Urban growth and its aggregate implications

Gilles Duranton[*¶]
University of Pennsylvania

Diego Puga[*§]
CEMFI

Revised, 16 June 2022

Abstract: We develop an urban growth model where human capital spillovers foster entrepreneurship and learning in heterogeneous cities. Incumbent residents limit city expansion through planning regulations so that commuting and housing costs do not outweigh productivity gains from agglomeration. 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 the effect of cities on economic growth and aggregate output quantitatively.

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, Alba Miñano Mañero, 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 the editor, Dave Donaldson, three anonymous referees, Morris Davis, Vernon Henderson, David Nagy, Diego Restuccia, Matthew Turner, 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: https://diegopuga.org).

1. Introduction

Urbanisation and economic growth are tightly linked. The process of economic growth and development leads to increases in the number and population sizes of cities (Bairoch, 1988; Henderson, 2005; Desmet and Henderson, 2015). However, some urban scholars have argued that causation could go, in part, in the opposite direction, with cities and urbanisation being a primary engine of economic growth and prosperity (Marshall, 1890; Jacobs, 1969; Lucas, 1988; Glaeser, 2011). Despite widespread interest, isolating the aggregate implications of cities' number and population sizes on economic growth and aggregate income has proved elusive.

We propose a new model of how cities and urbanisation interact with aggregate income and economic growth. Our model relies on solid microfoundations to represent individual cities, matches fundamental empirical regularities at the city and economy-wide levels, and generates novel predictions for which we provide evidence. This model is amenable to a quantification that relies on a small number of parameters and remains transparent regarding the mechanisms at work. Based on key model equations, we estimate the parameters determining the magnitude of urban costs and benefits. These parameter estimates then allow us to assess the effect of cities and urbanisation on economic growth and aggregate income quantitatively and examine various 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*,*c*; Gennaioli, La Porta, Lopez-de-Silanes, and Shleifer, 2013). In addition to these direct productivity benefits, spillovers also indirectly affect the aggregate output level through the population size of cities. 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 a larger city population fosters agglomeration and thus increases output, it also leads to higher urban costs. We pay particular attention to the characterisation and quantification of these costs, a topic neglected by extant research. Thus, our microfoundations are also helpful in distinguishing between the gross and net benefits of larger cities and making welfare pronouncements.

More central locations in a city provide better accessibility, which gets reflected in higher house prices (Alonso, 1964; Muth, 1969). As cities grow in population, they expand outward, which leads to longer, more congested typical commutes and higher average house prices (Couture, Duranton, and Turner, 2018; Combes, Duranton, and Gobillon, 2019). Travel costs also evolve with technology and rising incomes. Because of differences in their natural geography, cities also differ in their ability to expand outward (Saiz, 2010; Nagy, 2020). All these elements affect the relationship between urban costs and city population in the cross-section of cities. They also drive the long-term evolution of the urban system.[1]

[1]We do not model the relative geographical position of cities (Fujita, Krugman, and Mori, 1999; Puga, 1999; Nagy, 2020) 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; 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). We focus instead on inequalities between incumbent residents and potential newcomers across cities.

With both benefits and costs to city size, our model features what Fujita and Thisse (2002) call the 'fundamental tradeoff' of urban economics. To resolve this tradeoff, we propose a political economy mechanism where each city uses planning regulations to set its population and balance the greater commuting and housing costs associated with larger cities against agglomeration benefits.[2] This mechanism is socially inefficient as 'incumbent residents' limit entry into their city to maximise their own welfare but do so at the expense of potential newcomers who remain stuck in less productive cities.

While our modelling of city formation through a local political process is intuitively appealing, it also implies novel empirical predictions. First, to avoid seeing their higher productivity dissipated in urban costs, incumbent residents in more productive and larger cities tend to impose more restrictive planning regulations. Regulations are also more stringent in cities that are more geographically constrained in their ability to expand. In turn, regulations open a wedge between house prices in the periphery and their replacement costs (calculated as construction costs plus the cost of a vacant agricultural land parcel). The magnitude of this wedge increases with the restrictiveness of regulations and thus with city population. Finally, the systematic variation in planning regulations with the productivity and population size of individual cities implies that there should be little relationship between the housing price-cost wedge in the periphery of cities and new housing construction. These predictions are in contrast with 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). Using US data, we find empirical support for all these predictions.

Our equilibrium replicates other key stylised facts about urban systems. As the economy develops and aggregate population grows, new cities appear and a dwindling proportion of the population remains in rural areas. This is consistent with the situation in the United States and 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 agreement with our model, past literature attributes much of the population growth of individual cities 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).

Because they are fundamental to establishing the contribution of cities to aggregate output and growth, we directly estimate the intensity of urban costs and agglomeration benefits. Regarding

[2]Models of urban systems in the tradition of Henderson (1974) often resolve this tradeoff by relying 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 consumption. 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).

urban costs, we implement a series of novel and complementary approaches. We successively rely on 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). These approaches all result in a similar elasticity of urban costs with respect to city population of about 0.07. These urban costs are amplified by congestion with a population elasticity, which we also estimate, of about 0.04.

For agglomeration economies, our model leads us to 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 0.04, and an elasticity in the longer term, incorporating learning effects, of close to 0.08. This is in line with previous estimates for other countries (Combes and Gobillon, 2015; De la Roca and Puga, 2017).

Armed with our parameter estimates, we first quantify the importance of cities for the level of aggregate output and consumption by running a thought experiment where we relax planning regulations in seven large cities where the wedge between house prices at the periphery and their replacement cost is above 200,000 dollars. This relaxation allows for more constructions and, in turn, leads to a counterfactual increase in population of up to 38% in New York and 27% on average in the seven cities. Overall, this counterfactual implies an increase in aggregate output of 7.95% and an increase in aggregate consumption of 2.16%.

Next we assess the effects of cities and urbanisation on economic growth. Agglomeration effects in cities and average city population growth magnify output growth. Weakening agglomeration effects to zero only has a modest effect on the growth rate of aggregate output, about 0.14 percentage point per year. Output growth also results from the better spatial allocation of population associated with the expansion of more productive cities in response to human capital accumulation, productivity growth, and transport improvements. Overall, we find that the reallocation of population towards and across cities increases the growth rate of aggregate output by about 0.66 percentage points per year.

Our framework builds on the extensive literature on systems of cities initiated by Henderson (1974) and reviewed in Behrens and Robert-Nicoud (2015). Of particular relevance to our work is the landmark model of Black and Henderson (1999*b*) which links urban and economic growth through human capital externalities in production. To assess the effect of cities and urbanisation on economic growth quantitatively, our model differs from theirs in three important ways. First, we rely on different microeconomic foundations, including more detailed modeling of the commuting technology to match both micro-estimates of commuting and housing costs and the empirical relationship between aggregate income growth and urban growth. Second, inspired by the insights in Albouy, Behrens, Robert-Nicoud, and Seegert (2019), we resolve the tradeoff between the benefits and costs of cities through a political economy mechanism with endogenously-determined planning regulations, instead of relying on an efficient market for cities (Becker and Henderson, 2000). Third, our model features two other drivers of growth in addition to endogenous human capital accumulation: transport costs and total factor productivity. By allowing for city-specific productivity shocks, we bridge the gap between models of random

urban growth like those proposed by Gabaix (1999) and Eeckhout (2004) and models of systematic drivers of urban growth like Black and Henderson (1999*b*).[3]

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 a rare dynamic analysis which complements ours, Davis, Fisher, and Whited (2014) use a neoclassical model of growth with physical capital. In their model, urban growth requires 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. Davis, Fisher, and Whited (2014) find that these channels boost aggregate growth by about 10%. In work which overlaps with ours, Hsieh and Moretti (2019) consider the misallocation of labour across cities that can occur because of planning regulations. They consider a static model where cities differ in their productivity and availability of land for production. 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.

2. An urban growth model

Locations and timing

Time is discrete, with periods subindexed by t. There is a continuum of potential sites for cities, subindexed by i. Potential sites for cities are heterogeneous, differing in time-varying underlying productivity and time-invariant geographical constraints to development. At any point in time, only a subset of potential sites hosts a city.

Total population in the economy, N_t, evolves exogenously. The number of cities, which sites they occupy, their population sizes, N_{it}, and the population of an alternative rural sector, N_{rt}, are all determined endogenously and vary over time.

Individuals live for two periods. In their first period, they are children cohabiting with their adult parents. In their second period, upon reaching adulthood, they can choose to remain in

[3]Gabaix (1999) and Eeckhout (2004) focus on a growth process resulting from the accumulation of city-specific shocks. This type of model generates realistic city size distributions, but leaves aside the systematic determinants of growth that are empirically important. These models must also impose a fixed number of cities since they assume that a reduction in population in any given city always increases output per person. On the other hand, Black and Henderson (1999*b*) and, more generally, the literature that focuses on systematic drivers of urban growth, does not naturally generate realistic city size distributions. 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*). 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.

their late parents' residence or move elsewhere. In every period t, the following sequence of events takes place.

First, the adult generation of the previous period passes away, while the children of the previous generation become adults and have one offspring each.

Next, the idiosyncratic production amenity in each city location i is updated to a new level by a multiplicative shock g_{it}, independently drawn from some common distribution with support $(1, \infty)$, so that the new level is $A_{it} = g_{it}A_{it-1}$.[4]

Adult residents in each location decide whether and by how much the housing stock in the city should expand. They do so by establishing more or less stringent planning regulations that create a nuisance regulatory cost on potential newcomers, which we refer to as a housing permitting cost (what Glaeser, Gyourko, and Saks, 2005, call a 'regulatory tax').

Adult residents in all cities and the rural sector can choose to either remain at their current residence or move, taking their children with them.[5] Any adult interested in becoming a new resident in a city can do so by incurring this city's permitting cost p_{it}, in addition to bidding for a one-period lease on one plot of land in this 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]

Having chosen a location for their adult life, each individual accumulates human capital through a process described next. Urban workers then commute between their residence and their job, receive their income, and consume housing and the numéraire good. Rural workers obtain their income at their place of residence and consume.

Human capital accumulation

We now turn to the human capital accumulation process. This takes place in three steps: compulsory education, voluntary further education, and on-the-job experience.

During childhood, all individuals receive compulsory education and achieve the average level of human capital attained after further education by the previous generation, $\bar{h}_t$.

During adulthood, each individual j chooses what share δ_t^j of the unit of time of her adult life to devote to further education. This raises her human capital level to $b(\delta_t^j)\bar{h}_t$, where the learning function $b(\delta_t^j)$ captures how further education raises the worker's human capital, and the average level of human capital of the previous generation is $\bar{h}_t \equiv \left(\int b(\delta_{t-1}^j)\bar{h}_{t-1}\mathrm{d}j \right) / \int \mathrm{d}j$. It is natural to assume that $b'(\delta_t^j) > 0$ and $b(0) = 1$. As more advanced knowledge becomes part of the standard

[4]The support $(1, \infty)$ implies non-negative shocks. 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.

[5]When population grows, the additional individuals also choose where to locate. We can think of population growth as resulting from international migration.

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

curriculum for the next generation, the same individual investment in further education results in higher human capital over time.

Having completed further education, each worker acquires some initial work experience in the city where they have chosen to spend their adult life. Early experience raises her human capital from the post-education level $b(\delta_t^j)\bar{h}_t$ to $b(\delta_t^j)\bar{h}_t(N_t^j)^\beta$.[7] Note that the proportionate increase is larger the bigger the city where she acquires this initial experience. This assumption 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. We assume that this knowledge is of a more personal nature and does not get passed on to the next generation.

After gathering this early experience, each individual works for the remaining share $(1-\delta_t^j)$ of her adult life. The amount of effective human capital provided by worker j to her employer in city i during her adult career in period t can then be expressed as

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

In appendix A, we show that the human capital accumulation process described by equation (1) results in a constant rate of human capital accumulation so that $b(\delta_t^j) = b(\delta)$. Then, the equilibrium level of human capital resulting from education and early job experience is the same for all workers in a given city and an increasing iso-elastic function of city size:

$$h_{it} = h_t N_{it}^\beta \ , \tag{2}$$

where $h_t = (1-\delta)b(\delta)\bar{h}_t = b(\delta)h_{t-1}$.[8]

City-size benefits: output and individual earnings

Suppose final output is produced under constant returns to scale and perfect competition by combining non-tradable intermediate inputs with a constant elasticity of substitution $\frac{1+\sigma}{\sigma}$, where $\sigma > 0$. Final output in city i at time t is then given by

$$Y_{it} = A_{it}\left\{\int_0^{m_{it}} [q_{it}(\omega)]^{\frac{1}{1+\sigma}} \mathrm{d}\omega\right\}^{1+\sigma} \ , \tag{3}$$

where ω indexes intermediate inputs, $q_{it}(\omega)$ denotes the quantity of intermediate ω used in final production, m_{it} denotes the endogenous mass of intermediates available in city i at time t, and A_{it} measures the local level of production amenities. Final output is freely tradable across cities and we use it as numéraire.

Intermediate inputs are produced using human capital as an input:

$$q_{it}(\omega) = H_{it}(\omega) \ , \tag{4}$$

where $H_{it}(\omega)$ is human capital employed by the firm producing intermediate ω. Since intermediate producers are symmetric, they each employ the same levels of human capital. Let H_{it} denote

[7] To simplify notation, we set the duration of this apprenticeship period to zero. Alternatively, we could increase the total length of adult life by its duration.

[8] This relationship between human capital and city population implies that cities of different sizes differ in terms of their levels of human capital but not in terms of their rate of growth of human capital. We document the empirical relevance of this implication in section 4.

the total level of local human capital after further education. This is further amplified by a factor $(N_{it})^{\beta}$ as a results of early job experience, as per equation (2). Thus, we can express intermediate output as

$$q_{it}(\omega) = q_{it} = \frac{H_{it}(N_{it})^{\beta}}{m_{it}} \ . \tag{5}$$

Substituting equation (5) into (3) yields:

$$Y_{it} = A_{it}\left[m_{it}(q_{it})^{\frac{1}{1+\sigma}}\right]^{1+\sigma} = A_{it}\left(m_{it}\right)^{\sigma} H_{it}(N_{it})^{\beta} \ . \tag{6}$$

Entrepreneurial ideas arise in proportion to the total local human capital after further education, H_{it}, 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:

$$m_{it} = \rho H_{it} \ , \tag{7}$$

Combining $H_{it} = h_t N_{it}$ with equations (6) and (7), we can express output per worker as:[9]

$$y_{it} = \frac{Y_{it}}{N_{it}} = \rho^{\sigma} A_{it}(h_t)^{1+\sigma}(N_{it})^{\sigma+\beta} \ . \tag{8}$$

There is ample evidence regarding the productivity benefits of bigger cities, and urban economists have reached a broad consensus about their magnitude (see Rosenthal and Strange, 2004; Combes and Gobillon, 2015; Ahlfeldt and Pietrostefani, 2019, for reviews). While many mechanisms can give rise to such agglomeration economies (Duranton and Puga, 2004), existing studies attribute a crucial role to human capital externalities and entrepreneurship (Moretti, 2004*b*; Glaeser, Kerr, and Kerr, 2015). In our framework, bigger cities concentrate more human capital, which 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 more local intermediate producers. These advantages are amplified by the greater value of early job experience in bigger cities.

City-size costs: Housing and transportation within each city

Bigger cities feature not only stronger agglomeration economies but also higher urban costs. To characterise these costs, we next 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.

We capture heterogeneity across cities in their geographic ability to expand by assuming that each raw unit of land only provides $\frac{1}{z_i}$ units of land suitable for housing where $z_i > 1$ is a city-specific parameter. Hence, cities with a higher z_i are more geographically constrained.

[9]In appendix A, while deriving the individually optimal allocation of time, we disentangle relative rewards to human capital and entrepreneurial ideas. However, to determine the 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, with perfect competition in the final good sector, is the aggregate value of final city output. We assume that early job experience is valuable for production but not for generating new ideas merely to simplify notation.

All city dwellings provide one unit of floorspace built on one unit of land suitable for development, thus requiring z_i units of raw land each.[10] 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.

City residents must commute to access their jobs. The commuting costs of a resident who resides at a distance x from the city centre are given by

$$T_{it}(x) = \tau_{it} x^{\gamma} \, . \tag{9}$$

The length of each city resident's commute increases with elasticity $\gamma > 0$ with the distance x between her dwelling and the city centre.[11] Individual commuting costs are then the result of multiplying the distance travelled, x^{γ}, by the cost per unit of distance, τ_{it}, where

$$\tau_{it} = \tau_t (N_{it})^{\theta} \, . \tag{10}$$

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 can vary over time, allows us to consider changes in commuting technology, altering, for instance, how much travellers value time in vehicles or the speed at which they travel.

The rural sector

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 a rural area, in which case they attain a level of individual income

$$y_{rt} = A_{rt} (N_{rt})^{-\lambda} \, , \tag{11}$$

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 some specific factor in fixed supply, such as arable land, in a rural production function with constant aggregate returns to scale.[12]

3. The number and sizes of cities

Thriving cities in the United States tend to restrict population growth through planning regulations that are widely seen as protecting the interests of incumbent residents at the expense of potential

[10] With fixed housing consumption, maximizing utility is equivalent to maximizing final good consumption.

[11] This specification is more general than the usual linear commuting technology of the monocentric model. Our generalisation has an empirical motivation: in section 5 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. We can also think of this formulation as a reduced-form way to account for features that the monocentric model abstracts from, including secondary employment centres. We can still recover the classic specification where $\gamma = 1$ as a particular case.

[12] More specifically, equation (11) 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. An even simpler alternative would be to have an outside option with a fixed rural income y_r, but this would mean no consumption growth for rural dwellers and the eventual disappearance of the rural sector.

newcomers. We now derive the equilibrium number and sizes of cities arising from our modelling of this local political process.

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 permitting cost p_{it} anticipating she will have to bid $z_i R_{it}(x)$ for z_i raw units 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} and a participation R_{it}/N_{it} in total land rent in the city R_{it}. The maximum bid per unit of raw land $R_{it}(x)$ 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:

$$y_{it} - T_{it}(x) - z_i R_{it}(x) + \frac{R_{it}}{N_{it}} - p_{it} = c_t = y_{rt} \ , \quad \forall x \ . \tag{12}$$

At the spatial equilibrium within a city, residents must be indifferent across city locations. Equating expression (12) valued at $x = 0$ with the same expression valued at any other distance and simplifying, implies that the sum of commuting costs and land rents is independent of location within a city and equal to land rents at the centre, where no commuting is necessary:

$$T_{it}(x) + z_i R_{it}(x) = z_i R_{it}(0) \ . \tag{13}$$

Let us use $P_{it}(x) \equiv z_i R_{it}(x)$ to denote more succinctly the price of a dwelling at a distance x from the centre of city i at time t. Differentiating equation (13) with respect to x shows that a marginal increase in housing costs must be offset by a marginal decrease in commuting costs to preserve the indifference of residents choosing across locations within the city:

$$\frac{\mathrm{d}P_{it}(x)}{\mathrm{d}x} = -\frac{\mathrm{d}T_{it}(x)}{\mathrm{d}x} \ . \tag{14}$$

Note this is the standard Alonso-Muth condition in the monocentric city model (Alonso, 1964; Muth, 1969), and arguably one of its greatest theoretical insights.[13] In the process of estimating our model's parameters in section 5, we show its empirical validity.

The edge of the city, denoted by $\bar{x}_{it}$, 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 $\underline{R}$: $R_{it}(\bar{x}_{it}) = \underline{R}$. To simplify notation, we set to zero the value of land in the best alternative use: $\underline{R} = 0$.[14] Substituting equation (9) into (13), valued at $x = \bar{x}_{it}$, and using $R_{it}(\bar{x}_{it}) = 0$ and $P_{it}(0) = z_i R_{it}(0)$, we can express the equilibrium price of a dwelling at the city centre as

$$P_{it}(0) = \tau_{it} (\bar{x}_{it})^{\gamma} \ . \tag{15}$$

Combining equations (9), (10), (13), and (15), the bid-rent for land at a distance x from the city centre is:

$$R_{it}(x) = \frac{\tau_t}{z_i} (N_{it})^{\theta} \left(\bar{x}_{it}^{\gamma} - x^{\gamma} \right) \ . \tag{16}$$

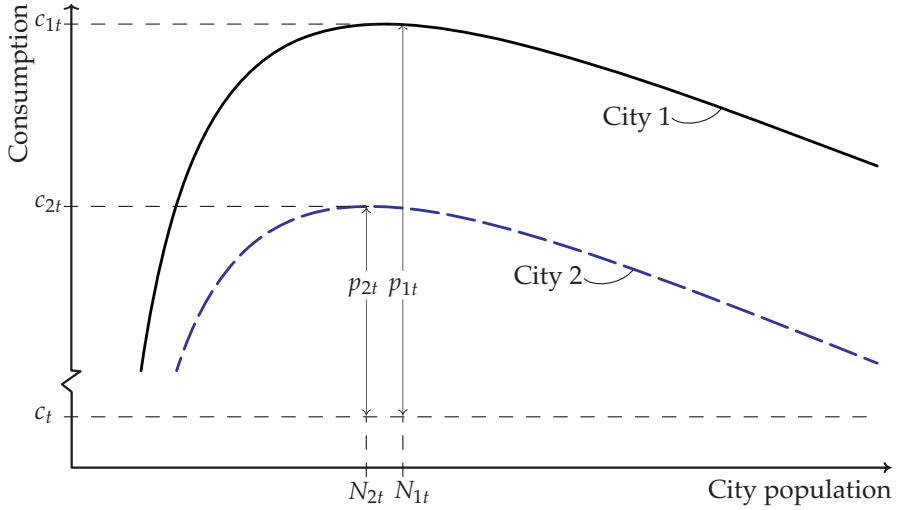
Under our simplifying assumptions of a linear city, fixed-sized housing, and geographical constraints uniformly distributed through the city, we obtain:

$$\bar{x}_{it} = z_i N_{it} \ . \tag{17}$$

[13] By the envelope theorem, the 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).

[14] In the United States, the value of land while in agricultural use is fairly homogeneous and low (Burns, Key, Tulman, Borchers, and Weber, 2018). We provide further details in section 6.

Figure 1: Final consumption as a function of city size



Notes: Consumption for incumbents, c_{it}, plotted as a function of population, N_{it}, using equation (19) and parameter values estimated in section 5 ($\gamma = 0.07$, $\theta = 0.04$, $\sigma = 0.04$, and $\beta = 0.04$), for two cities, 1 and 2, that differ only in production amenities, with $A_{1t} > A_{2t}$ set to give populations of 10 and 9.5 million.

Substituting this last expression into equation (16) and integrating over the extent of the city yields total land rents:

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

Incumbent residents set the population size of their city through planning regulations to maximise their final consumption, $c_{it} = y_{it} - T_{it}(x) - P_{it}(x) + R_{it}/N_{it}$. Using equations (8), (10), (13), (15), (17), and (18) to simplify this expression, we can write incumbents' programme as[15]

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

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 $(N_{it})^{\sigma+\beta}$) 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:

$$N_{it} = \left(\frac{\rho^{\sigma}(\sigma+\beta)(\gamma+1)}{\gamma+\theta} \frac{A_{it}(h_t)^{1+\sigma}}{\tau_t(z_i)^{\gamma}} \right)^{\frac{1}{\gamma+\theta-\sigma-\beta}} . \tag{20}$$

The second-order condition requires $\gamma + \theta - \sigma - \beta > 0$ and positive city sizes $\sigma + \beta > 0$. Below we show both restrictions hold empirically.

Figure 1 illustrates the relationship between final consumption for incumbents and city size for two cities with different production amenities. The concavity of final consumption reflects

[15]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 where endogenous planning regulations keep the marginal resident indifferent across cities.

the tradeoff between agglomeration economies and crowding described in equation (19). For each city, the maximum of its consumption curve corresponds to the population size defined by equation (20). Incumbent residents achieve their maximum consumption for a larger population size in city 1 than in city 2, $N_{1t} > N_{2t}$, because we have assumed a higher level of idiosyncratic productivity in city 1 than in city 2, $A_{1t} > A_{2t}$. These population sizes, N_{1t} and N_{2t}, are optimal from the perspective of incumbent residents. However, residents in smaller and less productive cities would like to join incumbent residents of more productive cities, thereby further increasing their cities' populations, were it not for the permitting cost.

While final consumption for incumbent residents is higher in bigger cities, final consumption for the marginal incoming resident is equated across cities through the cost of permitting at the spatial equilibrium across cities.[16] Equivalently, the sum of consumption for the marginal resident in the marginal populated city and permitting costs in city i equals consumption for incumbents in city i $p_{it} = c_{it} - c_t$. Isolating $\rho^\sigma A_{it}(h_t)^{1+\sigma}$ from equation (20) yields:

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

Substituting equation (21) into the first term on the right-hand side of equation (19) yields c_{it} as a function of N_{it}, z_i, and parameters:

$$c_{it} = \frac{\gamma+\theta-\sigma-\beta}{(\sigma+\beta)(\gamma+1)}\tau_t\,(z_i)^\gamma\,(N_{it})^{\gamma+\theta} \ . \tag{22}$$

Substituting equation (22) into $p_{it} = c_{it} - c_t$, equilibrium permitting costs can be written as

$$p_{it} = \frac{\gamma+\theta-\sigma-\beta}{(\sigma+\beta)(\gamma+1)}\tau_t(z_i)^\gamma(N_{it})^{\gamma+\theta} - c_t \ , \tag{23}$$

which are increasing in the city's population N_{it} and in natural geographical constraints on development z_i. The positive relationship between planning regulations and geographical constraints arises because the same population increase creates greater crowding for incumbents in more geographically constrained cities without improving agglomeration economies. Thus, incumbents react by setting stricter planning regulations to balance the trade-off to suit them. Below, we provide empirical evidence regarding this complementarity.

Permitting costs equate the private, but not the social, returns to the marginal resident across cities. Aggregate consumption would increase by vacating the least productive sites and allocating more residents to the remaining cities.[17] In section 7, we quantitatively explore the consequences of relaxing locally-imposed planning regulations and thus lowering permitting costs.

[16]Permitting costs reflect only the consumption differential between a city and the best alternative for the current generation but do not capitalize the gains for future generations, as we ignore bequest motives. Incorporating these would affect the value of p_{it}, but not the equilibrium city population size given by equation (20). This population size maximizes consumption period by period —a necessary condition to maximize consumption across generations in our context if we introduce bequest motives.

[17]We can think of three alternative micro-foundations for sub-optimally small cities. First, perhaps the nuisance arising from additional housing construction and increases in crowding is experienced with much greater intensity locally while the gains from greater agglomeration economies diffuse through the metropolitan area. Then, to the extent that planning barriers are also more local, they may place 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.

Having derived equilibrium city sizes, the final step to characterise the equilibrium urban system is to determine which sites attract population at any given time. Sites for potential cities are heterogeneous, differing in production amenities and geographical constraints to development. A sufficient statistic of a site's overall attractiveness is the price of a dwelling at the city centre when this site hosts a city of equilibrium size. To see why this is the case, note that from equation (15), with τ_{it} given by equation (10) and $\underline{R} = 0$, this price is $P_{it}(0) = \tau_t (z_i)^{\gamma} (N_{it})^{\gamma+\theta}$. Using this last equation to replace $\tau_t (z_i)^{\gamma} (N_{it})^{\gamma+\theta}$ in equation (22), we can write:

$$c_{it} = \frac{\gamma + \theta - \sigma - \beta}{(\sigma + \beta)(\gamma + 1)} P_{it}(0) \ . \tag{24}$$

Thus, in equilibrium, the final consumption level for a city's incumbent residents is proportional to the price of a dwelling at that city's centre. This implies that, when maximizing c_{it} in the programme of equation (19), incumbent city residents vote for planning regulations that effectively maximize the value of their individual homes, as in Fischel's (2001) 'homevoter hypothesis'. We can think of this as a local "golden rule" of planning regulation, similar to the "golden rule" of public good provision in Flatters, Henderson, and Mieszkowski (1974).

Cities attract residents as long as they offer newcomers a level of final consumption that leaves them no worse than in rural areas. The marginal populated city satisfies two conditions. First, suppose we order potential city sites from most to least attractive (in a sense just discussed). In that case, the marginal city is where incumbent residents must impose no planning restrictions to maximize their own consumption while matching consumption for newcomers to consumption in rural areas. More formally, this is the city for which $c_{it} = c_t = y_{rt}$, where c_{it} and y_{rt} are given by equations (24) and (11) respectively. Second, with all sites at least as attractive as the marginal city populated at the level given by equation (20), their populations and the rural population must add up to the total population at time t, N_t.

Before we put our model to use further, we now provide intuition for its equilibrium by representing the urban system of the United States as seen through the lens of the model.

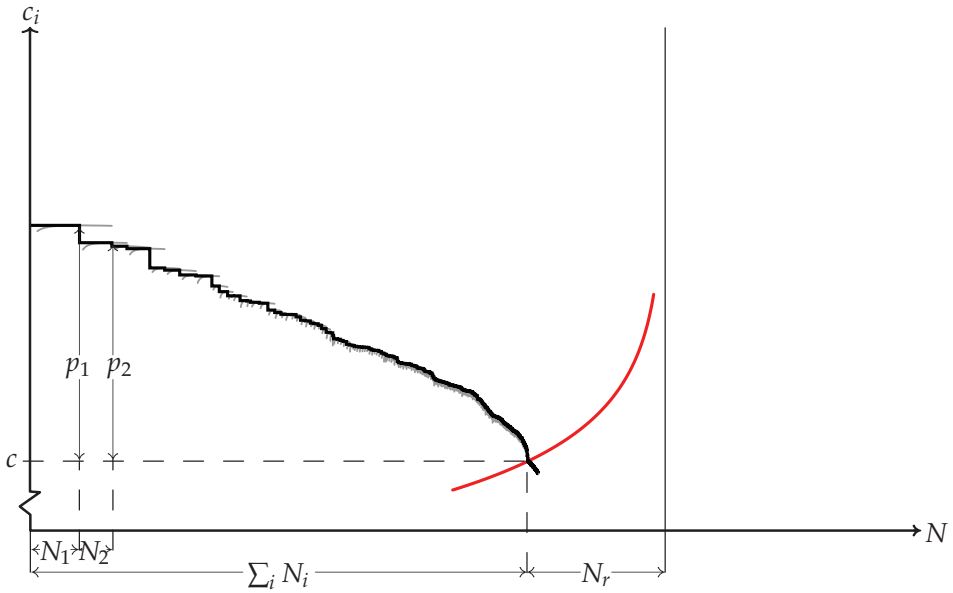
Illustrating the equilibrium with the urban system of the United States

In panel A of figure 2 we represent the urban system of the conterminous United States in 1980. 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).[18] The thick segment on the top left corresponds to New York. Because of its attractiveness, this city's location would be the first to be populated. Incumbent New Yorkers then set a permitting cost to maximise their consumption, which is represented by the thin curve below the thick segment and tangent to it for a population N_1. The permitting cost that achieves this population level for New York is p_1 and it corresponds to the vertical gap between consumption for incumbent New Yorkers and consumption for newcomers everywhere and rural residents, marked as c).

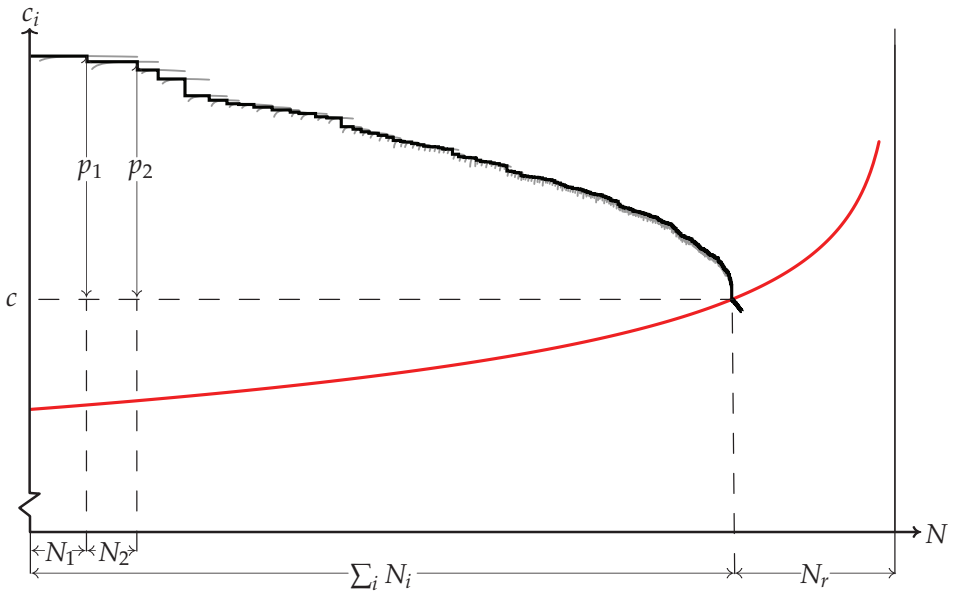
[18] Throughout the paper, we define US cities using 1999 county-based metropolitan area definitions.

Figure 2: Equilibrium allocation of population

Panel A: United States, 1980



Panel B: United States, 2010



Notes: Panels A and B depict the allocation of population across US metropolitan and non-metropolitan areas in 1980 and 2010 as an equilibrium of the model. Subindex t omitted from notation in the figure. The thick horizontal segments represent equilibrium consumption for incumbents in each city, c_i (segment height), and population N_i (segment length). Horizontal axis length is total US population, N. Total urban population $\sum_i N_i$ can be read as the horizontal distance to the left-side axes origin and rural population, $N_r = N - \sum_i N_i$ can be read as the distance to the right-side axes origin, with rural consumption as a function of rural population given by the smooth long curve. Rural consumption and consumption in the marginal populated city are equated at the intersection point marked c (consumption for city newcomers and rural residents). The thin curves tangent to each thick segment plot consumption for incumbents in each city when population differs from its equilibrium level. Incumbents set permitting costs at $p_i = c_i - c$ to achieve the consumption at the maximum of the curve for their city while keeping newcomers indifferent. To draw this figure, we use parameter values estimated or calibrated in section 5 ($\gamma = 0.07$, $\theta = 0.04$, $\sigma = 0.04$, $\beta = 0.04$, and $\lambda = 0.18$), the actual distribution of population in each year, the share of the area within 30 kilometres of the centre of each city that is not geographically constrained (to determine z_i), and the equations detailed in appendix B, changing τ over time to exactly match the growth in average gross domestic product per capita between 1980 and 2010.

The thick segment drawn starting where New York's segment ends corresponds to Los Angeles. 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. Incumbents in Los Angeles set permitting cost at p_2 to achieve this population. We can then continue this process for every metropolitan area.

The horizontal axis also measures the population outside of metropolitan areas, but in this case represented from right to left, similarly to the diagram used to analyze diagrammatically the specific-factors model in international trade (Mussa, 1974). The smooth curve extending along the entire length of the graph represents rural consumption as a function of the rural population, as given by equation (11).

We set the length of the horizontal axis to match the total population of the United States. The point where the step-wise thick schedule for the urban sector and the smooth curve for the rural sector intersect defines the marginal populated city. On the horizontal axis, we can read the total urban population as the distance between the left origin and the intersection point. Rural population is the horizontal distance between the right origin and the intersection point. On the vertical axis, this intersection point indicates the level of final consumption for rural residents as well as for new residents in every city.[19]

Panel B of figure 2 represents the urban system of the conterminous United States in 2010. The increased distance between the two vertical axes in panel B relative to panel A represents the growth in total population from 225 million in 1980 to 307 million in 2010. The population outside metropolitan areas grows somewhat in absolute terms but falls as a share of the total population as urbanization advances.

Between panels A and B, curves in the sequence representing the urban sector move up vertically and expand horizontally with individual city growth. This growth stems partly from human capital accumulation and partly from the accumulation of idiosyncratic shocks to each city's level of production amenities. Incumbent residents adapt the cost of permitting to let cities expand up to the new, larger, locally-optimal level.

However, the shocks are heterogeneous across cities, so their evolution is different and their relative positions change. While the four most attractive cities, New York, Los Angeles, San Francisco, and Chicago remained unchanged between 1980 and 2010, Washington DC, Boston, and Miami overtook Detroit.

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. We first show that our model generates a growth process for cities that satisfies Gibrat's law where the rate of population growth is independent from initial population size. Then, we prove that, under standard additional conditions, Gibrat's law results in a steady-state city-size

[19]We represent unpopulated potential sites for additional cities to the right of the marginal city, following the strategy explained in appendix B. The alternative to replacing our rural sector with a fixed outside option mentioned above would correspond to replacing the curve for the rural sector in figure 2 with a horizontal line. Note that since agglomeration economies amplify the effects of productivity growth in the urban sector but not in the rural sector, this creates a tendency of the economy to urbanise with overall productivity growth. Our modelling of the rural sector ensures that it never disappears even if its relative size evolves over time.

distribution that approximates Zipf's law, i.e. steady-state city sizes follow a Pareto distribution with shape parameter approaching 1. This is a desirable feature for model since a vast literature argues that Zipf's law is a good empirical approximation to the city-size distribution in the United States and other countries (Gabaix and Ioannides, 2004; Duranton and Puga, 2014). Importantly, we can obtain these results in a model where cities arise endogenously, are subject to agglomeration economies and crowding costs, and experience population growth driven by both idiosyncratic productivity shocks and systematic growth in human capital which, in turn, is systematically related to city population.

To compute log population change between two consecutive periods in city i, if a city exists at that location, we take the log of equation (20) 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:

$$\Delta \ln(N_{it}) = \frac{1}{\gamma + \theta - \sigma - \beta} \left[\Delta \ln(A_{it}) + (1+\sigma) \Delta \ln(h_{it}) - \Delta \ln(\tau_t) \right] . \tag{25}$$

A first component of city population growth arises from the evolution of idiosyncratic productivity at each location. Taking logs and time differencing the evolution of production amenities through multiplicative shocks, $A_{it} = g_{it} A_{it-1}$, implies $\Delta \ln(A_{it}) = \ln(g_{it})$. 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 can produce more. In addition, there is a human capital externality which expands output per worker further —hence the factor $1+\sigma$ multiplying $\Delta \ln(h_{it})$ in equation (25). As already discussed above, our model features 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.[20] Let us assume that this evolves at some constant rate, reflecting, for instance, changes in commuting technology or in the value of travel time: $\Delta \ln(\tau_t) = \Delta \ln(\tau)$.

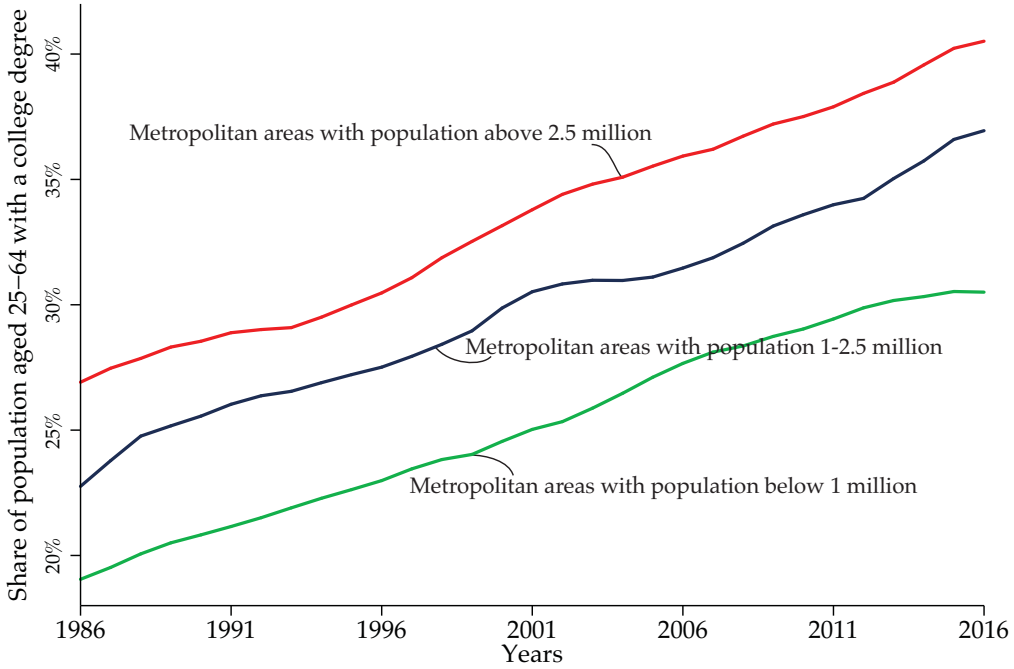
We can now rewrite equation (25) describing the population growth of a city at location i between time $t-1$ and time t as

$$\Delta \ln(N_{it}) = \frac{1}{\gamma + \theta - \sigma - \beta} \left[\ln(g_{it}) + (1+\sigma) \Delta \ln(h) - \Delta \ln(\tau) \right] . \tag{26}$$

The population growth process of (26) satisfies Gibrat's law (after Gibrat, 1931): since g_{it} is identically and independently distributed for every city, so is the growth rate of urban population on the right-hand side of equation (26). This growth rate has a systematic component arising from human capital accumulation and from the evolution of commuting that is common to all cities, captured by $(1+\sigma)\Delta \ln(h) - \Delta \ln(\tau)$. It also has an idiosyncratic random component arising from local productivity shocks, captured by $\ln(g_{it})$. Thus, cities experience parallel population growth in expectation but are subject to idiosyncratic ups and downs relative to this common trend, consistent with the empirical evidence (Henderson, 2005).

[20] We restrict time-varying heterogeneity across cities to their production amenities to keep the exposition simple. While we also allow for heterogeneous geographical barriers to development, these are time-invariant so only have a level effect. We could nonetheless readily extend our model to idiosyncratic shocks in, say, transport infrastructure by enriching equation (10) and explicitly considering shocks to roadway expansion.

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



Notes: Data from the Current Population Survey Annual Social and Economic Supplement. Each metropolitan area is assigned to one of the three curves based on its 2010 population.

Satisfying Gibrat's law in our context is far from obvious. Relative to Gabaix (1999), a first complication is that our cities feature agglomeration economies and crowding costs, which could potentially magnify or dampen productivity shocks. We obtain Gibrat's law nevertheless because, in equilibrium, all cities operate at the point where agglomeration economies and crowding costs balance out (see Duranton, 2007; Rossi-Hansberg and Wright, 2007).

A second complication is that, in addition to random determinants of urban growth, our model also features systematic determinants that depend on city size. More concretely, bigger cities enjoy greater average human capital per worker and human capital is a key driver of urban growth. Nonetheless, individual city population growth rates will remain independent of city sizes because the growth rate of human capital is the same across cities of all sizes. While agglomeration effects magnify the effects of human capital growth, they do so equally across all cities, as made clear in equation (26).

Hence, an important property of our model is that, while human capital levels can differ across cities of different size, human capital growth rates should not vary systematically. How realistic is this property? Figure 3 shows that it accords well with the evidence for the United States. This figure plots the evolution of the share of the population aged 25–64 who hold a college degree in metropolitan areas of different sizes over the 1986–2016 period in the United States. For each of these three decades, there is always 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 about 1.6 between 1986 and 2016 in all three city-size classes, keeping the relative magnitude of their college shares stable.[21] 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 0.10.[22]

To obtain approximately Zipf's law from Gibrat's law, two additional conditions are required. First, there must be some mechanism that prevent cities from shrinking indefinitely (Champernowne, 1953; Gabaix, 2009). Like Gabaix (1999), we assume a reflexive lower bound on city sizes such that, when a city reaches this minimum size, further shocks can only bring size up and not further down.[23]

Second, we must be able to normalise city sizes so that their normalised mean 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), we normalise city sizes relative to their average by defining $\tilde{N}_{it} \equiv \frac{N_{it}}{\bar{N}_t}$, where $\bar{N}_t$ denotes the average population at time t of all potential cities and the normalised mean size of all potential cities is equal to 1. We then assume that the reflexive lower bound on normalised sizes, η, is constant.

Champernowne's (1953) insight is that in steady state:

$$F(\tilde{N}) = \begin{cases} 1 - \left(\frac{\tilde{N}}{\eta}\right)^{-\zeta} & \text{if } \tilde{N} \geqslant \eta \,, \\ 0 & \text{if } \tilde{N} < \eta \,, \end{cases} \tag{27}$$

where $F(\tilde{N})$ denotes the share of potential cities with a normalised population size $\tilde{N}$ or lower. The probability density function corresponding to this cumulative distribution function is then $f(\tilde{N}) = \frac{\mathrm{d}F(\tilde{N})}{\mathrm{d}\tilde{N}} = \eta^{\zeta}\zeta\tilde{N}^{-\zeta-1}$. The mean normalised size of all potential cities can be calculated as

$$\int_{\eta}^{+\infty} \tilde{N} f(\tilde{N}) \mathrm{d}\tilde{N} = \frac{\eta^{\zeta}\zeta}{1-\zeta}\left[\tilde{N}^{1-\zeta}\right]_{\eta}^{+\infty} = -\frac{\eta\zeta}{1-\zeta} \,, \tag{28}$$

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

$$\zeta = \frac{1}{1-\eta} \,. \tag{29}$$

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.

Unlike in Gabaix (1999), the set of sites that host a city at any point is endogenously determined in our model. Thus, the relevant distribution is that for the absolute sizes of actual cities rather than the normalised sizes of all potential cities. To show that this distribution also converges

[21] 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.

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

[23] In practice, such a reflexive force can arise from the durability of housing (Glaeser and Gyourko, 2005). See Saichev, Malevergne, and Sornette (2009) for alternatives.

over time to a Pareto distribution with shape parameter $\frac{1}{1-\eta}$, we must take two additional steps. First, we must recover absolute city sizes as $N = \bar{N}_t \tilde{N}$. Multiplying a variable distributed Pareto by a constant results in a transformed variable that follows a Pareto distribution with the same shape parameter. Second, we must restrict ourselves to city sites that are actually populated. This involves left-truncating the distribution of potential city sites. Left-truncating a Pareto distribution also leaves its shape parameter unchanged.[24]

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. It is not easy to distinguish empirically between a Pareto distribution and a truncated log-normal distribution for city sizes. What matters for us is that they are very similar and good approximations of reality.

5. Empirical estimates of the model's key parameters

We now turn to the empirical estimation of the key parameters of the model regarding urban costs and agglomeration benefits. Details regarding data sources and variable definitions are provided in appendix B.

Population elasticity of urban costs

Past research 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.[25] Our theoretical framework suggests three alternative approaches to estimate γ, which we now implement. The parameter γ first appears in the commuting cost equation (9) as the elasticity of a resident's commute with respect to the distance x between her dwelling and the city centre. Taking the natural logarithm (hereafter log), of this equation and differentiating with respect to $\ln(x)$ yields:

$$\frac{\mathrm{d}\ln(T_{it}(x))}{\mathrm{d}\ln(x)} = \gamma \,. \tag{30}$$

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 (NHTS), we estimate a regression at the household level of the log of vehicle-kilometres travelled by members of household j, T_i^j, on the log of the distance between

[24] The scale parameter, however, is no longer given by the reflexive lower bound but by the (higher) truncation point associated with the marginal viable city.

[25] As we discuss in Duranton and Puga (2015), there is a large literature estimating various urban gradients associated with population and housing but we know of no attempt to link findings about urban gradients to deeper structural parameters. An exception is Combes, Duranton, and Gobillon (2019), but their approach to estimating urban costs does not provide a direct equivalent of our model parameters. See also Couture, Duranton, and Turner (2018) for estimates of how congestion varies with city population.

the household's residence and the centre of their metropolitan area i, x_i^j:

$$\ln(T_i^j) = \gamma \ln(x_i^j) + a_i + \boldsymbol{X^j b} + \epsilon_i^j \,, \tag{31}$$

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 0.0728.

Once we solve for a spatial equilibrium within each city, the Alonso-Muth condition of equation (14) implies that, within each city, variation in commuting costs should be offset by variation in housing costs. It follows from this equation that $\frac{\mathrm{d}[P_{it}(0)-P_{it}(x)]}{\mathrm{d}x} = \frac{\mathrm{d}T_{it}(x)}{\mathrm{d}x}$. Then, equation (13) can be rewritten as $[P_{it}(0) - P_{it}(x)] = T_{it}(x)$. Dividing the previous equation by this one, multiplying both sides by x, and using natural logs to simplify leads to:

$$\frac{\mathrm{d}\ln[P_{it}(0) - P_{it}(x)]}{\mathrm{d}\ln(x)} = \frac{\mathrm{d}\ln[T_{it}(x)]}{\mathrm{d}\ln(x)} = \gamma \,. \tag{32}$$

The intuition is both straightforward and of fundamental importance: in equilibrium, indifference across locations within a city 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 (Duranton and Puga, 2015), to the best of our knowledge it has not been tested before.

Based on equation (32), a second approach to estimate γ is to exploit 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 log of the difference between the median rent in the most expensive block group in the city, $\bar{P}_i$, and the median rent in block group j, P_i^j, on the log of the distance between block group j and the centre of its metropolitan area i, x_i^j:

$$\ln(\bar{P}_i - P_i^j) = \gamma \ln(x_i^j) + a_i + \boldsymbol{X^j b} + \epsilon_i^j \,, \tag{33}$$

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 (33), $\ln(\bar{P}_i - P_i^j)$, is our empirical counterpart to $\ln(P_{it}(0) - P_{it}(x))$ in the equilibrium equation (32). Results are shown in column (2) of table 1 and they imply a value for γ of 0.0769, very similar to the 0.0728 value we obtained using transport data. Beyond providing reassurance regarding the estimated value of γ, we regard this as important empirical evidence in support of the Alonso-Muth trade-off.

Our theoretical framework also suggests a third approach to estimate γ, relying on cross-city variation and the following relationship:

$$P_{it}(0) + p_{it} + c_t = P_{it}(0) + c_{it} = \left(1 + \frac{\gamma + \theta - \sigma - \beta}{(\sigma + \beta)(\gamma + 1)}\right) \tau_{it} (\bar{x}_{it})^{\gamma} \,. \tag{34}$$

In this expression, the left-hand side is total expenditure for newcomers, including urban costs (housing and commuting combined) summarised by $P_{it}(0)$, permitting costs p_{it}, and final consumption c_t. The difference between consumption for newcomers and consumption for incumbents is permitting costs: $p_{it} + c_t = c_{it}$. In turn, the "golden rule" of planning regulation of

Table 1: Estimation of urban costs (model parameters γ and θ)

	(1)	(2)	(3)	(4)	(5)
Dependent variable:	Log household miles travelled	Log block-group differential median house price	Log augmented city-centre house price	Log city travel speed (NHTS)	Log city travel speed (Google)
Log distance to city centre	0.0728*** (0.0100)	0.0769*** (0.0145)			
Log city-periphery distance			0.0726*** (0.0248)		
Log city population				-0.0388*** (0.0033)	-0.0386*** (0.0034)
Log city travel speed			-1.3444*** (0.2100)		
City indicators	Yes	Yes			
Controls	Block-group & household	Block-group & dwelling	For house price and speed variable construction	For speed variable construction	
Observations	107,492 households	127,518 block-groups	182 cities	182 cities	180 cities
R^2	0.319	0.205	0.361	0.438	0.416

Notes: In column (1) units of observation are households and the dependent variable is the (natural) log of the household's annual miles travelled; the estimate of γ is the coefficient on the log of distance to the city centre.

In column (2) units of observation are city block-groups and the dependent variable is the log of the difference between the median housing rental price in the most expensive block-group in the city and the median price in the block group under consideration; the estimate of γ is the coefficient on the log of distance to the city centre.

Columns (1) and (2) include city indicators and, 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, an indicator for waterfront location, and ruggedness. Column (1) also controls for the following household characteristics: the log of household size, the log 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: the percentage dwellings in block group by type of structure, by number of bedrooms, and by construction decade.

In column (3) units of observation are cities and the dependent variable is the log of the sum of house prices in the centre of the city and a reference measure of consumption for marginal residents common across cities; the estimate of γ is the coefficient on the log of the median distance to the city centre.

The city-centre house prices in column (3) are estimated in a previous step by predicting the value of a national-reference house at the centre of each city for city-average block-group characteristics based on a regression of the log of the block-group median house rental price on a third-degree polynomial of distance to the city centre, and the same dwelling characteristics and block-group characteristics as in column (2).

The log of city travel speed in column (3) is estimated in a previous step 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 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.

In columns (4) and (5) units of observation are cities and the dependent variable is the log of travel speed, estimated as in column (3) for column (4) and the log of travel speed computed by Akbar, Couture, Duranton, and Storeygard (2022) using Google Maps data for column (5); the estimates of θ are (minus) the coefficients on the log of city population in columns (4) and (5).

All regressions include a constant term and standard errors are clustered at the city level in columns (1) and (2). ***, **, and * indicate significance at the 1, 5, and 10 percent levels. The R^2 reported in columns (1) and (2) is within city.

equation (24) requires consumption for incumbents, c_{it}, to be proportional to differential house prices at the centre, $P_{it}(0)$. Thus, the combined left-hand side of equation (34) is proportional to $P_{it}(0)$. Then, the spatial equilibrium within cities of equation (15) requires $P_{it}(0)$ to equal commuting costs at the edge, which are equal to $\tau_{it}\,(\bar{x}_{it})^{\gamma}$ and thus increase with the spatial extent of the city with elasticity γ.

Empirically, equation (34) maps into a regression of the log of the sum of total house price at the center, including permitting costs, and consumption in the marginal populated city on the log of the spatial extent of the city and the log of commuting cost per unit of distance:

$$\ln[\hat{P}_i + c(\gamma)] = a + \gamma \ln(\bar{x}_i) + \psi \ln(\hat{\tau}_i) + \epsilon_i \,. \tag{35}$$

While Appendix B provides full details of how we implement this regression, the following comments are in order. First, for the dependent variable, we note that house prices at the center of some cities can be unusual due to specific dwelling or neighbourhood characteristics. Rather than actual prices, we measure total house prices at the center of cities using a prediction obtained from a regression of all house prices on a polynomial of distance to the center and local dwelling and neighbourhood characteristics. Second, the dependent variable also contains a value of marginal consumption that is unknown. Equation (24) indicates that this value should be proportional to the price of housing at the centre of the cheapest city with a proportionality constant which depends on our key parameter of interest, γ. To go round this problem, we estimate equation (35) iteratively and update the value of γ to compute $c(\gamma)$ until convergence. Third, we do not observe the cost of commuting per unit of distance but expect its variation across cities to be negatively related to travel speed. We obtain our measure of travel speed from a regression of the speed for individual trips by private car on city indicators, while controlling for driver and trip characteristics. Fourth, in the model, the spatial extent of the city $\bar{x}$ is given by the distance between the centre and the periphery of the city. Since metropolitan area definitions are county-based, instead of defining the city periphery relying on county borders, we implement a consistent definition across cities. We take the city periphery to be at the longest distance from the city centre within the metropolitan area boundaries that is within the 95th percentile of dwelling distances and has at least 500 dwelling units per square mile in the block group.

Our estimation results are shown in column (3) of table 1 and they imply a value for γ of 0.0726, not statistically different from the 0.0728 and 0.0769 values that we obtained using, respectively, within-city variation in distance travelled and housing prices.

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 parameter appears in equation (10) which, after taking logs, maps directly into the following regression:

$$-\ln(\tau_i) = a - \theta \ln(N_i) + \epsilon_i \,. \tag{36}$$

In the process of implementing our third approach to estimate γ, we have already obtained an estimate of travel speed in each city, $-\ln(\hat{\tau}_i)$. If we use this estimate in place of $-\ln(\tau_i)$ in equation (36) and estimate this regression, we obtain an estimated value of θ of 0.0388, as shown in column (4) of table 1.

To validate the self-reported travel duration estimates of NHTS respondents, we turn to data from Akbar, Couture, Duranton, and Storeygard (2022), who query Google Maps over an extended time period for the estimated travel duration of trips taken by NHTS respondents. This results in a very similar estimated value of θ of 0.0386, as shown in column (5) of table 1. 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 (8). As a result, two parameters, σ and β, must be estimated.

While we could attempt to estimate equation (8) 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 workers 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:

$$y_{it}^{j} = a_i + a_j + a_t + \sum_i b_i e_{it}^{j} + \boldsymbol{X_t^j b} + \epsilon_{it}^{j} \,, \tag{37}$$

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, $\boldsymbol{X_t^j}$ 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.

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

$$\hat{a}_i = \sigma N_i + \epsilon_i \,. \tag{38}$$

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 $\bar{e}$:

$$\hat{a}_i + \hat{b}_i \bar{e} = (\sigma + \beta) N_i + \epsilon_i \,. \tag{39}$$

This gives us a value for $\sigma + \beta$ and, subtracting from this the value of σ estimated from (38), yields an estimate for β.

If we were just interested in σ, we could also estimate this relationship in a single step by replacing a_i in equation (37) with the right-hand-side of equation (38).[26] Column (1) in table 2

[26] 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.

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

	(1)	(2)	(3)	(4)	(5)
Estimation method:	OLS	TSLS		OLS	
Dependent variable:	Log earnings			Initial premium (city indicator coefficients column (3))	Medium-term premium (initial + 8.4 years local experience)
Log city population	0.0443*** (0.0053)	0.0421*** (0.0057)		0.0452*** (0.0045)	0.0770*** (0.0063)
City indicators			Yes		
Worker fixed-effects	Yes	Yes	Yes		
Experience in cities $\geqslant$ 5 million	0.0194*** (0.0067)	0.0196*** (0.0067)	0.0192*** (0.0069)		
Experience in cities $\geqslant$ 5 million $\times$ exp.	-0.0004** (0.0002)	-0.0004** (0.0002)	-0.0004** (0.0002)		
Experience in cities 2-5 million	0.0068* (0.0037)	0.0070* (0.0037)	0.0075** (0.0036)		
Experience in cities 2-5 million $\times$ exp.	-0.0002 (0.0001)	-0.0002 (0.0001)	-0.0002 (0.0001)		
Experience	0.0633*** (0.0045)	0.0633*** (0.0045)	0.0629*** (0.0046)		
Experience^2	-0.0007*** (0.0001)	-0.0007*** (0.0001)	-0.0007*** (0.0001)		
Observations	50,393	50,393	50,393	63	63
R^2	0.3416		0.3440	0.4870	0.6692

Notes: In columns (1), (2), and (3) units of observation are individual worker-year pairs (annually 1980–1994 and biannually 1996–2012) and the dependent variable is the (natural) log of earnings. Columns (1), (2), and (3) include firm tenure and its square, and indicators for two-digit sector, occupation, and year; worker values of experience expressed in years.

The estimate of γ in columns (1) and (2) is the coefficient on the log of city population. In the TSLS estimation of column (2) we instrument the (natural) log of city population in 2010 with the the percentage of the area within 30-kilometres of the city centre that has slopes greater than 15% and the percentage covered by wetlands, the arsinh of city population in 1850 and 1920, the arsinh of distance to Eastern Seaboard, and heating degree days. First-stage results reported in Appendix C.

In column (3), instead of log city population, we include city indicators, aggregating 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. The estimated coefficients on these city indicators in column (3) become the dependent variable in column (4). The estimate of γ in column (4) is the coefficient on the log of city population.

In column (5), the dependent variable is the city medium-term earnings premium, calculated as the sum of the estimated coefficients on city indicators in column (3) (capturing the earnings premium a worker obtains immediately upon getting a job in that city relative to the smallest city) and the additional value of workers' experience when accumulated in that city according to the estimated coefficients on experience in column (3) applied to the average experience (8.4 years) that workers in the sample accumulate in one city (capturing the additional earnings premium a worker gets over time by accumulating experience locally instead of in the smallest city). The estimate of $\gamma + \beta$ in column (5) is the coefficient on the log of city population.

All regressions include a constant term. Coefficients are reported with robust standard errors in parenthesis, which are clustered by worker and city in columns (1)-(3). ***, **, and * indicate significance at the 1, 5, and 10 percent levels. The R^2 reported in columns (1) and (3) is within workers.

shows results for this one-step estimation. The table uses panel data from the US 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 (8) of our model, while treating A_{it} as exogenous. However, equation (20) points to a systematic relationship between A_{it} and N_{it}. This suggests instrumenting for city size.[27] Equation (20) also indicates as potential instruments for N_{it} determinants of z_i that do not affect A_{it} directly. We use as instruments the percentage of the area in a 30-kilometre radius around the city centre that has slopes greater than 15% and the percentage covered by wetlands.[28]

We also incorporate other common instruments for city population in the estimation of urban agglomeration economies. In the spirit of Ciccone and Hall (1996), we use the inverse hyperbolic sine of the city's population in 1850 and 1920 and of distance to the Eastern Seaboard.[29] 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. Inspired by the redistribution of population towards areas with nice weather documented by Rappaport (2007), our final instrument is heating degree days (a measure of the coldness of climate).

In the first stage of our instrumental variable estimation (reported in Appendix C), all the instruments are significant, jointly and individually (except for distance to the Eastern Seaboard). 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 TSLS 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.[30]

In column (3), we turn to the first step of the two-step estimation, corresponding to equation (37). Relative to the one-step procedure of columns (1) and (2), this replaces the log of city size

[27]We do not instrument when estimating γ and θ in table 1 because the specifications we use do not leave a residual containing an unobserved variable correlated with our explanatory variables. Instead we think of the residuals in our various estimations of γ and θ as the result of imperfect measurement.

[28]We do not use our full set of geographical constraints to urban expansion as instruments because the ocean, in addition to acting as a barrier, can also facilitate transportation thus violating the exclusion restriction, while the probability of land being protected can be endogenous to the dynamics of city population

[29]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 $\text{arsinh}(0) = 0$.

[30]See Combes and Gobillon (2015) for a discussion of instrumentation and alternative approaches to deal with endogeneity in this context.

with city fixed effects, which are then regressed on the log of city size in the second step shown in column (4).[31] 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). Based on these empirical results, we will use $\sigma = 0.04$ for our quantitative analysis below.

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 (39) 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. Rounding this value to 0.08, and given our estimate of $\sigma = 0.04$, we use $\beta = 0.04$ for our quantitative analysis.

Population elasticity of rural income

The final parameter of our model is λ, which appears in equation (11) and corresponds to the population elasticity of rural income. This plays a relatively minor role for our quantification of the aggregate output effects of cities.[32] As already noted when introducing this equation, it is natural 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.

6. Planning regulations and new constructions in US cities

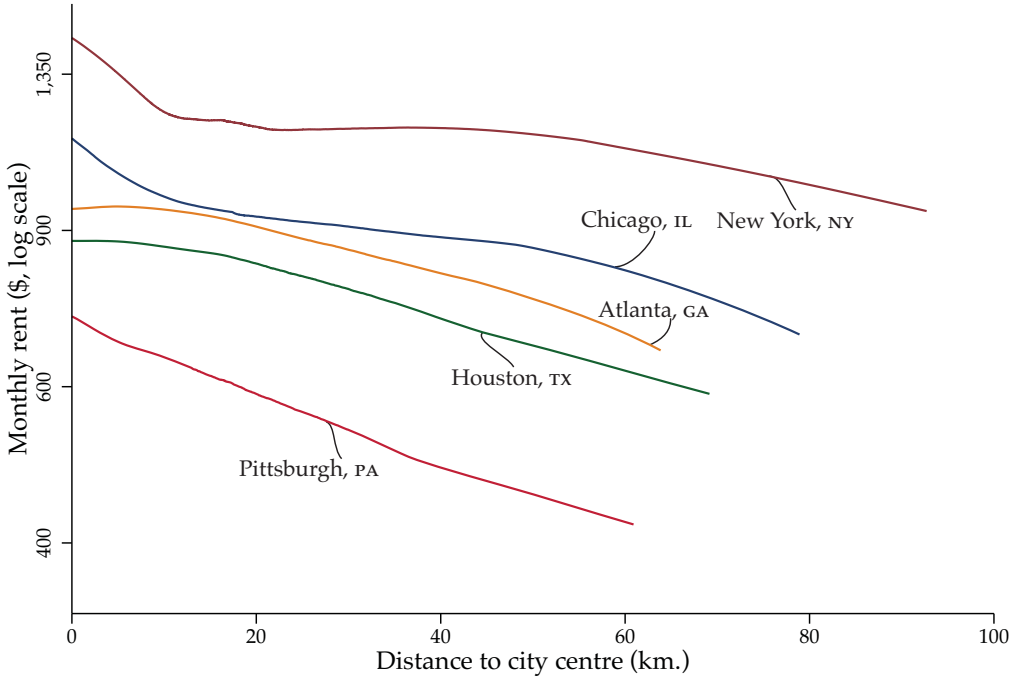
We now provide further empirical evidence about key predictions of our model. Our model shares many features with the standard monocentric city model going back to Alonso (1964) and Muth (1969) and with models of urban systems building on Henderson (1974). Within each city, there is gradient of house prices decreasing in distance to the centre to offset higher commuting costs (equation 16). Equilibrium city sizes result from a tradeoff between agglomeration economies and crowding costs (equation 19), and are also increasing in local productivity, human capital, weaker geographical constraints, and travel speed (equation 20). More populated cities feature higher house prices at the centre, all else equal (equations 15 and 17) and higher earnings (equation 8).

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. When a city experiences a positive shock that attracts new residents, a competitive construction industry supplies new housing at prices that equal replacement costs. These replacement costs are the price of land in

[31] 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.

[32] 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, consumption must be equated between the marginal city and rural areas, this barely affects aggregate output changes.

Figure 4: House price gradients in selected cities



Notes: Monthly rent of a dwelling with average national characteristics in a neighbourhood with average city characteristics as a function of distance to the centre in each city. Estimated with a semilinear regression at the block-group level for each city using 2008–2012 American Community Survey data and Yatchew's (1998) difference estimator. The dependent variable is the median contract rent in the block group. The linear component includes the same dwelling and neighbourhood controls as column (2) of table 1 while distance to the city centre is treated nonparametrically.

the best alternative use (generally, agriculture) plus constructions costs, which these models treat as common across cities and empirically show little spatial variation (as we document below).

Instead, in our framework, 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 to the point where the additional crowding costs imposed on them by the marginal migrant offset additional agglomeration benefits they bring. From the point of view of the marginal migrant, the higher earnings of more populated cities must be enough to compensate not just for a longer commute, as in standard monocentric models, but also for the higher cost of permitting (equation 23).

This key difference leads to testable implications from our framework that do not hold in the standard framework. Through the political-economy mechanism that determines the number and sizes of cities in section 3, planning regulations should be more stringent in more populated cities and the resulting permitting costs should be higher following equation (23). In turn, higher permitting costs imply higher house prices in the periphery of more populated cities.

To illustrate the empirical reality which underlies our data, figure 4 provides a flexible representation of housing price gradients for five US cities. From highest to lowest population, these are New York, Chicago, Atlanta, Houston, and Pittsburgh. For housing units of comparable characteristics, each curve gives their rental price as a function of distance to the centre of each

city.[33] The figure illustrates the empirical relevance of several features that our model shares with standard urban models. There is gradient of house prices within each city 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 in the periphery of New York are much higher than those in the periphery of Chicago, which themselves are higher than in the periphery of Atlanta, and so forth. Put differently, more populated cities tend to extend to a larger distance from their centre but not to the point where housing prices at their periphery are equalised.

While figure 4 provides an illustration for only a few cities, panel A of figure 5 systematically plots periphery housing prices against the 2010 population of all US metropolitan areas for which we can perform the same calculation (see appendix B for details). As predicted by our model, we observe a clear positive relationship between city population and housing prices in the periphery.

In our framework, the initial cause of higher periphery housing prices is stricter planning regulations in more populated cities. Panel B of figure 5 provides direct evidence for this relationship. We measure the strictness of planning regulations using the Wharton Residential Land Use Regulatory Index of Gyourko, Saiz, and Summers (2008) and Gyourko, Hartley, and Krimmel (2021), interpolating 2006 and 2018 values to 2010. Plotting this against the 2010 population of US metropolitan areas, we can observe that planning regulations are more stringent in larger cities.

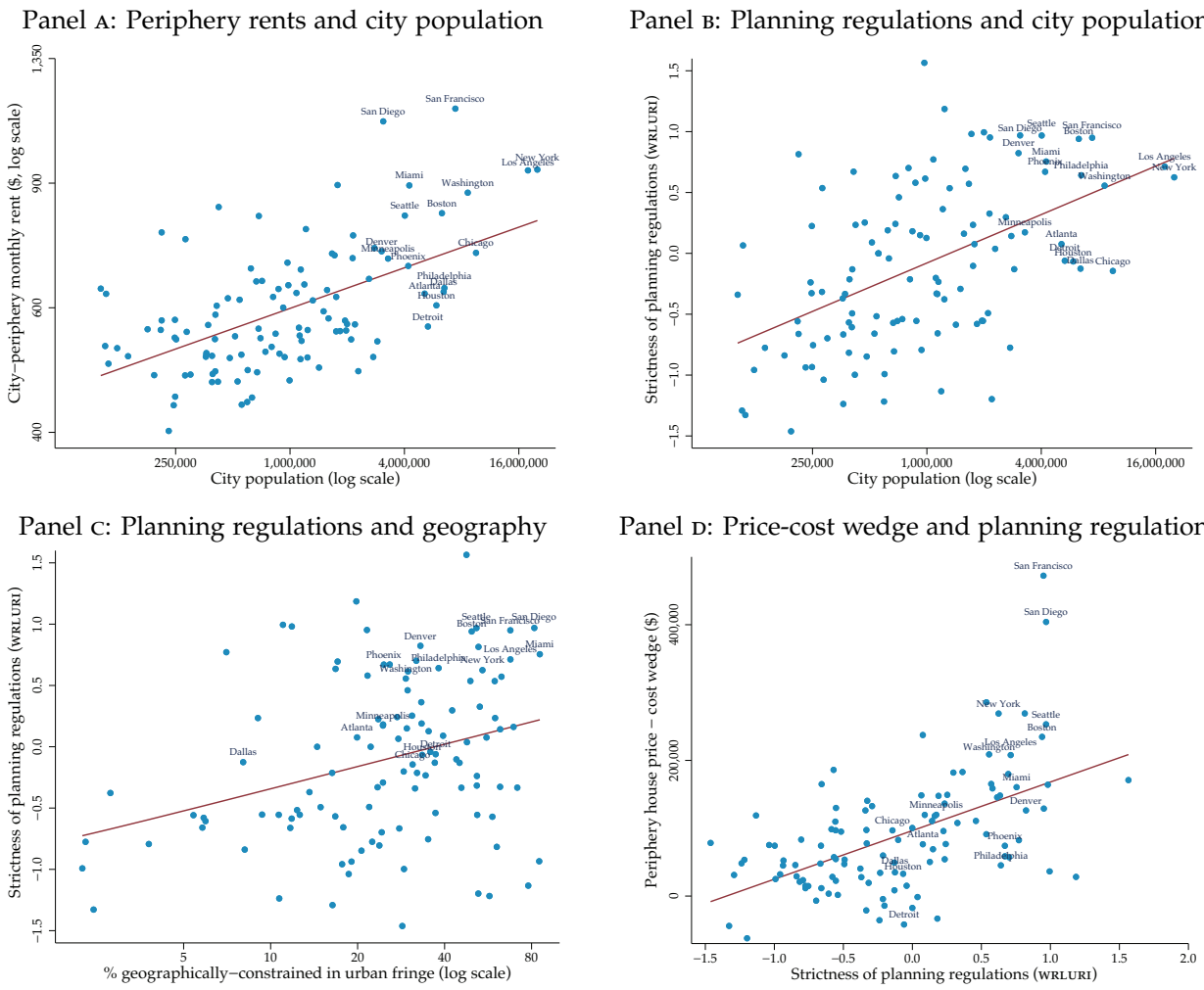
In the model, the strictness of planning regulations depends positively not just on the city's population but also on geographical constraints on the city's ability to expand (equation 23). This complementarity between regulatory and natural constraints occurs because, in more geographically-constrained cities, the same population increase creates greater crowding for incumbents without improving agglomeration economies. As a result incumbent residents impose tighter planning regulation in more geographically-constrained cities. In panel C of figure 5, we plot the strictness of planning regulations against geographical constraints to urban expansion (as we did against population in panel B).[34] The plot in this panel supports the notion of a complementarity between natural and regulatory constraints on urban expansion. Note that we only show the raw relationships with both variables to keep our illustrations simple.[35]

[33] We let the distance vary between zero and the periphery of each city where, as in table 1, we take the city periphery to be at the longest distance from the city centre within the metropolitan area boundaries that is within the 95th percentile of dwelling distances and has at least 500 dwelling units per square mile in the block group.

[34] We measure geographical constraints to urban expansion using the percentage of the area at the fringe of the city covered by slopes steeper than 15%, water, wetlands, or land permanently protected from land cover conversion with a mandate to conserve its natural state. We use the term urban fringe for the area where the city would likely expand next and define this as the area within 20km of land developed at medium or high intensity in 2011 that is undeveloped or developed at low intensity, based on land cover data from the 2011 slice of the 2019 version of the National Land Cover Database.

[35] Residual-plus-component plots isolating the effect on planning regulations of either population or geographical constraints, while controlling for the other variable, yield similar positive and statistically-significant relationships as the raw uni-variate plots of figure 5.

Figure 5: Planning regulations, periphery prices, and new construction in the United States



Notes: City population corresponds to the 2010 Census. The same 112 metropolitan areas for which all data is available appear in the four panels and all metro areas with population above 3 million are labelled.

City-periphery monthly rent is the monthly rent of a dwelling with average national characteristics in a neighbourhood with average city characteristics located at the city periphery. Estimates based on a regression of the log of the median contract rent in the block group on a third-degree polynomial of distance to the city centre and the same dwelling and neighbourhood controls as column (2) of table 1 using 2008–2012 American Community Survey (ACS) data. City periphery is defined the longest distance from the city centre that is within the 95th percentile of dwelling distances and has at least 500 dwelling units per square mile in the block group.

Strictness of planning regulations is measured using the Wharton Residential Land Use Regulatory Index of Gyourko, Saiz, and Summers (2008) and Gyourko, Hartley, and Krimmel (2021), interpolating 2006 and 2018 values to 2010.

Percentage of geographically-constrained land is the percentage of the area in the urban fringe covered by slopes steeper than 15%, water, wetlands, or land permanently protected from land cover conversion with a mandate to conserve its natural state. The urban fringe is defined as the area within 20km of land developed at medium or high intensity in 2011 that is undeveloped or developed at low intensity, based on land cover data from the 2011 slice of the 2019 version of the National Land Cover Database.

Periphery house price - cost wedge is the difference between the value of a house and its replacement cost in the periphery of the city. The house value corresponds to a four-bedroom single-family detached house built 2000–2009 in a neighbourhood with average city characteristics located at the city periphery. This is estimated based on the same regression as that for city-periphery monthly rent, but with median house value instead of the median contract rent in the block group as dependent variable. The replacement costs is the sum of city-specific construction costs for an economy-quality single-family detached house of 2000 square feet (using RSMeans data for 2010 obtained from Glaeser and Gyourko, 2018) and the price of a quarter-acre vacant plot of land used for agriculture at the urban fringe (using land value data from Nolte, 2020 and land cover data from the 2011 slice of the 2019 version of the National Land Cover Database).

In turn, more stringent planning regulations result in higher permitting costs to build new housing in the model. Empirically, we can measure permitting costs as a wedge between the price of housing in the periphery of cities and its replacement cost in the same location, including the value of undeveloped land and the cost of construction.[36]

Prices of undeveloped land are low relative to house prices and vary little across the periphery of different cities. Our calculations show that the price of a quarter-acre vacant lot of agricultural land in the periphery of US cities is about 2,700 dollars, and the coefficient of variation across the 112 cities in figure 5 is 0.93, for the same sample of cities as in figure 5 (see appendix B for details). This price represents less than one percent of the price of a single-family four-bedroom economy-quality house on average.[37] Construction costs are also fairly homogenous (Gyourko and Saiz, 2006). The mean construction cost in 2010 for a typical economy-quality house of 2,000 square feet is about 136,000 dollars, and the coefficient of variation across the 112 cities in figure 5 is 0.10.

To calculate the periphery house price-cost wedge in each city, we subtract from the price of a typical four-bedroom single-family detached house built 2000–2009 in the city periphery the city-specific construction costs for such a house, the city-specific price of a quarter-acre vacant lot of agricultural land in its periphery, and a gross profit margin of 17 percent. Panel D of figure 5 plots this price-cost wedge against planning regulations, measured again by the WRLURI index. As predicted by the model, more stringent planning regulations effectively result in a higher price-cost wedge.

In standard urban models, a positive productivity shock in a city may open a wedge between peripheral house prices and the cost of developing new houses. However, new constructions take place and eventually arbitrage this wedge away. A sluggish response of new construction, and not regulations, could thus be at the source of a positive periphery house price-cost wedge. But even with a sluggish construction response, we should see more new constructions in cities where a positive wedge opens up. In our model, instead, planning regulations are used to limit new construction, and a periphery house price-cost wedge is a feature of the equilibrium.[38] Importantly, this wedge in our model should not lead to differences in the expected growth rate of the housing stock: we expect Gibrat's law to hold as per equation (26), and it applies to both population and the housing stock. Figure C.1 in Appendix C confirms the absence of a

[36]For simplicity in our model, we ignore the cost of construction and normalise the value of land in alternative uses to zero so that, in equilibrium, permitting costs equal the housing price in the periphery of cities.

[37]When calculating replacement costs for housing, it is important to use the price of a vacant agricultural plot and not the price being paid for land by housing developers because the latter will include the effect of zoning and other regulations that limit where one can build. At the same time, it is also important to focus on the periphery of cities since agricultural land is more scarce there. Absent planning regulations, we would expect prices of undeveloped land in 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). 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). However, Plantinga, Lubowski, and Stavins (2002) show that variation in the value of land after its conversion to urban use and this option value contribute very little to the current value of agricultural land in the United States.

[38]We still expect new housing built in cities that experience a large positive shock, but only up to the point that incumbents find it to their best interest.

systematic relationship between permits for new residential units relative to the housing stock and the periphery house price-cost wedge of metropolitan areas in the United States.

7. The aggregate consequences of planning regulations

The evidence presented in the previous section indicates that planning regulations are a prevalent feature of the urban system in the United States, that they are systematically more stringent in more populous and geographically-constrained cities, and that, by opening a substantial wedge between the price and the cost of providing housing, they restrain the expansion of the most productive cities.

We now use our model to examine the aggregate implications of these planning regulations. We treat the actual distribution of population across cities and rural areas in the United States in 2010 as the equilibrium of our model, where planning regulations are chosen by incumbents in each city. This will be the baseline against which we compare counterfactual scenarios. We then change housing permitting costs in a subset of cities to some counterfactual level and derive the set of populated city sites, their population and physical sizes, individual consumption levels for incumbents and newcomers, and aggregate output under this counterfactual.

Counterfactual evaluation

Equation (22) gives consumption at the baseline equilibrium for incumbents in each city, which can be expressed as a function of N_{it}, z_i, and parameters. Since, in the model, each unit of raw land in city i provides $1/z_i$ units of developable land, our empirical counterpart to z_i is one over the share of the area within 30 kilometres of the centre of each city that is geographically unconstrained. We consider geographically unconstrained any area that is not covered by slopes steeper than 15%, water, wetlands, or land permanently protected from land cover conversion with a mandate to conserve its natural state, and it does not belong to a foreign country. Our empirical counterpart to N_{it} is the city's total population. We use the parameter values estimated in our empirical analysis in section 5: $\gamma = 0.07$, $\theta = 0.04$, $\sigma = 0.04$, $\beta = 0.04$, and $\lambda = 0.18$.

From equations (11) and (12), consumption for rural residents and city newcomers is given by

$$c_t = A_{rt}\left(N_t - \textstyle\sum_{i \in I} N_{it}\right)^{-\lambda}, \tag{40}$$

where I is the full set of populated cities, which corresponds to US metropolitan areas. We obtain the value of A_{rt} by equating $c_{rt} = c_{it}$ for the marginal populated city.

In the counterfactuals, we consider exogenously changing housing permitting costs in a subset of cities. Because incumbent residents no longer set permitting costs at their preferred level, their consumption is no longer given by equation (22) and we must use instead equation (19). Using a hat to denote the value of a variable under the counterfactual scenario that we wish to evaluate, we can rewrite equation (19) as

$$\hat{c}_{it} = \rho^{\sigma} A_{it} (h_{it})^{1+\sigma} (\hat{N}_{it})^{\sigma+\beta} - \frac{1}{\gamma+1} \tau_t (z_i)^{\gamma} \left(\hat{N}_{it}\right)^{\gamma+\theta}. \tag{41}$$

Then, we can eliminate $A_{it}(h_t)^{1+\sigma}$ from this equation using equation (21). Combining the resulting expression further with equations (17) and (22) gives a closed-form solution for the consumption for incumbents under the counterfactual relative to the baseline:

$$\frac{\hat{c}_{it}}{c_{it}} = \frac{\gamma+\theta}{\gamma+\theta-\sigma-\beta}\left(\frac{\hat{N}_{it}}{N_{it}}\right)^{\sigma+\beta} - \frac{\sigma+\beta}{\gamma+\theta-\sigma-\beta}\left(\frac{\hat{N}_{it}}{N_{it}}\right)^{\gamma+\theta}. \tag{42}$$

Changes in permitting costs in a subset of cities will also change which city sites are populated. We obtain the set of populated cities through the condition that $i \in \hat{I} \iff \hat{c}_{it} \geqslant \hat{c}_t$. Since counterfactual rural consumption is endogenous, we need an additional equation for this: equation (40) but with the set of populated city sites and their population sizes changed to those that correspond to the counterfactual, $\hat{I}$ and $\hat{N}_{it}$.

Allowing large and productive cities to expand further

Restrictions on urban expansion coupled with productivity differences across locations 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 are enacted by incumbent residents to maximise the benefits of a larger local population against its costs. These regulations, however, 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.

To quantify the gains that might be attained by allowing the most productive cities to expand further, we now examine a counterfactual where we relax their planning regulations. Our counterfactual targets the seven large cities (population above three million) in which there is a substantial wedge between house prices and their replacement costs at the periphery (above 200,000 dollars), indicating that planning regulations are significantly curtailing urban expansion in these locations. These are New York, Los Angeles, San Francisco - Oakland - San Jose, Washington DC, Boston, Seattle, and San Diego.

In our counterfactual, we allow for greater population growth in these seven cities between 1980 and 2010 through lower permitting costs. More specifically, we assume that permitting costs are forced to fall enough for housing permits to reach the 75th percentile level seen across all US cities over this 30-year period, 0.808 permits per initial housing unit. This compares with between 0.219 permits per initial housing unit for New York and 0.794 in Seattle, with a median across the seven cities of 0.445.[39]

The counterfactual increase in permits relative to the baseline is listed in column (1) of table 3. Under the assumption that each additional permit will, on average, facilitate the same additional inhabitants as each actual permit in that same city, columns (3) lists 2010 population in the counterfactual scenario, which can be compared against the actual 2010 population in column (2).

[39]Interestingly, this figure of 0.445 permits per initial housing unit is not far from the median level for all US cities of 0.490 permits per initial housing unit over this period. This is in accordance with our model and the evidence discussed in the previous section, indicating that, with incumbents setting regulations at their preferred level, we expect Gibrat's law to hold as per equation (26) for population and for the housing stock.

Table 3: Relaxing planning regulations in seven large and productive cities

	(1) Additional permits 1980–2010	(2) Baseline population 2010 (thousands)	(3) Counterf. population 2010 (thousands)	(4) Change in output per person	(5) Change in consumption per person	(6)
					Incumbents	Newcomers
New York, NY	271.1%	20,043	27,586	2.59%	-0.046%	6.55%
Los Angeles, CA	81.6%	17,877	23,083	2.07%	-0.029%	6.55%
San Francisco, CA	122.2%	7,413	9,912	2.35%	-0.038%	6.55%
Washington, DC	27.4%	8,615	9,389	0.69%	-0.003%	6.55%
Boston, MA	162.9%	6,300	7,869	1.80%	-0.022%	6.55%
Seattle, WA	1.8%	4,022	4,050	0.06%	-0.000%	6.55%
San Diego, CA	26.6%	3,095	3,423	0.81%	-0.004%	6.55%
Rural areas		57,499	40,414	6.55%		

Notes: The counterfactual targets the seven large cities (population above three million) in which there is a substantial wedge between house prices and their replacement costs at the periphery (above 200,000 dollars). The 1999 metropolitan area definition of San Francisco encompasses Oakland and San Jose. We assume that permitting costs in these seven cities are lowered enough for housing permits to reach the 75th percentile level seen across all US cities, corresponding to 0.808 permits per initial housing unit. Counterfactual 2010 population assumes that each additional permit will, on average, facilitate the same additional inhabitants as each actual permit in that same city. Values of output and consumption per person are those implied by the model with N_{it} given by counterfactual versus actual 2010 population, with z_i unchanged at one over the share of the area within 30 kilometres of the centre of each city that is geographically unconstrained, and parameter values estimated in section 5 ($\gamma = 0.07$, $\theta = 0.04$, $\sigma = 0.04$, $\beta = 0.04$, and $\lambda = 0.18$).

New York experiences the largest increase in its population, from 20 to 27.6 million, and other cities also grow substantially, with the exception of Seattle.

All of these changes would bring gains in output per person of between 0.06% in Seattle and 2.59% in New York through stronger agglomeration economies. However, rising house prices, longer average commutes, and greater congestion imply a modest fall in consumption for incumbents. Incumbent New Yorkers would experience consumption losses of -0.046%, with losses for incumbents in other cities being even smaller. 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 real gains of 6.55% (slightly smaller in the case former incumbents in relatively unproductive cities). Rural areas would see their population fall from 57.5 million to 40.4 million while enjoying an increase in consumption of 6.55%. The ten least attractive cities for incumbents would also be vacated by these, who could now do better in the seven cities that expand.

The other key source of aggregate gains in our framework is the fall in the cost of regulation for incumbent city residents. Our counterfactual exercise exogenously relaxes planning regulations in only seven cities. However, our model predicts that incumbent residents in other cities will endogenously lower planning regulations to keep themselves unaffected when the expansion of the seven targeted cities weakens pressure on their own housing market. Lower regulatory costs everywhere, not just in the seven cities directly affected, are an important source of aggregate gains in our framework, with newcomers everywhere also seeing consumption gains of 6.55%.

Overall, relaxing planning regulations in these seven large and restricted cities increases output per person by 7.95%. This effect is larger than any of the effects for pre-existing residents because new residents relocating from relatively unproductive locations gain even more. The overall change in average consumption per person is 2.16%.[40] Relaxing planning regulations would also substantially decrease inequalities. On average, consumption for city newcomers and rural residents would rise from 63.6% to 67.7% of the consumption of city incumbents.[41]

8. City population growth and aggregate output growth

We now turn to quantifying the contribution of cities to aggregate output and consumption growth. The ability to do this is a unique feature of our model.[42] We consider two different exercises. In the first, we perform some 'comparative dynamics' to quantify how much of aggregate output growth is accounted for, on average, by agglomeration economies. In the second, we counterfactually shut down population changes in cities. This allows us to quantify the contribution to aggregate output growth of the average population growth which occurs in cities as output evolves as well as of the reallocation of population across cities in response to differences in their productivity evolution.

The effect of agglomeration economies on aggregate output growth

We can compute expected growth in log output per person by taking the log of equation (8) and time differencing it. With i.i.d. shocks on productivity, which in turn affect city population

[40] While the empirical results in section 5 make us confident about the estimated values of our key parameters, the findings of our counterfactual exercise are robust to changes in these. Our estimate of total urban agglomeration benefits $\sigma + \beta = 0.04 + 0.04 = 0.08$ is higher than some estimates in the literature because it incorporates immediate benefits and learning benefits that accumulate over time. We have tried alternative values as low as $\sigma + \beta = 0.04$. For our estimate of total urban costs $\gamma + \theta = 0.07 + 0.04 = 0.11$, there is scarce literature for comparison, but we have tried alternative values in the range 0.09 – 0.13. Results are shown in table C.2 in appendix C. Depending on the combination of parameters, the overall change in consumption per person ranges between 2.07% and 2.21%. Since cities in our model operate where local urban benefits and costs are equated at the margin, moderate changes in city population have small net effects in consumption in each city, and most of the aggregate benefits arise from the spatial reallocation of population towards more productive locations. The overall increase in output per person varies more with parameters, but remains in the range 5.72% – 8.19%.

[41] In a related exercise, Hsieh and Moretti (2019) predict an 8.9% increase in US aggregate output 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. Despite the similar magnitude of their results and ours, the mechanics underlying them are quite different. Most importantly, Hsieh and Moretti (2019) do not consider the tradeoff between agglomeration benefits and urban costs. Instead, in their framework an increase in a city's population is always detrimental to existing residents and the optimal size of a city for an incumbent resident is zero. This has three implications. First, in their framework, decreasing returns dissipate most of the gains for migrants moving to more productive cities while population losses in less productive cities greatly benefit those left behind. Second, cities and their planning regulations are exogenous. In our framework, the endogenous relaxation of planning regulations in untargeted cities and the extensive margin of urbanisation are important sources of additional gains. Third, since the exogenous level of planning regulations affects housing costs for all local residents identically, their quantification does not distinguish between incumbent residents and newcomers.

[42] See Duranton and Puga (2014) for discussion of the difficulties of making growth endogenous in urban models. As mentioned above, Davis, Fisher, and Whited (2014) is a partial exception which focuses on housing and physical accumulation in a feedback loop with agglomeration economies. However, sustained growth does not occur in their neoclassical framework.

through equation (20), we can take expectations to obtain

$$\mathbb{E}(\Delta \ln(y_{it})) = \mathbb{E}(\Delta \ln(A_{it})) + (1+\sigma)\Delta \ln(h_t) + (\sigma+\beta)\mathbb{E}(\Delta \ln(N_{it})) \,. \tag{43}$$

We can similarly derive the evolution of expected city population from equation (25) as

$$\mathbb{E}(\Delta \ln(N_{it})) = \frac{1}{\gamma+\theta-\sigma-\beta}\left[\mathbb{E}(\Delta \ln(A_{it})) + (1+\sigma)\Delta \ln(h_t) - \Delta \ln(\tau_t)\right] \,. \tag{44}$$

An important feature of these equations is the absence of dynamic scale effects, in the sense that the growth of neither aggregate output, 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. The lack of dynamic scale effects is a desirable property. For output 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 output growth. The constant multiplying $\Delta \ln(h_t)$ in equation (43), which in the absence of cities would be 1, becomes $1+\sigma$ with cities.

In a second effect, city population growth contributes to output growth through the agglomeration economies that lead parameters σ and β to multiply $\Delta \ln(N_{it})$ in equation (43). 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 also foster city population growth). This indirect component can be seen in equation (44) where, if we let the agglomeration parameters σ and β become very small, cities grow very slowly (aside from also becoming small in population levels, as per equation 20). As shown by equation (43), any slowdown in city growth impacts output growth negatively.

In contrast to σ and β, the parameters related to the costs of cities, γ and θ, do not affect the growth of output directly in equation (43) since they play no direct role in production. They nonetheless affect output growth indirectly through their role in city population growth in equation (44).

To assess the quantitative contribution of agglomeration economies to output growth, we consider a thought experiment where we decrease agglomeration economies until they disappear. In light of equation (43), we need to know about the aggregate evolution of output per person, city populations, and human capital. Growth in output 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 $\Delta \ln(h_t) = \ln(1.006)$.

Starting with the magnification of individual human capital accumulation, $\Delta \ln(h_t)$, is multiplied by $1 + \sigma = 1 + 0.04 = 1.04$ in equation (43). In the absence of agglomeration economies, it would be multiplied by 1 instead. Since growth in output 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 output in the US from 0.60 to 0.62 percentage points. Expressed as a fraction of the total, this represents slightly more than 1% of the overall rate of output growth.

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 output 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 $\sigma + \beta = 0.04 + 0.04 = 0.08$ in the last term of equation (43). The product of these two terms, capturing the effect of larger cities on the rate of growth in output per person, is thus equal to an annual 0.12 percentage points. Expressed as a fraction of the total, this represents about 6% of the overall rate of output growth.

Combining the 0.12 percentage points in annual growth in output per person from average city population growth with the annual 0.02 percentage points from the magnification of individual human capital accumulation, we obtain a 0.14 percentage point difference or a modest 7% of the growth rate. Over the period of 60 years used for our calculations, such difference nevertheless adds up to 7.9% lower output per person in the absence of agglomeration effects.

The 0.6% annual growth in human capital over 1950–2010 implies that this factor directly accounts for 29% of output growth in equation (43). With cities accounting for another 7% though agglomeration effects, equation (43) implies that nearly two-third of output growth is accounted for by total factor productivity growth $\mathbb{E}(\Delta \ln(A_{it}))$. 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.[43]

In Appendix D, we use equation (44) to show that if agglomeration economies vanished, this would have a much larger effect on the population growth of cities than on aggregate output growth: if we brought σ and β towards zero from their estimated values, cities would grow on average 0.2% annually instead of 1.5%. In the process of performing these calculations, we also use (44) to back out a value for the evolution of commuting costs: an annual increase of 1.9% per year. In the same appendix, we show this implies an elasticity of the value of travel time with respect to aggregate income of 0.92 and discuss how this is consistent with the transportation literature on the subject.

[43] 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).

Spatial reallocation and aggregate output growth

The above quantification relates average city population growth to aggregate output growth. Importantly, by implicitly having all cities grow at the same expected rate, this quantification leaves aside the contribution of the extensive margin of city growth present in our model. As the most productive locations expand, they draw workers away from less productive cities and rural areas, leading to further aggregate gains.

To bring these additional gains into our quantitative analysis, we turn again to examining counterfactual scenarios in our model. The equations and procedure for evaluating these two counterfactuals are the same we use in section 7. First, imagine keeping the population of every US city at its 1950 level with no new city being created. At the same time, let total factor productivity in cities, transportation costs, and human capital evolve just as they did between 1950 and 2010. Between 1950 and 2010, the population of the conterminous United States grew by 156 million, with 141 million going into cities and only 15 million into rural areas. Absent the possibility of going into cities, average annual growth rate in output per person between 1950 and 2010 would drop from the actual 2.1% to 0.9%. Consumption losses would be lower, but after 60 years would still add up to 26.8%.

These losses from freezing city population growth that we have just described would be partly due to a falling urbanisation rate in a context of rising total US population. However, losses would also arise in part from not being able to accommodate more people into the most productive cities, as even incumbents would want, with rising human capital and total factor productivity. To isolate this last channel, let us perform the same thought experiment, but now without any aggregate population growth. Thus, instead of comparing actual with counterfactual outcomes, we now compare two counterfactual scenarios. In both scenarios, we keep aggregate population in the United States unchanged at its 1950 level. We also let total factor productivity in cities, transportation costs, and human capital evolve just as they did between 1950 and 2010. We then ask, what is the difference in this context between letting cities grow freely versus keeping the population of every city at its 1950 level with no new city being created.

If we allow cities to grow, incumbents in each city will be happy to let this happen to a varying extent, as rising productivity and human capital levels increase the population level where higher urban costs offset higher agglomeration economies. Since the evolution of productivity has been heterogeneous across cities, the most productive cities will expand even more. At the same time, because we are keeping aggregate US population unchanged at its 1950 level, the expansion of the most productive cities will draw population away from less productive cities and rural areas. Relative to being stuck with the 1950 US system (increasingly inefficient with evolving fundamentals even for unchanged total population), the reallocation of population towards and across cities brings an additional 0.66 percentage points of output growth annually.

9. 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. The number and size of cities are endogenous. Differences in productivity and geographic constraints across locations lead cities to differ in their population size. As households seek to live in the most attractive places, this heterogeneity represents an essential source of urban gains in addition to agglomeration economies. Driven by these potential gains, residents of less attractive locations would be willing to move to more attractive locations to the point of dissipating their advantage into longer and more congested commutes. For this reason, incumbent residents choose to limit the arrival of newcomers through planning regulations. The barriers imposed on newcomers represent another source of urban costs for them and a source of deadweight loss for society.

By modelling heterogeneity across locations as, in part, the outcome of cumulative productivity shocks, we also bring together random and systematic urban growth. In addition to rising total factor productivity, human capital accumulation and the evolution of commuting costs also drive aggregate growth. This combination allows matching key empirical features of modern urban systems, including city size distributions that follow Zipf's law, and ongoing urbanisation through a combination of gradual population growth of existing cities 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 output and consumption by relaxing planning regulations in seven particularly restricted large US cities. We counterfactually allow them to build as much as cities in the 75th percentile of permitting over a 30 year period. After these seven cities receive 18 million additional inhabitants among them, aggregate output increases by 7.95% and aggregate consumption by 2.16%. The expansion of the seven targeted cities weakens pressure on other housing markets, leading to endogenously laxer regulation elsewhere, and substantially reducing inequalities between incumbents and new residents.

Next, we assess the effects of cities and urbanisation on economic growth. Having cities expand on average amplifies aggregate income growth modestly 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 reallocation of population towards and across cities accounts for 0.66 percentage points of output growth annually.

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 taller buildings and smaller dwellings when land and housing get more expensive would capture relevant additional aspects of the urbanisation process, since these two margins further determine urban costs.

Third, allowing for labour mobility to affect endogenous urban consumption amenities beyond the agglomeration effects that we already consider would be an important step and may alter some of our quantitative conclusions. Another channel through which cities contribute to aggregate growth is by facilitating innovation. Modelling the process of creation and diffusion of new ideas and products would provide underpinnings for the evolution of total factor productivity, and amplify the role of cities further.

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. 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 further education. The worker devotes the remaining share $1-\delta_t^j$ of her time to working.

Human capital also generates entrepreneurial ideas. Since entrepreneurial ideas arise in proportion to the total local human capital after further education, H_{it}, it is natural to assume that rewards to these ideas 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 (3) yields conditional intermediate input demand:

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

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 (4), 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

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

The returns to each entrepreneurial idea can then be computed as:

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

where the final expression makes use of equations (A.2) and (5).

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

$$y_t^j = m_{it}\pi_{it}\frac{(1-\delta_t^j)b(\delta_t^j)\bar{h}_t}{H_{it}} + w_{it}(1-\delta_t^j)b(\delta_t^j)\bar{h}_t(N_{it})^{\beta} . \tag{A.4}$$

Worker j of the generation born at time t chooses how much time δ_t^j to devote to further education to maximise her income. Substituting equation (A.3) into equation (A.4), we see that the privately-optimal education will be defined by the following decision:

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

Simplifying and re-arranging the first-order condition yields

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

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 further education.

Appendix B. Data sources and treatments

City definitions. Our empirical and quantitative analysis focuses on the conterminous United States during the period 1950–2010. 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.

Population. We use county-level population data from the US decennial censuses for 1850, 1920, 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 Manson, Schroeder, Riper, Kugler, and Ruggles (2021) for 2010.

City centre and city periphery. We define the city centre as the location indicated by Google Maps for the core city of the metropolitan area.

In addition to defining centres, we need a measure for the spatial extent of the city, corresponding to $\bar{x}_{it}$ in the model. Since, in practice, cities cover two dimensions, there will be different distances between the city periphery and the city center depending on the direction

we follow. When cities have irregular shapes, using the maximum distance to the centre or a very high percentile of the distribution of distances to the centre can be problematic. Also, since metropolitan area definitions are county-based and some urban counties, particularly in the West of the country, extend well into rural areas, a few scattered dwellings very far away from the center in a county that is part of a city can increase the measured distance between the city periphery and the city center artificially. To address all of these difficulties, we implement a consistent definition of the city periphery. We take the city periphery to be the longest distance from the city centre within the metropolitan area boundaries that is within the 95th percentile of dwelling distances and has at least 500 dwelling units per square mile in the 2012 American Community Survey data described below.

Urban fringe. We use the term urban fringe for the area where the city would likely expand next. This is the area where we measure agricultural land prices when calculating replacement costs for housing at the city periphery and where we measure geographical constraints to urban expansion when relating these to the strictness of current planning regulations. The urban fringe is defined as the area within 20km of land developed at medium or high intensity in 2011 that is undeveloped or developed at low intensity, based on land cover data from the 2011 slice of the 2019 version of the National Land Cover Database.

Geographical constraints to urban expansion. To obtain an empirical counterpart to z_i, we calculate the share of the area within 30 kilometres of the centre of each city that is geographically unconstrained. This corresponds to $1/z_i$. We characterise this area with a 30-metre resolution. Each 30-metre cell is classified as geographically unconstrained if is not covered by slopes steeper than 15%, water, wetlands, or land permanently protected from land cover conversion with a mandate to conserve its natural state, and it does not belong to a foreign country. Slope is calculated on the basis of 1 arc-second Digital Elevation Models from the 3D Elevation Program of the US Geological Survey (2018). Water and wetlands cover is based on the 2019 National Land Cover Database (NLCD2019). The NLCD2019 (Dewitz and US Geological Survey, 2021) offers land cover for years 2001, 2003, 2006, 2008, 2011, 2013, 2016, 2019. We use the 2011 slice of the NLCD2019, since our estimations are all centred around 2010. Protected land is identified based on the Protected Areas Database of the United States (US Geological Survey, 2020). This database maps protected areas and assigns them a GAP status code as a measure of intent to permanently protect its natural state. We use GAP status codes 1 and 2 used to isolate land permanently protected from land cover conversion with a mandate to conserve its natural state. As to foreign land, this is identified from the official boundary files of Statistics Canada and Mexico's Instituto Nacional de Estadística y Geografía.

When calculating an empirical counterpart to z_i, we are interested in geographical constraints on the long-term expansion of a city over the course of its history. However, when thinking about geographical constraints to urban expansion as a determinant of the strictness of current planning regulations in panel C of figure 5, it is more appropriate to focus on the area where the city would

likely expand next. This is the urban fringe defined above, so we also calculate the share of the urban fringe that is geographically unconstrained for the same set of geographical constraints.

Illustrating the equilibrium with the urban system of the United States Panels A and B of figure 2 depict the allocation of population across US cities and rural areas in 1980 and 2010 as an equilibrium of the model. To draw this figure, we use parameter values estimated or calibrated in section 5 ($\gamma = 0.07$, $\theta = 0.04$, $\sigma = 0.04$, $\beta = 0.04$, and $\lambda = 0.18$), the actual population in each US metropolitan area and outside metropolitan areas in each year to assign values to N_{it} and N_{rt}, and the share of the area within 30 kilometres of the centre of each city that is geographically unconstrained as the empirical counterpart to $1/z_i$. We normalise $\tau_t = 1$ in 1980 (amounts to a choice of numérarire). We set τ_t in 2010 so that population-weighted average growth in y_{it} in the model matches the actual growth in average Gross Domestic Product per person in the United States 1980–2010. For this purpose, we use equation (21) to obtain $\rho^\sigma A_{it}(h_t)^{1+\sigma}$ for each city as a function of τ_t from its values of N_{it}, z_i, and parameters. Substituting this into equation (8) then yields y_{it} for each city as a function of τ_t. We obtain y_{rt} by equating income in rural areas with income in the marginal populated city, where the latter is given by equation (22). We find that increasing τ_t from 1 to 1.569 between 1980 and 2010 makes output per person in the model increase by a factor of 1.658, which matches the ratio of 2010 to 1980 Gross Domestic Product per person in the United States. Note that the numerical computation of τ_t is straightforward since every value of y_{it} is proportional to τ_t (equations 8 and 21). Note also that the ratio $1.569/1.658 = 0.94$ can be interpreted as the elasticity of the value of travel time with respect to income, which is a reasonable value in light of the transport literature on the subject discussed in appendix D.

In the figure, horizontal axis length is total US population, N_t. Total urban population $\sum_i N_{it}$ can be read as the horizontal distance to the left-side axes origin and rural population, $N_{rt} = N_t - \sum_i N_{it}$ can be read as the distance to the right-side axes origin. The thick horizontal segments in the figure represent equilibrium consumption for incumbents in each city, c_{it} (segment height), obtained from equation (40), and population N_{it} (segment length). The thin curves tangent to each thick segment plot consumption for incumbents in each city when population differs from its equilibrium level, as given by equation (41). Incumbents set permitting costs at $p_{it} = c_{it} - c_t$ to achieve the consumption at the maximum of the curve for their city while keeping newcomers indifferent. Rural consumption as a function of rural population is given by the smooth long curve, corresponding to equation (40), where we obtain the value of A_{rt} by equating $c_{rt} = c_{it}$ for the marginal populated city.

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 assign individual observations to a specific metropolitan area 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 population 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.

About one-third of the individual observations for residents in metropolitan area cannot be assigned to a specific area. We assign these observations to the same three size population 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.

National Household Travel Survey and Census. Column (1) of table 1 estimates γ as the elasticity of distance travelled with respect to the distance between her dwelling and the city centre. Data on household travel behaviour 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 log of household miles travelled on the log 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. 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. For consistency with the specifications using housing data, 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.

The controls for household characteristics, all based on the same NHTS data, are the log of the household size, the log 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 percentages of Hispanic, Black, and Asian population (based on 2010 Census data, obtained from the IPUMS-NHGIS project, Manson, Schroeder, Riper, Kugler, and Ruggles, 2021, since the 2008–2009 NHTS records block groups using 2010 boundaries), 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, with variation at the tract level), an indicator for waterfront location (constructed by combining the 2010 block-group boundaries provided in the IPUMS-NHGIS 2010 Census 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).

The measure of travel speed in each city included as a control in the regression in column (3) of table 1 and as the dependent variable in column (5) 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 log 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 log 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 log 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.

To validate the self-reported trip duration estimates of NHTS respondents, we turn to data from Akbar, Couture, Duranton, and Storeygard (2022). They query Google Maps over an extended time period about the duration of a trip with the same origin, destination, day of the week, and departure time as each trip reported by NHTS respondents. Using this alternative trip duration, they recompute an alternative measure of speed that we use in column (6) of table 1.

American Community Survey. All our estimations regarding housing rental prices and values use 5-year 2008–2012 data from the 2012 American Community Survey (ACS), obtained from the IPUMS-NHGIS project (Manson, Schroeder, Riper, Kugler, and Ruggles, 2021). The unit of observation is the block group. We use all block groups between the centre and the periphery of every metropolitan area except for college towns, as defined above, given their distinct housing markets.

All block-group housing regressions use the same controls for housing and 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 2008–2012 ACS data. The controls for block-group characteristics are the same as in the travel regressions, but re-computed for 2012 ACS block groups: the percentages of Hispanic, Black, and Asian population, the performance in standardised tests of the closest public school relative to the city average, an indicator for waterfront location, an indicator for riverfront location, and terrain ruggedness.

Column (2) of table 1 estimates γ based on variation in house prices across locations within a city as a function of distance to the city centre. The dependent variable is the log of the difference between the median rent in the most expensive block group in the city and the median rent in block group under consideration, from the 2008–2012 ACS data. We regress this on the log of the distance between the block group and the centre of its metropolitan area, city fixed effects, and the dwelling and neighbourhood characteristics described above.

Column (3) of table 1 estimates γ based on variation in house prices at the center of cities as a function of the spatial extent of the city. The first component of the dependent variable for this regression is estimated from an auxiliary regression at the block group level of the log of the median monthly contract rent on city indicators, a third-degree polynomial of distance between the block-group centroid and the city centre, and the aforementioned controls for housing and block-group characteristics. We use this regression to predict the rental price of a national-reference house for city-average neighbourhood characteristics at the centre of each city —i.e. when $x_i^j = 0$. This corresponds to $\hat{P}_i$ on the left-hand side of the empirical specification of equation (35). On the right-hand side of that expression, we have the spatial extent of the city, $\bar{x}_{it}$, and travel speed $\hat{\tau}_i$. We measure $\bar{x}_{it}$ using the distance between the centre and the periphery of each city as defined above, i.e. the longest distance from the city centre within the metropolitan area boundaries that is within the 95th percentile of dwelling distances and has at least 500 dwelling units per square mile in the 2012 ACS data. Our estimate of speed is the predicted speed of a 15km commuting trip on a Tuesday at 8:00AM by a driver with average characteristics in each city, using NHTS data as described above.

The final component of equation (35) that we need to measure is c_t. Equation (24) tells us this should be proportional to the price of housing at the centre of the cheapest city. Unfortunately, the proportionality constant is itself a function of our key parameter of interest, γ. Since γ, appears on both sides of equation (35), we estimate this iteratively. Given a starting value of γ, the values of θ, σ and β obtained below, and the estimated city-centre house price in the cheapest city $\underline{\hat{P}}$, we obtain a value for $c(\gamma) = \frac{\gamma+\theta-\sigma-\beta}{(\sigma+\beta)(\gamma+1)}\underline{\hat{P}}$. This value allows us to compute our dependent variable

in regression (35), $\ln[\hat{P}_i + c(\gamma)]$. Estimating this regression by ordinary least squares yields an updated value of γ, which allows recomputing $c(\gamma)$ and thus $\ln[\hat{P}_i + c(\gamma)]$. We then re-estimate regression (35), and so on until convergence is achieved.

Figure 4 plots housing price gradients for five US cities. We predict the monthly rent of a dwelling with average national characteristics in a neighbourhood with average city characteristics as a function of distance to the city centre with a semilinear regression at the block-group level for each city using Yatchew's (1998) difference estimator. The dependent variable is the median contract rent in the block group. The linear component includes the same dwelling and neighbourhood controls as column (2) of table 1 while distance to the city centre is treated nonparametrically.

Panel A of figure 5 plots the city-periphery monthly rent against 2010 city population. City-periphery monthly rent is the monthly rent of a dwelling with average national characteristics in a neighbourhood with average city characteristics located at the city periphery. This is estimated from the same regression used to estimate the city-centre monthly rent used in column (3) of table 1, but valued at a distance from the centre corresponding to the periphery of each city instead of at a distance zero.

Planning regulations. The strictness of planning regulations in each metropolitan area plotted in panels B, C, and 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 2006 nationwide survey of residential planning regulations in over 2,600 communities across the United States. Gyourko, Hartley, and Krimmel (2021) construct an updated index based on a 2018 survey with some differences with respect to the 2006 survey, both in terms of questions and responding communities. To aggregate the 2006 index to the level of metropolitan areas, we retain data on the 1896 responding communities that are part of a 1999 MSA/NECMA, average their index to the level of primary metropolitan statistical areas weighting by their population with a correction for community response probability provided by Gyourko, Saiz, and Summers (2008), and then average these values to the level of metropolitan areas weighting by population. To aggregate the 2018 index to the level of metropolitan areas, we retain data on the 1877 responding communities that are part of a 1999 MSA/NECMA, and then average their index to the level of metropolitan areas using the weights provided for large metropolitan areas by Gyourko, Hartley, and Krimmel (2021) and population weights for the rest. Finally, we interpolate the 2006 and 2018 values of the index to obtain a value for 2010 to match the timing of our other data.

Housing replacement costs and price-cost wedges. The periphery house price - cost wedge plotted against the strictness of planning regulations in panel D of figure 5 is the difference between the value of a house and its replacement cost in the periphery of the city. The house value corresponds to a four-bedroom single-family detached house built 2000–2009 in a neighbourhood with average city characteristics located at the city periphery. This is estimated based on a regression of the log of the median house value in the block group on a third-degree polynomial

of distance to the city centre and the same dwelling and neighbourhood controls as column (2) of table 1 using 2008–2012 American Community Survey (ACS) data. City periphery is defined the longest distance from the city centre that is within the 95th percentile of dwelling distances and has at least 500 dwelling units per square mile in the block group.

The replacement costs is the sum of city-specific construction costs for an economy-quality single-family detached house of 2000 square feet (using RSMeans data for 2010 obtained from Glaeser and Gyourko, 2018) and the price of a quarter-acre vacant plot of land used for agriculture at the urban fringe (using land value data from Nolte, 2020 and land cover data from the 2011 slice of the 2019 version of the National Land Cover Database). The urban fringe is defined as the area within 20km of land developed at medium or high intensity in 2011 that is undeveloped or developed at low intensity, based on land cover data from the 2011 slice of the 2019 version of the National Land Cover Database.

Building permits. Data about the number of building permits plotted in figure C.1 and used for our counterfactuals are from the US Department of Housing and Urban Development (HUD). The source data is at the county level and we aggregate this up to the 1999 MSA/NECMA level. The variable annual permits relative to housing stock on the vertical axis of figure C.1 divides for each city the total number of residential construction permits during the period 2008–2012 (to match the timing of the ACS housing data) over the total number of housing units in the city for that period as recorded in the ACS data.

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) 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 log of city size with the percentage of the area in a 30-kilometre radius around the city centre that has slopes greater than 15% and the percentage covered by wetlands (both computed as in our geographical constraints to urban expansion), the inverse hyperbolic sine of the city's population in 1850 and 1920 (from Schroeder, 2016), the inverse hyperbolic sine of the distance to the Eastern Seaboard (computed using coastline shapefiles from the National Hydrography Dataset of the US Geological Survey), and heating degree days (from Burchfield, Overman, Puga, and Turner, 2006).

Appendix C. Supplementary tables and figures

Table C.1: First-stage of TSLS estimation of agglomeration economies (model parameter σ)

Dependent variable:	(1) Log city population
% slope $>$ 15% 30km from city centre	-0.0073***
	(0.0024)
% wetlands 30km from city centre	-0.0171***
	(0.0054)
Arsinh population 1850	-0.0898***
	(0.0144)
Arsinh population 1920	0.9605***
	(0.0288)
Arsinh distance to Eastern Seaboard	-0.0086
	(0.0113)
Heating degree days	-0.0003***
	(0.0000)
Observations	50,393
R^2	0.8599
P-value LM test (H_0: model underidentified)	0.0000
F-test weak ident. (H_0: instruments jointly insignificant)	346.993
P-value J test (H_0: instruments uncorr. with error term)	0.4444
P-value endog. test (H_0: exogeneity of instrumented var.)	0.3065

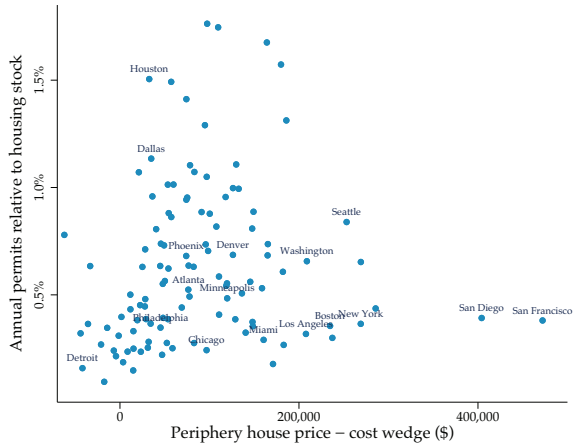
Notes: The table reports the first stage of the two-stage least squares specification in column (2) of table 2 in the main text. As in that column, units of observation are individual worker-year pairs (annually 1980–1994 and biannually 1996–2012). The dependent variable in the first stage (and instrumented variable in the second stage) is the natural log of city population in 2010.

We report coefficients for all the excluded instruments: the percentage of the area within 30-kilometres of the city centre that has slopes greater than 15% and the percentage covered by wetlands, the arsinh of city population in 1850 and 1920, the arsinh of distance to Eastern Seaboard, and heating degree days. All other regressors in column (2) of table 2 are also included: worker fixed-effects, experience in cities $\geqslant$ 5 million, experience in cities $\geqslant$ 5 million $\times$ experience, experience in cities 2-5 million, experience in cities 2-5 million $\times$ experience, experience and its square, firm tenure and its square, and indicators for two-digit sector, occupation, and year.

All regressions include a constant term. Coefficients are reported with robust standard errors in parenthesis, which are clustered by worker and city. ***, **, and * indicate significance at the 1, 5, and 10 percent levels. The R^2 reported in columns is within workers.

The F-statistic (or Kleinberger-Papp Wald statistic) reported on the weak instruments identification test exceeds all thresholds proposed by Stock and Yogo (2005) for the maximal relative bias and maximal size.

Figure C.1: New construction and the housing price-cost wedge in the United States



Notes: The same 112 metropolitan areas as in the four panels of figure 5 are represented and all metro areas with population above 3 million are labelled.

Periphery house price - cost wedge is the difference between the value of a house and its replacement cost in the periphery of the city. The house value corresponds to a four-bedroom single-family detached house built 2000–2009 in a neighbourhood with average city characteristics located at the city periphery. This is estimated based on a regression of the log of the median house value in the block group on a third-degree polynomial of distance to the city centre and the same dwelling and neighbourhood controls as column (2) of table 1 using 2008–2012 American Community Survey (ACS) data. City periphery is defined the longest distance from the city centre that is within the 95th percentile of dwelling distances and has at least 500 dwelling units per square mile in the block group. The replacement costs is the sum of city-specific construction costs for an economy-quality single-family detached house of 2000 square feet (using RSMeans data for 2010 obtained from Glaeser and Gyourko, 2018) and the price of a quarter-acre vacant plot of land used for agriculture at the urban fringe (using land value data from Nolte, 2020 and land cover data from the 2011 slice of the 2019 version of the National Land Cover Database). The urban fringe is defined as the area within 20km of land developed at medium or high intensity in 2011 that is undeveloped or developed at low intensity, based on land cover data from the 2011 slice of the 2019 version of the National Land Cover Database.

Annual permits relative to the housing stock divides 2008-2012 building permits from the US Department of Housing and Urban Development by the total number of housing units in the city from the 2008–2012 ACS.

Table C.2: Sensitivity analysis: relaxing planning regulations in seven large and productive cities

		$\sigma+\beta$		
		0.04	0.06	0.08
$\gamma+\theta$	0.09	2.21% 5.72%	2.21% 6.92%	2.21% 8.19%
	0.11	2.17% 5.96%	2.16% 6.93%	2.16% 7.95%
	0.13	2.08% 6.27%	2.07% 7.08%	2.07% 7.92%

Notes: The counterfactual targets the seven large cities (population above three million) in which there is a substantial wedge between house prices and their replacement costs at the periphery (above 200,000 dollars). We assume that permitting costs in these seven cities are lowered enough for housing permits to reach the 75th percentile level seen across all US cities. The top number in every cell is the increase in average consumption per person and the bottom number is the increase in average output per person for each pair of values for $\sigma+\beta$ and $\gamma+\theta$. Values of output and consumption per person are those implied by the model with N_{it} given by counterfactual versus actual 2010 population, with z_i unchanged at one over the share of the area within 30 kilometres of the centre of each city that is geographically unconstrained. Our baseline results correspond to $\sigma+\beta=0.08$ and $\gamma+\theta=0.11$.

Appendix D. The effect of agglomeration economies on city population growth and the implied changes in travel costs

We now assess the effect of agglomeration economies on city population growth using equation (44). We reproduce this equation for convenience:

$$\mathbb{E}(\Delta \ln(N_{it})) = \frac{1}{\gamma + \theta - \sigma - \beta} \left[\mathbb{E}(\Delta \ln(A_{it})) + (1+\sigma)\Delta \ln(h_t) - \Delta \ln(\tau_t) \right] . \tag{44}$$

In this equation, the growth rates in population $\Delta \ln(N_{it})$ and human capital $\Delta \ln(h_t)$ can be read again directly from the data, with annualized values of 1.5% and 0.6% respectively. The change in total factor productivity, $\Delta \ln(A_{it})$, is easily obtained by difference from the quantification of equation (43). It is equal to 1.4% annually. We can then use equation (44) to back out the change in unit commuting costs $\Delta \ln(\tau_t)$ which is equal to 1.9% annually.

Before considering what would happen to city population growth if agglomeration economies vanished, it is worth asking whether this annual increase in commuting costs of 1.9% implied by the structure of the model is reasonable. 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.

Now, as in the main text, we consider again a thought experiment where we decrease agglomeration economies until they disappear. As made clear by equation (20), any increase in total factor productivity or human capital and any decrease in transportation costs is exponentially magnified by the inverse ratio of urban costs minus agglomeration effects: $\frac{1}{\gamma+\theta-\sigma-\beta}$. Setting $\sigma + \beta$ to zero instead of $\sigma + \beta = 0.08$ dramatically alters the fundamental tradeoff between agglomeration economies and urban costs and reduces the magnification ratio $\frac{1}{\gamma+\theta-\sigma-\beta}$ by a factor of nearly four. The disappearance of agglomeration economies also implies the end of the magnification of human capital by a factor of $1 + \sigma$. Overall, disappearing agglomeration effects lead to an expected growth rate of city population, $\mathbb{E}(\Delta \ln(N_{it}))$, of 0.2% per year instead of its historical value of 1.5%. While agglomeration economies only have a minor impact on the growth of aggregate income, they have a large impact on the population growth of cities.

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.

Ahlfeldt, Gabriel M. and Elisabetta Pietrostefani. 2019. The economic effects of density: A synthesis. *Journal of Urban Economics* 111: 93–107.

Akbar, Prottoy A., Victor Couture, Gilles Duranton, and Adam Storeygard. 2022. Mobility and congestion in urban India. *American Economic Review* (forthcoming).

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*. Spatial evolution of population and industry in the United States. *American Economic Review* 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.

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 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. *Review of Economics and Statistics* 100(4): 725–739.

Couture, Victor and Jessie Handbury. 2019. Urban revival in America. Preprint, 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. 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. City of dreams. Preprint, 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 5. 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.

Dewitz, Jon and US Geological Survey. 2021. *National Land Cover Database (NLCD) 2019 Products: Version 2.0, June 2021*. Sioux Falls, SD: United States Geological Survey.

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

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

Duranton, Gilles and Diego Puga. 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.

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.

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

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

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. and Joseph Gyourko. 2018. The economic implications of housing supply. *Journal of Economic Perspectives* 32(1): 3–30.

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, Jonathan S. Hartley, and Jacob Krimmel. 2021. The local residential land use regulatory environment across US housing markets: Evidence from a new Wharton index. *Journal of Urban Economics* 124: 103337.

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: Vintage.

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, Tracy Kugler, and Steven Ruggles. 2021. *Integrated Public Use Microdata Series, National Historical Geographic Information System: Version 16.0*. Minneapolis: IPUMS.

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

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

Mussa, Michael. 1974. Tariffs and the distribution of income: The importance of factor specificity, substitutability, and intensity in the short and long run. *Journal of Political Economy* 82(6): 1191–1203.

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

Nagy, Dávid Krisztián. 2020. Hinterlands, city formation and growth: Evidence from the US westward expansion. Preprint, Centre de Recerca en Economia Internacional.

Nolte, Christoph. 2020. High-resolution land value maps reveal underestimation of conservation costs in the united states. *Proceedings of the National Academy of Sciences* 117(47): 29577–29583.

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.

Puga, Diego. 1999. The rise and fall of regional inequalities. *European Economic Review* 43(2): 303–334.

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.

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

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.

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

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.

Stock, James H. and Motohiro Yogo. 2005. Testing for weak instruments in linear IV regression. In James H. Stock and Donald W. K. Andrews (eds.) *Identification and Inference for Econometric Models: Essays in Honor of Thomas J. Rothenberg*. Cambridge: Cambridge University Press, 109–120.

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 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.

US Geological Survey. 2020. *Protected Areas Database of the United States (*PAD-US*): Version 2.1, December 2020*. 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.

Yatchew, Adonis. 1998. Nonparametric regression techniques in Economics. *Journal of Economic Literature* 36(2): 669–721.