# Efficient Urban Sprawl and Alternative Public Finance Policies in a System of Replicable Cities[*]

by

Alex Anas[#] and David Pines[# #]

November 29, 2007

## ABSTRACT

We study a system of replicable (identical) cities with un-priced traffic congestion where each city requires a minimum investment in infrastructure setup costs. We consider alternative planning regimes (or, equivalently, market with profit-maximizing developers), differing from each other according to the public policy for financing the infrastructure. These policies play two roles: 1) providing the revenues that defray the costs of city setup; 2) internalizing the congestion externality. When congestion is unpriced, there are two resource allocation distortions. In the intensive margin, too many workers-residents live in the suburbs; while in the extensive margin, the number of cities is too few. The tax menu we consider includes differential land rent tax, congestion toll, head tax, and suburban land tax. The first-best policy consists of a confiscatory tax on differential land rent (a Henry George tax), together with a toll on traffic congestion. When tolls are not available, the head tax and suburban land tax, should together supplement the Henry George tax. If, in addition to tolling, one of these two taxes is also unavailable, the other should supplement the Henry George tax. We rank these policies according to efficiency, economic sprawl (total travel cost) and geographic sprawl (aggregate land used).We identify conditions under which the efficient policies, reduce congestion by spreading the population among more but smaller cities, inducing more land use and lower average densities on aggregate which is an essential process for reducing the externality of congestion.

**Keywords:** Sprawl, congestion, congestion tolls, taxes, urban growth boundaries, system of cities.

**JEL classification:** D61, D62, H23, H44, R13, R14, R41, R48, R52.

[*] Paper prepared for presentation at the Taipei Conference on Urban and Regional Economics, National Taipei University, Taipei, Taiwan, December 3-5, 2007.

[#] Department of Economics, State University of New York at Buffalo, Amherst, New York, 14260, United States of America. alexanas@buffalo.edu. Tel: 716-688-5816,Fax: 801-749-7805.

[# #] The Eitan Berglass School of Economics, Tel-Aviv University, Ramat-Aviv, Israel. pines@post.tau.ac.il. Tel: 972-3-640-9904. Fax: 972-3-640-9908.

# Efficient Urban Sprawl and Alternative Public Finance Policies in a System of Replicable Cities

by

Alex Anas and David Pines

November 29, 2007

## 1. Introduction

Urban planners and economists alike have viewed urban sprawl as a purely geographic measure. The cost of human interaction resulting from geographic dispersal (an economic measure) has either been ignored completely, or it has been assumed explicitly or implicitly, that such costs must be positively correlated with the geographic dispersal. Such a conclusion is valid in the context of the traditional analytical tool of urban economics, the monocentric model of a single city in isolation. It has been known since the 1970s, for example, that the negative externality of congestion which remains un-priced causes the monocentric city to become more dispersed. The most recent demonstration of this well-known result was by Wheaton (1998) who also showed that significant increases in land use densities near the center of such a city would occur if congestion were to be optimally priced. Much earlier results by Kanemoto (1977), Arnott (1979b) and Pines and Sadka (1985) went a step further and demonstrated that in th absence of tolls, restricting urban expansion by means of an urban growth boundary (UGB) would be a second-best policy that could achieve higher densities in the centers of cities. Although these insights come from a narrow setting, the single monocentric city, it has been assumed that they must be more general. In this spirit, Brueckner (2000) extrapolated the intuition to any city and advocated UGBs and congestion tolls as policy instruments that would raise central densities.

An appropriate theoretical perspective for understanding the relationship between the geographic and economic implications of urban sprawl has emerged only recently. In three earlier papers by Anas and Rhee (2006, 2007) and Anas and Pines (2007), the conclusions reached are quite contrary to the commonly held beliefs summarized above. We put it as follows:

> "*Geographic and economic measures of sprawl are often inversely related. To reduce the economic measure of sprawl planners may need to implement policies that increase the geographic measure of sprawl. And if planners limit geographic sprawl, they often end up increasing economic sprawl.*" (Anas and Pines, 2007).

Indeed, the claim of the above quotation has already been shown to be valid in two different settings. In both settings the geographic spread of urban areas due to the presence of an un-priced congestion externality is less than what would be efficient. Hence, policies exist that can increase geographic sprawl and thus improve efficiency.

The first setting where this was shown is that of Anas and Rhee (2007) in which – contrary to the monocentric setting – workers can be employed in either the urban core or the suburbs. When congestion tolls are levied in order to internalize the negative externality of congestion, more population can switch its workplaces to the suburbs increasing the geographic sprawl of the urban area. Such an increase in geographic sprawl is efficient, not inefficient as assumed by planners or most economists, because the increased geographic sprawl is accompanied by lower average travel time, namely lower economic sprawl, and higher welfare. The authors also considered urban growth boundaries (UGBs) as second-best policies when congestion tolls cannot be implemented, and used simulations to show that when tolls expand the urban area by causing more workers to work in the suburbs, then the second-best UGB policy is to zone land away from agriculture and into urban use, not to restrict urban use as is commonly assumed.

The second setting is that of Anas and Pines (2007) who consider an urban system of two cities that differ in an exogenous amenity and therefore in size, and in which all jobs are located in the urban cores but population is free to relocate between the core and the periphery of any one of the cities or from one city to the others. We showed analytically that when the un-priced congestion externality is internalized, then provided that the two cities are sufficiently unequal in the exogenous amenity, or when the elasticity of substitution between the housing and composite good is sufficiently small, or both, optimal tolling entails more extensive land use than the laissez faire with un-priced congestion. In general, the tolls have two opposing relocation effects. The first, the intra-city effect, is population relocation from the suburb to the core which tends to reduce aggregate urban land use; the second, the inter-city effect, is population relocation from the larger and more congested city to the smaller and less congested city which increases aggregate urban land use. When the two cities are sufficiently unequal in their exogenous amenity, or when the elasticity of substituting housing for the composite good is sufficiently small, or both, the inter-city positive effect dominates the intra-city negative effect on aggregate land consumption. Only when the cities are nearly equal in size because they are nearly equal in their amenities or when the elasticity of substituting housing for the composite good is sufficiently large, or both, the intra-city population reallocation from the suburbs to the cores dominates the intercity reallocation of population and optimal tolling entails lower aggregate land use than does the laissez-faire urban system without congestion tolls. The authors also concluded that UGBs are a second-best policy when tolls are not used, and that the UGB on the larger city should be restrictive while that on the smaller city should be expansive. Subject to the above qualifications, the optimal UGB policy again increases the geographic sprawl, that is the aggregate land use, *above* and reduces the economic sprawl, that is the aggregate transportation cost, *below* their laissez-faire levels.

*The setup*: *a system of identical congested cities*

In the current article, we examine the relationship between geographic and economic sprawl in a new setting: a system of identically sized cities with given system-wide population, in which the cities are linked by intercity migration, and the number of cities is endogenously determined. We call this “a system of replicable cities”. In fact, our setup is a synthesis of the single-monocentric-city land use model of urban economics with congestion, and the theory of local public good (LPG).[1] In this synthesis, the formation of a city (jurisdiction), either by a central planner or a profit-maximizing developer, is conditional on spending a fixed cost on basic infrastructure (LPG).

[1] The LPG theory is a version of a general economic approach to social groups formed for enjoying the benefit and sharing the cost of collective action. The theory of LPG (see, e.g., Tiebout (1956), Stiglitz (1977), Arnott (1979a), and Arnott and Stiglitz (1979)) is closely related to the theory of clubs (see, e.g., Buchanan (1964) Berglas (1976), Berglas and Pines (1981), and Scotchmer and Wooders (1987)). The main difference between them, however, is that the club theory is concerned with collective use of a single facility, whereas the LPG theory is concerned with multiple facilities as well as with production. The LPG theory is also closely related to the issue of optimal city size when the benefit of the collective action is not necessarily cost sharing but other source of scale economies (to local population size), like information spillover (see Fujita (1989) and Fujita and Thisse (2002)). In all the above three versions of the theory of collective action, the extended version of the Henry George rule and other theorems apply.

Each city's area consists of a core with exogenously fixed area size, in which all jobs and some residents are located, and suburbs with an area that is endogenously determined and in which only residents can be located. Congestion occurs as suburban residents travel to their jobs in the core and can be avoided by relocating to the core. A city attains a finite optimal size by balancing the marginal social costs of transportation against the marginal benefits of sharing the cost of the infrastructure. In this setup, there are two important margins through which economic and geographic sprawl are determined. On the one hand, in the intensive margin, various policies can be used to gain welfare improvements by allocating population from the suburbs to the core of each city which reduces congestion. On the other hand, in the extensive margin, which is important in the long run, congestion can also be alleviated by creating new cities, which spreads the population among more and less congested cities. *Our central question is whether improving economic efficiency by alleviating congestion results in less or more sprawl.*

*The policy regimes*

In particular, in this article, we investigate a variety of alternative economic regimes that differ from each other according to the policy tools available for financing the LPG and controlling the externality. On the one end of the spectrum, we discuss the allocation determined centrally by a benevolent and omnipotent planner that controls all resource allocation; on the other end of the spectrum, we discuss resource allocation under pure laissez faire (which can be achieved by relatively passive planners or by profit-maximizing developers); and, in between, we discuss mixed regimes where both a centralized planner and the markets play roles. The policy-tools menu conceived in the current paper includes: (i) a confiscatory tax on differential land rent (Henry George (single) tax); (ii) tolling of congestion; (iii) a head tax or subsidy; (iv) a unit tax or subsidy on suburban land rent (which is equivalent to an urban growth boundary (UGB)). In classifying these regimes, we assume, however, that the Henry George tax is available to local finance under any regime. Depending on the regime, it can be supplemented by one or more tools included in the above menu.

Our main results about the optimal policies can be summarized as follows:

- We show that when a central planner or the decentralized profit-maximizing developers use the same policy tools, then the resulting allocations are identical, a result that also holds in the literature on local public-goods, clubs, and optimal city size. A special case of this result was illustrated in part 1 of Proposition 4.3 of Fujita and Thisse (2002) which states that "*the equilibrium city system resulting from competition between profit-maximizing firms or developers is identical to the optimum system in which the common utility is maximized and conversely*." We show that our version of the local public good model yields the same result not only in the first-best (examined by Fujita and Thisse) but also in each of the lower-bests (when not all the tools in the menu are available). That is, given any policy tool set, the same allocation would be achieved in a long-run equilibrium by profit-maximizing competitive developers as by an omnipotent benevolent planner.

- Applying any set of policy tools, the aggregate imputed profits of planners (or the aggregate profits of developers) are zero after paying for the cost of the infrastructure needed to sustain the cities. This is a general version of the Henry George rule which applies to lower bests as well as to the first-best optimum. This differs from Anas and Rhee (2007) an Anas and Pines (2007) where the tax revenues were distributed to the population since there was no public good that needed financing.

- The congestion externality causes two resource allocation distortions when it remains un-priced. In the intensive margin, too many workers-residents live in the suburbs; while in the extensive margin, the number of cities is too few. Under general conditions, the first-best policy tool set for alleviating these distortions are a confiscatory tax on differential land rent (a Henry George tax), together with a toll on traffic congestion. When the congestion toll instrument is not available, the second-best policy is that the two remaining taxes, the head tax and suburban land tax, should together supplement the Henry George tax. If, in addition to tolling, one of these two taxes is also unavailable, the other alone should supplement the

Henry George tax and the corresponding regimes are third- and fourth-bests. Finally, the laissez-faire regime in which only the Henry George tax is available is a fifth best.

*Efficiency of urban sprawl*

With regard to the efficiency of urban sprawl, our results show the following:

- When the elasticity of substitution between the composite good and lot size is zero, the first best allocation is attainable by the Henry George tax supplemented by any of the other taxes (congestion toll, head tax, suburban land tax) or by any combination of the these taxes. The reason is that when the elasticity of substitution between the composite good and lot size is zero, consumers are insensitive to rents. Hence, only an inter-city effect exists. Furthermore, under the above restrictive qualification, lot size is equal everywhere and depends only on utility (income effect). It, then, follows that laissez-faire (where none of the taxes is available, and utility is therefore less than the first-best) generates a smaller geographical sprawl than the first-best which is achievable by supplementing the Henry George tax by any of the other taxes. In other words, subject to the above qualification and in contrast to the conventional belief, the optimal land use zoning or other optimal policy should expand the total urban land use.

- Congestion tolls which induce the first-best optimum can under a broad set of conditions when the elasticity of substitution between the composite good and lot size is not zero, reduce overall congestion and improve efficiency, while at the same time expanding aggregate urban land use. This pattern can occur under several circumstances;

  a) When the consumer's rent elasticity of the demand for lot size is sufficiently small, then the first-best optimal land use is more sprawled than the laissez-faire land use. This corresponds to the case of zero elasticity of substitution between lot size and the composite commodity, mentioned above.

b) When the land areas devoted to the cores of cities are small enough (recall that these are fixed in our model), then again the first-best optimal land use is more sprawled than is the laissez-faire land use.

c) When cities are highly congested in the laissez-faire regime.

Furthermore, in the transition from laissez-faire to the first-best optimum, *the rent per unit of land in the core of a city decreases as congestion tolls are levied* (Proposition 2), which, contrary to all the theoretical results in urban economics based on a single monocentric city, implies that congestion tolls would result in lower not higher urban densities in the cores of cities, because congestion tolls would, in the long run, depopulate city cores and spread core populations to the cores and suburbs of new urban areas.

The article is organized as follows. In section 2, we present in detail the setup of our model which, as mentioned above, combines features from urban economics and the theory of local public-goods. We show the feasibility constraints that must be satisfied by any policy that will allocate resources. In sections 3, each of the policy regimes is formulated as an optimization problem and solved. The analysis determines the efficiency ranking of the policy regimes discussed above under perfectly general assumptions about consumer preferences. Appendix A provides the basic analytical framework and shows how the same regimes can be decentralized if profit maximizing city-developers compete with each other. In section 4 we examine the special case of the zero elasticity of substitution between lot size and the composite commodity. Under this assumption we can unambiguously show that the four policy regimes achieve the same utility and number of cities. This utility is higher than laissez-faire as is also the number of cities. The four policies also achieve the same geographic sprawl which is higher than the laissez-faire geographic sprawl. In section 5, we investigate the transition from the laissez-faire to the first-best regime (i.e. the two ends of the policy spectrum), and we establish the conditions under which geographic sprawl (the aggregate land area of the city system) expands as congestion tolls approach their first-best values. In section 6 we present a framework for the numerical analysis of the model under short run and long run conditions. Section 7 provides some tentative conclusions.

## 2. The Setup

We treat a closed economy with $N$ (exogenous) urban consumer-workers who are identical in preferences as well as in their initial endowment and who are distributed among $m$ identical cities. Hence, each city accommodates $n = N/m$ resident-workers. Each worker is endowed with one unit of labor and labor is used as the only factor to produce a composite good. Each worker is employed in the city in which he resides. Each city consists of two areas: a core in which all jobs (production) are located and in which residents can also be located, and a suburb in which only residences can be located. It is convenient to think of the cities as circular with the core contained within a central circle and the suburb being an annulus, concentric with the core. In each city, $n_1$ consumer-workers reside in the core, while the remaining $n_2 = n - n_1$ reside in the suburb. Commuting within the core is costless, but suburban residents commuting into the core incur a transportation cost that exhibits congestion when they cross from the suburb into the core. This cost is then measured in units of the economy's composite good and is given by the function $t(n_2)$ with $t'(n_2) > 0$ and $t''(n_2) > 0$.

Establishing a core, in which transport is costless, requires a minimum investment in resources. In particular, we assume a technology of core production in which land and non-land inputs are perfectly complementary, so that a core costs $k + H_1 r$ units of the composite good to set up, where $H_1$ is the exogenous land area of the (circular) core required for a city to exist, and $k$ is the exogenous cost of some non-land infrastructure, also required for a city to exist. $r$ is the exogenous opportunity cost of land in non-urban uses such as agriculture (or, equivalently, the cost of converting each unit of freely available raw land to make it suitable for urban use). While the land area of the core is limited by assumption, the aggregate land developed for residential use in the suburbs will be endogenous and also costs $r$ per unit.

Suppose that one unit of labor is required to produce one unit of the composite good, then each city's output will be $n$ and the system of cities will produce $N$ units in the aggregate. The preferences of each consumer-worker is given by the strictly increasing and concave utility function $u(x_i, h_i)$ , where $x$ is the quantity of the consumer's consumption of the composite good and $h$ is the consumer's lot size and $i = 1$ designates residence in any core whereas $i = 2$ designates residence in any suburb.

We consider five alternative mixed economic regimes where the economic decisions are made in part by a benevolent planner and in part by utility-maximizing consumers and profit-maximizing entrepreneurs. In each of the five regimes, the planner establishes cities by developing a core which is essential for a functioning city. The rest of the economic decisions are made by individuals and firms. Developing the core costs, $k + rH$ , as explained above and is financed by a set of taxes available to the planner. The menu of taxes considered in this paper includes the following: a 100% tax on differential land rent (or a Henry George tax, denoted by HG), a congestion toll, $\tau$ , levied on the suburban residents crossing the bridge into the core, a head tax, $\Omega$ , levied on all residents and a suburban land tax, $s$ , per unit of land in the suburbs converted from agricultural to urban use. The Henry George tax is available to the planner in all five regimes. The regimes differ from one another, however, by whether the other taxes of the menu are available to the planner who chooses the available ones optimally in order to maximize the common welfare of the urban population:

- *First best* (*fb*): The entire tax menu $\{HG, \tau, \Omega, s\}$ is available to the planner;
- *Second best* (*sb*): The available tax menu includes the Henry George tax, *HG*, suburban land tax, $s$ , and the head tax, $\Omega$ , but congestion tolls are not available;
- *Lower best 1*(*lb1*): The available tax menu includes only the Henry George tax, HG, and the head tax, $\Omega$ ;
- *Lower best 2* (*lb2*) : The available tax menu includes only the Henry George tax, *HG*, and the land tax, $s$ , and this regime is equivalent to placing urban growth boundaries around each city;

- *Laissez faire* (*lf*): The single available instrument is the Henry George tax, *HG*.

The common features of the equilibriums defined by the five regimes are represented by the following five equations in $\left\{u, R_1, n, n_1, m \middle| \tau, \Omega, s\right\}$ :

$$n_1 x(R_1, u) + (n - n_1)\left(x(r+s, u) + h(r+s, u)r + t(n - n_1)\right) + H_1 r + k - n = 0 \quad (1)$$

$$n_1 h(R_1, u) - H_1 = 0, \quad (2)$$

$$E(R_1, u) + \Omega - 1 = 0, \quad (3)$$

$$E(r+s, u) + t(n - n_1) + \tau + \Omega - 1 = 0, \quad (4)$$

$$mn - N = 0, \quad (5)$$

where, $x(\bullet), h(\bullet), E(\bullet)$ are, respectively, the consumer's compensated demands for the composite good and housing, and the minimum expenditure function.

Equation (1) is the market clearing condition of the composite good (or resource) and shows that the city output $n$ is exhausted by the demand for direct consumption of the composite good, paying for commuting by the suburban residents, the cost agricultural land for the core and the suburb, and creating the infrastructure. (2) is the market clearing condition for the core's land. (3) is the budget constraint of a household living in the core. (4) is the budget constraint of a household living in the suburbs. (5) states that each urban resident is accommodated in one of the identical $m$ cities.

Given the set of available tools under his control, the planner maximizes $u$ subject to (1)-(5). More precisely, (1)-(5) are the constraints on the first best allocation; (1)-(5) and $\tau = 0$ are the constraints on the second best; (1)-(5), $\tau = 0$ , and $s = 0$ on the lower best 1; (1)-(5), $\tau = 0$ , and $\Omega = 0$ on the lower best 2; and (1)-(5), $\tau = 0$ , $s = 0$, and $\Omega = 0$ on the laissez fair.

Notice, however, that the solution procedure can be simplified by first maximizing $u$ subject to (1)-(4) and the additional specific constraints of the relevant regime to solve for $\left\{u, n_1, n, R_1 \middle| \tau, \Omega, s\right\}$.Then, this solution can be used to solve (5) for $m$ .

It can easily be shown that (1)-(4) imply that the planner's budget is balanced for *any* choice of the policy instruments that would be active in a particular regime:

$$n\Omega + H_1\left(R_1 - r\right) + \left(n - n_1\right)\left(h\left(R_2, u\right)s + \tau\right) - k = 0, \tag{6}$$

where the revenue comes from Henry George tax, congestion toll, head tax, and the suburban land tax.

# 3. The alternative regimes and their characterization

## 3.1 First Best

As specified earlier, in this case, the planner can use the Henry George tax supplemented by any tax of the set $\{\tau, s, \Omega\}$ and their quantitative levels in maximizing the common utility level $u$ subject to (1)-(4). The corresponding normalized Lagrangian of (1)-(4) is

$$\begin{aligned} \Im = {} & \frac{u}{\lambda} - n_1 x\left(R_1, u\right) - \left((n - n_1)\left(x\left(r + s, u\right) + h\left(r + s, u\right)r + t(n - n_1)\right) + H_1 r + k - n\right) \\ & -\rho_1\left(n_1 h(R_1, u) - H_1\right) - \theta_1\left(E(R_1, u) + \Omega - 1\right) - \theta_2\left(E(r + s, u) + t(n - n_1) + \tau + \Omega - 1\right) \end{aligned}$$

where $\rho_1$ is the shadow rent on land in the core. The first-order conditions for the first-best regime are, therefore:

$$\begin{aligned} \frac{\partial \Im}{\partial u} = {} & \frac{1}{\lambda} - n_1 \frac{\partial x(R_1, u)}{\partial u} - n_2\left(\frac{\partial x(r + s, u)}{\partial u} + \frac{\partial h(r + s, u)}{\partial u} r\right) \\ & -\rho n_1 \frac{\partial h(R_1, u)}{\partial u} - \theta_1 \frac{\partial E(R_1, u)}{\partial u} - \theta_2 \frac{\partial E(r + s, u)}{\partial u} = 0, \end{aligned} \tag{7}$$

$$\frac{\partial \Im}{\partial \tau} = -\theta_2 = 0, \tag{8}$$

$$\frac{\partial \Im}{\partial \Omega} = -\theta_1 - \theta_2 = 0, \tag{9}$$

$$\frac{\partial \Im}{\partial s} = -n_2\left(\frac{\partial x(r + s, u)}{\partial R_2} + r \frac{\partial h(r + s, u)}{\partial R_2}\right) - \theta_2 h(r + s, u) = 0, \tag{10}$$

$$\frac{\partial \Im}{\partial R_1} = -n_1\left(\frac{\partial x(R_1, u)}{\partial R_1} + \rho_1 \frac{\partial h(R_1, u)}{\partial R_1}\right) - \theta_1 h(R_1, u) = 0, \tag{11}$$

$$\frac{\partial \Im}{\partial n} = 1 - \left(x\left(r+s,u\right) + h\left(r+s,u\right)r + t(n_2)\right) - \left(n_2 + \theta_2\right)t'(n_2) = 0, \tag{12}$$

$$\frac{\partial \Im}{\partial n_1} = -\left(x(R_1,u) + \rho_1 h(R_1,u)\right) + \left(x\left(r+s,u\right) + h\left(r+s,u\right)r + t(n_2)\right)$$
$$+\left(n_2 + \theta_2\right)t'(n_2) = 0, \tag{13}$$

(1)-(4) and (7)-(13) are 11 equations in the unknowns $\{u, n_1, n, R_1, \tau, \Omega, s, \theta_1, \theta_2, \rho_1, \lambda\}$.

The properties of the first-best regime are: $(8)\ \mathit{and}\ (10) \Rightarrow R_2^{fb} = r \Rightarrow s^{fb} = 0;$

$(8),(9)\ \mathit{and}\ (11) \Rightarrow R_1^{fb} = \rho_1^{fb}$, $(8),(3),(12)\ \mathit{and}\ (13) \Rightarrow E(R_1^{fb},u^{fb}) = 1 \Rightarrow \Omega^{fb} = 0;$ $(8),(4)\ \mathit{and}\ (12)\ \mathit{with}\ s^{fb} = 0, \Omega^{fb} = 0 \Rightarrow \tau^{fb} = n_2 t'\left(n_2\right)$. In deriving the first and second, we use the derivative property of the expenditure function: $\dfrac{\partial E(R_i,u)}{\partial R_i} = \dfrac{\partial x_i}{\partial R_i} + R_i \dfrac{\partial h_i}{\partial R_i} = 0.$ Finally, it is straightforward to show that the same results follow from maximizing the utility, $u$, directly with respect to $\{x_1, x_2, h_1, h_2, n_1, n\}$ without imposing any policy instruments or prices and subject only to material balance constraints of the composite good and housing. Hence, the first-best solution described by the above results is indeed a way to obtain a Pareto efficient allocation.

### 3.2 Lower bests

The common feature of the four lower bests is that $\tau = 0$. Since equation (8) no longer applies, this implies that $\theta_2$ and, hence, $\theta_1$ from (9), need not vanish at a lower best optimum. (7) and (11)-(13), remain unchanged in the four lower-best regimes. In the case of second best, (7) and (9)-(13) are satisfied; in the case of lower best 1, ($s = 0$), only (7 ), (9), and (11)-(13) are satisfies; in the case of lower best 2, ($\Omega = 0$), (7 ) and (10) - (13) are satisfied; and in the laissez-faire case, only (7) and (11) – (13) are satisfied.

Before proceeding to explore in more detail each of the lower-best regimes, we state and prove the following:

**Lemma 1**: *In a any of the 5 regimes,* $\Omega\ = h(R_1\ ,u\ )\left(\rho_1 - R_1\right)$ *and, therefore,* $\rho_1 = R_1$ *in any regime in which* $\Omega = 0$.

**Proof:** It follows from (3), (12) and (13) which continue to hold in any of the lower best regimes.●

### 3.2.1 Second best $(\tau = 0)$

Since $\Omega$ is an available policy instrument, making use of (9), we can define

$$\theta \equiv -\theta_2 = \theta_1 . \qquad (14)$$

Then, adding $n_2\left(\partial x\left(r+s,u\right)/\partial R_2 + R_2 \partial h\left(r+s,u\right)/\partial R_2\right) = 0$ to the left side of (10) and $n_1\left(\partial x\left(R_1,u\right)/\partial R_1 + R_1 \partial h\left(R_1,u\right)/\partial R_1\right) = 0$ to the left side of (11) and using (14), we obtain:

$$n_2 s h_{R_2} + \theta h(R_2,u) = 0, \qquad (15)$$

$$n_1(\rho_1 - R_1)h_{R_1} + \theta h(R_1,u) = 0, \qquad (16)$$

where, for abbreviation sake, we use $h_{R_i}$ for $\partial h\left(R_i,u\right)/\partial R_i$, hereinafter. Now we subtract and add $\sum_{i=1}^{2} R_i h(R_i,u)$ to the left side of (13) and obtain

$$-(\rho_1 - R_1)h(R_1,u) - sh(R_2,u) + \left((n-n_1)-\theta\right)t'(n-n_1) = 0 . \qquad (17)$$

Solving (15) for $s$ and (16) for $\rho_1 - R_1$ and substituting the results into (17), we obtain:

$$\theta^{sb} = \frac{t'}{t' + \left(\dfrac{h_1^2}{n_1\left(-h_{R_1}\right)} + \dfrac{h_2^2}{n_2\left(-h_{R_2}\right)}\right)} n_2 \ , \qquad (18)$$

where the abbreviated notation is used and the functions are evaluated at the second best.

Since the compensated demands are declining functions of their own prices, the denominator of (18) is positive and larger than the numerator. Hence, $0 < \theta^{sb} < n_2 = (n - n_1)$. It then follows from (15) that $s^{sb}$ >0. Next, subtracting and adding $\left(R_2 - r\right)h_2$ to the left side of (12), substituting (15) into the result, and using (18):

$$\begin{aligned}\Omega &\equiv 1 - E(R_1,u) = (n_2 - \theta^{sb})t' - \left(R_2 - r\right)h_2 \\ &= \left(\frac{h_1^2}{n_1\left(-h_{R_1}\right)} + \frac{h_2^2}{n_2\left(-h_{R_2}\right)}\right)\theta^{sb} - \frac{h_2^2}{n_2\left(-h_{R_2}\right)}\theta^{sb} = \frac{h_1^2}{n_1\left(-h_{R_1}\right)}\theta^{sb} > 0 \end{aligned} \qquad (19)$$

Finally, with $\tau$ =0, (6) implies

$$n\Omega + H_1\left(R_1 - r\right) + n_2 h_2 s - k = 0 . \qquad (20)$$

The above results deserve an intuitive interpretation with references to the public economics literature. To that end, we further investigate (15) and (16) in the light of the existing tax literature. In order to analyze the effect of the policy instruments, we denote an equilibrium, conditional on a given policy instrument set, $\{\Omega, s\}$, as $\{n^e, n_1^e, R_1^e, u^e\} \equiv$ $\{n(\Omega,s), n_1(\Omega,s), R_1(\Omega,s), u(\Omega,s)\}$ that, together with $\{\Omega, s\}$ solves (1)-(4). Accordingly, (3) and (4) become:

$$E\left(R_1(s,\Omega), u(s,\Omega)\right) + \Omega - 1 = 0. \tag{21}$$

$$E\left(r+s, u(s,\Omega)\right) + t\left(n_2(s,\Omega)\right) + \Omega - 1 = 0 \tag{22}$$

Differentiating (21) and (22) totally, we obtain (again using abbreviated notation):

$$h_1\left(\frac{\partial R_1}{\partial s}ds + \frac{\partial R_1}{\partial \Omega}d\Omega\right) + \frac{\partial E_1}{\partial u}\left(\frac{\partial u}{\partial s}ds + \frac{\partial u}{\partial \Omega}d\Omega\right) + d\Omega = 0, \tag{23}$$

$$h_2 ds + \frac{\partial E_2}{\partial u}\left(\frac{\partial u}{\partial s}ds + \frac{\partial u}{\partial \Omega}d\Omega\right) + t'\left(\frac{\partial n_2}{\partial s}ds + \frac{\partial n_2}{\partial \Omega}d\Omega\right) + d\Omega = 0. \tag{24}$$

At the second-best optimum, $\frac{\partial u}{\partial s} = \frac{\partial u}{\partial \Omega} = 0$ by definition, and these facts should be imposed on (23), (24). Also, setting $d\Omega = 0$, (24) reduces to

$$h_2 = -t'\frac{\partial n_2}{\partial s}. \tag{25}$$

Substituting (25) into (15) yields

$$n_2 s h_{R_2} = \theta t' \frac{\partial n_2}{\partial s}. \tag{26}$$

(26) expresses the equality of marginal social cost to the marginal benefit of marginally increasing $s$ at the optimum. In order to see this, observe that the left side is the marginal aggregated social excess burden (dead weight loss). This is illustrated in Figure 1, where the area of the triangle *abc* is the excess burden (dead weight loss) resulting from increasing the suburban land rent from $r$ to $r+s$; $\overline{ac}$ is the marginal excess burden (dead weight loss) in terms of the composite good's units; and $\overline{ab}$ is the excess burden (dead weight loss) in terms of land units. This can be verified by observing that $\overline{ab}/\overline{ac} = \overline{ab}/s = \cot\left(\sphericalangle abc\right) = h_{R_2} \Rightarrow \overline{ab} = sh_{R_2}$. Multiplying $\overline{ab}$ by $n_2$ the left side of (26)

is the aggregate social excess burden (dead weight loss) resulting from the increase of the suburban land rent by the init tax $s$.

The right side of (26) is the welfare gain resulting from the decline in the suburban population induced by $s$, and the resulting decline in congestion cost, once again, in terms of land area. More precisely, $t'\partial n_2 / \partial s$ is the decline in the transportation cost (in terms of land units) resulting from imposing the land tax $s$. Because $s$ is a second best instrument, $\theta$ is, then, the fraction of the potential benefit that would occur from the decline in transportation costs in the case of the first best, recalling that $\theta = 0$ in the first-best but $0 < \theta < n_2$ in the second-best.

Returning to (15), we can multiply and divide the first term by $R_2$ and obtain

$$\frac{s}{R_2} = -\theta \frac{h_2}{n_2 h_{R_2} R_2} = -\theta \frac{1}{n_2 \eta_{h_2:R_2}}, \qquad (27)$$

where $\eta_{h_2:R_2}$ is the own-price compensated demand elasticity for lot size in the suburb. (27) is the standard Ramsey pricing rule when there is no cross elasticity between lot size in the suburb and the core and when $n_2$ identical suburban individuals consume the taxed good, $h_2$. The absence of cross elasticity, as in Ramsey (1927), although in our case the prices are interrelated by migration, is puzzling. It can, however, be explained by the assumed non-convexity of the consumption set: the price of lot size in the core does not directly affect the demand for lot size in the suburbs.[2] (27) expresses the result that the more inelastic the compensated demand for suburban land, the higher the tax rate on land, and the larger the suburban population, the lower the tax rate on land.

Turning now to the head tax, note first that even though the head tax applies equally to the core and the suburbs it does affect the allocation because in our model, there are income effects which are different in city and suburb. Also, the head tax paid by the suburban residents would play the role of the congestion toll in the first-best. If only the suburban residents were levied a head-tax, then the optimal value of such a tax would be $n_2^{fb} t'(n_2^{fb}) = \tau^{fb}$ and since this achieves the first-best optimum, no other tax would be

[2] Observe that if we include several goods, say $x_i$, $i = 1, 2, ..., n$ in the utility function, then the cross elasticities could not be overlooked as we are able to do in the present case.

needed. However, in our case, the head tax is also paid by the core residents and this introduces a distortion away from the first-best allocation which is partially offset by the land tax in the suburbs as we saw above. To show the marginal effects of the head-tax, we analyze its marginal cost and benefit (when $s$ is kept constant). To that end we recall that, at the (second-best) optimum, $\partial u / \partial s = \partial u / \partial \Omega = 0$ and we now set $ds = 0$ in (23) and (24) to obtain:

$$\frac{\partial R_1}{\partial \Omega} = -\frac{1}{h_1}, \tag{28}$$

$$\frac{\partial n_2}{\partial \Omega} = -\frac{1}{t'}. \tag{29}$$

Substituting (28) and (29) into (16) and using Lemma 1, we obtain:

$$\frac{n_1 \Omega h_{R_1} \partial R_1 / \partial \Omega}{h_1} = \theta = -\theta t' \frac{\partial n_2}{\partial \Omega}. \tag{30}$$

The left side represents the marginal cost which is the deadweight loss associated with the deviation of the core's land price, $R_1$, at the optimum from its opportunity cost, $\rho_1$, resulting from the marginal increase in $\Omega$; the right side represents the benefit which follows from the decrease in $n_2$, induced by the head tax $\Omega$. 16) can also be rewritten as

$$\frac{\rho_1 - R_1}{R_1} = \theta \frac{h_1}{n_1 R_1 h_{R_1}} = \theta \frac{1}{n_1 \eta_{h_1:R_1}}, \tag{31}$$

which, together with (27) gives,

$$\frac{s}{R_2} / \frac{\rho_1 - R_1}{R_1} = \frac{n_1 \eta_{h_1:R_1}}{n_2 \eta_{h_2:R_2}}. \tag{32}$$

(32) is again the Ramsey rule, this time with two different groups each composed of ex-ante identical individuals where the left side is the relative deviation of the second-best pricing from the market opportunity costs.

Observe that (32) modifies the original Ramsey rule in two peculiar ways. In contrast to Ramsey, in our case, there is more than one individual. More precisely, there are many individuals who are ex ante identical but, ex post are distributed by self-selection into two distinct groups: $n_1$ individuals who live in the core and consume only the core's lots (the suburban lot size is missing not in their utility) and $n_2$ individuals who live in the suburb and consume only a suburban lot (the core's lot size is not in their utility). Therefore, as in the original version of Ramsey, the cross elasticities of the demand for locating in the core with respect to the suburban land rent (or in the suburb with respect to the core rent) are irrelevant and do not appear in (32). Second, the deviation of the market price in the core is from the shadow rent on land, not the fixed producer's price as it would be in the original Ramsey rule. The reason for this is that the deviation must be from socially optimal prices. Since in our case an externality is present, the socially optimal and the market price of land in the core and in the suburb diverge.

### 3.2.2 Head tax ($\tau = s = 0$)

Solving (16) and (17) for $\theta$, where $s = 0$, yields

$$\theta^{ht} = \frac{-t' n_1 h_{R_1}}{h_1^2 - t' n_1 h_{R_1}} n_2 , \qquad (33)$$

which, as in the case of second best, satisfies $0 < \theta^{ht} < n_2$. It, then, follows from Lemma 1, (17), $s = 0$, and (33) that

$$\Omega^{ht} = \left(\rho_1 - R_1\right) h_2 = \frac{h_1^2}{n_1 h_{R_1}} \theta^{ht} = \frac{h_1^2}{h_1^2 - n_1 h_{R_1}} n_2 t' > 0 , \qquad (34)$$

where the variables are evaluated at the lower-best 1 or *ht*.

### 3.2.3 Suburban land tax–Urban Growth Boundary ($\tau = \Omega = 0$)

That the suburban land tax is perfectly equivalent to a UGB was proved in Anas and Pines (2007).[3] The planner's problem is that of the second-best with the additional constraint that $\Omega = 0$. Using Lemma 1, (15), and (17), the solution for $\theta^{ugb}$ is:

$$\theta^{ugb} = \frac{-n_2 h_{R_2} t'}{h_2^2 - n_2 h_{R_2} t'} n_2 , \qquad (35)$$

implying, as in the second-best and head tax, that $0 < \theta^{ugb} < n_2$. We now substitute (35) into (15) and get,

[3] The clear intuition is as follows: A restrictive UGB rations suburban land away from the urban use and into farming. Thus it reduces the supply of suburban land raising the suburban land rent. Since the suburban land tax is a wedge between the farming rent and the UGB-induced suburban rent, a planner could achieve the same result by directly instituting such a tax instead of instituting a UGB.

$$s^{ugb} = R_2 - r = \frac{h_2}{h_2^2 - n_2 h_{R_2} t'} n_2 t' > 0, \qquad (36)$$

where all the expressions are evaluated at the optimum UGB. This result then says that in the second-best regime a restrictive UGB must be used in each suburb. It contrasts with Anas and Rhee (2007) where, in the single monocentric city with job mobility, the UGB could be restrictive or expansive depending on whether the effects of tolls (in the first-best) were centralizing or decentralizing on the location of jobs. It also contrasts with Anas and Pines (2007), where with the mobility of the population between the two cities, the UGB in the larger monocentric city had to be restrictive while that in the smaller one had to be expansive.

### 3.2.4. Fifth-best allocation: laissez-faire $(\tau = 0,\ \Omega = 0,\ s = 0)$

In this case, the planner is most passive. He uses only the Henry George tax to finance the core's formation that costs $H_1 r + k$. Accordingly, the planner maximizes $u$ subject to (1)-(4) where the unknowns are $\{u, n, n_1, R_1\}$. Since the number of unknowns equals the number of constraints, the optimization problem is degenerate but there is still a benefit to formulating it as an optimization problem, if one needs to derive shadow prices (see Arnott (1979b)). In this case, the Henry George rule collapses to

$$H_1(R_1 - r) - k = 0. \qquad (37)$$

Also, from (10), since $R_2 = r$ under laissez-faire, the derivative property of the expenditure function implies that $\theta_2^{lf} = -n_2$.

## 3.3 Comparison of the regimes

### 3.3.1 General

Using the results of sections 3.2.1-3.2.4, we can state the following proposition:

**Proposition 1:**

**(i)** *With the exception of the first best regime, the available policy instruments are chosen at a positive level; in the case of the first-best regime, however, where tolling is available,* $\Omega$ *and* $s$ *that are also available, but are chosen to be zero (i.e. not used by the planner).*

**(ii)** *In a less than first-best regime j,* $\Omega^j + s^j h(r + s^j, u^j) < n_2^j t'(n_2^j),\ \ j = sb, ht, ugb, lf$ , *the total tax paid by each suburban resident falls short of the externality caused by that resident.*

**(iii)** *The regimes can be ranked so that* $u^{fb} > u^{sb} > \max\left(u^{ht}, u^{ugb}\right) \geq \min\left(u^{ht}, u^{ugb}\right) > u^{lf}$.

**Proof**: **(i)** Proved in sections 3.2, 3.2.1-3.2.4. **(ii)** For $j = ht, ugb, lf$ the claim follows directly from 3.2.2, 3.2.3 and, trivially, from 3.2.4. We only have to prove it for *sb*. We use Lemma 1,(17), and $n_2^{sb} > \theta^{sb}$ which imply $\Omega^{sb} + s^{sb}h(r + s^{sb}, u^{sb}) = -\left(n_2^{sb} - \theta^{sb}\right)t' < 0$.

**(iii)** The ranking of the regimes is proved as follows. Let $u^{fb}$ be the optimal utility achieved when all tax instruments are simultaneously available. This optimum is found by maximizing $u$ subject to the constraints (1)-(5) with respect to $\{u, R_1, n, n_1, \tau, \Omega, s, m\}$. The second-best problem is obtained by adding the constraint $\tau = 0$ to the constraint set. Since, however, in the first-best regime $\tau > 0$, as we showed in 3.2.1, the new constraint $\tau = 0$ is binding and thus $u^{sb} < u^{fb}$. The regime *ht* is obtained by adding the constraint $s = 0$ to the second-best regime. Since, as we showed in 3.2.1, in the second-best regime, $s^{sb} > 0$, the new constraint is binding and thus $u^{ht} < u^{sb}$. Similarly, the regime *ugb* is obtained by adding the constraint $\Omega = 0$ to the second-best regime. Since we showed in 3.2.1 that in the second-best regime's optimum $\Omega > 0$, the new constraint is binding and thus $u^{ugb} < u^{sb}$. Therefore, $u^{sb} > \max(u^{ht}, u^{ugb})$. Finally, the laissez-faire regime is obtained either by adding the constraint $\Omega = 0$ to the *ht* regime or $s = 0$ to the *ugb* regime. Either way, the new constraints are binding, because, $\Omega > 0$ in the *ht* optimum as we showed in 3.2.2**,** and $s$ > 0 in the *ugb* regime as we showed in 3.2.3. Hence, $\min(u^{ht}, u^{ugb}) > u^{lf}$. ●

Notice that Lemma 1 and Proposition 1 imply that $\rho_1^{fb} - R_1^{fb} = \rho_1^{ugb} - R_1^{ugb} = \rho_1^{lf} - R^{lf} = 0$ whereas $\rho^{sb} - R^{sb}, \rho^{ht} - R^{ht} > 0$.

### 3.3.2 Laissez-faire versus first-best

Several qualitative properties that emerge from the comparison of the laissez-faire and the first best-regimes are easy to establish:

**Proposition 2:** *In the laissez-faire regime, compared to the first-best regime, the following hold*: **(a)** *Rents are higher in the cores,* $R_1^{lf} > R_1^{fb}$; **(b)** *Lot sizes are smaller in* *the cores and the suburbs and less composite commodity is consumed in the suburbs,* $h(R_1^{lf},u^{lf}) < h(R_1^{fb},u^{fb}),\ h(r,u^{lf}) < h(r,u^{fb}),\ x(r,u^{lf}) < x(r,u^{fb})$. **(c)** *There are more residents in each core,* $n_1^{lf} > n_1^{fb}$; **(d)** *There are more residents in each suburb,* $n_2^{lf} > n_2^{fb}$; **(e)** *Each city is larger and there are fewer cities,* $n^{lf} > n^{fb}$, $m^{lf} < m^{fb}$.

**Proof:** **(a)** The budget constraints of the residents are $E(R_1^{lf},u^{lf}) = 1,\ E(R_1^{fb},u^{fb}) = 1$. By Proposition 2, $u^{lf} < u^{fb}$. Since expenditure rises with both utility and rent, $R_1^{lf} > R_1^{fb}$ is necessary for $E(R_1^{lf},u^{lf}) = E(R_1^{fb},u^{fb}) = 1$. **(b)** From Proposition 1 it follows that $h(r,u^{lf}) < h(r,u^{fb})$ and $x(r,u^{lf}) < x(r,u^{fb})$, because both goods are assumed normal and, therefore, demanded quantity rises with utility, keeping rent constant. Since lot size is a declining in rent, it follows that $h(R_1^{lf},u^{lf}) < h(R_1^{fb},u^{fb})$ both from the utility and price effects. **(c)** From (b), and the land market constraint (2): $n_1^{lf} = \dfrac{H_1}{h(R_1^{lf},u^{lf})} > \dfrac{H_1}{h(R_1^{fb},u^{fb})} = n_1^{fb}$. **(d)** The suburban budget constraints can be rewritten as follows: $t(n_2^{lf}) = 1 - E(r,u^{lf}),\quad t(n_2^{fb}) + n_2^{fb}t'(n_2^{fb}) = 1 - E(r,u^{fb})$. By $u^{lf} < u^{fb}$, $E(R_1,u^{lf}) < E(R_1,u^{fb})$. Therefore, $t(n_2^{lf}) > t(n_2^{fb}) + n_2^{fb}t'(n_2^{fb})$. Since $n_2 t'(n_2)$ increases with $n_2$, the indicated inequality requires $n_2^{lf} > n_2^{fb}$. **(e)** From (c) and (d),

$n^{lf} = n_1^{lf} + n_2^{lf} > n^{fb} = n_1^{fb} + n_2^{fb}$. Then, $m^{lf} = \dfrac{N}{n^{lf}} < m^{fb} = \dfrac{N}{n^{fb}}$. ●

## 3.4 Decentralization with profit maximizing developers

We now examine how the optima described above can be achieved not by a planner operating in a mixed economy but by profit maximizing, utility-taking, developers who are free to establish or abandon cities. Such a developer must incur the minimum costs for establishing a city. As we know this requires buying the land area $H_1$ at rent $r$, and spending $k$ to create the core's infrastructure. Beyond this minimum investment, the developer is free to use the same taxes and subsidies (or any subset of these) that our planner used. More precisely, given any planning regime, we define a corresponding complete decentralized regime with developers, each using the same instruments as the planner to maximize profits in a single city. For example, corresponding to the second best planning regime, we examine decentralization when the developers are unable to impose congestion tolls but can impose head tax and suburban land tax. We can show (the proof is available upon request) that each of the five regimes can be implemented by its corresponding type of decentralization (see Appendix A for the case of the first-best regime). The reason is straightforward because the net profit a developer realizes is the surplus of the city, that is, the aggregate supply of the composite good, $n$, minus the aggregate consumption of the composite good minus the cost of the aggregate land (city area times $r$) minus aggregate transport costs minus $k$. Given the maximum utility the planner can achieve, the maximum surplus of the city is zero. For, otherwise, the utility could not have been maximized. But, then, given the maximum utility, the maximized profits are zero. Furthermore, it must also be true that if the utility is lower than what the planner can achieve, the developer can earn positive profits; and if the utility is higher than what the planner can achieve, the developer must incur a loss. In the first case, the profits will attract new developers forming new cities who will raise the real wage to attract labor and thus the utility will increase until the profits from city-development vanish; in the second case, developers will exit from the development business and the competition of workers on the fewer jobs will reduce the real wage and, consequently, the utility will decline until the developers can break even.

# 4. Special case: zero elasticity of substitution

When the elasticity of substitution between the composite good and lot size is exactly zero, namely $u^j = \min\left\{\frac{x_i^j}{a}, \frac{h_i^j}{b}\right\}$ with $a, b > 0$, then we can fully rank the five regimes in terms of $u, GS$ and $ES$. Note that the individual demands for the composite commodity and the lot size are the same regardless of location in the core or suburb. Hence, $x_1^j = x_2^j = x^j = au^j$ and $h_1^j = h_2^j = h^j = bu^j$. Then, $GS^j = Nbu^j$. It follows directly that the geographic sprawl is positively correlated with efficiency (level of utility), which is contrary to conventional wisdom as explained in the introduction.

**Proposition 3:** *When the utility function* $u(x,h)$ *exhibits zero elasticity of substitution, then*:

**(i)** $u^{fb} = u^{sb} = u^{ht} = u^{ugb} > u^{lf}$; **(ii)** $GS^{fb} = GS^{sb} = GS^{ht} = GS^{ugb} > GS^{lf}$;

**(iii)** $m^{fb} = m^{sb} = m^{ht} = m^{ugb} > m^{lf}$;**(iv)** $ES^{fb} = ES^{sb} = ES^{ht} = ES^{ugb} < ES^{lf}$.

**Proof: (a)** Since $x_1^j = x_2^j$ and $h_1^j = h_2^j$ do not depend on $R_1, \tau, s$ and,$\Omega$, the market clearing conditions (1) and (2) are the same for all the regimes, that is $n^j(a+br)u^j + \left(n^j - n_1^j\right)t\left(n^j - n_1^j\right) + k - n = 0,$ and $n_1^j bu^j - H_1 = 0$, respectively, for $j = fb, sb, lb1$ and $j = lb2$. **(b)** Let $\Omega^{sb}$ and $s^{sb}$ be chosen such that $s^{sb}$=0 and $\Omega^{sb} = n_2^{fb}t'\left(n_2^{fb}\right)$. Then, (4) of the second best regime becomes; $0 = au^{sb} + bu^{sb}r + t\left(n_2^{sb}\right) + \Omega^{sb} - 1 = \left(a+br\right)u^{sb} + t\left(n_2^{sb}\right) + n_2^{sb}t'\left(n_2^{sb}\right) - 1$. We thus have, by **(a)** and **(b),** three equations in $\left\{u, n_1, n_2\right\}$ for the second-best regime that are identical to the corresponding equations of the first-best regime. It, then, follows, that $\left\{u^{sb}, n_1^{sb}, n_2^{sb}\right\} = \left\{u^{fb}, n_1^{fb}, n_2^{fb}\right\}$. **(c)** A similar choice of$\Omega^{ht}$, that is, $\Omega^{ht} = n_2^{fb}t'\left(n_2^{fb}\right)$ leads to $\left\{u^{ht}, n_1^{ht}, n_2^{ht}\right\} = \left\{u^{fb}, n_1^{fb}, n_2^{fb}\right\}$. **(d)** Under $ugb$, we choose $s^{ugb} = n_2^{fb}t'\left(n_2^{fb}\right)/bu^{fb}$ such that (4) is : $0 = \left(a + b\left(r + s^{ugb}\right)\right)u^{ugb} + t\left(n_2^{ugb}\right) - 1 = \left(a+br\right)u^{ugb} + t\left(n_2^{ugb}\right) + s^{ugb}u^{ugb} - 1$ $= \left(a+br\right)u^{ugb} + t\left(n_2^{ugb}\right) + n_2^{fb}t'\left(n_2^{fb}\right)u^{ugb}/bu^{fb} - 1$. It, then, follows that $\left\{u^{ugb}, n_1^{ugb}, n_2^{ugb}\right\} = \left\{u^{fb}, n_1^{fb}, n_2^{fb}\right\}$ solves equations (1), (2), and (4) for *ugb*. Since in the case of zero elasticity of substitution (ZES, hereinafter), $GS^j \overset{ZES}{=} Nbu^j$**,** **(ii)** is then immediately clear. To prove **(iii)** we note that $n^j = \dfrac{H_1}{bu^j} + n_2^j$. By part **(d)** of Proposition 2, we know that $n_2^{lf} > n_2^{fb}$ and we known that $u^{lf} < u^{fb}$. Therefore, $n^{fb} > n^{lf}$. From $N - m^j n^j = 0,$ therefore, $m^{fb} = m^{sb} = m^{ht} = m^{ugb} > m^{lf}$. To prove **(iv)** we use the output's market clearing aggregated over cities: $N\left(a+br\right)u^j + m^j k + ES^j - N = 0.$ This, (i) and (iii) yield the result. ●

Recall that although regimes $j = fb, sb, ht, ugb$ are identical in $\left\{u, n_1, n_2,\right\}$, they differ in$R_1^j$ ,found from $E(R_1^{fb}, u) - 1 = 0,$ $E(R_1^{ht}, u) - (1 - \Omega^{ht}) = 0,$ $E(R_1^{ugb}, u) - 1 = 0.$

Since $\Omega^{ht} > 0$, it is immediately clear that $R_1^{fb} = R_1^{ugb} > R_1^{ht}$. In the head tax regime, core residents pay the same tax as suburban residents, whereas in the first-best and UGB regimes, suburban residents pay the same equivalent tax but core residents pay no tax. Because there are no substitution effects, paying the tax regardless of location under the head tax regime, induces a pure shift of the population to the suburbs as compared to the first-best or UGB, which causes rents in the core to fall to restore equilibrium between core and suburb. It also follows that the second-best policy can be achieved by using one of the following two taxes: either uniform head tax, $\Omega^{sb} = n_2 t'(n_2)$ and $s = 0$, or suburban land tax $s^{sb} = n_2 t'(n_2)/h_2$, and $\Omega^{sb} = 0$. Under the former, $R_1^{sb} = R_1^{ht}$. Under the latter, $R_1^{sb} = R_1^{ugb}$.

## 5. Transition: laissez-faire to first-best

In order to examine how the first-best characteristics differ from the laissez-faire under more general conditions, without assuming a zero elasticity of substitution, we follow a different strategy. A formal comparative statics analysis of the following equations allows us to derive precise quantitative expressions for the derivatives of the endogenous variables with respect to the congestion toll $\tau \in \left[0, n_2^{fb} t'(n_2^{fb})\right]$, i.e. as the toll varies from its laissez-faire value of zero towards its first-best value. The equations that are differentiated are (19)-(21) and (23). After setting $\Omega = s = 0$, they become:

$$\begin{aligned} &n_1 h(R_1, u) - H_1 = 0, \\ &E(r, u) + t(n_2) + \tau - 1 = 0, \\ &E(R_1, u) - 1 = 0, \\ &R_1 H_1 - r H_1 + n_2 \tau - k = 0. \end{aligned}$$

The linear algebra takes the form:

$$\begin{bmatrix} h_1 & 0 & n_1\left(\partial h_1 / \partial R_1\right) & n_1\left(\partial h_1 / \partial u\right) \\ 0 & t'(n_2) & 0 & E_{2u} \\ 0 & 0 & h_1 & E_{1u} \\ 0 & \tau & H_1 & 0 \end{bmatrix} \begin{bmatrix} dn_1 / d\tau \\ dn_2 / d\tau \\ dR_1 / d\tau \\ du / d\tau \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \\ 0 \\ -n_2 \end{bmatrix}$$

with $Det. = -h_1\left(h_1 \tau E_{2u} + H_1 t'(n_2) E_{1u}\right) < 0.$ The results are:

$$\frac{dn_1}{d\tau} = \left(\tau - n_2 t'(n_2)\right) n_1 \left( \underbrace{h_{R1}E_{1u} - h_1 h_{1u}}_{<0} \right) / \, Det. \leq \; 0 \Leftrightarrow \tau \leq n_2^{fb} t'(n_2^{fb}), \; n_2^{fb} t'(n_2^{fb}) \quad (38a)$$

$$\frac{dn_2}{d\tau} = h_1 \left( n_2 h_1 E_{2u} + H_1 E_{1u} \right) / \, Det. < 0 \quad (38b)$$

$$\frac{dR_1}{d\tau} = \left( n_2 t'(n_2) - \tau \right) h_1 E_{1u} \, / \, Det. \leq \; 0 \Leftrightarrow \tau \leq n_2^{fb} t'(n_2^{fb}), \quad (38c)$$

$$\frac{du}{d\tau} = \left(\tau - n_2 t'(n_2)\right) h_1^2 \, / \, Det. \geq \; 0 \Leftrightarrow \tau \leq n_2^{fb} t'(n_2^{fb}). \quad (38d)$$

With the help of Proposition 2 and the above expressions, we will now investigate how economic and geographic sprawl change as the congestion toll, $\tau$ , increases from zero toward its first-best optimal value $\tau^* = n_2^{fb} t'(n_2^{fb})$. Economic and geographic sprawl are $ES = mn_2 t(n_2),$ and $GS = m\left(H_1 + n_2 h(R_2, u)\right)$ where $R_2 = r$ both under laissez-faire and first-best. We will also define number of cities, $m = N(n_1 + n_2)^{-1}$, and aggregate suburban population. $SP \equiv mn_2 = \dfrac{N}{n_1 + n_2} n_2$, which is itself of interest and could, according to a possible definition, be used as an alternative measure of sprawl. Then, $ES = SP \times t(n_2), \;\; GS = mH_1 + SP \times h(r,u)$. From these, expressions we can directly state and prove the following lemma.

**Lemma 2: (a)** *If the congestion toll causes the aggregate suburban population* $(SP)$ *to decrease, then aggregate economic sprawl* $(ES)$ *decreases;* **(b)** *For aggregate geographic sprawl to decrease it is necessary but not sufficient that aggregate suburban population decrease;* **(c)** *If the congestion toll causes the aggregate suburban population not to decrease or causes it to increase, then aggregate geographic sprawl increases.*

**Proof: (a)** $\dfrac{d(SP)}{d\tau} < 0 \Rightarrow \dfrac{d(ES)}{d\tau} = SP \times t'(n_2) \dfrac{dn_2}{d\tau} + t(n_2) \dfrac{d(SP)}{d\tau} < 0$ **,**which follows because by (32b) we know that $\dfrac{dn_2}{d\tau} < 0$; **(b)** From $GS = mH_1 + SP \times h(r,u)$**,** we can see that by the sum of the expressions (32a), (32b), $\dfrac{dm}{d\tau} > 0$; and since by (32d), $\dfrac{du}{d\tau} > 0$, $h_2 = h(r,u)$ increases with $\tau$ . Hence, for $\dfrac{d(GS)}{d\tau} < 0,$ to be possible, suburban population must decrease, namely $\frac{d(SP)}{d\tau} < 0$ is necessary; **(c)** All three effects, $\frac{dm}{d\tau} > 0$, $\frac{d(SP)}{d\tau} \geq 0$, $\frac{du}{d\tau} > 0 \Rightarrow \frac{dh_2}{d\tau} > 0$ and are aligned, working together to increase $GS$ .●

The marginal effect of $\tau$ on the number of cities is decomposed by:

$$\frac{dm}{d\tau} = \underbrace{\frac{1}{n}m\left(-\frac{dn_1}{d\tau}\right)}_{\substack{\textit{New cities formed}\\ \textit{to accommodate those}\\ \textit{leaving the cores of}\\ \textit{existing cities.}}} + \underbrace{\frac{1}{n}m\left(-\frac{dn_2}{d\tau}\right)}_{\substack{\textit{New cities formed}\\ \textit{to accommodate those}\\ \textit{leaving the cores of}\\ \textit{existing cities.}}} > 0 \Rightarrow \frac{-dn_1 / d\tau}{-dn_2 / d\tau} > -1 \qquad (39)$$

Since the left side of this inequality is positive, it holds regardless of any assumption or other values. Thus the number of cities increases continuously as the toll is raised continuously from its laissez-faire value of zero to its first-best optimal value.

The following expression decomposes the marginal effect of $\tau$ on the aggregate suburban population:

$$\frac{d(SP)}{d\tau} = \underbrace{\frac{n_2}{n_1+n_2}m\left(-\frac{dn_1}{d\tau}\right)}_{\substack{\textit{Population that leaves}\\ \textit{an existing city's core for a new}\\ \textit{suburb.}}} - \underbrace{\frac{n_1}{n_1+n_2}m\left(-\frac{dn_2}{d\tau}\right)}_{\substack{\textit{Population that leaves}\\ \textit{an existing suburb for a}\\ \textit{new city's core.}}}. \qquad (40)$$

Recall from the comparative statics outcome, (38b), that imposing the congestion toll on suburban residents causes residents to relocate out of existing suburbs. From (38c), the toll also causes the rent in the city core to fall and thus core densities to fall implying that city core residents also relocate out of the existing cores. These two effects cause total population in existing cities to fall. To accommodate these movers with total population declining in each city, new cities are created. However, in this process, total suburban population will decrease only if those relocating out of existing suburbs and into the cores of existing cities are more numerous than those relocating out of existing city cores and into the suburbs of new cities. Hence, (40) implies,

$$\frac{d(SP)}{d\tau} < 0 \;\Rightarrow\; \frac{-dn_1/d\tau}{-dn_2/d\tau} < \frac{n_1}{n_2}. \qquad (41)$$

Using the expressions (38a) and (38b):

$$\frac{\left(\tau - n_2 t'(n_2)\right) n_1 \left( \underbrace{h_{R1} E_{1u} - h_1 h_{1u}}_{<0} \right)}{h_1 \left( n_2 h_1 E_{2u} + H_1 E_{1u} \right)} < \frac{n_1}{n_2} \tag{42a}$$

$$\frac{\left(n_2 t'(n_2) - \tau\right) n_2 \left( \underbrace{h_1 h_{1u} - h_{R1} E_{1u}}_{>0} \right)}{h_1^2 \left( n_2 E_{2u} + n_1 E_{1u} \right)} < 1 \tag{42b}$$

Suppose that the expenditure function is linear in $u$ (indirect utility linear in income). Then, $E_{1u} = E_1 / u = 1/u$, $E_{2u} = E_2 / u = (1 - t_2 - \tau)/u$. Also, under this assumption, $\eta_{h_1 u}$ (elasticity of $h_1$ with respect to $u$) is 1. Then,

$$\frac{\left(n_2 t'(n_2) - \tau\right) n_2 \left( \underbrace{1 - \frac{\eta_{h_1 R_1}}{h_1 R_1}}_{>0} \right)}{\underbrace{n_2 \underbrace{u E_{2u}}_{=E_2} + n_1 \underbrace{u E_{1u}}_{=E_1 = 1}}_{=n - n_2 t(n_2) - \tau n_2}} < 1. \tag{42c}$$

Further comments on the above inequality (41) and its somewhat specialized version of (42c) are as follows:

(i) The inequality (41) clearly holds for any utility function, as the toll approaches its optimal level since the numerator on the left side goes to zero, $n_2 t'(n_2) - \tau$=0. This can be seen by visual inspection of (42a).

(ii) It also holds for any utility function, as the rent elasticity of compensated demand for core lot size approaches zero. In that case, we can see from (42c), that the inequality reduces to:

$$\left(n_2 t'(n_2) - \tau\right) n_2 < n_1 E_1 + n_2 E_2 \Rightarrow$$

$$\left(n_2 t'(n_2) - \tau\right) n_2 < n_1 + n_2 (1 - t(n_2) - \tau) \Rightarrow$$

$$n_2 \left[ t(n_2) + n_2 t'(n_2) \right] < n.$$

The last line simply says that the total product of a city must be large enough to cover the total social cost of transportation. Therefore, we conclude that,

$$\lim_{\tau \to n_2 t'(n_2)\ \mathit{or}\ \eta_{h_1 R_1} \to 0} \frac{d(SP)}{d\tau} < 0 . \tag{43}$$

Next we turn to the behavior of economic and geographic sprawl, $ES$ and $GS$. In the case of $ES$ , the key expression for decomposing the overall marginal effect of $\tau$ is:

$$\frac{d(ES)}{d\tau} = \underbrace{-m\left(-\frac{dn_2}{d\tau}\right) n_2 t'(n_2)}_{\substack{\textit{Travel cost saved by residents}\\ \textit{of existing suburbs becuse others}\\ \textit{relocated out of those suburbs}\\ \textit{reducing congestion}}} \underbrace{\underbrace{-\frac{n_1}{n_1+n_2} m\left(-\frac{dn_2}{d\tau}\right) t(n_2)}_{\substack{\textit{Travel cost saved by residents}\\ \textit{of existing suburbs relocating to}\\ \textit{the cores of newly formed cities}}} + \underbrace{\frac{n_2}{n_1+n_2} m\left(-\frac{dn_1}{d\tau}\right) t(n_2)}_{\substack{\textit{Added travel cost incurred by residents}\\ \textit{relocating from existing city cores to}\\ \textit{newly formed suburbs}}}}_{=\frac{d(SP)}{d\tau} t(n_2)} \tag{44}$$

In the case of $GS$, the expression that decomposes the overall marginal effect of $\tau$ is:

$$\frac{d(GS)}{d\tau} = \underbrace{\frac{\partial h_2}{\partial u}\frac{du}{d\tau} m n_2}_{\substack{\textit{Aggregate lot size}\\ \textit{increases by those}\\ \textit{remaining in existing}\\ \textit{suburbs.}}} + \underbrace{\hat{h} m\left(-\frac{dn_1}{d\tau}\right)}_{\substack{\textit{Aggregate land added}\\ \textit{by residents of existing}\\ \textit{cores moving to a new}\\ \textit{city's core or suburb.}}} \underbrace{-\underbrace{(h_2 - \hat{h})}_{>0} m\left(-\frac{dn_2}{d\tau}\right)}_{\substack{\textit{Aggregate land saved by}\\ \textit{suburbanites moving out}\\ \textit{of existing suburbs to a}\\ \textit{new city's core or suburb.}}}, \tag{45}$$

where $\hat{h} = \dfrac{H_1 + n_2 h_2}{n_1 + n_2}$ is the average lot size and $h_2 \equiv h(r,u)$ is the suburban lot size. The first effect in (45) occurs because land is a normal good. By (38d) the toll increases utility monotonically in the range $\tau \in \left[0, \tau^*\right]$, hence the suburban residents who are in the suburbs, $mn_2$, increase their consumption of land because the land rent, $r$ , does not change while their utility has increased due to the efficiency improving nature of the Pigouvian toll. The second term in (45) refers to those moving out of existing cores. Recall that by (38c) the rent in the cores falls and lot sizes increase, but because overall land in each core is fixed at $H_1$, this means that some core residents, $m\left(-\dfrac{dn_1}{d\tau}\right)$, will be forced out. On average, this second group of residents will add a lot size of $\hat{h}$ on average, by relocating either to a new city's core or to a suburb. The third term in (45) refers to $m\left(-\dfrac{dn_2}{d\tau}\right)$, those who move, at the margin, out of existing suburbs because of the direct effect of $\tau$ on suburban residents. Each such relocating suburban resident, gives up a lot size $h_2$ and rents, on average, a lot size $\hat{h}$ (because a fraction $n_1 / n$ relocates to a new city core and adds a lot of $h_1 \equiv h(R_1, u)$ while the fraction $n_2 / n$ relocates to a new suburb and adds a lot of $h_2 \equiv h(r, u)$. Thus, on average, sprawl increases by $h_2 - \hat{h} > 0$ per each resident in this third group. In order for congestion tolls to decrease aggregate land use as in the case of the standard monocentric analysis, the third effect is crucial and must dominate the other two effects. In order to understand the third effect we note that,

$$\begin{pmatrix} \textit{Aggregate land saved by} \\ \textit{suburbanites moving out} \\ \textit{of existing suburbs to a} \\ \textit{new city's core or suburb.} \end{pmatrix} = (h_2 - \hat{h})m\left(-\frac{dn_2}{d\tau}\right) = \frac{n_1}{n}(h_2 - h_1)m\left(-\frac{dn_2}{d\tau}\right). \qquad (46)$$

This shows under what circumstances the third effect in (45) tends toward vanishing.

(i) First, it tends toward vanishing when the population residing in a core is small, $n_1 = \dfrac{H_1}{h_1} \to 0$. This can happen, for example, because the core land area, $H_1$, which is exogenous, is small. It could also happen because the demand for land is strong enough that $h_1$, the lot size consumers choose in the core, is large enough. Because the remaining two effects are positive, then $GS$ continually increases with the level of the congestion toll, $\tau$.

(ii) The effect also vanishes when $h_2 - h_1$ is close to zero. This would occur when the rent elasticity of lot size demand is close to zero. In such a case there would be no substitution effect and consumers, being insensitive to rent, would choose identical lot sizes whether they reside in a core or a suburb.

Another way in which sprawl can increase with $\tau$, is for the second effect in (45) to dominate the third. More precisely, such a requirement implies that,

$$\frac{-dn_1/d\tau}{-dn_2/d\tau} > \frac{n_1(h_2 - h_1)}{H_1 + H_2} \qquad (47a)$$

Given any $n_1$, $h_2$, $h_1$, the above condition will hold depending on other effects on the left side. (47a). By using (38a) and (38b), (47a) can also be written as,

$$\frac{\left(n_2 t'(n_2)-\tau\right)\left(\underbrace{h_1 h_{1u}-h_{R1}E_{1u}}_{>0}\right)}{h_1^{\,2}\left(n_2 E_{2u}+n_1 E_{1u}\right)}>\frac{h_2-h_1}{H_1+H_2} \tag{47b}$$

Suppose, just for example, that the utility function is Cobb-Douglas, then (47b) becomes,

$$\frac{2\left(n_2 t'(n_2)-\tau\right)}{\left(n-n_2 t(n_2)-n_2\tau\right)}>\frac{h_2-h_1}{H_1+H_2} \tag{47c}$$

It is easy to see that (47c) is likely to hold especially when the toll, $\tau$ , is close to zero (near the laissez-faire regime, and especially for highly congested cities (i.e. sufficiently high $t'(n_2)$). Furthermore, the circumstances discussed here ignore the income effect (first effect in (45)). Therefore, they are valid when there is no income effect, but a positive income effect would only add to the strength by which *GS* increases with $\tau$ .

# 6. Plan for numerical analysis

Additional insight into the properties of the model can be gained more efficiently by resorting to numerical analysis. Our analysis so far treated the number of cities as a continuous variable. This means that it is approximately valid only when the number of cities is sufficiently large or the value of *k* is sufficiently small. Instead, numerical analysis should be done by treating the number of cities as integer. This immediately leads to an important difference between the short run and long run equilibria of the model.

More precisely, in the short run the number of cities is fixed. Consider the laissez-faire equilibrium in the short run. In this regime the confiscatory land tax in the core is used to finance the infrastructure cost *k* in each city. Now consider levying a congestion toll on suburban residents to achieve the first-best regime, but suppose that congestion and, hence, the optimal toll is not so high that the number of cities under the first-best will be higher than under laissez-faire. To put it differently, for one more city to be created, because the number of cities is integer, congestion in existing cities must be high enough so that the tolls will induce consumers to move to a new city in numbers sufficient to allow the financing of one more *k* and, thus, the creation of one more city. If this is not the case, i.e. congestion is not that high, then the effect of the suburban tolls can only be to induce intra-city relocations. Some consumers in each city will indeed want to move to the core of their own city to avoid the effect of the toll. This would cause the aggregate rents in each core to rise. Since core size is fixed and the infrastructure budget was balanced before the tolls were levied, and since toll revenues are now also available, each city would now show a surplus equal to the increase in the core's land rent plus the aggregate tolls in that city. Since the aggregate surplus from all existing cities is not, according to our assumption above, sufficient to finance a new city, then the surplus in each city should be redistributed to the residents in that city. For congestion tolls to be first-best, as expected, the redistribution together with the congestion reduction should raise utility above its laissez-faire level. For this to be possible for core residents, the income effect in the core from the per-capita redistributed amount net of the rent increase in the core should be high enough to outweigh the substitution effect from the rise in core rents. Otherwise utility could not increase.

Now consider that congestion is high enough in the short run equilibrium. Then, imposing tolls may generate enough aggregate surplus as to make it possible to create one more city. Assume that in fact the surplus is exactly sufficient to create one more city. Then, one more city will be created and in that case the results we have established in the preceding sections, for the long run equilibrium will hold.

## 7. Conclusions

Recent studies on the relationship between traffic congestion and geographic urban sprawl pose a serious challenge to the belief, previously held by planners and economists alike that, without tolling congestion, the market forces induce excessive geographic urban sprawl relative to the efficient allocation (see simulations results obtained by Pines and Sadka (1981), Wheaton (1998), and extrapolated to real cities by Brueckner (2000)). Furthermore, rigorous analytical studies in the 1970s by Kanemoto (1977) and Arnott (1979b) indicated that a policy designed to contain urban growth (that is, a restrictive UGB) may be used as a second best policy and this implication is used by Pines and

Sadka (1985) and Brueckner (2000) to explicitly recommend such a policy, for a monocentric city in the former case and for any city in the latter.

The models used in the aforementioned literature are confined to a single monocentric city. When Anas and Rhee (2006, 2007) explored the issue in a single city with dispersed employment, they obtained an opposite result: optimal zoning and tolling could require the expansion of the urban area rather than restricting it. Furthermore, when Anas and Pines (2007) explored the case of two monocentric cities, they found (this time analytically) that the second-best optimal UGB policy may be expansive rather than restrictive on the aggregate land use while requiring an expansive UGB in the small and a restrictive UGB in the larger city, thus corroborating Anas and Rhee's (2006, 2007) finding.

In the current article, we extended Anas and Pines' (2007) findings to the new context of a city system consisting of identical cities that are run either by a central planner or by profit-maximizing developers. Thus, our analysis is the first that examines the urban sprawl phenomenon by synthesizing the spatial land use model of urban economics with the theory of local public goods in a system-of-cities. In this context, as well, the under-pricing of congestion may result in less not more aggregate urban land use and in long run densities in city cores that are too high not too low, in sharp contrast to conventional beliefs. In addition, as we showed, optimally or sub-optimally alleviating congestion, results in more not less geographic sprawl and in lower not higher densities which, contrary to widely held beliefs, is efficient not inefficient.

Analytically, this paper adapts the insight of Anas and Rhee (2006, 2007) to a system of cities. It turns out that evolving a single monocentric to a single polycentric urban setup is qualitatively very similar to evolving a single monocentric city to a system of monocentric cities. For, the evolution of a monocentric into a polycentric city can be conceived as fragmenting the city into an intertwined layout of separate subareas, each consisting of agglomerated places of work and the neighborhoods that accommodate their employees.

# REFERENCES

A. Anas and D. Pines, 2007. Anti-sprawl policies in a system of congested cities, Working Paper submitted for publication.

A. Anas and R.J. Rhee, 2006. Curbing urban sprawl with congestion tolls and urban boundaries *Regional Science and Urban Economics*, 36, 510-541.

A. Anas and R.J. Rhee, 2007. When are urban growth boundaries not second-best policies to congestion tolls? *Journal of Urban Economics*, 61, 263-286.

R.J. Arnott and J.E. Stiglitz, 1979. Aggregate land rents, expenditure on public goods and optimal city size, *Quarterly Journal of Economics,* 93, 471-500.

R.J. Arnott,1979a. Optimal city size in a spatial economy, *Journal of Urban Economics*, 6, 65-89.
R.J. Arnott, 1979b. Unpriced transport congestion, *Journal of Economic Theory* 21, 294-316.

E. Berglas, 1976. On the theory of clubs, *American Economic Review, Papers and Proceedings,* 66, 116-121.

E. Berglas and D. Pines, 1981. Clubs, local public goods and transportation models: a synthesis, *Journal of Public Economics,* 15, 141-162.

J. K. Brueckner, 2000. Urban sprawl: diagnosis and remedies, *International Regional Science Review* 23, 160-179.

J.M. Buchanan, 1965. An economic theory of clubs, *Economica,* 32, 1-14.

M. Fujita, 1989. *Urban Economic Theory*, Cambridge University Press, Cambridge, UK..

M. Fujita and J-F. Thisse, 2002. *Economics of Agglomeration*, Cambridge University Press, Cambridge, UK.

Y. Kanemoto, 1977. Cost-benefit analysis and the second-best land use for transportation, *Journal of Urban Economics* 4, 483-503.

D. Pines and E. Sadka, 1981. Optimum, second best, and market allocations of resources within an urban area, *Journal of Urban Economics*, 9, 173-189.

D. Pines and E. Sadka, 1985, Zoning, first-best, second-best and third-best criteria for allocating land to roads, *Journal of Urban Economics* 17, 167-183.

F.P Ramsey, 1927, A contribution to the theory of taxation, *Economic Journal*, 37, 47-61.

S.A. Scotchmer and M.H. Wooders, 1987. Competitive equilibrium and the core in club economics with anonymous crowding, *Journal of Public Economics*, 34, 159-174.

J.E. Stiglitz, 1977. The theory of local public goods, in M.S. Feldtsein and R. P. Inman (eds.) The Economics of Public Services, MacMillan, London, 274-333.

C.M. Tiebout, 1956. A pure theory of local public goods, *Journal of Political Economy,* 64, 416-424.

W.C. Wheaton, 1998. Land use and density in cities with congestion, *Journal of Urban Economics* 43, 258-272.

## Appendix A: Complete decentralization of the first-best regime

Formally, the developer solves;

$$\underset{\{\tau,\Omega,s,R_1,R_2,n,n_1,H_2|u,H_1,r,k,t(\bullet)\}}{Maximize} \qquad \pi = H_1(R_1 - r) + n_2 h(R_2,u)s + n_2\tau + n\Omega - k \quad \text{(B1)}$$

subject to (2) –(4). Using (3) and (4), the right side of (B1) can be rewritten as

$$\begin{aligned}\pi &= H_1(R_1 - r) + n_2 h(r+s,u)s + n_2\tau + n_1\left(1 - E\left(R_1,u\right)\right)\\ &\quad + n_2\left(1 - E\left(r+s,u\right) - t\left(n_2\right) - \tau\right) - k \qquad \text{(B2)}\\ &= n - n_1 x\left(R_1,u\right) - n_2\left(x\left(r+s,u\right) + h\left(s+r,u\right)r + t\left(n_2\right)\right) - H_1 r - k.\end{aligned}$$

(B2) states that the city developer's profit is in fact the surplus output of the city, that is, the aggregate output of the composite good minus the output required to pay for the resources in order to provide the $n$ residents-workers with the prevailing utility $u$. The above developer's profits maximization problem is, therefore, equivalent in the short run to choosing $\left\{R_1,n,n_1,s|u\right\}$ that maximizes (B2) subject to (2). The corresponding normalized Lagrangian is

$$\begin{aligned}\Im &= n - \left(n_1 x(R_1,u) + n_2\left(x(r+s,u) + h\left(r+s,u\right)r + t(n_2)\right)\right)\\ &\quad - H_1 r - k - \rho_1\left(n_1 h(R_1,u) - H_1\right).\end{aligned} \qquad \text{(B3)}$$

The first order conditions are:

$$s: \quad -n_2\left(\frac{\partial x(r+s,u)}{\partial R_2} + r\frac{\partial h(r+s,u)}{\partial R_2}\right) = 0, \qquad \text{(B4)}$$

$$R_1: -n_1\left(\frac{\partial x(R_1,u)}{\partial R_1} + \rho_1\frac{\partial h(R_1,u)}{\partial R_1}\right) = 0, \qquad \text{(B5)}$$

$$n:1-\big(x(r+s,u)+rh(r+s,u)+t(n-n_1)+(n-n_1)t'(n-n_1)\big)=0, \qquad \text{(B6)}$$

$$n_1:\big(x(r+s,u)+h\big(r+s,u\big)r+t(n_2)+n_2t'(n_2)\big)-\big(x(R_1,u)+\rho_1 h(R_1,u)\big)=0, \qquad \text{(B7)}$$

that are identical to (8)-(13) with one exception: the value of short run $u$ underlying (B4)-(B7) may differ from the optimized $u$ in (8)-(13). However, it follows from the envelope theorem that,

$$\frac{d\pi}{du}=\frac{\partial \Im}{\partial u}=-\left(n_1\frac{\partial E\left(R_1,u\right)}{\partial u}+n_2\frac{\partial E\left(r,u\right)}{\partial u}\right)<0. \qquad \text{(B8)}$$

We also know that the surplus is zero for the planner's utility maximization problem, and so $\pi$ must be zero also. Hence, $\pi$ is positive (negative) if the short run utility is lower (higher) than the optimal utility. Positive profits attract new developers forming new cities, which drives the utility up and development profits down; losses induce developers to exit the development business which reduces the number of cities which drives utility down and profits up. In the long run profits and losses are eliminated and the prevailing utility achieved by the competitive developers is at its optimal level that the planner could have achieved.