DISCUSSION PAPER SERIES

DP14215

URBAN GROWTH AND ITS AGGREGATE IMPLICATIONS

Diego Puga and Gilles Duranton

INTERNATIONAL TRADE AND REGIONAL ECONOMICS

MACROECONOMICS AND GROWTH

ISSN 0265-8003

URBAN GROWTH AND ITS AGGREGATE IMPLICATIONS

Diego Puga and Gilles Duranton

Discussion Paper DP14215
Published 19 December 2019
Submitted 16 December 2019

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

This Discussion Paper is issued under the auspices of the Centre's research programmes:

- International Trade and Regional Economics
- Macroeconomics and Growth

Any opinions expressed here are those of the author(s) and not those of the Centre for Economic Policy Research. Research disseminated by CEPR may include views on policy, but the Centre itself takes no institutional policy positions.

The Centre for Economic Policy Research was established in 1983 as an educational charity, to promote independent analysis and public discussion of open economies and the relations among them. It is pluralist and non-partisan, bringing economic research to bear on the analysis of medium- and long-run policy questions.

These Discussion Papers often represent preliminary or incomplete work, circulated to encourage discussion and comment. Citation and use of such a paper should take account of its provisional character.

Copyright: Diego Puga and Gilles Duranton

URBAN GROWTH AND ITS AGGREGATE IMPLICATIONS

Abstract

We develop an urban growth model where human capital spillovers foster entrepreneurship and learning in heterogenous cities. Incumbent residents limit city expansion through planning regulations so that commuting and housing costs do not outweigh productivity gains. The model builds on strong microfoundations, matches key regularities at the city and economy-wide levels, and generates novel predictions for which we provide evidence. It can be quantified relying on few parameters, provides a basis to estimate the main ones, and remains transparent regarding its mechanisms. We examine various counterfactuals to assess quantitatively the effect of cities on economic growth and aggregate income.

JEL Classification: C52, R12, D24

Keywords: urban growth, agglomeration economies, Urban costs, planning regulations, city size distributions

Diego Puga - diego.puga@cemfi.es
CEMFI and CEPR

Gilles Duranton - duranton@wharton.upenn.edu
University of Pennsylvania and CEPR

Acknowledgements
Puga gratefully acknowledges funding from the European Research Council under the European Union's Horizon 2020 Programme (ERC Advanced Grant agreement 695107 – DYNURBAN) and from Spain's Ministry of Science, Innovation and Universities (grants ECO2013-41755-P, ECO2016-80411-P and PRX19-00578), as well as the support and hospitality of the Wharton School's Department of Real Estate during his visit as Judith C. and William G. Bollinger Visiting Professor. We are grateful to Xinzhu Chen, Yan Hu, Junhui Yang, and Jungsoo Yoo for research assistance, to Jorge De la Roca for advice on the NLSY79 and CPS data, to Matt Kahn and Giacomo Ponzetto for very helpful discussions, and to Morris Davis, Vernon Henderson, Diego Restuccia and seminar and conference participants for useful comments.

Urban growth and its aggregate implications

Gilles Duranton[*‡]
University of Pennsylvania

Diego Puga[*§]
CEMFI

16 December 2019

Abstract: We develop an urban growth model where human capital spillovers foster entrepreneurship and learning in heterogenous cities. Incumbent residents limit city expansion through planning regulations so that commuting and housing costs do not outweigh productivity gains. The model builds on strong microfoundations, matches key regularities at the city and economy-wide levels, and generates novel predictions for which we provide evidence. It can be quantified relying on few parameters, provides a basis to estimate the main ones, and remains transparent regarding its mechanisms. We examine various counterfactuals to assess quantitatively the effect of cities on economic growth and aggregate income.

Key words: urban growth, agglomeration economies, urban costs, planning regulations, city size distributions

JEL classification: C52, R12, D24

[*]Puga gratefully acknowledges funding from the European Research Council under the European Union's Horizon 2020 Programme (ERC Advanced Grant agreement 695107 – DYNURBAN) and from Spain's Ministry of Science, Innovation and Universities (grants ECO2013-41755-P, ECO2016-80411-P and PRX19-00578), as well as the support and hospitality of the Wharton School's Department of Real Estate during his visit as Judith C. and William G. Bollinger Visiting Professor. We are grateful to Xinzhu Chen, Yan Hu, Junhui Yang, and Jungsoo Yoo for research assistance, to Jorge De la Roca for advice on the NLSY79 and CPS data, to Matt Kahn and Giacomo Ponzetto for very helpful discussions, and to Morris Davis, Vernon Henderson, Diego Restuccia and seminar and conference participants for useful comments.

[‡]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/).

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

1. Introduction

Urbanisation and economic growth are tightly linked. Much of the increase in number and population sizes of cities results from the process of economic growth and development (Bairoch, 1988, Henderson, 2005, Desmet and Henderson, 2015). However, some urban scholars have made the claim that causation could go, in part, in the opposite direction, with cities and urbanisation being a primary engine of economic growth (Marshall, 1890, Jacobs, 1969, Lucas, 1988, Glaeser, 2011). This claim relies on the notion that the learning and human capital spillovers occurring in cities are fundamental to the creation of new ideas and the entrepreneurship which underpin higher incomes and economic growth.

Despite widespread interest, isolating the aggregate implications of the number and population sizes of cities on economic growth and aggregate income has proved elusive. Given the general equilibrium nature of the problem, the micro-mechanisms that generate productivity and innovation advantages of cities and the empirical estimates we have for them do not immediately map into aggregate implications (see Carlino and Kerr, 2015, Combes and Gobillon, 2015, for reviews). An alternative approach would be to estimate the contribution of cities to aggregate outcomes directly from aggregate data. Unfortunately, attempts such as regressing the rate of output growth of countries on characteristics of their cities, while suggestive, have fallen into the usual pitfalls of cross-country regressions (Durlauf, Johnson, and Temple, 2005).

In this paper, we propose a new model of how cities and urbanisation interact with aggregate income and economic growth. Our model relies on strong microfoundations to represent individual cities, matches key empirical regularities at the city and economy-wide levels, and generates novel predictions for which we provide evidence. Most importantly, this model is amenable to a quantification that relies a small number of parameters and remains transparent regarding the mechanisms at work. We directly estimate important parameters, which determine the magnitude of urban benefits and costs. This then allows us to assess quantitatively the effect of cities and urbanisation on economic growth and aggregate income and examine a variety of counterfactuals. Let us develop these points in more detail.

Consistent with suggestions from the empirical literature, we model the agglomeration benefits of cities as arising from human capital spillovers. These spillovers foster entrepreneurship which, in turn, leads to higher city productivity (Moretti, 2004*a*,*b*, Gennaioli, La Porta, Lopez-de-Silanes, and Shleifer, 2013). As cities grow in population, they facilitate learning and further human capital accumulation (Glaeser and Maré, 2001, Baum-Snow and Pavan, 2012, De la Roca and Puga, 2017), magnifying economic growth.

While research characterising and quantifying the costs of larger cities is scarcer, our modelling pays particular attention to the disadvantages of city population growth. We consider several components that vary in strength within and across cities, as well as over time. Transportation costs inside cities have been found to be an important determinant of urban growth (Duranton and Turner, 2012). Within each city, as emphasised by the standard monocentric city model (Alonso, 1964, Muth, 1969), more central locations feature better accessibility in exchange for more expensive homes. In practice, however, not everyone works centrally, so transport requirements

do not increase in proportion with distance to the city centre. Our modelling takes this into account, as well as the fact that congestion causes transport costs to increase with the number of travellers. Across cities, the cost of housing with a given accessibility also increases significantly with city population (Combes, Duranton, and Gobillon, 2019). Over time, travel speed evolves with technology, which, together with rising incomes, also changes the value of travel time. These elements affect the relationship between urban costs and city population in the cross section of cities and also the long-term evolution of the urban system.

With both benefits and costs to city size, our model incorporates what Fujita and Thisse (2002) call the 'fundamental tradeoff' of urban economics.[1] To resolve this tradeoff, models of urban systems in the tradition of Henderson (1974) often rely on city developers to deliver the socially optimal number and population sizes of cities. Becker and Henderson (2000) show that the equilibrium outcome with developers would also be obtained if local governments actively set local population levels to maximise local incomes. However, the equivalence between what is delivered by city developers, local governments, and a social planner breaks down once we allow for heterogeneity across cities (Albouy, Behrens, Robert-Nicoud, and Seegert, 2019).

Taking this into consideration, we propose a political economy mechanism where endogenously-determined planning regulations balance the greater commuting and housing costs associated with larger cities against agglomeration benefits. 'Incumbent residents' of more productive cities regulate land use to limit entry into their city, thereby maximising their own welfare at the expense of potential newcomers.

Our modelling of city formation through a local political process is intuitively appealing and implies novel predictions regarding patterns of planning regulations, land prices, and housing development in cities. More specifically, more productive and larger cities tend to impose more restrictive planning regulations to avoid seeing their higher productivity dissipated in urban costs. In turn, more stringent regulations translate into higher land prices at the periphery of more populated cities, unlike in standard models of land use where cities are allowed to expand until the best use for land is no longer urban (Alonso, 1964, Muth, 1969). Finally, the systematic variation in planning regulations with the productivity and population size of individual cities implies that there should be little relationship between housing prices at the periphery of cities and new housing construction. Using US data, we find empirical support for all these predictions.

Beyond providing a realistic basis for our modelling of cities, the microfoundations on which we build our framework help us distinguish between static and dynamic effects of agglomeration. In our model, human capital spillovers, which are at the root of urban agglomeration, affect the level of aggregate income, directly through their effect on city productivity, and indirectly through the population size of cities. Human capital spillovers also affect the rate at which aggregate income grows. Our microfoundations allow to disentangle these different effects. Stronger agglomeration

[1] We do not model the relative geographical position of cities (Fujita, Krugman, and Mori, 1999, Nagy, 2017) nor their sectoral specialisation (Becker and Henderson, 2000, Duranton and Puga, 2001, 2005). We also leave aside consumption amenities and their role in urban development (Glaeser, Kolko, and Saiz, 2001, Cheshire and Magrini, 2006, Rappaport, 2007, Carlino and Saiz, 2019, Couture and Handbury, 2019). Finally, we do not consider sorting across cities by skills or occupation (Behrens, Duranton, and Robert-Nicoud, 2014, Davis and Dingel, 2019), instead focusing our exploration of inequalities on those that arise between incumbent residents and potential migrants and across cities.

effects lead to larger cities and increased output but some of that output growth is dissipated into higher urban costs. Thus, our microfoundations are also useful to distinguish between the gross and net benefits of larger cities and to make welfare pronouncements.

Because the magnitudes of urban costs and agglomeration benefits are fundamental to establish the contribution of cities to aggregate growth, we directly estimate parameters capturing this tradeoff in our model.

Regarding urban costs, we implement three novel and complementary approaches based on equations of the model at different levels of aggregation and using different sources of variation, all of which yield almost identical estimates. These approaches amount to estimating our initial commuting cost equation (using within-city variation in travel distance across individuals), the spatial equilibrium within each city (using within-city variation in house prices across locations), and the spatial equilibrium across cities (using cross-city variation in city-centre house prices). All three approaches result in a similar elasticity of urban costs with respect to city population of about 7%. These urban costs are then further amplified by congestion with a population elasticity, that we also estimate, of about 4%.

Regarding agglomeration economies, we implement the approach of De la Roca and Puga (2017) using US microdata. We estimate a short-term elasticity of earnings with respect to city population close to 5%, and an elasticity in the longer term, incorporating learning effects, of close to 8%. This is in line with previous estimates for other countries (Combes and Gobillon, 2015, De la Roca and Puga, 2017).

The remaining parameters are the population elasticity of income in rural areas, and the rates of change in output per person, in city populations, and in transportation costs per unit of distance. When they are not directly observable, we obtain them from the literature or calibrate them to some specific moments of the data.

We also show our model can match key regularities and magnitudes at the individual, city, and economy-wide levels, and help us assess quantitatively the effect of cities on economic growth and aggregate income. Our equilibrium replicates key stylised facts about systems of cities. As the economy develops and aggregate population grows, new cities appear, while a dwindling proportion of the population remains in rural areas. This is consistent with the situation in the United States and many other countries (Black and Henderson, 1999*a*, Henderson and Wang, 2007, Sánchez-Vidal, González-Val, and Viladecans-Marsal, 2014). As existing cities become more productive and their residents accumulate human capital, they also grow in population. In the United States for instance, cities have seen their population grow on average by 1.5% per year since 1950. In agreement with our model, much of the population growth of individual cities is attributed by past literature to their human capital and entrepreneurship (Glaeser and Saiz, 2004, Shapiro, 2006, Glaeser, Kerr, and Kerr, 2015). While cities experience parallel growth in expectation, each has its own ups and downs around a common trend (Black and Henderson, 2003, Ioannides and Overman, 2003, Duranton, 2007). This idiosyncratic component of city growth also results in the size distribution of cities following Zipf's law and thus resembling the size distribution of cities observed in the United States and other countries (see Duranton and Puga, 2014, for a discussion of the evidence). In addition, some cities hit by a sequence of negative shocks will exit despite net

entry (Sánchez-Vidal, González-Val, and Viladecans-Marsal, 2014, Michaels and Rauch, 2018).

Armed with our parameter estimates, we first quantify the importance of cities for the level of aggregate income and consumption by running a thought experiment where we cap city sizes. Constraining the two largest US cities, New York and Los Angeles, to be no larger than the third largest city, Chicago, would reduce average output per person by about 16%. Despite savings on urban costs, aggregate consumption would also be 3% lower. Capping the population of US cities at 5 million would raise the losses to about 25% in terms of output per person and 8% in terms of aggregate consumption.

Because incumbent residents prevent entry in the most productive cities through planning regulations, these cities are inefficiently small in equilibrium, as suggested by Hsieh and Moretti (2019). In turn, these regulations push part of the population into poorly productive cities and rural areas. We find that relaxing planning regulations in the three most productive cities, by reducing the misallocation of population, might generate large aggregate real gains of about 8%.

Next, we assess the effects of cities and urbanisation on economic growth. Specifically, we ask how much slower growth would be if agglomeration effects were weaker and if city population growth was smaller or absent. Agglomeration effects in cities and average city population growth magnify income growth. In addition, as more productive cities expand in response to human capital accumulation, productivity growth, and transport improvements, they draw workers away from less productive cities and rural areas, improving the spatial allocation of population. Overall, we find that preventing city population growth from 1950 onwards would lower the average growth rate in US income per person from 2.1% to 0.8%, with accumulated consumption losses by 2010 of 19%.

Our framework builds on the large literature on systems of cities initiated by Henderson (1974) and reviewed in Behrens and Robert-Nicoud (2015). As discussed in Duranton and Puga (2014), some of that literature has sought to embed models of systems of cities into a growth framework by focusing on factors that systematically influence city growth empirically. Of particular importance is the landmark model of Black and Henderson (1999*b*) which links urban and economic growth through human capital externalities in production. This is the closest parent to our work.

Relative to Black and Henderson (1999*b*), our main contribution is to assess quantitatively the effect of cities and urbanisation on economic growth. To do this, we need to write down a model that differs from theirs in three important ways.

First, our microeconomic foundations differ from theirs. These changes are guided by our desire to map empirical estimates directly into the model, while being able to accommodate the other quantitatively-relevant features that we introduce. For instance, Black and Henderson (1999*b*) work with fixed commuting costs. We find that it is important to consider congestion and allow for the commuting technology to evolve over time, be it only to capture the fact that the opportunity cost of the time spent travelling increases as wages increase. This matches micro-estimates and, more importantly, this feature is also crucial to match the empirical relationship between urban growth and aggregate income growth. As income grows, so does the value of travel time, dampening city growth in line with what we observe in the data.

Second, as already noted, instead of relying on competitive city developers, we determine equi-

librium city populations through a political economy mechanism with endogenously-determined planning regulations. This mechanism, where incumbent resident choose to prevent entry to maximise their welfare, can be seen as providing microfoundations for the local population allocation proposed by Albouy, Behrens, Robert-Nicoud, and Seegert (2019), although moving to a dynamic setting brings in additional complications. Endogenous planning regulations, in combination with differences in city productivity, also allow us to consider the possibility that the most productive cities might be too small.

Third, in addition to endogenous human capital accumulation, our model features two other drivers of growth. Related to our more detailed modelling of urban transport, we consider the evolution of transport technology over time. A third engine of city growth is the evolution of total factor productivity, featuring a common exogenous component that affects all cities as well as idiosyncratic productivity shocks affecting individual cities differently. This last element is related to the random growth models proposed by Gabaix (1999) and Eeckhout (2004), which focus on a growth process resulting from the accumulation of city-specific shocks. Like the literature that focuses on systematic drivers of urban growth, the model of Black and Henderson (1999*b*) does not naturally generate realistic city size distributions. Random growth models like Gabaix (1999) and Eeckhout (2004) generate realistic city size distributions, but leave aside the systematic determinants of growth that have been found to be empirically important. These models also counterfactually impose a fixed number of cities where production is subject to decreasing returns. An important contribution of our work is to combine these two distinct approaches to model urban growth. These approaches have been so far disconnected. An exception is Rossi-Hansberg and Wright (2007), who also consider city creation and the tradeoff between agglomeration benefits and urban costs in a model inspired by Black and Henderson (1999*b*). As with Black and Henderson (1999*b*), the main differences relative to Rossi-Hansberg and Wright (2007) are the empirical and quantitative components of our framework, which in turn require a different and rich modelling of urban costs and city formation, including endogenous planning regulations.

Our work is also related to a small number of recent quantitative assessments of the implications of cities on the level or the growth rate of aggregate income. These assessments are more partial than ours or explore other channels. Desmet and Rossi-Hansberg (2013) develop a static framework where city residents incur both real frictions (e.g. commuting) and fiscal frictions (e.g. taxes to maintain the local infrastructure) that distort their labour supply choice. Cities are larger because of their higher productivity, better amenities, or better ability to reduce frictions. For US cities, they find that reducing differences between cities in productivity, amenities, or frictions has large effects on their (counterfactual) population sizes but small welfare effects. In another static model where cities differ in their productivity and availability of land for production, Hsieh and Moretti (2019) focus on the misallocation of labour across cities that can occur because of planning regulations. Their findings suggest potentially large effects of planning regulations on aggregate income. We discuss the results of Hsieh and Moretti (2019) at greater length and how they relate to ours below. Davis, Fisher, and Whited (2014) use a neoclassical model of growth with physical capital and no human capital. In their model, urban growth requires rising physical investments in infrastructure and housing. This form of decreasing returns depresses growth. At the same

time, cities also become denser and this fosters agglomeration benefits. Overall, Davis, Fisher, and Whited (2014) find a modest contribution of about 10% of cities and agglomeration to aggregate growth percolating through these channels.

2. Technology, entrepreneurship and cities

There is a continuum of potential sites for cities, identified by subindex i. Time is discrete, with periods identified by subindex t. Final output is produced under constant returns to scale and perfect competition by combining intermediate inputs with a constant elasticity of substitution $\frac{1+\sigma}{\sigma}$, where $\sigma > 0$. Final output is freely tradable across cities and used as numéraire, whereas intermediates are non-tradable. Final output in city i at time t is then given by

$$Y_{it} = A_{it} \left\{ \int_0^{m_{it}} [q_{it}(\omega)]^{\frac{1}{1+\sigma}} \mathrm{d}\omega \right\}^{1+\sigma} , \tag{1}$$

where ω indexes intermediate inputs, $q_{it}(\omega)$ denotes the quantity of intermediate ω used in final production, and m_{it} denotes the endogenous mass of intermediates available in city i at time t. Potential city sites are heterogeneous, and A_{it} measures the level of production amenities in city i at time t.

Intermediate inputs are produced using human capital as an input:

$$q_{it}(\omega) = H_{it}(\omega) , \tag{2}$$

where $H_{it}(\omega)$ is the amount of human capital employed by the firm producing intermediate ω. Let H_{it} denote total human capital in the city. Since intermediate producers are symmetric, they each employ the same levels of human capital: $H_{it}(\omega) = \frac{H_{it}}{m_{it}}$. Using this and equation (2), equation (1) can be rewritten as:

$$Y_{it} = A_{it} \left\{ m_{it} \left[\frac{H_{it}}{m_{it}} \right]^{\frac{1}{1+\sigma}} \right\}^{1+\sigma} = A_{it} \left(m_{it} \right)^{\sigma} H_{it} . \tag{3}$$

Entrepreneurial ideas arise in proportion to the total local human capital, with proportionality constant $\rho > 0$. Each idea allows either to set up a new intermediate producer or to update the technology of an existing producer. Intermediate producers that do not update their technology in any given period become obsolete and exit. Thus, the total number of intermediate producers is:

$$m_{it} = \rho H_{it} . \tag{4}$$

Substituting equation (4) into equation (3) yields aggregate production:

$$Y_{it} = \rho^{\sigma} A_{it} (H_{it})^{1+\sigma} . \tag{5}$$

Note that, despite constant returns in final production and also in intermediate production, equation (5) exhibits increasing returns at the city level. Local aggregate increasing returns arise due to the relationship between human capital and entrepreneurship. A higher level of human capital in a city, everything else being equal, results in more entrepreneurial ideas and therefore in more input-producing firms. With a constant elasticity of substitution in final production, there are gains

from variety that imply greater aggregate output when there are many small local intermediate producers instead of a few large ones. Thus, the relationship between entrepreneurship and human capital produces an externality in aggregate human capital at the city level that raises its exponent from 1 in equation (2) to $1+\sigma$ in equation (5).

Assume each worker j chooses what share δ_t^j of her unit of available time to invest in education prior to working for the remaining time share $1-\delta_t^j$. Denote by h_t^j the amount of effective human capital worker j provides to her employer. This can be expressed as

$$h_t^j = (1-\delta_t^j) b(\delta_t^j) \bar{h}_t^j \,, \tag{6}$$

where the learning function $b(\delta_t^j)$ captures how education raises the worker's human capital building on the level of human capital she inherits from the previous generation, $\bar{h}_t^j$. It is natural to assume that $b'(\delta_t^j) > 0$ and $b(0) = 1$. Suppose that the level of human capital inherited by a given generation is the average level achieved after education by the previous generation:

$$\bar{h}_t^j = \frac{1}{\int \mathrm{d}j} \int b(\delta_{t-1}^j) \bar{h}_{t-1}^j \mathrm{d}j \,. \tag{7}$$

In appendix A, we show that, subject to some weak regularity conditions for the learning function $b(.)$, this set-up results in a constant rate of human capital accumulation over time so that $b(\delta_t^j) = b(\delta)$. Then, at any period t, workers in all cities will provide the same level of human capital $h_t = h_{it} = (1-\delta) b(\delta) \bar{h}_{it} = (1-\delta) b(\delta) h_{t-1} / (1-\delta) = b(\delta) h_{t-1}$. In section 5, we relax this feature of the model and allow human capital levels to vary across cities and to be systematically related to city size.

The revenue of local intermediate producers is used to pay for both their human capital input and entrepreneurial ideas. In appendix A, while deriving the individually optimal allocation of time, we disentangle relative rewards to human capital and entrepreneurial ideas. However, for the purpose of determining total income for each worker, this is not necessary since all workers in city i at time t are symmetric. Individual income then results from dividing between workers the revenue of local intermediate producers, which in turn, with perfect competition in the final good sector, is the aggregate value of final city output. Let N_{it} denote the population of city i at time t. Total local human capital can then be written as

$$H_{it} = h_{it} N_{it} \,. \tag{8}$$

Using equations (5) and (8), we can express income per worker as

$$y_{it} = \frac{Y_{it}}{N_{it}} = \rho^{\sigma} A_{it} (h_{it})^{1+\sigma} (N_{it})^{\sigma} \,. \tag{9}$$

Bigger cities concentrate more human capital and foster entrepreneurship, and this makes individual income increase with city population with elasticity σ. However, bigger cities feature not only stronger agglomeration economies but also higher urban costs. To characterise these costs, we need to look into the internal structure of cities.

Cities are linear and monocentric. Land in each city extends along the positive real line, but only a segment of endogenous length is built-up and inhabited at any given point in time. All city dwellings are built on equally-sized land plots and have identical floorspace z.[2]

The commuting costs of a worker who resides at a distance x from the city centre are given by

$$T_{it}(x) = \tau_{it} x^{\gamma} \, . \tag{10}$$

The length of each city resident's commute increases with elasticity $\gamma > 0$ with the distance x between her dwelling and the city centre. Note that this is slightly different than the standard monocentric model, where everyone is assumed to commute to the city centre. Here, we instead assume that an individual's commute increases non-linearly with distance to the centre —we can think of this as a reduced-form way to account for features that the monocentric model abstracts from, including the existence of secondary employment centres within cities.[3] Individual commuting costs are then the result of multiplying the distance travelled, x^{γ}, by the cost per unit of distance, τ_{it}. This, we specify in turn as

$$\tau_{it} = \tau_t (N_{it})^{\theta} \, . \tag{11}$$

The term $(N_{it})^{\theta}$, where $0 < \theta < 1$, captures congestion, which makes travel over a given distance slower in more populous cities. Parameter τ_t, which we allow to change over time, allows us to consider changes in commuting technology, altering for instance how much travellers value time in vehicle or the speed at which they travel.

To allow for changes in the degree of urbanisation over time, we assume that, as an alternative to living in one of the existing cities, workers can choose to reside in rural areas, in which case they attain a level of individual income

$$y_{rt} = A_{rt} (N_{rt})^{-\lambda} \, , \tag{12}$$

where N_{rt} denotes the rural population at time t, A_{rt} allows rural productivity to change over time, and $0 < \lambda < 1$. We can think of decreasing returns to rural labour as arising from the presence of some specific factor in fixed supply, such as arable land, in a rural production function with constant aggregate returns to scale.[4]

3. The number and sizes of cities

To characterise the number and sizes of cities, we need to specify how cities are created and managed. In section 6 we show that an important feature of the urban system of the United States is that local governments impose planning regulations that prevent bigger cities from further population

[2]The advantage of this simplification is that, with fixed housing consumption, utility maximisation is equivalent to maximising final good consumption.

[3]This generalisation also has an empirical motivation, since in section 7 we estimate the elasticity of an individual's travelled distance with respect to the distance between her residence and the city centre to be well below one. Nevertheless, we can still recover the classic specification where $\gamma = 1$ as a particular case.

[4]More specifically, equation (12) corresponds to a Cobb-Douglas rural production function for the numéraire good with a coefficient λ for arable land. Since our focus is not on structural transformation, we do not complicate derivations by introducing a separate rural good.

growth. In our model, incumbent homeowners only want their city to grow as long as the extra agglomeration benefits from new migrants dominate rising commuting and housing costs. New migrants, on the other hand, are willing to enter as long as income net of commuting and housing costs is higher in this city than elsewhere. To capture this tension between incumbent homeowners and potential new migrants into a city, we provide a simple model of the local political process.[5] The empirical evidence that we provide in section 6 about planning regulations, city populations, and house prices at the periphery of cities in the United States is consistent with the implications of this political process.

At the beginning of every period t, the idiosyncratic production amenity in each city location i is updated to a new level, given by A_{it}. A new generation is born and replaces the previous generation at their place of residence. Incumbent residents in each location vote in a local election, where a decision is made by simple majority on whether and by how much the housing stock in the city should expand. They do so by establishing more or less stringent planning rules that create a nuisance regulatory cost on potential newcomers (what Glaeser, Gyourko, and Saks, 2005, call a 'regulatory tax'). Any worker interested in becoming a new resident in the city can do so by incurring this cost of planning regulations p_{it}, in addition to bidding for a one-period lease on one plot of land in the city. The local government rents land at the going rate in the best alternative use, subleases it to the highest bidder at each location, and redistributes the difference among the local population.[6] Finally, workers commute between their residence and their job, engage in human capital accumulation, in the generation of entrepreneurial ideas and in production, obtain their income, and consume housing and the numéraire good.

Consider a new resident moving to city i from a rural area and choosing to locate at a distance x from the city centre. She incurs the cost p_{it} of planning regulations anticipating she will have to bid $R_{it}(x)$ per unit of land to successfully lease the plot on which her residence is built and incur a commuting cost $T_{it}(x)$ to access her job and obtain income y_{it}. For simplicity, we abstract from any other costs of building new homes, so that leasing a land plot and satisfying planning regulations is enough to build a new home in the city. The maximum bid $R_{it}(x)$ this new city resident is able to place while attaining the level of consumption available to rural residents $c_t = y_{rt}$ must therefore satisfy:

$$c_{it}(x) = y_{it} - T_{it}(x) - zR_{it}(x) - p_{it} = c_t = y_{rt} \ , \quad \forall x \ . \tag{13}$$

Equating expression (13) valued at $x = 0$ with the same expression valued at any other distance x from the city centre and simplifying, we can see that within each city the sum of commuting costs and land rents is independent of x and equal to the land rents at the city centre, where no commuting is necessary:

$$T_{it}(x) + zR_{it}(x) = zR_{it}(0) \ . \tag{14}$$

[5] An alternative would be to directly assume that local governments restrict city growth to maximise average consumption in their city (as, e.g., Albouy, Behrens, Robert-Nicoud, and Seegert, 2019).

[6] The three possibilities regarding land ownership commonly used in the literature are local public ownership, national public ownership, and absentee ownership (see Fujita, 1989, chapter 3). Assuming national public ownership or absentee ownership instead of local public ownership would reduce all equilibrium city sizes in the same proportion, equivalently to rescaling A_{it} everywhere. We prefer the assumption of local public ownership because it avoids introducing an additional distortion for which we see no strong empirical basis. A richer version of our assumption would have local governments supply public goods instead of the numéraire with the proceeds from local land taxation.

Differentiating equation (14) with respect to x shows that a marginal increase in land rents must be offset by a marginal decrease in commuting costs to preserve the indifference of residents choosing across locations within the city:

$$z\frac{\mathrm{d}R_{it}(x)}{\mathrm{d}x} = -\frac{\mathrm{d}T_{it}(x)}{\mathrm{d}x} \,. \tag{15}$$

Note this is the standard Alonso-Muth condition in the monocentric city model (Alonso, 1964, Muth, 1969).[7] The same marginal condition applies to incumbent city residents, although being exempt from the cost of planning regulations allows them to attain a higher level of final consumption $c_t + p_{it}$.

Without loss of generality, let us choose land and floorspace units so that the floor area ratio and z are both 1. The physical extent of each city is then the same as its population N_{it}. The edge of the city, $x = N_{it}$, is endogenously determined as the point beyond which urban residents are not willing to bid for a plot of land more than the rent this can fetch in the best alternative use, denoted by $\underline{R}$: $R_{it}(N_{it}) = \underline{R}$. Substituting equation (10) into (14) and valuing the resulting expression at the city edge $x = N_{it}$, we can express the equilibrium price of a dwelling at the city centre as

$$R_{it}(0) = \tau_{it}(N_{it})^{\gamma} + \underline{R} \,. \tag{16}$$

To simplify notation, we assume there is no alternative use for urban land, so that $\underline{R} = 0$. Combining this with equations (10), (11), (14), and (16), we can express the bid-rent price for a land plot at a distance x from the city centre as

$$R_{it}(x) = \tau_t(N_{it})^{\theta}\left[(N_{it})^{\gamma} - x^{\gamma}\right] \,. \tag{17}$$

Integrating $R_{it}(x)$, as given by (17), over the extent of the city yields total land rents as

$$R_{it} = \int_0^{N_{it}} R_{it}(x)\mathrm{d}x = \frac{\gamma}{\gamma+1}\tau_t\left(N_{it}\right)^{\gamma+\theta+1} \,. \tag{18}$$

Incumbent residents, through planning regulations voted in local elections, set the population size of their city to maximise their final consumption, $c_{it} = y_{it} - T_{it}(x) - R_{it}(x) + R_{it}/N_{it}$. Replacing y_{it}, $T_{it}(x)$, $R_{it}(x)$ and R_{it} with, respectively, equations (9), (10), (17) and (18), we can write the corresponding programme as[8]

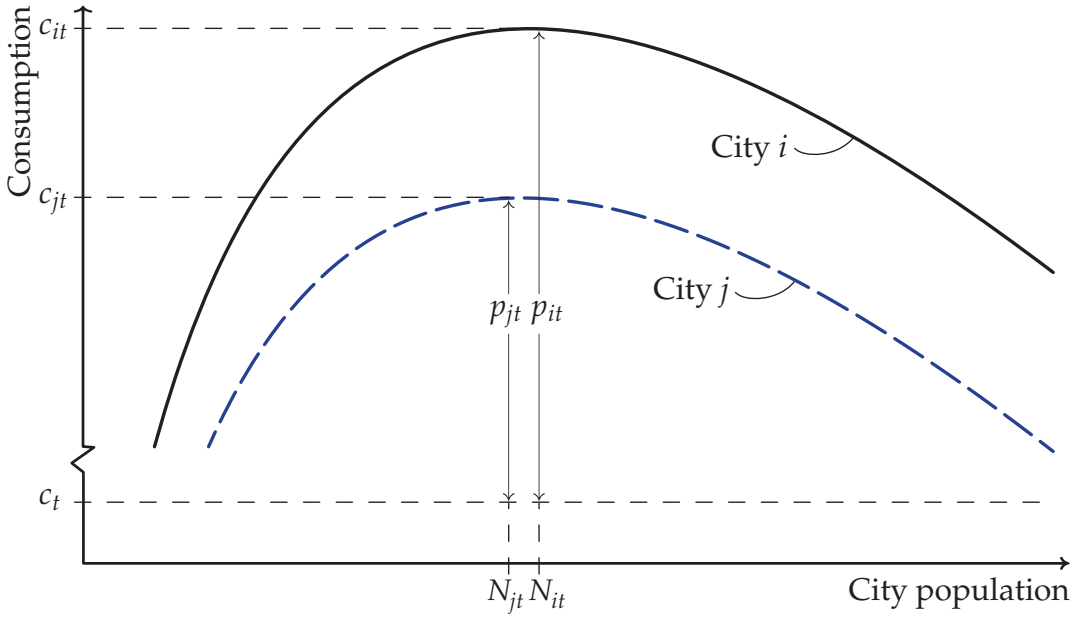
$$\max_{\{N_{it}\}} c_{it} = \rho^{\sigma} A_{it}(h_{it})^{1+\sigma}(N_{it})^{\sigma} - \frac{\tau_t}{\gamma+1}(N_{it})^{\gamma+\theta} \,. \tag{19}$$

Since incumbent residents who already have a house in the city do not need to incur the cost p_{it} of planning regulations, they are able to obtain a higher level of final consumption $c_{it} = c_t + p_{it}$

[7] By the envelope theorem, the exact same condition holds if we allow residents to choose heterogeneous amounts of housing consumption in different locations within the city (see Duranton and Puga, 2015).

[8] The same programme applies if we simply assume each city has a local government that decides independently of others how many residents to take with the aim of maximising their individual utility, as in Albouy, Behrens, Robert-Nicoud, and Seegert (2019). The modelling proposed here can be seen as developing microfoundations for that reduced-form assumption. In addition, it restores a spatial equilibrium in which the marginal migrant is kept indifferent between moving or not through the additional cost of new housing introduced by endogenous planning regulations. Note that, with a continuum of potential city sites, changes in N_{it} for any one city i have no effect on r_t.

Figure 1: Final consumption as a function of city size



compared with c_t for newcomers. It follows that the value of owning a house in city i from the outset of period t is given by

$$p_{it} = c_{it} - c_t \, . \tag{20}$$

Thus, to maximise c_{it} in the programme of equation (19), incumbent city residents vote for planning regulations that maximise the value of their individual homes, as in Fischel's (2001) 'homevoter hypothesis'.[9]

When solving the programme of equation (19), incumbent city residents are willing to let the city expand only if the marginal benefit of doing so in terms of agglomeration economies that raise earnings (captured by the term in $(N_{it})^{\sigma}$) outweighs the marginal cost in terms of increased crowding (captured by the term in $(N_{it})^{\gamma+\theta}$). The first-order condition yields equilibrium city sizes as

$$N_{it} = \left(\frac{\rho^{\sigma} \sigma(\gamma+1)}{\gamma+\theta} \frac{A_{it}(h_t)^{1+\sigma}}{\tau_t} \right)^{\frac{1}{\gamma+\theta-\sigma}} \, . \tag{21}$$

The second-order condition requires $\gamma + \theta - \sigma > 0$, which we show below holds empirically. For positive city sizes, we require $\sigma > 0$, which also holds empirically ($\sigma = 0$ implies $N_{it} = 0$).

Figure 1 illustrates the relationship between final consumption for incumbents and city size given by equation (19) for two cities with different levels of production amenities. The concavity of final consumption in the figure reflects the tradeoff created by an increase in a city's population between agglomeration economies and crowding. For each city, the population size defined by

[9]The cost of planning regulations reflects only the consumption differential between a city and the best alternative for the current generation but does not capitalise the gains for future generations, as we ignore bequest motives. Incorporating these would affect the value of p_{it}, but not the city population size given by equation (21). This size maximises consumption period by period —a necessary condition to maximise consumption across generations in our context if we introduce bequest motives.

equation (21) corresponds to the maximum of the corresponding curve. Incumbent residents achieve their maximum consumption for a larger population size in city i than in city j, $N_{it} > N_{jt}$, because we have assumed a higher level of idiosyncratic productivity in city i than in city j, $A_{it} > A_{jt}$. This size is optimal from the perspective of incumbent local residents. However, residents in smaller and less productive cities would like to join them, thereby making the city's population increase further, were it not for the excessive cost of planning regulations. While final consumption for incumbent residents is higher in city i than in city j, $c_{it} > c_{jt}$, final consumption for the marginal resident is equated across cities at c_t by the different level of planning regulations in each, with $p_{it} > p_{jt}$ and $c_{it} - p_{it} = c_{jt} - p_{jt} = c_t$.

In absence of the cost imposed by planning regulations ($p_{it} = p_{jt} = 0$), reallocating a small number of residents from the smaller and less productive city j to city i would lead to a first-order gain for the reallocated residents. This gain would only be partially offset by a second-order loss for the remaining residents of both cities —the marginal reallocation would move city i's population above N_{it} and city j's population below N_{jt}. Since the rents associated with this additional expansion cannot be captured by incumbent residents, the city size they implement using planning regulations fails to equalise the social returns to the marginal resident across cities. As a result, the equilibrium city size of equation (21) is too small relative to what would be socially desirable. Aggregate consumption would increase by vacating the least productive sites and allocate more residents to the remaining cities.[10] In section 8, we quantitatively explore the consequences of relaxing locally-imposed planning regulations.

Cities attract residents as long as they offer newcomers a level of consumption of the numéraire good that leaves them no worse than in rural areas. The marginal populated city (i.e. the city location with the lowest level of production amenities to be populated) satisfies two conditions. First, the marginal populated city has a value of production amenities $A_{it} = \underline{A}_t$ such that incumbent residents can only just attain a local population that maximises their individual final consumption while matching consumption for newcomers to consumption in rural areas by imposing no planning restrictions with $p_{it} = 0$. This condition can be obtained by equating the consumption level for incumbent city residents, c_{it}, as given by equation (19), and the rural consumption level, as given by equation (12). This yields $\rho^\sigma \underline{A}_t (h_{it})^{1+\sigma} (N_{it})^\sigma - (\tau_t/(\gamma+1))(N_{it})^{\gamma+\theta} = A_{rt}(N_{rt})^{-\lambda}$. Using equation (21) to set N_{it} at the level that maximises local individual consumption then yields a first equation in $\underline{A}_t$ and N_{rt}: $[(\gamma+1)/\tau]^\sigma \left\{ [\sigma/(\gamma+\theta)]^\sigma - [\sigma/(\gamma+\theta)]^{\gamma+\theta} \right\} \left[\rho^\sigma \underline{A}_t (h_{it})^{1+\sigma} \right]^{\gamma+\theta} = \left[A_{rt}(N_{rt})^{-\lambda} \right]^{\frac{1}{\gamma+\theta-\sigma}}$. Second, the level of production amenities of the marginal populated city, $\underline{A}_t$, must be such that the combined population of all cities with $A_{it} \geqslant \underline{A}_t$ and the rural population add up to the total population at time t, N_t. If we use equation (21) to express equilibrium city population as a function of the level of production amenities, $N_{it} = N(A_{it})$, this second equation in $\underline{A}_t$ and N_{rt}

[10] We can think of three alternative micro-foundations for sub-optimally small cities. First, if the nuisance arising from additional housing construction and increases in crowding are experienced with much greater intensity locally while the gains from greater agglomeration economies are diffused through the metropolitan area, to the extent that planning barriers are also more local in nature, they may be set placing undue weight on the costs of urban expansion relative to the benefits. Second, as highlighted by Fischel (2001), city population growth may entail some risks for a majority of risk-averse incumbent residents. Third, with strong idiosyncratic location preferences, incumbents may use planning regulations to extract rents from potential newcomers with a high willingness to pay for their city.

can be written as $N_{rt} + \int_{\underline{A}_t}^{+\infty} N(A) g_t(A) \mathrm{d}A = N_t$, where $g_t(A)$ is the probability density function of city production amenities at time t.

Our framework shares many features with the standard monocentric city model going back to Alonso (1964) and Muth (1969) and with models of urban systems building on Henderson (1974). Within each city there is gradient of house prices decreasing in distance to the centre to offset higher commuting costs (equation 17). Equilibrium city sizes result from a tradeoff between agglomeration economies and crowding costs (equation 19), and are also increasing in local productivity, human capital, and travel speed (equation 21). Bigger cities feature higher house prices at the centre (equation 16) and higher earnings (equation 9).

However, there is also one fundamental difference. In standard monocentric city and urban system models, house prices at the city edge are equated across cities. The marginal migrant sustains a longer commute in bigger cities but this is exactly offset by higher earnings. When a city experiences a positive shock that attracts new residents, new construction takes place freely until the equality of house prices at the city edge is restored. In our framework, however, incumbent residents use local planning regulations to curb new construction in reaction to a local positive shock. They allow the city to expand, but only up the the point where the additional crowding costs imposed on them by the marginal migrant exactly offset additional agglomeration benefits they bring. The higher earnings of the marginal migrant in bigger cities must offset not just a longer commute but also the cost of stricter planning regulations.

This key difference leads to two testable implications from our framework that do not hold in the standard framework. Planning regulations should be more stringent in bigger cities. House prices at the edge should also be higher in bigger cities. These two predictions can be seen by combining equations (13) and (19)-(21) to express the costs of local planning regulations as well as the difference between house prices at the city edge and the common price of land in the best alternative use as

$$p_{it} = \frac{\gamma + \theta - \sigma}{\sigma(\gamma + 1)} \tau_t (N_{it})^{\gamma+\theta} - y_{rt} \ , \tag{22}$$

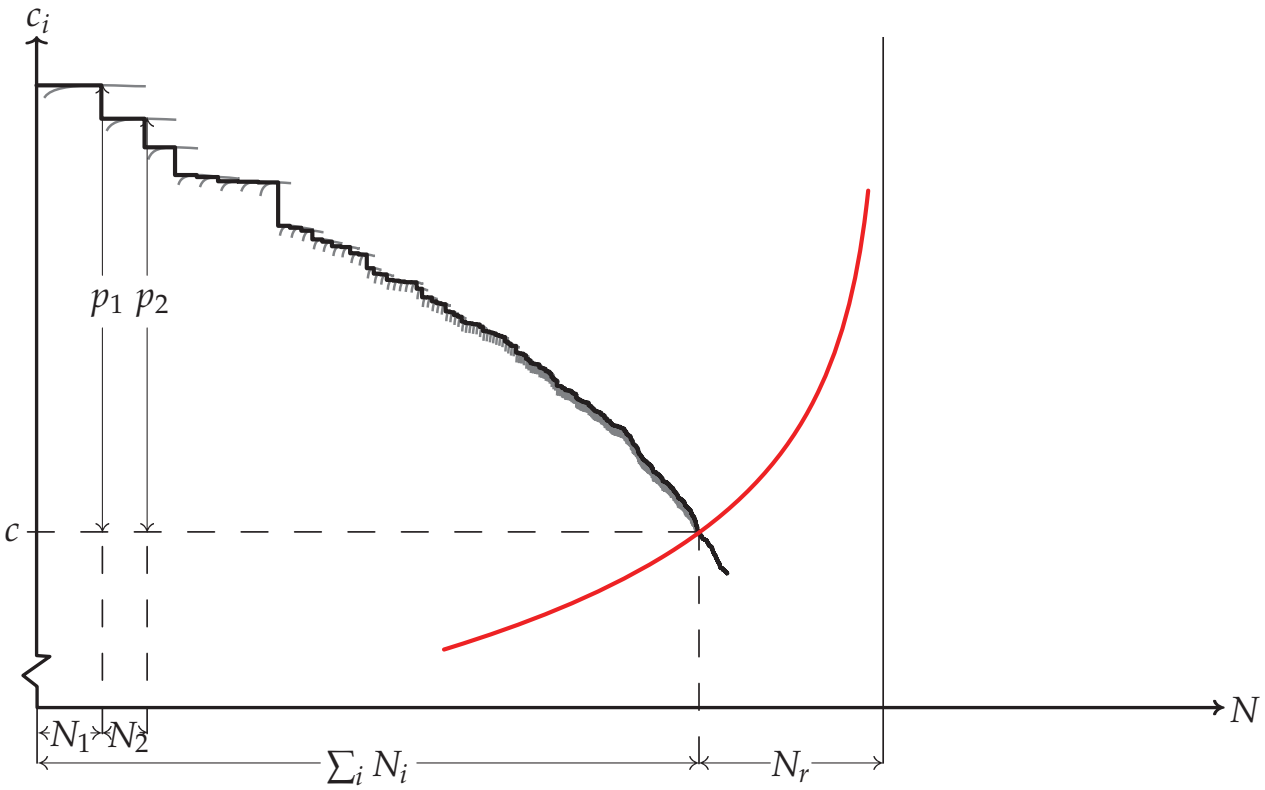
which is an increasing function of the city's population N_{it}. In section 6 we show that these predictions are first-order features of the data for the cities of the United States.

Illustrating the equilibrium with the urban system of the United States

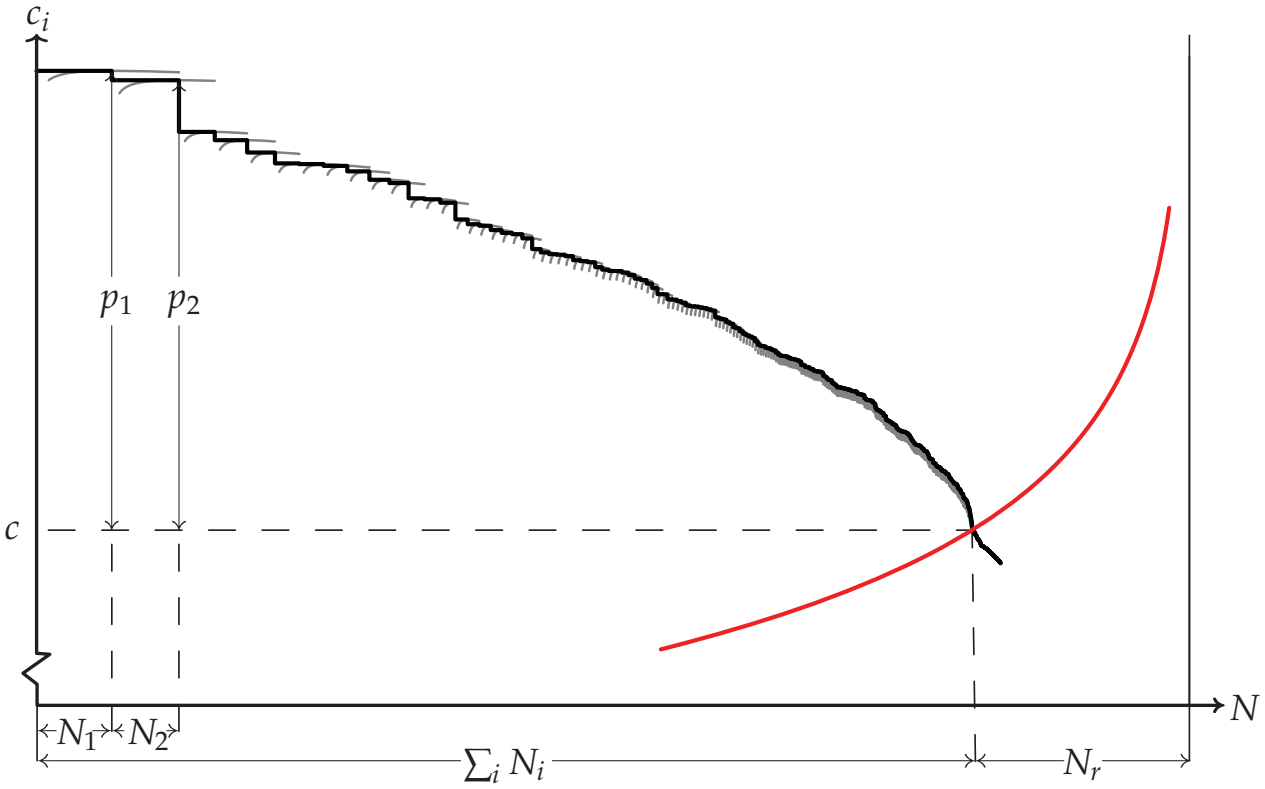
To help build intuition for the equilibrium of our model, in panel A of figure 2 we represent the urban system of the conterminous United States in 1980 as seen through the lens of the model. Throughout the paper we define US cities using 1999 county-based metropolitan area definitions, which gives 275 cities. We leave discussion of the parameter values used for section 7, were we present the details of their estimation. The sequence of thick segments represents consumption for incumbent residents in each metropolitan area (measured on the vertical axis) as a function of its population (measured along the horizontal axis). The thick segment on the top left corresponds to New York. This would be the first location to be populated and incumbent residents use planning regulations to keep its population at the level beyond which additional residents inflict greater crowding costs than the agglomeration benefits they bring. This population level corresponds to

Figure 2: Equilibrium allocation of population

Panel A: United States, 1980



Panel B: United States, 2010



the maximum of the thin curve mapping final consumption which is depicted just below each thick segment and tangent to it to the left of its maximum. With city population determined by the maximum of the final consumption curve, we can read New York's population as N_1 on the horizontal axis. The thick segment immediately to the right of New York's corresponds to Los Angeles and this is drawn shifted to the right by the population level of New York. Thus, the horizontal distance between New York's population, N_1, and the point at which the second final consumption curve reaches its maximum gives the population of Los Angeles, N_2. We can then continue this process for every metropolitan area, following the sequence in decreasing population sizes.

To draw these curves, given the actual population size of each metropolitan area in the United States, we can use equation (21) to determine the level of idiosyncratic productivity A_{it} that would sustain that size in equilibrium. We can then substitute this value of A_{it} into equation (19) to calculate consumption for incumbent residents in each metropolitan area, c_{it}, as a function of its population for values between 0 and the equilibrium level at which c_{it} is maximised.

Population outside of metropolitan areas is also measured on the horizontal axis, but from right to left, setting the length of the horizontal axis to match the total population of the United States, similarly to the diagram used to represent the specific-factors model in international trade. The curve extending along the full length of the diagram represents rural consumption as a function of the rural population, as given by equation (12). Parameter A_{rt} in this equation is calculated so that the rural consumption level implied by equation (12) for the actual level of non-metropolitan population in the United States equals consumption for incumbent residents in the smallest city implied by equation (19). The point where the curves for the urban sector and the curve for the rural sector intersect gives on the horizontal axis the total urban population as the distance between the left origin and the intersection point, and rural population as the distance between the right origin and the intersection point. This intersection point gives on the vertical axis the level of consumption, net of housing and commuting costs, for new residents in every city. The level of planning regulations is then given by the vertical distance between this common consumption level for new residents and the higher level of consumption for incumbent residents who own a house in the city from the outset (e.g. p_1 for New York and p_2 for Los Angeles).

We have represented additional potential city sites to the right of the marginal city at the intersection point between the curves for the urban sector and the curve for the rural sector. In our model, these additional cities correspond to unpopulated urban sites drawn from the lower tail of the distribution of production amenities, below the amenities of the marginal city, i.e. with $A_{it} \leqslant \underline{A}_t$. We explain in section 8 how we construct the values of A_{it} for these unpopulated city locations.

Panel B of figure 2 represents the urban system of the conterminous United States in 2010. The increased distance between the two vertical axes represents the growth in the total population of the conterminous United States between 1980 and 2010 from 225 million to 307 million people. Population outside of metropolitan areas grows somewhat in absolute terms but falls as a share of total population, as urbanisation keeps advancing in the nation. Each curve in the sequence representing the urban sector moves up vertically and expands horizontally. This captures that

every single metropolitan area sees its population grow over this period, partly through a systematic urban growth component arising from human capital accumulation and partly through the accumulation of idiosyncratic shocks to each city's level of production amenities. Incumbent residents adapt planning regulations to let cities expand up to the new, larger, locally optimal level. However, the shocks are heterogeneous across cities, so not all of them grow at the same rate. Comparing changes for the first two curves between 1980 and 2010, we can see that Los Angeles grows more than New York. Looking at the eight and ninth curves in 1980, we can see that Detroit was much larger than Houston at that point. Small growth in Detroit (population in the central city fell but still rose in the metropolitan area as a whole) and large growth in Houston brought these two cities to almost the same level in 2010. Dallas grew much more than both Detroit and Houston, so its curve appears as the twelfth in 1980 and as the eighth in 2010, displacing Detroit and Houston to ninth and tenth.

4. Urban growth and the size distribution of cities

We now examine city population growth in our model and its implications for the size distribution of cities. To compute the log population change between two consecutive periods in city i, if a city exists at that location, we take the natural logarithm of equation (21) and subtract the resulting expression valued at time $t-1$ from the same expression valued at time t. Let us use the Δ operator to denote the difference in a variable with respect to the previous period, e.g. $\Delta \ln(N_{it}) \equiv \ln(N_{it}) - \ln(N_{it-1})$. We can then write

$$\Delta \ln(N_{it}) = \frac{1}{\gamma + \theta - \sigma} \left[\Delta \ln(A_{it}) + (1+\sigma) \Delta \ln(h_{it}) - \Delta \ln(\tau_t) \right] . \tag{23}$$

A first component of city population growth arises from the evolution of idiosyncratic productivity at each location. We assume this productivity evolves through the accumulation of random multiplicative shocks: $A_{it} = g_{it} A_{it-1}$, where the shocks g_{it} are identically and independently distributed across locations. For simplicity, we assume non-negative shocks.[11] Taking natural logarithms and time differencing yields, $\Delta \ln(A_{it}) = \ln(g_{it})$.

The accumulation of human capital over time also makes cities grow. When workers have a greater level of human capital, they impose the same crowding on other workers in the city but are able to produce more. In addition, there is an externality whereby greater human capital promotes entrepreneurship and new firm birth which expands output per worker further —hence the $1+\sigma$ multiplying $\Delta \ln(h_{it})$ in equation (23). As already discussed above and formally shown in appendix A, subject to some weak regularity conditions for the learning function, there is a constant rate of human capital accumulation over time: $\Delta \ln(h_{it}) = \Delta \ln(h)$.

Finally, a third potential component of city growth arises from the evolution of τ_t.[12] Let us assume that this evolves at some constant rate, reflecting, for instance, changes in commuting

[11] Negative shocks can be incorporated if we allow for sufficient depreciation of the housing stock so that some additional construction is always needed and planning regulations remain relevant.

[12] We make cities heterogeneous only in their production amenities to keep the exposition simple. We could nonetheless readily extend our model to idiosyncratic shocks in, say, transport infrastructure by enriching equation (11) and explicitly consider shocks to roadway expansion.

speed: $\Delta \ln(\tau_t) = \Delta \ln(\tau)$.

We can now rewrite equation (23) describing the population growth of a city at location i between time $t-1$ and time t, provided a city exists at this location in both periods, as

$$\Delta \ln(N_{it}) = \breve{g}_{it} \ , \tag{24}$$

where

$$\breve{g}_{it} = \frac{1}{\gamma + \theta - \sigma} \ln(g_{it}) + \frac{(1+\sigma)}{\gamma + \theta - \sigma} \Delta \ln(h) - \frac{1}{\gamma + \theta - \sigma} \Delta \ln(\tau) \ . \tag{25}$$

The urban population growth rate $\breve{g}_{it}$ has a systematic component arising from human capital accumulation and from the evolution of commuting speed that is common to all cities, captured by the last two terms in the right-hand side of equation (25). It also has an idiosyncratic random component arising from local productivity shocks, captured by the first term in the right-hand side of equation (25). Thus, cities experience parallel growth in expectation but are subject to idiosyncratic ups and downs relative to this common trend.

The growth process of (24) satisfies Gibrat's law (after Gibrat, 1931): since g_{it} is identically and independently distributed for every city, by equation (25) so is $\breve{g}_{it}$. As shown by Gabaix (1999), Gibrat's law results in a steady-state city-size distribution that approximates Zipf's law, i.e. steady-state city sizes follow a Pareto distribution with shape parameter approaching 1. This requires two additional conditions. First, there must be some mechanism preventing cities from shrinking indefinitely. In Gabaix (1999) this mechanism is a reflexive lower bound on city sizes such that when this minimum size is reached further shocks can only bring size up and not further down.[13]

The second condition we need to obtain approximately Zipf's law is to be able to normalise city sizes so that their mean normalised size and the reflexive lower bound are both time-invariant (this is what Saichev, Malevergne, and Sornette, 2009, call the 'balance condition'). Following Gabaix (1999), a simple normalisation is to express city sizes relative to their average size. We thus define normalised city sizes as $\tilde{N}_{it} \equiv \frac{N_{it}}{\bar{N}_t}$, where $\bar{N}_t$ denotes the average population at time t of all potential cities. Mechanically, the mean normalised size of all potential cities is then constant and equal to 1.

Champernowne (1953) was the first to study such a random growth process with a lower bound and showed that it generates a Pareto distribution (see also Gabaix, 2009). Let $F(\tilde{N})$ denote the share of potential cities with a normalised population size $\tilde{N}$ or lower in steady-state, i.e. the cumulative distribution function. Champernowne's (1953) insight is that

$$F(\tilde{N}) = \begin{cases} 1 - \left(\frac{\tilde{N}}{\eta}\right)^{-\zeta} & \text{if } \ \tilde{N} \geqslant \eta \ , \\ 0 & \text{if } \ \tilde{N} < \eta \ , \end{cases} \tag{26}$$

where η denotes the reflexive lower bound on normalised sizes. The corresponding probability density function is then $f(\tilde{N}) = \frac{\mathrm{d}F(\tilde{N})}{\mathrm{d}\tilde{N}} = \eta^{\zeta} \zeta \tilde{N}^{-\zeta-1}$ and the mean normalised size of all potential

[13] In practice, such a reflexive force can arise from the durability of housing (Glaeser and Gyourko, 2005). Following Saichev, Malevergne, and Sornette (2009), an alternative mechanism preventing cities from shrinking indefinitely would be to have cities exit when their productivity falls below some threshold and to have new potential sites for cities arise stochastically with some initial productivity level that is higher than the threshold but bounded from above. This could happen, for instance, if over time agricultural land is being converted to potential urban use.

cities can be calculated as

$$\int_{\eta}^{+\infty} \tilde{N} f(\tilde{N}) \mathrm{d}\tilde{N} = \frac{\eta^{\zeta}\zeta}{1-\zeta} \left[\tilde{N}^{1-\zeta}\right]_{\eta}^{+\infty} = -\frac{\eta\zeta}{1-\zeta} , \tag{27}$$

provided $\zeta > 1$ (otherwise the mean normalised size is infinite). As noted above, this mean normalised size equals 1, so solving $-\frac{\eta\zeta}{1-\zeta} = 1$ for ζ yields

$$\zeta = \frac{1}{1-\eta} . \tag{28}$$

Hence, the steady-state distribution of normalised sizes for all potential cities follows a Pareto distribution with shape parameter $\frac{1}{1-\eta}$ and scale parameter η.

We have characterised the steady-state distribution of normalised population sizes of potential cities. Some of these potential cities are actually populated while others, with lower values of A_{it}, are left vacant. For the latter, the normalised population values we consider are the ones we would find at those locations if they were populated —for instance, if total population was larger. Of course, the distribution we are interested in is the distribution of absolute sizes of actual cities. To characterise this, starting from the steady-state distribution of normalised population sizes of potential cities, we first remove the size normalisation and then focus on those potential cities that are actually populated.

Since normalised sizes, $\tilde{N}$, express city sizes relative to their average size, absolute city sizes can be recovered as $N = \bar{N}_t \tilde{N}$. Multiplying a variable that is distributed Pareto by a constant results in a transformed variable that follows a Pareto distribution with the same shape parameter as the original distribution and simply has its scale parameter multiplied by that constant. Thus the distribution of absolute population sizes of potential cities still converges over time to a Pareto with shape parameter $\frac{1}{1-\eta}$.

Finally, we can move from the size distribution of potential cities to the size distribution of actual cities. Potential sites for cities will be occupied in order determined by their level of production amenities A_{it}, starting with the highest. As detailed in section 3, the marginal populated city at any point in time is that which has a value of $A_{it} = \underline{A}_t$ such that it can only just provide the same level of final consumption as rural areas when populated at the level that maximises local individual consumption while satisfying the overall population constraint. Thus, at any point in time, the size distribution of actual cities is the results of taking the size distribution of potential cities and left-truncating it to remove cities with values of $A_{it} \leqslant \underline{A}_t$. Left-truncating a Pareto distribution changes its scale parameter but not its shape parameter. Therefore, the distribution of city sizes in the model converges over time to a Pareto distribution with shape parameter $\frac{1}{1-\eta}$. As the reflexive bound on city sizes η approaches 0, this distribution approximates Zipf's law: it becomes a Pareto distribution with a shape parameter approaching 1 (although it can never become exactly 1 for the expected population size of cities to remain finite).

We have just shown that with a small lower bound on city sizes that evolves in parallel with mean city size, we obtain approximately Zipf's law.[14] If we eliminate the lower bound altogether

[14] If we have a lower bound but this evolves at a different rate than mean city size, the steady state city size distribution will still be Pareto but its shape parameter will differ from 1 (this follows from proposition 3.4.1 in Saichev, Malevergne, and Sornette, 2009).

and allow cities to keep shrinking indefinitely, the steady-state distribution of city sizes converges over time to a left-truncated log-normal, the variance of which keeps increasing indefinitely. To see this, simply note that equation (24) implies that after T periods the size of a potential city at location i is

$$\ln(N_{iT}) = \ln(N_{i0}) + \sum_{t=1}^{t=T} \breve{g}_{it} \, . \tag{29}$$

Since the multiplicative shocks $\breve{g}_{it}$ are identically and independently distributed for every city, by the central limit theorem, over time $\ln(N_{it})$ approaches a normal distribution and the distribution of N_{it} thus becomes log-normal. As before, potential sites for cities will be occupied in order determined by their level of production amenities A_{it}, starting with the highest and ending once all population is allocated to cities. The size distribution of actual cities is thus the result of truncating the size distribution of all potential cities for values of $A_{it} \leqslant \underline{A}_t$, so without a reflexive lower bound over time the distribution of city sizes approaches a left-truncated log-normal distribution.

Previous models in which independent and identically distributed random shocks affect an exogenously given number of cities obtain approximately Zipf's law if they assume a lower bound on city sizes (e.g. Gabaix, 1999) and a log-normal distribution if they do not (e.g. Eeckhout, 2004). In our framework, with an endogenous time-varying number of cities, the distinction becomes more subtle: it is no longer Pareto versus log-normal, but (truncated) Pareto versus truncated log-normal.

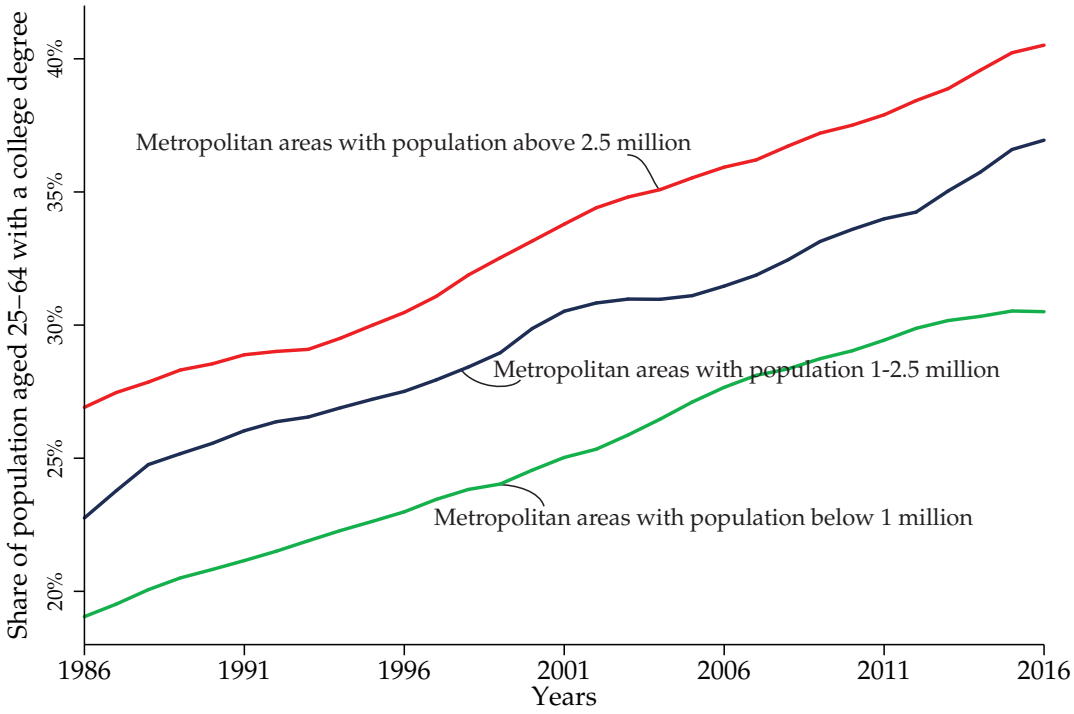
While this distinction is important theoretically, separating between a Pareto distribution and a truncated log-normal distribution for city sizes is extremely difficult empirically. A Pareto distribution has a thicker upper tail relative to a log-normal distribution. With few observations in the upper tail even in the largest countries, attempts to distinguish between the two typically lack statistical power. Pareto and log-normal distributions also differ with respect to their lower tail but truncating the log-normal will mask this difference. Rather than testing for differences in outcome, one might be tempted to test the key assumption driving the difference between these two outcomes, that is the existence of a reflexive lower bound. This is also problematic since this assumption affects sites at the lower tail of the productivity production, which are expected to remain unpopulated in equilibrium.

Leaving aside the distinction between Pareto and truncated log-normal distributions, what is important to keep in mind is that our model is able to replicate the approximate shape of observed city size distributions regardless of whether we impose a reflexive lower bound. After all, the reason why it is difficult to determine which of the two distributions, a Pareto or a truncated log-normal, best approximates the actual size distribution of cities is that they are both very similar and good approximations of reality.

5. Urban growth when systematic determinants are correlated with city sizes

In our framework, the urban population growth rate has a systematic component arising from human capital accumulation and the evolution of urban transport. It also has an idiosyncratic component arising from local productivity shocks. We have thus brought together systematic and random urban growth models while obtaining realistic stable city-size distributions. However, we

Figure 3: Evolution of college-educated population shares in the United States



have so far achieved this under restrictive assumptions that ensure that in the model the systematic determinants of urban growth are equally strong in all cities. In practise, there are factors that have been found to be important empirical drivers of individual city growth that are distributed across cities in a way that is systematically related to their population sizes.

The most obvious example is human capital. This is a fundamental determinant of city growth and there is a concentration of human capital in bigger cities (Glaeser and Saiz, 2004, Moretti, 2004*c*, Shapiro, 2006). Figure 3 plots the evolution of the share of population aged 25–64 who hold a college degree in metropolitan areas of different sizes over the period 1986–2016 in the United States, using data from the Current Population Survey (CPS). Over these three decades, there has always been a larger share of college educated individuals in bigger cities. In 1986, the college share was 19.2% in metropolitan areas with less than one million inhabitants, 24.0% in metropolitan areas with between 1 and 2.5 million inhabitants, and 26.4% in metropolitan areas with over 2.5 million inhabitants. The share of individuals holding a college degree has also increased by a factor of 1.6 between 1986 and 2016 in all three city-size categories, keeping the relative magnitude of the college share across city-size categories stable.[15] If instead of splitting cities into size classes, we estimate an elasticity of the share of college educated individuals with respect to city population based on the same CPS data, we obtain a stable elasticity over the period 1986–2016 of around

[15]The ratio of the 2016 to the 1986 college share is 1.56 in metropolitan areas with less than one million inhabitants, 1.55 in metropolitan areas with between 1 and 2.5 million inhabitants, and 1.56 in metropolitan areas with over 2.5 million inhabitants.

0.10.[16] We now show that our model can easily be extended to allow for such patterns.

Suppose that after choosing how much time to devote to education, each worker acquires some initial work experience. This experience raises their human capital from their post-education level $b(\delta_t^j)\bar{h}_t^j$ to $b(\delta_t^j)\bar{h}_t^j(N_t^j)^\beta$. Note that the proportionate increase is larger the bigger the city where they acquire this initial experience. This is consistent with the findings of De la Roca and Puga (2017), who show that the value of early job experience increases with the city where this is acquired. They also show that differences across cities in the value of experience are large relative to differences in intrinsic skills (as captured by a worker fixed effect in an earnings regression). The amount of effective human capital provided after this initial work experience then becomes

$$h_t^j = (1-\delta_t^j)b(\delta_t^j)\bar{h}_t^j(N_t^j)^\beta \,. \tag{30}$$

In appendix A, we show that using equation (30) instead of (6) still results in a constant rate of human capital accumulation over time across cities so that $b(\delta_t^j) = b(\delta)$. Recall this common rate across cities of different sizes is what figure 3 reflects for the United States over the period 1986–2016. Then, the equilibrium level of human capital resulting from education and early job experience is an increasing iso-elastic function of city size:

$$h_{it} = h_t N_{it}^\beta \,, \tag{31}$$

where, as before, $h_t = b(\delta)h_{t-1}$. Keeping the assumption that entrepreneurial ideas arise in proportion to the total local post-education level of human capital implies $m_{it} = \rho h_{it} N_{it}$. Combining this expression with equations (3) and (8), yields the following alternative expression for individual earnings:

$$y_{it} = \frac{Y_{it}}{N_{it}} = \rho^\sigma A_{it}(h_t)^{1+\sigma}(N_{it})^{\sigma+\beta} \,. \tag{32}$$

With earnings given by equation (32) instead of (9), equilibrium city sizes are determined by solving the following programme instead of (19):

$$\max_{\{N_{it}\}} c_{it} = \rho^\sigma A_{it}(h_{it})^{1+\sigma}(N_{it})^{\sigma+\beta} - \frac{\tau_t}{\gamma+1}(N_{it})^{\gamma+\theta} \,. \tag{33}$$

Compared with the expression in our baseline model, given by equation (19), the only difference is that the exponent on city population in the first term is $(\sigma+\beta)$ instead of σ. What this shows is that if systematic drivers of growth, such as human capital in this example, increase in magnitude with city size with some constant elasticity, they simply become an additional magnification effect, akin to having stronger agglomeration economies.

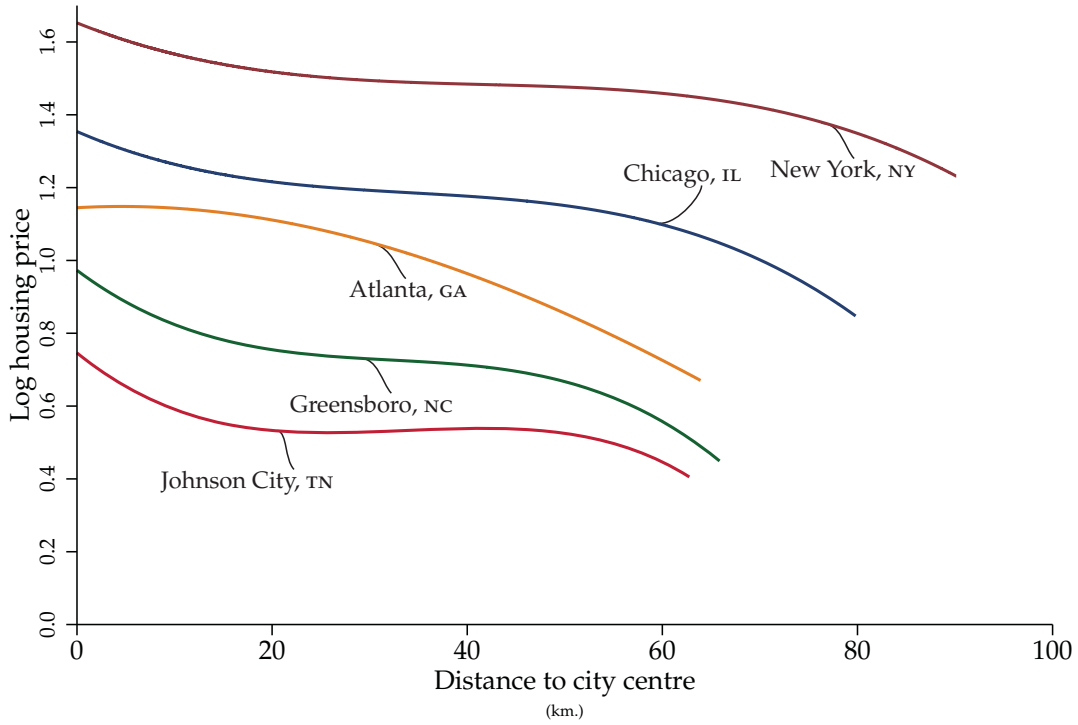
The first-order condition for (33) yields equilibrium city sizes as

$$N_{it} = \left(\frac{\rho^\sigma(\sigma+\beta)(\gamma+1)}{\gamma+\theta}\frac{A_{it}(h_t)^{1+\sigma}}{\tau_t}\right)^{\frac{1}{\gamma+\theta-\sigma-\beta}} \,. \tag{34}$$

The second-order condition requires $\gamma+\theta-\sigma-\beta > 0$, which we show below holds empirically. For positive city sizes, we require $\sigma+\beta > 0$, which also holds empirically ($\sigma+\beta = 0$ implies $N_{it} = 0$). Thus, this extended version of our model still exhibits the same qualitative behaviour and is also consistent with realistic city-size distributions.

[16] The estimated elasticity is 0.114 in 1986, 0.108 in 1996, 0.100 in 2006, and 0.100 in 2016.

Figure 4: House price gradients in selected cities



6. Housing regulation and new construction in US cities

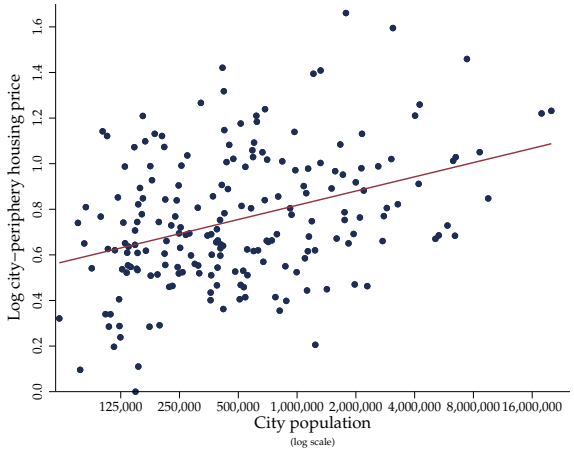
The political-economy mechanism that determines the number and sizes of cities in section 3 highlights the importance of planning regulations. More specifically, this mechanism implies that planning regulations will be stricter in larger cities. As a result, housing prices at the edge of cities will increase with their population, as demonstrated by equation (22). We now examine this implication empirically for the United States.

Figure 4 shows housing price gradients for five US cities. These are, from highest to lowest population, New York, Chicago, Atlanta, Greensboro and Johnson City. For housing units of comparable characteristics, each curve gives their price at distances ranging from zero to their 95th percentile in each metropolitan area (we give further details of how we construct this figure in section 7 below). The figure illustrates the empirical relevance of several features that our model shares with standard urban models. Within each city there is gradient of house prices that typically decreases in distance to the centre to offset higher commuting costs. More populated cities tend to experience higher house prices at the centre. They also tend to extend over larger distances.

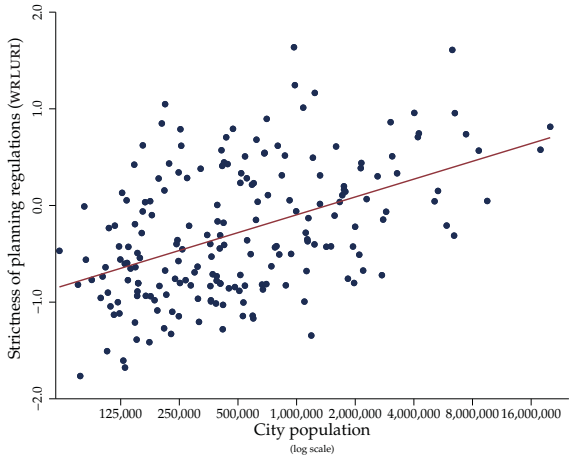
Figure 4 also shows the empirical relevance of the positive relationship between a city's population and housing prices at its periphery, a feature that is specific to our framework. Taking the rightmost value of each curve as the housing price at the edge or periphery of the metropolitan area, we can observe that prices at the periphery of New York are much higher than those at the periphery of Chicago, which themselves are higher than at the periphery of Atlanta, etc. Put differently, more populated cities tend to extend to a larger distance from their centre but not to

Figure 5: Housing regulation, periphery prices, and new construction in the United States

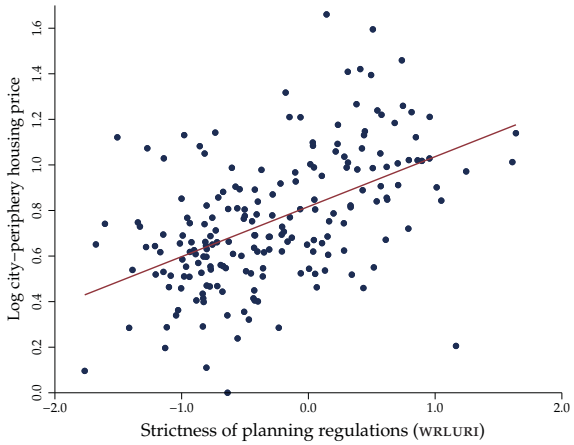
Panel A: Periphery prices and city population



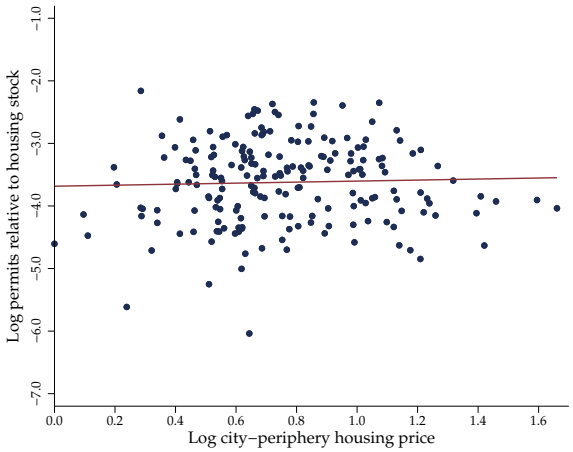
Panel B: Planning regulations and city population



Panel C: Periphery prices and planning regulations



Panel D: Residential permits and periphery prices



the point where housing prices at their periphery would be equalised.

While figure 4 provides an illustration for only a few cities, panel A of figure 5 systematically plots periphery housing prices against city population for all US metropolitan areas for which we we can perform the same calculation (see appendix B for details). As predicted by our model, we observe a clear positive relationship between city size and housing prices at the periphery.

The price of housing at the periphery of cities reflects, in principle, the price of undeveloped land at the periphery plus construction costs and the costs of planning regulations. Prices of undeveloped land vary little across the periphery of different cities (Burns *et al.*, 2018).[17] Construction

[17]In absence of planning regulations, we expect prices of undeveloped land at the periphery to equal the net present value of the return to land in the best alternative use, often agricultural, until the date of conversion to urban use plus the net present value of the return to land in urban use after that date minus conversion costs (Capozza and Helsley, 1989, Duranton and Puga, 2015). Empirically, however, variation in the value of land after its conversion to urban use contributes very little to the current value of agricultural land in the United States (Plantinga, Lubowski, and Stavins, 2002). The value of land while in agricultural use is also fairly homogenous and low, mostly below 8,000 dollars per acre (Burns *et al.*, 2018). The literature recognises that the irreversibility of housing development and uncertainty about future house prices also imply an option value for the price of land at the urban fringe (Capozza and Helsley, 1990, Duranton and Puga, 2015). The empirical study by Plantinga, Lubowski, and Stavins (2002) also considers this component.

costs are also fairly homogenous (Gyourko and Saiz, 2006).[18] Therefore, higher housing prices at the periphery or large metropolitan areas should be an indication of stricter planning regulations in these cities. Panel B of figure 5 provides direct evidence to that effect. We plot the Wharton Residential Land Use Regulatory Index (WRLURI) from Gyourko, Saiz, and Summers (2008), which measures the strictness of planning regulations, against the population of US metropolitan areas. As predicted by our model, we can observe that planning regulations are more stringent in larger cities. In turn, more stringent planning regulations effectively result in higher housing prices at the periphery, as can be seen in panel C, which plots the price of housing at the periphery of US metropolitan areas against the WRLURI measure.

In standard urban models, a city that receives a large positive shock may experience a rise in peripheral housing prices until sufficient new construction takes place. Then we should see more new constructions in cities where peripheral housing prices are (temporarily) higher. In our model, instead, planning regulations are used to prevent this expansion. Differences in the price of housing at the periphery are a feature of the steady-state and should not lead to differences in the expected growth rate of the housing stock. Instead, Gibrat's law is expected to hold as per equation (24), which applies to both population and the housing stock. Panel D of figure 5 confirms the absence of a systematic relationship between permits for new residential units relative to the housing stock and the price of housing at the periphery of metropolitan areas in the United States.

7. Empirical estimates of the model's key parameters

We now turn to the empirical estimation of the key parameters of the model regarding agglomeration and urban costs. In doing so, we develop three new approaches to estimate urban costs using different equations from our model and different sources of empirical variation. Our estimation of agglomeration is more standard but, rather than borrow directly from the literature, we connect the approach of De la Roca and Puga (2017) with our framework and use micro data for US metropolitan areas in similar time periods as in our estimations of urban costs. This has the benefit of maintaining consistency both with our modelling and across different parameters. Details regarding data sources and variable definitions are provided in appendix B.

Population elasticity of urban costs

The empirical urban economics literature has devoted little attention to estimating the elasticity of urban costs with respect to city size, focusing instead on the the elasticity of urban agglomeration benefits with respect to size (see Combes and Gobillon, 2015, for a review).[19] Our theoretical framework suggests three alternative approaches to estimate this elasticity, which we now implement.

The key urban cost parameter γ in our model first appears in the commuting cost equation (10) as the elasticity of a resident's commute with respect to the distance x between her dwelling and

[18] Gyourko and Saiz (2006) report a construction cost of about 116,000 dollars in year 2003 for a typical economy-quality house of 2,000 square feet. The coefficient of variation across 140 markets is 0.14.

[19] An exception is Combes, Duranton, and Gobillon (2019), but their approach to estimating urban costs does not provide a direct equivalent of γ.

the city centre. Taking natural logs of this equation and differentiating with respect to $\ln(x)$ yields:

$$\frac{\mathrm{d}\ln(T_{it}(x))}{\mathrm{d}\ln(x)} = \gamma \,. \tag{35}$$

We can estimate this equation exploiting variation in travel distance across individuals within a city as a function of how far they live from the city centre. Using the 2008–2009 US National Household Travel Survey, we estimate a regression at the household level of the natural log of vehicle-kilometres travelled by members of household j, $\ln(T_i^j)$, on the natural log of the distance between the household's residence and the centre of their metropolitan area i, $\ln(x_i^j)$:

$$\ln(T_i^j) = \gamma \ln(x_i^j) + a_i + \boldsymbol{X^j b} + \epsilon_i^j \,, \tag{36}$$

where a_i is a city fixed effect, $\boldsymbol{X^j}$ is a vector of household and neighbourhood characteristics that we control for, $\boldsymbol{b}$ is a vector of parameters, and ϵ_i^j is an error term. Results are shown in column (1) of table 1 and they imply a value for γ of 0.0729.

As we make progress towards solving our model, the same parameter γ reappears in key equilibrium relationships. In particular, once we solve for a spatial equilibrium within each city, the Alonso-Muth condition of equation (15) implies that, within each city, variation in commuting costs should be offset by variation in housing costs. It follows from this equation that $z\frac{\mathrm{d}[R_{it}(0)-R_{it}(x)]}{\mathrm{d}x} = \frac{\mathrm{d}T_{it}(x)}{\mathrm{d}x}$. Then, equation (14) can be rewritten as $z[R_{it}(0) - R_{it}(x)] = T_{it}(x)$. Dividing the previous equation by this one, multiplying both sides by x, and using natural logs to simplify leads to:

$$\frac{\mathrm{d}\ln(R_{it}(x) - R_{it}(0))}{\mathrm{d}\ln(x)} = -\frac{\mathrm{d}\ln(T_{it}(x))}{\mathrm{d}\ln(x)} = -\gamma \,. \tag{37}$$

The intuition is both straightforward and of fundamental importance: in equilibrium, indifference across locations requires that as individuals move to less central locations within their city and travel costs increase, housing costs fall in the same proportion. While this relationship is not new, and in fact is one of the key implications from the classic Alonso-Muth framework, it has surprisingly barely received any empirical attention (Duranton and Puga, 2015).

Thus, we can also estimate γ exploiting variation in house prices across locations within a city as a function of distance to the city centre. Using the 2008–2012 US American Community Survey, we estimate a regression at the block-group level of the natural log of the median rent for renter-occupied housing units in block group j, $\ln(R_i^j)$, on the natural log of the distance to the centre of its metropolitan area i, $\ln(x_i^j)$:

$$\ln(R_i^j) = a_i - \gamma \ln(x_i^j) + \boldsymbol{X^j b} + \epsilon_i^j \,, \tag{38}$$

where a_i is a city fixed effect, $\boldsymbol{X^j}$ is a vector of dwelling and neighbourhood characteristics that we control for, $\boldsymbol{b}$ is a vector of parameters, and ϵ_i^j is an error term. Note that the dependent variable in equation (38), $\ln(R_i^j)$, corresponds to $\ln(R_{it}(x))$ in the equilibrium equation (37), while the city fixed effect a_i absorbs $\ln(R_{it}(0))$. Results are shown in column (1) of table 1 and they imply a value for γ of 0.0734, almost identical to the 0.0729 value we obtained using transport data.

Our theoretical framework, in addition to suggesting two approaches for estimating the magnitude of urban costs from within-city variation in, respectively, travel and housing data, also

Table 1: Estimation of urban costs (model parameter γ)

	(1)	(2)	(3)
Dependent variable:	Ln household miles travelled	Ln block-group median house price	Ln city-centre house price
Ln distance to city centre	0.0729*** (0.0100)	-0.0734*** (0.0092)	
Ln city population			0.0721*** (0.0188)
Ln city travel speed			-1.0870*** (0.3199)
City indicators	Yes	Yes	
Controls	Block-group & household	Block-group & dwelling	For previous step
Observations	107,492 households	134,985 block-groups	182 cities
R^2	0.3186	0.2992	0.3021

Notes: All regressions include a constant term. Columns (1) and (2) include as block-group controls the percentages of hispanic, black, and asian population, the performance in standardised tests of the closest public school relative to the city average, indicators for waterfront and riverfront locations, and ruggedness. Column (1) also controls for the following household characteristics: natural log of size and of number of drivers, the share of drivers that are male, and indicators for a single-person household, for the presence of small children, for the household respondent being hispanic, white, black and asian, and for being a renter. Column (2) also controls for the following block-group dwelling characteristics: percentage dwellings in block group by type of structure, by number of bedrooms, and by construction decade. Column (3) involves a previous step to estimate the natural log of city-centre house prices for each city (the dependent variable) by regressing the natural log of block-group median house prices on a third-degree polynomial of distance to the city centre, and the same dwelling characteristics and block-group characteristics as column (2). Block-group characteristics are centred at the city mean, so that we predict the value of a national-reference house at the centre of each city for city-average block-group characteristics. In addition to this, column (4) involves previously estimating city travel speed by regressing travel speed for individual trips by private car on city indicators, including the same controls as column (1) in addition to the household's distance to the city centre and the following trip characteristics: the natural log of trip distance and indicators for day of the week, departure time in 30-minute intervals, and trip purpose. We use this to predict for each city the speed of a 15km commuting trip on a Tuesday at 8:00AM by a driver with average characteristics. ***, **, and * indicate significance at the 1, 5, and 10 percent levels. The R^2 reported in columns (1) and (2) is within city.

suggests a third approach relying on cross-city variation. Equation (14) implies that, while overall urban costs are split differently between the cost of commuting and the cost of housing depending on location within the city, they can be summarised for each city by the cost of housing at the centre, as given by equation (16). Taking logs of the latter:

$$\ln(R_{it}(0)) = \ln(\tau_{it}) + \gamma \ln(N_{it}) \,. \tag{39}$$

Thus, once we solve for the spatial equilibrium across cities, a larger population will get reflected in higher urban costs, increasing with the same elasticity γ.

To estimate γ from equation (39), we need a value for $\ln(R_{it}(0))$ in each city. A first possibility would be to use the rental price of a particular dwelling at the city centre. However, this approach is problematic since both the dwelling and the city centre may be unusual —and perhaps more so in large cities. A second possibility is to obtain an empirical counterpart to $\ln(R_{it}(0))$ from

equation (38).[20] Our focus when estimating (38) was to obtain an average housing price gradient, whereas now we must place more emphasis on measuring variation in $\ln(R_{it}(0))$ appropriately to relate it to differences in city size. Since there is potentially heterogeneity in house price gradients which is systematically related to city size, we prefer to allow for a more flexible estimation of the housing price gradient, both in terms of functional form and variation across cities. Thus, using the 2008–2012 American Community Survey as before, we estimate the following regression of the log of the median rent for renter-occupied housing units in block group j, $\ln(R_i^j)$:

$$\ln(R_i^j) = r_i - \boldsymbol{P}(x_i^j)\boldsymbol{c_i} + \boldsymbol{X^j b} + \epsilon_i^j \,, \tag{40}$$

where r_i is a city fixed effect, $\boldsymbol{P}(x_i^j)$ is a polynomial of distance to the city centre x_i^j, $\boldsymbol{X^j}$ is a vector of dwelling and neighbourhood characteristics that we control for, $\boldsymbol{c_i}$ and $\boldsymbol{b}$ are vectors of parameters, and ϵ_i^j is an error term. If we centre neighbourhood characteristics at the city mean, the estimated city fixed effects $\hat{r}_i$ capture the value of a national-reference house for city-average neighbourhood characteristics at the centre of each city —i.e. when $x_i^j = 0$. In section figure 4 of section 6 we already displayed for five cities the housing price gradients that result from estimating equation (40). Each curve shows the price of housing units of comparable characteristics at distances ranging from zero to their 95th percentile in each metropolitan area. The value of each curve at distance 0 corresponds to $\hat{r}_i$ and is the price of housing at the city centre that we then use for our cross-city analysis.

In addition to the price of housing at the city centre, estimating γ from equation (39) also requires a measure for $\ln(\tau_{it})$. Recall that τ_{it} is the value of the time spent travelling over a unit of distance in each city. The literature on the value of travel time does not provide a good guidance of how precisely to measure this for our purposes. However, it does suggest that τ_{it} should be related to travel speed (since at higher speeds, covering the same distance will take less time). To a first approximation we can take τ_{it} to be close to inversely proportional to travel speed. To estimate travel speed in each city, we use the 2008–2009 National Household Travel Survey and regress travel speed for individual trips by private car on city indicators, while controlling for driver and trip characteristics. We use this regression to predict for each city the speed of a 15km commuting trip on a Tuesday at 8:00AM by a driver with average characteristics, as we use this predicted value as our measure of trip travel speed in each city.

Equation (39) then maps into a regression of the estimated price of housing at the centre of city i, $\hat{r}_i$, on the city's log population, $\ln(N_i)$, and the city's log travel speed:

$$\hat{r}_i = a + \gamma \ln(N_i) + \psi \ln(\hat{\tau}_i) + \epsilon_i \,, \tag{41}$$

where the term $\psi \ln(\hat{\tau}_i)$ lets τ_{it} not be exactly inversely proportional to estimated travel speed $\hat{\tau}_i$ (it would be with $\psi = -1$), a is a constant, and ϵ_i is an error term. This is shown in column (3) of table 1 and it implies a value for γ of 0.0721, not statistically different from the 0.0729 and 0.0734

[20] While equation (38) cannot be used to value housing at $x_i^j = 0$, this could be sidestepped by using $\text{arsinh}(x_i^j)$ instead of $\ln(x_i^j)$.

values that we obtained using, respectively, within-city variation in distance travelled and housing prices.[21]

Based on the very similar estimates for γ obtained from within-city variation in travel distance and housing prices and from cross-city variation in urban costs, we will use $\gamma = 0.07$ for our quantitative analysis below.

The other urban cost parameter in our model is θ, best interpreted as (minus) the elasticity of travel speed with respect to city population. This first appears in equation (11) which, after taking logs, maps directly into the following regression:

$$\ln(\tau_i) = a + \theta \ln(N_i) + \epsilon_i \ . \tag{42}$$

In the process of implementing our third approach to estimate γ, we have already obtained an estimate of travel speed in each city, $-\ln(\hat{\tau}_i)$, measured as the speed of a 15km commuting trip on a Tuesday at 8:00AM by a driver with average characteristics. If we replace $\ln(\tau_i)$ with $\ln(\hat{\tau}_i)$ in equation (42) and estimate this regression, we obtain an estimated value of θ of 0.0388 with a standard error of 0.0032 and an R^2 of 0.438. We will therefore use $\theta = 0.04$ for our quantitative analysis.

Population elasticity of urban benefits

The magnitude of agglomeration economies in our model is reflected in the relationship between earnings and city population in equation (9) and, more richly, in equation (32), where we allow workers' human capital to reflect both education and initial work experience and to be systematically related to city size. As a result, two parameters, σ and β, must be estimated.

While we could attempt to estimate equation (32) from average city earnings, we prefer to use longitudinal worker-level information. This offers two important benefits. First, it allows us to condition out individual heterogeneity in initial human capital, as well as heterogeneity in occupations and sectors, which are absent from our model. Second, and most importantly, we can estimate σ and β separately. While our model simplifies the life-cycle of individuals to a single period and location, in practice they move and acquire experience in different cities. Thus, using longitudinal information we can separate the agglomeration economies associated with working in a bigger city at a given point in time (reflected in σ) from the additional value of early work experience when this is acquired in a bigger city (reflected in β).

Following De la Roca and Puga (2017), we first estimate the following individual earnings regression:

$$y_{it}^j = a_i + a_j + a_t + \sum_i b_i e_{it}^j + \boldsymbol{X}_{\boldsymbol{t}}^{\boldsymbol{j}} \boldsymbol{b} + \epsilon_{it}^j \ , \tag{43}$$

where a_i is a city fixed effect, a_j is a worker fixed effect, a_t is a time fixed effect, e_{it}^j is the experience acquired by worker j in city i up until time t, $\boldsymbol{X}_{\boldsymbol{t}}^{\boldsymbol{j}}$ is a vector of time-varying individual and job characteristics, the scalar b_i and the vector $\boldsymbol{b}$ are parameters, and ϵ_{it}^j is an error term.

[21] The coefficient on travel speed is not precisely estimated but the point estimate is close to -1, suggesting that having τ_{it} inversely proportional to travel speed is a good approximation.

Table 2: Estimation of agglomeration economies (model parameters σ and β)

	(1)	(2)	(3)	(4)	(5)
Estimation method:	OLS	TSLS		OLS	
Dependent variable:	Ln earnings			Initial premium (city indicator coefficients column (3))	Medium-term premium (initial + 8.4 years local experience)
Ln city size	0.0443*** (0.0045)	0.0423*** (0.0050)		0.0452*** (0.0045)	0.0770*** (0.0063)
City indicators			Yes		
Worker fixed-effects	Yes	Yes	Yes		
Experience in cities $\geqslant$ 5 million	0.0195*** (0.0027)	0.0196*** (0.0027)	0.0193*** (0.0028)		
Experience in cities $\geqslant$ 5 million $\times$ exp.	-0.0004*** (0.0001)	-0.0004*** (0.0001)	-0.0004*** (0.0001)		
Experience in cities 2-5 million	0.0068** (0.0032)	0.0069** (0.0032)	0.0075** (0.0032)		
Experience in cities 2-5 million $\times$ exp.	-0.0002 (0.0001)	-0.0002 (0.0001)	-0.0002* (0.0001)		
Experience	0.0633*** (0.0040)	0.0632*** (0.0040)	0.0628*** (0.0039)		
Experience^2	-0.0007*** (0.0001)	-0.0007*** (0.0001)	-0.0007*** (0.0001)		
Observations	50,733	50,394	50,733	63	63
R^2	0.3416	0.3416	0.3439	0.4858	0.6678

Notes: All regressions include a constant term. Columns (1), (2), and (3) include firm tenure and its square, and two-digit sector, occupation, and year indicators. Worker values of experience expressed in years. In the TSLS estimation of column (2) we instrument the natural log of city size with the arsinh of city size in 1850 and 1920, the arsinh of distance to Eastern Seaboard, heating degree days, and the mean terrain ruggedness index and range in elevation within 30km of the city centre. The city indicators in column (3) aggregate the 261 metropolitan areas included in the panel into 63 groups, with individual indicators for all metropolitan areas with population above 2 million and additional indicators for groups of similar-size metropolitan areas with population below 2 million. City medium-term premium calculated for workers' average experience in one city (8.4 years). Coefficients are reported with robust standard errors in parenthesis, which are clustered by worker in columns (1)-(3). ***, **, and * indicate significance at the 1, 5, and 10 percent levels. The R^2 reported in columns (1)-(3) is within workers.

Then, in a second step, we regress the estimated city fixed effects on city population to obtain a value for σ:

$$\hat{a}_i = \sigma N_i + \epsilon_i \ . \tag{44}$$

We can incorporate the additional advantages of larger cities arising from a greater value of job experience by re-estimating this regression after adding to the same city fixed effects the differential value of local experience, valued at the average local experience $\bar{e}$:

$$\hat{a}_i + \hat{b}_i \bar{e} = (\sigma + \beta) N_i + \epsilon_i \ . \tag{45}$$

This gives us a value for $\sigma + \beta$ and, subtracting from this the value of σ estimated from (44), yields an estimate for β.

If we were just interested in σ, we could also estimate this relationship in a single step by re-

placing a_i in equation (43) with the right-hand-side of equation (44).[22] Column (1) in table 2 shows results for this one-step estimation. The table uses panel data from the National Longitudinal Survey of Youth 1979 (NLSY79), which allows us to track individuals' location and labour market activities over their entire careers. The estimation yields a value for σ of 0.0443. This captures the elasticity of earnings with respect to city population upon moving to a different-sized city. The regression also shows that work experience is significantly more valuable when acquired in bigger cities. A first year of experience in a city of over 5 million people increases earnings by about one-third more relative to the baseline value outside cities with over 2 million (0.0195 coefficient for experience in cities above 5 million compared with 0.0633 coefficient for experience).

Our empirical approach to estimate σ is motivated by equation (32) of our model, while treating A_{it} as exogenous. However, equation (34) points to a systematic relationship between A_{it} and N_{it}. This suggests instrumenting for city size. Inspired by Ciccone and Hall (1996), we use as instruments for current city size the inverse hyperbolic sine of the city's population in 1850 and 1920 and of distance to the Eastern Seaboard.[23] The logic behind using historical population as an instrument is that there is substantial persistence in the spatial distribution of population (which ensures the relevance of the instruments), but the drivers of high productivity today greatly differ from those in the distant past (which helps satisfy the exclusion restriction). The distance to the Eastern Seaboard captures the Westward historical expansion of urbanisation in the United States.

Following Combes, Duranton, Gobillon, and Roux (2010), we also use geographical instruments likely to have affected agricultural productivity when this mattered for city sizes but unlikely to have important effects on city productivity today other than through the persistence of relative population differences: heating degree days (a measure of the coldness of climate), and the mean terrain ruggedness index and range in elevation within 30km of the city centre.

In the first stage of our instrumental variable estimation (not reported) all the instruments are significant, jointly and individually. They are also strong, as shown by the F-statistic for weak identification. In column (2) of table 2 we show results for the second stage of a 2SLS re-estimation of column (1), which yields very similar parameter estimates. In fact, according to the endogeneity test, the data do not reject the use of OLS. This is consistent with results in the literature, where instrumenting for current city sizes rarely makes much of a difference when estimating agglomeration economies.[24]

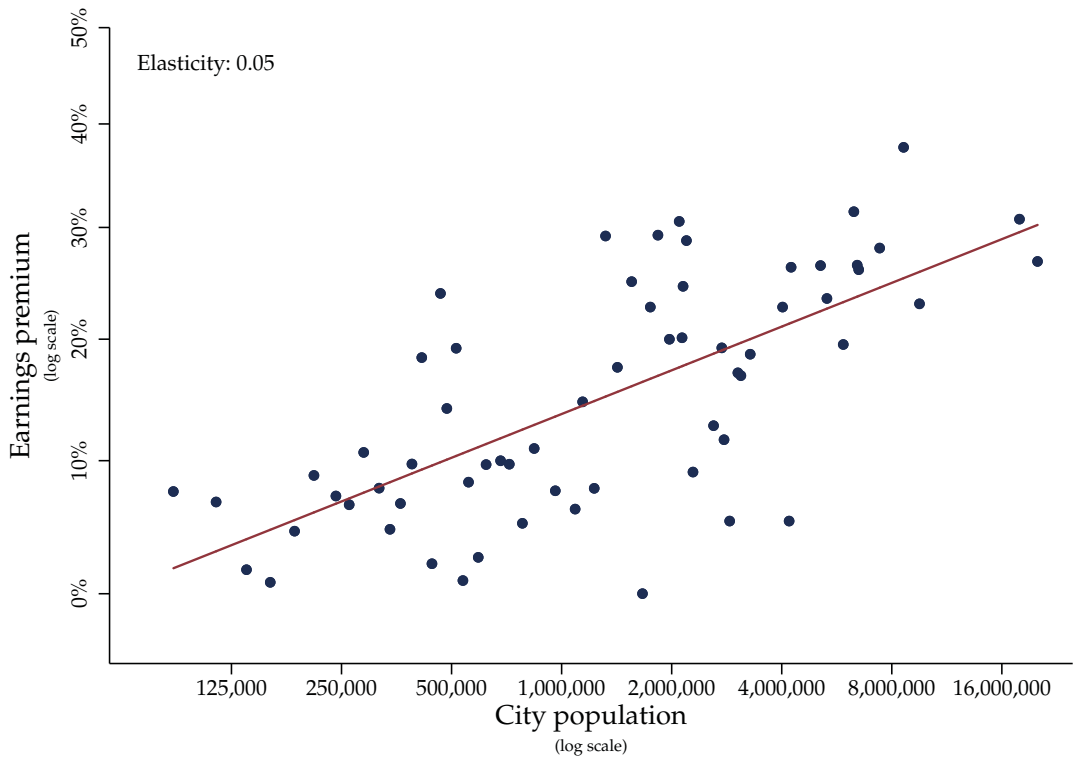
In column (3) we turn to the first step of the two-step estimation, corresponding to equation (43). Relative to the one-step procedure of columns (1) and (2), this replaces the natural log of city size with city fixed effects, which are then regressed on the natural log of city size in the second

[22] See Combes and Gobillon (2015), p. 258–260, for a discussion of why a two-step estimation is often preferable to a single-step estimation in this context.

[23] Since a few current US cities are in areas that were unpopulated back in 1850, we cannot take logarithms of historical population without losing observations. Thus, we use the inverse hyperbolic sine of the city's population, $\text{arsinh}(N_i) = \ln(N_i + \sqrt{(N_i)^2 + 1})$, which converges to $\ln(N_i) + \ln(2)$ and has the advantage that $\text{arsinh}(0) = 0$.

[24] In addition, the *LM* test of over-identifying restrictions (analogous to a Sargan test) rejects that the model is under-identified. That is, all our instruments estimate statistically similar parameters for the natural log of city size. See Combes and Gobillon (2015) for a discussion of instrumentation and alternative approaches to deal with endogeneity in this context.

Figure 6: The earnings premium of bigger cities in the United States



step shown in column (4).[25] Note that coefficients are almost identical across columns (1)-(3). The R^2 is also 0.34 in all three cases. Column (4) yields a value for σ of 0.0452, which is statistically indistinguishable from the estimate in column (1).

The additional earnings attained upon moving to a different-sized city are illustrated in figure 6. This plots city fixed effects from the individual earnings regression against instrumented city population, with both represented on a natural logarithmic scale. The slope of the fitted line corresponds to the estimate of σ from column (4).

Column (5) repeats the estimation of column (4) after adding to the same city fixed effects the differential value of local experience of column (3), valued at the average local experience in the sample, which is 8.4 years. This corresponds to equation (45) and allows us to incorporate the additional advantages of larger cities arising from a greater value of job experience. This yields a value for $\sigma + \beta$ of 0.0770. Subtracting from this the value for σ of 0.0452 from column (4), gives an estimate for β of 0.0318. Based on these empirical results, we will use $\sigma = 0.05$ and $\beta = 0.03$ for our quantitative analysis below.

[25]The sample size of the NLSY79 panel does not allow estimating a city fixed effect for smaller cities. Thus, when constructing city indicators for table 2, we aggregate the 261 metropolitan areas included in the panel into 63 groups, with individual indicators for all metropolitan areas with population above 2 million and additional indicators for groups of similar-size metropolitan areas with population below 2 million.

Population elasticity of rural income

The final parameter of our model is λ, which appears in equation (12) and corresponds to the population elasticity of rural income. This plays a relatively minor role for our quantification of the aggregate income effects of cities.[26] As already noted when introducing this equation, a natural way to fundament this is to think of λ as the income share of arable land in the rural sector. Based on this, we take $\lambda = 0.18$ from the estimate in Valentinyi and Herrendorf (2008) for agriculture in the United States.

8. The contribution of cities to aggregate income

One of our main goals is to evaluate the contribution of cities and the parameters that govern them to aggregate income. In this section, we investigate the implications of cities for aggregate income levels. Then, in the next section, we consider their effects on aggregate income growth.

We examine the importance of cities by restricting their sizes to some counterfactual level.[27] When assessing such counterfactuals, we treat the current distribution of population across cities and rural areas in the United States as the equilibrium of our model, where planning regulations are chosen by incumbents in each city through the local political process. We think of counterfactual city populations as being implemented by removing part of the local housing stock and changing the restrictiveness of the city's planning regulations. The size and planning regulations of other cities, the set of city locations that are populated, and overall urban population then adjust endogenously to new equilibrium levels.

We base the quantitative analysis that follows on the generalised version of the model that allows human capital to vary systematically with city population. This is because our estimation of β in the previous section confirms that the learning advantages of bigger cities that we introduced into our framework in section 5 are empirically relevant, in accordance with existing findings (Glaeser and Maré, 2001, Baum-Snow and Pavan, 2012, De la Roca and Puga, 2017). We use the parameter values estimated in our empirical analysis in section 7: $\gamma = 0.07$, $\theta = 0.04$, $\sigma = 0.05$, $\beta = 0.03$, and $\lambda = 0.18$.

Our model helps us to both quantify the aggregate consequences of smaller cities and structure them into several important components. Restricting city sizes by itself has a direct effect on the *nominal* income levels of the remaining city dwellers through a loss of agglomeration effects. This loss in nominal income also implies a loss in *real* income, but smaller in magnitude since agglomeration losses are partly offset by lower urban costs. In addition, those newcomers who remain in the city experience an additional loss from the stricter planning regulations used to implement city size restrictions. Some of those who are pushed away move to the rural sector. As they do so, rural income falls due to decreasing marginal returns to labour. This makes feasible the

[26] Changing the value of λ mainly affects the extent to which, in counterfactuals where we prevent more productive cities from expanding or reduce their populations, workers are pushed into infra-marginal cities as opposed to rural areas. Since, in equilibrium, income must be equated between the marginal city and rural areas, this barely affects aggregate income changes.

[27] We could also consider eliminating cities altogether, but a fully rural economy with no cities at all seems too extreme for a reasonable quantification.

development of new cities in relatively unproductive locations that, absent any restrictions, would remain unpopulated. Thus, people who are prevented from locating in existing cities also see their income fall.

Quantification procedure

Throughout our counterfactual exercises, we keep intrinsic differences across locations unaltered, but let agglomeration benefits and urban costs move with changing population levels. While such intrinsic differences are not directly observable, we can use equation (34) to solve for the combination of idiosyncratic productivity A_{it}, human capital h_t and commuting costs τ_t that would yield the actual population of each city as an equilibrium in our model:[28]

$$\frac{A_{it}(h_t)^{1+\sigma}}{\tau_t} = \frac{\gamma+\theta}{\rho^{\sigma}(\sigma+\beta)(\gamma+1)}(N_{it})^{\gamma+\theta-\sigma-\beta}. \tag{46}$$

The main complication is that our counterfactual exercise will give rise to additional cities on inframarginal sites that are not observed in the baseline. We must develop a procedure to infer what their equivalent level of $A_{it}(h_t)^{1+\sigma}/\tau_t$ might be. We rely on the fact that the distribution of city sizes is well approximated by a Pareto distribution and on the distribution of populated cities being a truncated version of the same distribution for all possible sites. We can thus estimate the shape parameter of a Pareto distribution fitted to the distribution of actual city populations, which will have the same shape parameter as the distribution for all sites. To obtain the equilibrium population of these additional potential locations for cities if they were populated, we take random draws from a Pareto distribution with the shape parameter estimated from the data.[29] Equation (46) then gives the value of $A_{it}(h_t)^{1+\sigma}/\tau_t$ that corresponds to each of these additional potential locations.

For our quantification, we also need to classify residents in every city into incumbents and newcomers. In keeping with the conventional generation length of 30 years, we consider the number of incumbents in each city equal to the city's population in 1980 and the number of newcomers equal to the increase in the city's population between 1980 and 2010.

Capping population in the largest cities

Our first set of counterfactuals evaluates the importance of cities for aggregate income levels by restricting the population of the largest cities in the United States. We begin with a moderate

[28] To simplify exposition, we have chosen to set up our model loading all differences across cities into productivity A_{it}. However, our quantification is agnostic about how differences across cities are split between productivity, human capital, and transportation. In a more general formulation, we would still back out the value of $A_{it}(h_{it})^{1+\sigma}/\tau_{it}$ from the distribution of city sizes using a version of equation (34) generalised to allow initial human capital and transportation to vary across cities.

[29] The shape parameter, estimated as in Gabaix and Ibragimov (2011), is equal to -0.86. The scale parameter needs to be low enough that we have enough potential additional locations at smaller sizes, and we set this to 1,000. Finally, since the number of cities is discrete, the number of draws we take also matters: a larger number of draws results in more cities within any given size interval. We take 11,000 draws because in a typical run this yields approximately the same number of draws with population above 75,000 as in the actual 2010 US distribution, 274. We can then extend the actual city-size distribution with a set of additional locations that correspond to the draws we obtain with a population below 75,000.

version of this thought experiment, where we constrain the population of the two largest US cities, New York and Los Angeles, to be no larger than that of the third largest city, Chicago. Implicitly, this exercise involves removing part of the housing stock in these two cities and ramping up local planning regulations so that it is not rebuilt.

Restricting city sizes by itself has a direct effect on nominal income levels. Equation (34) expresses the equilibrium size of each city N_{it} as a function of parameters and we take this equilibrium size to be the actual size that we observe. Let us denote by $\hat{N}_{it}$ the size of each city under the counterfactual scenario that we wish to evaluate. The level of nominal income an individual gets in city i under the counterfactual population level $\hat{N}_{it}$ relative to the baseline level under the city's actual population level N_{it} is easily calculated dividing equation (32) valued at $\hat{N}_{it}$ by the same equation valued at N_{it}:

$$\frac{\hat{y}_{it}}{y_{it}} = \left(\frac{\hat{N}_{it}}{N_{it}}\right)^{\sigma+\beta}. \tag{47}$$

Agglomeration economies imply that, in each city whose size changes, differences in the natural log of income per person are equivalent to differences in the natural log of this city's population multiplied by a factor $\sigma + \beta$. With a value of $\sigma + \beta = 0.05 + 0.03 = 0.08$, equation (47) implies that lowering New York's population from 20 million to Chicago's 9.5 million reduces the nominal income of incumbent New Yorkers who remain in the city by 5.8%.[30] In the case of incumbent Angelenos, the nominal income loss is 4.9%.

Equation (47) gives only a partial view of the effects of changes in city size because urban costs also depend on population. A smaller population in the largest cities lowers land rents (although, as we discuss below, general equilibrium housing costs for newcomers nevertheless increase because of stricter planning regulations). It also results in shorter average commutes and, by alleviating congestion, speeds up travel. Thus, to get a more complete measure of the aggregate implications of changing city sizes we must quantify the consequences for real income levels, i.e., for the level of individual consumption income net of housing and commuting expenditures. From equation (33), we can write the value of individual consumption in each city under its counterfactual population $\hat{N}_{it}$ as

$$\hat{c}_{it} = \rho^{\sigma} A_{it} (h_{it})^{1+\sigma} (\hat{N}_{it})^{\sigma+\beta} - \frac{\tau_t}{\gamma+1} (\hat{N}_{it})^{\gamma+\theta}. \tag{48}$$

Note that this value applies to incumbent residents who remain in the city. For newcomers who remain, we need to subtract from $\hat{c}_{it}$ the new level of planning regulations $\hat{p}_{it}$ required to implement the counterfactual city size. Combining equations (34) and (48) yields

$$\hat{c}_{it} = \left[(\gamma+\theta)\left(\frac{\hat{N}_{it}}{N_{it}}\right)^{\sigma+\beta} - (\sigma+\beta)\left(\frac{\hat{N}_{it}}{N_{it}}\right)^{\gamma+\theta}\right] \frac{\tau_t (N_{it})^{\gamma+\theta}}{(\sigma+\beta)(\gamma+1)}. \tag{49}$$

[30] This is larger than the actual 12.6% difference in median earnings between New York and Chicago. However, it is important to note that the counterfactual exercise limits New York's population, thus attenuating agglomeration economies, but maintains its intrinsic productivity advantage over Chicago.

Dividing equation (49) by the same equation valued at $\hat{N}_{it} = N_{it}$ then gives a closed-form solution for the relative real income loss for incumbents from restricting city sizes:

$$\frac{\hat{c}_{it}}{c_{it}} = \frac{\gamma + \theta}{\gamma + \theta - \sigma - \beta} \left(\frac{\hat{N}_{it}}{N_{it}} \right)^{\sigma + \beta} - \frac{\sigma + \beta}{\gamma + \theta - \sigma - \beta} \left(\frac{\hat{N}_{it}}{N_{it}} \right)^{\gamma + \theta}. \tag{50}$$

We can use equation (50) to evaluate the real income losses for incumbent residents who remain in New York and Los Angeles once we restrict their sizes to be no larger than Chicago. The first term on the right-hand side captures the negative effect of restricting the size of some cities through the loss of agglomeration economies that already appeared in equation (47). The second term captures the countervailing effect of reducing urban costs. Of course, given that $\hat{N}_{it} = N_{it}$ maximises $\hat{c}_{it}$, it follows that $\hat{c}_{it}/c_{it} \leqslant 1$. Overall, the real income losses for incumbent New Yorkers and Angelenos are much smaller than their nominal losses, only about 0.2%. The small real changes are partly a consequence of the small difference in magnitude between the population elasticities of urban benefits ($\sigma + \beta = 0.05 + 0.03 = 0.08$) and costs ($\gamma + \theta = 0.07 + 0.04 = 0.11$). They also partly reflect that we are evaluating population changes starting from the levels that balance out marginal benefits and costs from the point of view of local incumbents.

Real income losses are larger for remaining newcomers than for remaining incumbents in New York and Los Angeles, since they experience an additional loss from the stricter planning regulations used to keep the local populations at Chicago levels. In particular, the new level of regulation will make them experience the same real income loss experienced by those in rural areas. Of the 18.9 million displaced from New York and Los Angeles, 13 million end up in rural areas in the new general equilibrium under this counterfactual. With a value of $\lambda = 0.18$, equation (12) implies a 3.6% reduction in rural output per person and real income as the rural population increases from 57.5 to 70.5 million, due decreasing marginal returns to labour.

More stringent planning regulations also affect newcomers elsewhere. Incumbents in existing cities see no reason to let housing supply increase, since local fundamentals remain unchanged. However, to maintain the same level of population they now need more stringent planning regulations, so as to fend off pressure from those displaced from New York and Los Angeles. We regard this regulatory spillover as empirically relevant. When a city restricts its housing supply, this displaces people towards other cities, which may ramp up their own barriers to new development just to keep the status quo. The new level of planning regulations is given by equation (22), which shows that the cost of planning regulations everywhere needs to increase by as much as rural output per person has fallen. Thus, newcomers in every city experience the same 3.6% reduction in real income as a result.

The fall in rural income makes it feasible to develop new cities in relatively unproductive locations that, had the size of New York and Los Angeles not been restricted, would remain unpopulated. The 5.9 million people displaced from New York and Los Angeles to these infra-marginal city locations see their nominal incomes drop by between 45.3% and 85.8%, depending on where they end up. The 85.8% drop in nominal income for those who end up in the new marginal city is the same as for the 13 million people displaced to rural areas.

Real income loses for those who end up in a different location because of the counterfactual restrictions are smaller than the nominal loses, since they save on urban costs. Newcomers in

New York and Los Angeles were kept indifferent by planning regulations relative to locating in any other city or in rural areas. Population restrictions affect them negatively, but not by more than the 3.6% experienced by rural residents. The biggest individual losers, are those who could have been incumbents in New York and Los Angeles. With their homes in those cities torn down and regulations making it prohibitive to rebuild, and with more stringent regulations in other cities keeping them out, they end up as incumbents in previously infra-marginal cities at best. The real income losses for them are between 45.3% and 47.9%. However, since these 2.8 million New Yorkers and 6.4 million Angelenos are a relatively small fraction of the 307 million total 2010 population, they have a limited impact on average aggregate losses.

Pulling all of the above changes together, constraining the populations New York and Los Angeles to be no larger than that of Chicago results in an average reduction in income per person of 16.3% and an average reduction in real income per person of 3.4%.

We next consider more extreme versions of the counterfactual exercise that restricts the size of the largest US cities. If we restrict all cities to be no larger than 5 million people, this would force 11 cities with populations above that level (New York, Los Angeles, Chicago, Washington DC, San Francisco, Philadelphia, Dallas, Boston, Houston, Detroit and Atlanta) to become not much larger than Miami. This results in an average reduction in nominal income per person of 24.5% and an average reduction in real income per person of 7.9%. The main reasons why the losses are much larger than when we only restrict New York and Los Angeles to be the size of Chicago is the that there are many more people displaced to previously infra-marginal cities and rural areas and that a higher fraction of the displaced would have been incumbents in the most productive cities. If we restrict all cities to be no larger than 3 million people, the average income losses would be 32.7% in nominal terms and 12.5% in real terms.

Relaxing planning regulations

An important feature of our model is that incumbent city residents can limit the arrival of newcomers through a local political process that determines the level of planning regulations. This process implies distinct predictions for planning regulations, land prices, and housing development, which we showed in section 6 are first-order features of the US urban system. Planning regulations also fundamentally affect which potential city locations are populated and to what level.

Our model combines a tradeoff between agglomeration benefits and urban costs with productivity differences across locations. Models of urban systems often assume that the tradeoff between the benefits and costs of larger cities is resolved by competitive city developers. In the absence of productivity differences across locations, relying on competitive developers is not crucial, as the equilibrium with developers is the same as would be obtained if local governments actively set local population levels to maximise local incomes (Becker and Henderson, 2000). However, once we allow locations to differ in their productivity for a given population level, the equivalence between the population that maximises developer profits and the population that maximises local welfare breaks down (Albouy, Behrens, Robert-Nicoud, and Seegert, 2019).

Productivity differences across locations also create a potential for spatial misallocation. This is an important point brought to general attention by Hsieh and Moretti (2019). In our framework,

planning regulations benefit the incumbents who enact them focusing on the tradeoff between the local benefits and costs of a larger population, but represent an additional urban cost for newcomers and a source of deadweight loss for society. The most productive cities are inefficiently small in equilibrium and too many small and relatively unproductive cities remain in operation.

We now evaluate the aggregate consequences of planning regulations by examining a counterfactual where we eliminate them. This could raise population in the largest cities well beyond the empirically observed range. Worldwide, only Tokyo, Delhi and Shanghai have significantly larger populations than New York's roughly 20 million. This suggests that, above some population level, congestion (quantified by parameter θ in our model) makes urban costs rise very rapidly and this places a natural limit on how large cities can become. We therefore assume that the largest feasible city with current technology is only slightly more populated than the largest city observed in the world, and that above that congestion is prohibitive.

Under this counterfactual where planning regulations are completely lifted, New York would reach the largest feasible size we contemplate, 40 million people. The seven next largest cities in the United States would also be substantially larger than New York currently is, with their sizes determined by unrestricted migration in response to differences in productivity and the tradeoff between agglomeration benefits and urban costs. For instance, Philadelphia reaches 38 million but Boston stops at just under 30 million. These large population inflows would vacate less productive cities and also make population in rural areas fall sharply from 75 to 4 million.

All of these changes would bring gains in output per person of between 5.7% in New York and 13.3% in Boston through stronger agglomeration economies. However, rapidly rising urban costs would more than offset these gains for incumbent residents in these cities. Incumbent New Yorkers would experience large net real consumption losses of 13% while Bostonian would see only a minimal decline of 1.1%. The big winners would be those who, following the lifting of regulatory barriers to entry into the most productive cities, could now afford to move into these. Former residents of less productive locations and rural areas would see gains of between 0.5% (for former incumbents in Houston) and 60.9% (for rural residents). On average, income gains would amount to 34.9% in nominal terms and 25.7% in real terms.[31]

In a related exercise, Hsieh and Moretti (2019) conclude that aggregate income in the United States would rise by 8.9% if three of the most productive cities raised their housing supply elasticity (implicitly, by relaxing planning regulations) to the level of the median US city. We find a similar magnitude using our model for a comparable thought experiment. If we relax planning regulations in the three most productive cities to same level as in the median US city, average real income in our framework would rise by 8.2%.[32]

[31] The limit of 40 million we set on the population of the largest city is arbitrary, but does not have a major impact on these magnitudes. Setting the limit at 30 million instead would only lower the income gains to 31.1% in nominal terms and 23.8% in real terms. Setting it at 50 million would raise them to 37.2% in nominal terms and 26.3% in real terms.

[32] In their counterfactual, Hsieh and Moretti (2019) change the housing supply elasticity in three of the most productive cities. This is akin to changing the population elasticity of urban costs for the three most productive cities in our model. We instead keep the population elasticity of urban costs unaltered and change the level of planning regulations in these three cities from the endogenous level set by incumbent residents to the level of the median US city, while letting regulations everywhere else adjust to new equilibrium levels.

Despite our results being of similar magnitude, the mechanics that generate them and the distribution of gains and losses across groups differ significantly from Hsieh and Moretti (2019). Their formulation does not consider the tradeoff between agglomeration benefits and urban costs.[33] This has four implications. First, in their framework an increase in a city's population is always detrimental for existing residents. Second, since the optimal size of a city for an incumbent resident is zero, they take planning regulations as exogenous. Third, since this exogenous level of planning regulations affects housing costs for all local residents identically, their quantification does not distinguish between incumbent residents and newcomers. Fourth, changes in local population have larger effects on real income in their framework, where cities are subject to strong decreasing returns. In our framework, the tradeoff between benefits and costs brings them close to constant returns at the margin. An additional difference is that the number of cities is fixed in their model so, unlike in our framework, there is no adjustment at the extensive margin in terms of how many cities are populated and how population is distributed between cities and rural areas.

In Hsieh and Moretti (2019), increasing the housing supply elasticity in three particularly productive cities lets more workers enjoy the higher total factor productivity of these locations. However, with free mobility and large decreasing returns, much of the resulting aggregate gains are dissipated when these highly productive cities become much larger The marginal product of labour falls by 25% on average in these three cities, not only reducing the gains for new migrants but also hurting existing residents very substantially. In our framework, relaxing planning regulations in the three most productive cities increases output per worker for existing residents through agglomeration economies. However, rising house prices, longer average commutes, and greater congestion turn real income changes for incumbents into a modest fall. Residents of these three cities who were incurring the costs of stringent regulations in the baseline see much larger gains under the counterfactual, precisely because of these lower regulatory costs. New migrants into these cities also experience substantial gains in our framework as they move out of relatively unproductive cities and rural areas.

The increase in aggregate income in Hsieh and Moretti (2019) arises mainly because population losses in all but the three highly productive cities raise the marginal product of their remaining residents. In our model, incumbent residents outside the three most productive cities endogenously lower planning regulations to keep themselves unaffected when the expansion of the three most productive cities weakens pressure on their own housing market. Lower regulatory costs everywhere, not just in the three most productive cities, are an important source of aggregate gains in our framework. The other key source of aggregate gains in our framework is the equilibrium reallocation of population from relative unproductive cities and rural areas towards the three most productive cities. The gains from this reallocation remain large because worker inflows into those three locations affect both agglomeration benefits and urban costs, making their real income levels decline only modestly. In addition, the remaining rural residents in our framework also gain as population outflows raise their marginal product.

[33] In the extension where Hsieh and Moretti (2019) consider agglomeration economies, these are equivalent to lowering the magnitude of decreasing returns to labour. Thus, they do not create a tradeoff between the benefits and costs of a larger city in the sense that existing city residents always prefer a smaller population.

9. City population growth and aggregate income growth

Having studied the importance of cities for aggregate income levels, we now turn to quantifying their contribution to aggregate income growth. We conduct this exercise in four steps. First, we derive closed-form equilibrium expressions for the average growth rates of income and city population in our model. Next, we obtain values for these growth rates from US data, as well as for the evolution of human capital and travel costs. Third, we combine those values with the derived expressions to quantify the extent to which agglomeration economies and the intensive margin of city population growth amplify aggregate income growth. Finally, we consider a thought experiment where we prevent both the population growth of existing cities and the creation of new cities, and use this to examine the contribution of the extensive margin of city population growth to aggregate income growth.

The amplification effect of cities on aggregate income growth

We can compute the growth in log income per person by taking the natural logarithm of equation (32), time differencing it, and taking expectations to obtain

$$\mathbb{E}(\Delta \ln(y_{it})) = \mathbb{E}(\Delta \ln(A_{it})) + (1+\sigma)\Delta \ln(h_t) + (\sigma+\beta)\mathbb{E}(\Delta \ln(N_{it})) \,. \tag{51}$$

We can similarly derive the evolution of city population from equation (34) as

$$\mathbb{E}(\Delta \ln(N_{it})) = \frac{1}{\gamma+\theta-\sigma-\beta}\left[\mathbb{E}(\Delta \ln(A_{it})) + (1+\sigma)\Delta \ln(h_t) - \Delta \ln(\tau_t)\right] \,. \tag{52}$$

An important feature of these equations is the absence of dynamic scale effects, in the sense that the growth of neither aggregate income, human capital nor city population depends on their respective initial level. This is in contrast with the important static scale effects in city population associated with agglomeration effects and urban costs that we previously highlighted and explored.[34] The lack of of dynamic scale effects is a desirable property. For income and human capital, scale effects would either prevent growth or, on the contrary, lead to explosive growth. For city population, scale effects would eventually imply the concentration of the economy in a single city or convergence towards a single population size. Importantly, the lack of dynamic scale effects also implies that economic growth depends on changes, but not on levels, of city populations.

Turning to the role of the various parameters of our model, we first note that the agglomeration parameter σ magnifies the effect of human capital accumulation on aggregate income growth. The constant multiplying $\Delta \ln(h_t)$ in equation (51), which in the absence of cities would be 1, becomes $1+\sigma$ with cities.

In a second effect, city population growth contributes to income growth through the agglomeration economies that lead parameters σ and β to multiply $\Delta \ln(N_{it})$ in equation (51). This second effect incorporates both a direct component (city population growth matters for aggregate growth only if there are agglomeration economies) and an indirect component (agglomeration economies

[34] In the growth literature, the distinction is often framed in terms of strong versus weak scale effects (Jones, 2005). For a discussion of how city structure can prevent dynamic scale effects that would lead to explosive growth, see Rossi-Hansberg and Wright (2007).

also foster city population growth). This indirect component can be seen in equation (52) where, if we let the agglomeration parameters σ and β become very small, cities not only become very small, as per equation (34), but also grow very slowly. As shown by equation (51), any slowdown in city growth impacts negatively on income growth.

In contrast with σ and β, the parameters related to the costs of cities, γ and θ, do not affect the growth of income directly in equation (51) since they play no direct role in production. They nonetheless affect income growth indirectly through their role in city population growth in equation (52).

Empirical magnitude of changes in average income, city population and human capital

In light of equations (51) and (52), to assess the effects of cities and agglomeration on income growth, we need to know about the aggregate evolution of income per person, city populations, human capital and travel costs. Growth in income per person for the United States was 2.1% per year on average over the period 1950–2010 (US Bureau of Economic Analysis, 2019). Regarding city population growth, US metropolitan areas grew on average by 1.5% over the period 1950–2010. To a first rough approximation, we can measure the growth rate of human capital through changes in average years of schooling using the Current Population Survey. Between 1950 and 2010, average years of schooling grew at an average annual rate of 0.6%. Thus, in what follows, we use $\mathbb{E}(\Delta \ln(y_{it})) = \ln(1.021)$, $\mathbb{E}(\Delta \ln(N_{it})) = \ln(1.015)$, and $\Delta \ln(h_t) = \ln(1.006)$. Changes in travel costs are more difficult to quantify, so we deal with them separately in the next subsection.

Inferring changes in travel costs

Because τ_t parameterises the cost of commuting holding distance and city population fixed, there is no simple way to estimate changes in τ_t directly from the data. We use instead the structure of our model. More specifically, we can substitute the productivity and human capital terms in equation (32) and use the expression for y_{it} in equation (34) before taking the natural logarithm of the resulting expression, time differencing it, and taking expectations to obtain

$$\Delta \ln(\tau_t) = \mathbb{E}(\Delta \ln(y_{it})) - (\gamma + \theta)\mathbb{E}(\Delta \ln(N_{it})) \, . \tag{53}$$

This expression highlights how changes in commuting costs, income growth, and city population growth are related through the two parameters which characterise urban costs, θ and γ.

In section 7, based on three alternative estimations relying on different sources of variation, we consistently obtain values for γ very close to 0.07. We also estimate $\theta = 0.04$. Plugging $\mathbb{E}(\Delta \ln(y_{it})) = \ln(1.021)$, $\mathbb{E}(\Delta \ln(N_{it})) = \ln(1.015)$, $\gamma = 0.07$, and $\theta = 0.04$ into equation (53) leads to $\Delta \ln(\tau_t) = \ln(1.019)$, or about a 1.9% annual increase.

Does the structure of our model give us a reasonable value for $\Delta \ln(\tau_t)$? To a first approximation, we can think of changes in τ_t as resulting from the combination of changes in the time it takes to travel over a given distance and changes in the value individuals attach to that time.

Let us consider the speed of travel first. Using data from the 1995–1996 National Personal Transportation Survey and the 2001–2002 and 2008–2009 National Household Transportation Surveys

for the United States, and after factoring out the change in trip duration measurement following the 1995–1996 survey, Couture, Duranton, and Turner (2018) show that travel speed has remained roughly constant in US metropolitan areas.

With no change in travel speed, values of $\Delta \ln(\tau_t) = \ln(1.019)$ and $\mathbb{E}(\Delta \ln(y_{it})) = \ln(1.021)$ imply an elasticity of the value of travel time with respect to aggregate income of $\Delta \ln(\tau_t)/\mathbb{E}(\Delta \ln(y_{it})) = 0.92$. This implied elasticity is consistent with the findings of the large transport literature on the value of travel time. For instance, the meta-analysis of Abrantes and Wardman (2011) suggests an income elasticity of the value of travel time of 0.90 for all travel modes and 0.96 when focusing on car travel. Fosgerau (2005) also finds an after-tax income elasticity of the value of travel time of 0.90, while the panel-data approach of Swüardh (2008) yields an elasticity of 0.94.

The contributions of agglomeration and average city growth to aggregate income growth

We can now assess the quantitative contribution of cities to income growth. Agglomeration creates human capital externalities that magnify the returns to individual human capital accumulation and also makes city population growth matter for aggregate income growth. To get a sense of the magnitude of these effects, consider a thought experiment where we decrease agglomeration economies until they disappear.

Starting with the magnification of individual human capital accumulation, $\Delta \ln(h_t)$, is multiplied by $1 + \sigma = 1 + 0.05 = 1.05$ in equation (51). In the absence of agglomeration economies, it would be multiplied by 1 instead. Since growth in income per person in the United States is 2.1% per year on average in 1950–2010 and growth in human capital (proxied by years of education) over the same period is 0.6% per year, it follows that urban agglomeration economies raise the contribution of individual human capital to the annual growth rate of income in the US from 0.60 to 0.63 percentage points. Expressed as a fraction of the total, this represents less than 2% of the overall rate of income growth.

An important caveat here is that our model only considers agglomeration effects that percolate through the accumulation of human capital. Outside of our model, cities arguably foster innovation (Carlino and Kerr, 2015, Moretti, 2019). While we treat the dynamics of total factor productivity $\Delta \ln(A_{it})$ as exogenous here, a richer model that also considers innovation explicitly would have cities fostering the common component of total factor productivity, $\mathbb{E}(\Delta \ln(A_{it}))$, through innovation.[35]

The total stock of human capital in a city grows partly through accumulation at the individual level and partly through population growth bringing the human capital of more workers together. Thus, agglomeration economies also make the population growth of cities matter for aggregate income growth. Average city population growth, $\mathbb{E}(\Delta \ln(N_{it}))$, which is equal to an annual 1.5% for US cities in 1950–2010, is multiplied by $\sigma + \beta = 0.05 + 0.03 = 0.08$ in the last term of equation (51). The product of these two terms, capturing the effect of larger cities on the rate of growth in

[35]For city productivity shocks to remain independent and identically distributed, innovations would need to either diffuse very fast across cities (Desmet and Rossi-Hansberg, 2009) or be exploited in locations other than where they are created (Duranton and Puga, 2001).

income per person, is thus equal to an annual 0.12 percentage points. Expressed as a fraction of the total, this represents about 8% of the overall rate of income growth.

Note that part of this effect of city population growth is indirect: agglomeration economies make city population growth matter for aggregate income growth but they also make city population growth itself larger. As we let agglomeration economies disappear, average city population growth approaches 0.4% per year instead of the actual 1.5%.[36]

Combining the 0.12 percentage points in annual growth in income per person from average city population growth with the annual 0.03 percentage points from the magnification of individual human capital accumulation, we obtain a 0.15 percentage point difference. Over the period of 60 years used for our calculations, such difference adds up to 9.6% lower output per person in the absence of cities.

The extensive margin of city growth, spatial reallocation, and aggregate income growth

The above quantification relates average city population growth to aggregate income growth. By implicitly having all cities grow at the same expected rate, it misses the contribution of the extensive margin of city growth present in our model. In our framework, human capital accumulation and productivity increases make incumbents in existing cities willing to let local populations expand substantially. Since migrants go by preference to more productive locations, the expansion of these draws workers away from less productive cities and rural areas, leading to further aggregate gains. To bring these additional gains into our quantitative analysis, imagine stopping US city population growth in 1950. This implies keeping the population of every US city capped at its 1950 level and no new cities being created.[37] At the same time, we let total US population, total factor productivity, and human capital accumulation evolve just as they did between 1950 and 2010. The equations and procedure for evaluating this counterfactual are the same we used in section 8, but now focusing on the growth aspects.

Total US population doubled between 1950 and 2010, with an additional 140 million people going into cities and another 15 million going into rural areas. According to our model, if the majority went into cities it was because improving fundamentals allowed equilibrium city sizes to increase. Under the thought experiment where there is neither existing city growth nor new city creation after 1950, all 155 million would go into rural areas instead. This would lead to large losses both from not being able to accommodate more people into productive cities, as even incumbents would want, and from running rural areas deeply into decreasing returns. Under this counterfactual, the average annual growth rate in income per person between 1950 and 2010 would drop from the actual 2.1% to 0.8%. Real income losses would again be lower than nominal loses due to urban costs not increasing with city populations unchanged, but after 60 years of stagnant

[36]To calculate this figure, note that if we let the agglomeration parameters σ and β become very small in equation (52), at the limit $[\mathbb{E}(\Delta \ln(A_{it})) + (1+\sigma)\Delta \ln(h_t) - \Delta \ln(\tau_t)]$ becomes divided by $\gamma + \theta = 0.07 + 0.04 = 0.11$ instead of by $\gamma + \theta - \sigma - \beta = 0.07 + 0.04 - 0.05 - 0.03 = 0.03$. Recall that $\Delta \ln(h_t) = \ln(1.006)$ and $\Delta \ln(\tau_t) = \ln(1.019)$. $\mathbb{E}(\Delta \ln(A_{it}))$ can be calculated from equation (51) as $\ln(1.013)$. Pulling all of this together gives the 0.4% counterfactual value for $\mathbb{E}(\Delta \ln(N_{it}))$ as $\sigma + \beta$ approaches 0.

[37]It does not matter whether we assume the population cap at 1950 levels is implemented through planning regulations or by other means, since with no new construction in any city no-one actually incurs the costs of regulations.

city population growth they would still add up to 18.7% of individual consumption in 2010.[38] City growth in response to human capital accumulation, productivity growth, and the evolution in transportation is a powerful force for improving the spatial allocation of population.

10. Conclusions

We propose a new model of how cities and urbanisation interact with aggregate income and economic growth. In our framework, cities result from a tradeoff between agglomeration economies and urban costs. We model the agglomeration benefits of cities as arising from human capital spillovers, which foster entrepreneurship and learning. Unlike most of the past literature, we also pay close attention to the modelling of urban costs. A greater population in a city leads to its physical expansion and longer average commutes. A greater population also worsens congestion, slowing down travel. As we show, the relative magnitude of the benefits and costs of cities is fundamental to establish the contribution of cities to aggregate income and aggregate growth.

In our framework, the number and size of cities is endogenous. Productivity differences across locations lead cities to differ in their population size. As households seek to live in the most productive locations, this heterogeneity represents an important source of urban gains in addition to agglomeration economies. Driven by these potential gains, residents of less productive locations would be willing to move to more productive locations to the point of dissipating their advantage. For this reason, incumbent residents choose to limit the arrival of newcomers through planning regulations. The barriers imposed to newcomers represent another source of urban costs for them and a source of deadweight loss for society.

By modelling heterogeneity across locations as the outcome of cumulative productivity shocks, we also bring together in the same framework random urban growth and systematic urban growth driven by human capital accumulation and agglomeration effects. This combination allows our model to match key empirical features of modern urban systems, including city size distributions that follow Zipf's law, as well as ongoing urbanisation through a combination of gradual population growth of existing cities as more human capital is accumulated and new city creation. Our model also leads to novel predictions regarding planning regulations, house prices, and new constructions at the fringe of cities for which we provide empirical support.

We estimate key parameters that pertain to the costs and benefits of cities. To estimate urban cost parameters, we implement three novel approaches based on equations of the model at different levels of aggregation and using different sources of variation, which yield almost identical estimates. Using these parameter estimates and equilibrium conditions of our framework, we provide quantifications for various thought experiments.

We first quantify the importance of cities for the level of aggregate income and consumption by considering various caps on city sizes. For instance, capping the population size of the largest cities

[38] These large magnitudes are not driven by the doubling of total US population between 1950 and 2010. Repeating the same thought experiment under the assumption that total US population had not changed, still produces real income losses from 60 years of stagnant city population growth adding up to 16.9% of real income compared with 18.7% when total population doubles. In terms of the real income losses from capping city growth, leaving many workers in relatively unproductive cities is not very different from leaving them in rural areas.

to 5 million would reduce aggregate nominal income by about 25% and consumption per person by nearly 8%. The tradeoff between agglomeration economies and urban costs is at the root of large differences between the nominal and the real effects of limiting city populations. While capping the size of the largest cities leaves remaining residents with reduced agglomeration benefits, much of that loss is offset by lower urban costs. Our framework also allows us to quantify the potential gains from improving the spatial allocation of population by relaxing planning regulations in the three most productive cities at around 8% of real US income.

Next, we assess the effects of cities and urbanisation on economic growth. Having cities expand on average amplifies aggregate income growth through agglomeration economies. In addition, some cities grow more than others and this helps the spatial allocation of population follow heterogenous changes in fundamentals. Since the population growth of more productive cities draws workers away from cities with lower productivity levels and rural areas, this helps alleviate spatial misallocation further. Overall, we find that the US average growth rate in income per person would have been only 0.8% instead of 2.1% per year between 1950 and 2010 if city populations had not grown from their 1950 levels.

We envision several directions for further work. First, and quite obviously, our framework could be applied to other countries beyond the United States. There are both similarities and important differences across countries, which our framework can shed light on. For instance, our model predicts a weaker relationship between income growth and city population growth in developing countries that have seen traffic slow down substantially with urbanisation. Second, our modelling of housing production and consumption sacrifices realism for the sake of tractability and transparency. Allowing for higher buildings and smaller dwellings when land gets more expensive would capture relevant additional aspects of the urbanisation process, since these two margins further determine urban costs. Third, an important reason behind the large economic effects of restricting city populations is the exogenous determination of production amenities through the accumulation of random shocks. Allowing for labour mobility to affect productivity beyond the agglomeration effects that we already consider would be an important step and may alter some of our quantitative conclusions. Prior to that, strong empirical evidence on the subject would be needed to be able to quantify these effects. Finally, our analysis points to large costs associated with planning regulations and barriers to entry into highly productive cities. Further work should help with articulating policy solutions to this important problem.

Appendix A. Endogenous human capital accumulation

In this appendix, we show that, subject to some weak regularity conditions for the learning function, privately-optimal investments in human capital result in a constant rate of human capital accumulation over time. The learning function $b(\delta_t^j)$ gives the rate at which human capital increases as a result of investing a share δ_t^j of time into education. The worker devotes the remaining share $1 - \delta_t^j$ of her time to working.

Human capital also generates entrepreneurial ideas in proportion to the total local post-education level of human capital. It is natural then to assume that rewards to entrepreneurial

idea accrue to individuals in proportion to their human capital. Let π_{it} denote the reward to each of the m_{it} entrepreneurial ideas generated in city i at time t. To derive an expression for π_{it}, we must calculate the difference between the revenue and the cost of an intermediate producer. Let $s_{it}(\omega)$ denote the equilibrium sales price of intermediate variety ω produced in city i at time t. The minimisation of final production costs $\int_0^{m_{it}} s_{it}(\omega) q_{it}(\omega) \mathrm{d}\omega$ subject to the technological constraint of equation (1) yields conditional intermediate input demand:

$$q_{it}(\omega) = \frac{[s_{it}(\omega)]^{-\frac{1+\sigma}{\sigma}} Y_{it}}{\left\{ \int_0^{m_{it}} [s_{it}(\omega')]^{-\frac{1}{\sigma}} \mathrm{d}\omega' \right\}^{1+\sigma}} . \tag{A.1}$$

It follows from this expression that each intermediate firm faces an elasticity of demand with respect to its own price of $-(1+\sigma)/\sigma$. Marginal revenue can then be expressed as $s_{it}/(1+\sigma)$, where, due to symmetry across all intermediate producers in city i at time t, we have dropped the ω index for intermediate varieties. Given the intermediate production technology of equation (2), the marginal cost is simply the price per unit of human capital employed, denoted by w_{it}. Intermediate prices can be obtained by equating marginal revenue and marginal cost to obtain

$$s_{it} = (1+\sigma) w_{it} . \tag{A.2}$$

Each intermediate producer hires $\frac{H_{it}}{m_{it}}$ units of human capital and produces the same units of output. The returns to each entrepreneurial idea can then be computed as:

$$\pi_{it} = (s_{it} - w_{it}) q_{it} = \sigma w_{it} \frac{H_{it}}{m_{it}} . \tag{A.3}$$

With rewards to entrepreneurial ideas accruing to individuals in proportion to their human capital, individual income can be expressed as

$$y_t^j = m_{it} \pi_{it} \frac{h_t^j}{H_{it}} + w_{it} h_t^j = (1+\sigma) w_{it} h_t^j . \tag{A.4}$$

Worker j of the generation born at time t chooses how much time to devote to education to maximise her income. In the extended model of section 5, where human capital increases with city size with some constant elasticity, h_t^j is given by equation (30). Substituting this into equation (A.4), we see that the privately-optimal education will be defined by the following decision:

$$\max_{\{\delta_t^j\}} y_t^j = (1+\sigma)(1-\delta_t^j) b(\delta_t^j) \bar{h}_t^j (N_t^j)^{\beta} w_{it} . \tag{A.5}$$

Simplifying and re-arranging the first-order condition yields

$$\frac{b'(\delta_t^j)}{b(\delta_t^j)} = \frac{1}{1-\delta_t^j} . \tag{A.6}$$

Note that this same solution applies to the baseline model, where human capital does not vary across cities, since the human capital decision is a particular case of (A.5) where $\beta = 0$, which is absent from the first-order condition. To ensure the existence of a unique solution for δ_t^j such that $0 < \delta_t^j < 1$, we restrict $b(.)$ to be log-concave and sufficiently increasing such that $b'(0) > 1$. Then, regardless of their city of residence and the time t, all workers invest the same share of their time $\delta_t^j = \delta$ into education.

Appendix B. Data sources and treatments

City definitions. Our empirical and quantitative analysis focuses on the conterminous United States during the period 1980–2010. To define cities, we use Metropolitan Statistical Area and Consolidated Metropolitan Statistical Area (MSA) definitions outside of New England and New England County Metropolitan Area (NECMA) definitions in New England, as set by the Office of Management and Budget on 30 June 1999. This defines 275 metropolitan areas.

City centres and distances. We define the city centre as the location indicated by Google Maps for the core city of the metropolitan area. We measure the distance to the centre as the haversine distance between the centroid of each block-group and the centre of each metropolitan area.

Population. We use county-level population data from the US decennial censuses for 1850, 1950, 1980, and 2010, that we aggregate to the 1999 MSA/NECMA level. The sources are Schroeder (2016) for 1850 and 1920, Forstall (1996) for 1950 and 1980, and US Bureau of the Census (2012) for 2010.

Current Population Survey. Figure 3 plots the evolution of the share of population aged 25–64 who hold a college degree in metropolitan areas of different sizes over the period 1986–2016 in the United States. It uses data from the Annual Social and Economic (ASEC) supplement of the Current Population Survey (CPS), obtained from the IPUMS-CPS project (Flood, King, Rodgers, Ruggles, and Warren, 2018).

We are able to assign two thirds of individual-year observations in the source data to a specific metropolitan area. We make this assignment based on their county of residence, when available, which we then match to the corresponding 1999 MSA/NECMA; when the county of residence is unavailable, the state of residence is outside of New England, and the CPS source data contains the 1999 MSA of residence, we use this; alternatively, we use a purposely-built crosswalk (available with the replication code for this paper) between alternative metropolitan area codes contained in the CPS source data and 1999 MSA/NECMA codes.

We then group metropolitan areas into three size categories based on their 2010 population (below 1 million, between 1 and 2.5 million, and above 2.5 million), so that each line in the figure corresponds to the same set of metropolitan areas throughout.

For the one third of individual-year observations in the source data that we cannot assign to a specific metropolitan area, we can still assign them to the same three metropolitan area size categories based first on the metropolitan area size variable and next on the core-based statistical area size variable in the CPS. The downside of this procedure relative to be able to assign individual-year observations to a specific metropolitan area is that some observations may be assigned to different curves over time despite corresponding to the same metropolitan area if the population of this area crosses the 1 million or the 2.5 million thresholds.

Up until 1991, the CPS contains information on the years of college completed but not on whether the individual has obtained a bachelor's degree, so we classify individuals as having a college degree if they have completed at least 4 years of college. From 1992 onwards, we use the

information on whether they have a bachelor's degree or higher. We plot the figure using the ASEC person-level weights.

American Community Survey. Figure 4 plots housing price gradients for five US cities. Panels A, C and D of figure 5 relate city-periphery house prices to, respectively, city population, the strictness of planning regulations, and permits. Column (3) of table 1 estimates a regression for which the dependent variable is the city-centre house price. These housing price gradients, city-centre and city-periphery house prices are all based on the same regression. This regression uses 5-year 2008–2012 data from the 2012 American Community Survey (ACS), obtained from the IPUMS-NHGIS project (Manson, Schroeder, Riper, and Ruggles, 2018). The unit of observation is the block group. We use all block groups from all metropolitan areas except for college towns, defined as the 46 metropolitan areas with under one million inhabitants in 2010 where at least 10% of them are college students, since the high concentrations of students make housing markets in such college towns very distinct.

We regress the natural logarithm of the median contract rent for renter-occupied housing units in each census block group on a third-degree polynomial of distance to the centre, controls for housing and block-group characteristics, and metropolitan area fixed effects. The coefficients on the third-degree polynomial of distance to the centre are allowed to vary across cities, so that we can construct city-specific gradients. Block-group characteristics are centred at the city mean but housing characteristics are not, so that we predict the value of a national-reference house for city-average block-group characteristics.

The controls for housing characteristics are the percentage of dwellings in the block group by type of structure, by number of bedrooms, and by construction decade, all based on the same ACS data. The controls for block-group characteristics are the percentages of hispanic, black, and asian population (also from the ACS), the performance in standardised tests of the closest public school relative to the city average (from De la Roca, Gould Ellen, and O'Regan, 2014*a*, with variation at the tract level), an indicator for waterfront location (constructed by combining the 2012 block-group boundaries provided in the IPUMS-NHGIS ACS data with the coastline shapefiles from the National Hydrography Dataset and the Great Lakes and watersheds shapefiles from the Great Lakes Restoration Initiative of the US Geological Survey), an indicator for riverfront location (constructed by combining the same block-group boundaries with the major rivers within the United States shapefile included with Esri Data & Maps), and terrain ruggedness (measured by the Terrain Ruggedness Index of Riley, DeGloria, and Elliot, 1999, calculated on the basis of 1 arc-second Digital Elevation Models from the 3D Elevation Program of the US Geological Survey, 2018, and then averaged at the block-group level).

In such a regression, the estimated metropolitan area fixed effects correspond to the natural log price of a housing unit of nationally-comparable characteristics with city-average neighbourhood characteristics located at the centre of the city. We use this city-centre house price as the dependent variable in the specification in column (3) of table 1 and regress this on the natural logarithms of the city's 2010 population and a measure of its travel speed described below. Assigning values to distance to the city centre from zero to the 95th percentile distance across all housing units in the

metropolitan area and using the city-specific estimated coefficients on the distance polynomial, provides, in combination with the city-centre house price, the gradient of house prices in the metropolitan area. Figure 4 plots this gradient for five specific cities: New York, Chicago, Atlanta, Greensboro and Johnson City. The value of this price gradient at the 95th percentile distance across all housing units in the metropolitan area is our city-periphery house price, which we plot for all available metropolitan areas in panels A, C and D of figure 5.

Column (2) of table 1 uses the same data and variables to estimate the population elasticity of urban costs by exploiting variation in house prices across locations within a city. We regress the natural logarithm of the median contract rent for renter-occupied housing units in each census block group on the natural logarithm of distance between the household's block-group of residence and the city centre, the same controls for household and block-group characteristics we have just described, and metropolitan area fixed effects.

Planning regulations and building permits. The level of planning regulations in each metropolitan area plotted in panel D of figure 5 is measured using the Wharton Residential Land Use Regulatory Index (WRLURI). This index is constructed by Gyourko, Saiz, and Summers (2008) applying factor analysis to responses from a nationwide survey of residential planning regulations in over 2,600 communities across the United States. We keep data on the 1990 communities that are part of a 1999 MSA/NECMA. The measure we use is the average value of the WRLURI for each metropolitan area.

Data about the number of building permits plotted in panel D of figure 5 are from the US Department of Housing and Urban Development (HUD) from 2008 until 2012 (to match the timing of the ACS housing data). The source data is at the county level and we aggregate this up to the 1999 MSA/NECMA level. The variable log permits relative to housing stock on the vertical axis of panel D of figure 5 divides for each city the total number of units given residential construction permits 2008–2012 over the total number of housing units in the city for that period as recorded in the ACS data and then calculates the natural logarithm of this ratio.

All four panels of figure 5 are plotted for the same 198 metropolitan areas for which our estimate of the city-periphery house price, the level of planning regulations, and building permits are all available.

National Household Travel Survey. Column (1) of table 1 estimates the population elasticity of urban costs by exploiting variation in travel distance across households within a city. Data on household travel behavior come from the 2008–2009 US National Household Travel Survey (NHTS). The survey is sponsored by various agencies at the US Department of Transportation. For a nationally-representative sample of households, the NHTS provides a travel diary kept by every member of each sampled household where we observe the distance, duration, mode, purpose, and start time for each trip taken on a randomly-assigned travel day. It also includes household and individual demographics.

Household miles travelled are measured using the best estimate of household annual miles computed by the survey administrators, which is their preferred measure. We regress the nat-

ural logarithm of household miles travelled on the natural logarithm of distance between the household's block-group of residence and the city centre, controls for household and block-group characteristics, and metropolitan area fixed effects. The controls for household characteristics, all based on the same NHTS data, are the natural logarithms of the household size and of the number of drivers in the household, the share of drivers that are male, and indicators for a single-person household, for the presence of small children, for the household respondent being hispanic, white, black and asian, and for being a renter. The controls for block-group characteristics are the same as in the housing regression described above: the percentages of hispanic, black, and asian population in the block-group, the performance in standardised tests of the closest public school relative to the city average, indicators for waterfront and riverfront location, and terrain ruggedness.

The measure of travel speed in each city included as a control in the regression in column (3) of table 1 is based on the same NHTS data. We keep data on trips in a household vehicle, where this vehicle is a car, van, SUV, or pick-up, and is driven by the survey respondent. Following Couture, Duranton, and Turner (2018), we exclude all trips by households where either the respondent does not recall if they were the driver, or they report one or more trips in top or bottom 0.5% of all trips by distance, time or speed. As they note, removing all trips by the affected household and not only the odd ones is important to avoid biasing the calculations. Since speed varies very substantially depending on trip, individual and household characteristics, we need a minimum number of trips to compute a reliable measure of distance. We restrict our sample to the 182 cities where we have at least 100 trips recorded. We first calculate the speed of individual trips dividing trip miles by trip duration. We then regress the natural logarithm of travel speed for individual trips on metropolitan area fixed effects, controls for trip characteristics the same controls for household and block-group characteristics as in the regression in column (1) of table 1, and the natural logarithm of distance between the household's block-group of residence and the city centre. The controls for trip characteristics, all based on the same NHTS data, are the natural logarithm of trip distance and indicators for day of the week, departure time in 30-minute intervals, and trip purpose. We use the estimated regression coefficients to predict, for each city, the speed of a 15km commuting trip on a Tuesday at 8:00AM by a driver with average characteristics.

Census. While the controls for block-group characteristics in the transport regression in column (1) of table 1 and in the auxiliary regression to estimate travel speed for the regression in column (3) of table 1 are the same as in the housing regression in column (2) of table 1, the block-group definitions in the 2008–2009 NHTS correspond to the Census 2000 instead of to the 2012 ACS. We therefore re-compute these controls on the basis of Census 2000 data, obtained from the IPUMS-NHGIS project (Manson, Schroeder, Riper, and Ruggles, 2018).

National Longitudinal Survey of Youth. Our estimation of the parameters governing agglomeration economies in table 2 uses panel data from the "cross-sectional sample" of the National Longitudinal Survey of Youth 1979 (NLSY79). The survey, conducted by the US Department of Labor's Bureau of Labor Statistics, follows a nationally representative sample of 6,111 men and women who were 14–22 years old when they were first surveyed in 1979. These individuals

were interviewed annually through 1994 and were interviewed on a biennial basis since 1996. We use data for the period 1979–2012. The NLSY79 contains information on a rich set of personal characteristics and tracks individuals' labour market activities. Our starting panel is the same as in De la Roca, Ottaviano, and Puga (2014*b*) and we refer the reader to that paper for further details. For each respondent, the confidential geocoded portion of the NLSY79 reports the county and state where they were located at birth, at age 14, and at each interview date since 1979. We use that location information both to record the 1999 MSA/NECMA where each worker is currently employed and to split work experience accumulated until then into work experience in cities with populations equal or greater than 5 million, in cities with populations equal or greater than 2 million but below 5 million, and elsewhere. Since we need a reasonable number of observations to estimate city fixed effects, we include indicators for all metropolitan areas with population above 2 million and additional indicators for groups of similar-size metropolitan areas with population below 2 million. In particular, we have a common indicator for cities in groups that start at 75,000 people in increments of 25,000 until 600,000, then in increments of 50,000 people until 800,000, and then in increments of 100,000 people until 2 million. This aggregates the 261 metropolitan areas included in the panel into 63 groups. In the TSLS estimation of column (2) in table 2, we instrument the natural logarithm of city size with the arsinh of city size in 1850 and 1920 (from Schroeder, 2016), the arsinh of distance to Eastern Seaboard from the centre of each city (computed using coastline shapefiles from the National Hydrography Dataset of the US Geological Survey), heating degree days (from Burchfield, Overman, Puga, and Turner, 2006), and the mean terrain ruggedness index and range in elevation within 30km of the city centre (calculated on the basis of 1 arc-second Digital Elevation Models from the 3D Elevation Program of the US Geological Survey, 2018).

References

Abrantes, Pedro A. L. and Mark R. Wardman. 2011. Meta-analysis of UK values of travel time: An update. *Transportation Research Part A* 45(1): 1–17.

Albouy, David, Kristian Behrens, Frédéric Robert-Nicoud, and Nathan Seegert. 2019. The optimal distribution of population across cities. *Journal of Urban Economics* 110: 102–113.

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

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

Baum-Snow, Nathaniel and Ronni Pavan. 2012. Understanding the city size wage gap. *Review of Economic Studies* 79(1): 88–127.

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. 2014. Productive cities: Sorting, selection, and agglomeration. *Journal of Political Economy* 122(3): 507–553.

Behrens, Kristian and Frédéric Robert-Nicoud. 2015. Agglomeration theory with heterogeneous agents. In Gilles Duranton, J. Vernon Henderson, and William Strange (eds.) *Handbook of Regional and Urban Economics*, volume 5A. Amsterdam: Elsevier, 171–245.

Black, Duncan and J. Vernon Henderson. 1999*a*. Urban evolution of population and industry in the United States. *American Economic Review Papers and Proceedings* 89(2): 321–327.

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

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

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.

Burns, Christopher, Nigel Key, Sarah Tulman, Allison Borchers, and Jeremy Weber. 2018. *Farmland Values, Land Ownership, and Returns to Farmland, 2000–2016*. Washington DC: Economic Research Service, United States Department of Agriculture.

Capozza, Dennis R. and Robert W. Helsley. 1989. The fundamentals of land prices and urban growth. *Journal of Urban Economics* 26(3): 295–306.

Capozza, Dennis R. and Robert W. Helsley. 1990. The stochastic city. *Journal of Urban Economics* 28(2): 187–203.

Carlino, Gerald A. and William R. Kerr. 2015. Agglomeration and innovation. In Gilles Duranton, J. Vernon Henderson, and William Strange (eds.) *Handbook of Regional and Urban Economics*, volume 5A. Amsterdam: Elsevier, 349–404.

Carlino, Gerald A. and Albert Saiz. 2019. Beautiful city: Leisure amenities and urban growth. *Journal of Regional Science* 59(3): 369–408.

Champernowne, David G. 1953. A model of income distribution. *Economic Journal* 63(250): 318–351.

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

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

Combes, Pierre-Philippe, Gilles Duranton, and Laurent Gobillon. 2019. The costs of agglomeration: House and land prices in French cities. *Review of Economic Studies* 86(4): 1556–1589.

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 and Laurent Gobillon. 2015. The empirics of agglomeration economies. In Gilles Duranton, J. Vernon Henderson, and William Strange (eds.) *Handbook of Regional and Urban Economics*, volume 5. Amsterdam: Elsevier, 247–348.

Couture, Victor, Gilles Duranton, and Matthew A. Turner. 2018. Speed. *The Review of Economics and Statistics* 100(4): 725–739.

Couture, Victor and Jessie Handbury. 2019. Urban revival in America. Processed, University of California Berkeley.

Davis, Donald R. and Jonathan I. Dingel. 2019. A spatial knowledge economy. *American Economic Review* 109(1): 153–170.

Davis, Morris A., Jonas D. M. Fisher, and Toni M. Whited. 2014. Macroeconomic implications of agglomeration. *Econometrica* 82(2): 731–764.

De la Roca, Jorge, Ingrid Gould Ellen, and Katherine M. O'Regan. 2014*a*. Race and neighborhoods in the 21st century: What does segregation mean today? *Regional Science and Urban Economics* 47: 138–151.

De la Roca, Jorge, Gianmarco I.P. Ottaviano, and Diego Puga. 2014*b*. City of dreams. Processed, CEMFI.

De la Roca, Jorge and Diego Puga. 2017. Learning by working in big cities. *Review of Economic Studies* 84(1): 106–142.

Desmet, Klaus and J. Vernon Henderson. 2015. The geography of development within countries. In Gilles Duranton, J. Vernon Henderson, and William Strange (eds.) *Handbook of Regional and Urban Economics*, volume 5B. Amsterdam: Elsevier, 1457–1517.

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

Desmet, Klaus and Esteban Rossi-Hansberg. 2013. Urban accounting and welfare. *American Economic Review* 103(6): 2296–2327.

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

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. 2005. From sectoral to functional urban specialisation. *Journal of Urban Economics* 57(2): 343–370.

Duranton, Gilles and Diego Puga. 2014. The growth of cities. In Philippe Aghion and Steven N. Durlauf (eds.) *Handbook of Economic Growth*, volume 2B. Amsterdam: Elsevier, 781–853.

Duranton, Gilles and Diego Puga. 2015. Urban land use. In Gilles Duranton, J. Vernon Henderson, and William Strange (eds.) *Handbook of Regional and Urban Economics*, volume 5. Amsterdam: Elsevier, 467–560.

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: Elsevier, 555–677.

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

Fischel, William A. 2001. *The Homevoter Hypothesis*. Cambridge, MA: Harvard University Press.

Flood, Sarah, Miriam King, Renae Rodgers, Steven Ruggles, and J. Robert Warren. 2018. *Integrated Public Use Microdata Series, Current Population Survey: Version 6.0*. Minneapolis: University of Minnesota.

Forstall, Richard L. 1996. *Population of States and Counties of the United States: 1790 to 1990*. Washington DC: US Bureau of the Census.

Fosgerau, Mogens. 2005. Unit income elasticity of the value of travel time savings. In *Proceedings of the Association for European Transport European Transport Conference*.

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

Fujita, Masahisa, Paul R. Krugman, and Tomoya Mori. 1999. On the evolution of hierarchical urban systems. *European Economic Review* 43(2): 209–251.

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

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

Gabaix, Xavier. 2009. Power laws in Economics and Finance. *Annual Review of Economics* 1: 255–293.

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.

Gennaioli, Nicola, Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer. 2013. Human capital and regional development. *Quarterly Journal of Economics* 128(1): 105–164.

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.

Glaeser, Edward L. 2011. *Triumph of the City: How Our Greatest Invention Makes Us Richer, Smarter, Greener, Healthier, and Happier*. London: MacMillan.

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., Sari Pekkala Kerr, and William R. Kerr. 2015. Entrepreneurship and urban growth: An empirical assessment with historical mines. *Review of Economics and Statistics* 2(97): 498–520.

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

Gyourko, Joseph and Albert Saiz. 2006. Construction costs and the supply of housing structure. *Journal of Regional Science* 46(4): 661–680.

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.

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

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

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.

Hsieh, Chang-Tai and Enrico Moretti. 2019. Housing constraints and spatial misallocation. *American Economic Journal: Macroeconomics* 11(2): 1–39.

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.

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

Jones, Charles I. 2005. Growth and ideas. In Philippe Aghion and Steven N. Durlauf (eds.) *Handbook of Economic Growth*, volume 1B. Amsterdam: Elsevier, 1063–1111.

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

Manson, Steven, Jonathan Schroeder, David Van Riper, and Steven Ruggles. 2018. *Integrated Public Use Microdata Series, National Historical Geographic Information System: Version 13.0*. Minneapolis: University of Minnesota.

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

Michaels, Guy and Ferdinand Rauch. 2018. Resetting the urban network: 117–2012. *Economic Journal* 128(608): 378–412.

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. 2004*c*. Human capital externalities in cities. In J. Vernon Henderson and Jacques-François Thisse (eds.) *Handbook of Regional and Urban Economics*, volume 4. Amsterdam: Elsevier, 2243–2291.

Moretti, Enrico. 2019. The effect of high-tech clusters on the productivity of top inventors. Processed, University of California Berkeley.

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

Nagy, Dávid Krisztián. 2017. City location and economic development. Processed, Centre de Recerca en Economia Internacional.

Plantinga, Andrew J., Ruben N. Lubowski, and Robert N. Stavins. 2002. The effects of potential land development on agricultural land prices. *Journal of Urban Economics* 52(3): 561–581.

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

Riley, Shawn J., Stephen D. DeGloria, and Robert Elliot. 1999. A terrain ruggedness index that quantifies topographic heterogeneity. *Intermountain Journal of Sciences* 5(1–4): 23–27.

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

Saichev, Alexander I., Yannick Malevergne, and Didier Sornette. 2009. *Theory of Zipf's Law and Beyond*. Heidelberg: Springer.

Sánchez-Vidal, María, Rafael González-Val, and Elisabet Viladecans-Marsal. 2014. Sequential city growth in the US: Does age matter? *Regional Science and Urban Economics* 44: 29–37.

Schroeder, Jonathan P. 2016. *Historical Population Estimates for 2010* US *States, Counties and Metro/Micro Areas, 1790–2010*. Minneapolis: University of Minnesota.

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.

Swüardh, Jan-Erik. 2008. Is the intertemporal income elasticity of the value of travel time unity? Working Paper 2008:3, Swedish National Road & Transport Research Institute.

US Bureau of the Census. 2012. *Intercensal Estimates of the Resident Population for Counties and States: April 1, 2000 to July 1, 2010*. Washington DC: United States Bureau of the Census.

US Bureau of Economic Analysis. 2019. *Real gross domestic product per capita*. Washington, DC: United States Bureau of Economic Analysis. Retrieved from FRED, Federal Reserve Bank of St. Louis.

US Geological Survey. 2018. *1 Arc-second Digital Elevation Models –* USGS *National Map* 3DEP *Downloadable Data Collection*. Reston, VA: United States Geological Survey.

Valentinyi, Ákos and Berthold Herrendorf. 2008. Measuring factor income shares at the sectoral level. *Review of Economic Dynamics* 11(4): 820–835.