Second–best Congestion Pricing Schemes in the Monocentric City

Erik T. Verhoef

Faculty of Economics and Business Administration, Vrije Universiteit Amsterdam.

SECOND-BEST CONGESTION PRICING SCHEMES IN THE MONOCENTRIC CITY

Erik T. Verhoef[*]
Department of Spatial Economics
Free University Amsterdam
De Boelelaan 1105
1081 HV Amsterdam
The Netherlands
Phone: +31-20-4446094
Fax: +31-20-4446004
Email: everhoef@econ.vu.nl

This version: 01/10/04

Abstract

This paper considers second-best congestion pricing in the monocentric city, with endogenous residential density and endogenous labour supply. A spatial general equilibrium model is developed that allows consideration of the three-way interactions between urban density, traffic congestion and labour supply. Congestion pricing schemes are analyzed that are second-best 'by design' (and not because distortions exist elsewhere in the spatial economy), like cordon charging and flat kilometre charges. Both for Cobb-Douglas utility and for CES utility, the analyses suggest that the relative welfare losses from second-best pricing, compared to first-best pricing, are surprisingly small.

[*]Affiliated to the Tinbergen Institute, Roetersstraat 31, 1018 WB Amsterdam.

1.   Introduction

Practical applications of traffic congestion pricing typically involve so-called second-best pricing regimes, which fail to charge every individual road user his or her exact marginal external congestion costs. With pay-lanes, to an increasing extent employed in the US, unpriced congestion remains existent on parallel highway lanes. In case of cordon charges, such as used in Singapore, every road user passing the cordon pays the same charge independent of the distance travelled and route followed before and after passing the cordon, and users who remain within or outside the cordon do not pay the charge. Area charges, as recently introduced in London, impose the same charge on every user who drives within the area independent of the distance travelled, and leaves congestion outside the area uncharged. And flat kilometre charges, as currently considered for The Netherlands, would not differentiate over time and route followed.

A substantial literature has recently emerged on the economics of second-best congestion charges (*e.g.* Lindsey and Verhoef, 2001, provide an overview). Most of these studies employ partial equilibrium approaches, in which only the transport (network) market is considered explicitly. An exception is the work by authors such as Mayeres and Proost (2001) and Parry and Bento (2001), who study traffic congestion and road pricing for commuters in general equilibrium settings, allowing for distortions on the labour market. Their results suggest that these interactions can be of significant importance for the efficiency impacts of both congestion pricing and the use of the associated revenues.

Another non-transport market that is of importance when evaluating congestion pricing strategies for urban areas is the (spatial) housing market. Already in the 1970's, a number of studies looked into the interactions between traffic congestion and urban structure in the context of the monocentric model (Solow and Vickrey, 1971; Solow, 1972; Kanemoto, 1976; Arnott, 1979). Anas and Kim (1996) and Anas and Xu (1999) extended this line of research by allowing for multicentric configurations, endogenizing the emergence of centres through the explicit consideration of agglomeration forces.

The present paper considers second-best congestion pricing in the monocentric city, with endogenous residential density and endogenous labour supply. A spatial general equilibrium model is developed that allows consideration of the three-way interactions between urban density, traffic congestion and labour supply. The model would therefore, for example, allow an investigation of second-best congestion pricing with distorted spatial labour markets. This matter, however, will be addressed in a companion paper to the present one (Verhoef, 2004). The present paper is instead concerned with congestion pricing schemes that are second-best 'by design', like the examples mentioned above, and not because distortions exist elsewhere in the spatial economy.

Prior studies of traffic congestion in the monocentric model have typically looked at first-best congestion pricing measures, although second-best issues arising from non-optimal allocations of land to road capacity have been considered (*e.g.* Arnott, 1979). The recent contribution by Mun, Kunishi and Yoshikawa (2003) is an exception. They focus on second-

best optimal cordon pricing in a monocentric city. The policy appeared to perform unexpectedly well: when optimizing both the location of the cordon and the charge, welfare gains of around 94% of the gains from first-best pricing were found (computed from their Table 2). This is remarkably well when realizing that with a cordon charge, some road users will not face a congestion charge at all (those who live inside the cordon), some will face a charge that exceeds the marginal external costs that they create over their full trip (those who live outside but relatively close to the cordon), and a third group faces a charge below their marginal external costs (those who live outside and relatively far from the cordon).

Given the potentially far-reaching policy conclusions of this finding, an important question is to what extent the result depends on the assumed monocentric spatial configuration as such, and to what extent it is the result of other specific features of their model, such as the facts that urban densities are assumed given, and that a partial spatial equilibrium model was used. One might for instance hypothesize that an important difference between cordon charge and first-best tolls would be that the former provides a smaller marginal incentive to move closer to the city centre, as there is no reward in terms of a reduced congestion charge. At the same time, however, a cordon tax provides a relatively strong non-marginal, discrete, incentive to choose a location inside the cordon. The question arises whether, as a result of these opposing forces, the average density in the city increases or decreases under cordon charges compared to first-best tolling, and to which extent the discreteness of the charge and the likely resulting discontinuities of land rents and densities at the cordon location induce additional welfare losses. One objective of this paper is to explore these questions by using a spatial general equilibrium model of a monocentric city. Compared to the model of Mun, Kunishi and Yoshikawa (2003), urban density will be made endogenous, trips will be assumed to involve commuting rather than other purposes (such as shopping), and only simultaneous equilibria of the transport market, the urban land market and the labour market are considered. But as in Mun, Kunishi and Yoshikawa (2003), the monocentric urban structure is imposed exogenously, which means in the present model that all production is assumed to take place in a (spaceless) CBD. A model that endogenizes the formation of agglomerations within the urban area is planned for future work.

An alternative second-best policy is considered as well, and this involves flat kilometre charges. This means that a single, 'flat' per-kilometre congestion charge can be imposed throughout the city, also when marginal external congestion costs per kilometre driven vary over space. Like cordon charges, such a policy could be motivated by excessive transaction and implementation costs for first-best congestion charges, typically requiring per-kilometre charges that vary continuously over space. Whereas the cordon tax does imply spatial variation of per-kilometre charges but at the cost of creating a discontinuity, the flat kilometre charge is in some sense its counterpart by avoiding discontinuities while preventing spatial variation of per-kilometre charges.

The two benchmarks against which both policies will be evaluated are the no-toll equilibrium on the one hand, and first-best congestion charging on the other.

2.   The analytical model[1]

This section presents the details of the analytical model. Before turning to a detailed description of consumers' behaviour, the congestion technology, firms, and a characterization of general equilibrium, some introductory remarks are in order. First, z will be used to denote a one-dimensional continuous urban space. The location of the spaceless CBD is at $z=0$, and the residential area stretches from $z=0$ to $z=z^*$, with z^* being the endogenous city boundary. At the boundary of the city, the equilibrium residential bid-rent $r(z^*)$ should be equal to the exogenous and constant agricultural bid-rent r_A. A closed city is considered, meaning that the population size N is treated as fixed and given.

It is assumed that all excess land rents above r_A are redistributed in a lump-sum manner among the city's population.[2] Some share of the urban production will therefore not be consumed in the urban area, but will be 'exported' in exchange for the purchase of land against the agricultural rent.

All consumers and producers are price-takers. Households are identical, and so are firms. The industrial product can be transported costlessly, and the given world-market price of the industrial good p is used as the *numéraire*. We now turn to the various actors in the city and the associated equilibrium conditions.

Consumers

The closed city has N households, which are treated as a continuum of utility-maximizing economic entities. A household's utility depends on the consumption of the industrial good y, on the consumption of space or the size of the residence s, and on the consumption of free time or leisure T_f. A household's financial budget then consists of the net wage rate $w-\tau_L$ (w is the gross wage, τ_L the labour tax) times the amount of hours worked T_w, plus the redistributed excess land rents (R in total, R/N per household), plus – possibly – a lump-sum government transfer G (the government's budget is always balanced). In equilibrium, the household's budget is fully spent on the consumption of y and s, and – if levied – on road tolls and labour taxes. A household's given time budget is denoted by T, and is exhausted by allocation over leisure (T_f), work (T_w) and commuting (T_c). All prices and taxes are treated parametrically by the price-taking households.

Commuting therefore does not require financial outlays other than possibly a total toll (*i.e.*, over the full trip) $\tau_R(z)$, but does take time: $T_t(z)$ per return trip (the underlying travel time function will be discussed below). The number of commuting trips made by a household

---

is assumed to be equal to the amount of effective working time supplied (T_w). Hence, T_w is, as it were, expressed in terms of numbers of days worked, each of a fixed duration in hours. The total time spent commuting can therefore be written as:

$$T_c(z) = T_w(z) \cdot T_t(z) \tag{1}$$

In summary:

$$T = T_w(z) + T_c(z) + T_f(z) = T_w(z) \cdot (1 + T_t(z)) + T_f(z) \quad \Leftrightarrow \quad T_w(z) = \frac{T - T_f(z)}{1 + T_t(z)} \tag{2}$$

A household's simultaneous labour supply and consumption decisions can be modelled by using the 'gross budget' $M(z)$, that would be available under the maximum possible amount of time worked, and to let the household 'buy back' leisure time against the prevailing shadow price which is given by $(w - \tau_L - \tau_R(z))/(1 + T_t(z))$. Observing that the household's optimization problem is dependent on the residential location z, it can namely be written as:

$$\underset{y(z),s(z),T_f(z)}{Max} U\big(y(z), s(z), T_f(z)\big)$$

$$s.t. \ \frac{R}{N} + G + \frac{w - \tau_L - \tau_R(z)}{1 + T_t(z)} \cdot \big(T - T_f(z)\big) - p \cdot y(z) - r(z) \cdot s(z) = 0 \tag{3}$$

so that the 'gross budget' $M(z)$ available at location z is defined as:

$$M(z) = \frac{R}{N} + G + \frac{w - \tau_L - \tau_R(z)}{1 + T_t(z)} \cdot T \tag{4}$$

The earlier assumption on redistribution of excess land rents implies:

$$R = \int_0^{z^*} r(z) - r_A \, dz \tag{5a}$$

Both labour taxes and road tolls are collected by the local government, and the revenues are lump-sum redistributed in equal shares *G*. A balanced government budget therefore implies:

$$G = \frac{1}{N} \cdot \left(\int_0^{z^*} n(z) \cdot T_w(z) \cdot (\tau_L + \tau_R(z)) dz \right) \tag{5b}$$

where $n(z)$ gives the density of households at *z*.

A spatial equilibrium requires that utility $U(z)$ be constant over *z* for all $0 < z \leq z^*$ (and exceeds $U(z)$ for $z > z^*$). This implies a particular equilibrium pattern of land-rents. We can be explicit about this when postulating a specific form for the utility function. Two types of utility function will be considered in this paper: Cobb-Douglas (with a unitary elasticity of substitution) and CES (constant elasticity of substitution). In this analytical section, only the Cobb-Douglas function is considered, which allows for an analytical expression for equilibrium land rents. It is expressed as:

$$U(z) = y(z)^{\alpha_y} \cdot s(z)^{\alpha_s} \cdot T_f(z)^{\alpha_f}$$
$$\text{with}: \quad \alpha_y + \alpha_s + \alpha_f = 1 \tag{6}$$

The unitary elasticity of substitution implies that the gross budget shares spent on y, s and T_f are constant, and given by the parameters α. Specifically, the conditional demands for y, s and T_f are:

$$y(z) = \frac{\alpha_y \cdot M(z)}{p} \tag{7a}$$

$$s(z) = \frac{\alpha_s \cdot M(z)}{r(z)} \tag{7b}$$

$$T_f(z) = \frac{\alpha_f \cdot M(z)}{\left(w - \tau_L - \tau_R(z)\right)/\left(1 + T_t(z)\right)} \tag{7c}$$

and the indirect utility – for analytical convenience defined as the logarithm of the maximum utility achievable under given prices and wage – can be written as:

$$V(z) = \alpha_y \cdot \ln\alpha_y + \alpha_s \cdot \ln\alpha_s + \alpha_f \cdot \ln\alpha_f + \ln\left(\frac{R}{N} + G + \frac{w - \tau_L - \tau_R(z)}{1 + T_t(z)} \cdot T\right)$$
$$- \alpha_y \cdot \ln p - \alpha_s \cdot \ln r(z) - \alpha_f \cdot \ln\left(\frac{w - \tau_L - \tau_R(z)}{1 + T_t(z)}\right) \tag{8}$$

The condition that V in (6) be constant over space implies:

$$V'(z) = \frac{\left(-\dfrac{\tau_R'(z)}{1 + T_t(z)} - \dfrac{\left(w - \tau_L - \tau_R(z)\right) \cdot T_t'(z)}{\left(1 + T_t(z)\right)^2}\right) \cdot T}{\dfrac{R}{N} + G + \dfrac{w - \tau_L - \tau_R(z)}{1 + T_t(z)} \cdot T}$$
$$- \alpha_s \cdot \frac{r'(z)}{r(z)} + \alpha_f \cdot \left(\frac{\tau_R'(z)}{w - \tau_L - \tau_R(z)} + \frac{T_t'(z)}{1 + T_t(z)}\right) = 0 \tag{9}$$

where a prime denotes a 'space derivative' (with respect to location). Equation (9) gives a first-order differential equation for $r(z)$ that can be solved to yield:

$$r(z) = K \cdot \left(\frac{w - \tau_L - \tau_R(z)}{1 + T_t(z)}\right)^{-\frac{\alpha_f}{\alpha_s}} \cdot \left(R + G \cdot N + N \cdot T \cdot \frac{w - \tau_L - \tau_R(z)}{1 + T_t(z)}\right)^{\frac{1}{\alpha_s}} \tag{10}$$

where K is a constant of integration. Invoking the equilibrium boundary condition that $r(z^*) = r_A$, we can solve for K:

$$K = \frac{r_A}{\left(\dfrac{w - \tau_L - \tau_R(z^*)}{1 + T_t(z^*)}\right)^{-\frac{\alpha_f}{\alpha_s}} \cdot \left(R + G \cdot N + N \cdot T \cdot \dfrac{w - \tau_L - \tau_R(z^*)}{1 + T_t(z^*)}\right)^{\frac{1}{\alpha_s}}} \tag{11}$$

We conclude this part of the analysis with a few identities. We can find the local population density $n(z)$ as the inverse of the 'lot-size' $s(z)$:

$$n(z) = \frac{1}{s(z)} \tag{12}$$

The total population is given, so that:

$$\int_0^{z^*} n(z)\,dz = \int_0^{z^*} \frac{1}{s(z)}\,dz = N \tag{13}$$

Total labour supplied equals:

$$L = \int_0^{z^*} n(z) \cdot T_w(z)\,dz \tag{14}$$

Total local consumption of the city's product equals:

$$Y = \int_0^{z^*} n(z) \cdot y(z)\,dz \tag{15}$$

The total amount of land consumed in the city must be equal to z^*, which is by definition true:

$$z^* = \int_0^{z^*} n(z) \cdot s(z)\,dz = \int_0^{z^*} 1\,dz = z^* \tag{16}$$

Travel times

A single radial road of a given, constant capacity is used jointly by all households when commuting. The per-unit-of-distance travel time $t(z)$ for a return trip at each point along the road depends on the local (traffic) density of commuters, defined by the cumulative labour supply between z and z^*. A linear travel time function is used, defined by two constants t_0 (the free-flow travel time for one unit of distance) and t_1 (the function's slope):

$$t(z) = t_0 + t_1 \cdot \int_z^{z^*} n(\zeta) \cdot T_w(\zeta)\,\mathrm{d}\zeta \tag{17}$$

Total commuting time, for a return trip between z and the CBD, can then be written:

$$T_t(z) = z \cdot t_0 + t_1 \cdot \int_0^z \int_\psi^{z^*} n(\zeta) \cdot T_w(\zeta)\,\mathrm{d}\zeta\,\mathrm{d}\psi = z \cdot t_0 + t_1 \cdot \int_0^{z^*} \mathrm{Min}\{\zeta, z\} \cdot n(\zeta) \cdot T_w(\zeta)\,\mathrm{d}\zeta \tag{18}$$

The congestion externality is clearly reflected in (18): $T_t(z)$ depends on labour supply at every location in the city. The same therefore holds for the maximum utility attainable at z; compare equations (3) and, specifically for Cobb-Douglas utility, (8). Because labour supply at z in turn depends on $T_t(z)$ (compare (7c) for Cobb-Douglas utility), the congestion externality induces complex direct and indirect spatial interactions throughout the city, in terms of both equilibrium utility obtained and in terms of labour supply decisions.

Absent economic distortions other than the traffic congestion externality, first-best road pricing involves spatially differentiated per-unit-of distance charges equal to per-unit-of-distance marginal external congestion costs. A problem equivalent to determining these is to find the marginal external congestion costs from supplying one additional unit of labour at every location z, and to derive the total (over the full trip) optimal road prices for trips as a function of trip origin z.

An additional unit of labour supplied at z increases $T_t(\zeta)$ by $t_1 \cdot z$ for $\zeta \geq z$, and by $t_1 \cdot \zeta$ for $\zeta < z$. The additional travel time loss for the household at ζ is therefore $T_w(\zeta) \cdot t_1 \cdot z$ for $\zeta \geq z$, and $T_w(\zeta) \cdot t_1 \cdot \zeta$ for $\zeta < z$. The relevant shadow price of (both leisure and work) time is equal to $(w - \tau_L - \tau_R(\zeta))/(1 + T_t(\zeta))$ (note that the other two consumer prices, p and $r(\zeta)$, are not directly dependent on $T_t(\zeta)$). This shadow price is therefore in part directly dependent on the government's use of the two tax instruments. Because $n(\zeta)$ households will be affected at location ζ, the marginal external costs of supplying one additional unit of labour at z, $mec(z)$ can therefore be written:

$$mec(z) = t_1 \cdot \int_0^{z^*} \text{Min}\{\zeta, z\} \cdot n(\zeta) \cdot T_w(\zeta) \cdot \frac{w - \tau_L - \tau_R(\zeta)}{1 + T_t(\zeta)} \, d\zeta \tag{19}$$

It is verified in the numerical model in Section 3 below that a policy of Pigouvian road taxes, defined by $\tau_R(z) = mec(z)$, with the revenues redistributed in a lump-sum manner via G, indeed maximizes equilibrium utility in the city when the labour τ_L is zero and hence no distortions in the labour market are present.

Producers

Probably the simplest possible production structure is assumed to apply. There is a continuum of firms, each of which is infinitesimally small relative to the market and takes all prices as given. The industrial output is homogeneous. All firms are located in the spaceless CBD, but the agglomeration forces that induce this spatial clustering are not modelled explicitly. This also means that no market distortions through agglomeration externalities are assumed to exist; internal consistency could be achieved by, for example, assuming that zoning regulation prohibits firm location outside the CBD. The assumption of exogenous, central firm locations is clearly an unattractive feature of the present model, which is, however, made solely to allow us to concentrate on the performance of second-best congestion pricing in a monocentric city without introducing additional market distortions arising from agglomeration externalities. Because these market distortions are expected to be relevant in reality, they are considered explicitly in the companion paper Verhoef (2004).

Firms have a simple linear production technology with one input (labour). A firm's production function thus exhibits constant returns to scale, and therefore qualifies for application of Euler's theorem. The following aggregate production function applies:

$$Q = A \cdot L \tag{20}$$

Perfect competition drives profits to zero, with the result that the following equality holds:

$$p \cdot A = w \tag{21}$$

General spatial equilibrium

The model described above has 19 unkowns, some of which are functions of z. These unkowns are $T_c(z)$, $T_w(z)$, $M(z)$, R, G, $y(z)$, $s(z)$, $T_f(z)$, $V(z)$, $r(z)$, K, $n(z)$, L, Y, z^*, $t(z)$, $T_t(z)$, Q, and w (recall that r_A, p and N are given; tax levels are treated as exogenous and all other scalars are parameters). The 19 equations needed to solve this system are (1), (2), (4), (5ab), (7a-c), (8), (10)-(15), (17), (18), (20) and (21). For other types of utility and production functions, as long as they imply unique conditional (factor) demands, a similar equality of numbers of equations and unkowns should in principle hold. We refrain from a formal analysis of existence, uniqueness and stability of equilibria and optima in our model.

In our list of equations, we did not include the 'aggregate demand equals aggregate supply' relation, which in our partly open system reads:

$$p \cdot (Q - Y) = r_A \cdot z^* \tag{22}$$

Equation (22) states that the value of the city's production in excess of its local consumption should be just sufficient to pay for the purchase of land against the exogenous terms of trade r_A/p. The share of local production not exported is consumed locally. The reason for not including this equilibrium condition explicitly is that it will be automatically satisfied under the zero profit condition and exhaustion of consumers' financial budgets – as in fact dictated by Walras' Law. To see why, first observe that zero profits imply that:

$$p \cdot Q = w \cdot L \tag{23}$$

The exhaustion of consumers' total financial income, in combination with the balanced government budget, implies (in aggregate terms) that the sum of redistributed land rents and wage income should be equal to the sum of expenses on the local product and rents:

$$R + w \cdot L = p \cdot Y + r_A \cdot z^* + R$$
$$\Rightarrow w \cdot L = p \cdot Y + r_A \cdot z^* \tag{24}$$

Substitution of (24) into (23) immediately yields (22).

The model does not produce any insightful closed-form (equilibrium) results. We therefore now move to the results of a numerical illustration,[3] to study the comparative static properties of the free-market and some second-best (and first-best) equilibria.

---

[3] The numerical model was written in Mathematica 5.0, and finds spatial equilibria by using a repeated nested approach, in which the various markets of interest are successively brought into equilibrium while keeping other

3. First-best and second-best road pricing: numerical results

3.1. Calibration

The basic numerical model deploys the Cobb-Douglas utility function also used above. Although the model is of course a rather strong abstraction from reality, some effort was put in calibrating the model such that the main endogenous results bear resemblance to what is observed in reality. We therefore start this section by briefly discussing this calibration. It should be borne in mind that the calibration aimed to produce reasonable base-case equilibrium outcomes for a model that is in the first place rather abstract, and that in the second place is calibrated under the conceptually 'clean' but practically unrealistic assumption of zero (labour and transport) taxes.

We start with the normalizations used. As stated, the given world-market price of the good produced in the city is used as a *numéraire*, and p is set equal to 1. Also the agricultural land rent is assumed to be given, and units of space are chosen such that also the agricultural rent $r_A=1$. Next, units of time are chosen such that the total time endowment $T=1$. And finally, the number of households is set at 1000.

The parameters α of the Cobb-Douglas utility function determine the equilibrium (gross) budget shares of the industrial good, housing and leisure. The values were set at $\alpha_y=0.2$, $\alpha_s=0.15$ and $\alpha_f=0.65$. Because the monetary budget is fully spent on the industrial good and on housing, the first two α's imply that some 43% of total monetary income is spent on housing and 57% on other consumption. This seems reasonable for urban areas; 43% on rents or mortgages alone would be on the high side, but when considering other costs that vary directly with house size (energy, decoration, maintenance) it seems a rather plausible figure. The value of α_f, in combination with equilibrium city size and commuting times, leads to an average T_w of 0.237 (see also Figure 1 and Table 1 below). For an average week, consisting of 7·16 hours (excluding 8 hours sleep per day), this means on average 26.5 hours working time per week throughout the year. This includes (bank) holidays, part-time workers, temporary unemployment, *etc.* The most central worker has an average labour supply of around 36.5 hours per week.

---

prices fixed, until convergence is reached (the convergence criterion used in the different loops was set at $1 \cdot 10^{-7}$ for relative changes in key variables between successive iterations within each loop). Provided reasonable starting values are used, this takes (for given policy instrument levels) a few minutes on a modern lap-top computer. The flatness of the utility plots $U(z)$ in Figure 1 provides a graphical illustration of the fact that the model succeeds in producing a spatial equilibrium in which no incentives for relocation remains. Note that $U(z)$ in Figure 1 was calculated by substituting endogenously found (spatial patterns of) consumption levels into (6), so that the flatness is a genuine confirmation of consistency of equilibrium ($U(z)$ itself was not directly equalized over space by the algorithm used, which instead solves for (10) and (11)). Other consistency checks, carried out for every equilibrium found, involvethe zero profit condition for firms; the balanced budget conditions for consumers, the government, and the city as a whole (*i.e.* condition (22)); and the equality of residential and agricultural rent at the city fringe z^*.
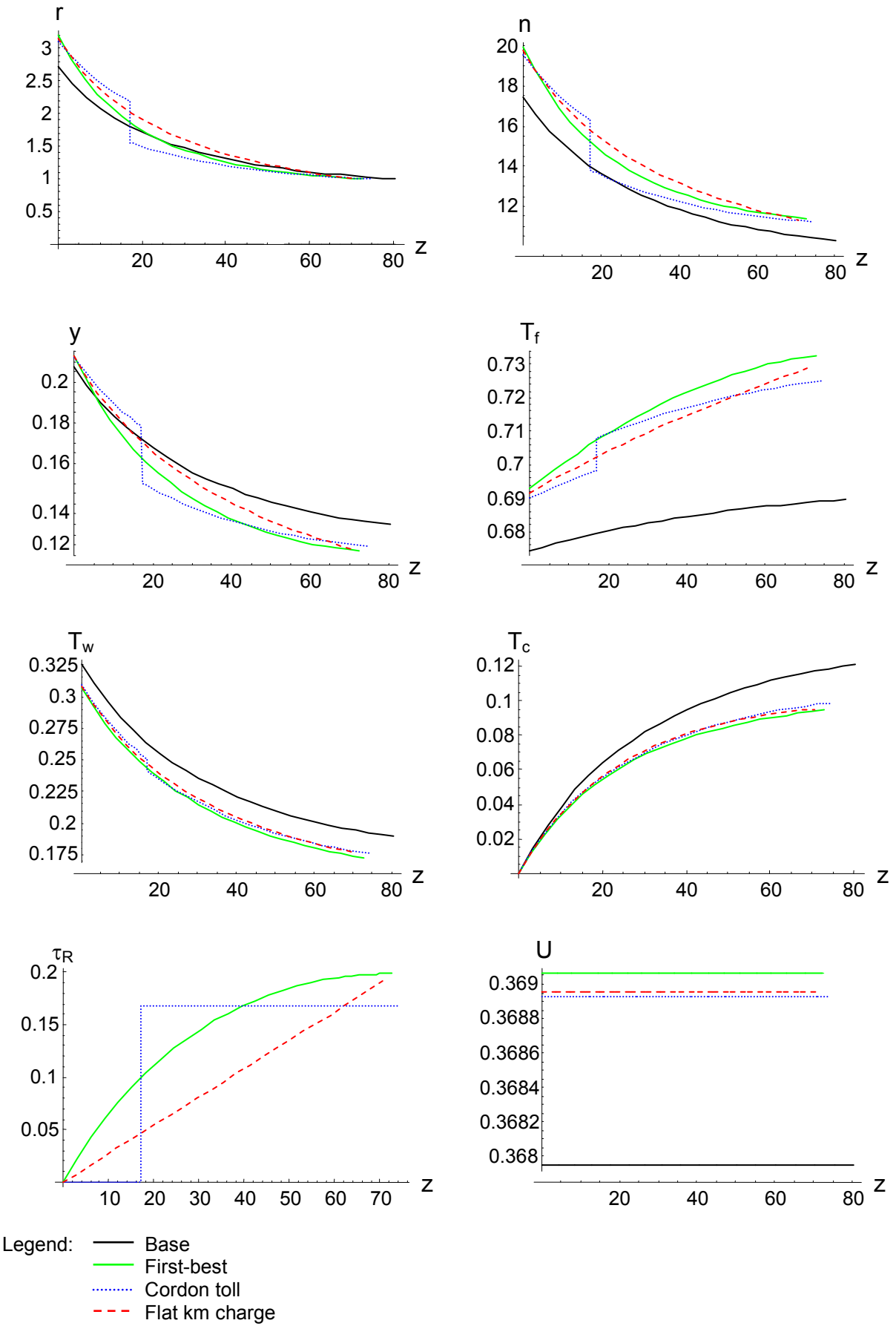
Figure 1. Key results for Cobb-Douglas utility

The final two parameters to be chosen are t_0, set at $3 \cdot 10^{-3}$, and t_1, set at $5 \cdot 10^{-5}$. The ratio of these two parameters causes the equilibrium speed near the CBD to be just over 20% of the free-flow speed (as applying at the city fringe). The absolute sizes of these parameters – together with the further parametrization – cause the equilibrium city size z^* to be such that the person living at z^* has a commuting time $T_c(z^*)$ of $0.64 \cdot T_w(z^*)$. With labour supplied in units of 8-hour working days, this would mean a maximum commute of around 5 hours per working day (for a return trip), or 2.5 hours for a single trip. The average commute takes around one third of average labour supply; implying around one hour and 20 minutes for a single trip.

Finally, in the base equilibrium, wage income forms some 86.5% of total monetary income (the remainder being redistributed excess rents).

Figure 1 show the equilibrium spatial patterns of some further variables of interest. Note for example that residential land rent near the CBD is around three times as high as the agricultural rent and density is nearly twice as high as near the fringe. The convex shape of the equilibrium land rent is caused by both the possibility of substitution in consumption, and the fact that equilibrium per-unit-of-distance transport costs increase towards the CBD.

3.2. *First-best and second-best road pricing: Cobb-Douglas utility*

The model is used to evaluate the performance of two second-best road pricing schemes, relative to the base equilibrium and the benchmark of optimal Pigouvian congestion charging. For each of the three road pricing schemes, it is assumed that toll revenues are redistributed as a lump-sum benefit G, which is the optimal type of redistribution (at least for first-best road pricing) given that no initial labour taxes are existent (the alternative of recycling through negative labour taxes would distort the labour-leisure trade-off). Table 1 summarizes the equilibrium levels of the model's main endogenous non-spatial variables, while Figure 1 compares spatial patterns for the main spatially differentiated variables. Before turning to these, it is useful to discuss the procedures used to find the optimal levels for the tax instruments.

Finding first-best and second-best tax levels

The first-best equilibrium was found by consistently applying space-varying per-unit-of-distance Pigouvian congestion taxes as given in (19):

$$\tau_R(z) = mec(z) \tag{25a}$$

The optimality of this policy – when revenues are redistributed in a lump-sum manner – was verified by investigating equilibrium utility levels for four perturbations of (25a). The first two perturbations applied tax rates $\tau_R(z)=0.9 \cdot mec(z)$ and $\tau_R(z)=1.1 \cdot mec(z)$. The other two tested 'tilted' tax schedules: $\tau_R(z)=(0.9+0.2 \cdot z/z^*) \cdot mec(z)$ and $\tau_R(z)=(1.1-0.2 \cdot z/z^*) \cdot mec(z)$. All perturbations led to utility levels below the first-level ($U=0.369067$), with deviations occurring only from the fifth digit onwards (the fifth and sixth digits became 59, 60, 53, and

52, respectively (in order of appearance of the perturbations above), compared to 67 for first-best pricing).

The second-best cordon tax requires the optimal choice of two instruments: the location of the cordon (z_{cor}) and the toll level (τ_{cor}). The implied road toll becomes:

$$\tau_R(z) = \begin{cases} \tau_{cor} & \text{if } z \geq z_{cor} \\ 0 & \text{otherwise} \end{cases} \tag{25b}$$

As was the case for the model in Mun, Kunishi and Yoshikawa (2003), no closed-form solutions for the second-best optimal levels of z_{cor} and τ_{cor} could be found, and a heuristic grid search method was used to identify these. The method entails two stages. First, 4×4 combinations of z_{cor} and τ_{cor} were tested, and a 'utility hill' as shown in Figure 2 was constructed from the results by means of third-order polynomial interpolation. Its maximum defines the first-round prediction of z_{cor} and τ_{cor}. Next, the same procedure was applied to again 4×4 combinations of z_{cor} and τ_{cor}, where for both a narrower range was chosen, namely between –20% and +20% of the first-round predictions. The prediction of optimal values in this second round was taken to give the second-best optimal instrument levels. Due to the flatness of the 'utility hill' near the second-best optimum, further refinement seems redundant.

Figure 2 shows that equilibrium utility appears to be relatively speaking more sensitive to deviations in τ_{cor} than to deviations in z_{cor} (a similar pattern was also found for the CES utility function). This is in some sense good news for the design of cordon toll schemes when the regulator is uncertain about the second-best optimal levels of τ_{cor} and z_{cor}. Whereas the location of the cordon will often involve relatively large fixed costs due to installation of necessary equipment, toll levels are in principle more flexible. The pattern shown in Figure 2 suggests that a relatively small mistake in the location of the cordon need not cause large relative welfare losses. The instrument for which mistakes are relatively speaking more important – the toll level– is also the one that is probably less costly to adjust in reality.
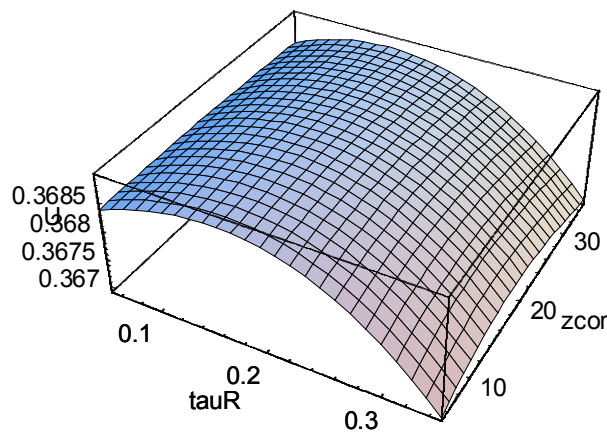

Figure 2. 'Utility hill' for cordon charging

The third congestion tolling scheme considered involves 'flat' (*i.e.*, not differentiated over space) kilometre charges. The tax rate, τ_{km} was again found by a heuristic procedure; a single-nested one-dimensional variant of the procedure used for cordon charges. The implied road tolls now simply becomes:

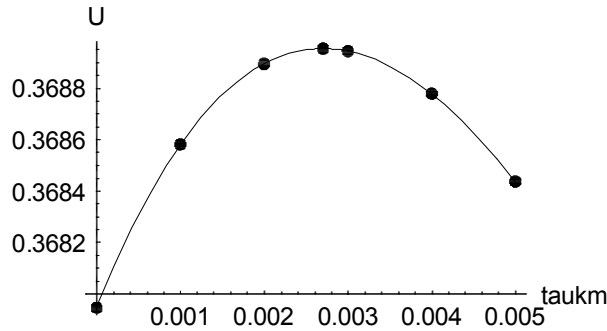$$\tau_R(z) = z \cdot \tau_{km} \tag{25c}$$

Figure 3. 'Utility hill' for flat kilometre charges

Figure 3 depicts the 'utility hill' for this policy, with the central point being the second-best optimum. Again, flatness of the utility hill near the second-best optimum secures that small deviations from the truly second-best optimal tax level will not affect the results in any significant way.

The impacts and relative performance of first-best and second-best congestion pricing

Figure 1 and Table 1 present the main findings. It is instructive to first consider the main impacts of first-best charging. This policy leads to reductions in aggregate kilometrage (*KM*) of around 17% and in aggregate commuting time (*TC*) of nearly 21%. These reductions are achieved partly by a 7.6% decrease in labour supply (and production), and a 9.7% decrease in city size (the latter helps explaining why *KM* decreases more strongly than *TC*). The result is that utility increases by 0.30%; *i.e.*, introduction of optimal congestion pricing raises utility by the same amount as equiproportional increases of *y*, *s* and T_f by 0.30% would. This may seem a modest increase at first sight, but it appears to be reasonably in line with recent predictions of surplus gains from optimal road pricing in urban areas.[4]

Figure 1 confirms that residential density increases throughout the city, while central rents rise and rents near the fringe fall (the latter is consistent with the fringe rent for the smaller optimal city still being equal to the agricultural rent). The consumption of *y* falls at

---

[4] Lindsey *et al.* (2004) provide estimates of annual per capita social surplus gains from first-best road pricing for four European cities (Paris, Brussels, Helsinki and Oslo), which vary from € 111 – 403. The average of these extremes, of around € 257, corresponds, in terms of our model, with 0.30% gain in a household's gross budget if the latter amounts to around € 85 700. If, as in our model, 35% of this gross budget is monetary, the households monetary income should be around € 30 000 to make the welfare gains from the present study comparable to those in Lindsey *et al.* (2004), which appears a reasonable order of magnitude.

nearly all locations (although close to the CBD the redistributed revenues dominate the toll payments and consumption increases), while the consumption of leisure increases as both labour supply and time spent commuting fall. The concave spatial pattern of $\tau_R(z)$ confirms the intuitive notion that per-unit-of-distance tolls are zero at the city fringe and rise towards the CBD. Equilibrium utility of course remains constant over space, albeit that the level increases.

	Base equilibrium	First-best charging[a]	Cordon charging[a]	Flat km charging[a]
Policy instruments				
τ_L	0	0	0	0
τ_R	0	see Fig. 1		
τ_{cor}			0.167	
z_{cor}			17.13	
τ_{km}				0.00271
G	0	0.027	0.023	0.018
Endogenous variables				
L	237.40	93.37%	94.61%	95.31%
Q	237.40	93.37%	94.61%	95.31%
Y	156.86	94.93%	95.70%	99.20%
z^*	80.53	90.34%	92.49%	87.74%
R	37.12	104.88%	102.69%	124.08%
KM^b	7889.58	83.13%	85.57%	82.78%
TC^c	79.65	78.80%	81.32%	80.04%
TF^d	682.95	104.78%	104.05%	103.96%
Toll revenues	0	27.00	23.48	17.66
Tax revenues	0	0.00	0.00	0.00
U	0.367946	100.30%	100.27%	100.27%
ω	0	1	0.880	0.899

Notes:
[a] Percentages are relative to base equilibrium levels
[b] Aggregate kilometrage, defined as $\int z \cdot n(z) \cdot T_w(z)\,\mathrm{d}z$
[c] Aggregate commuting time, defined as $\int n(z) \cdot T_c(z)\,\mathrm{d}z$
[d] Aggregate leisure time, defined as $\int n(z) \cdot T_f(z)\,\mathrm{d}z$

Table 1. The relative impacts of first-best and second-best congestion pricing schemes

Both second-best policies perform rather well in terms of relative welfare gains. Both accomplish nearly 90% of first-best gains, as shown by the efficiency indicator ω (defined as the proportion of the first-best equilibrium utility increase that the policy achieves).

The relative performance of the cordon charge, with ω=0.880, is well in line with the findings of Mun, Kunishi and Yoshikawa (2003), who found ω=0.940 for their basic model. Of course, the relative welfare losses have doubled compared to those reported by Mun, Kunishi and Yoshikawa (2003). Nevertheless, it appears that consideration of neither induced changes in residential densities, nor in labour supply decisions, would undermine to any great extent the significant welfare gains from cordon charging in the monocentric city that were found by Mun, Kunishi and Yoshikawa (2003). Neither the sharp kinks that the policy produces at z_{cor} (see Figure 1), nor the inherent distortions resulting from the inability of charging all road users and from applying imperfect charges to (nearly) all others, apparently

matter too much, provided both the cordon location and toll levels are set in a second-best optimal way.

What causes cordon charges to be so efficient? Part of the explanation lies in the fact that in aggregate terms, the tax induces both a reduction in labour supply and a reduction in city size, which were the main two changes also induced by first-best taxes. The aggregate density increases (compared to the base equilibrium) for two reasons. First, land inside the cordon becomes relatively attractive, which drives up land rents and hence density (see Figure 1). Secondly, the toll discourages labour supply outside the cordon, which translates – via reduced budgets – into lower land consumption and hence a higher density. Labour supply decreases for two reasons. Inside the cordon, the redistributed toll revenues increase the gross budget, which encourages the consumption of leisure. Outside the cordon, the toll discourages labour supply as the shadow price of leisure decreases. Therefore, the cordon tax does induce the same two aggregate responses as first-best tolls do. Of course, the kinks introduced by the cordon toll – illustrated in Figure 1 – lead to welfare losses compared to first-best prices. But the diagrams show that, provided the cordon location and toll level are set optimally, spatial patterns of key variables under cordon charging are nevertheless relatively close to first-best results, with 'too large' levels inside the cordon often compensated by 'too low' levels outside the cordon – or *vice versa*.

The flat kilometre charge, with ω=0.899, performs even better than the cordon charge. Again, the policy succeeds in reducing both labour supply and city size. The modesty of welfare losses compared to first-best pricing is now intuitively explained by the fact that the per-unit-of-distance tax rates are too low near the CBD, and too high near the fringe. Because households only consider the full-trip toll, there is a natural tendency for the two errors to compensate for each other. Certainly, they can not cancel exactly for all z (which is why ω<1), but the optimal flat rate does a good job at finding a reasonably efficient compromise. In aggregate terms (Table 1), most results for flat km charges are comparable to those of cordon charging. The main exceptions are that optimal toll revenues are significantly lower, and aggregate excess land-rents are significantly higher. The latter also exceed first-best rents, which is due to the fact that the equilibrium bid-rent is, as expected, less convex under flat charges than under first-best pricing, while the city size and the central rent are relatively close.

All in all, when comparing the two second-best instruments, the efficiency losses from the inability to differentiate tolls over space (under flat km charges) are apparently somewhat lower than the efficiency gains from avoiding kinks. When comparing both to first-best pricing, it is especially the high relative efficiency that remains surprising. In combination with the results of Mun, Kunishi and Yoshikawa (2003), this raises the hypothesis that the regularity of the spatial lay-out of the monocentric city probably yields good opportunities for minimizing the inherent distortions from which second-best congestion charging mechanisms suffer. This raises the question of whether the monocentric model is the appropriate model for studying the relative performance of such policies in reality. If anything, a generalization of these favourable results to polycentric cities seems premature.

3.3.   First-best and second-best road pricing: CES utility

The Cobb-Douglas utility function deployed up to this point may be critized for its restrictive assumption of unitary elasticity of substitution. It is therefore of some interest to investigate the impacts of the same policies when a somewhat more general constant elasticity of substitution (CES) utility function applies. This leads to changes in equations (6)–(11) above. Using primes to denote the relevant 'CES' equations, the relevant expressions become:

$$U(z) = \left(\left(\delta_y \cdot y(z) \right)^\rho + \left(\delta_s \cdot s(z) \right)^\rho + \left(\delta_f \cdot T_f(z) \right)^\rho \right)^{\frac{1}{\rho}} \tag{6$'$}$$

The elasticity of substitution, σ, is equal to $1/(1-\rho)$; a convenient parameter when working with this type of utility function is $\chi = \rho/(\rho-1)$. The conditional demands for y, s and T_f become:

$$y(z) = M(z) \cdot \frac{\left(\dfrac{p}{\delta_y{}^\rho} \right)^{-\sigma}}{\left(\dfrac{p}{\delta_y} \right)^\chi + \left(\dfrac{r(z)}{\delta_s} \right)^\chi + \left(\dfrac{\left(w - \tau_L - \tau_R(z) \right)/\left(1 + T_t(z) \right)}{\delta_f} \right)^\chi} \tag{7a$'$}$$

$$s(z) = M(z) \cdot \frac{\left(\dfrac{r(z)}{\delta_s{}^\rho} \right)^{-\sigma}}{\left(\dfrac{p}{\delta_y} \right)^\chi + \left(\dfrac{r(z)}{\delta_s} \right)^\chi + \left(\dfrac{\left(w - \tau_L - \tau_R(z) \right)/\left(1 + T_t(z) \right)}{\delta_f} \right)^\chi} \tag{7b$'$}$$

$$T_f(z) = M(z) \cdot \frac{\left(\dfrac{\left(w - \tau_L - \tau_R(z) \right)/\left(1 + T_t(z) \right)}{\delta_f{}^\rho} \right)^{-\sigma}}{\left(\dfrac{p}{\delta_y} \right)^\chi + \left(\dfrac{r(z)}{\delta_s} \right)^\chi + \left(\dfrac{\left(w - \tau_L - \tau_R(z) \right)/\left(1 + T_t(z) \right)}{\delta_f} \right)^\chi} \tag{7c$'$}$$

The indirect utility can be written as (while writing $M(z)$ in full):

$$V(z) = \left(\frac{R}{N} + G + \frac{w - \tau_L - \tau_R(z)}{1 + T_t(z)} \cdot \left(T - T_c(z) \right) \right) \cdot \left(\left(\frac{p}{\delta_y} \right)^\chi + \left(\frac{r(z)}{\delta_s} \right)^\chi + \left(\frac{\dfrac{w - \tau_L - \tau_R(z)}{1 + T_t(z)}}{\delta_f} \right)^\chi \right)^{\frac{-1}{\chi}} \tag{8$'$}$$

The space-derivative of V in (8$'$) is a straightforward but tedious expression and is therefore suppressed, while for the implied first-order differential equation for $r(z)$ – to obtain a constant utility over space – no closed-form analytical solution could be found. Numerical solutions, however, could be obtained, and these will be reported below.

Calibration

To obtain sufficient contrast with the Cobb-Douglas utility function, the elasticity of substitution σ was set equal to 0.5 (with the corresponding ρ and χ following as indicated below (6′)). To maximize comparability with the Cobb-Douglas model, all non-utility parameters were kept unchanged. Only the parameters δ therefore had to be calibrated, and these were set such that with weighted average prices from the Cobb-Douglas equilibrium ($\bar{p} = 1$, $\bar{r} = (R + z^* \cdot r_A)/z^*$ and a shadow price of leisure $\bar{p}_f = (w - \tau_L -$ [total toll revenues over total labour supply]) / (1 + total time spent commuting over total labour supply)), the budget shares from the CES utility function are equal to those from the Cobb-Douglas function. Some basic manipulations reveal that this is achieved when, for consumption good x with a weighted average price $\bar{p}_x$, δ_x from the CES function is related to α_x from the Cobb-Douglas function according to $\delta_x = \bar{p}_x / \alpha_x^{1/\chi}$. This yielded δ_y=25.0; δ_s=64.9275 and δ_f=1.77225. The base-case results in Table 2 confirm that the aggregate levels of equilibrium variables with the CES utility function are indeed close to those with the Cobb-Douglas function in Table 1. The spatial patterns shown in Figure 4 also reflect a close correspondence with the Cobb-Douglas results, albeit that the lower elasticity of substitution now causes a somewhat less pronounced differentiation of quantities consumed over space.

	Base equilibrium	First-best charging[a]	Cordon charging[a]	Flat km charging[a]
Policy instruments				
τ_L	0	0	0	0
τ_R	0	see Fig. 2		
τ_{cor}			0.185	
z_{cor}			18.86	
τ_{km}				0.00277
G	0	0.030	0.0274	0.020
Endogenous variables				
L	235.01	95.74%	96.29%	96.74%
Q	235.01	95.74%	96.29%	96.74%
Y	155.12	96.57%	96.77%	98.81%
z^*	79.88	94.13%	95.37%	92.71%
R	38.50	112.08%	108.73%	133.60%
KM^b	8187.35	89.55%	90.94%	89.39%
TC^c	83.93	86.66%	88.05%	87.40%
TF^d	681.06	103.11%	102.75%	102.68%
Toll revenues	0	30.40	27.40	20.25
Tax revenues	0	0.00	0.00	0.00
U	0.780405	100.20%	100.18%	100.18%
ω	0	1	0.897	0.911

Notes:
[a] Percentages are relative to base equilibrium levels
[b] Aggregate kilometrage, defined as $\int z \cdot n(z) \cdot T_w(z) \, dz$
[c] Aggregate commuting time, defined as $\int n(z) \cdot T_c(z) \, dz$
[d] Aggregate leisure time, defined as $\int n(z) \cdot T_f(z) \, dz$

Table 2. The relative impacts of first-best and second-best congestion pricing schemes: CES utility

First-best and second-best congestion pricing

Table 2 shows the results of first-best and second-best congestion pricing for the CES utility function. The main conclusions are (1) that relative gains from congestion pricing, compared to the base equilibrium, are lower than under Cobb-Douglas utility due to the reduced sensitivity of households to price differences; and (2) the relative welfare gains of the two second-best policies, compared to first-best welfare gains, are nearly identical to those under Cobb-Douglas utility. In other words, the lower elasticity of substitution affects the size of welfare gains from congestion pricing, but not the relative welfare gains from different pricing schemes.

An intuitive explanation is as follows. The higher elasticity of substitution for the Cobb-Douglas utility function than for the CES function leads to larger behavioural adjustments (in terms of the three conditional demands $y(z)$, $s(z)$ and $T_j(z)$) under optimal pricing, which boosts the relative size of welfare gains for all pricing schemes. Second-best pricing gives imperfect incentives to individuals to adjust their demands. But the relative deviations from first-best behavioural adjustments under second-best pricing for each consumption good are comparable for both utility functions, because both have constant and symmetric elasticities of substitution. And this causes the relative performance of the two second-best instruments to be similar for both types of utility function.

4.   Conclusion

The results presented in this paper suggest that the surprisingly optimistic conclusions that Mun, Konishi and Yoshikawa (2003) reach on the relative performance of cordon pricing in the monocentric city are robust with respect to the inclusion of residential land markets, endogenous labour supply and general spatial equilibrium formulation conditions. Moreover, the result is obtained both for a Cobb-Douglas and for a CES utility function (with a lower elasticity of substitution). This raises the suspicion that the regular monocentric configuration, which moreover ignores transport network effects, may be the responsible factor for this counter-intuitive result. The analysis furthermore showed that also flat kilometre charges perform surprisingly well in this setting – even slightly outperforming cordon charges.

A future research agenda is easily sketched. One line of research would endogenize the formation of (sub-)centres, by endogenizing firm location decisions and agglomeration advantages. A second line of research enabled by the model proposed in this paper concerns the investigation of second-best distortions in congestion pricing as arising from the existence of distortionary labour taxes in a spatial general equilibrium setting.
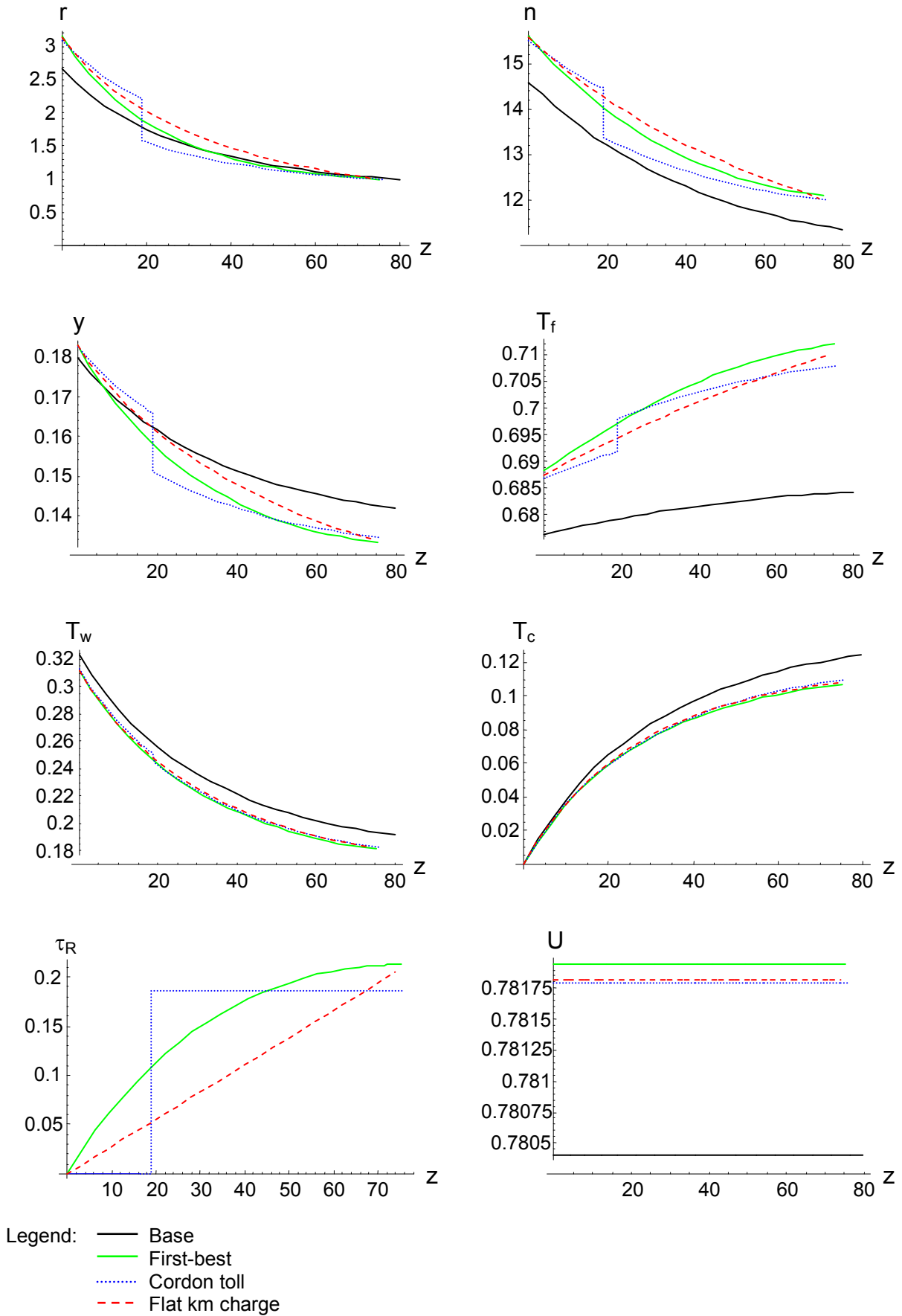
Legend:
— Base
— First-best
‥‥ Cordon toll
— — Flat km charge

Figure 4. Key results for CES utility

References

Anas, A., and I. Kim (1996) "General equilibrium models of polycentric urban land use with endogenous congestion and job agglomeration" *Journal of Urban Economics* **40** 232-256.

Anas, A., and R. Xu (1999) "Congestion, land use and and job dispersion: a general equilibrium model" *Journal of Urban Economics* **45** 451-473.

Arnott, R. (1979) "Unpriced transport congestion" *Journal of Economic Theory* **21** 294-316.

Kanemoto, Y. (1976) "Cost-benefit analysis and the second-best land use for transportation" *Journal of Urban Economics* **4** 483-503.

Lindsey, C.R. and E.T. Verhoef (2001) "Traffic congestion and congestion pricing". In: D.A. Hensher and K.J. Button (eds.) (2000) *Handbook of Transport Systems and Traffic Control, Handbooks in Transport 3* Elsevier / Pergamon, Amsterdam, pp. 77-105.

Lindsey, C.R., E. Niskanen, E.T. Verhoef, A. de Palma, P. Moilanen, S. Proost and A. Vold (2004) "Implementation paths for marginal-cost-based pricing in urban transport: theoretical considerations and case study results" unpublished manuscript.

Mayeres, I. and S. Proost (2001) "Marginal tax reform, externalities and income distribution" *Journal of Public Economics* **79** 343-363.

Mun, S., K. Konsihi and K. Yoshikawa (2003) "Optimal cordon pricing" *Journal of Urban Economics* **54** 21-38.

Parry, I.W.H. and A.M. Bento (2001) "Revenue recycling and the welfare effects of congestion pricing" *Scandinavian Journal of Economics* **103** 645-671.

Solow, R.M. (1972) "Congestion, density and the use of land in transportation" *Swedish Journal of Economics* **74** 161-173.

Solow, R.M. and W. Vickrey (1971) "Land use in a long, narrow city" *Journal of Economic Theory* **3** 430-447.

Verhoef (2004) "Traffic congestion and spatial labour markets" Dept. of Spatial Economics, Free University Amsterdam (currently in progress).

Verhoef, E.T. and P. Nijkamp (2002) "Externalities in urban sustainability: environmental versus localization-type agglomeration externalities in a general spatial equilibrium model of a single-sector monocentric industrial city" *Ecological Economics* **40** 157-179.