The optimal distribution of population across cities*

David Albouy† Kristian Behrens‡ Frédéric Robert-Nicoud§ Nathan Seegert¶

13 February 2017

Abstract: We upend the received economic wisdom that cities are too big. This wisdom assumes that city sites are homogeneous, migration is unfettered, land is given to incoming migrants, and federal taxes are neutral. In a more general city system with heterogeneous sites, we demonstrate that cities may be inefficiently small with local governments or unrestricted migration. A quantitative simulation suggests that cities may be too numerous, with the best sites underpopulated, for a wide range of parameter values that resemble developed countries. Welfare costs from free migration equilibria appear small, whereas they appear substantial when local governments control city size.

Keywords: City size; heterogeneous sites; local governments; land value; federal taxation.
JEL Classification: R12; J61; H73.

---

*This paper merges and supersedes Albouy and Seegert (2012) and Behrens and Robert-Nicoud's (2014) pieces. We are grateful to Vernon Henderson for his detailed and extremely valuable comments. We are also grateful to Costas Arkolakis, Richard Arnott, Spencer Banzhaf, Morris Davis, Klaus Desmet, Pablo Fajgelbaum, Patrick Kline, David Pines, Esteban Rossi-Hansberg, Bernard Salanié, William Strange, Jacques Thisse, Tony Venables, Wouter Vermeulen, Dave Wildasin, and numerous conference and seminar audiences for discussions and feedback. Albouy would like to thank the Lincoln Institute for Land Policy for generous assistance on this project. Behrens and Robert-Nicoud gratefully acknowledge financial support from the CRC Program of the Social Sciences and Humanities Research Council (SSHRC) of Canada for the funding of the *Canada Research Chair in Regional Impacts of Globalization*. The study has been funded by the Russian Academic Excellence Project '5-100' and the Michigan Center for Local, State, and Urban Policy (CLOSUP).

†University of Illinois at Urbana Champaign, USA; and NBER, USA. Email: albouy@illinois.edu

‡Université du Québec à Montréal, Canada; National Research University Higher School of Economics, Russian Federation; and CEPR, UK. E-mail: behrens.kristian@uqam.ca

§HEC Lausanne, Switzerland; GSEM, Switzerland; SERC, UK; and CEPR, UK. E-mail: frederic.robert-nicoud@unil.ch

¶University of Utah, USA. E-mail: nathan.seegert@business.utah.edu

1. Introduction

Cities define civilization and epitomize modernity, and yet the received economic wisdom is that they are too big. Positive urban externalities — from better matching, greater sharing, and quicker learning — give rise to agglomeration economies that create a centripetal attraction to cities. These are countered by negative externalities — congestion, crime, pollution, and disease — that create a centrifugal repulsion from cities. In the standard argument, negative externalities come to dominate positive ones with city size. Free migration then causes cities to become inefficiently large because migrants to cities do not pay for their increasingly negative externalities. This view that "cities are never too small" is presented as fact in any first course in urban economics, e.g., O'Sullivan (2011), and it is easily accepted as it reinforces ancient negative stereotypes of cities. Ultimately, this view legitimizes policies that limit urban growth, such as land-use restrictions and disproportionate governmental transfers towards rural areas.

We upend the received wisdom and argue that *large cities are likely to be too small* for either of two simple, but profound, reasons. First, *fiscal externalities* from federal taxes and land purchases — which arise as individuals do not internalize the consequences of their location decisions on revenues from taxes and land ownership — increase with city size, and generally benefit non-residents.[1] Second, city sites are of *heterogeneous quality*, and incentives are generally poor at allocating individuals efficiently across those sites, especially when local authorities or interest groups have some control over in-migration.[2] As a result, both the intensive (number of cities) and extensive (size of cities) margins of urbanization are generally inefficient: large cities on high-quality sites will be too small, and cities will be developed on sites of quality inferior to that of an efficient urban system.[3]

Our theoretical analysis characterizes the general properties of an efficient urban system. Cities have a minimum efficient scale and should be more populated on high-quality sites than on lower-quality ones, especially when agglomeration economies are strong relative to urban

[1]Fiscal externalities from land purchases are mentioned in the work of Helpman and Pines (1980) and are connected to the much discussed GHV (Henry George) Theorem in Vickrey (1977), Stiglitz (1977), and Arnott and Stiglitz (1979). For reviews, see Vickrey (2002) and Arnott (2004), who states (p. 1072) that "Ricardian differences in land" have not to his knowledge "been investigated in the literature." Externalities from federal taxation are discussed in Hochman and Pines (1997), and Albouy (2009, 2012). Ades and Glaeser (1995) argue that migrants to capital cities, in particular, tend to absorb federal funds rather than contribute to them.

[2]Heterogeneous sites are a first-order feature of the world according to the work of Haurin (1980), Roback (1982), Redding and Sturm (2008), Bleakly and Lin (2012), Davis and Weinstein (2002), Behrens, Mion, Murata, and Suedekum (2011), Desmet and Rossi-Hansberg (2013), Allen and Arkolakis (2014), and Albouy (2016). However, first nature is only one aspect that determines the location of cities. See, e.g., Powell (2012) for a detailed description of how local interest-group thinking and colonial settlement policy jointly influenced the location of New Orleans on what is arguably an inferior site.

[3]For arguments that cities are too large, see Harris and Todaro (1970), Tolley (1974), Arnott (1979), Upton (1981), Abdel-Rahman (1988), and Fenge and Meier (2002). Formal reasoning on optimal systems of regions was pioneered by Buchanan and Goetz (1972) and Flatters, Henderson, and Mieszkowski (1974); developed extensively by Henderson (1974a); and given comprehensive treatments by Kanemoto (1980), Henderson (1988), Fujita (1989) and Abdel-Rahman and Anas (2004).

dis-economies. Sites that do not achieve a minimally efficient scale remain undeveloped. The better the site, the more it should be crowded past the point that would be optimal if all sites were identical, as good sites are scarce. System-wide, aggregate land values should exceed the value of agglomeration economies by an amount proportional to the dispersion of urban wage premia. This finding generalizes the "George-Hotelling-Vickrey" (GHV) Theorem — a.k.a, the "Henry George" Theorem — of Vickrey (1977) and Stiglitz (1977), and opens the case for a land tax at the federal level as opposed to a strictly local one.

When cities are given local control over their populations, cities on the best sites are prone to be under-populated. Without side payments (e.g. impact fees), residents on good sites will halt immigration as soon as it causes them to suffer in the least, no matter how great the migrants' benefit. As a result, the excluded population inhabits low quality sites that would not be developed optimally, as well as rural areas beyond what is optimal. With fiscal externalities, this inefficiency is exacerbated, as residents ignore how their own city being made larger benefits the greater economy.

Under free-migration, the see-saw of over or under-urbanization may swing in either direction. Without fiscal externalities, better sites have inefficiently high populations, and sub-optimally few sites are inhabited. This confirms the prevailing sense that the developing worlds' mega-cities are over-crowded. In developed countries, where fiscal externalities are strong, the opposite situation occurs: migration to the best sites is suboptimal, and as may be urbanization overall.

We illustrate our model and gauge its quantitative implications by applying it to U.S. data. A quick test based on our Generalized HGV Theorem indicates that the American urban system is less congested than an optimal system with similar urban wage dispersion, but more congested than a system determined solely by local politics. More detailed simulations, based on precise numbers, imply that large American cities are undersized by about a third, that the number of cities is twice the optimum, and that more than half of the urban population is misallocated. Despite that sizable misallocation, the ensuing welfare costs are equal to only around 1% of real consumption in the free-migration equilibrium. The main reason for this low elasticity of welfare costs to the scope of urban misallocation is that the urban system operates at close to constant returns.[4] Misallocation costs are substantially higher if migration is impeded by local governments. In that case, local politics may generate welfare costs of about 18% of real consumption. While the data suggest that large U.S. cities may be too small, urban systems in developing countries — where fiscal externalities appear slight and coordination problems more rampant — may be more prone to over-urbanization, with cities on the best sites suffering from the greatest overcrowding.

Our paper contributes to several strands of the literature. First, our approach yields a

[4]Behrens et al. (2011), Desmet and Rossi-Hansberg (2013), and Behrens, Duranton, and Robert-Nicoud (2014) also find that the elasticity of welfare costs to the scope of urban misallocation is fairly small with free migration.

comprehensive characterization of the full urban system, thereby revisiting and encompassing the canonical work of Buchanan and Goetz (1972), Flatters, Henderson, and Mieszkowski (1974), and the extensive literature that followed these pioneering work (see footnote 3). Second, we complement work on the consequences of externalities and size restrictions on the fabric of urban systems pioneered by Vickrey (1977), Stiglitz (1977), and others (see footnote 1) and contemporaneously revisited by Eeckhout and Guner (2016) and Hsieh and Moretti (2016). In the latter models, the efficient urban scale is zero, with diminishing returns to urban scale at all city sizes. As a result, these settings feature only the intensive margin of urbanization (how to allocate population among an exogenously given number of cities). Like these authors, we find that externalities and size restrictions may severely distort urban systems. Our distinctive contribution is to show how modeling the extensive margin (allowing the number of cities to vary) adds new insights and qualifies several positive and normative results. This modeling approach is also more amenable to developing countries, where many challenges of urbanization lie today.

Finally, variable returns to city size and the presence of an extensive margins limit prevent us in using the tools pioneered by Allen and Arkolakis (2014) to characterize and solve quantitative economic geography models such as Fajgelbaum, Morales, Suarez Serrato, and Zidar (2015) or Redding (2016) that do not feature that margin. Our solution is to simplify the geography by lumping any local production advantages of a city into a single parameter. Under this more classical assumption, we can solve for different allocations of people to cities and characterize the positive and normative properties of those allocations.

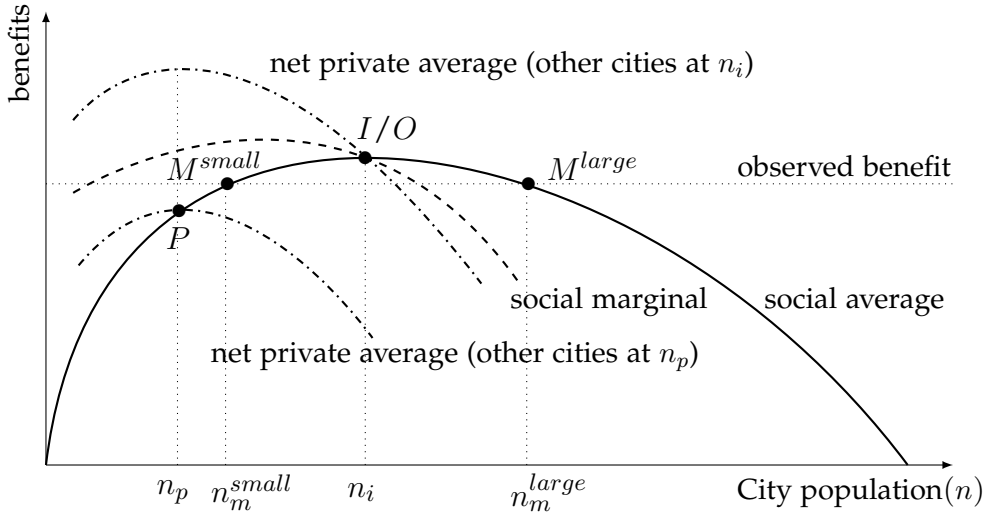
The rest of the paper is structured as follows. Section 2 builds intuition and introduces the model. Section 3 characterizes three different spatial allocations: (i) the optimal federal allocation, which prevails when choices are made at the federal level; (ii) the local politics allocation, which prevails when choices are made locally by city governments; and (iii) the free migration allocations, which prevail when choices are made by unconstrained individuals. Section 4 discusses how the different externalities can (or cannot) be internalized using federal fiscal instruments and derives the optimal policy. Section 5 discusses our baseline calibration, quantifies distortions in the city size distribution, and puts numbers on the welfare costs of population misallocation across cities, with either free migration or city governments. Section 6 summarizes and concludes. A collection of appendices contains proofs, extensions, and data descriptions.

2. An urban system with heterogeneous sites and fiscal externalities

2.1 Preview of the model

We develop our argument using a parsimonious model of urban systems that extends the seminal work of Henderson (1974b). We depart from the canonical setting by adding heterogeneous sites and fiscal instruments, including land purchases, federal taxes, and discounts to congestion and housing costs. To understand how adding either fiscal externalities or heterogeneous sites to the canonical model can overturn a central result in urban economics, consider first the basic argument explaining why cities are too large.

Figure 1: Benefits curves and coordination problem with homogeneous sites.

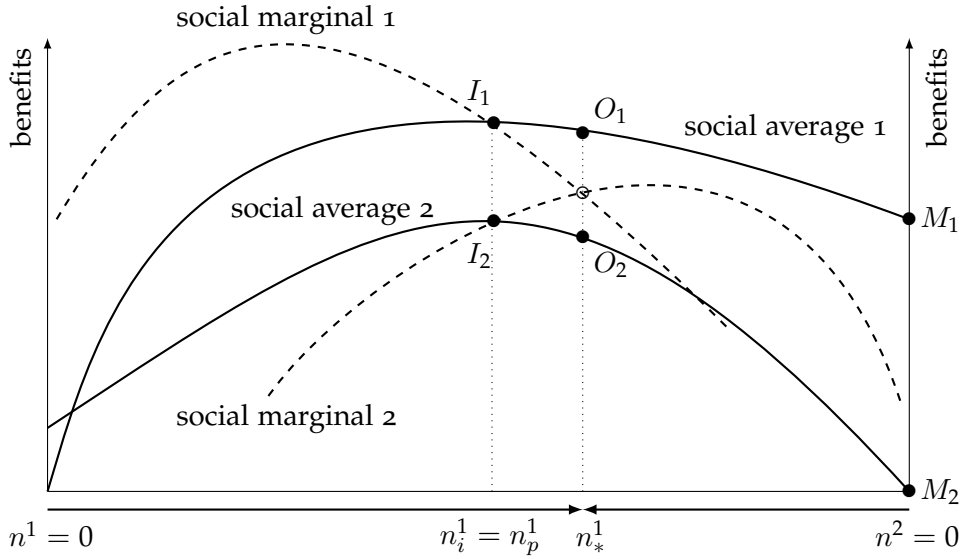


As drawn in Figure 1, the benefits migrants receive from entering a city are equal to the average, rather than the marginal, benefit associated with urban life. Analogously to the efficient scale of a firm, the efficient population scale of a city is attained at point I/O, where the social marginal and social average benefits coincide. Yet, migrants who respond to the (social) average benefit enter a city until it equals that of an outside option — possibly from the countryside or another city — shown here by the 'observed benefit' equilibrium line. As equilibrium population levels are only stable when benefits are falling with city size (to the right of n_i), population levels such as n_m^{large} are possible, while levels below the optimum, such as n_m^{small}, are ruled out.[5] If all potential sites were identical, local governments would optimally reduce their respective populations to n_i, thereby raising the benefits of residents everywhere. All existing cities are hence too large in equilibrium, and there are too few of them, which is the standard result. In this paper, we extend that basic model in two directions and show how each of them, and their interactions, give rise to fundamentally different results.

[5]Using a similar reasoning, Knight (1924) already pointed out almost a century ago that free-access highways tend to become overly congested.

Consider first the presence of fiscal externalities. With fiscal externalities — or any other kind of cross-city externality — migrants respond to incentives described by the dash-dotted net private average benefits curve instead of the (social) average benefit in Figure 1. Assume that the fiscal externalities are zero on net when all cities are the same size (through budget balance), and that these externalities increase with city size. Therefore, private average benefits are below social benefits when a city is larger than the others. As externalities increase with city size, the private average benefit curve peaks to the left of point I/O. Hence, each local government has an incentive to shrink its city population to the new peak n_p — which is to the left of the optimum n_i — thereby free-riding off the benefits from other cities. If all city governments lower their population to n_p, the reduction in fiscal externalities shifts the net private average benefit curve down. At the resulting equilibrium, point P, which achieves budget balance, residents everywhere are worse off. Free-riding of local governments causes cities to be too small and too numerous. If cities can enforce cooperation, they would all be better off by enforcing a *higher* population level, n_i, which runs counter to the canonical argument that cities are too large. With free migration by atomistic agents and fiscal externalities, all population levels to the right of n_p are potentially stable, provided other cities are of the same size. While cities may be too large, at n_m^{large}, they may also be too small, at size n_m^{small}. *A priori* the social average benefit curve is unobserved: we do not know if it is upward or downward sloping at an observed population and benefit level. Without further information, we do not know if the population is to the left or right of the optimum.

Figure 2: Population split between superior (1) and inferior (2) sites.



Consider next heterogeneous city sites. Heterogeneity in site quality can also lead to cities being too small, particularly when city populations are constrained by local governments. Consider an example with two cities — and no fiscal or other cross-city externalities — in

a classic 'bucket' diagram in Figure 2. Distance from the left vertical axis represents the population in city 1, whereas the remaining population — given by the distance to the right vertical axis — lives in city 2. As seen by its higher benefit curve, city 1 is located on a superior site to that of city 2, e.g., a natural harbor on an ocean shore versus a landlocked location in a scorching desert. Say that, by coincidence, with city governments, the cities are each at the peak of their social average benefit (denoted sab) at I_1 and I_2, with the population divided equally between the two: $n_p^1 = n_i^1 = n_p^2 = n_i^2$. Yet, *the social optimum is where the social marginal benefit (denoted smb) curves cross* at points O_1 and O_2, where city 1 is larger than city 2: $n_*^1 > n_*^2$. The optimum balances the supramarginal gains of migrants with the inframarginal losses of residents. The marginal losses for residents at sites 1 and 2 (the decrease along the benefits curve due to in- or out-migration) are more than offset by the discrete gains of the migrants moving from the inferior site 2 to the superior site 1 (the upwards jump between the benefits curves due to site switching). From the point of view of local voters, the better site is over-populated at O_1, whereas the worse site is under-populated at O_2. Thus, benefit-maximizing voters in city 1 would vote to curb migration since they bear the costs of those moves. Note, however, that free migration could still lead to city 1 being overcrowded and city 2 to be uninhabited, at points M_1 and M_2, respectively.

The foregoing simple examples consider separately fiscal externalities or two heterogeneous sites. Moving to a general city-systems model that allows for both fiscal wedges and numerous heterogeneous sites is a demanding task (Henderson, 1988), particularly when cities have bell-shaped benefit curves. We approach the problem by taking site heterogeneity as continuous — following the pioneering work of Aumann (1964) — and using concrete functional forms.[6] The latter involve reduced forms of urban externalities, with economies and diseconomies changing with the scale of a city at constant rates. Specific functional forms aid our analysis in several ways. First, they transparently reveal the core economic mechanisms at work and uncover the subtle interactions among them. Second, they enable us to easily compare different economic parametrizations, reflecting a range of estimates on agglomeration economies and diseconomies in the literature. They provide a unified framework within which to revisit findings in urban and public economics, such as how efficient population allocations equalize social marginal returns across space. The simplified but varied policy parameters allow us to incorporate competing effects of federal taxation, payments to land, benefits to owner-occupied housing, and congestion charges, all of which can vary tremendously across urban systems in different countries. The presence of the extensive margin of urbanization also makes our model relevant for applications to developing countries, where both the growth of existing cities and the formation of new cities are important drivers of current urban change.

[6]On his assumption of a continuum of traders, Aumann (1964, p.41) writes: "The idea of a continuum of traders may seem outlandish to the reader. Actually, it is no stranger than a continuum of prices or of strategies or a continuum of 'particles' in fluid mechanics." We assert that a continuum of cities is conceptually no stranger than a continuum of traders.

2.2 *Urban production, site heterogeneity, and agglomeration*

Having distilled the intuitions of our key results, we now more fully lay out the model. The economy comprises a mass $\overline{N}$ of homogeneous worker-households, or agents, to be allocated in a spatial economy. The economy is made of various cities with a total urban population of N, and a rural area, with a population N^R. By definition, $\overline{N} \equiv N + N^R$. Sites that can host cities are given a measure of one, and are *heterogeneous* in that they are not all equally amenable to urban production.[7] More precisely, we denote by $a \in A \equiv [\underline{a}, \overline{a}] \subseteq \mathbb{R}_+$ the local exogenous production amenity. It is distributed according to the twice differentiable cumulative distribution function $G(\cdot)$ over the interval A, which may be unbounded. The (largely unexplored) extensive margin of urban systems is important for the analysis when the worst developed site is better than $\underline{a}$. In that case, there is room to develop additional cities. Note that our analysis does not require an upper bound, i.e., $\overline{a} = \infty$, accommodating many distributions, such as the Pareto and Log-normal. While differentiability of the distribution rules out mass points — as implied by Figure 2 — our results cover multi-peaked distributions and the limiting homogeneous site case in Figure 1 where $\underline{a} = \overline{a}$ so that A reduces to a singleton.

In what follows, we refer to a city with productivity a as 'city a' for short. The population of a single city is denoted n. Each city is small relative to the size of the economy, and acts as a price-taker in output and input markets. The gross output of a worker in city a is given by $w(n, a) = an^\epsilon$, where n^ϵ is a scale effect external to the representative firm, parameterizing the elasticity of output with respect to city size. We are agnostic about the precise microeconomic foundations of these *agglomeration economies*.[8] The representative firm uses labor only, is perfectly competitive, produces under constant returns to scale, and trades freely across cities. This makes the traded good a natural choice as the numéraire. Total city production is given by $an^{1+\epsilon}$, implying a social marginal product of $(1+\epsilon)an^\epsilon$. The firm only captures the average social product, an^ϵ.

2.3 *Urban costs and land*

Urban dwellers work and consume two goods: the traded good, x, and land, l. One unit of land is essential and utility increases linearly in the consumption of x. Thus, utility is $u(l, x) = \mathbb{I}(l)x$, where $\mathbb{I}(l) = 1$ if $l \geq 1$ and $\mathbb{I}(l) = 0$ if $l < 1$. Hence, in equilibrium, utility is $u(x) = x$.

[7] We can extend the model to allow for sites that differ in (additive) amenities and congestion cost levels. As this is not necessary to our main argument, we alleviate notation by focusing just on heterogeneous sites in terms of productivity. Results with heterogeneous quality-of-life are available upon request.

[8] Agglomeration economies include knowledge spillovers and human capital externalities that are the engine of modern economic growth (Lucas, 1988; Romer, 1990). Duranton and Puga (2004) survey a wide class of models that deliver this reduced form via sharing, matching, and learning mechanisms. The evidence for agglomeration economies is surveyed in Combes and Gobillon (2015).

Costly commuting and scarce land mean that urban costs increase with city size.[9] We assume that total urban costs are given by $n^{1+\gamma}$, where the parameter $\gamma > 0$ characterizes urban diseconomies of scale. The difference, γn^{γ}, between social marginal costs $(1+\gamma)n^{\gamma}$ and social average costs n^{γ} provides the per-capita (differential) shadow-price of land in the city.[10] We impose $\gamma > \epsilon$ in order to ensure that urban costs come to dominate agglomeration economies as cities grow large. This implies that optimal and equilibrium city sizes are finite.

Taking the social product net of urban costs, we characterize the social average benefit as $sab(n,a) = an^{\epsilon} - n^{\gamma}$, and the social marginal benefit as:

$$smb(n,a) = (1+\epsilon)an^{\epsilon} - (1+\gamma)n^{\gamma}. \tag{1}$$

2.4 Land ownership and federal taxation

To determine equilibrium allocations, an important (and often hidden) assumption addresses who claims land rents generated in the city. We let ρ define the share of the land value γn^{γ} that a migrant rents or purchases to inhabit the city. Absent (uninteresting) income effects, it is innocuous to assume that land sales or rent payments accrue to the federal government, and are rebated lump-sum. The case of $\rho = 1$ is standard in a Roback (1982) equilibrium, while $\rho = 0$ is the most frequent assumption in the optimal-city size literature. As migrants are rarely given land in the location they move to, a value of ρ closer to one seems realistic for modeling migration. A lower value may be justified if property rights for land are weak — whereby migrants 'squat' on land, as in many developing countries (Jimenez, 1984) — or if land rents are collected through local property taxes at a rate $1-\rho$, and redistributed through perfectly

[9] We follow the seminal work of Alonso (1964), Muth (1969), and Mills (1967) in taking urban costs as a combination of costly commuting and competition over accessible land (see Duranton and Puga, 2015, for a modern synthesis). It is restrictive, yet standard in the literature, to assume that all urban diseconomies are related to land values. Pollution and noise are, for example, pure diseconomies that affect the city as a whole but do not directly show up in land values (other than by influencing city size via migration).

[10] The Alonso-Muth-Mills monocentric model is the classic way to deliver such urban costs (Fujita, 1989; Duranton and Puga, 2015). Our expression for urban costs fixes a choice of units. Consider a radial monocentric city, with radius $\sqrt{n/(d\Theta)}$, where d is density and Θ is the arc of expansion. Set per-unit commuting costs as t, and assume that commuting costs at distance z from the central business district are $tz^{2\gamma}$. Land rent leaves agents indifferent between locations within a linear city. The differential land rent — normalizing land values at the city fringe to zero — is given by $R(z) = t\{[n/(d\Theta)]^{\gamma} - z^{2\gamma}\}$. This implies Aggregate Urban Costs, Aggregate Land Rent, and Aggregate Commuting Costs of:

$$\text{AUC} = \frac{t}{d^{1+\gamma}\Theta^{\gamma}} n^{1+\gamma}, \quad \text{ACC} = \frac{1}{1+\gamma}\text{AUC} \quad \text{and} \quad \text{ALR} = \frac{\gamma}{1+\gamma}\text{AUC},$$

with ALR + ACC = AUC. We set $t = d^{1+\gamma}\Theta^{\gamma}$ by choice of units to obtain our expression for aggregate urban costs. The normalization of a in numéraire production causes it to be in proportion to such transportation costs.

rival public goods.[11] A key observation is that the purchase of land, $\rho\gamma n^\gamma$, is a private cost to migrants but, unlike commuting, is not a social cost. They are a transfer to another party whose receipt of the transfer does not depend on her residence, making them a positive fiscal externality in the urban system.

We assume the federal government taxes nominal wages at a uniform rate $\tau \leq 1$.[12] We further assume that urban costs are discounted at a uniform rate $\delta \leq 1$. This accounts for how housing and commuting both receive implicit (and sometimes explicit) subsidies, e.g., from the non-taxation of commuting time and implicit rental income of owner-occupiers. We use a single rate, noting that discount rates to commuting are often similar to those for land.

We assume that all federal tax and land revenues are rebated lump-sum, so that everyone receives a net payment T. These payments are independent of location, although they could be indexed. A benefit of our closed system — where all fiscal revenues are redistributed — is that all social benefits are internal to the system. Our normative results hence do not depend on how welfare weights are assigned to absentee landlords or others.

2.5 Externalities within and across cities

Assembling the ingredients laid out in the foregoing, the utility an individual receives from residing in city a of size n is equal to

$$u(n,a) = (1-\tau)an^\epsilon - (1-\delta)n^\gamma - \rho(1-\delta)\gamma n^\gamma + T \equiv pab(n,a) + T. \tag{2}$$

The first term is the after-tax wage; the second, the after-discount average commuting cost; the third, the after-discount land payment; and the fourth, a uniform rebate given lump-sum to each urban resident. The sum of the first three terms in (2) is the *(gross) private average benefit* (denoted pab) of residing in city a of size n. The transfer transforms this into a net measure and does, by definition, not affect location choices. It is determined by the entire urban system.

Externalities per capita are found by differencing the social marginal benefit of being in city a from the private average benefit of being there:

$$smb(n,a) - pab(n,a) = \underbrace{\epsilon an^\epsilon - \gamma n^\gamma}_{smb-sab} + \underbrace{\tau an^\epsilon + \rho(1-\delta)\gamma n^\gamma - \delta n^\gamma}_{sab-pab}.$$

[11] If squatters tend to live near the urban fringe, a positive value of ρ may still be justified. The literature on the GHV Theorem is in fact predicated on values of $\rho = 0$ and perfectly non-rival public goods. Yet, most public services such as schools, roads, and police seem largely rival, in accordance with a central assumption of Tiebout (1956). In the presence of local politics, we may alternatively interpret ρ as the political strength of specific special interest groups, as discussed in subsection 3.2.

[12] We follow Albouy (2009) in assuming that the progressivity of the tax schedule is a secondary concern for earners. The numeric model of Eeckhout and Guner (2015) pursues the importance of tax progressivity and takes a different stance on landownership.

The first term, $smb - sab$, expresses the standard *urban externalities* from agglomeration and congestion within cities, independent of policies. This is the wedge considered in most of the literature, and our approach acknowledges how it varies with amenities a and population n.

The second term, $sab - pab$, expresses *fiscal externalities* across cities due to federal taxes and land payments, net of discounts. This externality increases with n for most standard values, meaning that urban growth provides positive externalities to the economy. It exists even without explicit federal policy, so long as migrants must purchase some land, i.e., if $\tau = \delta = 0$ and $\rho > 0$.[13]

Cities in isolation exhibit positive but finite efficient scales. The fiscal parameters create a wedge between the sizes, n_i and n_p, that maximize social and private average benefits, respectively (see Figure 1). Some basic properties of efficient city scales that will be useful in what follows are summarized in the following lemma:

Lemma 1 (Properties of efficient city scales) *In the social (equation* (1)*) and private (equation* (2)*) frameworks above, cities exhibit unique social and private efficient scales, $n_i(a)$ and $n_p(a)$, determined by $smb\left[n_i(a), a\right] = sab\left[n_i(a), a\right]$ and $pmb\left[n_p(a), a\right] = pab\left[n_p(a), a\right]$, respectively. These scales exhibit the following properties:*

(i) Population size: *The private efficient scale is a multiple of the social efficient scale. More precisely:*

$$n_i(a) = \left(\frac{\epsilon}{\gamma}a\right)^{\frac{1}{\gamma-\epsilon}} \quad \text{and} \quad n_p(a) = n_i(a)\Phi^{\frac{1}{\gamma-\epsilon}}, \quad \text{where} \quad \Phi \equiv \frac{1-\tau}{(1-\delta)(1+\rho\gamma)} > 0. \tag{3}$$

(ii) Utility benefits: *Efficient scales yield benefits $sab_i(a) \equiv sab[n_i(a), a]$ and $pab_p(a) \equiv pab[n_p(a), a]$ given by*

$$sab_i(a) = \left(\frac{\gamma}{\epsilon} - 1\right)[n_i(a)]^{\gamma} \quad \text{and} \quad pab_p(a) = (1-\delta)(1+\rho\gamma)\left(\frac{\gamma}{\epsilon} - 1\right)[n_p(a)]^{\gamma}. \tag{4}$$

(iii) Relative size: *The elasticity of both efficient scales to site productivity, a, are constant, positive, and equal to $1/(\gamma - \epsilon)$. Using hat notation, $\widehat{x} = dx(a)/x$, we have*

$$\widehat{n}_i = \widehat{n}_p = \frac{1}{\gamma - \epsilon}\widehat{a}, \quad \text{and} \quad \widehat{sab}_i = \widehat{pab}_p = \frac{\gamma}{\gamma - \epsilon}\widehat{a} > 0. \tag{5}$$

Proof The proof is immediate from $smb\left[n_i(a),a\right] = sab\left[n_i(a),a\right]$ and from $pmb\left[n_p(a),a\right] = pab\left[n_p(a),a\right]$, and by using the definitions in the text. □

Part (i) of Lemma 1 establishes that efficient scales naturally increase with site quality a. They also increase with agglomeration economies, ϵ, and decrease with urban diseconomies, γ.

[13] If there are specific interest groups, such as incumbent homeowners, and if city size is determined by a political voting process, $1 - \rho$ may be viewed as the share of local voters who (may) benefit from higher land values, as in Fischel's (2002) 'homevoter hypothesis.' Under this interpretation, the larger is ρ, the more powerful are migrants relative to incumbents, and the smaller is the private efficient scale for a city.

The ratio of these elasticities, γ/ϵ, equals the ratio of average urban benefits to costs, either an^{ϵ}/n^{γ} or $\Phi an^{\epsilon}/n^{\gamma}$. The parameter bundle Φ — which collects fiscal and landownership parameters — reflects the private benefit-to-cost ratio relative to the social one. In the typical developed-world case where $\Phi < 1$, the private efficient scale, $n_p(a)$, falls short of the socially efficient scale, $n_i(a)$. Part (ii) shows that benefits are a multiple of urban costs, n^{γ}. Finally, part (iii) shows that n_p and u_p increase at constant rates in a: better sites host larger cities and offer larger benefits to their residents.

2.6 The non-urban sector

Our model features an extensive margin of urbanization by allowing for an endogenous number of cities. Yet, if all agents are assumed to live in cities, then the model is silent on the issue of 'urbanization' in general, which is especially important when thinking of developing countries. We introduce a rural sector in order to address the issue of *overall urbanization*, and to close the model in a general fashion. The rural population, $N^R \equiv \overline{N} - N$ produces the traded good (from agriculture) using a concave technology $X^R = F(N^R)$. Rural workers earn the competitive wage $w_R = F'(N^R)$. The omitted factor, agricultural land, L^R, receives the remaining product, $F - N^R F'(N^R)$. A fraction ρ^R of land rents from agriculture is collected by a 'federal land trust', while $1 - \rho^R$ is distributed to agricultural workers. Let τ^R be the tax on the agricultural wages, and T^R be a transfer. The net-of-tax income a rural migrant receives is then $u^R(N^R) = (\rho^R - \tau^R)F' + (1 - \rho^R)F/N^R + T^R$.

We assume weakly positive but diminishing returns of adding people to the countryside: $F'(N) \geq 0$ and $F''(N) \leq 0$ for $N \in \left[0, \overline{N}\right]$. For concreteness, assume the agricultural production function takes a constant-elasticity-of-substitution (CES) form:

$$X^R = a^R N^R \left[(1-\alpha) + \alpha (L^R/N^R)^{(1-\sigma)/\sigma}\right]^{\sigma/(1-\sigma)}, \tag{6}$$

where $a^R \geq 0$, $\alpha \geq 0$, and $\sigma \geq 0$ are scale, distribution, and substitution parameters, respectively. Two important limiting cases are covered with $\rho_R = 1$. The first case is when the rural population is fixed, corresponding to perfect complementarity, i.e., $\sigma \to 0$. The second is that of an outside option u_R which is of a constant value, corresponding to perfect substitution $\sigma \to \infty$. Imperfect substitution between the consumption of rural and urban goods may also be subsumed in this elasticity. In general, the rural sector provides a supply curve of urban population, and may be as elastic as circumstances merit.

3. Three urban allocations

We now turn to the allocation of people between the urban system and the rural area, as well as their allocation within the urban system. The question is to determine: (i) the rural-urban

population split (the *extent of urbanization*); (ii) which sites host cities (the *extensive margin* of urban development); and (iii) how many urban dwellers are allocated to each of these cities (the *intensive margin* of urbanization). We characterize three different allocations:

1. The *centralized optimum (indexed with $*$)*, i.e., the one that a federal central planner would choose, taking into account site heterogeneity. This allocation can be implemented, under certain conditions, by *competitive land developers*;

2. The *local politics allocation (indexed with p)*, i.e., when city sizes are chosen at the city level by uncoordinated local governments who can restrict in-migration; and

3. The *free migration equilibria (indexed with m)*, i.e., when households make individual location decisions based on private incentives. Although there are multiple equilibria in that case, we focus on a constrained-efficient equilibrium.

In what follows, the subscripts for the centralized optimum, the local politics allocation, and free-migration equilibria are given by '$*$', 'p', and 'm', respectively. Each of these three allocations $z \in \{*, p, m\}$ is a mapping $n_z : a \in A \to \mathbb{R}_+$ that satisfies the population adding-up constraint:

$$N = \int_{\underline{a}}^{\overline{a}} n_z(a) \mathrm{d}G(a) \quad \text{and} \quad N^R + N = \overline{N}. \tag{7}$$

3.1 Centralized optimum allocation

The federal planner's problem is to maximize aggregate consumption of the numéraire good in the spatial economy by setting the intensive, $n_*(a)$, and extensive, a_*, margins of the urban system, as well as the degree of urbanization, N_*. All agents have a constant and identical marginal utility of income. Hence, utility is transferable, and uniform transfers do not affect location (Mirrlees, 1982). The problem of the federal planner is to optimize the following Lagrangian:

$$\mathcal{L} \equiv F(N^R) + \int_{a_*}^{\overline{a}} n(a) \left[a n(a)^{\epsilon} - n(a)^{\gamma} \right] \mathrm{d}G(a) + \mu \left[\overline{N} - N^R - \int_{a_*}^{\overline{a}} n(a) \mathrm{d}G(a) \right]. \tag{8}$$

The first-order condition with respect to μ yields the population adding-up constraint in equation (7). The first-order condition for the optimal rural population, N^R, is given by

$$\mu_* \geq F'(N_*^R), \quad N_*^R \geq 0, \tag{9}$$

while the first-order conditions for the optimal city sizes, $n(a)$, are characterized by

$$smb[n_*(a), a] \equiv a(1+\epsilon) n_*(a)^{\epsilon} - (1+\gamma) n_*(a)^{\gamma} \leq \mu_*, \quad n_*(a) \geq 0. \tag{10}$$

Equation (10) states that the social marginal benefit of residing in any city must be equal across all occupied sites (Flatters et al., 1974). It equals μ_* — the Lagrange multiplier evaluated at the (highest) optimal value — which itself equals the marginal benefit in the rural area from (9). Each case has complementary slackness. While complementary slackness is not crucial in (9) because we assume that there is an interior rural-urban split, it is more interesting and important in (10) since not all sites need to develop cities. The first-order condition for the optimal extensive margin of urban development, a_*, yields

$$\mu_* = a_* n_*(a_*)^\epsilon - n_*(a_*)^\gamma \leq \left(\frac{\gamma}{\epsilon} - 1\right) n_i(a_*)^\gamma = u_i(a_*). \tag{11}$$

In addition, observe that by the envelope theorem,

$$\mu_* = \frac{\partial}{\partial \overline{N}} \mathcal{L}\left(n_*(a), \mu_*\right) \tag{12}$$

holds at the optimal allocation. In words, the social marginal value of population equals both agricultural productivity and urban productivity net of urban costs.

We may now show our first set of results describing the centralized optimum allocation:

Lemma 2 (Structure of the centralized optimum allocation) *There exists a unique solution to equations* (9), (10), *and* (11), *which characterizes the optimal allocation. In particular:*

(i) Extent of urbanization: *There exists a unique urban population size $N_* \in [0, \overline{N}]$;*

(ii) Urban extensive margin: *There exists a unique threshold $a_* \in A$ such that*

$$N_* = \int_{a_*}^{\overline{a}} n_*(a) \mathrm{d}G(a), \quad n_*(a) > 0 \quad \text{for all} \quad a \geq a_*, \quad \text{and} \quad n_*(a) = 0 \text{ otherwise;}$$

(iii) Minimum city size: *the optimal size of the smallest city is no smaller than its efficient scale, i.e.,*

$$n_*(a_*) = k_* \left(\frac{\epsilon}{\gamma} a_*\right)^{\frac{1}{\gamma - \epsilon}} = k_* n_i(a_*), \quad k_* \geq 1, \tag{13}$$

with $k_* = 1$ if $a > \underline{a}$. The worst site developed is at its efficient scale unless all sites are occupied;

(iv) Urban intensive margin: *For all $a > a_*$, the optimal city size $n_*(a)$ is increasing in a, with*

$$\frac{\widehat{n}_*}{\widehat{a}} = \frac{1}{\gamma} \frac{\Gamma}{\left(n_*/n_i\right)^{\gamma - \epsilon} - \Phi_*} > 0, \quad \text{where} \quad \Gamma \equiv \frac{\gamma}{\epsilon} \Phi_* \quad \text{and} \quad \Phi_* = \frac{1+\epsilon}{1+\gamma}; \tag{14}$$

(v) Implicit solution: *The optimal solution $n_* = n_*(a)$ is implicitly determined by the equation*

$$\left(\frac{a}{a_*}\right)^{\frac{\gamma}{\gamma - \epsilon}} = \frac{\Gamma k_*^\epsilon - k_*^\gamma}{\Gamma \left(n_*/n_i\right)^\epsilon - \left(n_*/n_i\right)^\gamma}, \tag{15}$$

for $n_*/n_i \in \left(1, \Gamma^{1/(\gamma - \epsilon)}\right)$. For all $a > a_*$, $n_*(a)$ is (weakly) larger than its efficient scale, $n_i(a)$, and lower than the size with zero marginal net production: $n_0(a) \equiv (a\Phi_*)^{1/(\gamma - \epsilon)}$.

Proof See Appendix A.1. □

Part (i) of Lemma 2 first establishes that there is a unique rural-urban population split that pins down the extent of urbanization. Part (ii) states that there is a minimal site quality, a_*, with all inferior sites undeveloped. Unless all sites are occupied, the inferior site is at its efficient scale $n_i(a_*)$ in (3) by part (iii), in analogy with producer theory. Part (iv) establishes the relative size of cities on sites of quality superior to a_*, expressed in elasticity form. It also shows that, as expected, city size is increasing in site quality a. Last, the implicit solution for $n_*(a)$ is seen in (v), expressed as a ratio to $n_i(a)$ as introduced in Lemma 1.

Figure 3: Centralized optimum city sizes and socially efficient scales.

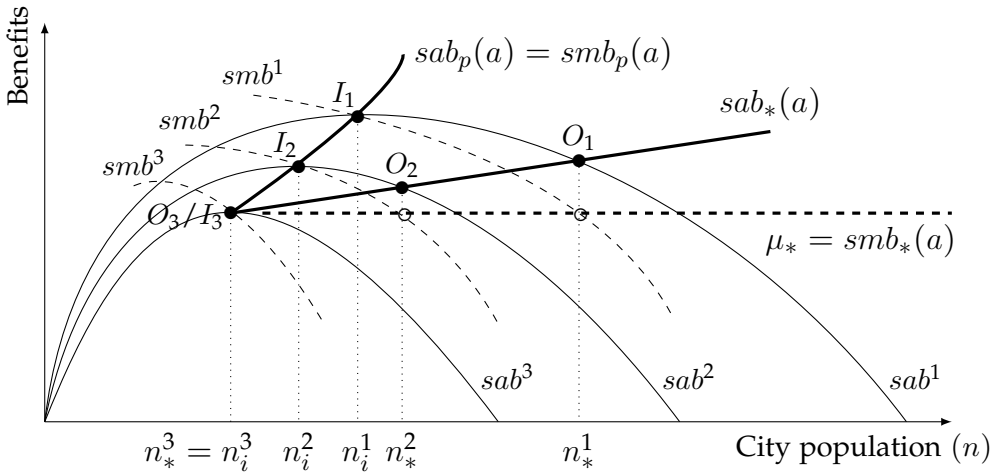
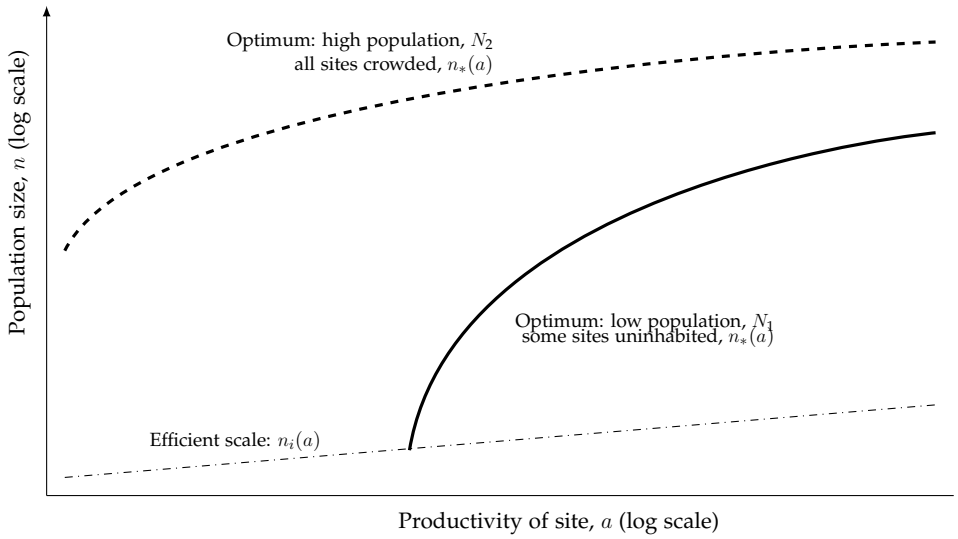


Figure 3 illustrates properties of an optimal allocation with productivity $a_1 > a_2 > a_3$. For each city, j, the social marginal benefit curve, smb^j, intersects the average benefit curve, sab^j, at the social efficient scales n_i^j. The optimal sizes, n_*^j with equal social marginal benefits are (weakly) larger than n_i^j. Furthermore, the gap between n_i^j — the top of the $\cap$-curves — and n_*^j — which equalizes the social marginal benefits across cities — increases with a: *the agglomeration distortion is worse for better sites*. In the centralized optimum allocation, superior sites are pushed beyond their efficient scale to benefit outsiders. From an isolated point-of-view, everyone believes their city is "too big" — the more so the bigger is the city — although globally it is not.

Figure 4 further illustrates properties of the optimum. The socially efficient scale, $n_i(a)$, increases with a at a constant rate of $1/(\gamma - \epsilon)$. The optimal population, $n_*(a)$, increases at a much greater, albeit declining rate. At reasonably low population levels, N_1, some sites are uninhabited. As N rises, more sites are inhabited, and populations on all inhabited sites rise: urbanization proceeds along the intensive and extensive margins. With a very high urban population $N_2 > N_1$, all sites become occupied, and all are crowded beyond $n_i(a)$, as the extensive margin is shut down. Observe that the elasticity of city size to amenities a falls

Figure 4: Optimal populations, $n_*(a)$, and productivity, a, with and without an extensive margin.



throughout the urban system as worse sites are progressively put into use. As the urban system runs out of sites, the city size distribution tends to become more even.

We now turn to the welfare properties of the optimal allocation.

Proposition 1 (Welfare in the centralized optimum allocation) *The normative properties of the optimal allocation characterized by equations* (9), (10), *and* (11) *when* $a_* > \underline{a}$ *are the following:*

(i) Urban benefits: *For* $a \geq a_*$, *the social marginal benefit is constant at*

$$smb\left[n_*(a), a\right] = \mu_* = u_i(a_*) = F'(N_*^R), \tag{16}$$

while social average benefits increase with site quality a *as follows:*

$$sab_*(a) \equiv sab\left[n_*(a), a\right] = \mu_* \frac{1}{1+\epsilon} \frac{\Gamma - \Phi_* \left(n_*/n_i\right)^{\gamma-\epsilon}}{\Gamma - \left(n_*/n_i\right)^{\gamma-\epsilon}} \geq \mu_*, \quad sab'_*(a) \geq 0;$$

(ii) Decreasing returns of the economy: *The economy as a whole features decreasing returns with respect to population*

$$\frac{\mathcal{L}_*}{\overline{N}} > \frac{\partial \mathcal{L}_*}{\partial \overline{N}} = \mu_*;$$

(iii) Decreasing returns of the urban system: *For any given rural population* N^R, *the urban system exhibits decreasing returns, i.e., the social marginal benefit of urban dwellers is below the social average benefit. The decreasing returns occur at both the intensive and the extensive margins:*

$$\frac{\mathrm{d}a_*}{\mathrm{d}N} < 0 \quad \text{and} \quad \frac{\mathrm{d}n_*(a)}{\mathrm{d}N} > 0 \quad \text{for all} \quad a > a_*.$$

(iv) Generalized Goerge-Hotelling-Vickrey Theorem: *The ratio of the value of urban land to marginal agglomeration externalities is weakly greater than one and equal to*

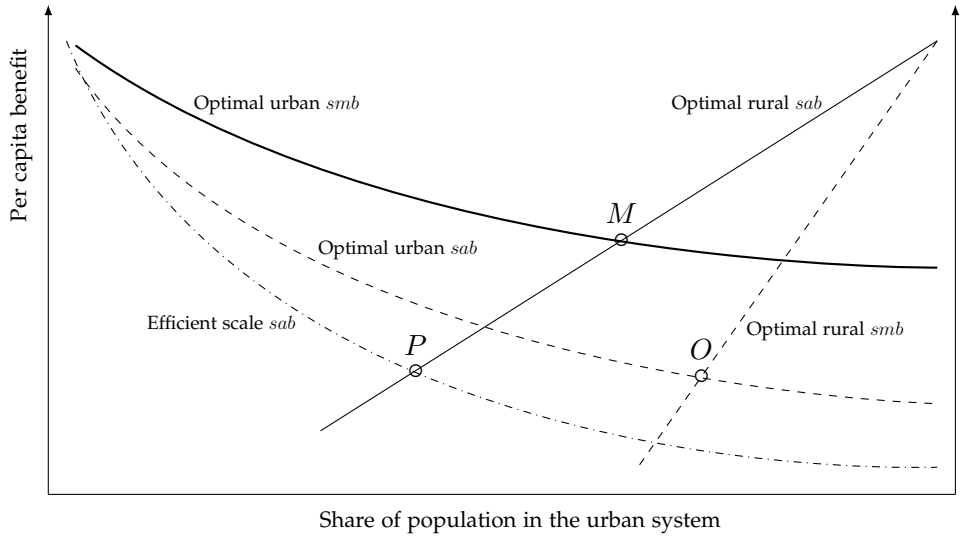
$$\bar{v}_* \equiv \frac{\gamma \int_{a_*}^{\overline{a}} n_*^{1+\gamma} dG(a)}{\epsilon \int_{a_*}^{\overline{a}} a n_*^{1+\epsilon} dG(a)} = 1 + \frac{\bar{w}_* - w_*(a_*)}{\bar{w}_*} \left(\Gamma - 1\right) \tag{17}$$

where $\bar{w}_*$ is the average urban wage, and $w_*(a_*)$ is the lowest urban wage.

Proof See Appendix A.2. □

Part (i) of Proposition 1 establishes that urban benefits and congestion increase in a, as described in Figure 4. Parts (ii) and (iii) establish that there are decreasing returns to the economy in general — and to the urban system in particular — for all values of $\overline{N}$ and N, which is illustrated in Figure 5. This bucket diagram plots the urban share of the population, $N/\overline{N}$, as the distance from the left axis and the rural share, $N^R/\overline{N}$, as the distance from the right. Social average and marginal benefits of the urban system fall with the urban share, while rural average and marginal benefits rise, as derived from the CES function (6). The intersection of urban and rural marginal benefit curves determines the optimal degree of urbanization, N_*, and the marginal benefit of urbanization, μ_*.

Figure 5: Benefits of an optimal urban system, and the extent of urbanization.



Lastly, (iv) provides a generalization of the GHV Theorem based on potentially observable wage differences. When cities are homogeneous, the average and lowest urban wage are the same, so that the ratio $\bar{v}_* = 1$, meaning that urban land values equal the value of urban agglomeration externalities. When cities are heterogeneous, the average wage is higher than the lowest wage, and land values exceed urban agglomeration benefits, by a ratio as high as Γ.[14]

3.2 Local politics allocation

Consider next the allocation that arises if city governments maximize the average utility of their residents, ignoring the consequences of their choices for potential migrants. Local authorities

[14] Another theory that makes use of the mean/min wage ratio is Hornstein et al. (2011) in their analysis of wage dispersion in search models.

have the power to exclude people either directly — by using urban growth boundaries and other controls — or indirectly — by using land-use regulations that impose a 'regulatory tax' on potential newcomers (Glaeser, Gyourko, and Saks, 2005). Then, local authorities expand their city only as long as the benefits of doing so outweigh the costs, meaning that they choose the privately efficient scale, $n_p(a)$, described in Lemma 1.

We assume that only the best sites are populated in the local politics allocation, namely, there exists $a_p \in A$ such that $n_p(a) > 0$ if and only if $a \geq a_p$ and $n_p(a) = 0$ otherwise. With migration limited, cities offer different returns, and if all sites are occupied (i.e., $a_p = \underline{a}$), the rural sector may offer a lower return than the worst city. The following conclusions ensue:

Proposition 2 (Normative properties of the local politics allocation) *Assume that local governments maximize the average utility of their residents. If some sites are unoccupied at the optimum, i.e., $a_* > \underline{a}$, then:*

(i) *If $\Phi < 1$, large cities are undersized and there are too many cities. The excess small cities are oversized by virtue of existing;*

(ii) *If $\Phi = 1$, the optimum is achieved if sites are homogeneous. With site heterogeneity, there are too many cities, with large cities being undersized and small cities being oversized;*

(iii) *If $1 < \Phi < \Gamma$, there are too few cities that are all oversized if sites are homogeneous. With site heterogeneity, small cities are oversized, there are (generically) too many or too few cities, and large cities are undersized if $\overline{a} < \infty$.*

(iv) *If $\Phi \geq \Gamma$, there are too few cities and all cities are oversized.*

(v) Urban benefits: *In all cases, urban benefits increase with a as a constant multiple of urban costs. Private average benefits increase with a according to $pab_p(a)$ defined in* (4)*.*

(vi) Land values: *the ratio of land values to urban agglomeration externalities is $\bar{v}_p = \Phi$.*

Proof See Appendix A.3. □

Several comments are in order. First, Proposition 2 is stated for the case where some sites are left unoccupied. If all sites are occupied at the optimum, $a_* = \underline{a}$, then Proposition 2 still holds if we replace 1 with $k_*^{1/(\gamma-\epsilon)} < 1$ in the cases above. In words, when all sites are occupied, larger cities are quite naturally less likely to be undersized than when some sites are left vacant.

Second, in the absence of fiscal externalities, $\Phi = 1$, local politics allocate the efficient scale to each city, i.e., $n_i(a) = n_p(a)$. With homogeneous sites, local politics causes the original GHV conditions to hold, as $\bar{v}_* = 1$. But with heterogeneous sites, local politics causes land values to be too low, as $\bar{v}_* > 1$. At their minimum efficient scale, $n_i(a)$, large cities with $a > a_*$ are undersized, while sites with $a = a_*$ are just the right size. The excess population due to restricted entry into larger cities is put into sites that would not otherwise exist, and those cities are therefore 'oversized' by virtue of existing. Thus site heterogeneity is sufficient to overturn

a central result in urban economics. This is generally made worse with the extensive margin or fiscal externalities.

Which case of Proposition 2 is the most plausible one? The case where $\Phi < 1$ appears to describe modern OECD economies, where tax and land payments resulting from agglomeration are high relative to congestion discounts. This then reduces the private benefit of urbanization below its social value. In that case, the private efficient scale is below the social one, i.e. $n_p(a) < n_i(a)$, and thus by transitivity, the local politics population is lower than the optimal one, $n_*(a)$. Additional population is then pushed out onto inferior sites, which are over-sized by virtue of existing. If $1 < \Phi < \Gamma$, local incentives to stay small on the best sites are dominated by generous fiscal benefits. The incentives to 'go big' only dominate for the few best sites, whereas the bulk of the other sites remains too small. For $\Phi > \Gamma$, an unlikely case in any economy, cities are always too big.

As Figure 3 shows in the simplest case of $\Phi = 1$, i.e., no externalities, the $sab_p(a) = pab_p(a)$ schedule is much steeper than the $sab_*(n)$ schedule of the optimum. Local governments at the best sites keep out migrants to preserve benefits for their constituents, thereby leading to undersized large cities and a proliferation of cities on inferior sites. These types of NIMBY-istic policies used to control local populations appear commonplace. They can take the form of urban containment in some North-American cities such as Portland, Oregon, and Vancouver, British Columbia, and in the United Kingdom (Cheshire and Sheppard, 2002), or of restrictive land use regulations (Glaeser et al., 2005; Hilber and Robert-Nicoud, 2013). The normative implication of the model is that policies that heavily restrict urban development should not be designed by local authorities alone because they fail to internalize the benefits of these policies to outsiders.[15]

3.3 Free-migration equilibria

The other important, and often more realistic, case is that of decentralized free-migration equilibria. In that case, migrants move to the city that offers the highest utility. Consequently, with identical households, utility is the same across all inhabited sites. Formally, the mobility condition is

$$pab(n, a) = \mu_m = u_m - T_m \tag{18}$$

for all sites $a \in A$ with $n_m(a) > 0$, and $u_m \geq u^R(N_m^R)$. Equilibria must satisfy the additional requirement of each city's population being stable, which is guaranteed if $\partial pab(n, a)/\partial n \leq 0$ and $\partial u^R(N^R)/\partial N^R < 0$. The foregoing conditions imply that urban population levels in a free-migration equilibrium are at least as large as with local politics, $n_m \geq n_p$. Without further refinement, u_m may take different values, corresponding to different free-migration equilibria.

[15] Vermeulen (2016) reaches a similar conclusion in a very different setup.

While multiple equilibria are interesting, we focus on a constrained-efficient free-migration equilibrium, and compare it to the optimum. Besides free mobility, the additional constraint this equilibrium imposes, is that households will seek out new sites if they can profit from them by later selling the land to other migrants.[16] It produces the smallest stable cities in a free-migration equilibrium, characterized below.

The constrained-efficient free-migration equilibrium is characterized by the following lemma, which mirrors Lemma 2, with equations (9) and (11) being replaced by

$$u_m \geq u^R(N_m^R), \quad N_m^R \geq 0 \tag{19}$$

(with complementary slackness) and

$$u_m = (1-\tau)a_m n_m(a_m)^\epsilon - (1-\delta)(1+\rho\gamma)n_m(a_m)^\gamma, \tag{20}$$

respectively:

Lemma 3 (Structure of the constrained-efficient free-migration allocation) *There exists a unique solution to equations* (18), (19), *and* (20) *which characterizes the free-migration allocation. In particular:*

(i) Extent of urbanization: *There exists a unique urban population size* $N_m \in \left[0, \overline{N}\right]$*;*

(ii) Urban extensive margin: *There exists a unique threshold* $a_m \in A$ *such that*

$$N_m = \int_{a_m}^{\overline{a}} n_m(a)\mathrm{d}G(a), \quad n_m(a) > 0 \quad \textit{for all} \quad a \geq a_m, \quad \textit{and} \quad n_m(a) = 0 \quad \textit{otherwise;}$$

(iii) Minimum city size: *The equilibrium size of the smallest city is no smaller than its private efficient scale*

$$n_m(a_m) = k_m \left(\Phi \frac{\epsilon}{\gamma} a_m\right)^{\frac{1}{\gamma-\epsilon}} = k_m n_p(a_m), \quad k_m \geq 1, \tag{21}$$

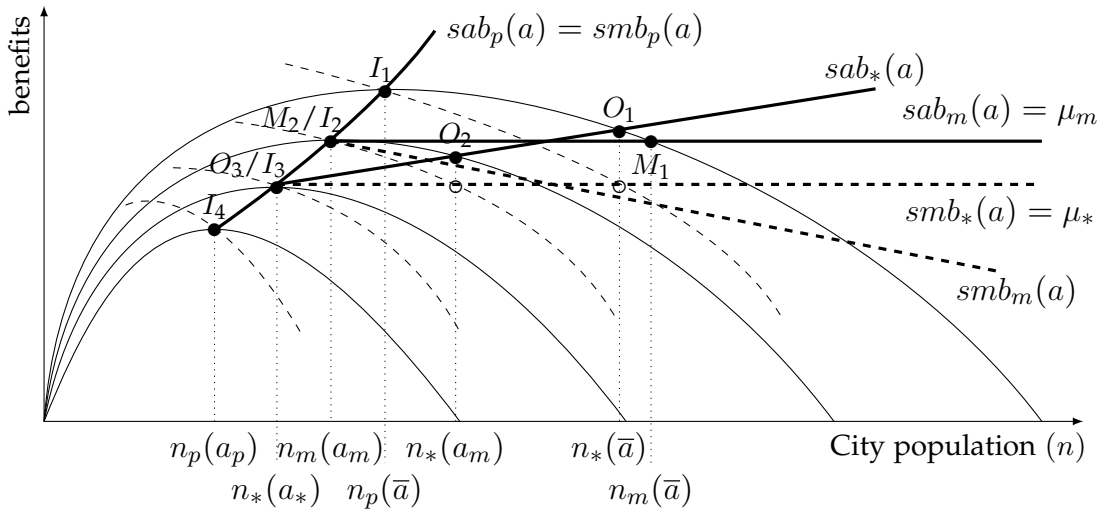
with $k_m = 1$ *if* $a_m > \underline{a}$*, i.e., if some sites are uninhabited;*

(iv) Urban intensive margin: *For all* $a > a_m$*, the equilibrium city size* $n_m(a)$ *is increasing in* a*, with*

$$\frac{\widehat{n}_m}{\widehat{a}} = \frac{1}{\gamma} \frac{\frac{\gamma}{\epsilon}\Phi}{(n_m/n_i)^{\gamma-\epsilon} - \Phi} = \frac{1}{\epsilon} \frac{1}{(n_m/n_p)^{\gamma-\epsilon} - 1}; \tag{22}$$

[16] The free-migration equilibrium, which has the smallest city at its privately efficient scale, can be rationalized as the result of forward looking residents and potential developers in a multi-stage game (Seegert 2011, 2013). Potential developers do not choose smaller sizes because of stability problems, which results in no profits being made and lower welfare. They do not choose larger sizes, as any migrant is better off in a new city at $a \geq a_m$, acquiring land for free, than joining an existing city. This is an extreme equilibrium that Milgrom and Roberts (1994) suggest is most useful for comparisons. In our simulations below, assuming cities are larger by a factor of $k_m \in (1,1.5]$ does not result in substantially different results, especially for welfare. See also recent work by Henderson and Venables (2009) that discusses the dynamics of city formation without large agents and comes close to solving the coordination failure.

Figure 6: Centralized optimum, local politics, and free migration city sizes with $\Phi = 1$.



(v) Implicit solution: *The free-migration solution satisfies the equation*

$$\left(\frac{a}{a_m}\right)^{\frac{\gamma}{\gamma-\epsilon}} = \frac{\frac{\gamma}{\epsilon}k_m^{\epsilon} - k_m^{\gamma}}{\frac{\gamma}{\epsilon}(n_m/n_p)^{\epsilon} - (n_m/n_p)^{\gamma}}, \tag{23}$$

where $k_m \equiv n_m(a_m)/n_p(a_m)$ equals 1 *if $a_m > \underline{a}$. For all $a > a_*$, the equilibrium city size $n_m(a)$ is (weakly) larger than its privately efficient scale, $n_p(a)$, and lower than the size with zero marginal net production: $n_0(a) \equiv (a\Phi)^{1/(\gamma-\epsilon)}$.*

Proof See Appendix A.4. □

A free-migration equilibrium is contrasted to the optimal and local politics allocations in Figure 6, for the special case with $\Phi = 1$. Disregard sites 3 and 4 for now and consider only sites 1 and 2. The smaller city at M_2/I_2 is at its private efficient scale, i.e., $n_m(a_m) = n_p(a_m)$. City 1 offers the same sab_m as this city, which determines its population at point M_1. The marginal site with city 2 is superior to the optimal one a_*. Yet, the free-migration population at a_m is too small: $n_m(a_m) < n_*(a_m)$, seen at O_2, although larger than the smallest city at the optimum $n_*(a_*)$ at O_3. However, the best site with city 1 has a higher population than the optimum at O_1. As Figure 6 shows, local politics — by restricting city sizes — forces residents onto more numerous and inferior sites, reaching down to a_p, with low benefits at I_4 because of diminishing returns in the rural sector. With free migration, social marginal benefits fall with the productivity of the city, as better sites are increasingly over-crowded from the global (not just the local) perspective.

With large positive fiscal externalities, $\Phi < \Phi_*$, for example when the tax rate on urban wages is higher than the discount rate on urban costs, the free-migration equilibrium becomes

generically suboptimal, including the case with homogeneous sites — recall Figure 1.[17] With $\Phi < \Phi_*$, the best sites are always underpopulated, which we can see as we consider welfare:

Proposition 3 (Normative properties of the constrained-efficient free-migration equilibrium) *Consider the constrained-efficient free-migration equilibrium with $n_m(a) > 0$ for all $a \geq a_m$, $n_m(a) = 0$ for all $a < a_m$, and $n_m(a_m) = n_p(a_m)$. This equilibrium is such that:*

(i) *If $\Phi \leq \Phi_*$, cities are too numerous. Large cities are too small, whereas small cities are too large by virtue of existing;*

(ii) *If $\Phi_* < \Phi < 1$, then cities are too small and numerous if sites are homogeneous. If sites are heterogeneous, there are (generically) too many or too few cities, and large cities are over-sized if $\overline{a} < \infty$;*

(iii) *If $\Phi = 1$, then the optimum is achieved with homogeneous sites. If sites are heterogeneous, there are too few cities, and large cities are oversized;*

(iv) *If $\Phi > 1$, then large cities are oversized, and there are too few cities.*

When all sites are occupied at the optimum, $a_* = \underline{a}$, then cities are optimally sized if $\Phi = \Phi_*$; large (small) cities are too small (big) if $\Phi < \Phi_*$; large (small) cities are too big (small) if $\Phi > \Phi_*$.

(v) Urban benefits: *The private average benefit received uniformly across cities is $\mu_m \equiv pab_m(a) = pab_p(a_m)$, defined in* (4)*, while the social marginal benefits vary with a according to*

$$smb_m(a) \equiv smb[n_m(a), a] = \mu_m \frac{1+\epsilon}{1-\tau} \frac{\Gamma - \Phi \left(n_m/n_p\right)^{\gamma-\epsilon}}{\Gamma - \Phi_* \left(n_m/n_p\right)^{\gamma-\epsilon}}, \tag{24}$$

which increases with a if $\Phi < \Phi_*$.

(vi) Land values: *The ratio of land to values urban agglomeration externalities equals*

$$\bar{v}_m = \Phi k_m^{\gamma-\epsilon} + \frac{\bar{w}_m - w_m(a_m)}{\bar{w}_m} \Phi \left(\frac{\gamma}{\epsilon} - k_m^{\gamma-\epsilon}\right) \tag{25}$$

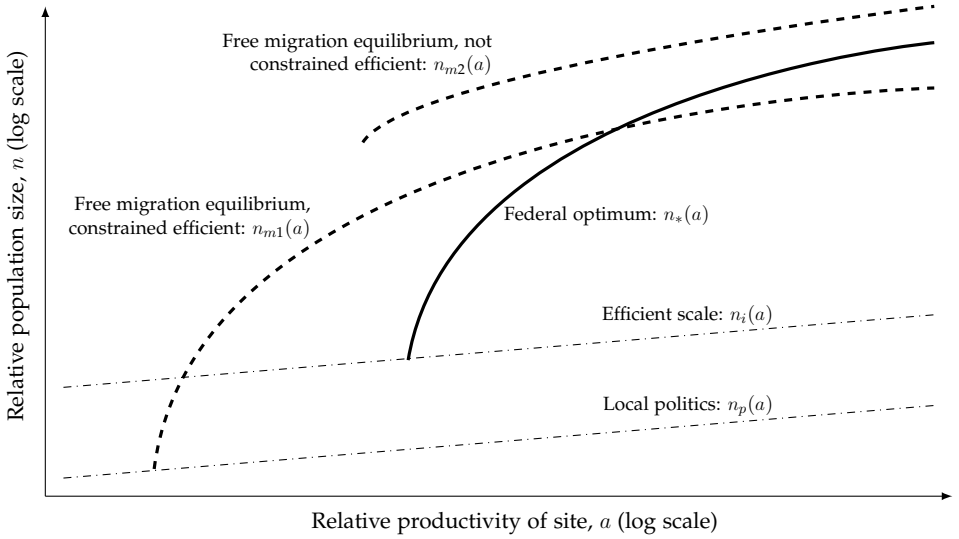
where $\bar{w}_m$ is the average equilibrium urban wage, and $w_M\left(a_m\right)$ is the lowest urban wage.

Proof See Appendix A.5. □

Relative equilibrium population sizes for case (i), when $\Phi < \Phi_*$, are covered in Figure 7. The smallest equilibrium city, a_m, is at the private efficient scale, $n_p(a)$, and below the social one, $n_i(a)$. The elasticity of size with respect to productivity is initially infinite, but then tapers off. In this case, there is a point where the equilibrium and optimal city sizes, $n_*(a)$ cross, after which city sizes are too smalll. While we argue this case holds for most OECD countries, it may exist with no explicit federal policies just from land purchases alone. When $\tau = \delta = 0$, but $\rho \geq (1 - \epsilon/\gamma)/(1+\gamma)$, large cities are underpopulated with free migration.

[17] The upper tail of the free-migration equilibrium city size distribution, $n_m(a)$, inherits the properties of the distribution of site characteristics in the same way that $n_p(a)$ does (see also Behrens and Robert-Nicoud, 2015). If very good sites are especially scarce — there are not many natural harbors of the same quality as New York's — and if those sites are developed in priority, then large cities are also scarce and the city size distribution displays properties consistent with Zipf's law.

Figure 7: Optimal, free-migration, and local politics city sizes.



The three other cases (ii), (iii), and (iv) under heterogeneity, as well as the case of all sites being populated, are illustrated further in the proof in Appendix A.5. With $\Phi = \Phi_*$, the optimum cannot be achieved since the city at a_* would be too small due to the fiscal externality. With the optimal population rate of change with respect to a, the population adding-up constraint would be violated. Therefore more cities will exist. With $\Phi = 1$ (in the absence of fiscal wedges), making the city at a_* optimally sized leads to all superior sites being too large. Therefore fewer cities will exist. When $\Phi \geq \Gamma$ (the fiscal system discounts urban costs more than it taxes wages), the optimal size is always below the private efficient scale, and very few cities exist. Finally, when all sites are occupied but $\Phi < \Phi_*$, then differences in city size are smaller than in the optimum: small cities are too large, while large cities are too small, similar to Albouy (2009).

Although we emphasize issues arising from heterogeneity and fiscal externalities, much of the work on urban systems is concerned about coordination failures related to inadequate decentralization — the so-called 'migration pathology'. The initial benefit of occupying a site in our $\cap$-shape formulation for net benefits is essentially zero. For a city to reach a stable population it must provide a threshold utility as good as the value of living elsewhere, which requires a quantum leap of population. In a system with growing N, existing cities risk becoming overcrowded if marginal (inferior) sites are not developed quickly enough. Multiple equilibria can be characterized in our formulation by setting $k_m > 1$ for a marginal site $a_m > \underline{a}$. This means some sites remain unoccupied, while even the worst-sited city is crowded beyond its private efficient scale.

The equilibrium that can arise by coordination failure of decentralized migration decisions is illustrated in Figure 7, where all cities are too big for the highest curve. While an equilibrium with smaller cities is reasonable, larger cities cannot be ruled out. Fortunately, part (vi) on

urban land values can provides us with a test of whether cities suffer from the migration pathology.

3.4 Optimal urban systems and the urban-rural fringe

The middle and bottom panels of Table 1 summarize the results of the foregoing subsections. The top panel completes the picture by emphasizing that — on top of what is going on in the urban system — urbanization itself may be excessive or insufficient. To see why this is so, recall that efficiency requires the marginal product of labor to be equal to the economy-wide social marginal benefit, i.e., $F'(N^R) = \mu_*$. When $\rho_R < 1$, i.e., rural land ownership is granted to migrants (at least partially), the rural private average benefit is greater than the social marginal benefit. A spatial allocation with free migration is generally inefficient in this case because the marginal benefit of being in the rural sector exceeds the marginal benefit of being in the urban sector. With fiscal wedges and urban externalities, the spatial allocation of agents is, in general, inefficient at the rural-urban margin. The top panel of Table 1 covers the various possibilities.

Table 1: Comparisons of allocations with the centralized optimum when $a_* > \underline{a}$.

		$\rho = 1$	$\rho < 1$		
Urbanization.					
$u_R(N^R) < \mu_*$		Too little	Too little		
$u_R(N^R) \geq \mu_*$		Too much	Ambiguous		
		$\Phi < 1$	$\Phi = 1$	$\Phi \in (1, \Gamma)$	$\Phi \geq \Gamma$
Local politics.					
Large cities		undersized	undersized§	undersized‡	oversized
Small cities		oversized†	oversized§	oversized	oversized
Number of cities		too many	too many§	ambiguous	too few
	$\Phi < \Phi_*$	$\Phi \in [\Phi_*, 1)$	$\Phi = 1$	$\Phi \in (1, \Gamma)$	$\Phi \geq \Gamma$
Free migration.					
Large cities	undersized	oversized‡	oversized§	oversized	oversized
Small cities	oversized†	undersized	undersized§	ambiguous	oversized
Number of cities	too many	ambiguous	too few§	too few	too few

Notes: §optimal if homogeneous; †by virtue of existing; ‡if $\overline{a}$ is sufficiently large.

Without fiscal externalities in either the urban or rural system, free migration between the rural and urban sector can result in under-urbanization even if the urban sector is efficiently organized. This case is represented by point M in Figure 5, where urban-rural migration equalizes (ex ante) average benefits in each sector, whereas the equalization of marginal benefits would entail a larger urban population at point O. Sub-optimal organization of the urban system can aggravate this problem. If urban populations are determined by local politics, urbanization may be even lower, at point P.[18]

[18]The assumption of free urban-rural migration equalizing average expected benefits is similar to that of Harris and Todaro (1970), who argue that this leads to over-urbanization, while we still find the opposite to be possible.

4. Internalizing wedges using federal fiscal instruments

Since fiscal wedges and urban externalities prevent the urban system from being efficient, a natural question is how policies or fiscal instruments may help neutralize distortions.

4.1 Implementing the optimal allocation through developers

With no fiscal wedges ($\rho = \tau = \delta = 0$, so that $\Phi = 1$), the optimum with heterogeneity may be implemented as an equilibrium outcome with perfectly competitive land developers, as in Henderson's (1974) work with homogeneous land (see Appendix B for the proof). Competitive developers offer subsidies to, and collect land rents from, urban dwellers. These payments internalize all urban externalities within cities, which could lead to inefficient migration as discussed before. What is most remarkable is that the developer result extends to our setting with heterogeneous land. The key assumption is that land developers are atomistic and behave competitively. Developers who own superior sites make strictly positive profits since better sites are in limited supply and hence command Ricardian rents. The major caveats are that developers lack incentives to create optimal city sizes if there are fiscal externalities or they have market power (market power also prevents land developers to implement the socially efficient allocation in the model with homogeneous land).

4.2 Internalizing urban externalities

The parameters that provide the optimal allocation with free-migration may be determined by finding values of k_m and Φ that allow $n_m = n_*$ to satisfy both (15) and (23):

Proposition 4 (Implementing the optimal allocation) *With free migration, the optimal allocation can be implemented using the following policy:*

(i) *Social marginal benefits are equalized across sites by setting $\Phi = \Phi_*$, thus neutralizing the wedge due to agglomeration economies and urban costs;*

(ii) *The smallest site is set at its socially efficient scale: if $a_* < \underline{a}$, $k_m = (\Phi_*)^{-\frac{1}{\gamma-\epsilon}} k_* \equiv k_m^*$.*

Proof See Appendix A.6. □

With three fiscal parameters in Φ, there are an infinite number of ways of equalizing social marginal benefits. Two particularly interesting solutions that allow to equalize Φ and Φ_* are

$$\tau^* = -\epsilon < 0, \quad \text{and either} \quad \rho^* = 1 \ \text{ and } \ \delta^* = 0, \quad \text{or} \quad \delta^* = -\frac{\gamma(1-\rho^*)}{1+\rho^*\gamma} \leq 0. \tag{26}$$

Setting $\tau = \epsilon$ provides a Pigouvian subsidy for agglomeration spillovers, so that workers internalize those benefits. Setting $\rho = 1$ requires that workers pay for land costs completely.

Combined, these two results suggest an alternative "Henry George" style policy, whereby the federal government fully taxes land, and uses it to subsidize agglomeration spillovers (which could be generalized to include public goods). At the optimum, this scheme generates a surplus with heterogeneous sites, as cities beyond their social efficient scales have land values greater than agglomeration benefits. The second alternative, with $\delta^* < 0$ raises revenue through what is essentially a congestion charge.[19]

As implied by Proposition 4, *equalizing social marginal benefits is generally insufficient when some sites are unoccupied*. The development externality at $\Phi = \Phi_*$ distorts the value of marginal sites. In the constrained-efficient migration equilibrium, the smallest site is undersized by the ratio $\Phi_*^{1/(\gamma-\epsilon)} < 1$. As a result, too many sites are occupied. Internalizing urban externalities by creating fiscal wedges distorts the extensive margin (see also Vermeulen, 2016). Achieving the efficient outcome requires additional coordination to abandon inferior sites to crowd better ones. The politics to coordinate such abandonment may be insurmountable.

This result is important for two reasons. First, it implies that any analysis of the consequences of fiscal policies on the spatial allocation of agents that does not allow for adjustments at the extensive margin is incomplete. This applies to existing models of 'optimal city systems and taxation' that work with a fixed number of cities (Eeckhout and Guner, 2015; Hsieh and Moretti, 2015; Fajgelbaum, Morales, Suarez Serrato, and Zidar, 2015). It is a corollary of a standard result in public economics and second-best theory: when there are multiple externalities, fixing one may exacerbate another (Tinbergen, 1952; Lipsey and Lancaster, 1956). Second, this policy entails *subsidizing* wages and *taxing* land and congestion. This is the opposite of what most current OECD tax systems do, and our results suggest that this may skew the population distribution away from the better sites.

4.3 Generalized optimal transfers and tax systems

To achieve the optimum in equilibrium, we consider two alternative policies, extending the model slightly.[20] The first is that of a system of city-specific transfers, $T(a)$, so that net private average benefits for each city are $u(n,a) = pab(n,a) + T(a)$. Two requirements must be met. First, social marginal benefits are equalized across cities. This means that $smb(n,a) - u[n(a_*),a_*] = \Delta$, where Δ is a constant. Second, the extensive margin is corrected by setting $\Delta = 0$. Together, these requirements mean that the optimal transfer is $T_*(a) = smb(n,a) - pab[n(a_*),a_*]$.

Altenatively, we can consider non-linear policies. We may solve for an optimal nonlinear tax, $\tau(a)$, which is proportional to the wedge between average and marginal benefits within

[19]By controlling two parameters, both of these schemes work to equalize social marginal benefits, even when cities vary in quality of life.

[20]To simplify, we abstract from rural-urban migration. It is straightforward to add a tax or transfer to the rural sector to implement the efficient rural-urban margin.

cities:

$$\tau_*(a) = \frac{sab_*(a) - smb_*(a)}{sab_*(a)} = \gamma \frac{[n_*(a)/n_i(a)]^{\gamma-\epsilon} - 1}{\frac{\gamma}{\epsilon} - [n_*(a)/n_i(a)]^{\gamma-\epsilon}}, \tag{27}$$

which is zero at the smallest city and increases with city size as the wedge grows. It does not depend on the other policy parameters. To not distort other behavioral responses, Φ should be kept equal to one. If we take ρ to be fixed by local decree, this implies

$$\delta_*(a) = 1 - \frac{1}{1+\rho\gamma} \frac{smb_*(a)}{sab_*(a)} = \gamma \frac{\frac{\gamma}{\epsilon} - 1 + (1-\rho)\left[an_*(a)^\epsilon - n_*(a)^\gamma\right]}{(1+\gamma)\left(\frac{\gamma}{\epsilon} - [n_*(a)/n_i(a)]^{\gamma-\epsilon}\right)}. \tag{28}$$

It is easy to show that $\delta_*(a) \geq \tau_*(a)$, with equality if $\rho = 0$. Thus, when residents own land, they are paid a subsidy to amortize their costs, counteracting the development externality. When all land values are appropriated locally, $\tau = \delta$, and larger cities pay more on net to the federal government. Using these, we can establish the following results.

Proposition 5 (Optimal fiscal policy) *The fiscal policy that implements the efficient allocation displays the following features:*

(i) *The (marginal) income tax rate is non-negative and increasing in city size;*

(ii) *The congestion discount rate is positive and increasing in city size;*

(iii) *If $\rho = 0$ then $\delta_*(a) = \tau_*(a)$.*

Proof In the text above. □

Several comments are in order. First, allowing for endogenous adjustments at the extensive margin of the urban system changes the qualitative features of the optimal fiscal policy in a fundamental manner. With an exogenous extensive margin, Proposition 4 suggests that the optimal fiscal policy is to subsidize labor earnings and tax urban congestion. By contrast, Proposition 5 establishes that the optimal tax and discount rates are both positive. Second, the optimal fiscal policy is progressive even though agents are homogeneous. This is because congestion dis-economies dominate agglomeration economies at the margin; and because even if agents are homogeneous ex ante, they are ex post heterogeneous in their location choices across sites of different quality. Finally, when all land is owned locally and the allocation is the one with local politics, the optimal policy is to set a common, city-specific tax rate on labor and land earnings net of congestion costs. In all cases, tax rates are increasing in nominal earnings but this does not necessarily violate the principle of treating equals equally: at the free-migration allocation, equals end up being equally well-off anyway, regardless of the tax scheme that is implemented.

5. Population (mis)allocations and welfare costs

We now explore some quantitative implications of the model. We are most interested in putting numbers on the extent of population misallocations — both at the extensive and the intensive margins — and to evaluate the welfare costs due to these misallocations. First, we briefly describe the model calibration — leaving a more detailed discussion for Appendix 6. Second, we consider how our inferred land values compare with those required under optimality under the Generalized GHV Theorem, and see under what circumstances some cities appear to be too small. Third, we simulate the system of cities under the three solution concepts (federal optimum, free migration, and local governments). Finally, we provide estimates of the welfare losses due to the misallocation of population. Using a system of cities calibrated to the United States, we find that the largest cities may be undersized by about a third, the smallest cities are too big, and that about twice too many sites may be developed. The welfare losses are fairly small under free migration, equal to only around 1% of real consumption. Yet, they can be substantial with local governments, reaching about 18% of real consumption.

5.1 Model calibration

Table 2 summarizes the range of urban and fiscal parameters we use. It also reviews we determine the level and disersion of the a values: see appendix 6 for greater detail.

Table 2: Parameter values used for the simulations.

Parameter	Baseline	Range	Source
Urban parameters			
Agglomeration ϵ	0.03	$[0.03, 0.06]$	Combes et al. (2008); Rosenthal and Strange (2004); Melo *et al.* (2009)
Congestion γ	0.25	$[0.25, 0.50]$	Combes et al. (2016); Glaeser and Gottlieb (2008); Saiz (2010); Desmet and Rossi Hansberg (2013)
Fiscal parameters			
Tax rate τ	0.34	$[0, 0.34]$	Albouy (2009)
Land rebate ρ	1	$[0, 1]$	Jimenez (1984)
Urban discount δ	0.17	$[0, 0.17]$	Albouy and Lue (2015)
Estimated distribution of amenities $G(a)$			
Wage moment	$a_j = w_j n_j^{-\epsilon}$		From wages & population in Seegert (2013)
Wage dispersion	$(\bar{w} - w_{min}) / \bar{w} = 0.13$		see Appendix 6
Urban cost moment	$(\sum_j n_j^{1+\gamma}) / (\sum_j w_j n_j) = 0.15$		Gross urban costs = 15 % of wages

Notes: This table reports our calibration. We vary the different parameters selectively within the indicated ranges as robustness checks. In our model, a parametrizes the productivity of the site. We estimate a using two moment conditions and data on wages and population from the American Community Survey. Wages are mincerized and populations combined into metropolitan areas, using the calculations in Seegert (2013).

We consider a range of parameter values for agglomeration and congestion from the literature. Our base estimates use $\epsilon = 0.03$ and $\gamma = 0.25$, which implies $\Phi_* = (1+\epsilon)/(1+\gamma) = 0.824$. Similarly, we consider a range of values for taxes, land rebates, and urban discounts to match the U.S. Our base estimates use $\tau = 0.34$, $\rho = 1$, and $\delta = 0.17$, which implies $\Phi_* > \Phi = (1-\tau)/[(1-\delta)(1+\rho\gamma)] = 0.636$. Theoretically, this parameter configuration

implies that either local politics or free-migration (absent coordination failures) should lead to undersized big cities and too many (oversized) small cities (see Table 1).

Realistically, other population distributions may prevail. Coordination failure could cause the free migration population numbers to be larger than in the constrained-efficient case, i.e., with $n_m(a)$ satisfying (23) $k_m > 1$ even with $a_m > \underline{a}$. Furthermore, incomplete enforcement of local politics could cause populations to tak on any value between the $n_p(a)$ and $n_m(a)$. One situation that seems least likely, however, is that cities are below $n_p(a)$, since this should be unstable.

Fortunately estimates of the amenity distribution do not rely on the solution concept. To determine the dispersion of amenities, we use data on wages and population from the American Community Survey. Finding a high level of dispersion will bias results towards finding cities are too small. To be conservative about this conclusion, we shrink hourly wage differences by controlling for observed worker characteristics. To control for possible unobserved characteristics, we reduce our wage differences by another third based on numbers suggested by Glaeser and Mare (2001). The productivity of each city is then given by $a_j = w_j n_j^{-\epsilon}$. We smooth the distribution, by fitting a Pareto distribution $G(a)$ for the top 150 cities. With $\epsilon = 0.03$ this produces a shape parameter of $\eta = 30$, so that the coefficient of variation for a is 0.035.

To scale the a parameters, we ensure that gross congestion costs equal a percentage of wage income: $\sum n_j^{\gamma} = 0.15 \sum_j a_j n_j^{\epsilon}$. This generous 15 percent figure is based on the conservative assumption that time costs of commuting are valued at the wage. The large pushes towards finding that cities are too big. Since the differences are known, this procedure sets the top value for $\overline{a}$ Smoothed values of a are then drawn from the power law $\ln(a_j) = -\beta_1 \ln(rank) + \beta_2$ with $\beta_1 = 1/\eta$ and $\beta_2 = \ln \overline{a}$.

Finally, to consider whether cities are congested to the level implied by the Generalized GHV Theorem, we compute a of how the minimum city wage relates to the mean. To be conservative against finding cities to be too small, we again use (shrunken) wage differences for only the 150 highest a cities. We use 1st percentile for w_{min}. This implies a narrow proportional difference of only $(\bar{w} - w_{min})/\bar{w} \equiv \Delta w/\bar{w} = 0.13$.

Table 3 is arranged to consider several other cases of Φ and Φ_*. The second and third rows of that table double either the agglomeration or congestion parameters. In both cases, $\Phi_* > \Phi$ remains true. The fourth row considers the benchmark where all of the fiscal parameters are set to zero ($\tau = \rho = \delta = 0$). This may characterize an extremely undeveloped economy with weak property rights and no taxation. In contrast to the others, $\Phi_* < \Phi$ holds here. Thus, it features oversized large cities in the free migration allocation, and may feature oversized large cities in the local politics allocation (see Table 1). The fifth row in Table 3 shows that even without federal involvement, $\tau = \delta = 0$, just having to purchase land fully upon migrating, which corresponds to $\rho = 1$, is sufficient for $\Phi_* > \Phi$ to hold. The last row considers the implications of also subsidizing the agglomeration externality, as mentioned after Proposition 4 .

Table 3: Parameter values, private-social wedges, size distortions, and urban cost-benefit ratios

Case	Urban Parameters		Fiscal Parameters			Optimal Pvt/Soc	Actual Pvt/Soc	Devel Shrink	Max Lnd/Agg	Optimal Lnd/Agg	Estimate Lnd/Agg	Free-Mig Oversize
	ϵ	γ	τ	ρ	δ	Φ_*	$\Phi = \bar{v}_p$	n_p/n_i	Γ	$\bar{v}_*$	$\bar{v}$	k_m
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Base	0.03	0.25	0.34	1	0.17	0.82	0.64	0.13	6.87	1.76	1.25	1.06
High Agglomeration	0.06	0.25	0.34	1	0.17	0.85	0.64	0.09	3.53	1.20	0.63	0.03
High Congestion	0.03	0.5	0.34	1	0.17	0.69	0.53	0.26	11.44	2.36	2.50	9.86
Developing Country	0.03	0.25	0	0	0	0.82	1	1	6.87			
Land Reform	0.03	0.25	0	1	0	0.82	0.80	0.36	6.87			
Land Ref + Agg Sub	0.03	0.25	-0.03	1	0	0.82	0.82	0.41	6.87			

The parameters $(\epsilon, \gamma, \tau, \rho, \delta)$ capture agglomeration, congestion, tax rate, land rebate, and urban discount, respectively. The optimal ratio $\Phi_* = (1+\epsilon)/(1+\gamma)$ and the private ratio $\Phi_* = (1-\tau)/((1-\delta)(1+\rho\gamma))$. The developer shrinkaage $n_p/n_i = \Phi^{1/\gamma-\epsilon}$ is how much local politics shrinks cities. The max value $\Gamma = \Phi^*(\gamma\epsilon)$ The optimal ratio of land values to agglomeration externalities is from (17) $\bar{v}_* = 1 + (\Gamma - 1)\Delta w/\bar{w}$ where $\Delta w/\bar{w} = 0.13$, the estimate of wage dispersion. The estimate of thsi ratio is $0.15\gamma/\epsilon$, where 0.15 reflects average urban costs. k_m is estimated from (25) by solving for k_m assuming $\bar{v}_m = \bar{v}$.

The ratio Φ is likely smaller for OECD countries other than the U.S. For example, France has a higher tax rate and no mortgage interest deduction, implying a lower Φ. Many countries, such as Canada, also have strong fiscal equalization systems that greatly favor smaller cities and non-urban areas (Albouy, 2012).

5.2 Urban costs, agglomeration benefits, and a simple test of optimality

The ratios in Column 8 of table 3 describe how much fiscal externalities shrink cities relative to their efficient scale under local politics, recalling equation (3): $n_p(a)/n_i(a) = \Phi^{\frac{1}{\gamma-\epsilon}}$. In the base case, the reduction is 87 percent! Now relative to the optimum allocation, this reduction applies only to the marginal optimal city at a_*. Since $n_*(a) > n_p(a)$ for $a > a_*$, higher-quality cities are even more undersized.

According to the ratio $\bar{v}_p = \Phi$ in column 7, land equals 64 percent the value of agglomeration benefits under local politics. According to the Genral GHV Theorem, with $\Delta w_*/\bar{w} = 0.13$, an optimal system would have the ratio of land values would be $\bar{v}_* = 1.76$, according to the number calculated in column 10 from equation (17).

As an inexact test of whether cities are too large, we attempt to calculate the actual ratio of land values to agglomeration benefits. Based on the assumption that urban costs are 15 percent of wages, the observed ratio of land values to agglomeration benefits is $\bar{v}_m = 0.15\gamma/\epsilon = 1.25$. This ratio is *below* the optimal ratio of 1.76, implying that American cities are less congested than an optimal system with the same exact wage dispersion.

It is possible that the ratio of 1.25 is too low: it implies that the income share of land is only 3.75 percent that of labor. Theoretically, however, we are only interested in "differential" land values due to commuting, and not other sources of land value such as the opportunity cost of agricultural land or from heterogeneity in neighborhood local good provision (e.g. from

school quality). Land values from these sources need to be subtracted away before applying this test.[21]

The calculations in column 12, operates under the assumption that cities are in a free-migration equilibrium. Thus, the equilibrium ratio is what is observed, i.e, $\bar{v}_m = \bar{v}$. This is possible, as $\bar{v}_m$ in 25 depends on the additional free variable k_m, describing how over-sized the smallest city is relative to $n_p(a_m)$. Thus, a free-migration equilibrium that produces the ratio $\bar{v}_m$ that is observed, implies a specific value of k_m. Under this calibration, the implied value of $k_m = 1.06$, as the observed $\bar{v}$ is very close to the value of $\bar{v}_m$ assuming no coordination failure, i.e, $k_m = 1$. The product $\Phi^{\frac{1}{\gamma-\epsilon}} k_m = 0.14$ implies that that the smallest city is undersized by 86 percent, while more productive cities are even more undersized, as the rate of change is suboptimal: $\Phi = 0.64 < 0.84 = \Phi_*$

Under the other two main cases, the conclusions are altered. With high agglomeration, the optimal land-ratio $\bar{v}_*$ exceeds the observed ratio, $\bar{v} = 0.15(0.25/0.06) = 0.63$, by an even larger amount. Furthermore, a free-migration equilibrium cannot be rationalized, as $k_m < 1$, which is incompatible with stability. Under this scenario, local politics must play a role in shrinking cities.

In the high congestion case the case for undersized cities grows weaker. The optimal ratio, $\bar{v}_* = 2.36$ is very close to the estimate, $\bar{v} = 0.15(0.50/0.03) = 2.50$. Under the assumption of free-migration, we have $k_m = 9.86$ and $\Phi^{\frac{1}{\gamma-\epsilon}} k_m = (0.26)(9.86) = 2.55$, meaning that the smallest city is 2.55 times its efficient scale. In this case, our basic moments suggest that coordination failure may be a major factor in determining city size. Here, at least some of the lower-quality cities are too large. However, since $\Phi = 0.53 < 0.69 = \Phi_*$ the highest-quality cities may still be too small. One must not forget at this point that our assumptions are biased against finding cities to be too small. Therefore, it is reasonable to conjecture that at least the largest (non-capital) cities are indeed too small in most developed countries.

5.3 Simulation of city sizes under the three solution concepts

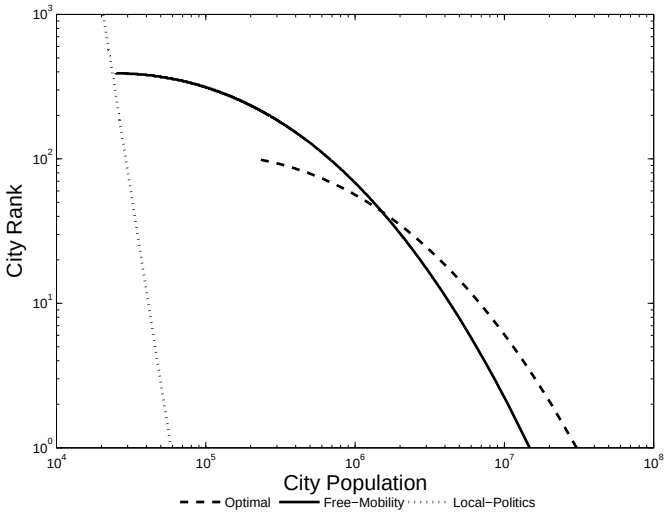
We simulate a system of cities with a total urban population of 280 million people using the three solution concepts — optimal allocation, free mobility, and local politics. Because the actual distribution of population is somewhere between the free mobility and local politics allocations, these solution concepts provide bounds on the misallocation of the actual distribution of population. Our simulations highlight three determining factors of the distribution of population. First, government policies such as taxes, land rebates, and urban cost discounts have a large impact on that distribution. Varying government policies in the simulations

[21] Moreover, we need only consider residential land values. If we take the exepnditure share on housing as 18 percent, and the cost-share of land from housing at one-third, the implied value of residential land is then 6 percent. In that case 2.25 percent of residential land values would need to stem from differences in schools, safety, and access to other amenities.

demonstrates that output is not a sufficient statistic for optimal government policies. In fact, *policies that increase total output can often decrease total welfare*. Second, the optimal distribution of population must balance both the tradeoff between agglomeration and congestion inside each city (intensive margin) and the tradeoff between overfilling productive cities and creating new, less productive cities (extensive margin). Finally, heterogeneity in city amenities is a fundamental feature of cities and interacts with both government policy and the extensive margin in ways that exacerbate the welfare consequences of population misallocations.

Figure 8: Distribution of city sizes and comparative statics.

(a) Three solution concepts.



(b) Varying fiscal parameters.

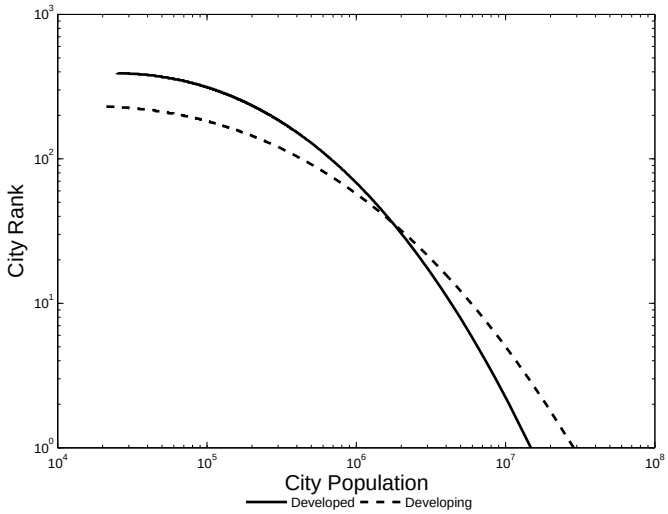


Figure 8 graphs the distribution of population in the traditional format, with log population on the horizontal axis and log rank (where the largest city is rank 1) on the vertical axis. As one can see, the city-size distributions match closely the empirical distribution in Gabaix (1999), where the size distribution of large cities, in particular, follows Zipf's law. In panel (a), the free-migration distribution (solid curve) features undersized large cities and oversized small cities relative to the optimal distribution (dashed curve), which can be seen by the free-migration distribution being a clockwise rotation of the optimal distribution. The local-politics distribution (dotted curve) consists of substantially more numerous and smaller cities, which again can be seen by the local-politics distribution being a clockwise rotation of the optimal distribution.

Panel (b) of Figure 8 graphs two free-mobility distributions for our baseline calibration. In the first one, we use the fiscal parameters of the U.S.; in the second one, we simulate a 'developing country' with no taxes and weak property rights. The developing country case features larger large cities and smaller (and fewer) small cities relative to the actual calibration using U.S. fiscal parameters. The magnitudes in this comparison suggest that *the fiscal parameters have the potential to distort the size distribution of cities substantially.* Put bluntly, with the current government policies in the U.S., New York and Los Angeles are quite substantially undersized.

In contrast, San Bernardino, California, is 60 percent too large in the free-migration allocation. As can be seen in Figure 8, an implication of under-populating the largest cities on superior sites is that more cities, in inferior sites, are populated than need be.

5.4 Welfare costs

Table 4 reports statistics on the differences in population distributions and the welfare consequences of the population misallocations. The first column reports statistics for the baseline estimates, which are calibrated to the U.S. urban system. The optimal distribution suggests the largest city of optimal size would host 30 million people. By contrast, the free-mobility and local-politics distributions have largest cities with populations of 14.8 million and 58 thousand, respectively. As a result, the free-mobility allocation populates almost four times as many cities as the optimal allocation, whereas the local-politics allocation populates a myriad of additional sites. In total, 61 percent of the population in the free-mobility allocation is misallocated relative to the optimal allocation. The welfare costs of these misallocations are 67 billion US dollars — about 1% — with free mobility, and 1.12 trillion US dollars — about 18% — with local politics.[22]

Table 4: Welfare calculations under the three solution concepts.

	(1) Baseline (Developed)	(2) Developing	(3) High agglomeration	(4) High congestion	(5) Low heterogeneity	(6) High heterogeneity
Largest city						
Optimal	30,525,000	30,525,000	50,000,000	2,760,000	11,370,000	30,525,000
Free mobility	14,790,000	49,915,000	44,410,000	1,740,000	4,870,000	40,900,000
Local politics	58,500	457,000	1,434,700	39,800	58,500	58,500
Smallest city						
Optimal	230,000	230,000	10,990,000	100,000	315,000	230,000
Free mobility	25,000	260,000	830,000	25,000	40,000	15,000
Local politics	13,300	150,200	556,600	20,400	30,300	3,000
Number of cities						
Optimal	99	99	9	804	199	99
Free mobility	392	43	33	1,633	796	165
Local politics	17,539	1,557	418	12,541	8,424	20,000
Welfare						
Optimal	22,200	22,200	25,114	21,548	23,530	22,200
Free mobility	21,960	22,060	24,737	21,427	23,333	20,358
Local politics	18,148	20,148	21,183	19,354	21,721	13,365
Output						
Optimal	28,977	28,977	35,495	24,630	28,688	28,977
Free mobility	27,002	30,313	32,955	23,682	27,099	27,444
Local politics	19,649	22,896	25,000	19,990	23,517	14,470

Notes: Baseline (Developed) has a calibration with $[\tau, \rho, \delta, \epsilon, \gamma] = [0.34, 1, 0.17, 0.03, 0.25]$. Developing calibration $[\tau, \rho, \delta, \epsilon, \gamma] = [0, 0, 0, 0.03, 0.25]$. High agglomeration $[\tau, \rho, \delta, \epsilon, \gamma] = [0.34, 1, 0.17, 0.06, 0.25]$. High congestion $[\tau, \rho, \delta, \epsilon, \gamma] = [0.34, 1, 0.17, 0.03, 0.5]$. Baseline heterogeneity $a = 19{,}550 \times rank^{-0.0333}$, comes from estimating the function $log(a_j) = \beta_0 + \beta_1 log(rank)$ from the calculated a values. We calculate the a values from the moment condition $a_j = w_j n_j^{-\epsilon}$ using data on wages, w_j, and population, n_j, from the American Community Survey, where skill differences have been controlled for in the wage. Low heterogeneity, $a = 19{,}550 \times rank^{-0.016}$. High heterogeneity, $a = 19550 \times rank^{-0.066}$.

[22] Remarkably, the welfare consequences in this simulation are comparable to those of Behrens et al. (2011), Desmet and Rossi-Hansberg (2013), and Behrens, Duranton, and Robert-Nicoud (2014), though the underlying microeconomic foundations of all four papers are entirely different and though Behrens et al.'s model features a more realistic geography than our own. Market access is encapsulated in the distribution of a's in our model. It depends on the distribution of bilateral trade costs in Behrens et al. (2011) and is thus endogenous.

The impact of government policy on the distribution of population is shown by contrasting the baseline estimates (Column 1 in Table 4), calibrated to the U.S. and its fiscal parameters, with the developing country estimates (Column 2 in Table 4). The latter is an extreme example with no taxes and weak property rights. In contrast to the baseline calibration, the latter allocation overcrowds the largest cities and populates too few sites. With weak property rights and no taxes, the largest city in the free-mobility allocation hosts almost 50 million people, 20 million more than in the optimal allocation and 35 million more than in the baseline calibration. As a consequence of overcrowding the largest cities, the free-mobility allocation populates only 43 sites, less than half the number of sites in the optimal allocation and substantially fewer than in the baseline calibration. Table 4 shows that the welfare consequences of these misallocations are even smaller than in the U.S. baseline case. It is worth noting that, because the free-mobility allocation overpopulates the cities with the highest production amenities, output at the free-mobility allocation is larger than at the optimal allocation. In other words, gross output is a poor measure of welfare because welfare is output net of congestion costs, and some allocations may feature massive excessive congestion. The presence of excessive congestion suggests a key role for taxes to correct for the congestion externality.

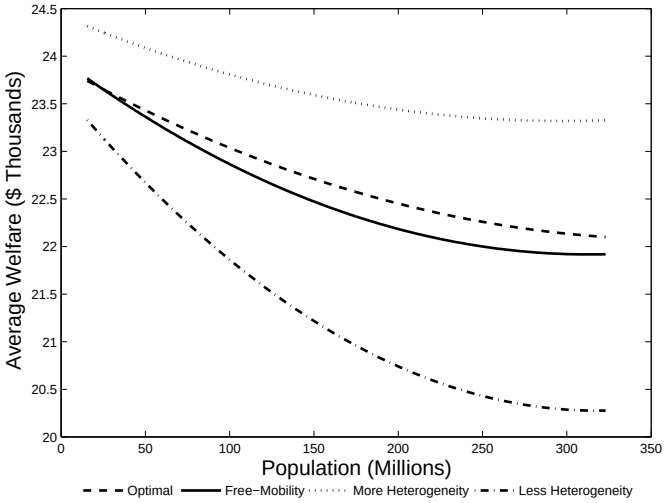
The impact of agglomeration economies and congestion diseconomies on the distribution of population is shown in Columns 3 and 4 of Table 4. Column 3 doubles the agglomeration parameter to 0.06 and Column 4 doubles the congestion parameter to 0.5. Not surprisingly, more population is allocated to the largest cities when agglomeration forces are stronger and/or congestion diseconomies are weaker.

The impact of heterogeneity in amenities across cities is shown in Columns 5 and 6 of Table 4, and its welfare effects are depicted in panel (a) of Figure 9. Changing the difference in amenities across cities changes the tradeoff between increasing congestion in cities with high amenity levels and populating cities with lower amenity levels. As expected, the impact of heterogeneity is large. The reason for this is that the urban system operates at close to constant returns to scale, i.e., the difference between agglomeration economies ϵ and urban costs γ is relatively small. Small amenity changes then translate into large population changes by giving better sites an additional advantage. To make the analogy with heterogeneous firms, small cost advantages imply large differences in market shares when products are good substitutes. In the case of cities, small differences in a imply large differences in population when $\gamma - \epsilon$ is small. Although the parameters we have chosen for the 'low' and 'high' heterogeneity cases are arbitrary, this exercise illustrates that getting those parameters right is of fundamental importance in any quantitative exercise.

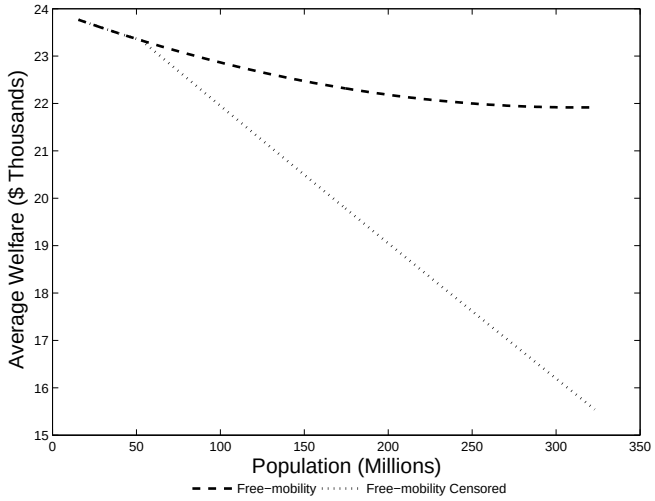
Last, to highlight the extensive margin tradeoff, panel (b) of Figure 9 compares the change in average welfare as total population increases when additional cities can and cannot be created. Average welfare decreases substantially faster when the number of cities is capped at 100 (dotted curve) than in the baseline case when the number of cities is allowed to grow (dashed

Figure 9: Welfare in the federal optimum and with free migration.

(a) Heterogeneity in amenities.



(b) Extensive margin.



curve). This suggests that the ability to create new cities is critical to maintaining high welfare as total population increases. This result has implications for urbanizing countries that tend to concentrate population in a few large cities (World Bank, 2009). In countries like China, India, or Brazil, the increasing urban population needs to be accommodated in an increasingly large number of cities to mitigate the decreasing returns that operate sharply in individual cities.

6. Conclusions

The theory and evidence in this paper suggest that the received economic wisdom — that cities are too big — is likely to fail in most developed countries. Fiscal externalities from taxes and land purchases generally discourage urbanization. They discourage urban development as the returns to land and labor are common resources, not fully given to migrants. Though not necessary for our key result, local politics naturally keeps the best areas from being sufficiently inhabited, thus resulting in the best areas being substantially undersized. Overall, we expect to see an inefficient urban system, characterized by lower welfare, over-ruralization, and too many minor cities developed on inferior sites. While the welfare costs of these misallocations are small with free migration, they are large under local politics. Hence, local control over land-use regulations leads to a seriously inefficient allocation of population across cities.

The circumstances involving urbanization in developing countries is less certain. Weak taxation and incomplete property rights over land may limit the effective tolls migrants pay for entering a city. This can easily cause the best sites in a country to become overcrowded. Local governments seem unlikely to have the power to stop this. Meanwhile, capital cities may be prone to gigantism due to the concerns of central governments. Thus, the conventional wisdom may still apply in some countries, with large cities being too big, and possibly too much urban-

ization overall. Nevertheless, our work highlights the importance of considering fiscal policy, site heterogeneity, and incentives to develop before making specific policy recommendations about whether urban development should generally be discouraged.

References

Abdel-Rahman, Hesham M. 1988. "Product differentiation, monopolistic competition and city size." *Regional Science and Urban Economics* 18**(1)**: 69–86.

Abdel-Rahman, Hesham, and Alex Anas. 2004. "Theories of systems of cities." In: Henderson, J. Vernon, and Jacques-François Thisse (eds.) *Handbook of Regional and Urban Economics, vol. 4.* North-Holland: Elsevier B.V., pp. 2293–2339.

Ades, Alberto F., and Edward L. Glaeser. 1995. "Trade and circuses: Explaining urban giants." *Quarterly Journal of Economics* 110**(1)**: 195–227.

Albouy, David. 2016. "What are cities worth? Local productivity, land values, and the total value of amenities." *Review of Economics and Statistics* 98**(3)**: 477–487.

Albouy, David. 2012. "Evaluating the efficiency and equity of federal fiscal equalization." *Journal of Public Economics* 96**(9)**: 824–839.

Albouy, David. 2009. "The unequal geographic burden of federal taxation." *Journal of Political Economy* 117**(4)**: 635–667.

Albouy, David and Andrew Hanson. 2014. "Are Houses Too Big or In the Wrong Place? Tax Benefits to Housing and Inefficiencies in Location and Consumption" *NBER Tax Policy and the Economy* 28: 63-96.

Albouy, David and Bert Lue. 2015. "Driving to opportunity: Local rents, wages, commuting, and sub-metropolitan quality of life." *Journal of Urban Economics* 89: 74-92.

Allen, Treb, and Costas Arkolakis. 2014. "Trade and the topography of the spatial economy." *Quarterly Journal of Economics* 129**(3)**: 1085–1140.

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

Arnott, Richard. 2004. "Henry George Theorem and optimal city size." *American Journal of Economics and Sociology* 63**(5)**: 1057–1090.

Arnott, Richard. 1979. "Optimal city size in a spatial economy." *Journal of Urban Economics*; 65–89.

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

Aumann, Robert J. 1964. "Markets with a continuum of traders." *Econometrica* 32**(1/2)**: 39–50.

Bleakly, Hoyt, and Jeffrey Lin. 2012. "Portage and path dependence." *Quarterly Journal of Economics* 127**(2)**: 587–644.

Behrens, Kristian, Giordano Mion, Yasusada Murata, and Jens Suedekum. 2011. "Spatial frictions." CEPR Discussion Paper #8572, *Centre for Economic Policy Research*, London, UK.

Buchanan, James M., and Charles J. Goetz. 1972. "Efficiency limits of fiscal mobility: An assessment of the Tiebout model." *Journal of Public Economics* 1**(1)**: 25–43.

Cheshire, Paul, and Stephen Sheppard. 2002. "The welfare economics of land use planning." *Journal of Urban Economics* 52**(2)**: 242–269.

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

Combes, Duranton, and Gobillon. 2016. "The costs of agglomeration: House and land prices in French cities." Mimeo, University of Pennsylvania.

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

Combes, Pierre-Philippe, and Laurent Gobillon. 2015. "The empirics of agglomeration economies." In: Duranton, Gilles, J. Vernon Henderson, and William C. Strange (eds.) *Handbook of Regional and Urban Economics, vol. 5*. North-Holland: Elsevier B.V., pp. 247–348.

Davis, Donald R., and David E. Weinstein. 2002. "Bones, bombs, and break points: The geography of economic activity." *American Economic Review* 92**(5)**: 1269–1289.

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

Duranton, Gilles, and Diego Puga. 2015. "Urban land use." In: Duranton, Gilles, J. Vernon Henderson, and William C. Strange (eds.) *Handbook of Regional and Urban Economics, vol. 5*. North-Holland: Elsevier B.V., pp. 476–560.

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

Eeckhout, Jan, and Nezih Guner. 2015. "Optimal spatial taxation: Are big cities too small?" CEPR *Discussion Paper* #10352, Centre for Economic Policy Research, London, UK.

Fajgelbaum, Pablo D., Eduardo Morales, Juan Carlos Suárez Serrato, and Owen M. Zidar. 2015. "State taxes and spatial misallocation." NBER Working Paper #21760, National Bureau for Economic Research, Cambridge, MA.

Fenge, Robert, and Volker Meier. 2002. "Why cities should not be subsidized." *Journal of Urban Economics* 52**(3)**: 433–447.

Fischel, William A. 2001. *The Homevoter Hypothesis: How Homevalues Influence Local Government Taxation, School Finance, and Land-Use Policies*. Harvard University Press, Cambridge, MA.

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

Fujita, Masahisa. 1989. *Urban Economic Theory*. MIT Press, Cambridge, MA.

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

Glaeser, Edward L., and Joshua D. Gottlieb. 2008. "The Economics of Place-Making Policies" *Brookings Papers on Economic Activity* Spring: 155-352.

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

Harris, John R., and Michael P. Todaro. 1970. "Migration, unemployment and development: A two-sector analysis." *American Economic Review* 60**(1)**: 126–142.

Haurin, Donald R. 1980. "The regional distribution of population, migration, and climate." *Quarterly Journal of Economics* 95 **(2)**: 293–308.

Helpman, Elhanan, and David Pines. 1980. "Optimal public investment and dispersion policy in a system of open cities." *American Economic Review* 70**(3)**: 507–514.

Henderson, J. Vernon. 1988. *Urban Development: Theory, Fact, and Illusion.* New York and Oxford: Oxford University Press.

Henderson, J. Vernon. 1974a. "Optimum city size: The external diseconomy question." *Journal of Political Economy* 82**(2)**: 373–388.

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

Henderson, J. Vernon, and Anthony J. Venables. 2009. "The dynamics of city formation." *Review of Economic Dynamics* 12**(2)**: 233–254.

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

Hochman and Pines. 1993. "Federal income tax and its effects on inter-and intracity resource allocation." *Public Finance Quarterly* 21**(3)**: 276–304.

Hornstein, Andreas, Per Krussell and Giovanni L. Violante. 2011. "Frictional wage dispersion in search models: a quantitative assessement." *American Economic Review* 101**(7)**: 2873-2898.

Hsieh, Chang-Tai, and Enrico Moretti. 2015. "Why do cities matter? Local growth and aggregate growth." NBER Working Paper #21154, National Bureau for Economic Research, Cambridge, MA.

Jimenez, Emmanuel, 1984. "Tenure security and urban squatting." *Review of Economics and Statistics*, 6675**(4)**: 556–67.

Kanemoto, Yoshitsugu. 1980. *Theories of Urban Externalities.* North-Holland, Amsterdam.

Knight, Frank H. 1924. "Some fallacies in the interpretation of social cost." *Quarterly Journal of Economics* 38**(4)**: 582–606.

Krugman, Paul R. 1996. "Confronting the mystery of urban hierarchy." *Journal of the Japanese and International Economies* 10**(4)**: 399–418.

Lee, Sanghoon, and Qiang Li. 2013. "Uneven landscapes and city size distributions." *Journal of Urban Economics* 78**(C)**: 19–29.

Lipsey, Richard G., and Kelvin Lancaster. 1956. "The general theory of second best." *Review of Economic Studies* 24**(1)**: 11–32.

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

Melo, Patricia C., Daniel J. Graham, and Robert B. Noland. 2009. "A meta-analysis of estimates of urban agglomeration economies." *Regional Science and Urban Economics* 39**(3)**: 332–342.

Milgrom, Paul, and John Roberts. 1994. "Comparing equilibria." *American Economic Review* 84**(3)**: 441–459.

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

Mills, Edwin S. and David M. Ferranti. 1971. "Market choices and optimum city size." *American Economic Association Papers and Proceedings* 61**(2)**: 340–345.

Mirrlees, James A. 1982. "Migration and optimal income taxes." *Journal of Public Economics* 18**(3)**: 319–341.

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

O'Sullvan, Arthur. 2011. *Urban Economics*, 8th edition. McGraw-Hill, New York.

Redding, Stephen J. 2016. Goods trade, factor mobility and welfare. *Journal of International Economics* 101**(C)**: 148–167.

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

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

Romer, Paul M. 1990. "Endogenous technological change." *Journal of Political Economy* 98**(5)**: 71–102.

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

Seegert, Nathan. 2013. "Rushing to opportunities: A model of entrepreneurship and growth." SSRN Discussion Paper #2857105.

Seegert, Nathan. 2011. "Land regulations and the optimal distribution of cities." SSRN Discussion Paper #2557399.

Stiglitz, Joseph E. 1977. "The Theory of Local Public Goods." in Feldstein, Martin S. and Robert P. Inman, eds. *The Economics of Public Services*. London, MacMillan Press.

Tiebout, Charles M. 1956. "A pure theory of local expenditures." *Journal of Political Economy* 64**(5)**: 416–424.

Tinbergen, Jan. 1952. *On the Theory of Economic Policy.* North-Holland, Amsterdam.

Tolley, George S. 1974. "The welfare economics of city bigness." *Journal of Urban Economics* 1**(3)**: 324–345.

Upton, Charles 1981. "An equilibrium model of city size." *Journal of Urban Economics* 10**(1)**: 15–36.

Vermeulen, Wouter. 2016. "Agglomeration externalities and urban growth controls." *Journal of Economic Geography*, forthcoming (`doi:10.1093/jeg/lbv047`).

Vickrey, William S. 1992. "Henry George, economies of scale, and land value taxation." Presented at a COPE meeting in Rio de Janeiro, Jan. 9, 1992, reprinted in Kenneth C. Wenzer (1999, ed.), *Land-Value Taxation: The Equitable and Efficient Source of Public Finance.*

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

Appendix material

This set of appendices is structured as follows. First, Appendix **A** contains the proofs of the propositions. Second, Appendix **B** shows that the optimal allocation can be implemented by perfectly competitive land developers, as in Henderson (1974a, b). Last, Appendix **C** contains details on our data and the procedure used to calibrate the model.

Appendix A: Proofs

Preliminaries. The proofs are simplified through a change in variables for a and $n(a)$, using the notation

$$n_i(a) \equiv \left(\frac{\epsilon}{\gamma}a\right)^{\frac{1}{\gamma-\epsilon}} \quad \text{and} \quad v(n(a)) \equiv \left[\frac{n(a)}{n_i(a)}\right]^{\gamma-\epsilon}.$$

Given the expressions in Lemma 1, we have $v_i \equiv v(n_i(a)) = 1$ and $v_p \equiv v(n_p(a)) = \Phi$. v gives us the ratio of the city's size relative to the social efficient scale, $n_i(a)$, raised to the power $\gamma - \epsilon$. Thus, v gives the ratio of urban costs to urban benefits at the given $n(a)$ relative to this ratio at $n_i(a)$. This latter term, replaces a as the underlying productivity parameter. We use this notation to re-express the social and private, average and marginal, benefits as follows:

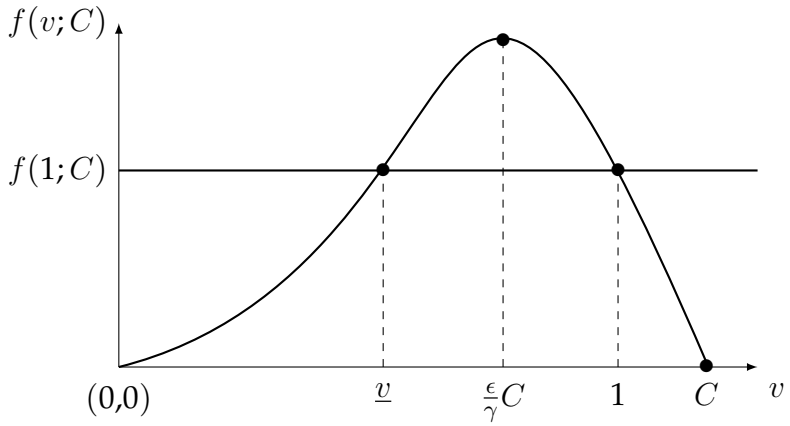
$$\begin{array}{lcllcl} sab(n,a) & = & v^{\frac{\epsilon}{\gamma-\epsilon}} n_i^{\gamma} \left(\frac{\gamma}{\epsilon} - v\right) & = & f\left(v; \frac{\gamma}{\epsilon}\right) \\ smb(n,a) & = & (1+\gamma)\, v^{\frac{\epsilon}{\gamma-\epsilon}} n_i^{\gamma} \left(\Gamma - v\right) & = & (1+\gamma)\, f\left(v; \Gamma\right) \\ pab(n,a) & = & \frac{1-\tau}{\Phi} v^{\frac{\epsilon}{\gamma-\epsilon}} n_i^{\gamma} \left(\frac{\gamma}{\epsilon}\Phi - v\right) & = & \frac{1-\tau}{\Phi} f\left(v; \frac{\gamma}{\epsilon}\Phi\right) \\ pmb(n,a) & = & \frac{(1+\gamma)(1-\tau)}{\Phi} v^{\frac{\epsilon}{\gamma-\epsilon}} n_i^{\gamma} \left(\Gamma\Phi - v\right) & = & \frac{(1+\gamma)(1-\tau)}{\Phi} f\left(v; \Gamma\Phi\right), \end{array} \tag{A-1}$$

where $f(v;C) \equiv v^{\frac{\epsilon}{\gamma-\epsilon}} n_i^{\gamma} \left(C - v\right)$ is a function of v, parametrized by (a bundle of) parameters C; and where (as defined in the main text)

$$\Phi \equiv \frac{1-\tau}{(1-\delta)\left(1+\rho\gamma\right)}, \quad \Phi_* \equiv \frac{1+\epsilon}{1+\gamma}, \quad \text{and} \quad \Gamma \equiv \frac{\gamma}{\epsilon}\Phi_*.$$

Figure 10 depicts the general, quasi-concave shape of f as a function of v for a fixed value of C. It is always a multiple of n_i^{γ}, the urban costs at the efficient scale. We are generally more interested in the positive values located between that of f at $v = 0$ and $v = C$.

Figure 10: Properties of $f(v;C)$ for $1 \leq C \leq \frac{\gamma}{\epsilon}$.



It is easy to show that at the maximum

$$\arg\max_v \;\; f(v;C) = \frac{\epsilon}{\gamma}C, \quad \text{and} \quad \max_v \;\; f(v;C) = \left(C\frac{\epsilon}{\gamma}\right)^{\frac{\gamma}{\gamma-\epsilon}} \left(\frac{\gamma}{\epsilon} - 1\right) n_i^{\gamma}.$$

Using hat notation, $\widehat{x} \equiv \mathrm{d}x/x$, then $\widehat{n}_i = \widehat{a}/\left(\gamma - \epsilon\right)$, and $\widehat{v} = \left(\gamma - \epsilon\right)\left(\widehat{n} - \widehat{n}_i\right) = \left(\gamma - \epsilon\right)\widehat{n} - \widehat{a}$. Taking into account the dependence of f on both v and implicitly on n_i, we have

$$\widehat{f} = -\frac{\frac{\gamma}{\epsilon}v - C}{\left(\frac{\gamma}{\epsilon} - 1\right)\left(C - v\right)}\widehat{v} + \gamma\widehat{n}_i = -\epsilon\frac{\frac{\gamma}{\epsilon}v - C}{C - v}\widehat{n} + \frac{C}{C - v}\widehat{a}. \tag{A-2}$$

where the second expression reintroduces the original population and productivity notation. Therefore, f is always increasing in a, for given n, if $v < C$; and increasing in n, for given a, if $0 < v < \frac{\epsilon}{\gamma}C$ and decreasing if $\frac{\epsilon}{\gamma}C < v < C$. We are generally interested in this latter range. Substituting in the proper value of C, the relevant ranges of our four functions in (A-1) are

	sab	smb	pab	pmb
$\frac{\epsilon}{\gamma}C$	1	Φ^*	Φ	$\Phi^*\Phi$
C	$\frac{\gamma}{\epsilon}$	Γ	$\frac{\gamma}{\epsilon}\Phi$	$\Gamma\Phi$

As far as values of C are concerned, the most relevant values obey $1 \leq C \leq \frac{\gamma}{\epsilon}$. For v this implies that $\frac{\epsilon}{\gamma} < v < \frac{\gamma}{\epsilon}$. For this range, the function $f(1;C) = 1$ has two real solutions, 1, and $\underline{v}$, a second value, which does not have a simple analytical expression, but by the continuity of $f(.;C)$ exists by the Intermediate Value Theorem. See Figure 10 for an illustration of $\underline{v}$'s existence.

Translating from (v,n_i) back to n,a gives us relevant ranges of n for social and private, average and marginal, benefits:

	sab $C = \gamma/\epsilon$	smb $C = \Gamma$	pab $C = (\gamma/\epsilon)\Phi$	pmb $C = \Gamma\Phi$
$n^0 \equiv \left(Ca\frac{\epsilon^2}{\gamma^2}\right)^{\frac{1}{\gamma-\epsilon}}$	$\left(a\frac{\epsilon}{\gamma}\right)^{\frac{1}{\gamma-\epsilon}}$	$\left(a\frac{\epsilon}{\gamma}\Phi_*\right)^{\frac{1}{\gamma-\epsilon}}$	$\left(a\frac{\epsilon}{\gamma}\Phi\right)^{\frac{1}{\gamma-\epsilon}}$	$\left(a\frac{\epsilon}{\gamma}\Phi_*\Phi\right)^{\frac{1}{\gamma-\epsilon}}$
$n^{\max} \equiv \left(Ca\frac{\epsilon}{\gamma}\right)^{\frac{1}{\gamma-\epsilon}}$	$a^{\frac{1}{\gamma-\epsilon}}$	$(a\Phi_*)^{\frac{1}{\gamma-\epsilon}}$	$(a\Phi)^{\frac{1}{\gamma-\epsilon}}$	$(a\Phi_*\Phi)^{\frac{1}{\gamma-\epsilon}}$

Values of n outside these parameter ranges are generally not pertinent to the analysis.

A.1. Proof of Lemma 2 Consider the problem given by the Lagrangian

$$\max_{N^R,a_*,n(a),\mu} \mathcal{L} \equiv F(N^R) + \int_{a_*}^{\overline{a}} n(a)\left[an(a)^\epsilon - n(a)^\gamma\right] \mathrm{d}G(a) + \mu\left[\overline{N} - N^R - \int_{a_*}^{\overline{a}} n(a)\mathrm{d}G(a)\right].$$

The first-order condition with respect to μ is simply the adding-up constraint. The first-order condition with respect to N^R yields $\mu = F'(N^R)$, the necessary condition for **(i)**. Using the continuum of a The first-order necessary condition with respect to a_* is

$$\mu_* \leq a_* n(a_*)^\epsilon - n(a_*)^\gamma = sab\left[n(a_*),a_*\right], \tag{A-3}$$

with equality if $a_* > \underline{a}$. Let $k_* = n_*(a_*)/n_i(a_*)$ be the ratio of the optimal size of the worst city relative to its efficient size. Solving for μ_* from the smb using the notation in (A-1), with $(1+\gamma)f\left[k_*^{\gamma-\epsilon},\Gamma\right]$, gives us the result of **(iii)**

$$\mu_* = (1+\gamma)\left(\Gamma k_*^\epsilon - k_*^\gamma\right)\left[n_i(a_*)\right]^\gamma. \tag{A-4}$$

The precise value of a_* needed for **(ii)** depends on the (unspecified) distribution G. Condition (A-3) shows that the Lagrange multiplier is pinned down by the social average benefit of the site at the extensive-margin (or worst-populated site), a_*, if the extensive margin of urban development is relevant. The multiplier also equals the marginal benefit of residing in the rural area. The latter decreases in N^R because F is concave in population.

The first-order condition with respect to $n(a)$ for the intensive margin is

$$\mu_* \geq a(1+\epsilon)n_*(a)^\epsilon - (1+\gamma)n_*(a)^\gamma, \tag{A-5}$$

with equality for all sites a such that $n(a) > 0$. Using the smb expression in A-1 again and our solution for μ_*, we have through a change of notation that

$$v_*^{\frac{\epsilon}{\gamma-\epsilon}} \left[n_i(a)\right]^\gamma \left[\Gamma - v_*\right](1+\gamma) = (1+\gamma)\left(\Gamma k_*^\epsilon - k_*^\gamma\right)\left[n_i(a_*)\right]^\gamma.$$

Using the fact that $\left[n_i(a)\right]^\gamma = (a/a_*)^{\frac{\gamma}{\gamma-\epsilon}}\left[n_i(a_*)\right]^\gamma$, this simplifies to

$$(a/a_*)^{\frac{\gamma}{\gamma-\epsilon}}\left[\Gamma v_*^{\frac{\epsilon}{\gamma-\epsilon}} - v_*^{\frac{\gamma}{\gamma-\epsilon}}\right] = \Gamma k_*^\epsilon - k_*^\gamma$$

for $\Phi_* < v_* < \Gamma$ by the second order condition. This range is tightened further into $k_* < v_* < \Gamma$ by our solution for a_*. These results translate immediately into result **(v)**. Note this solution may be written as $(a/a_*)^{\frac{\gamma}{\gamma-\epsilon}} = f[k_*^{\gamma-\epsilon}, \Gamma]/f[(n_*/n_i)^{\gamma-\epsilon}, \Gamma]$

To see that optimal city size is increasing in a, take the elasticity formula (A-2) for smb, using $C = \Gamma$ and $\hat{f} = 0$, which means

$$\epsilon\frac{\frac{\gamma}{\epsilon}v_* - \Gamma}{\Gamma - v_*}\hat{n}_* = \frac{\Gamma}{\Gamma - v_*}\hat{a}.$$

By cross-multiplying the result in **(iv)**, it becomes apparent

$$\frac{\hat{n}_*}{\hat{a}} = \frac{1}{\epsilon}\frac{\Gamma}{\frac{\gamma}{\epsilon}v_* - \Gamma}$$

once v is substituted in for. This expression is positive over the relevant range $v > 1 > \Phi_*$. Furthermore, it is possible to show that

$$\frac{\hat{v}_*}{\hat{a}} = \frac{\gamma}{\epsilon}\frac{\Gamma - v}{\frac{\gamma}{\epsilon} - v} \tag{A-6}$$

which is positive, implying that the ratio of n_* to n_i continues to grow, albeit at a declining rate. If A is unbounded above, then at $a \to \infty$, $v_* \to \Gamma$, and $n_*(a) \to \Gamma^{\frac{1}{\gamma-\epsilon}} n_i(a)$.

Clearly, $a_* > 0$ must hold if $F'(\overline{N}) > 0$, since $smb(n,0) = -(1+\gamma)n^\gamma < 0$ for all $n > 0$. In other words, it is always better to work in the countryside than live in a completely unproductive city due to urban costs. This establishes **(i)** for the extensive margin. This puts a limit on $k_* < (\gamma/\epsilon)^{\frac{1}{\gamma-\epsilon}}$.

As will be shown in Proposition 1, $\mathrm{d}a_*/\mathrm{d}N < 0$, which implies that $\mathrm{d}\mu_*/\mathrm{d}N < 0$. Since $N^R + N = \overline{N}$, then $\mathrm{d}\mu_*/\mathrm{d}N^R > 0$, there is a unique rural population $N^R \in (0, \overline{N})$.

This completes the proof. □

A.2. Proof of Proposition 1. To see **(i)**, take the ratios of sab to smb in (A-1) evaluated at v_* to obtain

$$sab_*(a) = \frac{1}{1+\gamma}\frac{\frac{\gamma}{\epsilon} - v_*}{\Gamma - v_*} smb_* = \frac{1}{1+\epsilon}\frac{\Gamma - \Phi_* v_*}{\Gamma - v_*}\mu_*.$$

Substituting in for v_* creates the required formula in **(i)**. Evaluated at $v_*(a_*) = k_*^{\gamma-\epsilon}$, we have $sab_*(a_*) = \left(\frac{\gamma}{\epsilon}k_*^{\epsilon} - k_*^{\gamma}\right)[n_i(a_*)]^{\gamma}$, which equals μ_* if $k_* = 1$, when $a_* > \underline{a}$. Taking the elasticity,

$$\widehat{sab}_* = \frac{\frac{\gamma}{\epsilon}(1-\Phi_*)v_*}{(\frac{\gamma}{\epsilon} - v_*)(\Gamma - v_*)}$$

from $\Phi_* < 1$ and $1 < v < \Gamma < \frac{\gamma}{\epsilon}$. The elasticity of sab with respect to v grows with v from the denominator. In logarithmic scales, the optimized $\ln[sab_*(a)]$ is convex in $\ln(n)$.

To establish **(ii)**, observe that the Lagrangian can be rewritten as follows:

$$\frac{\mathcal{L}_*}{\overline{N}} = \mu_* + \frac{F(N^R) - F'(N^R)N^R}{\overline{N}} + \frac{1}{\overline{N}}\int_{a_*}^{\overline{a}} n_*(a)\left[an_*(a)^{\epsilon} - n_*(a)^{\gamma} - \mu_*\right]\mathrm{d}G(a).$$

From the first-order condition (A-5), we have $an_*(a)^{\epsilon} - n_*(a)^{\gamma} = \mu_* + \gamma n_*(a)^{\gamma} - a\epsilon n_*(a)^{\epsilon}$. Inserting this into the Lagrangian above, and using $\mu_* = F'(N^R)$, yields

$$\frac{\mathcal{L}_*}{\overline{N}} = \mu_* + \frac{N^R}{\overline{N}}\left[\frac{F(N^R)}{N^R} - F'(N^R)\right] + \frac{1}{\overline{N}}\int_{a_*}^{\overline{a}} n_*(a)\left[\gamma n_*(a)^{\gamma} - a\epsilon n_*(a)^{\epsilon}\right]\mathrm{d}G(a).$$

The first expression in brackets is positive because of decreasing returns in the rural sector, whereas the second bracketed expression is increasing in $n(a)$ and zero at $n(a) = n_i(a)$. Since $n_*(a) > n_i(a)$, it follows that the integral term is positive. Then,

$$\frac{\mathcal{L}_*}{\overline{N}} > \mu_* = \frac{\partial \mathcal{L}_*}{\partial \overline{N}}, \tag{A-7}$$

where the last equality comes from the definition of the Lagrangian and the envelope theorem. Expression (A-7) shows that there are decreasing returns in the economy because the marginal Lagrangian is always below the average Lagrangian. Adding people to the system reduces average welfare. This occurs from both diminishing returns in the rural sector ($F'' < 0$) and urban system. To see the latter, and to prove **(iii)**, note that we can apply the same reasoning as above to the social average benefit in all the cities, given by

$$SAB = \mu_* + \frac{1}{N}\int_{a_*}^{\overline{a}} n_*(a)\left[\gamma n_*(a)^{\gamma} - a\epsilon n_*(a)^{\epsilon}\right]\mathrm{d}G(a) > \mu_* = smb.$$

Since the social average benefit exceeds the social marginal benefit, adding urban dwellers to the cities reduces welfare.

Holding rural population constant, the sources of decreasing urban returns in **(ii)** can be shown by differentiating the first-order condition (A-5) to obtain:[23]

$$\frac{\partial \mu_*}{\partial N} = \frac{\partial n_*(a)}{\partial N}\left[a(1+\epsilon)\epsilon n_*(a)^{\epsilon-1} - (1+\gamma)\gamma n_*(a)^{\gamma-1}\right].$$

[23]We hold N^R fixed and investigate the implications of an exogenous increase in N. This is different from a redistribution from N^R to N, which is more complicated to analyze.

The term in brackets is negative by the second-order condition of the optimization problem. Hence, $\partial\mu_*/\partial N$ and $\partial n_*(a)/\partial N$ have opposite signs. By differentiating the adding-up constraint for the urban population, we get:

$$1 = \int_{a_*}^{\overline{a}} \frac{\partial n_*(a)}{\partial N} \mathrm{dNd}G(a) - n_*(a_*)\frac{\partial a_*}{\partial N}. \tag{A-8}$$

Since the sign of $\partial\mu_*/\partial N$ is the same as that of $\partial a_*/\partial N$, it is then obvious that $\partial a_*/\partial N < 0$ and $\partial n_*(a)/\partial N > 0$ must hold, since opposite signs cannot yield a positive value as required by the left-hand side of (A-8).

Finally, the Generalized Henry George Theorem can be solved by finding $\bar{v}_*$ that satisfies equation (17). The expression for land values is found by solving

$$\bar{v}_* = \frac{\gamma}{\epsilon}\frac{\int_{a_*}^{\overline{a}} n_*^{1+\gamma} dG(a)}{\int_{a_*}^{\overline{a}} a n_*^{1+\epsilon} dG(a)}$$

First note that the denominator of the second term provides the sum of income.

$$\int_{\underline{a}}^{\overline{a}} a n(a)^{1+\epsilon} dG(a) = \int_{\underline{a}}^{\overline{a}} w(a) n(a) dG(a) = N\bar{w}$$

The numerator is solved by noting that we can write (10) and (11) together as

$$n_*(a)^\gamma = n_*(a_*)^\gamma + \Phi_* \left[w_*(a) - w_*(a_*)\right] \tag{A-9}$$

Furthermore, by (11), per-capita urban costs in the marginal city is

$$n_*(a_*)^\gamma = k^\gamma \left(a\frac{\epsilon}{\gamma}\right)^{\frac{\gamma}{\gamma-\epsilon}} = k_*^\gamma \frac{\epsilon}{\gamma} a_* n_*(a_*)^\epsilon = \frac{\epsilon}{\gamma} k_*^{\gamma-\epsilon} w_*(a_*) \tag{A-10}$$

Therefore, the numerator can be written in an easy to integrate form

$$\int_{a_*}^{\overline{a}} n_* \left[\Phi_* \left[w_*(a)\right] + \left(\frac{\epsilon}{\gamma}k_*^{\gamma-\epsilon} - \Phi_*\right) w_*(a_*)\right] dG(a) = N\left[\Phi_*\bar{w}_* + \left(\frac{\epsilon}{\gamma}k_*^{\gamma-\epsilon} - \Phi_*\right) w_*(a_*)\right]$$

Dividing by the numerator and multiplying by $\frac{\gamma}{\epsilon}$, produces the expression

$$\bar{v}_* = \Gamma - \left(\Gamma - k_*^{\gamma-\epsilon}\right)\frac{w_*(a_*)}{\bar{w}_*} = k_*^{\gamma-\epsilon} + \left(\Gamma - k_*^{\gamma-\epsilon}\right)\frac{\bar{w}_* - w_*(a_*)}{\bar{w}_*}$$

Substituting in $k_* = 1$ produces the result in (17). This completes the proof. □

A.3. Proof of Proposition 2. Part **(v)** follows directly from Lemma 1. Substituting in $v = \Phi$ into the smb equation in (A-1), it is quickly calculated that

$$smb_p = smb[n_p(a), a] = \Phi^{\frac{\epsilon}{\gamma-\epsilon}} \left[\Gamma - \Phi\right](1+\gamma)\left[n_i(a)\right]^\gamma.$$

Combining this with the solution for μ_*,

$$\frac{smb_p}{smb_*} = \frac{\Phi^{\frac{\epsilon}{\gamma-\epsilon}}\left[\Gamma-\Phi\right]}{\Gamma k_*^{\epsilon}-k_*^{\gamma}}\left(\frac{a}{a_*}\right)^{\frac{\gamma}{\gamma-\epsilon}} = \frac{f\left(\Phi,\Gamma\right)}{f\left(k_*^{\gamma-\epsilon},\Gamma\right)}\left(\frac{a}{a_*}\right)^{\frac{\gamma}{\gamma-\epsilon}} \tag{A-11}$$

which may hold for values $a < a_*$.

In order for the optimum to be achieved, this ratio needs to stay equal to one. This cannot happen if a varies. Therefore the optimum is only obtained if $a = a_*$ for all developed asites, and $\Phi = k_*^{\gamma-\epsilon}$, which will be one, unless all sites are occupied. This is case **(ii)**.

If $\Phi < k_*^{\gamma-\epsilon}$, case **(i)**, then $n_p(a_*) = \Phi\frac{1}{\gamma-\epsilon}n_i(a_*) < k_* n_i(a_*) = n_*(a_*)$, and therefore, the optimal marginal site is undersized. All cities with $a > a_*$, will be undersized, since $\hat{v}_*$ is positive for $0 < v < \Gamma$, meaning that the gap between the optimum and the political allocation will continue to grow. If $k_* = 1$, so that $a_* > \underline{a}$, then sites with $a < a_*$, are "too big" in the sense that they should not have been developed. For a given N, these will exist due to the adding up constraint.

If $\Phi \geq \Gamma$, case **(iv)**, then the smallest optimal site is oversized since $n_p(a_*) = \Phi^{\frac{1}{\gamma-\epsilon}}n_i(a_*) \geq \Gamma^{\frac{1}{\gamma-\epsilon}}n_i(a_*)$, while $n_*(a) < \Gamma^{\frac{1}{\gamma-\epsilon}}n_i(a)$ for any a. Or simply $v_*(a) < \Gamma^{\frac{1}{\gamma-\epsilon}} < \Phi^{\frac{1}{\gamma-\epsilon}} = v_p(a)$ for all a. Moreover, the value of smb_p is negative for all a, which never occurs at the optimum (better to not have a city at all), and hence, n_p is clearly too large.

In the intermediate case of $k_*^{\gamma-\epsilon}\Phi < \Gamma$, the marginal optimal city, a_*, is too large. Nonetheless, the optimal elasticity v_* is positive, so that the ratio of the v_* to v_p rises with a, and can exceed it, so that some cities are too small. If A is unbounded, then $v_* \to \Gamma$, and there will be a point a_p^*, where the political allocation is optimal. Cities with $a > a_p^*$, will be undersized with local politics. This covers case **(iii)** and finishes the proof.

A.4. Proof of Lemma 3 The proof for Lemma 3 is nearly identical to that of Lemma 2, using results based on f, with previous values of C multiplied by the parameter Φ. The rest of the mathematics is straightforward and does not need repeating.

It remains to be shown that the solution is constrained-efficient. Since relative values of n_m are determined by **(iv)**, the question involves the determination of a_m and k_m. The argument is that free entry in city development will continue so long as profits are positive. This will occur for any city with $n \geq n_p$. Therefore, migrants will continue to inhabit marginal sites, so long as they can.

A.5. Proof of Proposition 3. Substituting in the solution for v_m, the social marginal benefit evaluated at the decentralized equilibrium city size $n_m(a)$ is

$$smb_m(a) \equiv smb(n_m(a), a) = \mu_m \frac{1+\epsilon}{1-\tau}\frac{\Gamma - v_m}{\Gamma - (\Phi/\Phi_*)v_m}. \tag{A-12}$$

This establishes part **(v)** with elementary substitution. Taking the ratio of this to $smb_*(a)$,

$$\frac{smb_m}{smb_*} = \Phi^{\frac{\epsilon}{\gamma-\epsilon}} \frac{\Gamma - v}{\frac{\gamma}{\epsilon}\Phi - v} \left(\frac{a_m}{a_*}\right)^{\frac{\gamma}{\gamma-\epsilon}} \frac{\frac{\gamma}{\epsilon}k_m^\epsilon - k_m^\gamma}{\Gamma k_*^\epsilon - k_*^\gamma}. \tag{A-13}$$

This expression increases in v_m if $\Phi < \Phi_*$, which is the same condition for $\hat{n}_m/\hat{a} < \hat{n}_*$.

With these results in hand, the proof is straightforward, especially with the illustrations in Figures 7 and 11. First cover cases with $a_* > \underline{a}$, starting with case **(i)**.

When $\Phi \leq \Phi_*$ the population at the optimal extensive margin is too low $n_p(a_*) < n_m(a_*)$, while the elasticity of n_m is (weakly) less than that of n_*. Therefore cities from a_* to $\overline{a}$ will have $n_m(a) < n_*(a)$. By the population constraint, there is left-over population that must go into sites worse than a_*, and so $a_m < a_*$. Thus, too many sites are occupied, and all of those sites that should be occupied are underpopulated. Sites that should not be populated are trivially overpopulated.

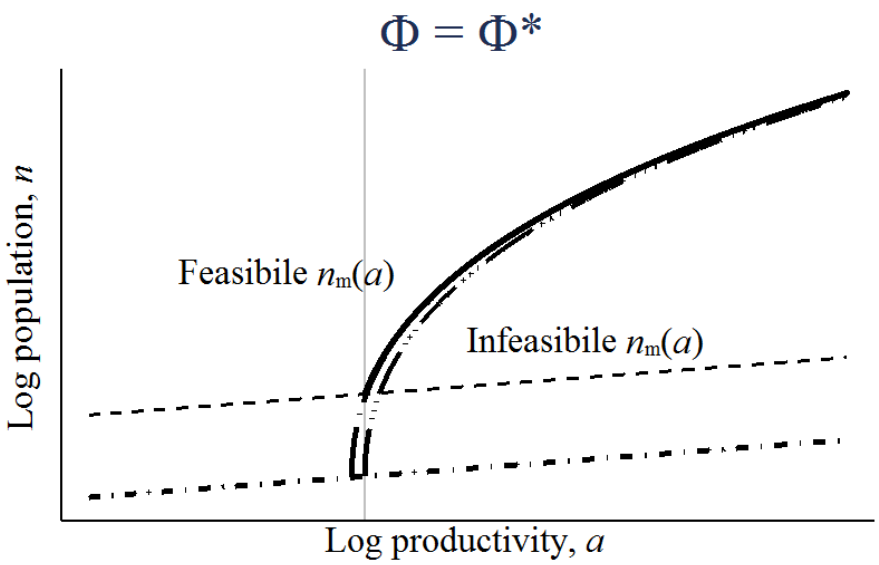
When $\Phi_* < \Phi < 1$ the population at the optimal extensive margin is too low $n_p(a_*) < n_m(a_*)$, while the elasticity of n_m is greater than that of n_*. Whether or not $a_m \gtrless a_*$, we have $n_m(a_m) = n_p(a_m) < n_*(a_m) \leq n_m(a_m)$. Therefore, the worst site is under-populated. If the range of $[a_m,\overline{a}]$ is sufficiently large, then eventually the greater elasticity of n_m with respect to a will cause it to overtake $n_m(a*)$. If it never does, then by the population constraint, there will be too many cities. If it does eventually, then it is possible that there will be too few cities if the best sites are sufficiently crowded.

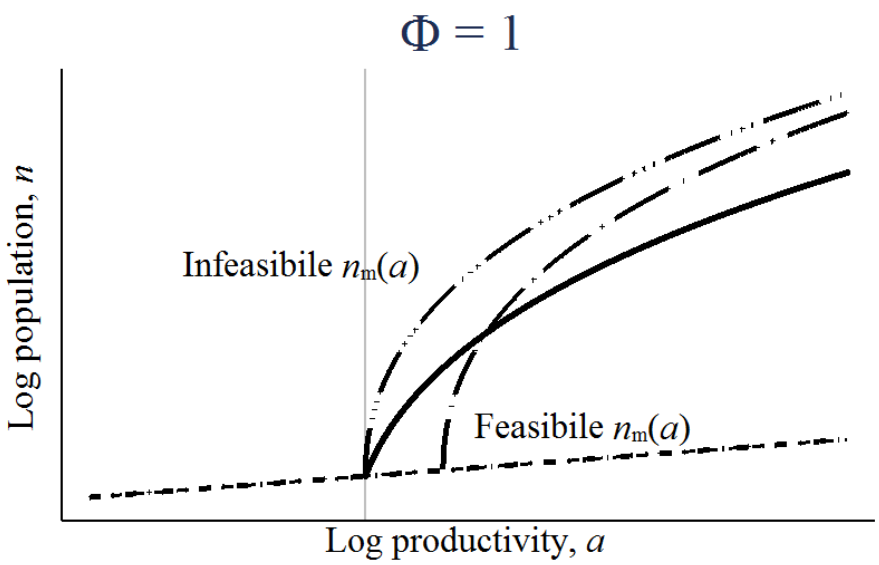
Take the all-important knife-edge case of $\Phi = 1$. When a is homogenous, $a_m = \underline{a} = \overline{a}$, the optimum is reached. If $\underline{a} \neq \overline{a}$, then $a_m \leq a_*$ is impossible since $n_m(a) > n_*(a)$ for $a > a_m$ because of the greater elasticity. This would violate the population constraint. Therefore, it must be that $a_m > a_*$ and that there are too few cities. However at a_m, $n_m(a_m) = n_i(a_m) < n_*(a_m)$ because of Lemma 2. The worst sites must be underpopulated. The population constraint requires that eventually $n_m(a) > n_*(a)$ for some value of $a > a_m$, and therefore the most productive cities will be too large.

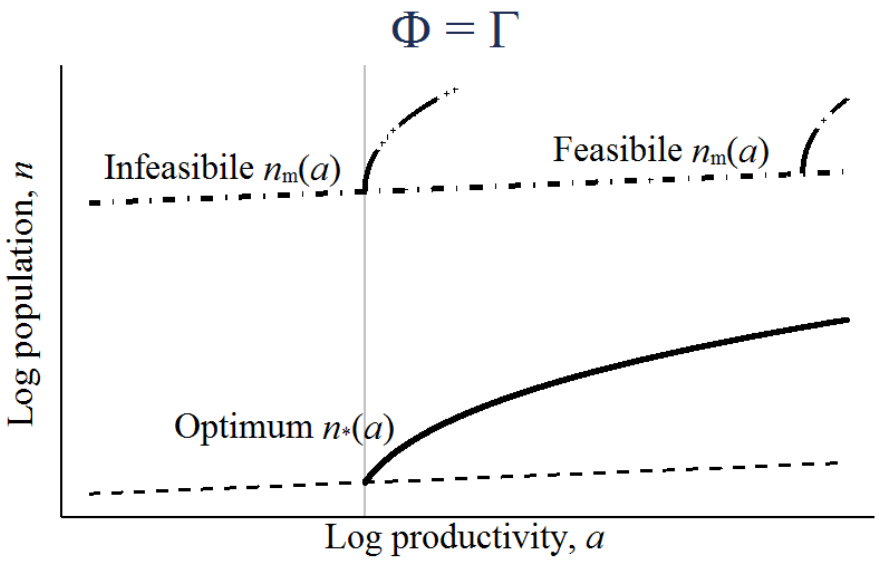
If $1 < \Phi < \Gamma$, it must still hold that $a_m \geq a_*$. However, whether $n_m(a_m) < n_*(a_m)$ depends on Φ and the distance between a_m and a_*. As with $\Phi = 1$, the population constraint eventually requires cities to be too large.

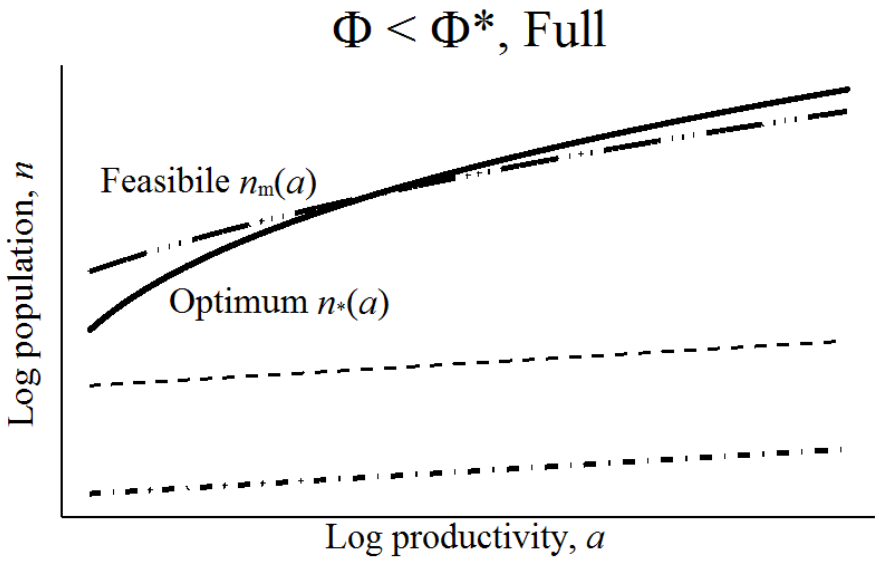
Finally, if $\Phi \geq \Gamma$ then the floor populations for $n_m(a)$ of $n_p(a)$ are always larger than $n_*(a)$, as in Propostion 2.

When all sites are filled in free-migration $a_m = \underline{a}$, then either the results of $a_m < a_*$ from above hold, or $a_* = a_m = \underline{a}$ and there is no extensive margin in either solution. In this case, if $\Phi < \Phi_*$, the elasticity of n_m is lower than that of n_a, so larger cities will be relatively smaller than what is optimal. By the population constraint, they will be absolutely smaller, while the smaller, inferior cities must be absolutely larger. The opposite holds true when $\Phi > \Phi_*$, while the optimum will hold with $\Phi = \Phi_*$.


Φ = Φ*
Log population, n
Feasibile $n_m(a)$
Infeasibile $n_m(a)$
Log productivity, a


Φ = 1
Log population, n
Infeasibile $n_m(a)$
Feasibile $n_m(a)$
Log productivity, a


Φ = Γ
Log population, n
Infeasibile $n_m(a)$
Feasibile $n_m(a)$
Optimum $n_*(a)$
Log productivity, a


Φ < Φ*, Full
Log population, n
Feasibile $n_m(a)$
Optimum $n_*(a)$
Log productivity, a

The derivation of $\bar{v_m}$ largely follows that of $\bar{v_*}$ except that (A-9) involves Φ instead of Φ_* and (A-10) involves $\Phi\frac{\epsilon}{\gamma}$ instead of $\frac{\epsilon}{\gamma}$.

This completes the proof.

A.6. Proof of Proposition 4. We see from (24) that the smb is equalized across sites if and only if $\Phi = \Phi_*$, which is a necessary condition for optimality as in (i). Yet, it is not sufficient since not all sites need to be occupied (i.e., the planner has to adjust optimally the extensive margin). This is done as follows. We have, by definition, $\Gamma/\Phi = \gamma/\epsilon$. Evaluating condition (15) at $a = a_*$ yields the condition

$$\Gamma\left(\frac{n_*}{n_i}\right)^{\epsilon} - \left(\frac{n_*}{n_i}\right)^{\gamma} = \Gamma k_*^{\epsilon} - k_*^{\gamma}.$$

Using condition (23), and the equality $n_p(a) = n_i(a)\Phi^{1/(\gamma-\epsilon)}$, and evaluating it at $a = a_m = a_*$, we also have

$$\frac{\frac{\Gamma}{\Phi}k_m^{\epsilon} - k_m^{\gamma}.}{\frac{\Gamma}{\Phi}\left(\frac{n_*}{n_i}\right)^{\epsilon}\Phi^{-\epsilon/(\gamma-\epsilon)} - \left(\frac{n_*}{n_i}\right)^{\gamma}\Phi^{-\gamma/(\gamma-\epsilon)}} = \frac{\frac{\Gamma}{\Phi}k_m^{\epsilon} - k_m^{\gamma}}{\left[\Gamma\left(\frac{n_*}{n_i}\right)^{\epsilon} - \left(\frac{n_*}{n_i}\right)^{\gamma}\right]\Phi^{-\gamma/(\gamma-\epsilon)}} = 1.$$

Substituting the foregoing expression yields

$$\frac{\frac{\Gamma}{\Phi}k_m^{\epsilon} - k_m^{\gamma}}{\Gamma k_*^{\epsilon} - k_*^{\gamma}} = \Phi^{-\gamma/(\gamma-\epsilon)}.$$

This equation has the solution $k_m = k_*\Phi^{-1/(\gamma-\epsilon)}$, which establishes (ii) since $\Phi = \Phi_*$ by (i).

Appendix B: Competitive land developers

Local land developers attract urban dwellers by offering (possibly negative) subsidies, s, such that the net utility is weakly larger than the economy-wide utility level, denoted by u_d, where subscript 'd' stands for 'developers'. They collect aggregate land rents in the site they develop. Each developer takes the federal fiscal parameters τ, δ, and ρ as given. A developer who owns site a solves the following maximization problem:

$$\begin{aligned}\max_{n,s}\ \pi(n,s;a,u_d) &= (1-\delta)\gamma(1-\rho)n^{\gamma+1} - ns \\ \text{s.t.}\quad u_d &\leq (1-\tau)an^{\epsilon} - (1-\delta)(1+\gamma)n^{\gamma} + s,\end{aligned}$$

where the first term in π is the aggregate land rent collected and the second term is the cost of the subsidy to attract n people to the site. Local land developers set agents at their reservation utility, i.e., $s = u_d - (1-\tau)an^{\epsilon} + (1-\delta)(1+\gamma)n^{\gamma}$. Plugging s into the profit and optimizing with respect to n, the first-order condition is given by

$$a(1-\tau)(1+\epsilon)n^{\epsilon} - (1-\delta)(1+\rho\gamma)(1+\gamma)n^{\gamma} - u_d\left[1+\mathcal{E}(u_d,n,\cdot)\right] = 0, \tag{B-1}$$

if $n(a)$ is positive (the term on the left-hand side is negative otherwise). In (B-1), $\mathcal{E}(\cdot)$ denotes the elasticity of u_d with respect to n, and the dot stands for the choices of the other developers. In our case with a continuum of sites, developers are atomistic and have no impact on the aggregate variable u_d (like in the case with monopolistically competitive firms). Hence, $\mathcal{E}(u_d, n, \cdot) \equiv 0$ for all developers, so that condition (B-1) simplifies to

$$a(1-\tau)(1+\epsilon)n^{\epsilon} - (1-\delta)(1+\rho\gamma)(1+\gamma)n^{\gamma} = u_d \tag{B-2}$$

for all developed sites. Observe that (B-2) is isomorphic to (A-5) if $\tau = \delta = \rho = 0$, i.e., in the absence of federal taxation, with u_d instead of μ_*. Then, $n_d(a) = n_*(a)$ if $u_d = \mu_*$. Note that $u_d > \mu_*$ is not possible (the allocation with developers cannot Pareto dominate the optimal allocation by definition). Assume hence that $u_d < \mu_*$. Then, cities are more populated and fewer than at the optimal allocation. This implies that land developers who own empty land with a slightly below a_* can attract workers by offering them a utility higher than u_d and yet make profits. To see this, use (B-2) to solve for u_d and substitute into the definition for land developer profit to see that the profit is positive for $n_d(a) > n_*(a)$. Thus, $u_d = \mu_*$ and the allocation with developers coincides with the federal optimum allocation.

It is easy to verify that the equilibrium profits of land developers are strictly positive for all $a > a_*$. When land is homogeneous, the equilibrium profits of land developers are zero. Here, positive rents remain because sites are (vertically) differentiated goods.[24]

Last, note that in the presence of federal taxation, (B-2) and (A-5) no longer coincide. In that case, the allocation with developers is no longer optimal. The reason is that the developers take into account the fiscal wedges when attracting people to their sites. If τ and δ are large, developers have to offer high subsidies to attract agents — especially on sites with high a — but are heavily taxed by the federal government on their profits (the land rents). Hence, large cities may end up being too small.

Appendix C: Parametrization

C.1. Economic parameters. For the elasticity of agglomeration economies, ϵ, we consider 0.03 from French estimates that control for sorting (Combes, Duranton, and Gobillon, 2008). For the elasticity of urban congestion, γ, we use 0.25. This corresponds roughly to an elasticity of 4

[24]It is useful to make the analogy between our problem at hand and that of imperfect competition between firms. In our model, developers differ by 'productivity' (i.e., the quality of the site they own), but varieties (i.e., sites) are viewed as perfect substitutes by mobile workers (consumers). The developer with the lowest cost (i.e., the best site) cannot capture the whole market because her production cost is convex in city size (because of increasing urban costs). This limits the size of any city. Since all agents eventually have to end up in some location — they can 'opt out' of the urban system, but there are decreasing returns in the rural area — this implies that some of the less efficient developers also end up developing sites. The outcome with developers is optimal when developers cannot strategically manipulate u_d, because for any site there always exists another one that is arbitrarily similar, which constrains developers since sites are viewed as perfect substitutes by mobile agents.

percent with respect to wage income (3 percent with respect to all income) from Combes et al. (2016), after dividing it by 15 percent. The 15 percent figure accounts for both time and material costs of commuting. According to the American Time Use Survey, time spend traveling to work is equal to 9 percent of time spent working. Valuing commuting at the same cost as working, then this constitutes a cost equal to 9 percent of the wage income. The remaining 6 percent is accounted for based by the observation that automobile users spend 4.9 percent of their total income on commuting, or close to 6 percent of wage income. These numbers are admittedly approximations. The price of gas varies substantially over time. Modes of public transportation involve considerable public subsidies. With $\gamma = 0.25$, the income share to land is $1/4$ of this 15 percent figure, or 3.75 percent of income. This value is close to that implied by Combes et al. of 4.1 percent.[25]

For robustness, we also consider $\epsilon = 0.06$, closer to traditional estimates (e.g., Ciccone and Hall, 1996), and $\gamma = 0.50$, based on the traditional mono-centric city model with linear transportation costs. See also Rosenthal and Strange (2004) and Melo, Graham, and Noland (2009), who review the literature and report estimates of ϵ in the range of $[0.01, 0.8]$. The value of $\gamma = 0.5$ is used recently in Saiz (2010) and Desmet and Rossi Hansberg (2013). For alternative values of ϵ and γ see also Glaeser and Gottlieb (2008), who consider dis-economies other than commuting. They cannot reject the hypotheses that all of the elasticities are constant.

C.2. Fiscal parameters. As our base case, meant to approximate the U.S., we choose a tax rate of $\tau = 0.34$, following Albouy (2009). This rate incorporates the average marginal income tax rate (net of federal transfers), but also the net-of-benefit burden of payroll tax rates (Medicare and Social Security), as well as an average sales tax rate applied to consumption.

The value of $\delta = 0.17$ accounts for similar tax advantages to owner-occupied housing and commuting. Note that the true discount may be separated into two components $\delta_1 \gamma n^\gamma + \delta_2 n^\gamma$, where δ_1 applies to land, which we assume is treated like housing, and δ_2 to transportation. Owner-occupied advantages include the mortgage interest deduction, limits on capital gains taxes, and the lack of sales taxes on housing consumption. For itemizers, taxes appear to lower the user cost of housing by 17 percent; or 24 percent when property taxes are excluded (Albouy and Hanson, 2014). This number is lower by 8 percentages points for non-itemizers. Furthermore, one third of households (accounting for about 27 percent of the value of the housing stock) is rented, and receives far less favorable tax treatment. Thus the lower value of $\delta_1 = 17$ percent appears reasonable. The discount for transportation costs is roughly similar

[25]As an additional check, note that the elasticity of population is n_p with respect to a is $(\gamma - \epsilon)^{-1} = (0.25 - 0.03)^{-1} = 4.55$ A demand shock from an increase in a increases wages by the elasticity $1 + \epsilon (\gamma - \epsilon)^{-1} = \gamma (\gamma - \epsilon)^{-1} = 1.13$. Thus the implicit supply elasticity of population with respect to wages is $1/\gamma = 4.0$. Theory implies that this elasticity for n_p is a lower bound for long-run labor supply elasticities. Most estimates of this elasticty are typically in this vicinity. Absent other frictions, this is consistent with higher, but not lower values of γ.

due to the fact that the time costs of commuting are not taxed. Since time costs, according to our numbers, account for 60 percent of transportation costs, this would result in a discount of $\delta_2 = (0.6)(0.34) = 0.20$. If commuting time is valued somewhat positively, this number would be closer to 0.17, and so we take this value as approximately similar to δ_1. See Albouy and Lue (2015) for more.

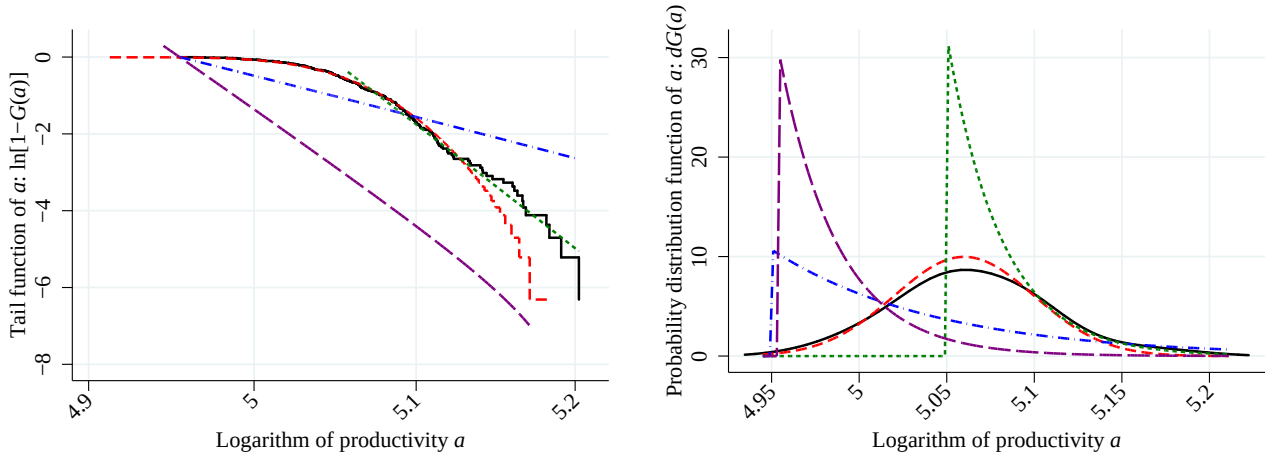
Finally, $\rho = 1$ is based on the assumption that migrants either pay rent or purchase land (i.e., housing) in the city they migrate, too. While we do not model the ownership of land, such ownership in a 'home city' would not negate the opportunity cost of not selling the home to another potential migrant. In other words, $\rho = 1$ applies just as well to owner-occupiers who sell their house in an origin city, while buying a house in the destination city. Turning to a stylized developing economy, we choose $\tau = \delta = \rho = 0$ based on the idea that: (i) income taxes are low, or tax evasion is ubiquitious, so that $\tau = 0$; (ii) transfers and subsidies for land and commuting are inexistent, i.e., $\delta = 0$; and (iii) property rights are weakly enforced so that many urban migrants can squat on land, i.e., $\rho = 0$. Note that both numbers are extreme cases, and that a more realistic parametrization would likely produce values strictly between 0 and 1.[26]

C.3. Distribution of productive amenity, a. We next choose a parametrization for the distribution $G(a)$ of sites. There appears to be considerable heterogeneity in productivity across U.S. cities. We estimate $G(a)$ from U.S. population and wage estimates which control for worker characteristics (see Albouy, 2016). We shrink the estimates by a factor of 2/3 to correct for possible unobserved sorting (Glaeser and Maré, 2001). The productivity parameters are inferred from $a = wn^{-\epsilon}$, and they are independent of any solution concept for city sizes (see Section 3). The empirical (productivity) distribution, depicted by the solid black line in Figure 12, appears to be close to log-normal, with a log-mean of 5.06 and a log standard deviation of 0.04. This empirical distribution should not be taken as the literal distribution $G(a)$ of potential sites, however, since we observe only a truncated distribution (the sites that are developed). There are likely to be an abundant number of sites with low productivity. Yet, most of them are not observed as there are no cities on them. Also, sites may be inhabited for advantages other than productivity, such as quality of life. The lower the productivity of a site is, the better it must be in other positive *unobserved* characteristics for it to be observed in our sample, i.e., for a city to be developed there. With this insight, we model $G(a)$ with a Pareto distribution, whereby the frequency of sites a diminishes with their quality at a constant rate. The rationale for using a Pareto distribution is that: (i) it is virtually indistinguishable from the

[26]We note that there is a possibility that some fraction of land value may be appropriated by property taxes and redistributed to local migrants in the form of valuable, congestible, public goods, such as public schools. For example, if property taxes account for 15 percent of housing user costs, this would imply a fraction of $\rho = 0.15$. How much differential land values are converted in practice to valuable local expenditures is an important question for future research.

empirical log normal distribution in the upper tail, thus providing a good approximation of the distribution of high-productivity sites where we observe cities; and (ii) it has a much larger mass on low-productivity sites which, as argued above, are clearly underrepresented in the observed sample of city sites. While we do not — by definition — observe the sites on which there are no cities, the idea that there are 'many' of them implies that the Pareto distribution provides a good approximation.

Figure 12: Fitted values of $\ln[1-G(a)]$ and $g(a)$ with $\epsilon = 0.03, \gamma = 0.25$.



Empirical distribution (solid) with ε = 0.03, γ = 0.25; Log normal (dash): mean = 5.06, s.d.= 0.04; Pareto (dash−dot): shape = 10.7, lnscale(min) = Pareto top 150 cities (shortdash) shape = 32.2, lnscale(min) = 5.05 Truncated Pareto (longdash) shape = 30.0, lnscale(min,max) = 4.95, 5.20

With a Pareto distribution, the tail distribution $\ln[1-G(a)]$ of the true distribution should decrease at a constant rate. Fitting the entire empirical distribution produces a distribution with a relatively small Pareto parameter of 10.7, thus implying that the distribution is relatively dispersed. But if we look instead at the upper 150 cities, located to the left of the mode, the estimated Pareto parameter is 32.2. This is likely to be more accurate, since superior sites are more likely to be inhabited and therefore observed. Furthermore, the fit of the upper tail appears to be far more accurate. We thus extrapolate a similar shape parameter of 30 for the entire distribution of a, using a truncated Pareto distribution that begins at the lowest value of $\ln a = 4.95$ and stops at $\ln a = 5.20$.[27] That distribution is depicted by the purple long-dashed distribution in Figure 12, and we henceforth take it as our benchmark. We also consider Pareto parameters that are twice or half the size — thereby implying different degrees of heterogeneity in the distribution of sites — for comparison.

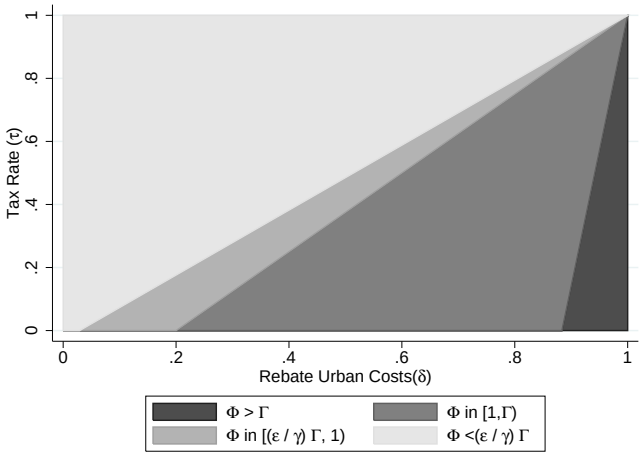
C.4. Fiscal parameters and city sizes. Figure 13 depicts the regions of fiscal parameter values for τ, δ, and ρ that generate the different cases summarized in Table 1 for our benchmark values

[27] By the central limit theorem, Lee and Li (2013) show that a is log-normally distributed when it is the product of numerous amenities that enter multiplicatively into a. The authors do not consider the case of truncated distributions, however.

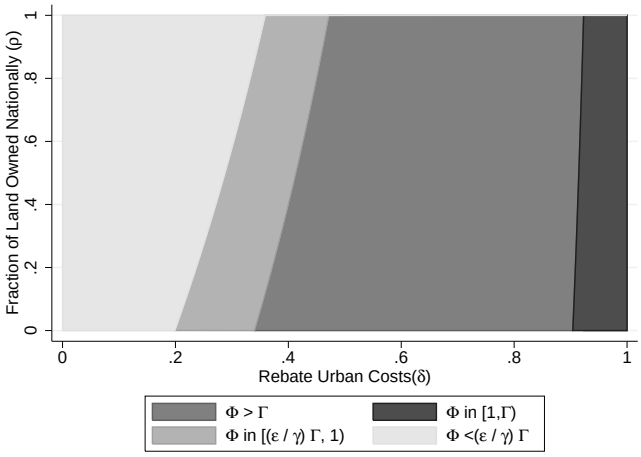
of $\epsilon = 0.03$ and $\gamma = 0.5$. The left panel graphs the congestion rebate parameter, δ, against the federal tax rate, τ, assuming $\rho = 1$. The unit square is divided into four gray-shaded regions: the two darkest delimit the parameter ranges for which free-migration leads to oversized cities.

Figure 13: Fiscal policies and expected solutions.

(a) Congestion rebate and federal tax rate.



(b) Land-ownership and congestion rebate.



Observe that the darkest shade of grey delimits the parameter space for which cities are too few and oversized at both the free migration equilibrium and the local politics allocation. This case occurs for a narrow range of fiscal parameters that does not seem to characterize any fiscal system, such as a discount rate that is higher than the tax rate. The lighter two gray zones characterize combinations for which local politics leads to too many cities, with undersized large cities and trivially oversized small cities. The free migration allocation also features these properties for parameters in the lightest area. This areas includes values $(\tau, \delta) = (0.34, 0.17)$ that apply to the U.S., and most developed OECD countries, which feature significant taxes and property rights over land. As Figure 13 shows, this configuration arises for a 'large' subset of parameters values and can, therefore, not be dismissed as a 'rare special case'.

The right panel of Figure 13 varies the land-ownership parameter ρ and the congestion rebate δ. It emphasizes the distorting role of local land ownership on the urban system at the local politics allocation. Local incentives to reduce the size of a city increase in δ and decrease in ρ. Thus, the local politics allocation is more likely to feature undersized (large) cities for parameter configurations that belong to the north-west space. Observe also that when rebates to urban costs are low — as seems empirically the case — undersized large cities are again more likely to occur.