Social Optimality of Cordon Area Congestion Pricing in a Monocentric City¹

Harya S. Dillon ²

[Diterima: 9 Juni 2014; disetujui dalam bentuk akhir: 4 November 2014]

Abstract. Traffic congestion is a global epidemic that at time has put cities to a state of paralysis. From a microeconomics point of view, congestion can be approached as a negative externality that merits a Pigovian tax to correct the suboptimal equilibrium. Externalities manifest in delays and wasteful urbanization. While the efficiency of congestion pricing has been well established since the late 1970s, policy adoption and implementation have been delayed due to technological constraints. One the most popular form of congestion pricing is cordon charging where inbound commuters are charged a certain fixed fee. The efficiency of cordon pricing has been explored by numerous empirical and simulation studies. However, theoretical explanations of results are largely missing. I consider a monocentric city model to explore the effect of cordon charging schemes on urban density and housing demand (equilibrium rent). Cordon pricing increases rents at the center and also the slope of the bid-rent curve. Most importantly, these quantities are functions of cordon parameters which suggest that efficiency is affected by planners’ choice of cordon parameters.

Keywords. Congestion, cordon pricing, monocentric city, urban structure

¹ An earlier version of this article was presented at Seminar Nasional 55 Tahun Pendidikan Planologi: Peran dan Pendidikan Perencanaan Wilayah dan Kota di Era Desentralisasi dan Demokrasi in Bandung on September 18, 2014.
² Department of Planning, Policy and Design, University of California, Irvine, email: hdillon@uci.edu
Introduction

Urban Traffic Congestion

Traffic congestion is one of the most challenging problems facing urban managers in both industrialized and developing cities. While some other problems such as public health and education tend to get better with income growth, absent strong intervention, traffic congestion tends to get worse. Since road capacity expansion is not always feasible, effective, or efficient (Wang, Yang, & Han, 2010; Small & Verhoef, 2007), alternative congestion relief measures are needed. Before we proceed, however, it is important to define congestion.

The most important measure of the quality of service in transportation is expected travel time (Small & Verhoef, 2007). Thus traffic congestion occurs when the quality of transportation service is decreasing in intensity of use (Small & Verhoef, 2007). As a public good problem, we can also define the urban road network as a congestible public good (access is non-excludable but congestion makes it partly rival). The level of service can be measured by travel speed; it is increasing in the capacity of the road network and decreasing in the number of vehicles.

Congestion has many causes. Vehicle queues can be formed by traffic signals at intersections, vehicles may be forced to travel at lower speed by slower vehicle ahead, and blockage can be caused by accidents or limitations in geometric highway design. Accordingly, there are many approaches to analyze congestion. Small & Verhoef (2007, p. 70) provide a useful distinction between stock and flow congestion. Stock congestion is characterized by bottlenecks or queuing which depend on prior condition (the arrival of prior vehicles in the system) while flow congestion is identified as lower travel speed caused by higher traffic intensity that is independent of prior traffic conditions. This essay focuses on the latter.

Congestion Relief

Many policy responses have been considered to tackle congestion. For brevity, I categorize these responses as quantity and pricing interventions. In addressing stock congestion, North American cities and transportation agencies have introduced ramp-metering to regulate vehicle entrance onto freeways. A sudden inflow of vehicles onto the freeway may create bottlenecks that would reduce travel speeds around on-ramps. Ramp-metering regulates vehicle inflow onto freeways and thus minimize the risk of slowdowns after freeway on-ramps. Ramp-metering may also discourage short freeway trips and it is effective when congestion is mild (Mohring, 1999, p. 190).

The crudest form of quantity intervention for congestion relief is road space rationing and access restrictions for heavy duty vehicles. Road rationing restricts vehicle use based on license plate enforcement, during specific times of the day (usually peak hours) and days of the week. This type of intervention is relatively common in South American cities such as Bogota, Medellin, Mexico City, Santiago and Sao Paulo (Wang, Yang, and Han, 2010). Similarly, the Singapore vehicle registration system charges lower registration fees for vehicles that can only operate during weekends and evenings.

Another popular congestion relief measure is the introduction of High Occupancy Vehicle (HOV) lanes. An HOV lane is a reserved lane with higher level of service but access is reserved

---

3 Commonly known as ‘Pico y Placa’ or peak (hours) and (license) plate in South American cities.
to vehicles with two or more occupants (at the minimum, the driver and a passenger). Many HOV lanes in California require 2 occupants but in some cases three occupants are required. This policy also incentivizes carpooling and thus reduces the number of vehicles on the road, which reduces travel times. In recent years, HOV lanes have expanded access to low emission vehicles and some have evolved into paid toll lanes where Single-Occupancy-Vehicles (SOVs) are allowed to buy underutilized space (Mohring, 1999, p. 188).

Peak travel demand is partly caused by firms’ coincident operating hours. The benefits of common operating hours comes at the cost of congestion (Mohring, 1999, p. 191). Staggered or even flexible working hours can help reduce peak congestion. Large employers can internalize the effects of their working-hour decisions by staggering business hours.

Many cities have also resorted to public transit to try to relieve congestion. Often this requires improving transit services and providing subsidies to lure peak-period passangers out of their automobiles (Mohring, 1999, p. 192). Subsidies to public transit imply that pricing schemes and quantity regulation are not mutually exclusive. However, quantity based intervention may not yield an efficient outcome and pricing may induce quantity responses. For example, if pricing applies during peak periods, some travelers may respond by delaying trips to off-peak hours. HOV lanes and transit subsidies are both subsidies in different currencies. The former uses time-saving while subsidies are monetary (Mohring, 1999). Lastly, transit can be the preferable alternative if driving is priced efficiently (Small & Verhoef, 2007, p. 160).

**Congestion Pricing in General**

Congestion can be understood as a form of market failure where unfettered markets fail to allocate goods and services efficiently. Intuitively, a road (a single road segment or a network) with fixed capacity is a congested public good where an additional road user (or motorist) increases the travel time of other road users. The decision by a single motorist to travel is based on his/her own valuation of travel trade-offs. Absent any corrective intervention, an additional motorist is only ‘paying’ for his/her own travel time and does not take into account his/her decision to use the road on the travel time of other road users, which creates a negative externality. This additional cost to society is called the marginal social cost (MSC) and it includes the marginal external congestion cost (MECC). The latter is the joint increase in travel time of all other motorists, multiplied by the value of time, caused by an additional motorist (Small & Verhoef, 2007). This cost is increasing faster than the average cost, which is the average time it takes to complete the travel multiplied by motorist’s value of time. The unfettered equilibrium where motorists’ behavior is guided solely by private cost-benefit trade-offs calls for more traffic than the social optimum. The excess traffic level creates a dead weight (welfare) loss to society. In this context, congestion relief is a policy response that aims to decrease this dead weight (welfare) loss, which represents excessive commute times due to congestion.

Like other market failures, congestion can be relieved by implementing a well-designed intervention\(^4\). Microeconomic theory provides a compelling case for marginal cost pricing in urban transportation network (Rouwendal & Verhoef, 2006). The basic principle is to levy a toll that equals the marginal external congestion cost that a motorist imposes on all motorists for

---

\(^4\) This paper focuses on congestion only. There are other market failures associated with urban transportation such as environmental externalities and accidents that can also be addressed by some sort of pricing instruments.
his/her use of the road. In the economic literature, this is known as ‘first-best’ pricing principle where prices equal marginal costs. From a welfare analysis point of view, the toll on a motorist ‘corrects’ the private cost that he/she faces under the unfettered/user equilibrium, resulting in lower traffic levels and lower social costs.

**Congestion pricing and urban spatial structure**

Cordon pricing is relatively easy to implement and it has been adopted in several cities including Singapore, Stockholm and London (Small & Verhoef, 2007). Other cities have mainly considered cordon charging schemes to implement congestion pricing, such as San Francisco. The absence of congestion pricing has also been analyzed as a cause of inefficient urban housing allocation. In this case, the welfare loss manifests itself as urban sprawl, i.e., in a spatial footprint larger than socially desirable. Brueckner (2007) and Brueckner & Helsley (2011) have shown that unpriced congestion leads to lower density in the central city and larger urban spatial footprint overall (urban sprawl). However, these authors only considered a ‘first-best’ congestion pricing schemes to make comparisons with the base-case (unfettered sprawling) and other anti-sprawl measure such as urban growth boundaries and development impact fees. Verhoef (2005) used simulation to demonstrate how equilibrium bid-rent curves and urban population gradients shift and tilt under various congestion pricing schemes.

In the context of urban sprawl, Anas (2012) independently devised a discrete urban space model setup identical to Brueckner & Helsley (2011) to extend the analysis to cities with two-modes of transportation: a congested urban road network and an un-congested public transit system with strong scale economies. The discrete urban space consists of a core city where land supply is fixed and a suburb where urban land competes with agricultural land; the two are connected by a bridge. In both papers, it is assumed that suburban commuters pay a toll for using the bridge (inbound to core city), which mimics cordon pricing with an exogenous location.

**Cordon pricing as a second-best pricing scheme**

Although economists have suggested first-best pricing schemes that apply network-wide, their implementation lacks political support because voters tend to view congestion pricing as a new tax and are thus unwilling to approve its implementation. Borger and Proost (2012) built a voting model to show how heterogeneous voters may vote against congestion pricing proposals. Russo (2013) pointed out how regional government structure may explain why some cities have voted for congestion pricing while others have opposed it. Both papers considered the use of revenue in the voting equilibrium. The case of New York is instructive. In 2008, New York’s Mayor Bloomberg proposed a congestion pricing scheme that passed in the City Council but was blocked by New York State Assembly. Since then New York has implemented other physical restrictions on vehicular traffic and parking fees in Manhattan have increased substantially (Russo, 2013).

Until recently, technology was also an obstacle because the implementation of first-best pricing entails pricing by network link and time of day as traffic fluctuates, which requires sophisticated technology. Thanks to transponders and other on-board devices that can track vehicles movement, this obstacle is no longer binding.

Given these difficulties, only second-best schemes in the form of cordon pricing have been
Social Optimality of Cordon Area Congestion Pricing in a Monocentric City

implemented to-date.\(^5\) Cordon pricing charges inbound commuters a fixed fee upon entry in a delineated area; they are second-best congestion pricing schemes “by-design” (Verhoef, 2005). Their earliest implementation took place in Singapore and dates back to 1975; it relied on a simple access licensing system enforced using stickers on vehicles (voting is not an issue in Singapore). Other cities where cordon pricing was implemented successfully (both politically and technically) include London and Stockholm.

Recent studies by Russo (2013) and de Borger & Proost (2012) have incorporated voting equilibrium in analyzing the ratification of congesting pricing schemes. The former considered a regional urban governance structure, while the latter considered the case of two transport modes and revenue recycling schemes. Another extension on congestion pricing models treats household labor supply as endogenous to capture the extent to which households may trade-off earned wages with tolls (Verhoef, 2005; Anas & Xu, 1999; Anas & Hiramatsu, 2013).

Modeling Framework

The design of a cordon pricing scheme includes two parameters; the toll level \((\tau_{cor})\) and the cordon location \((x_{cor})\). Previous theoretical studies on cordon pricing schemes have yet to find a closed-form solution for the decision pair \((x_{cor}, \tau_{cor})\) (Verhoef, 2005; Mun et al., 2003; Mun et al., 2005). They have all used numerical simulations to show that the welfare loss is quite small, and that traffic volumes inside the cordon are slightly larger than the first-best optimum while traffic outside the cordon is slightly larger than the first-best optimum (Mun et al., 2003). Moreover, the effect of cordon pricing on the urban structure is not yet fully understood. Verhoef (2005) improved upon Mun et al. (2003, 2005) by allowing for endogenous city boundary, residential density and labor supply, but a closed-form analytical solution was not found. Their simulation results show a kink in the bid-rent curve and residential density gradient at the cordon location, as expected.

I therefore propose a model that is simple enough to yield a closed-form analytical solution but complex enough to capture key features of this problem. A closed-form analytical solution is useful to conduct a formal comparative statics analysis.

Spatial setup of a monocentric city model

This section presents a preliminary analytical model. For familiarity, I follow the notation of Brueckner (1987). Following the standard monocentric city model, the city is a disk with a fixed radius \(B>0\) (for modeling convenience) and city residents commute to a central (and space-less) job center (CBD). This framework implies strong economies of agglomeration whereby all firms/employers are better off locating in a high density urban cluster. While this assumption is reasonable for many urban areas, it has been criticized by those unconvinced about the presence of economies of agglomeration. However Brueckner (1987) reasons that lessons from a stylized monocentric model are largely unchanged in a polycentric setting. Holding everything else constant, firms that choose to move out from the CBD to a new location closer to households can offer a lower wage to equalize disposable income (after accounting for transportation costs) up to the point where households are indifferent between working in CBD firms and off-CBD firms.

\(^5\) Congestion pricing can also take the form of express lane tolling or ‘value pricing’ (see Small & Yan (2001) and flat mileage fee (km-charge). Both usually combines road use charges that is used to cover road construction and maintenance costs.
In the model, \( x \in (0, B] \) denote the areal distance between the CBD (located at \( x = 0 \)) and a point in the city; the area of the city is \( A = \pi B^2 \). I assume that city limits are set exogenously by physical boundaries (for example an island city), but it is clear that the spatial size of any city is bounded because (i) transportation costs are increasing with \( x \) and resident can’t live beyond the point where their income is spent entirely on transportation (Saphores & Boarnet, 2006); and (ii) urban land use competes with non-urban use (e.g., agriculture). Set exogenously, agriculture rent determines city limit where housing rent and agriculture rents are equal (see Brueckner (1987) for an exhaustive explanation).

**Households and congestion**

For simplicity, the city population is made up of \( N \) households assumed to be identical in terms of income, \( y \), and preferences. To get started, let us assume that \( N \) is fixed for now, following the “closed-city” case where migration does not occur (Brueckner, 1987).

First, let us consider a city without congestion pricing policy. The presence of congestion increases per-mile transportation cost, which is assumed to increase with the city’s total population, denoted by \( N \):

\[
l(N) = \lambda + \gamma N^\delta
\]

This functional form is a special case of the Bureau of Public Road (BPR) delay function where highway capacity is normalized to 1. In this equation, \( \lambda \) is the pecuniary per-mile cost of transportation (e.g. fuel, vehicle depreciation), and \( \gamma, \delta > 0 \) are congestion parameters. For modeling convenience, commuting is assumed to be the only available mode and vehicles are assumed to be alike. This transportation cost function assumes that congestion depends on total city population as in Saphores and Boarnet (2006). Another possibility Brueckner (2007) is to have congestion at \( x \) be a function of the population that lives beyond that point. However, it does not lead to an explicit solution.

With these assumptions, the budget constraint of a household living at a distance \( x \) from the CBD can be written:

\[
y - l(N)x = c + \tau q
\]

where:

- \( q \) is the amount of housing consumed (i.e., housing size);
- \( r \) is rent per unit housing; and
- \( c \) is the composite good other than housing (or a numeraire) where price is normalized to unity.

I also assume that household utility can be described by the Cobb-Douglass utility function:

\[
U(r, q) = c^\alpha q^\beta
\]

where \( \alpha, \beta > 0 \). It is noteworthy that in equilibrium \( r, q, \) and \( c \), vary with \( x \) because disposal income is declining due to transportation costs. Also, utility at equilibrium depends on \( N \) and \( B \) through demand and supply respectively. For brevity, other subscripts are suppressed.

**Demand for housing in equilibrium**
By properties of the Cobb-Douglass function, the demand for housing and numeraire of utility-maximizing households at location $x$ are given by:

$$
q(x) = \frac{\beta}{\alpha + \beta} \frac{y - t(N)x}{r(x)} \quad (4)
$$

$$
c(x) = \frac{\alpha}{\alpha + \beta} (y - t(N)x) \quad (5)
$$

In equilibrium, all residents have the same level of utility so they do not have an incentive to move elsewhere in the city. Writing that the utility of a household living at distance $x$ from the CBD equals the utility of a household living by the CBD gives

$$
r(x) = r(0) \left( \frac{y - t(N)x}{y} \right)^{\nu}, \quad (6)
$$

where $\nu = (\alpha + \beta)/\beta$ and $r(0)$ is the rent at the CBD, which is highest in the city. Since $\frac{y - r(N)x}{y} \in (0, 1)$ and $\nu > 1$ rent falls exponentially with $x$ at a rate proportional to transportation cost.

The second equilibrium condition requires that all $N$ households must fit in the city. Following (Brueckner, 1987), this condition implies that

$$
N = \int_0^B \frac{2\pi x}{q(x)} dx. \quad (7)
$$

Substituting equation (4) and (5) into (6) leads to:

$$
r(0) = \frac{(\nu - 1)^2 \nu^{\nu N}}{2\pi y} (y^{\nu + 1} - (y - t(N)B)^\nu (y + \nu t(N)B))^{-1}. \quad (8)
$$

If $\alpha = \beta$, a special case where housing and the numeraire are equally preferred, then $\nu = 2$ and the equation above simplifies to:

$$
r(0) = \frac{Ny^2}{2\pi B^2} \left( y - \frac{2}{3} t(N)B \right)^{-1}. \quad (9)
$$

where $N/\pi B^2$ is the city’s population density. This implies that CBD rents are increasing with total population and decreasing as the radius of the city grows larger. Substituting (8) into (5) yields:

$$
r(x) = \frac{N}{2\pi B^2} \frac{(y - t(N)x)^2}{y - \frac{2}{3} t(N)B} \quad (9)
$$

$$
\frac{\partial r(x)}{\partial x} = \frac{N}{2\pi B^2} \frac{-2t(N)(y - t(N)x)}{y - \frac{2}{3} t(N)B} < 0
$$
Let us call this result the laissez-faire equilibrium. The slope of bid rent function is negative and proportional to transportation cost. The next section introduces congestion pricing and compares results with the laissez-faire equilibrium.

**Effect of First-Best Congestion Pricing**

Let us now suppose that each household must pay a congestion price that equals the marginal external cost of commuting (Small & Verhoef, 2007, p. 122; Brueckner, 2007). Let us denote the first-best congestion toll by:

\[ \tau^{FB}(N) = N. t'(N) = \delta yN^6 \]  

(10)

This toll is levied as a flat per-mile rate\(^6\) on households. This changes the budget constraints equation (2) to:

\[ y - (t(N) + \tau^{FB}(N))x = c + r \]  

(11)

The per unit rent at the CBD becomes:

\[ r^{FB}(0) = \frac{Ny^2}{2\pi B^2} \left( y - \frac{2}{3}(t(N) + \tau^{FB}(N))B \right)^{-1} \]  

(12)

It is straightforward to check that \( r^{FB}(0) > r(0) \) and also that \( \frac{\partial r^{FB}(x)}{\partial x} > \frac{\partial r(x)}{\partial x} \). This implies that congestion pricing increases CBD rent and that the slope of the bid-rent curve is steeper under a pricing regime. Therefore, these preliminary analytical results show that the implementation of congestion pricing affects urban spatial structure as in Brueckner (2007) and Brueckner & Helsley (2011).

Numerical results in Brueckner (2007) show an increase of up to 500% in housing rents adjacent to CBD, depending on exogenous parameters. Brueckner & Helsley (2011) used a discretized urban model and quasi-linear utilities to show that the urban boundary shrinks as a result of congestion pricing. Furthermore, population in the central city is higher under this pricing regime. Population and density should be proportional to rents because density is inversely proportional to housing size and housing size is inversely proportional to rents. Thus higher CBD rents imply higher densities in the inner rings of the city. Simulations by Verhoef (2005) are also consistent with this result.

**Impact of Cordon Pricing on Rent**

Cordon pricing is implemented by charging inbound commuters a toll \( \tau_{cor} \) as they pass \( x_{cor} \). Thus the budget constraints of households leaving inside and outside of the cordon are now:

---

\(^6\) This is called a flat-mile charge in Verhoef (2005), because congestion is assumed to depend on total population and not on population distribution across the city.
\[ y - \tau_{cor} - t(N)x = c + rq, \text{ for } x \in (x_{cor}, B] \]
\[ y - t(N)x = c + rq, \text{ for } x \in [0, x_{cor}] \text{ as equation (2)} \]

Equation (13) describes how households who live just outside of the cordon have to pay the toll \( \tau_{cor} \), whereas households within the cordon are exempted. The simulated bid-rent curve in Verhoef (2005) shows a discontinuity/drop of rent at the cordon, which is to be expected as the disposable income between households who live just inside and just outside of the cordon differs by \( \tau_{cor} \).

Let us show this result analytically. To do so, let us find rents at a point just adjacent to the cordon ring. Let \( x_e \) be such that \( x_e = x_{cor} + \varepsilon \) for an infinitesimal \( \varepsilon \). Households at \( x_e \) pays \( \tau_{cor} \) for commuting. Using the budget constraints in (13) and proceeding as before gives:

\[ r(x_e) = r(0) \left( \frac{y - \tau_{cor} - t(N)x_e}{y} \right)^2 \]

Note that \( x_e \) is infinitesimally close to \( x_{cor} \) so that \( t(N)x_e \to t(N)x_{cor} \) as \( \varepsilon \to 0 \). Taking the limit gives the rent ‘gap’ at the cordon:

\[ \lim_{\varepsilon \to 0^+} r(x_{cor}) - r(x_e) = 2 \frac{\tau_{cor}^2}{y^2} \left( yx_{cor} + t(N)x_{cor} - \tau_{cor} - 1 \right) + \frac{1}{2} \tau_{cor}^2 \]  

Since rent is decreasing in \( x \), disposable income is positive, and by (13) the ‘gap’ is guaranteed to be positive. Households at the cordon are exempt from the toll \( \tau_{cor} \). Thus they can bid for higher rent using this saving. It is clear that under laissez-faire, where both cordon parameters are set to zero, the rent-bid curve is smooth.

**Discussion and Conclusion**

In this study, a monocentric city model with a Cobb-Douglass utility function is used to analyze the effect of first-best congestion pricing and cordon pricing on rents across the city. I have derived results for a special case where housing and the numeraire are equally preferred (\( \alpha = \beta \)). Preliminary results show that congestion pricing increases land rents at the CBD. It will be useful to also evaluate land rents at the city fringe; (i.e. at \( x = B \)). Simulations by Verhoef (2005) shows that CBD rents rise while rents near the fringe fall, resulting in a steeper bid-rent curve. This finding is also consistent with Brueckner (2007) and Brueckner & Helsley (2011).

My analysis of the ‘rent-gap’ brought about by cordon pricing formalizes simulated findings by Verhoef (2005). This result can be expanded to the analysis of optimal cordon pricing. Furthermore, I have shown that bid-rent curve is affected by cordon parameters. In application, cordon parameters are planning decisions. This result shows that planners’ decision on cordon pricing determines the overall efficiency of urban transportation and also land use development.
References


