Public goods and congestion in a system of cities: how do fiscal and zoning policies improve efficiency?†

Alex Anas*† and David Pines**

*Department of Economics, State University of New York at Buffalo, Amherst, New York, NY 14260, USA  
**The Eitan Berglas School of Economics, Tel-Aviv University, Ramat-Aviv, Israel
†Corresponding author. Alex Anas, Department of Economics, State University of New York at Buffalo, Amherst, New York, NY 14260, USA. email: alexanas@buffalo.edu

Abstract

How can fiscal and density-zoning instruments help finance local public goods, while mitigating congestion in a general equilibrium system of cities? The literature—based on the model of a single monocentric city with congestion but no public good—shows core densities increasing and total land area decreasing if Pigouvian tolls or urban growth boundaries (UGBs) are used. In the utility-improving regimes of our setup, including tolls or the UGB, core densities are lowered in each city but more and smaller cities emerge at long-run equilibrium. If lot size and other goods are sufficiently complementary in consumption, it is optimal that total land area increases in the aggregate even as congestion tolls or the UGB decrease the total area of each city. Density zoning is either redundant or achieves first-best efficiency depending on the fiscal instruments present.

Key words: Public goods, traffic congestion, optimal city size, fiscal policies, urban growth boundaries, zoning policies

JEL classifications: D61, D62, H23, H44, R13, R14, R41, R48, R52

Date submitted: 15 October 2011 Date accepted: 24 July 2012

1. Introduction

How should urban size, the number of cities and land use within cities be optimally regulated? The answer to this important question must recognize that economies of scale in local public goods (LPGs) fosters concentration, but the benefits of this are offset by costs that rise with city size and by negative externalities. Leaving the externalities un-priced or mitigating them with the wrong instruments or funding the public goods with the wrong instruments can be costly.

† An earlier version (Anas and Pines, 2007, unpublished manuscript) was presented at the 2007 annual meetings of the Urban Economics Association held within the 54th North American Regional Science Association Meetings in Savannah, 8–10 November; at the Taipei Conference on Urban and Regional Economics, National Taipei University, 3–5 December 2007; and at the applied economics seminar of the department of economics, University of California at Riverside in 2008. The authors thank Richard Arnott and Mason Gaffney of Riverside, and Courtney Lafountain the discussant in Savannah for useful comments; and the two anonymous referees of the journal and editor Kristian Behrens for their comments during the review process. Any residual shortcomings are the responsibility of the authors.
Planners frequently defend denser cities to reduce road congestion and encourage transit use (e.g., Ewing, 1997). In the USA, Portland, OR; Seattle, WA; Boulder, CO are subject to urban growth boundaries (UGBs). Similarly restrictive land use controls have become more popular recently in New Jersey and Maryland. UGB or UGB-like land use controls in the exurban or suburban peripheries have also been used in London, Seoul, Moscow, Ottawa, Tianjin and other places around the world.

Congestion pricing—frequently advocated by economists—has been implemented in London (see Leape, 2006), Milan (Rotaris et al., 2010) and Stockholm (Eliasson, 2009), in addition to the much older scheme in land-limited Singapore (see Santos et al., 2004). Aggressive land use policies directly and congestion pricing indirectly control city size in order to improve efficiency, a problem that will continue to remain important as perhaps a billion more people urbanize in the next decades, mostly in Asia. Congestion continues to increase in the world’s megacities located in fast-growing developing nations. Beijing, which is in the early or middle stages of a sustained car ownership boom, is extremely congested with an average road speed of 18 km/h (IAPT, 2001). The fact that congestion has remained un-priced may have caused Beijing to grow too large in population and area. Should the congestion be contained by investing in the densification of cities by building mass transit systems, by zoning or by reallocating population to smaller cities or by the creation of more cities? As half a billion Chinese become urbanized in the future, should China acquire 200, 300 or 400 more cities?

To provide a framework for studying the interaction of urban size, the number of cities, urban public investment, negative externalities and land use regulation, we synthesize several classical concepts. The scale economy in the provision of LPGs explains urban concentration (Stiglitz, 1977), but travel cost increasing with concentration (Fujita, 1989) explains why urban size is limited. Traffic congestion is a negative externality, and commuters bear the average cost of congestion not the full marginal social cost (Walters, 1961; Vickrey, 1963), inviting Pigouvian pricing (Pigou, 1920). This under-pricing of congestion is a major economic distortion (Keeler and Small, 1977); and responsible indirectly, for the overpopulation of urban areas. Urban expansion is also driven by the fact that land is a normal good (Alonso, 1964). Cities meanwhile are not fixed in number but new cities emerge to disperse the population, so that the scale economies achieved from adding population to each city, are balanced at the margin, by the diseconomies of expansion in each city (Henderson, 1974); and—under ideal conditions, when a city’s population is optimally determined—the aggregate profits from land would be sufficient to finance the cost of public expenditure in each city (George, 1879). Combining these classical observations into one model, we explore the welfare and allocation effects of applying alternative fiscal and density-zoning instruments to finance LPGs, while reducing the resource cost of the un-priced congestion externality, when the number of cities is endogenous and the cities are in general equilibrium.

The mixed economy regimes we will study differ by the set of fiscal and zoning instruments that are available to the social planner. In our setup, each city has just two

---

1 Road pricing is endorsed by both European and US government agencies (see, for example, European Comission, 1995; U.S. General Accounting Office, 2003).
2 For a modern exposition of the theorem see, for example, Arnott (2004).
residential areas: a core (where everyone also works) and a suburb. In general equilibrium, the aggregate revenue from all fiscal instruments fully finances the city’s LPG in the core. The fiscal instruments include a congestion toll, an excise tax/subsidy on land that is equivalent to implementing a UGB, a (lump-sum) tax on all profits from land development and a supplementary per-capita tax/subsidy. The zoning instruments are density zoning in the city cores or in the suburbs.

None of our fiscal instruments are ever optimal to use as a subsidy. A tax on the profits from land development is available in all regimes and regardless of the other instruments it is optimal to always extract 100% of these profits whether in lump-sum or ad valorem (the Henry George tax or HGT, hereafter). A first-best optimal allocation is achieved by three alternative sets of instruments: (i) by the HGT and congestion tolling; or (ii) by the HGT, the per-capita tax and density zoning in city cores; or (iii) by the HGT, an UGB that restricts the aggregate suburban land area, and density zoning within the suburbs. Our second-best optimal allocation is achieved by using together the HGT, the per-capita tax and the UGB. A third-best (or fourth-best) optimal allocation is achieved by the HGT and the per-capita tax or by the HGT and a UGB, and the lowest-best (fifth-best) allocation by using only the HGT. That the UGB is used for a second-best allocation sheds new light on the long-standing controversy regarding the merits of a UGB policy in mitigating the negative effect of un-priced congestion.

In the absence of Pigouvian tolling, a restrictive UGB in each city always improves welfare, corroborating the traditional view. But in the allocation of resources, the traditional result that UGBs should cause the densification of population toward the city centers is reversed. Funding the LPGs while eliminating the congestion externality in our first-best regime or partially mitigating it in the UGB regime reduces each core’s density rather than increasing it! The reason for this seemingly surprising result is as follows. The fixed cost of the LPG is the same in any regime. In the first-best regime, the HGT and congestion tolls should fund the LPG, whereas in the lowest best it is just the HGT. Hence, in general equilibrium, the HGT revenue must be lower in the first best than in the lowest best. In both regimes, all of the HGT comes from the city’s core. Since the core’s land area is the same in the two regimes, the rent in the core is lower in the first best. Otherwise a city would generate a fiscal surplus. It then follows from two reasons that core lot sizes are larger and densities are lower in the first-best regime. The first reason is the price effect, that is, the price of land is lower; the second is the welfare (real income) effect, due to welfare being higher in the first than in the lowest best and land being a normal good. Similar explanations apply to the UGB regime in which the excise tax on land and the HGT together fund the LPG.

In Section 2, the extant literature is contrasted with the present study. In Section 3, we present our assumptions, fiscal tools and general equilibrium structure of the model and explain the notation. In Section 4 (relying on proofs gathered in Appendix A), we analyze and welfare-rank the regimes arising from combinations of the fiscal instruments, and compare in more detail the allocations of the first-best, second-best, the UGB, the per-capita tax and the lowest-best HGT regimes. For each regime other than the first-best, we discuss the distortions that occur as the externality is partially mitigated while the public good is financed. In Section 5, we show how the socially optimal city system of each of our regimes is decentralized by utility-taking city developers (relying on proofs in Appendix B). Section 6 adds density zoning in the cores
or suburbs and finds combinations of fiscal and zoning instruments that are redundant. Implications for policy and extensions are outlined in Section 7.

2. Review of the literature

The literature has examined congestion mitigation and LPG finance with models that are partial in several respects. A group of models has dealt with congestion mitigation in a single city in partial equilibrium, ignoring the presence of public goods that cause the urban concentration, while models of city systems that have dealt with the allocation of public investment among cities have ignored negative externalities like congestion.

Strotz (1965) was the first to develop a single-city model showing the efficiency of pricing congestion and its effects on land use assuming all jobs were pinned at the city’s center. Repeated applications of similar models have shown or implied that in response to the first-best remedy of congestion tolls, a city will contract in area and increase in central densities. Wheaton (1998) demonstrated that the first-best allocation in such a partial equilibrium monocentric city without an LPG and open in population with land rents accruing to absentee landlords does indeed achieve densification of the central area and contraction of the city’s total area. He examined curing the distortive effect of un-priced congestion by means of a Pigouvian toll, or an equivalent tax on land development, or by density zoning directly. Joshi and Kono (2009) discussed structural density regulation as a second-best policy in a single congested monocentric city, but without an LPG or rent redistribution. Using this partial equilibrium framework, they argued that structural density regulation improves welfare. In their city, closed in population, optimal zoning requires a minimum (maximum) residential structural density near the center (the suburbs). This is consistent with the traditional view on UGBs because it induces densification toward the center. The authors concluded that in an open city, only a minimum structural density should be imposed.

The recommendation of a restrictive UGB, in a single monocentric city without an LPG was originally motivated by the observation that under un-priced transportation congestion the shadow rent at the monocentric city’s boundary is lower than the market rent (Solow, 1972; Kanemoto, 1977; Arnott, 1979; Pines and Sadka, 1985) such that, under market allocation, the city boundary extends beyond its (second-best) optimum. The above theoretical papers did not formally compare the densification of cities under second-best and first-best policies, but Pines and Sadka (1981) presented numerical calculations that illustrated the relative densification of city centers under first-best, second-best and market allocations. In a policy-oriented article, Brueckner (2000) advocated the UGB as a second-best congestion mitigation policy on the basis that it increased the land use densities in city centers reducing congestion in a monocentric city.

Anas and Rhee (2007) challenged the traditional view by their polycentric city in which jobs can suburbanize in response to congestion tolls. Such a city can have a longer radius under the efficient than under the market allocation if congestion tolls cause a sufficient number of jobs to move closer to suburban workers. Then, they showed that, for a UGB to improve welfare, land should be zoned away from agriculture and into urban use (an expansive UGB), not the other way around. Anas and Pines (2008) studied just two monocentric cities unequal in population due to amenity differences, aggregate land rent shared equally among the residents of both.
First-best optimal Pigouvian tolling reduces the large but increases the small city’s area, as it shifts population from the larger city to the smaller. In the absence of Pigouvian tolling, the second-best policy requires a restrictive UGB in the larger but an expansive UGB in the smaller city. Under the first- or second-best policy, the aggregate urban area of the two cities can increase and their average density can fall if the elasticity of substitution between lot size and other goods is low or if preference for the amenity is high.

Abdel-Rahman and Anas (2004), a survey of the systems of cities literature, find no contributions that treat congestion. Helpman and Pines (1981) treated a system of monocentric cities without congestion. Under free population mobility among the cities, each emerges at a site with a unique natural amenity augmented by public investment. In an optimal regime, the cities in the high amenity sites—if developed—invest more in an LPG and are more populated and higher in density.

This article is more general than these previous studies in several respects. First, instead of dealing with either a single open or closed city, the present paper deals with cities that are open to in- or out-migration, while the system of cities is closed in population and the number of cities is endogenous. Second, each city has an LPG to fund and a congestion externality to mitigate. Third, the labor and land markets are in general competitive equilibrium implying fiscal balance in the city system. Fiscal or density-zoning instruments affect the allocation between a city’s core and its suburb in the intensive margin, and work in the extensive margin to change the number of cities. Therefore, this article provides a more complete context for the ongoing debate on the merits of urban densification versus suburban dispersion.

3. The model setup

We treat a closed economy with \( N \) exogenously given urban consumer-workers (residents, hereafter), identical in preferences and initial endowments and distributed among \( m \) identical and endogenously determined cities. At equilibrium, each city has \( n = N/m \) residents, \( n_1 \) residing in the core (zone 1) with fixed land area \( H_1 \) and \( n_2 = n - n_1 \) residing in the suburb (zone 2) with an endogenous land area \( H_2 \). Residents work in the core of their city of residence but can freely change cities or change residence between core and suburb of the same. The unitary labor endowment of each resident is the only input in production. Each unit of labor produces a unit of a composite good, so a city produces \( n \), and all cities \( N \) units.

\( u(x_i, h_i) \) is differentiable (hence, continuous), and strictly quasi-concave utility function, where \( x_i \) is the quantity of the composite good (numeraire) and \( h_i \) is the resident’s lot size in zone \( i \) (\( i = 1 \) core, \( i = 2 \) suburbs). Both are normal goods. Commuting requires private and public transportation technologies. Within the core, it is rendered costless by an LPG which costs \( k \) and includes a mass transit system. Suburbanites incur road congestion cost only when they cross into the core, where they access the mass transit system. The congestion cost rises strictly convexly with suburban traffic, \( t(n_2) \), (with \( t'(n_2) > 0 \) and \( t''(n_2) > 0 \)).

The opportunity cost of land in non-urban uses (or the cost of converting each unit of free raw land to urban use) is \( r \) and exogenous. At this cost, there is a perfectly elastic supply of suburban land and, therefore, suburban land rent \( R_2 = r \) at equilibrium. But if a UGB is implemented using an excise tax on land at a rate \( s \), then the after-tax rent is
$R_2 = r + s$. In the core, because the supply of land is inelastic (and equal to $H_1$), the excise tax is perfectly capitalized into rent and neutral. Thus, $R_1 = \bar{R}_1 + s$, where $\bar{R}_1$ is the after-tax rent. We use $R_1$.

Residents move freely and reach equilibrium utility in any core or suburb: $u_1 = u_2 = u$. Equal $u$ implies $R_1 > R_2$ and $h_1 < h_2$ because suburban residence entails congestion, but the larger suburban lot (at the lower rent) compensates for this. In the core, using transit to avoid congestion puts land at a premium, and lots are smaller, making densities higher.

Perfectly competitive markets determine $u, R_1, R_2, n_1, n_2, n$ (and $m$) while the fiscal and zoning instruments are set by a benevolent central planner who maximizes the equilibrium utility, $u$ and funds the core LPGs with the funds from the fiscal instruments. Each resident has before-tax income that is equal to his unitary output equilibrium utility, $u$, and zoning instruments are set by a benevolent central planner who maximizes the total net per-capita tax paid by a resident will be $R_1 - r$.

We begin with fiscal regimes. We will add density zoning later. The fiscal instrument menu is: (i) a congestion toll, $\tau$, on each suburban resident; (ii) an excise tax/subsidy on all urban land at the rate $s$ to implement a UGB and (iii) a lump-sum per-capita tax/subsidy $\Psi + \Omega$, where $\Psi = ADLR/n$ and $\Omega$ is a supplementary per-capita tax or subsidy. This is equivalent to taxing 100% of the profits from the core’s development before or after shareholders (residents) receive them and then using $\Omega < 0$ to distribute some, all or more than all of the profits equally among the shareholders, or levying $\Omega \geq 0$ as a per-capita tax in addition to $\Psi$, the 100% tax on profits from land. Thus, the total net per-capita tax paid by a resident will be $\Psi + \Omega$, and if $\Omega = 0$ then only all of the profit shares from land are taxed (the HGT). If $\Omega > 0$, then more than the profit shares from land are taxed. If $-\Psi < \Omega < 0$ then only a part of the profit shares from land are taxed, and if $\Omega < -\Psi$, then not only the profit shares from land are not taxed but a subsidy per-capita is paid from revenues that must be raised by some other fiscal instrument(s).

Any feasible fiscal mixed regime will be \{u, R_1, n, n_1, n_2, n; m; \tau, s; \Psi + \Omega, \Psi = (R_1 - r)H_1 / n\}. $\tau, s$ and $\Omega$ are unrestricted in sign a priori. The general equilibrium conditions are

\begin{align*}
    n_1 x(R_1, u) + (n - n_1)(x(r + s, u) + h(r + s, u)r + t(n - n_1)) + H_1 r + k - n &= 0, \\
    n_1 h(R_1, u) - H_1 &= 0, \\
    E(R_1, u) + \Omega - 1 &= 0,^3 \\
    E(r + s, u) + t(n - n_1) + \tau + \Omega - 1 &= 0, \\
    mn - N &= 0.
\end{align*}

$x(., .), h(., .)$ are the compensated demands for the composite good and lot size, and $E(., .)$ is the minimum expenditure function: $E(R_1, u) = x(R_1, u) + R h(R_1, u)$. Equation (3.1) is market clearing for the composite good and shows that output $n$ is exhausted by the direct demand for the composite good, the units needed to pay for suburban commuting,

---

3 More precisely, Equation (3.3): $E(R_1, u) + \Psi + \Omega = 1 + ADLR/n$, where $\Psi = ADLR/n$. And similarly for Equation (3.4).
feasible allocations and fiscal instruments

6 The intuition about Walras’s Law holding is that in a closed system the economy’s aggregate income must be exhausted in paying for all private and public expenditures.

4 Since each unit of composite good is produced by one unit of labor, Equation (3.1) also expresses labor market clearing.

5 The suburban market always clears by \( n_2 h(r+s,u) - H_2 = 0 \) since \( H_2 \) is endogenous.

6 The intuition about Walras’s Law holding is that in a closed system the economy’s aggregate income must be exhausted in paying for all private and public expenditures.

\[
n_\Omega + H_1(R_1 - r) + (n - n_1) h(r + s, u)s + (n - n_1) \tau - k = 0. \tag{3.6}
\]

Equation (3.6) is a fiscal balance in each city given any \( \tau, \Omega, s \). In the next section (supported by the proofs of the Lemmas in Appendix A), we will prove that in any optimal regime, no fiscal instrument, if available, will ever be used as a subsidy. Then, Equation (3.6) implies that the planner taxes 100% of the ADLR, and in some regime may raise even more revenue by using as taxes, at least one of the remaining four equations can be solved for the four-unknown equilibrium allocation \( \{u, n, n_1, R_1\} \) for each feasible choice of \( \Omega, s \) and \( \tau \). We drop Equation (3.1). Then, the planner’s fiscal optimization problem will be to maximize utility, \( u \), subject to the constraint set given by Equations (3.6) and (3.2)–(3.4). These constraints define the set of feasible allocations and fiscal instruments \( \{u, n, n_1, R_1|\tau, s, \Omega\} \). By Equation (3.6) any such feasible allocation satisfies fiscal balance and by Equations (3.2)–(3.4), it satisfies spatial equilibrium. However, a set of fiscal instruments that is feasible does not necessarily eliminate distortions in the economic allocation and could even add new distortions. To eliminate distortions completely or partially, the planner must optimize the values of all or a subset of the feasible fiscal instruments, while the markets determine \( \{u, n, n_1, R_1\} \).

From now on, we will use abbreviated notation when convenient which we now define: \( E_i \equiv E(R_i, u), x_i \equiv x(R_i, u) \). By Shephard’s lemma \( h_i \equiv h(R_i, u) = \partial E(R_i, u)/\partial R_i \).

\[
E_{iu} = \partial E(R_i, u)/\partial u > 0. \text{ Since both goods are normal: } x_{iu} = \partial x(R_i, u)/\partial u > 0 \text{ and } h_{iu} = \partial h(R_i, u)/\partial u > 0. \text{ By the law of demand, } h_{R_i} = \partial h(R_i, u)/\partial R_i < 0. \]

We will normally denote the suburban population by \( n - n_1 (= n_2) \), except where there is no confusion in using \( n_2 \) or where it is simpler to use \( n_2 \).

4. The fiscal regimes

In this section, we will prove the most important properties and the welfare ranking of the alternative fiscal regimes. In each regime, the lump-sum HGT on per-capita profits from land, \( \Psi \), is available. The regimes differ from one another according to which ones...
of the other three fiscal instruments are not available. We will hereafter refer to the
regimes by \( hgt \), \( pct \), \( ugb \), \( pct + ugb \), \( ct \) defined as follows:

1. Henry-George regime \( (hgt) \): all three instruments are not available \((\tau = 0, \Omega = 0, s = 0)\). Because in this regime none of the fiscal instruments are available, utility has its lowest possible value.

   The next three regimes are constrained-optimal regimes because at least one fiscal instrument is not available in each regime. In these regimes, the available instruments can be optimized to partially correct the distortion and utility is improved but not to the full extent possible.

2. Per-capita tax/subsidy regime \( (pct) \): the supplementary per-capita tax or subsidy is available but the congestion toll and the excise tax on land are not available \((\tau = 0, s = 0)\).

3. Excise tax on land or UGB regime \( (ugb) \): the excise tax on land is available (and serves to implement the UGB) but the per-capita tax/subsidy and the congestion toll are not available \((\tau = 0, \Omega = 0)\).

4. Per-capita tax/subsidy and UGB regime \( (pct + ugb) \): the congestion toll is not available but the supplementary per-capita tax/subsidy and the excise tax on land are both available \((\tau = 0)\).

   Finally, there is an optimal regime in which all three fiscal instruments are available. In this regime, the distortion is completely corrected and utility reaches its maximum extent.

5. Congestion toll regime \( (ct) \): the congestion toll, the excise tax on land and the supplementary per-capita tax/subsidy are all available.

### 4.1. The Henry-George regime

Since in the \( hgt \) regime the three instruments \( \tau, s \) and \( \Omega \) are not available to the planner he must rely entirely on the HGT, \( \Psi \), to fund the LPG (setting \( \tau = s = \Omega = 0 \) in the constraint set defined by Equations (3.2)–(3.4) and (3.6)). Then, the four unknowns \( u, n_1, n, R_1 \) are uniquely found from Equations (3.6) and (3.2)–(3.4). After using the HGT, the planner has no degrees of freedom to improve utility.

#### 4.1.1. Uniqueness of the hgt regime.

That the equilibrium allocation of this regime is unique is proved by noting that the allocation’s variables are sequentially determined. Note that when \( \tau = s = \Omega = 0 \) then the fiscal balance condition implies that \( ADLR = k \) and this means \( R_1 = r + k/H_1 \) which determines \( R_1 \) uniquely and explicitly from Equation (3.6). Plugging this into Equation (3.3) and from the fact that \( E(R_1, u) \) increases monotonically in \( u \), the value of \( u \) is implicitly but uniquely determined. Next from Equation (3.2), since both \( R_1 \) and \( u \) have been determined, \( n_1 = H_1/h(R_1, u) \) is unique. And since \( u \) and \( n_1 \) are unique, using these in Equation (3.4) and from the fact that \( l(n - n_1) \) is monotonically increasing in \( n \), \( n \) is also unique.\(^7\)

---

\(^7\) The uniqueness of the other regimes is proved in part (ii) of the Supplementary Appendix, available on the Journal’s web site. In the last paragraph of the concluding section of this article we discuss a non-trivial extension that could yield multiple equilibria.
4.1.2. The effect of the LPG cost in the hgt regime

The effect of \( k \) on the equilibrium is now easily determined. From Equation (3.6), \( dR_1/dk = 1/H_1 > 0 \). From Equation (3.3), \( du/dk = -(h_1/E_{1u})(dR_1/dk) = -h_1/(H_1 E_{1u}) < 0 \). From Equation (3.4), \( dnt_2/dk = -(E_{2n}/t(n_2))(du/dk) = (h_1 E_{2n})/(t(n_2)H_1 E_{1u}) > 0 \) and from Equation (3.2), \( dn_1/dk = -(n_1/h_1)(h_1 R_i(dR_1/dk) + h_{1u}(du/dk)) > 0 \). Hence, since both the core’s and the suburb’s population increases, \( dn/dk > 0 \). In summary, the effect of a higher LPG cost is that each city becomes denser in its core and more congested in its suburbs. There are, therefore, fewer and larger cities given a constant total population \( N \). As suburban population and thus congestion increase, rents in the core increase and utility falls. These results reflect the economy of scale in public good provision: the higher the total cost of the LPG, the larger the population that must be accommodated in a city to raise the ADLR which is necessary to fund the LPG in the hgt regime, since other sources of revenue are not available in this regime.

4.2. Ranking and properties of the optimal regimes

If any of the three fiscal instruments \( \tau, s \) and \( \Omega \) are available, they are chosen to maximize \( u \) subject to fiscal balance (3.6) and the spatial equilibrium conditions (3.2)–(3.4). The Lagrangian normalized by \( \lambda \), the shadow price of \( k \), is

\[
\mathcal{Z} = \frac{u}{\lambda} + n\Omega + n_1 h(R_1, u)(R_1 - r) + (n - n_1)sh(r + s, u) + (n - n_1)\tau - k
\]

\[
- \rho(n_1 h(R_1, u) - H_1) - \theta_1(E(R_1, u) + \Omega - 1) - \theta_2(E(r + s, u) + t(n - n_1) + \tau + \Omega - 1),
\]

where the Lagrangian multipliers \( \rho, \theta_1, \theta_2 \) are the shadow price of the core’s land, and the core and suburban population, respectively, normalized by \( \lambda \). In Appendix A, we present the first-order conditions and using these we prove four lemmas characterizing the solutions of the four utility-improving regimes defined earlier.

Proposition 1 relies on the lemmas of Appendix A to prove the utility ranking of the regimes. The main results of the lemmas and Proposition 1 are summarized in Table 1. The table also summarizes the values of the fiscal instruments at each regime’s optimum. Proposition 2(i) proves that the HGT (a 100% tax on profit shares from land)—which is available in each regime—is used in each regime. The optimal regimes differ by which ones of the other available instruments are also used and at what values. Proposition 2(ii) compares the regimes in how the available instruments are used and the fraction of the externality that is internalized in each regime. Allocations under the hgt, \( ugb \) and \( ct \) regimes (ignoring the \( pct \) and \( pct + ugb \) regimes) are compared in Proposition 3. The welfare economics of all of these regimes are then discussed in the subsection 4.3 focusing on how each constrained optimal regime improves utility while causing other distortions.

**Proposition 1:** \( u^{ct} (a) > u^{pct+ugb} (b) \geq \max(u^{pct}, u^{ugb}) \geq \min(u^{pct}, u^{ugb}) (c) > u^{hgt} \).

**Proof:** At the optimum of some regime \( j \), some instrument(s) are restricted to zero and not available to that regime. Then, by the Envelope Theorem, the local change in utility from relaxing each such restriction at the regime’s optimum is measured by Equations (4.1)–(4.3), the signs of which follow from the proofs of Lemmas 1–4 in Appendix A.
<table>
<thead>
<tr>
<th>Regime name (Abbreviation)</th>
<th>Ranking (Prop. 1)</th>
<th>How obtained from higher ranked regimes</th>
<th>Solutions of the fiscal regimes (Lemmas 1–4, Appendix A)</th>
<th>Optimal taxes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Shadow prices</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Land in the core</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Core resident per capita output</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Suburban resident per capita output</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\rho$</td>
<td>$\theta_1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\theta_2$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Optimal taxes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\tau$ Congestion toll</td>
<td>$\bar{\delta}_2$ Excise tax on land</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\Omega$ Suppl. per-capita tax</td>
<td>$\Psi + \Omega$ Henry-George tax plus suppl. per-capita tax</td>
</tr>
</tbody>
</table>

|                           |                  |                                        | $\rho$                                               | $\theta_1$    |
|                           |                  |                                        |                                                      | $\theta_2$    |
|                           |                  |                                        | Each suburban resident pays                          |               |
|                           |                  |                                        | $\rho$                                               | $\theta_1$    |
|                           |                  |                                        | Each suburban resident pays                          |               |
|                           |                  |                                        | Each resident pays                                   |               |
|                           |                  |                                        | Each resident pays                                   |               |

| Congestion toll (ct)      | First-best       | All instruments available             | $R_1 - r = n_1$                                      | $n_2 \ell(n_2)$ |
|                           |                  |                                        |                                                      | $0$            |
|                           |                  |                                        |                                                      | $= ADLR/n$      |
| Per-capita tax + UGB (pct + ugb) | Second-best | Set $\tau = 0$ in ct regime           | $R_1 - r > n_1$                                      | n.a.           |
|                           |                  |                                        |                                                      | $> 0$          |
|                           |                  |                                        |                                                      | $= ADLR/n$      |
| Per-capita tax (pct)      | Third- or Fourth-best | Set $s = 0$ in pct + ugb regime       | $R_1 - r > n_1$                                      | n.a.           |
|                           |                  |                                        |                                                      | $= ADLR/n$      |
| Urban growth boundary (ugb)| Third- or Fourth-best | Set $\Omega = 0$ in pct + ugb regime | $R_1 - r = n_1$                                      | n.a.           |
|                           |                  |                                        |                                                      | $= ADLR/n$      |
| HGT (hgt)                 | Fifth-best       | Set $\Omega = 0$ in pct or set $s = 0$ in ugb regimes | $R_1 - r = n_1$                                      | n.a.           |
|                           |                  |                                        |                                                      | $= ADLR/n$      |

n.a., fiscal instrument is not available in the regime; HGT is $\Psi = ADLR/n$ in each regime.
where we found that the Lagrangian multipliers $\theta_j^i = n_j^i$ for $j = \text{hgt, ugb}$ and $\theta_j^i < n_j^i$ for $j = \text{ct, pct or ugb}$:

$$\frac{du^j}{ds^2} = \frac{\partial \Omega}{\partial \Omega} \bigg|_{\Omega = 0} = n_j^i - \theta_j^i - \theta_j^2 = n_j^i - \theta_j^2 + n_j^2 - \theta_j^2 > 0 \quad \text{for } j = \text{hgt, ugb}$$ (4.1)

$$\frac{du^j}{ds} = \frac{\partial \Omega}{\partial s} \bigg|_{s^j = 0} = (n_j^i - \theta_j^2)h(r, u^j) > 0 \quad \text{for } j = \text{hgt, pct};$$ (4.2)

$$\frac{du^j}{d\tau} = \frac{\partial \Omega}{\partial \tau} \bigg|_{\tau = 0} = n_j^2 - \theta_j^2 > 0 \quad \text{for } j = \text{hgt, ugb, pct, pct + ugb.}$$ (4.3)

To prove inequality (a) in the statement of the proposition, note that $t^\text{pct+ugb} = 0$, but $t^\text{ct} > 0$ from Lemma 1 and that therefore the sign of Equation (4.1) implies $t^\text{ct} > t^\text{pct+ugb}$. To prove inequality (b) note that $s^\text{pct} = 0$, but $s^\text{pct+ugb} > 0$ from Lemma 2 and that therefore the sign of Equation (4.2) implies $u^\text{pct+ugb} > u^\text{pct}$; and note also that $\Omega^\text{ugb} = 0$, but $\Omega^\text{pct+ugb} > 0$ from Lemma 2 and that therefore the sign of Equation (4.1) implies $u^\text{pct+ugb} > u^\text{ugb}$. To prove inequality (c) note that $s^\text{hgt} = \Omega^\text{hgt} = 0$, but $s^\text{ugb} > 0$. From Lemma 4 and the sign of (ii) implies $u^\text{ugb} > u^\text{hgt}$. Note also that $\Omega^\text{uct} > 0$ from Lemma 3 and that the sign of Equation (4.1) implies $u^\text{uct} > u^\text{hgt}$.

**Proposition 2:** (i) It is not optimal to use any of the available instruments $\Omega, s, \tau$ as a subsidy in any regime, and it is optimal to use the HGT in each regime; (ii) in the first-best regime of congestion tolling $\tau = n_j^i t(n_j^e)$ fully captures the externality, but in any other regime $\Omega^j + s^j h(r + s^j, u^j) = \theta_j^2 t(n_j^e) < n_j^2 t(n_j^e)$, the total tax on each suburban resident falls short of the externality caused by that resident at the regime’s optimum.

**Proof:** (i) Lemmas 1–4 in Appendix A proved that $\Omega^j, s^j, \tau^j \geq 0$ in regimes $j = \text{ct, pct + ugb, pct, ugb}$. Equation (4.1) implies that reducing $\Omega$ from zero at the constrained optima of the hgt and ugb regimes would decrease utility. Therefore, from Equation (3.6) it is not optimal to use any instrument to return any part of the ADLR to the residents, in any regime. Hence, $\Psi + \Omega \geq \text{ADLR}/n$ in each regime, meaning that it is optimal to use the HGT in each regime. (ii) From Appendix A, $\Omega^j + s^j h(r + s^j, u^j) + \tau = \theta_j^2 t(n_j^e)$. In the first-best regime, only $\tau$ is used and Lemma 1 proved $\theta_j^2 = n_j^e$. In the $\text{pct + ugb, pct, ugb}$ regimes $\tau = 0$ and $\theta_j^2 < n_j^i$ was proved in Lemmas 2–4. In the hgt, $\theta_j^2 = 0$.

**Proposition 3:** Comparing the lowest-best (hgt), third- or fourth-best (ugb) and first-best (ct) regimes, the following hold in the cores by the utility ranking of these regimes: (a) rents increase: $R_1^\text{hgt} > R_1^\text{ugb} > R_1^\text{ct}$; (b) lot sizes increase: $h_1^\text{hgt} < h_1^\text{ugb} < h_1^\text{ct}$; (c) core population and density decrease: $n_1^\text{hgt} > n_1^\text{ugb} > n_1^\text{ct}$. In the suburbs: (d) the first-best ct regime’s lot size exceeds the ugb regime’s or lowest-best hgt regime’s: $\max(h_2^\text{hgt}, h_2^\text{ugb}) < h_2^\text{ct}$; (e) the lowest-best regime’s suburban and total population exceeds the first-best’s or the ugb regime’s: $n_2^\text{hgt} > \max(n_2^\text{ct}, n_2^\text{ugb})$, $n_2^\text{hgt} > \max(n_2^\text{ct}, n_2^\text{ugb})$; (f) there are fewer cities in the lowest-best hgt regime than in the ugb regime or the first-best regime: $m^\text{hgt} < \min(m^\text{ct}, m^\text{ugb})$. 

Downloaded from http://jegc.oxfordjournals.org/ by guest on October 6, 2012
Proof: (a) By Proposition 1, $u^{hgt} < u^{ugb} < u^{et}$, and from Equation (3.3) $E(R_{1}^{hgt}, u^{hgt}) = E(R_{1}^{ugb}, u^{ugb}) = E(R_{1}^{et}, u^{et}) = 1$. Since expenditure rises in $R_{1}$ or $u$, it follows that $R_{1}^{hgt} > R_{1}^{ugb} > R_{1}^{et}$. (b) Since $u^{hgt} < u^{ugb} < u^{et}$, $R_{1}^{hgt} > R_{1}^{ugb} > R_{1}^{et}$, then $h(R_{1}^{hgt}, u^{hgt}) < h(R_{1}^{ugb}, u^{ugb}) < h(R_{1}^{et}, u^{et})$ by both the utility and rent effects. (c) From Equation (3.2), $n_{1} = \frac{1}{\rho_{1}}$. From this and (b), $n_{1}^{hgt} > n_{1}^{ugb} > n_{1}^{et}$ follows; (d) by the utility effect, $h(r, u^{hgt}) < h(r, u^{et})$. By both the utility and the rent effects, $h(r + s^{hgt}, t^{hgt}) < h(r, u^{et})$. (e) By Equation (3.4), $E(r, u^{hgt}) + \rho_{2} = E(R_{2}, u^{ugb}) + \rho_{2} = 1$. Since by Proposition 1, $u^{ugb} > u^{hgt}$, by Lemma 4, $R_{2} > r$, and since $E(R_{2}, u)$ increases in utility and rent, and $\rho_{2}$ increases in $n_{2}^{ugb} > n_{2}^{hgt}$. $E(r, u^{hgt}) + \rho_{2} = E(r, u^{et}) + \rho_{2} = 1$. (f) then, since $m = \frac{n}{n}$, $m^{hgt} < m^{et}$ and $m^{hgt} < m^{ugb}$.

As explained in Section 1, Proposition 3 has strong implications in the context of the existing literature. For a single city, this literature has shown repeatedly that pricing the congestion externality causes the densification of a city’s core, and a sharp increase in the core’s rents, as argued most recently in Wheaton (1998). Not only does this literature focus on congestion in a single city rather than a system of cities, but it also ignores LPGs. Proposition 3 shows that the opposite of densification occurs when a fixed LPG investment $k$ is funded in each city. In such a system of identical cities, mitigating the un-priced congestion externality causes more and smaller cities to emerge, each having lower core densities and rents. Proposition 3 proves that this is true whether the congestion externality is mitigated fully by the Pigouvian congestion toll or imperfectly by the excise tax on land to implement a UGB.

4.3. The welfare economics of the optimal regimes

Propositions 1–3 proved formally the most important properties of the utility-improving regimes. We proved that as the fiscal instruments become available, the planner must create more and smaller cities in the extensive margin (Proposition 3). An important property is that at the optimum fiscal balance holds for the optimally chosen planner must create more and smaller cities in the extensive margin (Proposition 3). An important property is that at the optimum fiscal balance holds for the optimally chosen planner must create more and smaller cities in the extensive margin (Proposition 3). An important property is that at the optimum fiscal balance holds for the optimally chosen planner must create more and smaller cities in the extensive margin (Proposition 3). An important property is that at the optimum fiscal balance holds for the optimally chosen planner must create more and smaller cities in the extensive margin (Proposition 3). An important property is that at the optimum fiscal balance holds for the optimally chosen planner must create more and smaller cities in the extensive margin (Proposition 3). An important property is that at the optimum fiscal balance holds for the optimally chosen planner must create more and smaller cities in the extensive margin (Proposition 3).

But observe that in the short run, when the city’s population is exogenously given, one cannot guarantee fiscal balance without transfer to the consumers. For example, in the hgt regime ADLR can exceed $k$ if the city is sufficiently large or fall short of $k$ if the city is sufficiently small. Since $n$ is exogenous, we have five Equations (3.1)–(3.5) or (3.2)–(3.6) to determine the four remaining variables $\{u, R_{1}, n_{1}, m\}$. All five equations cannot be satisfied by choosing values for these four variables. Consider Equations (3.2)–(3.6). Then, either Equation (3.6) must be violated or $\Omega$ must be either positive or negative to satisfy Equation (3.6). So in a short-run optimal allocation, one cannot exclude a situation with $\Omega < 0$ which by the proof of Proposition 2(i) cannot happen in the long-run optimal allocation.

Consider now an example of the transition from the short-run to the long-run optimal allocation. Suppose that the hgt regime holds initially and $n$ is such that ADLR $= k$. Suppose that, starting from this situation, the planner imposes the congestion toll $\tau$. If the added revenue generated by $\tau$ is retained by the planner, in the short run the city’s population being fixed, some suburban population must move to
the core which will raise core rent until a new spatial equilibrium is established with a higher rent in the core and higher commuting cost inclusive of the toll in the suburb. At this stage utility declines, as proved in part (iii) of the Supplementary Appendix on the Journal’s web site. The planner should use his financial surplus to establish a new city at the setup cost of $k + rH_1$ attracting population from both the cores and the suburbs of the preexisting cities, reducing rent in the cores and the congestion suffered by the population that remains in the suburbs. This relocation will more than compensate the population in the preexisting cities for the initial decline in utility caused by the imposition of $\tau$, as we know from the proof of Proposition 1.

Since the planner is interested in maximizing utility without keeping any surplus for himself, he would observe that utility can be improved by reducing the population of each city by some amount until the planner gathers enough people from preexisting cities to form a new city (the extensive margin of population adjustment). The planner does this by using the retained surplus in the short run to establish a new city to which consumers will move both from the cores and the suburbs of existing cities. This would reduce congestion and core rents, thus improving utility. Because core rents fell, the ADLR would also fall and the fiscal surplus would now be less. The planner would then increase the instrument even closer to its optimal value and would create another city, continuing in this fashion until the regime-optimal value of the instrument is reached. That would be the long-run general equilibrium regime in which the number of cities and the size of each city are optimal and fiscal balance holds in each city.

It is worth noting that even without establishing new cities, the utility can be improved by redistributing the toll revenue in the existing cities. The resulting allocation, though preferable to the original hgt regime, is inferior to the long-run one associated with the increase in the number of cities. What was illustrated here about $\tau$, also applies to the case of adding $s$ to the hgt regime though in this cases only a constrained optimum would be achieved. Adding the per-capita tax instrument $\Omega$ to the Henry-George regime and then redistributing its revenue equally cannot increase short-run utility (the net effect is zero), but long-run utility is again increased by using the revenue from $\Omega$ to set up new cities.

We will now discuss how, in each optimal regime, the fiscal instruments cause reallocations in the direction of efficiency and how the instruments also cause secondary distortions. We will use the above explanations about city creation in the extensive margin as we discuss the economics of each of the regimes. In doing so, we will understand better the results of Propositions 1–3.

### 4.3.1 Congestion toll (first-best; ct regime)

The Pigouvian toll achieves the maximum utility improvement by fully internalizing the congestion externality. The toll captures the difference between the marginal social and average private cost of total congestion, imposed by each commuter on all other commuters. Equation (3.6) shows that at the regime’s long-run optimum aggregate tolls $(n_{ct}^2)'t' (n_{ct}^2)$, supplement the ADLR to fund $k$. This result is a generalization of the Henry-George rule: the sum of the pure social profits, evaluated using marginal social costs, should vanish at a finite positive optimal population (Berglas and Pines, 1981). The sources of pure social profits are the LPG, urban land and transportation. The pure social profits of the LPG are $-k$. The revenue from land is $H_1R_1 + n_2h(r, u)r$ and the social cost is $(H_1 + n_2h(r, u))r$. Then, the pure social profit is the ADLR. Social revenue
from transport is $n_2(t(n_2) + n_2t'(n_2))$, social cost is $n_2t(n_2)$ and social profit is $(n_2)^2t'(n_2)$. Pure social profit $(R_1 - r)H_1 + (n_2)^2t'(n_2) - k$ vanishes at the optimum (by Equation (3.6)).

How does imposing Pigouvian tolling in each lowest-best city induce reallocation? In the short run, some suburban residents must avoid the congestion toll by relocating to the core of the same city since the number of cities and the size of each is fixed. This increases the demand for land in the core, and the core rents and densities. But, before the congestion toll the lowest-best city had fiscal balance by $(R_1 - r)H_1 = k$. By Equation (4.1), the toll causes a positive fiscal surplus to emerge in the short run. Once fiscal balance is established in each city, congestion tolls pay for some of the LPG cost. Hence, rents in city cores become lower than before the imposition of the congestion tolls in the hgt regime. A lower rent and a higher utility means that each core resident demands a bigger lot, and hence core densities become lower also, and since the core’s land is fixed, core population decreases. The new cities become populated not only by those who relocated from the suburbs of initially existing cities to avoid the toll but also by some of the residents that left the cores of these cities.

4.3.2. Per-capita tax (pct regime) versus the UGB (ugb regime)

When the congestion toll is not available, we proved in Proposition 1 that using either the supplementary per-capita tax on all residents or the suburban land tax to implement a UGB improves welfare. The impacts of the per-capita tax and the suburban land tax differ because of the income and substitution effects of these taxes. These instruments do not do as well as congestion tolling because they work on the wrong margins. Hence, on the one hand, they do not take care of the externality directly, on the other hand, they cause distortions in the markets in which they have direct effects.

The suburban land tax achieves its imperfect corrective action mainly through its substitution effect. Suburban land becomes too expensive inducing suburbanites to locate out of the suburbs, which reduces congestion by making suburbs smaller. The per-capita tax is like a club fee paid for gaining membership in a city (Scotchmer, 1986). The income effect of this tax makes residents poorer so as new cities are created the direct effect of the per-capita tax on income is offset by the reduced suburban congestion.

4.3.3. UGB by suburban land taxation (third- or fourth–best; ugb regime)

The suburban excise tax on land is a blunt tool for internalizing the congestion externality, as it does so only indirectly by working on the suburban land market which is not the source of market failure. In the short-run, some residents relocate to the cores of their cities to avoid the suburban land tax. This raises rents in the core and since land tax revenue from the suburb is also available, the result is a fiscal surplus. In the long run, the fiscal surplus of existing cities is dissipated by reallocating population to new UGB cities, but since in each city the suburban land tax revenue supplements the ADLR from the core, the long-run rents in city cores will be lower than before the imposition of the UGBs. The suburban land tax does not capture all of the congestion externality (Proposition 2(ii)) and consequently, the land tax revenue does not raise short-run fiscal surplus as much as the congestion toll revenue does. That is why in the long run, fewer new UGB cities are needed to dissipate the fiscal surplus and that is why
core rents and densities in the \textit{ugb} regime remain higher than those under the first-best tolling in the \textit{ct} regime (Proposition 3).

**4.3.4. Per-capita tax (third- or fourth–best; \textit{pct} regime)**

The per-capita tax also works to reduce suburban congestion, but again imperfectly because it does so by working on the margin of city membership rather than directly on the congestion externality.

Notably, the per-capita tax is not neutral since our general utility function allows income effects. Furthermore, the per-capita taxes paid by the resident of the core and the suburb have different income effects. The per-capita tax is more onerous on the suburban resident who is already encumbered by the commuting cost. In the short run, when the city’s population is fixed, the per-capita tax causes a population reallocation to the core, thus the core rent and densities rise and the city generates a fiscal surplus. Again, the fiscal surplus is dissipated in the long run by the planner creating more cities with per-capita taxes.

If the per-capita tax were paid only by the suburban residents, then since the externality is confined to them, the tax could be set to fully internalize the externality, that is it could be set as $\Omega^{pct} = n^2_t f(n^2_t)$, which is the exact level of the Pigouvian toll in the first-best regime. But because the per-capita tax is also paid by the core’s residents who do not cause congestion, a distortion is created. In order to see why, note that while the per-capita tax levied on suburbanites induces some of them to relocate to the cores in the short run, the same per-capita tax paid by the core residents inefficiently induces some of them to want to locate in the suburbs where they would add to congestion. Since the per-capita tax does not capture all of the congestion externality (Proposition 2(ii)), the per-capita tax revenue does not contribute to short-run fiscal surplus as much as the congestion toll revenue does. Therefore, in the long run, fewer cities with per-capita taxes are needed to dissipate the fiscal surplus and that is why core rents and densities under the per capita-tax remain higher than core rents under the first-best tolling regime.

**4.3.5. UGB and per-capita tax in the second-best (\textit{pct} + \textit{ugb} regime)**

One of our results (Proposition 1, Lemma 2 in Appendix A) is that using the suburban land tax and the per-capita tax together to supplement the HGT gives a second-best allocation, whereas using only one of these taxes results in a third- or fourth-best allocation. The reason is that using only one tax assumes the other is not available. Thus, the \textit{ugb} regime is a constrained optimum where constraints $\Omega = 0$ and $\tau = 0$ are both binding. The \textit{pct} regime is a constrained optimum where constraints $s = 0$ and $\tau = 0$ are both binding. Keeping $\tau = 0$, but removing the other binding constraint improves utility (Proposition 1).

That the two taxes working together improve efficiency is an instance of the general theorem of the second best (Lipsey and Lancaster, 1956): when a Pareto efficiency condition cannot be fulfilled, in our case when the tolling instrument $\tau = 0$, then a suboptimal allocation can be reached only by deviating from all the other conditions of Pareto efficiency. For us this means that at the second-best optimum, the market price of suburban land will deviate from its marginal cost, $r$, and that the market price of land
in the core will deviate from the shadow price of land in the core. These deviations will be measured by Ramsey formulas which we will now derive.

We denote an equilibrium of the second-best regime, conditional on \( \{\Omega, s\} \), as the allocation \( \{n^e, R_1^e, n^f, n_t^f\} \equiv \{u(\Omega, s), R_1(\Omega, s), n(\Omega, s), n_1(\Omega, s)\} \). This together with \( \{\Omega, s\} \) solves Equations (3.6) and (3.2)–(3.4). Accordingly, Equations (3.3) and (3.4) are now rewritten:

\[
E(R_1(\Omega, s), u(\Omega, s)) + \Omega - 1 = 0, \\
E(r + s, u(\Omega, s)) + t(n_2(\Omega, s)) + \Omega - 1 = 0.
\]

By differentiating Equations (4.4) and (4.5) totally:

\[
h_1 \left( \frac{\partial R_1}{\partial \Omega} d\Omega + \frac{\partial R_1}{\partial s} ds \right) + \frac{\partial E_1}{\partial u} \left( \frac{\partial u}{\partial \Omega} d\Omega + \frac{\partial u}{\partial s} ds \right) + d\Omega = 0,
\]

\[
h_2 ds + \frac{\partial E_2}{\partial u} \left( \frac{\partial u}{\partial \Omega} d\Omega + \frac{\partial u}{\partial s} ds \right) + t' \left( \frac{\partial n_2}{\partial \Omega} d\Omega + \frac{\partial n_2}{\partial s} ds \right) + d\Omega = 0.
\]

To isolate the marginal effect of the excise tax on land at the second-best optimum, we use \( \frac{\partial u}{\partial \Omega} = \frac{\partial u}{\partial s} = 0 \) and set \( d\Omega = 0 \) to keep \( \Omega \) constant. Then, Equation (4.7) reduces to

\[
h_2 = -t' \frac{\partial n_2}{\partial s}.
\]

Using Equation (4.8) in the first-order condition Equation (A.2) of Appendix A and using \( s = R_2 - r \), yields

\[
s n_2 h_2 R_2 = (n_2 - \theta_2) t' (n_2) \frac{\partial n_2}{\partial s}.
\]

The left side of Equation (4.9) is the marginal social excess burden (deadweight loss) at the regime optimum, while the right side is the welfare gain from the marginal decline in the congestion cost caused by the marginal decline in the suburban population induced by the tax \( s \). Note that if we were evaluating this in the first best, the same expression would produce zero marginal benefit since, in that case, the Lagrangian multiplier \( \theta_2^f = n_t^f \) (Lemma 1): with a fully internalized externality in the first-best optimum, imposing an additional tax on suburban land would not generate any marginal benefit. But because in the second-best optimum, there is no congestion toll, there is a marginal benefit to reducing the congestion externality indirectly by the tax on suburban land. That this tax is beneficial in the second best is reflected by \( \theta_2^{pec+ugb} < n_t^{pec+ugb} \) and \( 0 < n_t^{pec+ugb} - \theta_2^{pec+ugb} < n_t^{pec+ugb} \) is the part of the marginal congestion reduction benefit that the planner could still achieve by levying a Pigouvian congestion toll at the second-best regime.

Now imposing \( \frac{\partial u}{\partial \Omega} = \frac{\partial u}{\partial s} = 0 \) and \( d\Omega = 0 \) on Equation (4.6) gives \( \frac{\partial R_1}{\partial s} = 0 \) revealing that with the per-capita tax constant, the suburban land tax has no marginal effect on land rent in the core. This is also why core residents do not appear in Equation (4.9).

Next, dividing Equation (A.2) across by \( R_2 \), using \( s = R_2 - r \) and rearranging, we obtain

\[
\frac{s}{R_2} = \left( \theta_2 - n_2 \right) \frac{h_2}{n_2 h_2 R_2} = \frac{n_2 - \theta_2}{n_2 \eta_{h2;R2}},
\]
where $\eta_{h_2;R_2}$ is the own-price compensated demand elasticity for lot size in the suburbs. Equation (4.10) is the Ramsey (1927) rule when there is no cross elasticity between lot size in the suburb and the core. Equation (4.10) shows that the more inelastic is the compensated demand for suburban land, the higher is the optimal excise tax rate on land.

Turning now to the role of the per-capita tax in the second-best regime, we recall again that even though it applies equally to the core and the suburb, it does affect the allocation since the income effects are different in the city and in the suburb. Recall also that if only the suburban residents were levied such a tax, then its optimal value would be

$$\eta_{h_2} = \frac{\partial n_2}{\partial \Omega} = 0$$

and since this achieves the first-best optimum, no other tax would be needed. But because the per-capita tax is also paid by the core residents, this introduces a distortion away from the first-best allocation which is only partially offset by the excise tax on land in the suburbs. To show the marginal effects of the per-capita tax, we analyze its marginal cost and benefit (when the excise tax on land, $s$, is kept constant).

To that end we now set, $\frac{\partial u}{\partial \Omega} = \frac{\partial u}{\partial s} = 0$ and $ds = 0$ in Equations (4.6) and (4.7) to obtain:

$$\frac{\partial R_1}{\partial \Omega} = \frac{1}{h_1},$$

(4.11)

and

$$-t' \frac{\partial n_2}{\partial \Omega} = 1.$$ 

(4.12)

Dividing the first-order condition (A.6) by $h_1$ and substituting from Equations (4.11) and (4.12):

$$n_1((\rho + r) - R_1)h_1 \frac{\partial R_1}{\partial \Omega} = (n_1 - \theta_1)t'(n_2) \frac{\partial n_2}{\partial \Omega}.$$ 

(4.13)

The left side of Equation (4.13) is the marginal social cost of $\Omega$. It is the deadweight loss due to the deviation of the core’s market land rent, $R_1$, from its shadow value, $\rho + r$, at the optimum of the second-best regime, from a marginal increase in $\Omega$. The right side is the marginal benefit from the decrease in $n_2$, induced by the per-capita tax, $\Omega$. Dividing Equation (4.13) by $R_1$ and rearranging:

$$\frac{(\rho + r) - R_1}{R_1} = \frac{n_1 - \theta_1}{n_1} = \frac{n_1 - \theta_1}{n_1|\eta_{h_1;R_1}|}.$$ 

(4.14)

Equation (4.14) is again a Ramsey rule. This, with Equation (4.10) and $n_1 - \theta_1 = \theta_2 - n_2$ from Equation (A.3), gives

$$\frac{s}{R_2} \frac{(\rho + r) - R_1}{R_1} = \frac{n_1|\eta_{h_1;R_1}|}{n_2|\eta_{h_2;R_2}|}.$$ 

(4.15)

The left side is the relative deviation of the second-best rents from the opportunity costs of land. Observe that Equation (4.15) is similar to the original version of the Ramsey rule and unlike that in the extensions by Diamond and Mirrlees (1971) and Atkinson and Stiglitz (1980), cross elasticities are absent. The reason is that in our

---

8 This follows from the requirement in urban models that each consumer can consume land at only one location implying a non-convex consumption set. Hence, the rent on land in one location does not directly affect the demand for land in the other. If the utility included more goods, then cross elasticity could not be overlooked.
context there are many individuals who are ex ante identical but ex post they are distributed by self-selection into two groups: \( n_1 \) who reside in the core and consume lots only there and \( n_2 \) who reside and consume a lot in the suburb only. The deviation of the market price in the core is from the shadow rent on land in the core, not from the fixed (producer’s) rent as it would be in the original Ramsey rule. The reason for this is that deviations must be measured from the socially optimal prices, and in our case, the socially optimal and market rents on land in the core or in the suburb diverge because of the externality.

4.4. The case of Leontief preferences

Since the proof of Proposition 1 relied on a smooth and strictly concave utility function, which we assumed at the outset, Propositions 1–3 and all of the above discussions excluded the special case of Leontief preferences. In that special case, the first-best allocation can be achieved by any one of the instruments. Proposition 1 is modified to

\[ u^{ct} = u^{pct+ugb} = u^{pct} = u^{ugb} > u^{hgt}. \]

The basic intuition is that lot size and the composite good are perfect complements and so there are no substitution effects but only income effects. The per-capita tax is paid by all residents. Because there are no substitution effects and because the supply of land in the core is inelastic, the per-capita tax now becomes perfectly capitalized into core rents and is neutral on the core’s residents. Then, the per-capita tax paid by suburban residents can be set to have exactly the same income effect as did the congestion toll and works as a perfect instrument for correcting the externality. In the case of the suburban land tax, it does not cause any distortion in the suburban land market since there is no substitution effect. Again it can be set so that the total land tax per consumer has exactly the same income effect as did the congestion toll.

A question remains whether, in any of the welfare improving regimes, the reduction in the population of the cores and the suburbs established by Proposition 3, causes the aggregate land area of the city system to increase. Consider, for example, the \( ugb \) regime. As this regime raises the market rent of suburban land, there is no substitution effect under Leontief preferences. The higher rent makes a direct negative contribution to the income effect, but since the higher rent also causes relocation out of the suburbs, it indirectly reduces the congestion cost thus having a positive income effect, the net being positive for utility to increase. Again looking at the case of Leontief preferences

\[ u^{ct} = u^{pct+ugb} = u^{pct} = u^{ugb} > u^{hgt}, \]

\[ h^{ct} = h^{pct+ugb} = h^{pct} = h^{ugb} > h^{hgt}. \]

Aggregate land is

\[ A^{ct} = N h^{ct}, \quad A^{pct+ugb} = A^{pct} = A^{ugb} > A^{hgt}. \]

Thus, the utility-improving fiscal regimes always increase the aggregate land area while reducing the land area and the externality in each city. Departing from Leontief preferences, as long as income effects are strong enough relative to substitution effects, by continuity any regimes that are utility improving can increase the aggregate land area.

5. Decentralization with developers

In order to find the optimal allocations, we used the device of a benevolent social planner who allocates resources globally over the city system. It behooves us to

---

9 A full algebraic proof is in part (iv) of the Supplementary Appendix available on the Journal’s web site.
demonstrate that the optimal allocations can also be attained without such a planner, if each city is created and fiscally managed by a utility-taking developer who maximizes profits from his city, allocating resources only locally. This decentralization device has been used in the literature (Fujita, 1989, pp. 153–160), despite the fact that developers rarely build entire cities.

We make the following standard assumptions: (i) each developer can manage only one city; (ii) each developer is a utility-taker, that is too small to influence the equilibrium level of utility that prevails in the city system and equivalently there are enough cities/developers; (iii) each developer has access to the same menu of fiscal instruments that were available to the social planner according to regime; (iv) in the short run, the number of developers (cities) are fixed and (v) in the long run developers can enter or exit freely and the number of cities is endogenously determined.

In the short run any developer maximizes the left side of Equation (3.6), the city’s profit given $u$, with respect to $R_1, n_1, n, \tau, s, \Omega$ and subject to the constraints (3.2)–(3.4). Equivalently, the developer sets the fiscal instruments and lets the city’s competitive markets determine the rent in the core and the population split between core and suburb, while also setting the total population of the city so that profit is maximized at the given level of utility. The Lagrangian of this problem is identical to that of our social planner (see Appendix A). The only difference is that the term $u/\lambda$ drops out since the developer takes $u$ as given. The first-order conditions are therefore identical to those of the social planner at the same $u$. Hence, in the developer’s short run, under each fiscal regime, choices of the fiscal instruments are identical to those of the benevolent planner and so are the allocations and equilibrium rents.

As we saw in our analysis of the long-run social optimum, given the maximum utility the planner can achieve, the maximum surplus of the city is zero. For, otherwise, the utility could be increased by redistributing the surplus to the residents. It must also be true that if the utility is lower than what the benevolent planner can achieve ($u_1 < u^*$ in Figure 1) the developer earns positive profits and if the utility is higher than what the benevolent planner can achieve ($u_2 > u^*$ in Figure 1), then the developer incurs a loss. Both possibilities, however, are unsustainable. In the first case, two market forces drive utility up from $u_1$ toward $u^*$. The first market force is that the profits attract new developers and, moving costs being zero, more identical cities emerge with fewer residents in each, because residents from existing cities relocate to the new ones. This effect is a decrease in the supply of residents available to a developer. The second market force is that the residents needed to maximize the profit of the developer increase and this is an increase in the developer’s demand for city residents. These two complementary forces drive up utility until the optimal city size is established with optimum utility $u^*$ and zero profit. In the second case, the market forces will operate in the opposite direction, driving utility down from $u_2$ toward $u^*$ until once again, the optimal city size is established with optimum utility and zero profit. In Figure 1, these adjustments occur on the locus $aa'$ which connects the profit maximizing city sizes at various levels of equilibrium utility.

---

10 Fujita’s Proposition 5.2 on p. 155, deals with the equivalence of a developer city and an optimal city when each city requires a fixed cost to be created as in our case, but congestion and alternative fiscal instruments are not treated.
We need to prove two properties illustrated in Figure 1. The first is that as utility increases the profit curves in Figure 1 indeed shift downwards with $u$. This is easily proved for any of our regimes, using the envelope theorem on the Lagrangian function of the developer which, as stated earlier, is identical to the Lagrangian function of the planner without the term $u/C_2$:

\[
\frac{d\pi}{du} = \frac{\partial \Pi}{\partial u} = (R_1 - r - \rho)n_1h_{1u} + (R_2 - r)n_2h_{2u} - n_1(x_{1u} + R_1h_{1u}) - n_2(x_{2u} + R_2h_{2u})
\]

\[
= -(r + \rho)n_1h_{1u} - mn_2h_{2u} - n_1x_{1u} - n_2x_{2u} < 0.
\]

(5.1)

The sign holds as long as both lot size and the composite good are normal goods as assumed. The second property to prove is that, profit maximizing city size decreases as utility increases, that is the locus $aa'$ in Figure 1 is positively sloped. To prove this, a comparative statics analysis can be performed in each of our fiscal regimes to show that $\frac{dn}{du} < 0$. This is done in Appendix B.

6. Zoning

Under which of the fiscal regimes already examined, does adding density zoning in the core or in the suburbs improve utility? Adding zoning instruments is harmless, because the planner is free to ignore these instruments if they are redundant. Clearly, zoning is redundant in the first-best $ct$ regime as the Pigouvian toll leaves no market failure. But the lower ranked regimes need to be re-examined to see whether utility is improved by a lot size that deviates from the consumer’s free choice in that regime.

Introducing zoning requires modifications of the equilibrium conditions (3.6) and (3.2)–(3.4). When the planner’s chosen lot size is imposed on a consumer, we should allow that the equality at equilibrium between the market rent and the $MRS_{x,h}$ (marginal rate of substituting $x$ for $h$) need not hold. The implied rent on a land unit is the cost of the lot to the consumer divided by the lot size, and in general this deviates from the $MRS_{x,h}$. Our fiscal regimes entailed the free choice of lot size. In the case of zoning, the lot’s size $h_i, i = 1, 2$ decided by the planner is given to the consumer, who expresses his market choice by the bid rent, $C_i, i = 1, 2$ for the zoned lot. It is assumed
that a zoned lot is indivisible and that zoned lots cannot be combined. Given the equilibrium utility, \( u \), consumption, \( x \), is implicitly determined from \( u(x_i, h_i^2) = u \), as \( x_i = \hat{x}(h_i^2, u), i = 1, 2 \).

In one regime, zoning lots in the cores and in the other, zoning lots in the suburbs is available. All the fiscal instruments except the congestion toll are available in both cases. The first regime (zoning lots in the core) is a set \( \{ n_1, n_2, C_1, u | s, \Omega, h_1^2 \} \) that satisfies:

\[
(n_1 + n_2)\Omega + (n_1 C_1 - H_1 r) + n_2 h(r + s, u)s - k = 0, \tag{6.1}
\]

\[
n_1 h_1^2 - H_1 = 0, \tag{6.2}
\]

\[
\hat{x}(h_1^2, u) + C_1 + \Omega - 1 = 0, \tag{6.3}
\]

\[
E(r + s, u) + t(n_2) + \Omega - 1 = 0. \tag{6.4}
\]

The regime of zoning suburban lots is a set \( \{ n_1, R_1, n_2, u | C_2, \Omega, h_2^2 \} \), \( s = \frac{C_2 - r h_2^2}{n_2} \), that satisfies:

\[
(n_1 + n_2)\Omega + H_1(R_1 - r) + n_2 (C_2 - h_2^2 r) - k = 0, \tag{6.5}
\]

\[
n_1 h(R_1, u) - H_1 = 0. \tag{6.6}
\]

\[
E(R_1, u) + \Omega - 1 = 0. \tag{6.7}
\]

\[
\hat{x}(h_2^2, u) + C_2 + t(n_2) + \Omega - 1 = 0. \tag{6.8}
\]

**Proposition 4:** If all the fiscal and zoning instruments except congestion tolls are available, then the first-best allocation can be achieved by using the HGT and either: (i) zoning lots in the core to their sizes in the first-best allocation, while using the per-capita tax at the level of the Pigouvian congestion toll, and not using the excise tax on land: \( h_1^2 = h(R_1^t, u^t) \), \( \Omega = n_2^t t(n_2^t), s = 0 \) or by (ii) zoning the suburban lots to their sizes in the first-best allocation, using the excise tax on land at the level of the Pigouvian congestion toll per unit of land and not using the per-capita tax: \( h_2^2 = h(r, u^t), s = n_2^t t(n_2^t)/h_1^2, \Omega = 0 \).

**Proof:** (i) \( \left\{ n_1^t, n_2^t, h(R_1^t, u^t) R_1^t - n_2^t t(n_2^t), u^t | s = 0, \Omega = n_2^t t(n_2^t), h_1^2 = h(R_1^t, u^t) \right\} \) satisfies Equations (6.1)–(6.4). At this choice of instruments, equilibrium is violated by any \( u^t \) other than \( u^t \), because as \( E(r + s, u) \) with \( s = 0 \) and \( \hat{x}(h_1^2, u) \) increase in \( u \) the markets will lower \( n_2 \) and \( C_1 \) for Equations (6.3) and (6.4) to hold. Then the left side of Equation (6.1) will decrease, deviating from zero because \( s = 0, n_1 = H_1/h_1^2 \) is fixed while \( H_1 r + k \) is exogenous and \( \Omega = n_2^t t(n_2^t) \) and fixed.

(ii) \( \left\{ n_1^t, R_1^t, h(r, u^t) \left( r + \frac{n_2^t t(n_2^t)}{h(r, u^t)} \right), u^t | s = \frac{n_2^t r(n_2^t)}{h(r, u^t)}, \Omega = 0, h_2^2 = h(r, u^t) \right\} \) satisfies Equations (6.5)–(6.8). At this choice of instruments, equilibrium is violated by any utility level other than \( u^t \), because as \( E(R_1, u) \) and \( \hat{x}(h_2^2, u) \) increase in \( u \), the markets will
lower \( n_1, n_2, R_1 \) so that Equations (6.6)–(6.8) continue to hold. Then, the left side of Equation (6.5) will decrease and deviate from zero because \( R_1 \) and \( n_2 \) decreased while other terms in Equations (6.5) are either zero, constant or fixed at the regime’s optimum.\(^{11}\)

Part (i) implies that zoning lots in the core in either the \( pct + ugb \) or the \( pct \) regimes achieves the first-best allocation and the excise tax on land becomes redundant. Part (ii) implies that zoning lots in the suburbs in either the \( pct + ugb \) or the \( ugb \) regimes achieves the first best and the per-capita tax becomes redundant.

The intuition of why zoning helps achieve the first-best allocation is as follows. Consider first (i). In this case, by zoning the lot in the core, the planner achieves perfect capitalization of the per-capita tax paid by the core’s residents. That is, because the consumer is not free to adjust the lot’s size, the lot’s cost in the core adjusts downward by the exact amount of the per-capita tax and there is no distortive effect in the composite good market since the after-tax cost of the lot in the core is unchanged. Hence, the per-capita tax is neutral on the core’s residents. Meanwhile, the perfect capitalization of the per-capita tax in the core frees the per-capita tax in the suburb to play the role of the Pigouvian toll. Hence, the regime achieves first-best allocation if only the core’s lots are zoned to their first-best values and the per-capita tax is set to match the first-best Pigouvian toll.

Turning now to (ii), the intuition is similar. By fixing the lot size in the suburbs, the planner assures that when the suburban tax per lot is imposed there will be no substitution effect in the land market. Then all that is needed is to set the tax per lot at the same value as that of the Pigouvian toll and to zone suburban lots to the first-best value.

**Proposition 5:** If the HGT is available, but: (a) congestion tolls, suburban lot zoning and the per-capita tax are not, then zoning lots in the cores is redundant and cannot improve the \( ugb \) regime; (b) congestion tolls and zoning lots in the cores and the excise tax on land are not available, then zoning lots in the suburbs is redundant and cannot improve the \( pct \) regime. (c) if congestion tolling, the excise tax on land and the per-capita tax are unavailable, neither zoning in the cores or the suburbs can increase the utility of the hgt regime.

**Proof:** (a) Maximizing utility subject to Equations (6.1)–(6.4) and \( \Omega = 0 \),

\[
\frac{\Delta \ln x(h^1, u)}{\Delta h^1} = MRS_{x,h} = \frac{C_1}{h^1} \text{ which implies } h^1_j = h\left(\frac{C_1}{h^1}, u\right).
\]

Then zoned and demanded lot sizes are equal. Renaming \( \frac{C_1}{h^1} \) as \( R_1 \) and \( h^1_j = h(R_1, u) \) and using these in Equations (6.1)–(6.4), we see that they become the equilibrium conditions of the \( ugb \) regime. (b) We maximize utility subject to Equations (6.5)–(6.8) and \( s = 0 \) finding that

\[
\frac{\Delta \ln x(h^2, u)}{\Delta h^2} = MRS_{x,h} = \frac{C_2}{h^2}.
\]

The rest is similar to (a). (c) The result of (a) continues to hold when we add \( s = 0 \) and the result of (b) continues to hold when we add \( \Omega = 0 \).

To see the intuition we start with (c) and use the intuition behind Proposition 4 that zoning becomes beneficial because it eliminates key substitution effects allowing the suburban land tax and the per-capita tax to function as Pigouvian tolls. Since in

\(^{11}\) An alternative but longer method of proof of (i) is to maximize \( u \) with respect to \( \{n_1, n_2, C_1, u, s, \Omega, h^1\} \) and for (ii) to maximize \( u \) with respect to \( \{n_1, R_1, n_2, u, C_2, \Omega, h^2\} \).
(c) fiscal instruments are not available, zoning lots toward their first-best values cannot increase utility. In the case of (a) and (b), the available fiscal instruments are not those that can be made to work with the available zoning instruments. Thus in (a), if the planner zones lots in the core by deviating from the utility maximizing lot size, utility falls while the suburban land tax causes a distortion in the land market. In (b), if the planner zones suburban lots utility is again reduced, while the per-capita tax remains distorteive.

7. Implications for policy and extensions

We revised the commonly held notion that fully or partially mitigating congestion causes cities to become more compact and denser in city cores. Our cores never get denser when congestion is fully mitigated by Pigouvian tolls or partially mitigated by UGBs or by the per-capita tax. Each city shrinks in population and is less dense in the core, while residents of existing cities with initially un-priced congestion are accommodated in more cities with fully or partially mitigated congestion. Thus at the extensive margin congestion is reduced by incurring higher LPG costs to create more cities. In the face of congestion a dispersive land use policy improves efficiency. For a rapidly growing nation like China where large cities are extremely congested, such a dispersive policy may be appropriate.

We also showed that the aggregate urbanized land area increases when income effects are sufficiently stronger than substitution effects, even as each city’s urban area is reduced by an optimal policy. Urban sprawl has been used to refer to urban land area, to low urban densities or to the number (dispersion) of settlements. In our model, all three measures move together and it is efficient for urban sprawl to increase when income effects are strong enough. Since optimal policies are lacking in the real world, this suggests that more sprawl could and should come with more efficiently managed urban development. The correct path of action for planners, of course, is to aim policies at mitigating congestion and to financing the public good efficiently, accepting whatever consequences may follow regarding sprawl.

We generalized our results to an alternative model with decentralization, where the planner is replaced by numerous profit-maximizing developers each forming one city in the system. As long as these developers do not behave strategically but rather take prices and utility as given, the outcomes of the planning model are identical to that of the profit-maximizing developers under the same instrument set that is available to the planner in a particular optimal regime. Imperfect competition among developers is an important extension of the fundamental model provided in this article when the number of cities is small and each city developer possesses market power. In that case, the market failures of the LPG and the local congestion would be joined by the market failure arising from the imperfect competition among city developers. Then, in such a world, how can overall efficiency improve by national level social planning? How should such a planner make transfers among cities (fiscal federalism), regulate the number of cities or otherwise induce city developers to act more efficiently?

An important extension of our results may be obtained by assuming a more realistic LPG cost function. Suppose that a planner (or a city developer) can choose the level of investment in the LPG and that a higher level of investment benefits the consumers or the productivity of the economy. A high level of LPG investment would support a city
that is large in population and more congested and a low level of LPG investment would support a city that is smaller in population and less congested. There are several interesting questions about such a setup: (i) can such cities of different congestion and LPG levels co-exist in an equilibrium of unequal city sizes? (ii) If such an equilibrium exists is it unique and if there are multiple equilibria what does their stability and efficiency depend on? (iii) In such a system of unequal city sizes, what are the various optimal regimes, how should the fiscal and zoning instruments be used and what kinds of fiscal transfers must be implemented by the social planner to achieve the optimal allocation in the city system?

**Supplementary Data**

Supplementary data for this paper are available at *Journal of Economic Geography* online.

**References**


Appendix A: Solution of the optimal fiscal regimes: Lemmas 1–4

The first-order conditions are

(A.1) \[
\frac{\partial \bar{x}}{\partial \tau} = n_2 - \theta_2 = 0,
\]

(A.2) \[
\frac{\partial \bar{x}}{\partial s} = (n_2 - \theta_2)h(r + s, u) + s n_2 \frac{\partial h(r + s, u)}{\partial R_2} = 0,
\]

(A.3) \[
\frac{\partial \bar{x}}{\partial \delta_2} = n - \theta_1 - \theta_2 = n_1 - \theta_1 + n_2 - \theta_2 = 0,
\]
\[ \frac{\partial \hat{\theta}}{\partial n} = \Omega + sh(r + s, u) + \tau - \theta_2 t'(n_2) = 0, \]  
(A.4) 
\[ \frac{\partial \hat{\theta}}{\partial n_1} = (R_1 - r - \rho)h(R_1, u) - sh(r + s, u) - \tau + \theta_2 t'(n_2) = 0, \]  
(A.5) 
\[ \frac{\partial \hat{\theta}}{\partial R_1} = (n_1 - \theta_1)h(R_1, u) + n_1(R_1 - r - \rho) \frac{\partial h(R_1, u)}{\partial R_1} = 0. \]  
(A.6) 

Equations (A.1)–(A.6), \( \frac{\partial \hat{\theta}}{\partial u} = 0 \) \(^{12}\) (3.6) and (3.2)–(3.4) are 11 equations in \( \{u, n_1, n, R_1, \tau, \Omega, s, \rho, \theta_1, \theta_2, \lambda\} \). If, in a regime, one or more of \( \tau, \Omega, s \) are not available, they are set to zero and the corresponding first-order conditions are dropped. Note first that in the \( hgt \) regime none of the three instruments are available. Hence, Equations (A.1)–(A.3) drop out. From Equation (A.4) \( \theta_2 = 0 \) and then from Equation (A.5) \( R_1 = r + \rho \). Then, from Equation (A.6) \( n_1 = \theta_1 \).

**Lemma 1 (congestion tolling, ct):** When all fiscal instruments are at hand, then only the congestion toll and the HGT are used: \( \tau^{ct} = n_2^2 t'(n_2^2) \), \( s^{ct} = 0 \), \( \Omega^{ct} = 0 \).

**Proof:** From Equation (A.1), \( \theta_2 = n_2 \); then using this fact in Equation (A.2), \( s = 0 \) since \( \frac{\partial h(r+s,u)}{\partial R_2} < 0 \), and from Equation (A.3) \( \theta_1 = n_1 \), since \( \theta_2 = n_2 \). Then, from Equation (A.6), \( R_1 = r + \rho \), since \( \frac{\partial h(R_1, u)}{\partial R_1} < 0 \) and since \( \theta_1 = n_1 \). From Equation (A.5), since \( R_1 = r + \rho \), and \( \theta_2 = n_2 \), and \( s = 0 \), it follows that \( \tau = n_2^2 t'(n_2) \). Then, from Equation (A.4), \( \Omega = 0 \) since \( s = 0 \), and \( \tau = \theta_2 t'(n_2) \). Since \( \Omega = 0 \), \( \Psi = \text{ADLR}/n \), the HGT tax, is used without the supplementary per-capita tax/subsidy.

**Lemma 2 (per-capita tax and ubg, pct + ubg):** When congestion tolling is not an available instrument \( (\tau = 0) \), but the HGT, the excise tax/subsidy on land (for implementing the UGB) and the per-capita tax are available, all are used and the last two are used as taxes: \( s^{pct+ubg} > 0 \), \( \Omega^{pct+ubg} > 0 \).

**Proof:** Since \( \tau = 0 \), Equation (A.1) is dropped and we set \( \tau = 0 \) in the remaining equations. Suppose that \( s \leq 0 \). Then, since \( \frac{\partial h(r+s,u)}{\partial R_2} < 0 \), Equation (A.2) implies \( \theta_2 \geq n_2 \). Then, Equation (A.3) implies \( \theta_1 \leq n_1 \). From this, since \( \frac{\partial h(R_1, u)}{\partial R_1} < 0 \), Equation (A.6) implies \( R_1 \geq r + \rho \). But from Equation (A.5), using \( 0 < n_2 \leq \theta_2 \) and \( s \leq 0 \), we get that \( R_1 < r + \rho \) contradicting that Equation (A.6) implied \( R_1 \geq r + \rho \), hence contradicting our supposition that \( s \leq 0 \). Hence, \( s > 0 \) and Equation (A.2) imply \( \theta_2 < n_2 \); Equation (A.3) implies \( \theta_1 > n_1 \) and then Equation (A.6) implies \( R_1 - r - \rho < 0 \). Adding Equations (A.4) and (A.5), \( \Omega + (R_1 - r - \rho)h_1 = 0 \) which implies that \( \Omega > 0 \) because \( R_1 - r - \rho < 0 \). Therefore, \( \Psi + \Omega > \text{ADLR}/n \). Hence, the HGT is supplemented by the per-capita tax and since \( s > 0 \), a restrictive UGB is also used.

**Lemma 3 (per-capita tax, pct):** When only the HGT and the per-capita tax/subsidy are available \( (\tau = s = 0) \), then both are used, and the per-capita subsidy is used as a tax: \( \Omega^{pct} > 0 \).

---

\(^{12}\) We ignore \( \partial \hat{\theta}/\partial u = 0 \). While it holds, all of our results are derived without using it.
Proof: Since $\tau = s = 0$, we drop Equations (A.1) and (A.2) and we set $\tau = s = 0$, in the remaining equations. Then, solving Equations (A.3)–(A.6), we get $\theta_2 = h_2^2(\bar{h}_2^2 - n_1 h \bar{R}_t t'(n_2))^{-1} n_2 < n_2$, and $\Omega = \theta_2 t'(n_2) > 0$. Therefore, $\psi + \Omega > \ADLR/n$. Hence, the HGT is supplemented with the head tax. Also, since $\theta_2 < n_2$, Equation (A.3) implies $\theta_1 > n_1$, and Equation (A.5) implies $R_1 - r - \rho < 0$.

Lemma 4 (urban growth boundary, ugb): When only the HGT and the excise tax/subsidy on land are available, $(\tau = \Omega = 0)$, then the excise tax/subsidy to implement the UGB is used a tax $s^{ubg} > 0$ (the UGB is restrictive) and the HGT is also used.

Proof: Since $\tau = \Omega = 0$, we drop Equations (A.1) and (A.3) and we set $\tau = \Omega = 0$, in the remaining equations. Since $\Omega = 0$, the HGT is used without the supplementary per-capita tax. Then, solving Equations (A.2) and (A.4)–(A.6) $\theta_2 = h_2^2 (\bar{h}_2^2 - h_2 R t'(n_2))^{-1} n_2 < n_2$, and from Equation (A.4) $s h_2 = \theta_2 t'(n_2) > 0$. Also Equation (A.5) implies $R_1 - r - \rho = 0$, since $\tau = 0$ and $s h_2 = \theta_2 t'(n_2)$. Then, Equation (A.6) implies $n_1 = \theta_1$.

Appendix B: $\frac{dn}{du} < 0$ in the decentralization by developers

The developer chooses the fiscal instruments $\{\tau, \Omega, s\}$ that are available to maximize the city’s fiscal profit, while the markets determine the allocation (Figure 1) $\{R_1, n_1, n|u\}$:

$$E(R_1, u) + \Omega - 1 = 0,$$

$$E(r + s, u) + i(n - n_1) + \tau + \Omega - 1 = 0,$$

$$n_1 h(R_1, u) - H_1 = 0.$$

(i) In the lowest-best regime, each fiscal instrument is zero (not available for use). Hence, we set $\tau = \Omega = s = 0$ in Equation (B.1)–(B.3). Then, from Equation (B.1), as $u$ is increased, $R_1$ must decrease. From Equation (B.3), since $u$ increases and $R_1$ decreases, the compensated demand $h(R_1, u)$ increases. Therefore, $n_1$ decreases. From Equation (B.2), by the increase in $u, E(r, u)$ also increases. By the decrease in $n_1$, the function $i(n - n_1)$ shifts up for any $n$ because the function is increasing in its argument. Therefore, a lower $n$ is needed to satisfy Equation (B.2). Hence, $\frac{dn}{du} < 0$ in the lowest-best regime. (ii) In the first-best regime, $\Omega = s = 0$ and $\tau = (n - n_1) t'(n - n_1)$. The steps of (i) are repeated to see, in the last step, that the function $i(n - n_1) + (n - n_1) t'(n - n_1)$ shifts up for any $n$ recalling that $t'(n - n_1)$ is also increasing in its argument. Hence, in this regime too, $\frac{dn}{du} < 0$. (iii) In the UGB regime, $\Omega = \tau = 0$. Equations (B.1)–(B.3) satisfy the following relationships. Suppose that $u$ increases. Then from Equation (B.1), recalling that $\Omega = 0$, it is determined independently of any other equation that $\frac{dR_1}{du} < 0$, since the expenditure function is an increasing function of each one of its arguments. Then, from Equation (B.3), it follows that since $u$ is higher and $R_1$ is lower, the compensated demand for lot size $h(R_1, u)$ must be higher (recalling that lot size is a normal good). Therefore, Equation (B.3) implies that $\frac{dH}{du} < 0$. Now assume that starting from $s = 0$, as $u$ is increased monotonically toward $u^{ugb}$, $s$ increases monotonically toward $s^{ugb} > 0$. From the first-order condition with respect to $s$, this would be true when $(s^{ugb} - \bar{s}) > \frac{h_2}{\bar{h}_2/\bar{s}}$. That is when on the path $u^{ugt} \rightarrow u^{ugb}$, the compensated demand for housing in the
suburbs remains inelastic enough, then \( s \) increases monotonically with \( u \). Since \( \frac{dR_1}{du} < 0 \), and \( \frac{dn_1}{du} < 0 \), it then follows from Equation (B.2) that since \( s \) increases monotonically with \( u \), \( E(r + s, u) \) must increase monotonically which implies that \( t(n - n_1) \) must decrease monotonically. But since this function is increasing in its argument and since \( \frac{dn_1}{du} < 0 \), it follows that \( \frac{dt}{du} < 0 \). (iv) In the per-capita tax regime, \( s = \tau = 0 \). We assume again that starting from \( \Omega = 0 \), as \( u \) is increased monotonically toward \( u^{pct} \), \( \Omega \) increases monotonically toward \( \Omega^{pct} > 0 \). Then, from Equation (B.1), we see that \( \frac{dR_1}{du} < 0 \) is necessary to satisfy the equality. Then, since \( \frac{dn_1}{du} < 0 \), as before, it is implied from Equation (B.2), that \( \frac{dt}{du} < 0 \).