area. However, these jurisdictions are themselves arbitrary (and sometimes extremely large in the West of the United States). The threshold of commuters is also arbitrary and the set of resulting metropolitan areas might be sensitive to this. Definitions based on spatial continuity also need to rely on some arbitrary level of distance with no development (or close to none) to separate metropolitan areas. For cities with green belt, spatial contiguity may also restrict the metropolitan area to be the area within the green belt when, in many cases, workers may commute from outside this green belt in large numbers. See Duranton (2013) for further discussion.

to its number of parameters. The usefulness of this approach is questionable because the penalty associated with more parameters in those tests is arbitrary.

Even if one is willing to accept that city sizes are drawn from a Pareto distribution, the third issue is, whether the Pareto shape parameter is equal to one or not. This can readily be 'tested' using standard levels of statistical significance relative to unity for $\hat{\xi}$ as estimated in regression (65). This approach is nonetheless debatable. With enough data points, one can always reject any sharp hypothesis like $\xi = 1$. In practice, the standard errors around ξ are fairly large even for urban systems with many cities so that it is hard to reject Zipf's law. Of course, it is also hard to reject distributions which are quite far from Zipf's law.[45]

In the end, the more relevant question is not so much whether the distribution of city sizes satisfies Zipf's law or not, but whether looking at this distribution through the lens of Zipf's law is useful. We believe it is, for two reasons. First, Zipf's law provides a reasonable first approximation, at least for the upper tail of the distribution. Second, because both the regularities of Zipf's law and the observed empirical deviations from it can be used to guide the modelling of economic processes underlying city size distributions (Gabaix, 1999*a*).[46]

7.2 The statistics of Zipf's law

Let us now explore the statistical processes that lead to Zipf's law. There are two (related) avenues: multiplicative and additive processes.

Following Gabaix (1999*a*) and Gabaix (1999*b*), multiplicative processes have attracted a lot of attention. These processes are referred to as Kesten processes (after Kesten, 1973). We borrow from Gabaix and Ioannides (2004) and consider an economy where total population and the number of cities are both fixed. Between time $t-1$ and t, city i grows according to $N_{it} = (1+g_{it})N_{it-1}$. We impose Gibrat's law (after Gibrat, 1931): g_{it} is identically and independently distributed for every city with density $f(g)$. After T periods the size of city i is:

$$\begin{aligned} \log N_{iT} &= \log N_{i0} + \sum_{t=1}^{t=T} \log(1+g_{it}) \\ &\approx \log N_{i0} + \sum_{t=1}^{t=T} g_{it} \ . \end{aligned} \tag{66}$$

We note that the approximation in this equation holds only when the shocks are small enough. By the central limit theorem, over time $\log N_{iT}$ approaches a normal distribution and the distribution

[45] Gabaix and Ibragimov (2011) show that the standard error on ξ is asymptotically $\sqrt{2/n}\ \xi$ where n is the number of observations. With 100 cities, it is not possible to reject that $\hat{\xi} = 1.38$ statistically differs from unity at 5%. Even with 1,000 cities $\hat{\xi} = 1.09$ cannot be rejected as being different from unity.

[46] An alternative way to proceed is proposed in Duranton (2007) where the (non-Zipf) predictions of the model are measured directly against the empirical reality. This is in contrast with much of the extant literature, which often proposes a model that may or may not yield Zipf's law, compares it to this benchmark, and then in turn compares the benchmark to the empirical reality. Comparing the predictions of a model directly to the data is more straightforward and avoids the pitfalls mentioned above. However, this is not without problems either. Some of the results of a model may depend on a choice of auxiliary parameters about which not much is known. Consequently, too many degrees of freedom might be available for a meaningful assessment of what really matters for the model. There is also a risk of overextending conclusions reached based on a particular dataset or country that may not be representative of a broader tendency.

of N_{iT} thus becomes log normal. This distribution of city sizes does not admit a steady state and its variance keeps increasing.

To obtain a steady state, one needs to impose a lower bound to city sizes. This prevents cities from becoming too small. Let $M_t(N)$ denote the share of cities with population size N or higher at time t. This can be calculated as the share of cities that experience a growth rate g between time $t-1$ and time t from a size of at least N/g at $t-1$, aggregated over the different possible values of g:

$$M_t(N) = \int_0^{+\infty} M_{t-1}\left(\frac{N}{g}\right) f(g)\mathrm{d}g \,. \tag{67}$$

At the steady state (and it can be shown that there is one when cities cannot fall below a small threshold), we have:

$$M(N) = \int_0^{+\infty} M\left(\frac{N}{g}\right) f(g)\mathrm{d}g \,. \tag{68}$$

We can then verify that Zipf's law, that is $M(N) = a/N$ (where a is a constant), is the steady state we are looking for. Inserting this into equation (68) implies

$$\int_0^{+\infty} g f(g)\mathrm{d}g = 1 \,, \tag{69}$$

which must hold since aggregate population is constant.[47]

More intuitively, without a lower bound on city sizes, their distribution is single-peaked with thin tails at both ends. This is because very few cities consistently get positive or negative shocks. With a lower bound on city sizes, things change dramatically because the thin lower tail disappears and there is instead a maximum of the density function at the lower bound. Preventing cities from becoming too small also allows the upper tail to be fed by more cities. As a result, it is fatter. This lower bound also allows for the existence of a steady state instead of an ever-widening distribution.

The seemingly innocuous assumption of a lower bound on city sizes is enough to generate a very different outcome. Without a lower bound, city sizes follow a log normal distribution. With a lower bound, city sizes follow a Pareto distribution. This suggests a relative theoretical 'fragility' of these statistical processes since the final outcome depends heavily on an auxiliary assumption that will be extremely hard to test. In turn, this puts some of the empirical debates about whether the size distribution of cities is best described by a Pareto or by a log normal back into perspective.

The main alternative to the multiplicative process just described are additive processes. The first was proposed by Simon (1955). In essence, Simon's model assumes that aggregate population grows over time by discrete increments. With some probability, a new lump goes to form a new city. Otherwise it is added to an existing city. The probability that any particular city gets it is proportional to its population. This mechanism generates a Pareto distribution for city sizes. The Pareto exponent falls to one at the limit as the probability of new cities being created goes to zero.[48]

Despite important differences between them, multiplicative and additive processes both have some version of Gibrat's law at their core, either directly through multiplicative shocks or through increases of fixed size that occur proportionately to population.

[47] See Gabaix (1999*a*) for a complete proof. Note also that the same proof applies with non-constant total population if one normalizes city sizes to represent population shares instead of population numbers.

[48] For technical details, see the expositions of Krugman (1996) and Duranton (2006).

7.3 The economics of Zipf's law

Among existing models of random growth with an economic content, that proposed by Eeckhout (2004) is the simplest. There is a continuum of cities. Labour is the only factor of production. There are aggregate decreasing returns at the city level, which are modelled through congestion costs that make output decrease with elasticity $-\gamma$ with respect to city size and agglomeration economies that simultaneously make output increase with elasticity σ with respect to city size, with $\sigma < \gamma$. In addition city i experiences a labour productivity shock B_{it} at time t. Hence output per worker in city i is $B_{it}N_{it}^{\sigma-\gamma}$, where N_{it} denotes its population at time t. We note that this modelling of cities differs from what we have used so far. In the trade-off between agglomeration and dispersion, dispersion forces always dominate here and optimal city size is zero. The assumption of a fixed number of cities then becomes crucial: if workers could move to new sites, cities would disappear as larger cities only offer net disadvantages.

Free mobility then implies the equalization of output per worker across all cities. Even though each city faces shocks, the law of large numbers applies in aggregate so that output per worker is deterministic. After normalizing output per worker to unity, the equilibrium size of city i is given by:

$$N_{it} = B_{it}^{\frac{1}{\gamma-\sigma}} \ . \tag{70}$$

With small i.i.d. shocks, productivity evolves according to $B_{it} = (1 + g_{it})B_{it-1}$. It is easy to see that after T periods, we have

$$\log N_{iT} \approx \log N_{i0} + \frac{1}{\gamma - \sigma} \sum_{t=1}^{t=T} g_{it} \ . \tag{71}$$

Equation (71) is derived in the same way as equation (66). The main difference is that instead of imposing 'arbitrary' population shocks, the model assumes cumulative productivity shocks. In a setting where free mobility implies that population is a power function of productivity (equation 70), the log normal distribution of city productivity maps into a log normal distribution of city population. Consistent with the argument made above, adding a reflexive lower bound for city size to Eeckhout's (2004) model would imply a Pareto distribution instead of a log normal distribution for city sizes.

The model of Rossi-Hansberg and Wright (2007) also relies on multiplicative and cumulative productivity shocks.[49] A key difference with Eeckhout (2004) is that shocks occur for an entire industry and there are no idiosyncratic productivity differences between cities. The other main difference between this model and other random urban growth models is that it explicitly treats cities as an equilibrium between agglomeration and dispersion forces. This is important since it shows that random growth models can accommodate a standard modelling of cities. In fact, we can write a version of Rossi-Hansberg and Wright's (2007) model simply by adding random shocks to the productivity shifter in the systems of cities model we developed in section 5. If these shocks are multiplicative and cumulative, the productivity shifter in sector j evolves according to $B_{t+1}^j = (1 + g_{t+1}^j)B_t^j$, where the shocks g_t^j are identically and independently distributed. Adding

[49] Zipf's law is obtained in two cases by Rossi-Hansberg and Wright (2007). The first is the case described here with permanent industry shocks. The second is a situation with temporary shocks which affect factor accumulation. For alternative ways to generate Zipf's law with cumulative shocks see also Córdoba (2008).

a time subscript to equation (49), we can write optimal city size (and equilibrium city size in the presence of competitive developers) as

$$N_{it} = \left(B_t^j \frac{\sigma}{\tau} \right)^{\frac{1}{\gamma - \sigma}} . \tag{72}$$

Following the approach used to derive equation (66), after T periods, we have

$$\log N_{iT} \approx \log N_{i0} + \frac{1}{\gamma - \sigma} \sum_{t=1}^{t=T} g_t^j . \tag{73}$$

This is exactly as equation (71), except that now the cumulative productivity term is sector-specific instead of city-specific. Thus, again, the distribution of city sizes is log normal. Adding a lower bound for productivity by sector leads N_{iT} to instead be Pareto distributed.

Despite the similarity of equations (71) and (73), the underlying dynamics of Eeckhout (2004) and Rossi-Hansberg and Wright (2007) are quite different. In Rossi-Hansberg and Wright (2007), utility is a concave function of city size and productivity shocks are common to all cities specializing in the same sector. By equation (42), utility equalization across cities requires the value of output per worker net of commuting cost expenditures to be the same everywhere. Then, when a sector j experiences a small positive shock B_t^j, optimal size for cities specializing in that sector increases as a result. If all existing cities specializing in sector j increased their population to this new, larger, size, the resulting increase in aggregate output in that sector would lower its price so that developers in cities specializing in sector j could not compete for residents until some developers exited and output prices rose again. At the new equilibrium, there will be fewer but larger cities specializing in sector j. Sectors that have received a sequence of higher productivity shocks, have a larger optimal city size and thus fewer cities. More precisely, the Pareto distribution of sectoral productivity maps directly into a Pareto distribution for optimal city sizes through equation (73). We note that this Pareto outcome crucially relies on cities being of optimal size.

Gabaix (1999*a*) considers a model where workers are mobile only at the beginning of their life, when they need to pick a city. At time t population in city i is made up of the N_{it}^y young workers who choose to locate there and the fraction $1 - \delta$ of the previous period population who survive:

$$N_{it} = N_{it}^y + (1 - \delta) N_{it-1} . \tag{74}$$

Workers derive utility in a multiplicatively separable fashion from the consumption of a homogenous freely tradable good and from a local amenity:

$$u_{it} = A_{it} \, w_{it} . \tag{75}$$

The level of amenity in each city i, A_{it}, is independently drawn every period from a common distribution. This reduces the location choice for young workers to a static utility maximization problem. The production function is homogenous of degree one in young workers N_{it}^y and incumbent residents. For simplicity assume a Cobb-Douglas functional form: $Y_{it} = (N_{it}^y)^{\alpha} [(1 - \delta) N_{it-1}]^{1-\alpha}$. This implies the following wage for young workers:

$$w_{it} = \alpha \left(\frac{(1 - \delta) N_{it-1}}{N_{it}^y} \right)^{1-\alpha} . \tag{76}$$

The number of young workers going to each city in each period adjusts to equalize contemporaneous utility to some common level: $u_{it} = \overline{u}$. Combining this with equations (74), (75), and (76) yields the growth rate for city i between periods $t-1$ and t as

$$g_{it} \equiv \frac{N_{it} - N_{it-1}}{N_{it-1}} = (1-\delta)\left(\frac{\alpha A_{it}}{\overline{u}}\right)^{\frac{1}{1-\alpha}} - \delta \,. \tag{77}$$

This growth rate is identically and independently distributed for all cities, regardless of their size. Since we are back to the evolution of city sizes given by equation (66), city sizes follow a lognormal distribution. With a reflective lower bound for city sizes, Zipf's law applies instead. There are two differences with the previous two models. First, the shocks apply to amenities and not to technology. Second, the shocks are temporary, not permanent. An interesting part of Gabaix's model is how temporary shocks have permanent effects. This arises because the wage of young workers depends only on the ratio of young mobile workers to immobile incumbents because of constant returns in production and because young workers become immobile after their original location choice.

The models of Gabaix (1999*a*), Eeckhout (2004), and Rossi-Hansberg and Wright (2007) are the three main multiplicative random growth models. Duranton (2006, 2007) proposes two related economic mechanisms that lead to additive random growth.

Duranton (2006) builds on Romer's (1990) endogenous growth model. Discrete innovations occur with probability proportional to research activity. Local spillovers make research activity in each location proportional to the number of local products. With mobile workers and no cost nor benefits from cities, the number of local products is proportional to population. Hence, in equilibrium, small discrete innovations occur in cities with probability proportional to their population size. Note that innovations need to be discrete to avoid the law of large numbers from applying, which would eliminate the randomness from the urban growth process. Innovations lead either to local production of the new product or, with some probability, to production at a new location where some required natural resource is available. Cities grow in population as a result of the increase in labour demand for producing a new product that follows an innovation. In essence, this model puts a geographical structure on a discrete version of Romer (1990). As shown by Duranton (2006), this model maps directly into Simon (1955) and generates Zipf's law as a limit case when the probability of new city formation tends to zero.[50]

Duranton (2007) uses a related model which builds instead on the Schumpeterian growth model of Grossman and Helpman (1991*b*). In this framework, profit-driven research tries to develop the next generation of a product up a quality ladder. A success gives it a monopoly which lapses when the next innovation on the same product occurs. Products are discrete to ensure the necessary granularity for shocks to affects cities. Again, local spillovers tie research on a given product to the location of its production. The core of the model is that research might succeed in improving the products it seeks to improve (same-product innovation) or, sometimes, because of serendipity in the research process, it might succeed in improving another product (cross-product innovation).

[50]This modelling also avoids some pitfalls of Simon (1955) which converges slowly. The cumulative and exponential nature of the growth process in Romer (1990) ensures that shocks, although additive, occur more frequently as time passes, which leads to much faster convergence.

With same-product innovation, the location of activity is unchanged by innovation and successful new innovators only replace incumbent producers in the same city. With cross-product innovation, the old version of the improved product stops being produced where it used to be and the new version starts being produced in the city where the innovation took place. This typically leads to a relocation of production with a population gain for the innovating city and a loss for the city of the incumbent producer.

To prevent cities from disappearing forever, the model also assumes that there is a core product in each city that cannot move. Symmetry and the absence of other costs and benefits from cities also ensure that city population is proportional to the number of products manufactured locally.

In steady-state, this model does not quite lead to Zipf's law because the arrival of new products is not exactly proportional to city size. Because they already have more products, large cities have fewer of them to capture from elsewhere. On the other hand, the smallest cities with only one fixed product can only grow. Hence, growth is less than proportional to city size and this leads to a distribution of city sizes that is less skewed than Zipf's law. This somewhat lower expected growth at the upper end of the distribution is an empirically relevant feature of the US city size distribution (Ioannides and Overman, 2003). More generally, a calibration of the model does well at replicating the US city size distribution. Unlike other models of random growth, this model does not focus exclusively on the size distribution of cities. It also replicates the fast churning of industries across cities, a well documented fact (Simon, 2004, Duranton, 2007, Findeisen and Suedekum, 2008).

7.4 The tension between random urban growth models and other models of urban growth

The main difference between random urban growth models and the 'classical' urban growth models we considered in sections 2–6 regards the role of shocks. In the latter approaches, urban growth is driven by city characteristics and what is left unexplained is treated as a residual. In random growth models, the 'residual' is everything. Far from being a nice complementarity between two classes of models, this is a source of important tensions.

From a theoretical standpoint, it is possible to combine ingredients from traditional and random growth models to have urban growth driven by a combination of substantive determinants and random shocks. Following Duranton and Puga (2013), let us take our urban growth model of section 6.1, where human capital accumulation drives aggregate growth and urban growth, and add sector-specific random shocks. If these are multiplicative and cumulative, the productivity shifter in sector j evolves according to $B_t^j = (1 + g_t^j)B_{t-1}^j$. From equation (54), city size at time t is given by

$$N_{it} = \left(\frac{\sigma}{\tau} B_t^j (1-\delta)^{1-\alpha} h_{it}^{\alpha+\sigma}\right)^{\frac{1}{\gamma-\sigma}} \quad . \tag{78}$$

Dividing this equation valued at time T from the same equation valued at time 0, and taking logs, we obtain

$$\log N_{iT} \approx \log N_{i0} + \frac{\alpha+\sigma}{\gamma-\sigma}(\log h_{iT} - \log h_{i0}) + \frac{1}{\gamma-\sigma}\sum_{t=1}^{t=T} g_t^j \, . \tag{79}$$

Now urban growth has both a systematic component arising from human capital accumulation and a random component arising from sectoral shocks. If we assume, as in equation (52), that

human capital grows at the same rate in every city, we have $\log h_{iT} - \log h_{i0} \approx Tb\delta$. Then, imposing a lower bound to sectoral productivity results in a Pareto distribution for city sizes. At the same time, human capital accumulation makes cities experience parallel growth in expectation.[51] With this specification, there is no theoretical incompatibility between classical and random urban growth models.

However, when the rate at which human capital accumulates depends on the level of human capital in the city, the growth of a city becomes a function of its initial human capital. Assume for instance that workers can choose how much human capital to accumulate. Because of complementarities in learning, it can be that workers optimally invest more in human capital accumulation in more educated cities so that the fraction of time spent learning δ is now an increasing function of city average human capital: $\delta(\bar{h}_{it})$. Then, by the same argument as above, $\log h_{iT} - \log h_{i0} \approx Tb\, f(\bar{h}_{i0})$. We no longer obtain Zipf's law because in equation (79) the effect of this systematic driver of urban growth eventually dwarves the cumulative effect of the sectoral shocks $\frac{1}{\gamma-\sigma}\sum_{t=1}^{t=T} g_t^j$.

To understand better the tension between classical and random urban growth models, consider a simple urban growth regression:

$$\Delta_{t+1,t} \log N_i = \beta_0 + \beta_1 \log N_{it} + D_{it}\beta_2^{'} + \epsilon_{it}\ , \tag{80}$$

where the growth of city i between t and $t+1$ depends on its population size at time t, some drivers of urban growth D_{it}, and a random term ϵ_{it}. As starting point, it is useful to think of the classical urban growth models we considered in sections 2–6 as focusing on N_{it} and D_{it} whereas random growth models focus on ϵ_{it}. Formally, the question is whether Gibrat's law (and hence Zipf's law), as generated by random urban growth models, is compatible with $\beta_1 \neq 0$ or $\beta_2^{'} \neq 0$ and whether these situations are empirically relevant.

It is best to discuss the issues surrounding initial population size ($\beta_1 \neq 0$) and those regarding systematic drivers of city growth ($\beta_2^{'} \neq 0$) separately. Starting with initial population size we note that, while there is some disagreement in the literature about the importance of mean-reversion in city population data (e.g., Black and Henderson, 2003, vs. Eeckhout, 2004), past city population is more often than not a significant determinant of city growth and its coefficient often appears with a negative sign in urban growth regressions.[52]

A first source of mean-reversion could be found in measurement error. Taking the simplified version of Rossi-Hansberg and Wright (2007) presented above, 'true' population growth in city between $t-1$ and t is given by the unobserved sectoral shock g_t^j. The level of population is nonetheless observed with error so that we observe $N_{it}\, e^{\mu_{it}}$ instead of N_{it} as population at time t. If μ_{it} is i.i.d., this has two implications. First, over two consecutive periods, there is a negative correlation between growth and initial size since, for instance, a large positive measurement shock in $t-1$ makes for both a higher initial population at $t-1$ and a lower growth rate between $t-1$

[51] This is not perfectly parallel growth, since random shocks mean that expected growth rates are equal across all cities whereas actual growth rates are not.

[52] Black and Henderson (2003) find a highly significant coefficient for $\beta_1 = -0.02$ in the case of US cities across decades of the 20th century. Covering an even longer time period, both Glaeser, Ponzetto, and Tobio (2011) and Desmet and Rappaport (2013) also find significant departures from Gibrat's law.

and t. Second, the observed growth rate is $\epsilon_{it} = g_{it} + \mu_{it} - \mu_{it-1}$. In turn, this implies:

$$\log N_{iT} = \log N_{i0} + \beta_0 + \mu_{iT} - \mu_{i0} + \sum_{t=1}^{t=T} g_{it} \ . \tag{81}$$

This equation is compatible with equation (79). As argued by Gabaix and Ioannides (2004) if the tail of the summation in g is fatter than that of μ, Zipf's law should still occur in steady state. Intuitively, mean-reversion does not matter provided it is 'dominated' by the cumulated 'Gibrat's shocks'. A similar argument would hold if the population was not mismeasured but instead be subject to real temporary shocks around optimal city size. Hence, Zipf's law need not rely on a 'strong' version of Gibrat's law where $\beta_1 = 0$. Instead, it can hold with a weaker version of Gibrat's law where $\beta_1 \neq 0$. This said, much remains to be done on this issue. We need to know what is the weakest version of Gibrat's law compatible with Zipf's law. For instance, an AR(1) error structure like $\epsilon_{it} = g_{it} + \rho\epsilon_{it-1}$ does not converge to log normal for N_T without further (Gaussian) assumptions about g.[53]

Turning to the other determinants of urban growth, let us return to equation (80), assume for simplicity $\beta_1 = 0$, allow for β_2 to be time varying, and consider that $\epsilon_{it} = g_{it}$, which is i.i.d. After simplification, we obtain:

$$\log N_{iT} = \log N_{i0} + \beta_0 + \sum_{t=1}^{t=T} D_{it}\beta_{2t}^{'} + \sum_{t=1}^{t=T} g_{it} \ . \tag{82}$$

This equation corresponds to the predictions of the model of section 6.2 where the growth rate of human capital is not constant across cities but is instead driven by some city characteristic (e.g., the local presence of strong research universities). It is now easy to understand that any term $D_{it}\beta_{2t}^{'} = D_i\beta_2^{'}$ that is constant in magnitude over time and differs across cities would lead to divergence in the long-run and a distribution that differs from Zipf's law. This suggests a major incompatibility between classical and random urban growth models.

There are a number of ways around this incompatibility. First, the upper tail of the city size distribution may remain Pareto despite different growth trends. To understand this point, consider two groups of cities, fast- and slow-growing cities. Provided the lower bound city size for each group of cities grows with its trend, there is a Pareto distribution emerging for each group of cities and divergence between the two groups. At any point in time the overall distribution will be a mixture of two Pareto distributions, both with a slope coefficient minus one. Above the largest of the two lower bounds, this distribution will be Pareto.[54]

Second, classical and random urban growth models are also compatible when the effects of $D_{it}\beta_{2t}^{'}$ are short lived, that is when there is mean reversion in β_{2t} or in D_{it}. Mean-reversion in β_{2t} corresponds to the situation where a permanent characteristic has a positive effect over a period of time and negative effect over another. In the United States for instance, it is possible that hot summers deterred population growth before the development of air-conditioning but promoted

[53]Such autoregressive processes are important in this context given the strong persistence of population shocks (Rappaport, 2004).

[54]Skouras (2009) considers a different but related argument. Among groups of cities with the same constant average size, any group that follows Zipf's law will eventually dominate the upper tail.

it after this. Proximity to coal and iron was arguably a factor of growth during the late 19th and early 20th century that became irrelevant after. Glaeser, Ponzetto, and Tobio (2011) provide formal support for this argument looking at the growth of counties in the Eastern and Central United States over a two hundred year period. They show that many determinants of county population growth such as geography and climate are not stable over time.

Mean-reversion in D_{it} corresponds instead to the situation where the determinants of growth are temporary in cities. For instance, it could be that receiving roads is a factor of urban growth as suggested by Duranton and Turner (2012) but that new roads are allocated proportionately to population.[55] In a slightly different vein, a number of papers highlight the importance of specific one-off technology shocks that affect urban growth. Duranton and Puga (2005) emphasise the availability of communication technologies allowing firms to separate their management from their production activities leading cities to specialise by function and no longer by sector (see also Ioannides, Overman, Rossi-Hansberg, and Schmidheiny, 2008, for another take on the effect of communication technologies on cities). Desmet and Rossi-Hansberg (2009) focus on the maturation of economic activities which are concentrated when new and gradually diffuse as they mature. Under some conditions, a series of one-off shocks like these may be able to bridge the gap between classical and random urban growth models.

More generally, what growth regressions and classical urban models treat as explanatory variables in some cases need to be thought of as the shocks in random growth models. This observation suggests that shocks in the context random growth models need not be equated with residuals in urban growth regressions.

Different time horizons between classical and random growth models may go a long way towards making them compatible with each other. Classical urban growth models, which constitute the theoretical underpinning of standard urban growth regressions, may be looking at the growth of cities around a particular period whereas random growth models may have a much longer time horizon. In that case, classical urban growth models help us uncover short run proximate factors of urban growth whereas random growth models help us understand the fundamental mechanics that drive urban growth in the long run.

Two further possibilities can be entertained to reconcile random and classical urban growth models. The first is that there might be a number of city characteristics distributed such that the effect of the entire vector of characteristics is about the same in all cities. In that case, the underlying trend for all cities would be the same and Zipf's law would occur in steady state. While an exact equalization across the effects of all characteristics across cities would be highly unlikely, some negative correlations across drivers of urban growth are not unthinkable.[56]

The second possibility is that Zipf's law may occur as the outcome of a static model while parallel city growth occurs for entirely unrelated reasons. Hsu (2012) proposes a microfounded model of central place theory which can generate Zipf's law. Lee and Li (2013) propose a model

[55] Duranton and Turner (2012) show that the 1947 plan which guided the early development of the US interstate highway system allocated highways to cities on average proportionately to their population. More recent highway developments are clearly less than proportional to population.

[56] For instance, cities with nice landscapes are also likely to suffer from greater construction costs and more generally a greater scarcity of useable land.

where city population depends multiplicatively on their many characteristics which are i.i.d. This leads to the static counterpart of Eeckhout's (2004) model. Behrens, Duranton, and Robert-Nicoud (2012) also obtain Zipf's law in a model of sorting across cities.[57] These papers are interesting because they show that Zipf's law need not be the outcome of a random growth model but could arise for other reasons. Nevertheless, these static Zipf's law models imply strong restriction on urban growth since parallel growth is needed to retain Zipf's law. The recent empirical findings of Desmet and Rappaport (2013) are consistent with this type of argument. They find that the US city size distribution first settled to its current form and only then began to satisfy a mild form of parallel growth (through Gibrat's law).

Finally, it should be kept in mind that random growth models mainly offer theories of the growth of individual cities, not theories of the growth of all cities. For instance, random growth models have little to say about the increase in average city size over the last two hundred years. Classical urban growth models propose both theories of the growth of all cities as well as theories of the growth of particular cities. Even if random growth models turned out to be a good explanation of urban evolutions, that would not prevent better and cheaper commuting technologies to be one important driver of the growth of all cities.

8. Conclusion

We have identified four key drivers of the population growth of cities in developed economies. First, transportation and housing supply. Second, amenities. Third, agglomeration effects, in particular those related to human capital and entrepreneurship. And fourth, technology and shocks to specific cities or industries.

The empirical case for these drivers rests first on cross-city growth regressions. Identifying causal factors in regressions of city population growth in cross-section is fraught with difficulties. Applications of this type of methodology to cross-country growth in income per capita has rightfully come under attack in the past (Durlauf, Johnson, and Temple, 2005). The exercise is arguably easier in the case of city population since there is less heterogeneity in the data. For instance, data on educational achievement is more directly comparable between Baltimore and Miami and than between Belarus and Malawi. The number of explanatory factors for city population growth within a country is also much smaller than the number of possible causes of income growth across countries since many variables can be, as a first approximation, held constant within a unified country. In the last decade, the literature on city growth has also repeatedly tackled fundamental inference concerns heads on, relying in particular on instrumental variables.

The literature on drivers of city growth nicely ties into the main modelling approaches used to study the economics of cities: the monocentric model for housing and transportation, the model of cross-city compensating differentials for amenities, models of microfounded agglomeration

[57] In addition, Henderson and Venables (2009) can generate Zipf's law from an underlying power law in site quality. In a very different model, Berliant and Watanabe (2009) generate empirically relevant size distributions. In their model, cities receive shocks by industries and only the best will produce. This leads to city sizes being determined by extreme value distributions which can be parameterized to fit existing distributions. A detailed analysis of static Zipf's law models is outside our scope here

economies for agglomeration effects and human capital, and random growth models for technology and sectoral shocks. Hence, the literature reviewed in this chapter goes beyond isolating specific drivers of city growth. It also provides empirical support for the core theoretical models of cities.

In the work reviewed above, close links between theory and empirics have turned out to be very useful. They allow going beyond the estimation of the elasticity of city growth with respect to a specific driver to examine other implications of these theories. For instance, in monocentric models of cities lower transportation costs imply not only population growth but also greater suburbanisation, increased land consumption, etc. Many of these extra predictions have been examined in the literature and receive strong empirical support.

This said, the success of this literature is only partial and much remains to be done. We identify several areas of interest for future work. First, most of the theories we relied on are static and only offer predictions based on comparative statics. Related to this, many results depend crucially on workers being homogenous and perfectly mobile. Dynamic urban models with heterogeneous agents and explicit mobility costs should be a key priority for theory. This will provide new insights into the evolution of cities and help us consider adjustment processes explicitly. In turn, this will hopefully lead to new empirical approaches that push the study of urban dynamics beyond cross-city growth regressions and avoid the ambiguities that mar the interpretation of many results in the literature.

Furthermore, some potential drivers of the growth of cities are yet to be explored. The biggest gap is arguably studying the effects of municipal and city governments, local policies and public finance. In addition, many empirical results should be strengthened and alternative empirical strategies developed to confirm them. Although a convincing empirical framework that examines all existing drivers of urban growth at the same time is too ambitious a goal, exploring drivers of urban growth in isolation is not satisfactory. Some explanations need to be confronted. For instance, the links between human capital and entrepreneurship need to be clarified. Also, both infrastructure and amenities drive city growth but infrastructure-rich places are often amenity-poor and vice-versa. Engines of city growth might substitute for one another or instead, perhaps, complement each other. Understanding the relationships between drivers of urban growth is of academic interest but it could also be highly relevant to design urban growth strategies.

As argued in the introduction, the growth of cities potentially offers a unique window into the broader issue of the determinants of economic growth and technological progress. This is where the results have been least satisfactory. Little in the study of the growth of cities so far has really illuminated how growth and technological progress take place. For instance, as made clear in this chapter, there is good evidence that average education in cities has a causal effect on their subsequent population growth. While this is important and interesting to urban economists, providing useful insights for growth economists will require convincing evidence about a much more detailed causal chain looking into how innovation takes place in cities, how workers learn from each other, and how knowledge diffuses between workers.

References

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

Aghion, Philippe and Peter Howitt. 1992. A model of growth through creative destruction. *Econometrica* 60(2):323–351.

Agrawal, Ajay, Iain Cockburn, and John McHale. 2006. Gone but not forgotten: Knowledge flows, labor mobility, and enduring social relationships. *Journal of Economic Geography* 6(5):571–591.

Albouy, David. 2008. Are big cities really bad places to live? Improving quality-of-life estimates across cities. Working Paper 14472, National Bureau of Economic Research.

Alonso, William. 1964. *Location and Land Use; Toward a General Theory of Land Rent*. Cambridge, MA: Harvard University Press.

Altonji, Joseph G. and David Card. 1991. The effects of immigration on the labor market outcomes of less-skilled natives. In John M. Abowd and Richard B. Freeman (eds.) *Immigration, Trade and the Labor Market*. Chicago, IL: Chicago University Press, 201–234.

Anas, Alex, Richard Arnott, and Kenneth A. Small. 1998. Urban spatial structure. *Journal of Economic Literature* 36(3):1426–1464.

Angrist, Joshua D. and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Arnott, Richard J. and Joseph E. Stiglitz. 1979. Aggregate land rents, expenditure on public goods, and optimal city size. *Quarterly Journal of Economics* 93(4):471–500.

Arnott, Richard J. and Joseph E. Stiglitz. 1981. Aggregate land rents and aggregate transport costs. *Economic Journal* 91(362):331–347.

Auerbach, Felix. 1913. Das Gesetz der Bevölkerungskonzentration. *Petermanns Geographische Mitteilungen* 59:73–76.

Bairoch, Paul. 1988. *Cities and Economic Development: From the Dawn of History to the Present*. Chicago: University of Chicago Press.

Baldwin, Richard E. 2001. Core-periphery model with forward-looking expectations. *Regional Science and Urban Economics* 31(1):21–49.

Barro, Robert J. 1991. Economic-growth in a cross-section of countries. *Quarterly Journal of Economics* 106(2):407–443.

Bartik, Timothy. 1991. *Who Benefits from State and Local Economic Development Policies?* Kalamazoo (MI): W. E. Upjohn Institute for Employment Research.

Baum-Snow, Nathaniel. 2007. Did highways cause suburbanization? *Quarterly Journal of Economics* 122(2):775–805.

Baum-Snow, Nathaniel and Byron F. Lutz. 2011. School desegregation, school choice, and changes in residential location patterns by race. *American Economic Review* 101(7):3019–3046.

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

Behrens, Kristian, Gilles Duranton, and Frédéric Robert-Nicoud. 2012. Productive cities: Sorting, selection, and agglomeration. Processed, Wharton School, University of Pennsylvania.

Berliant, Marcus and Hiroki Watanabe. 2009. Explaining the size distribution of cities: X-treme economies. Processed, Washington University in St. Louis.

Black, Duncan and J. Vernon Henderson. 1999. A theory of urban growth. *Journal of Political Economy* 107(2):252–284.

Black, Duncan and Vernon Henderson. 2003. Urban evolution in the USA. *Journal of Economic Geography* 3(4):343–372.

Blomquist, Glenn C., Mark C. Berger, and John P. Hoehn. 1988. New estimates of quality of life in urban areas. *American Economic Review* 78(1):89–107.

Boustan, Leah Platt. 2010. Was postwar suburbanization "white flight"? Evidence from the black migration. *Quarterly Journal of Economics* 125(1):417–443.

Brainard, Lael S. 1997. An empirical assessment of the proximity-concentration trade-off between multinational sales and trade. *American Economic Review* 87(4):520–544.

Brueckner, Jan K. and Stuart S. Rosenthal. 2009. Gentrification and neighborhood housing cycles: will America's future downtowns be rich? *Review of Economics and Statistics* 91(4):725–743.

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

Card, David. 2001. Immigrant inflows, native outflows, and the local labor market impacts of higher immigration. *Journal of Labor Economics* 19(1):22–64.

Carlino, Gerald A. and William R. Kerr. 2013. Agglomeration and innovation. Processed, Harvard University.

Carlino, Gerald A. and Albert Saiz. 2008. City beautiful. Working Paper 08–22, Federal Reserve Bank of Philadelphia.

Cheshire, Paul C. and Stefano Magrini. 2006. Population growth in European cities: Weather matters - but only nationally. *Regional Studies* 40(1):23–37.

Chinitz, Benjamin. 1961. Contrasts in agglomeration: New York and Pittsburgh. *American Economic Review Papers and Proceedings* 51(2):279–289.

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

Cingano, Federico and Fabiano Schivardi. 2004. Identifying the sources of local productivity growth. *Journal of the European Economic Association* 2(4):720–742.

Combes, Pierre-Philippe. 2000. Economic structure and local growth: France, 1984–1993. *Journal of Urban Economics* 47(3):329–355.

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

Combes, Pierre-Philippe, Gilles Duranton, and Laurent Gobillon. 2012*a*. The costs of agglomeration: Land prices in French cities. Processed, University of Pennsylvania.

Combes, Pierre-Philippe, Gilles Duranton, Laurent Gobillon, Diego Puga, and Sébastien Roux. 2012*b*. The productivity advantages of large cities: Distinguishing agglomeration from firm selection. *Econometrica* 80(6):2543–2594.

Combes, Pierre-Philippe, Gilles Duranton, Laurent Gobillon, and Sébastien Roux. 2010. Estimating agglomeration effects with history, geology, and worker fixed-effects. In Edward L. Glaeser (ed.) *Agglomeration Economics*. Chicago, IL: Chicago University Press, 15–65.

Combes, Pierre-Philippe, Thierry Magnac, and Jean-Marc Robin. 2004. The dynamics of local employment in France. *Journal of Urban Economics* 56(2):217–243.

Córdoba, Juan-Carlos. 2008. On the distribution of city sizes. *Journal of Urban Economics* 63(1):177–197.

Cuberes, David. 2011. Sequential city growth: Empirical evidence. *Journal of Urban Economics* 69(2):229–239.

Cullen, Julie Berry and Stephen D. Levitt. 1999. Crime, urban flight, and the consequences for cities. *Review of Economics and Statistics* 81(2):159–169.

Davis, Morris, Jonas D.M. Fisher, and Toni M. Whited. 2011. Macroeconomic implications of agglomeration. Processed, University of Wisconsin.

De la Roca, Jorge and Diego Puga. 2012. Learning by working in big cities. Processed, CEMFI.

de Vries, Jan. 1984. *European Urbanization, 1500–1800*. London: Methuen.

Desmet, Klaus and Jordan Rappaport. 2013. The settlement of the United States, 1800 to 2000: The long transition towards Gibrat's law. Discussion Paper 9353, Centre for Economic Policy Research.

Desmet, Klaus and Esteban Rossi-Hansberg. 2009. Spatial growth and industry age. *Journal of Economic Theory* 144(6):2477–2502.

Diamond, Rebecca. 2013. The determinants and welfare implications of US workers' diverging location choices by skill: 1980–2000. Processed, Harvard University.

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

Duby, George. 1981–1983. *Histoire de la France Urbaine*. Paris: Le Seuil.

Duranton, Gilles. 2006. Some foundations for Zipf's law: Product proliferation and local spillovers. *Regional Science and Urban Economics* 36(4):542–563.

Duranton, Gilles. 2007. Urban evolutions: The fast, the slow, and the still. *American Economic Review* 97(1):197–221.

Duranton, Gilles. 2013. Delineating metropolitan areas: Measuring spatial labour market networks through commuting patterns. Processed, Wharton School, University of Pennsylvania.

Duranton, Gilles, Peter M. Morrow, and Matthew A. Turner. 2013. Roads and trade: Evidence from the US. Processed, University of Toronto.

Duranton, Gilles and Diego Puga. 2000. Diversity and specialisation in cities: Why, where and when does it matter? *Urban Studies* 37(3):533–555.

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

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

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

Duranton, Gilles and Diego Puga. 2013. Urban growth: systematic, idiosyncratic and random determinants and their aggregate implications. Processed, Wharton School, University of Pennsylvania.

Duranton, Gilles and Matthew A. Turner. 2011. The fundamental law of road congestion: Evidence from US cities. *American Economic Review* 101(6):2616–2652.

Duranton, Gilles and Matthew A. Turner. 2012. Urban growth and transportation. *Review of Economic Studies* 79(4):1407–1440.

Durlauf, Steven N., Paul A. Johnson, and Jonathan R. W. Temple. 2005. Growth econometrics. In Philippe Aghion and Steven N. Durlauf (eds.) *Handbook of Economic Growth*, volume 1. Amsterdam: North-Holland, 555–677.

Eaton, Jonathan and Zvi Eckstein. 1997. Cities and growth: Theory and evidence from France and Japan. *Regional Science and Urban Economics* 27(4–5):443–474.

Eeckhout, Jan. 2004. Gibrat's law for (All) cities. *American Economic Review* 94(5):1429–1451.

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

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

Findeisen, Sebastian and Jens Suedekum. 2008. Industry churning and the evolution of cities: Evidence for Germany. *Journal of Urban Economics* 64(2):326–339.

Fischel, William A. 2000. Zoning and land use regulations. In Bouckaert Boudewijn and Gerrit De Geest (eds.) *Encycolopedia of Law and Economics*, volume 2. Cheltenham: Edward Elgar, 403–442.

Flatters, Frank, J. Vernon Henderson, and Peter Mieszkowski. 1974. Public goods, efficiency, and regional fiscal equalization. *Journal of Public Economics* 3(2):99–112.

Fujita, Masahisa. 1988. A monopolistic competition model of spatial agglomeration: A differentiated product approach. *Regional Science and Urban Economics* 18(1):87–124.

Fujita, Masahisa. 1989. *Urban Economic Theory: Land Use and City Size*. Cambridge: Cambridge University Press.

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

Fujita, Masahisa and Jacques-François Thisse. 2002. *Economics of Agglomeration: Cities, Industrial Location, and Regional Growth*. Cambridge: Cambridge University Press.

Gabaix, Xavier. 1999*a*. Zipf's law for cities: An explanation. *Quarterly Journal of Economics* 114(3):739–767.

Gabaix, Xavier. 1999*b*. Zipf's law and the growth of cities. *American Economic Review Papers and Proceedings* 89(2):129–132.

Gabaix, Xavier and Rustam Ibragimov. 2011. Rank-1/2: A simple way to improve the OLS estimation of tail exponents. *Journal of Business Economics and Statistics* 29(1):24–39.

Gabaix, Xavier and Yannis M. Ioannides. 2004. The evolution of city size distributions. In Vernon Henderson and Jacques-François Thisse (eds.) *Handbook of Regional and Urban Economics*, volume 4. Amsterdam: North-Holland, 2341–2378.

Gibrat, Robert. 1931. *Les inégalités économiques; applications: aux inégalités des 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*. Paris: Librairie du Recueil Sirey.

Giesen, Kristian, Arndt Zimmermann, and Jens Suedekum. 2010. The size distribution across all cities — Double Pareto lognormal strikes. *Journal of Urban Economics* 68(2):129–137.

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

Glaeser, Edward L. and Joseph Gyourko. 2005. Urban decline and durable housing. *Journal of Political Economy* 113(2):345–375.

Glaeser, Edward L., Joseph Gyourko, and Raven Saks. 2005. Why is Manhattan so expensive? Regulation and the rise in housing prices. *Journal of Law and Economics* 48(2):331–369.

Glaeser, Edward L., Joseph Gyourko, and Raven E. Saks. 2006. Urban growth and housing supply. *Journal of Economic Geography* 6(1):71–89.

Glaeser, Edward L. and Matthew Kahn. 2001. Decentralized employment and the transformation of the American city. *Brookings-Wharton Papers on Urban Affairs* :1–47.

Glaeser, Edward L. and Matthew E. Kahn. 2004. Sprawl and urban growth. In Vernon Henderson and Jacques-François Thisse (eds.) *Handbook of Regional and Urban Economics*, volume 4. Amsterdam: North-Holland, 2481–2527.

Glaeser, Edward L., Matthew E. Kahn, and Jordan Rappaport. 2008. Why do the poor live in cities? The role of public transportation. *Journal of Urban Economics* 63(1):1–24.

Glaeser, Edward L., Heidi Kallal, José A. Scheinkman, and Andrei Schleifer. 1992. Growth in cities. *Journal of Political Economy* 100(6):1126–1152.

Glaeser, Edward L. and William R. Kerr. 2009. Local industrial conditions and entrepreneurship: How much of the spatial distribution can we explain? *Journal of Economics and Management Strategy* 18(3):623–663.

Glaeser, Edward L., William R. Kerr, and Giacomo A. M. Ponzetto. 2010. Clusters of entrepreneurship. *Journal of Urban Economics* 67(1):150–168.

Glaeser, Edward L., Jed Kolko, and Albert Saiz. 2001. Consumer city. *Journal of Economic Geography* 1(1):27–50.

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

Glaeser, Edward L., Sari Pekkala Kerr, and William R. Kerr. 2012. Entrepreneurship and urban growth: An empirical assessment with historical mines. Processed, Harvard University.

Glaeser, Edward L., Giacomo A. M. Ponzetto, and Kristina Tobio. 2011. Cities, skills, and regional change. Processed, Harvard University.

Glaeser, Edward L. and Albert Saiz. 2004. The rise of the skilled city. *Brookings-Wharton Papers on Urban Affairs* 5:47–95.

Glaeser, Edward L., José A. Scheinkman, and Andrei Shleifer. 1995. Economic-growth in a cross-section of cities. *Journal of Monetary Economics* 36(1):117–143.

Glaeser, Edward L. and Kristina Tobio. 2008. The rise of the sunbelt. *Southern Economic Journal* 74(3):610–643.

Glaeser, Edward L. and Bryce A. Ward. 2009. The causes and consequences of land use regulation: Evidence from Greater Boston. *Journal of Urban Economics* 65(3):265–278.

Greenstone, Michael, Richard Hornbeck, and Enrico Moretti. 2010. Identifying agglomeration spillovers: Evidence from winners and losers of large plant openings. *Journal of Political Economy* 118(3):536–598.

Grossman, Gene M. and Elhanan Helpman. 1991*a*. *Innovation and Growth in the World Economy*. Cambridge, MA: MIT Press.

Grossman, Gene M. and Elhanan Helpman. 1991*b*. Quality ladders in the theory of growth. *Review of Economic Studies* 58(1):43–61.

Gyourko, Joseph, Christopher Mayer, and Todd Sinai. 2013. Superstar cities. *American Economic Journal: Economic Policy* (forthcoming).

Gyourko, Joseph, Albert Saiz, and Anita A. Summers. 2008. A new measure of the local regulatory environment for housing markets: The Wharton residential land use regulatory index. *Urban Studies* 45(3):693–729.

Helpman, Elhanan. 1998. The size of regions. In David Pines, Efraim Sadka, and Itzhak Zilcha (eds.) *Topics in Public Economics*. New York, NY: Cambridge University Press, 33–54.

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

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

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

Henderson, J. Vernon. 2005. Urbanization and growth. In Philippe Aghion and Steven N. Durlauf (eds.) *Handbook of Economic Growth*, volume 1B. Amsterdam: North-Holland, 1543–1591.

Henderson, J. Vernon, Ari Kuncoro, and Matt Turner. 1995. Industrial development in cities. *Journal of Political Economy* 103(5):1067–1090.

Henderson, J. Vernon and Anthony J. Venables. 2009. The dynamics of city formation. *Review of Economic Dynamics* 39(2):233–254.

Henderson, J. Vernon and Hyoung Gun Wang. 2007. Urbanization and city growth: The role of institutions. *Regional Science and Urban Economics* 37(3):283–313.

Hilber, Christian and Frédéric Robert-Nicoud. 2013. On the origins of land use regulations: Theory and evidence from US metro areas. *Journal of Urban Economics* 75(1):29–43.

Holmes, Thomas J. 1998. The effect of state policies on the location of manufacturing: Evidence from state borders. *Journal of Political Economy* 106(4):667–705.

Hsu, Wen-Tai. 2012. Central place theory and city size distribution. *Economic Journal* 122(563):903–932.

Ioannides, Yannis and Spyros Skouras. 2013. US city size distribution: Robustly Pareto, but only in the tail. *Journal of Urban Economics* 73(1):18–29.

Ioannides, Yannis M. and Henry G. Overman. 2003. Zipf's law for cities: an empirical examination. *Regional Science and Urban Economics* 33(2):127–137.

Ioannides, Yannis M., Henry G. Overman, Esteban Rossi-Hansberg, and Kurt Schmidheiny. 2008. The effect of ICT on urban structure. *Economic Policy* 23(54):201–242.

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

Jovanovic, Boyan and Yaw Nyarko. 1995. The transfer of human capital. *Journal of Economic Dynamics and Control* 19(5–7):1033–1064.

Jovanovic, Boyan and Rafael Rob. 1989. The growth and diffusion of knowledge. *Review of Economic Studies* 56(4):569–582.

Kerr, William R. 2010. Breakthrough inventions and migrating clusters of innovation. *Journal of Urban Economics* 67(1):46–60.

Kesten, Harry. 1973. Random difference equations and renewal theory for products of random matrices. *Acta Mathematica* 131(1):207–248.

Krugman, Paul. 1996. Confronting the mystery of urban hierarchy. *Journal of the Japanese and International Economies* 10(4):1120–1171.

Krugman, Paul R. 1980. Scale economies, product differentiation, and the pattern of trade. *American Economic Review* 70(5):950–959.

Leamer, Edward E. and James Levinsohn. 1995. International trade theory: The evidence. In Gene M. Grossman and Kenneth Rogoff (eds.) *Handbook of International Economics*, volume 3. Amsterdam: North-Holland, 1339–1394.

Lee, Sanghoon and Qiang Li. 2013. Uneven landscapes and the city size distribution. *Journal of Urban Economics* (forthcoming).

LeRoy, Stephen F. and Jon Sonstelie. 1983. The effects of urban spatial structure on travel demand in the United States. *Journal of Urban Economics* 13:67–89.

Lucas, Robert E., Jr. 1988. On the mechanics of economic development. *Journal of Monetary Economics* 22(1):3–42.

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

Mankiw, N. Gregory and David Romer David N. Weil. 1992. A contribution to the empirics of economic growth. *Quarterly Journal of Economics* 107(2):407–437.

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

McMillen, Daniel P. 2001. Nonparametric employment subcenter indentification. *Journal of Urban Economics* 50(3):448–473.

McMillen, Daniel P. 2006. Testing for monocentricity. In Richard J. Arnott and Daniel P. McMillen (eds.) *A Companion to Urban Economics*. Oxford: Blackwell, 128–140.

McMillen, Daniel P. and Stefani C. Smith. 2003. The number of subcenters in large urban areas. *Journal of Urban Economics* 53(3):332–342.

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

Mills, Edwin S. 1967. An aggregative model of resource allocation in a metropolitan area. *American Economic Review Papers and Proceedings* 57(2):197–210.

Moretti, Enrico. 2004*a*. Estimating the social return to higher education: Evidence from longitudinal and repeated cross-sectional data. *Journal of Econometrics* 121(1):175–212.

Moretti, Enrico. 2004*b*. Workers' education, spillovers, and productivity: Evidence from plant-level production functions. *American Economic Review* 94(3):656–690.

Moretti, Enrico. 2011. Local labor markets. In Orley Ashenfelter and David Card (eds.) *Handbook of Labor Economics*, volume 4. Amsterdam: Elsevier, 1237–1313.

Murata, Yasusada, Ryo Nakajima, Ryosuke Okamoto, and Ryuichi Tamura. 2013. Localized knowledge spillovers and patent citations: A distance-based approach. Processed, Nihon University.

Muth, Richard F. 1969. *Cities and Housing*. Chicago: University of Chicago Press.

Ortalo-Magné, François and Andrea Prat. 2010. The political economy of housing supply. Processed, University of Wisconsin.

Ottaviano, Gianmarco I. P. and Giovanni Peri. 2006. The economic value of cultural diversity: evidence from US cities. *Journal of Economic Geography* 6(1):9–44.

Porter, Michael. 1990. *The Competitive Advantage of Nations*. New York: Free Press.

Puga, Diego. 2010. The magnitude and causes of agglomeration economies. *Journal of Regional Science* 50(1):203–219.

Rappaport, Jordan. 2004. Why are population flows so persistent? *Journal of Urban Economics* 56(3):554–580.

Rappaport, Jordan. 2007. Moving to nice weather. *Regional Science and Urban Economics* 37(3):375–398.

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

Redding, Stephen J. and Daniel M. Sturm. 2008. The costs of remoteness: Evidence from German division and reunification. *Journal of International Economics* 98(5):1766–1797.

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

Romer, Paul M. 1986. Increasing returns and long-run growth. *Journal of Political Economy* 94(5):1002–1037.

Romer, Paul M. 1990. Endogenous technological-change. *Journal of Political Economy* 98(5):S71–S102.

Rosen, Kenneth T. and Mitchel Resnick. 1980. The size distribution of cities — An examination of the Pareto law and primacy. *Journal of Urban Economics* 8(2):165–186.

Rosen, Sherwin. 1979. Wage-based indexes of urban quality of life. In Peter N. Miezkowski and Mahlon R. Straszheim (eds.) *Current Issues in Urban Economics*. Baltimore, MD: Johns Hopkins University Press, 74–104.

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

Rosenthal, Stuart S. and William C. Strange. 2010. Small establishments/big effects: Agglomeration, industrial organization and entrepreneurship. In Edward L. Glaeser (ed.) *Agglomeration Economics*. Chicago, IL: Chicago University Press, 277–302.

Rossi-Hansberg, Esteban and Mark L. J. Wright. 2007. Urban structure and growth. *Review of Economic Studies* 74(2):597–624.

Rozenfeld, Hernán D., Diego Rybski, Xavier Gabaix, and Hernán A. Makse. 2011. The area and population of cities: New insights from a different perspective on cities. *American Economic Review* 101(5):2205–2225.

Saiz, Albert. 2010. The geographic determinants of housing supply. *Quarterly Journal of Economics* 125(3):1253–1296.

Serck-Hanssen, Jan. 1969. The optimal number of factories in a spatial market. In Hendricus C. Bos (ed.) *Towards Balanced International Growth*. Amsterdam: North-Holland, 269–282.

Shapiro, Jesse M. 2006. Smart cities: Quality of life, productivity, and the growth effects of human capital. *Review of Economics and Statistics* 88(2):324–335.

Simon, Curtis J. 1998. Human capital and metropolitan employment growth. *Journal of Urban Economics* 43(2):223–243.

Simon, Curtis J. 2004. Industrial reallocation across US cities, 1977–1997. *Journal of Urban Economics* 56(1):119–143.

Simon, Curtis J. and Clark Nardinelli. 1996. The talk of the town: Human capital, information, and the growth of English cities, 1861 to 1961. *Explorations in Economic History* 33(3):384–413.

Simon, Curtis J. and Clark Nardinelli. 2002. Human capital and the rise of American cities: 1900–1990. *Regional Science and Urban Economics* 32(1):59–96.

Simon, Herbert. 1955. On a class of skew distribution functions. *Biometrika* 42(2):425–440.

Skouras, Spyros. 2009. Explaining Zipf's law for US cities. Processed, Athens University of Economics and Business.

Soo, Kwok Tong. 2005. Zipf's law for cities: A cross country investigation. *Regional Science and Urban Economics* 35(3):239–263.

Starrett, David A. 1974. Principles of optimal location in a large homogeneous area. *Journal of Economic Theory* 9(4):418–448.

Stiglitz, Joseph E. 1977. The theory of local public goods. In Martin S. Feldstein and Robert P. Inman (eds.) *The Economics of Public Services*. London: MacMillan Press, 274–333.

Storper, Michael and Allen J. Scott. 2009. Rethinking human capital, creativity and urban growth. *Journal of Economic Geography* 9(2):47–167.

Sutton, John. 1991. *Sunk Costs and Market Structure*. Cambridge, MA: The MIT Press.

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

Syverson, Chad. 2004. Market structure and productivity: A concrete example. *Journal of Political Economy* 112(6):1181–1222.

Thompson, Peter and Melanie Fox-Kean. 2005. Patent citations and the geography of knowledge spillovers: A reassessment. *American Economic Review* 95(1):450–460.

Thünen, Johann H., von. 1826. *Der Isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie*. Hamburg: Perthes. English Translation: *The Isolated State*, Oxford: Pergammon Press, 1966.

Vickrey, William S. 1977. The city as a firm. In Martin S. Feldstein and Robert P. Inman (eds.) *The Economics of Public Services*. London: MacMillan Press, 334–343.

Zipf, George Kingsley. 1949. *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology*. Cambridge: Addison Wesley.