Pricing Congestion with Heterogeneous Agglomeration Externalities and Workers

Daniel J Graham and Kurt Van Dender[♠]

December 2008

Abstract

We consider an urban economy with two types of workers and two types of firms. The workers jointly use a congestion-prone network to commute to work, and the firms are subject to different degrees of agglomeration externalities. We show that, when no separate instruments are available to handle congestion and agglomeration effects, the optimal toll on commuting trades off congestion and agglomeration effects, so that it differs among workers. We use a numerical model to show the impact of constraints on tolls and on toll revenue redistribution schemes. The latter are shown to have a particularly strong downward impact on the welfare potential of tolling commuters.

JEL Codes: H23, R41, R48

[♠] Graham: Centre for Transport Studies, Imperial College London, London SW7 2AZ, d.j.graham@imperial.ac.uk, +44 20 7594 6088; Van Dender (corresponding author): Joint Transport Research Centre of the (ITF and the OECD), kurt.vandender@oecd.org , + 33 (0)1 45 24 95 98. We are grateful to Bruno de Borger, Robin Lindsey, and Ken Small for comments on an earlier version. We also benefited from feedback on a presentation at the Third International Conference on Funding Transport Infrastructure and 10th "Journée Transport", Paris School of Economics, June 2008

1. Introduction

Traffic in urbanized areas around the world is characterized by heavy congestion. Economic activity in those same areas is subject to positive agglomeration effects. Congestion pricing has the goal of internalizing marginal external congestion costs in order to avoid excessive congestion. This paper asks how that policy is amended when charges reflect the trade-off between congestion and agglomeration. In other words, we consider a situation where there are two externalities, but only one instrument to manage them. The consequence is that optimal charges deviate from marginal external costs or benefits. The analysis explicitly accounts for heterogeneity among workers and among firms. Workers, or commuters, earn different wages and this translates into different opportunity costs of time. Firms have different technologies and differ in the size of agglomeration effects. In first-best, the presence of heterogeneity in our model requires that charges are differentiated among workers in as far as agglomeration effects differ between them. Allowing for heterogeneity also is of interest in the design of second-best policies, since ignoring it may produce misleading results on the performance of such policies (see, e.g., Small and Yan, 2001).

Economists have defended congestion pricing as a key component of congestion management policy in urbanized areas for over a century, and just when the research community became increasingly pessimistic about the chances of widespread adoption of congestion pricing mechanisms (Arnott et al., 2005), the successful implementation of the London Congestion Charging scheme in February 2003 has increased interest in similar policies in many cities, including Stockholm (where it has been adopted) and New York (where it was debated but will not be adopted soon). Although a considerable amount of research has been done on second-best congestion tolling[1], it is probably fair to say that the conceptual background for these congestion-pricing initiatives is the basic Pigouvian analysis, according to which tolls should reflect marginal external congestion costs, even if existing schemes are only a crude approximation to those costs.

[1] Some of this work analyzes the impact of constraints on congestion pricing instruments on the level of the tolls (e.g. Verhoef, 2000). Others focus on the link between tolls and distortions in other sectors of the economy (e.g. Mayeres and Proost, 1997 and Parry and Bento, 2001).

This paper assesses the validity of the Pigouvian approach, focusing on the lessons from recent work on agglomeration economies. There is a good deal of empirical evidence that production in urbanized areas is subject to agglomeration externalities[2]. Furthermore, the strength of agglomeration externalities appears to vary across sectors. The implications of these results for the evaluation of investments in transport infrastructure have been explored (e.g. Graham, 2007a), but the presence of heterogeneous agglomeration effects also has implications for how to charge for congestion costs. This paper explores some of these implications.

We start from the basic insight that, if the realization of agglomeration benefits requires the simultaneous presence of commuters on the network (e.g. because the benefits depend on equal work start and end times), then marginal congestion costs need to be weighed against the marginal benefits of agglomeration. Consequently, if congestion tolls are the only policy instrument, focusing on marginal congestion costs alone implies excessively high toll levels. This point is made by Arnott (2007), who goes on to investigate the extent to which relaxing the strict simultaneity between agglomeration and congestion modifies recommendations for congestion charging. We retain the strict simultaneity, and focus on the impact of heterogeneity in agglomeration effects on congestion tolls. The analysis proceeds in two steps: an analytical framework and a numerical illustration.

First, we derive first-best congestion tolls in an analytical setting where two types of firms (sectors) are characterized by different agglomeration effects. In the simplest case, each sector employs a single type of worker, and worker types differ between sectors. While the workers travelling to these firms use the road simultaneously, their opportunity costs of time differ because of the wage differences between the sectors. We find that the first-best charges differ among workers, because agglomeration effects differ between the sectors that they work in. In this simple setting, where a sector employs only one type of worker, the policy problem arises only because both types of workers use the road simultaneously. Simply staggering work-hours by sector removes the need for differentiating tolls among workers that use the road at the same moment. Hence,

[2] For reviews of the empirical literature on agglomeration see Eberts and McMillen (1999), Rosenthal and Strange (2004) and Graham (2007a).

we also present toll expressions for a slightly more realistic production technology where each firm requires both types of workers as inputs. The input intensities and agglomeration externalities may differ among sectors. To the extent that productivity depends on the simultaneous presence of workers in a firm, staggering work-hours among firms or types of workers is not a perfect solution anymore, although it likely remains useful (Mun and Yonekawa, 2006). In this more general setting, realization of the first-best requires tolls that differ between workers according to their type and the sector they work in. Such differentiation is not easily obtained. Hence, we emphasize second-best solutions, by using a numerical model to illustrate the effects of restricting the extent to which tolls can be differentiated and of constraints on how toll revenues can be redistributed.

The second step is a numerical illustration of the more general version of the analytical model. The results show the extent of toll differentiation required for first-best, for a reasonable parameterization. In addition, they illustrate the impact of additional constraints, such as the inability to differentiate tolls between workers or firms and the absence of worker-specific lump sum transfers (which are required for first-best to be feasible). The results suggest that restrictions on toll differentiation and on mechanisms to redistribute toll revenues limit the welfare potential of tolls strongly. These findings are of some policy relevance, as restrictions of the type mentioned are common in practical charging mechanisms.

The paper is structured as follows. Section 2 provides some context on the theory of agglomeration economies and explains why heterogeneity in agglomeration externalities could be relevant for pricing. Section 3 presents the analytical framework. Section 4 explains the structure of the more general model with two labor inputs in each sector, and shows numerical illustrations. Section 5 contains concluding remarks.

2. Heterogeneous agglomeration externalities and pricing of commuting

This section discusses the connection between sources of agglomeration and the generalized cost of transport. It provides a background for the model introduced in Section 3, which focuses on the interaction between the cost of commuting, congestion and agglomeration.

The theoretical treatment of sources of agglomeration was traditionally guided by the distinction between industry concentration and urban concentration. Economies of industry concentration, termed localization economies, are external to the firm but internal to the industry and are principally thought to be sourced from labour market pooling, the sharing of intermediate inputs, and knowledge sharing or 'technological spillovers'. Economies of urban concentration, or urbanization economies, are external to the firm and the industry but internal to the city with benefits arising from the existence of local public goods, the scale of markets, the proximity of input-output sharing, and other kinds of inter-industry interaction (see for example Fujita and Thisse 2002).

While the localization / urbanization distinction may be reasonable, it need not imply any fundamental differences in underlying sources. For instance, Duranton and Puga (2004) propose a general three-way classification of 'mechanisms' of agglomeration which is consistent with the existence of both cities and industrial concentrations:

i. *Sharing* – positive externalities are generated by sharing public goods, inputs (gains from variety), the gains from specialization, and labour markets.
ii. *Matching* – economies accrue from the increased scale of agents operating in the labour market which improves the quality of matching and improves the chance of matching taking place.
iii. *Learning* – spatial concentration facilitate interactions allowing for the transfer of skills, ideas, information and knowledge.

The outcome of each mechanism is more or less equivalent being expressed in higher productivity and lower average costs.

This more generic approach to agglomeration can usefully be related to the analysis of transport infrastructure provision and pricing. We can conceive of transport infrastructure as essentially supporting agglomeration in the sense that it influences "effective density", i.e. the density of activity available to any location, by determining travel times (e.g. Venables 2007, Graham 2007a, 2007b). It is then clear that a change in the generalised cost of travel can alter effective densities. However, the generalised cost of travel differs among journeys[3] and this can have important implications in assessing the effects of transport policy on densities. Road pricing offers one example of this problem. We can define two components to the generalised cost of making some trip by road: the money price and the time cost, which is the value of time multiplied by the time taken to make the trip. The introduction of a single toll on congestion for all road users will increase the money cost for all journey types but will reduce congestion and so reduce travel times. Consequently, the travel time component will fall but the extent to which this compensates for the increase in money cost will vary among journeys. For instance, road users making business related trips may be better off following the imposition of the charge because they have very high value of time, while commuters, who have lower values of time may be made worse off in the sense that their generalised cost of travel has increased in absolute terms.

The classification of mechanisms of agglomeration proposed by Duranton and Puga (2004) is useful in thinking through the relationship between different journey types and agglomeration externalities. Their definition of *sharing* would be expressed in the demand for non-purpose specific general travel but also freight trips (input sharing) and commuting (sharing labour markets). *Matching* is essentially concerned with labour markets and so would be manifest in commuting trips. *Learning* comprises inter-firm interactions and so would involve work based trips. The key point is that different types of trip may make different contributions to agglomeration economies and it is interesting to explore how recognition of this heterogeneity would affect the problem of tolling congested roads.

This paper focuses on commuting trips, and emphasizes that commuters make use of a congestible public facility (the road network) that contributes to effective densities and hence to agglomeration economies. Workers travel to workplaces in different sectors, and these sectors differ in the intensity with which they use different types of workers (at least in the numerical model; the theoretical analysis abstracts from these differences). The sectors also differ in the size of agglomeration effects (Appendix A provides some empirical evidence on the extent to which agglomeration

[3] For example, DfT (2005) suggests a value of 7.2 €/h for a car commuting trip, 40 €/h for a car business trip, and 11 €/h for other car trips (2002 prices and values; 2008 exchange rate).

externalities may be heterogeneously distributed across workers). Our representation of agglomeration is generic (returns to scale at the level of the sector), and can be seen to derive mainly from the matching and sharing mechanisms defined above. The key issue then is that workers contribute equally to congestion when using the road, but their opportunity costs of time differ according to their type, and their contribution to agglomeration economies differs according to their type as well as according to the sector they work in. In a second-best environment, this heterogeneity affects how to charge for road use.

3. Analytical guidance

3.1 One type of worker per sector

We develop a framework that characterizes optimal congestion tolls in a setting where two types of workers simultaneously use a congestion-prone transport network ("the road") to commute to two types of firms (two sectors). Individual firms in each sector operate under constant returns to scale, but there are agglomeration effects at the level of the sector.[4] Firms belonging to a different sector differ in terms of average productivity and in terms of agglomeration benefits. The firms each produce one output which is consumed entirely by the workers of both sectors. Output prices are exogenous, and firms make zero profits. Consequently, the wage sum of each firm and sector equals the market value of its output, and this determines the sector's wage.

Since marginal external congestion costs are the same for each road user but agglomeration effects differ by sector or worker, and since there are no separate instruments for dealing with congestion and agglomeration, the optimal tolls differ by type of worker. We derive its precise structure. The framework is as simple as possible, as each firm employs only one type of worker. There are two separate sectors *i={1,2}* that only interact through the simultaneous use of roads by their workers. Section 2.2 introduces a more general production technology.

[4] The set-up allows for agglomeration economies and diseconomies, but we consider economies throughout the paper. Diseconomies have been reported in the literature (e.g. Elhorst et al., 2004). However, they occur because increased density generates high congestion costs, and congestion is separated out in the present model.

Set-up of the model

There are $N = N_1+N_2$ workers. The preferences of both types ($i=1,2$) are defined over both sectors' ($j=A,B$) output (y_{ij}, $i=1,2$, $j=A,B$) and over leisure (l_i). Workers face money and time budget constraints. Income consists of wages (w_i) earned per workday x_i of fixed length L_i and of a transfer T_i, and is spent on consumption of the two goods, of which the price is normalized to one, and on a toll τ_i. Since each workday x_i is taken to require a commute of travel time t, the toll is equivalent to a tax on wages. Available time is normalized to one (a full day) and is used for working plus commuting, and leisure. Working and commuting are strict complements, so the full length of a workday is equal to $(L_i + t)x_i)$. Each worker solves the following program to maximize utility:[5]

$$\Im = U_i\left[y_{i1}, y_{i2}, l_i\right] + \kappa_i\left(\left(w_i - \tau_i\right)x_i + T_i - y_{i1} - y_{i2}\right) + \mu_i\left(1 - \left(L_i + t\right)x_i - l_i\right) \qquad (1)$$

By combining the first-order conditions (see Appendix B) with respect to y_1, y_2, l_i, and x_i, we obtain the following expression for the opportunity cost of leisure time, or "the value of time":

$$\partial U_l \big/ \partial U_{yj} = \frac{w_i - \tau_i}{L_i + t}, i, j = 1,2\,. \qquad (2)$$

The value of time equals the real wage, that is the after tax wage per working day, where the latter includes commuting time (cf. e.g. De Borger and Van Dender, 2003, for similar expressions). The first-order conditions can also be used to derive the indirect utility function, which takes the following general form for $i=1,2$:

$$V_i \equiv V_i\left[w_i, \tau_i, T_i, t\right] \qquad (3)$$

[5] Square brackets denote functional dependence, and round brackets are used for algebra.

Workers take travel time *t* as parametric, but in fact it increases by a convex function of traffic volume *f*, which in turn is determined by the number of workers:

$$t = t[f] = t\left[\sum_i x_i N_i\right], t' > 0, t'' \geq 0\,. \tag{4}$$

The material feasibility constraints require that all consumption is actually produced. Each sector employs only one type of labor. Without loss of generality we assume that sector *A* employs type 1 workers and sector *B* employs type 2 workers. Hence,

$$F_j[x_i N_i] = \sum_i y_{ij} N_i, F_j' > 0; (i, j) = \{(1, A), (2, B)\} \tag{5}$$

Production increases as the single labor input increases. We allow for agglomeration economies (positive second derivative). While the model also accommodates diseconomies of agglomeration (negative second derivative), e.g. because fixed factors other than road capacity get excessively crowded, we do not consider them in the analysis.

Lastly, the government budget constraint stipulates that transfers must equal tax revenues:

$$\sum_i \tau_i N_i x_i = \sum_i T_i N_i \tag{6}$$

Note that this budget constraint allows for worker-specific lump sum transfers, a feature that greatly facilitates the design of welfare improving tax-and-transfer packages, as is explained next.

First-best: tolls for a Pareto-efficient allocation

Since the model has two types of workers, using a social welfare objective implies assigning relative welfare weights to both households. We use equal weights, so that the analysis focuses on the conditions for an efficient allocation.[6] The objective function is $W \equiv \sum_i N_i V_i[.]$ Note that, in a decentralized solution, the tax functions related to efficiency can be separated from those related to distributional objectives as long as type specific lump sum transfers are feasible, but that with restrictions on transfers all instruments contribute imperfectly to both functions (cf. e.g. Van Dender, 2004). In the theoretical analysis, we characterize optimal tolls that are conditional on type-specific transfers. Expressions for cases where there are restrictions on tolls or on transfers can be derived, but do not provide much insight. Instead, we rely on numerical illustrations to shed light on the impact of introducing constraints.

Denoting the multiplier for the government budget constraint by β, the program for maximizing social welfare is as follows:

$$\Im = \sum_i N_i V_i[.] + \beta\left(\sum_i \tau_i x_i N_i - \sum_i T_i N_i\right) \tag{7}$$

In taking derivatives of the indirect utility function, the social planner recognizes that because of the zero-profit conditions $w_i = {F_i[N_i x_i]}\big/{N_i x_i}$ for all *i*. This is in contrast to workers, who take wages as parametric. The first-order condition with respect to τ_i reads:

[6] Nevertheless, maximizing one type's utility while requiring some exogenous utility level for the other type would produce similar but not identical results. Differences arise because marginal utilities of income differ among consumers with different incomes and possibly different preferences. A utilitarian objective makes maximal use of consumers' different marginal contributions to social welfare, at least when consumer-specific lump sum transfers are available.

$$\frac{\partial \Im}{\partial \tau_i} = 0 = N_i\left(-\kappa_i x_i + \left(\kappa_i\left(F_i' - w_i\right) - \mu_i t' x_i N_i\right)\frac{\partial x_i}{\partial \tau_i}\right) - N_j \mu_j t' x_j N_i \frac{\partial x_i}{\partial \tau_i} + \beta\left(N_i x_i + \tau_i N_i \frac{\partial x_i}{\partial \tau_i}\right), i, j = 1,2, i \neq j \tag{8}$$

Dividing by $\beta N_i \frac{\partial x_i}{\partial \tau_i}$ and rearranging produces an implicit equation, for the optimal toll for type *i* (see Appendix B for the derivatives of the indirect utility function):

$$\tau_i = t'\left(\frac{\mu_i}{\beta} N_i x_i + \frac{\mu_j}{\beta} N_j x_j \right) - \frac{\kappa_i}{\beta}\left(F_i' - w_i\right) - \left(\frac{\beta - \kappa_i}{\beta}\right)\frac{x_i}{\partial x_i \big/ \partial \tau_{i;}} ; i, j = 1,2, i \neq j \tag{9}$$

The toll consists of three components. The first component reflects the marginal external congestion cost caused by a trip by type *i*. Congestion costs are incurred by both types of commuters. The marginal time impact is the same for both types, but their values of time – which are equal to the ratio of the multiplier of the time and money budget constraints – differs between types. Note that the congestion component of the optimal toll is anonymous, as it does not depend on consumer type. This is expected as a driver’s marginal impact on congestion costs does not depend on her type. The second component reduces the toll by the social value of the gap between marginal agglomeration effects from increased labor supply by type *i* and that type’s wage. That gap is larger than zero as long as average product wages are paid and there are agglomeration economies. The third component reflects the revenue raising function of the toll, and it depends on the difference between the marginal social and private value of income, and on the elasticity of labor supply (and commuting) to tolls (or wages).

The toll expression is simpler when optimal worker-specific lump sum transfers are available. In that situation, tolls perform no revenue-raising function, and the private and social valuations of income are the same. Hence:

$$\tau_i = t'\left(\frac{\mu_i}{\kappa_i}N_i x_i + \frac{\mu_j}{\kappa_j}N_j x_j\right) - \left(F_i' - w_i\right); i, j = 1,2, i \neq j \qquad (10)$$

The optimal toll for each type now consists of just two components. The first component reflects marginal external costs of congestion. As before, it is the sum of marginal time losses experienced by each type, valued at their specific opportunity cost of leisure time. The second component is the difference between marginal and average productivity in the sector that the worker is employed in, so it differs among types.

Decentralization of the efficient equilibrium requires congestion tolls that reflect differences in agglomeration externalities. The direction of the differentiation is clear, as long as group sizes are similar: workers in sectors with larger positive agglomeration externalities pay lower tolls. While such differentiation might be achieved by combining a uniform congestion charge with sector-specific income tax deductions, the administrative costs of such a system may be high. Therefore, it is useful to find out what a second-best uniform charge looks like. But because analytical expressions don't provide much insight, this is done through numerical analysis, in Section 3. Before turning to the numerical exercise, we briefly discuss the structure of the model where technologies in both sectors require both types of labor input.

3.2. Two types of worker per sector

We generalize the model by considering production functions where each type of labour is required in each sector. We allow wages to differ between sectors, but not between workers. A worker's labour income hence depends on how much labour time is supplied and on its distribution over the two sectors. For this model, the production functions, the input feasibility constraint and the wage sum constraint (as implied by the zero-profit condition and exogenous output prices equal to one) look as follows:

$$F_j\left[x_{1j}N_1, x_{2j}N_2\right] = \sum_j y_{ij}N_j, F_j' > 0; j = A, B, \qquad (11)$$

$$\sum_j x_{1j} = x_i; i = 1,2 \tag{12}$$

$$w_j \sum_i x_{ij} N_i = \sum_i y_{ij}; j = A, B. \tag{13}$$

Cost minimization (alternatively, profit-maximization subject to the zero-profit condition) on firms' behalf requires that each factor's marginal productivity equals the wage:

$$\partial F_j \big/ \partial x_{ij} = w_j; i = 1,2; j = A, B \tag{14}$$

Introducing this more general technology has some consequences worth mentioning. First, tolls now could be differentiated by workers according to the sector they work in. That is, in first-best a worker of type 1 employed in sector *A* pays a different toll from a type 1 worker in sector *B*. This is because workers' contribution to agglomeration effects differs between sectors even if they are of the same type. Charging the same toll to all workers in the same sector or to all workers of the same type is a second-best solution. Here, we only consider the case where the toll depends on the type of worker but not on the sector. Furthermore, we restrict attention to the case where both groups of workers are of the same size (N_1=N_2). Under these assumptions, the structure of the optimal toll (conditional on optimal transfers) is entirely analogous to that of the simpler model. The only difference is that agglomeration effects are sector-specific, and their contribution to the optimal toll depends on how large labour supply responses to toll changes are in both sectors. Specifically:

$$\tau_i = t'\left(\frac{\mu_i}{\kappa_i} N_i x_i + \frac{\mu_j}{\kappa_j} N_j x_j\right) - \sum_{f=A,B} \left(F_{i,f}' - w_f\right) \frac{\partial x_{i,f} \big/ \partial \tau_i}{\partial x_i \big/ \partial \tau_i}; i, j = 1,2, i \neq j \tag{15}$$

4. Numerical illustration

The structure of the numerical model is identical to the model of section 3.2. We assume throughout that the groups are of equal size, and normalise group size to one. The utility and production functions take a Cobb-Douglas form and the congestion function is of the Bureau of Public Roads (BPR) type. Sector *A* is characterized by lower agglomeration economies than sector *B*. The parameters of these functions are found by calibrating to a consumer equilibrium, which implies production requirements and traffic volumes. Using Cobb-Douglas functions implies a range of well-known restrictions, but we note that the econometric specification for estimating the production functions is of the same type. The functions are as follows:

$$U_i\left[y_{i1}, y_{i2}, l_i\right] \equiv y_{i1}^{\alpha} y_{i2}^{\beta} l_i^{\gamma}, i = 1,2; \alpha + \beta + \gamma = 1 \tag{16}$$

$$F_j\left[x_{1j}N_1, x_{2j}N_2\right] = M_j\left(\left(x_{1j}N_1\right)^{s_j}\left(x_{2j}N_2\right)^{1-s_j}\right)^{\delta_j}; j = A, B, \delta_A = 1.05, \delta_B = 1.20 \tag{17}$$

$$t = t[f] = freeflow + a(flow)^b \tag{18}$$

Commuting is assumed to take an hour per workday in the reference equilibrium, and occurs at a speed equal to half the free-flow travel speed on the network.[7] The marginal external travel time is twice the average travel time.

Wages and production function parameters are found through combining the consumer equilibrium and the constraint that the wage in each sector is the same for both types. In addition, the inputs of each type of worker in each sector in the reference equilibrium are given. By construction, sector *A* makes relatively intensive use of workers of type 1, and sector *B* is relatively intensive in type 2 workers.[8] Sector *A* has lower agglomeration externalities than sector *B* (the values for δ are the same).

The reference equilibrium, to which the model is calibrated, is summarized in the following table.

[7] In the units of the model, this means $t = 0.03125 + 0.004768 * (1.6)^4 = 0.0625$.

[8] The reverse situation was analyzed as well, but results are similar and are omitted for reasons of brevity.

Table 1 **Reference equilibrium and parameters for the two-factor model**

Consumers			Producers		
	1	2		A	B
Parameters					
α	0.118	0.127	M	290.012	484.312
β	0.212	0.200	S	0.700	0.206
γ	0.670	0.673	1-S	0.300	0.794
L	0.333	0.333			
Variables					
INC x	140.000	180.000	Y	120.000	200.000
TRAx	8.000	8.000	xA	0.600	0.200
YA	50.000	70.000	xB	0.200	0.600
YB	90.000	110.000			
l	0.683	0.683			
t	0.062	0.062			
x	0.800	0.800			
wA	175.000	225.000			
wB	175.000	225.000			
U	3.185	3.392			
W	6.577	6.577			
MECT trip	0.125	0.125			
MECC trip	60.076	60.076			
TOLL	10.000	10.000			
MVOT day	417.369	543.845			
MVOT h (16h day)	26.086	33.990			
κ	0.007	0.006			
μ	3.125	3.343			

As can be seen, the two types of consumers have similar but not identical preferences. Type 2 enjoys a higher income, as its wage is equal in both sectors and exceeds that of type 1. The two types work equal amounts (so they also consume equal amounts of leisure time), pay the same toll, and receive equal shares of the toll revenues. Both types spend a larger share of their money budget on good A than on good B, and type 2 spends relatively (and absolutely) more on good A. The values of time differ between the types, with type 2 showing a higher value of about 34€/h, against about 26€/h for type 1. On the production side, the two sectors are each others' mirror images in terms of factor intensities, with sector A intensive in type 1 and sector B intensive in type 2.

The calibrated model is used to evaluate the impact on welfare, which is the sum of consumer groups' utilities, of different tolling and revenue redistribution rules. For revenue redistribution, we distinguish two situations: one where each consumer group receives an equal share of the toll revenues, and one where revenue is redistributed optimally. Optimal redistribution is required for attaining first-best, but as will be seen it also strongly improves the performance of non-optimal tolling rules. For tolling, we consider the following rules in addition to the reference situations: a differentiated toll[9], a toll that is uniform across sectors and across firms ("single toll"), a Pigouvian toll (toll equal to the marginal external congestion cost, ignoring agglomeration effects), and a zero toll. Table 2 summarises the welfare effects.

Table 2 Percentage welfare change from alternative toll and revenue redistribution scenarios

	Revenue redistribution:	
Toll:	Equal	Optimal
Reference	0.000	
Differentiated toll	0.321	0.643
Single toll	0.108	0.636
Pigouvian toll	0.039	0.561
Zero toll	-0.168	-0.168

Table 2 shows that simple tolling schemes, including a single toll and even a Pigouvian toll, perform well in the sense that they rival first-best tolls at least when optimal transfers are available. The reason is that these tolls generate sufficient revenue to allow welfare-improving revenue redistribution schemes to be implemented. In this example, optimal revenue redistribution always means that type 1 receives all the revenue. A zero toll generates no revenue, and this reduces welfare compared to the reference equilibrium. This can be seen as an extreme case of a toll that generates low revenues, and does not allow welfare-improving revenue redistribution. Given the strong welfare improvements from optimal revenue redistribution, it is not surprising that constraining revenue use strongly limits welfare. The welfare potential of a fully differentiated toll is cut by half, the

[9] Differentiation among firms and workers is obviously first-best, but differentiation along only one of these dimensions is sufficient as well, because the transfers can be adapted to accommodate uniformity of the toll along the remaining dimension.

potential of a single toll is lower by a factor of about six, and that of a Pigouvian toll falls to nearly zero.

This is not to say that a single toll with equal revenue sharing is useless, as the reference situation generates higher welfare than the zero-toll equilibrium. Note that welfare in the reference equilibrium is not very different from welfare at the Pigouvian toll under equal revenue sharing. This, however, should not be taken as support for the view that marginal external congestion costs are a good guideline for setting tolls. Instead, it indicates that the redistribution of toll revenues on a 50/50 basis goes some way in the direction of optimal revenue redistribution. Combining the Pigouvian toll with optimal redistribution performs much better still (and implies that all revenues are allocated to type 1).

Of course the model is just an example, and different results may be expected for alternative calibrations or functional forms. But the general insight that the welfare potential of tolling mechanisms depends at least as much on what is done with the revenue than on the toll itself is quite general. The social value of transfers differs between consumers, as long as their incomes and preferences differ. And of course a value judgment is implied: here we assumed that each consumer has the same welfare weight, but alternative judgments are possible. Taking account of other distortions, for example labour tax distortions that can be mitigated to some extent by tolls, affects the social value of revenue as well.

5. Concluding remarks

We have developed a simple model of an urban economy in which two sectors employ two types of workers. Firms in each sector are homogenous and are price-takers on the output market. The sectors differ in the intensity with which they employ the two types of workers as well as in terms of agglomeration economies. Workers commute to the firms on a congestion-prone network.

We show that a toll on commuting needs to be differentiated according to worker-sector combinations, because marginal agglomeration benefits differ among such combinations. When such differentiation is not possible, the effectiveness of tolls as instruments to internalize congestion and agglomeration effects, declines.

However, as long as worker-specific lump sum transfers are possible, tolls perform quite well. When there are constraints on transfers, the welfare potential of tolls is much lower.

The analysis shows how the presence of heterogeneous agglomeration benefits affects the case for congestion tolls. The representation of agglomeration effects is very simple, as they are captured by increasing returns to scale at the level of a sector. We have treated agglomeration in relation to industrial densities per se, and have not adopted the traditional distinction between urbanisation and localization economies. To some extent, our model exhibits characteristics akin to localisation in that it draws sharp boundaries between sectors and allows for increasing returns at the level of the sector. However, the basic intuition of the model, which relates charging simultaneously to agglomeration and congestion externalities, is equally valid for either class of externalities, particularly since these tend to have equivalent outcomes. A useful extension of our model could attempt to model urbanisation and localisation explicitly by allowing for endogenous population size.

Appendix A: Empirical estimates of agglomeration economies

In our model the contribution that workers make to agglomeration economies differs depending on the industry and job type that they participate in. In this appendix, we provide some empirical evidence to test the extent to which agglomeration externalities may be heterogeneously distributed across workers from different industries.

Theory tells us that agglomeration economies shift the productivity of firms which in turn has consequences for the wages that firms pay to workers. Let,

$$Y_{it} = \left[\delta K_{it}^{\theta\varsigma} + (1-\delta) L_{it}^{\theta\varsigma}\right]^{1/\varsigma} U_{it}^{\beta_U} e^{(\eta_t + f_i + \nu_{it})}, \qquad \text{(A1)}$$

be the production function for the *i*th firm (*i = 1,...,N*) producing output *Y* at time *t* (*t = 1,...,T*). The firm uses a production technology which is homogenous of degree θ in labour (*L*) and capital (*K*) inputs and is located in an environment with a level of agglomeration measured by *U*. The input shares are determined by δ and the degree of substitution between them by ζ. The term η_t is a time specific effect that allows for unobserved shocks which are common across firms, f_i represents unobserved individual firm level time-invariant heterogeneity, and ν_{it} is an error term.

Differentiating the production function with respect to labour gives a Cobb Douglas type relationship of the form

$$\frac{\partial Y_{it}}{\partial L_{it}} = Y_{it}^{\,1-\varsigma} L_{it}^{\,(-1+\theta\varsigma)} U_{it}^{\,\beta_U}\,\theta(1-\delta)e^{(\eta_t + f_i + \nu_{it})}. \qquad \text{(A2)}$$

Assuming correspondence between the wages that firms pay (*w*) and the marginal product of labour and taking logs we have

$$\log w_{it} = \beta_Y \log Y_{it} + \beta_L \log L_{it} + \beta_U \log U_{it} + \eta_t + \varphi_i + \varepsilon_{it} \ , \qquad \text{(A3)}$$

where β_Y = (1 - ζ), β_L = (-1 + $\theta\zeta$), ψ_i is a firm level individual effect that subsumes the term log [θ(1-δ)], and ε_{it} is an error term which contains the 'transmitted error' ν_{it} from the production function model but also other sources of measurement error and misspecification associated with the dependent variable w_{it}. . The elasticity of substitution (γ) of the original production function can be derived from the wage equation as γ= - (1/β_L), and an estimate of economies of scale can also be recovered as θ= -[(β_L + 1)/ζ] = (β_L + 1)/(1-β_Y). Note that the inclusion of firm level individual effects, in addition to representing unobserved heterogeneity, also allows some deviation from the marginal productivity rule.

It is unlikely that the wage rate will adjust instantaneously to changes in productivity or agglomeration and so it useful to allow for a period of adjustment. We introduce dynamics by specifying a potentially autoregressive productivity shock in the error term, $\varepsilon_{it} = \rho\, \varepsilon_{it-1} + \phi_{it}$, with ρ<1 and ϕ_{it} ~ IID (0, σ^2) representing serially uncorrelated white noise error. Thus, we can rewrite () as an ADL(1,1) dynamic model

$$\log w_{it} = \rho \log w_{it-1} + \beta_Y \log Y_{it} - \rho\beta_Y \log Y_{it-1} + \beta_L \log L_{it} - \rho\beta_L \log L_{it-1} + \beta_U \log U_{it} \\ - \rho\beta_U \log U_{it-1} + (\eta_t - \rho\eta_{t-1}) + \varphi_i (1-\rho) + \phi_{it} \qquad . \tag{A4}$$

Equation (A4) cannot be estimated by ordinary least square because there are sources of endogeneity that affect all right hand side variables arising from correlation with the individual effects and with the error term. Following Blundell and bond (2000) we apply a Generalized Method of Moments (GMM) estimator for dynamic panel data which exploits the time series nature of the data to derive instruments that are correlated with the endogenous regressors but orthogonal to the errors. We use the system GMM estimator which differences equation (A4) to remove the individual effects and uses contemporaneous values of the regressors as instruments for the differenced equation, but also specifies an equation in levels with lagged differences of the endogenous variables as instruments[10]. Equation (A4) is estimated with non-linear common factor restrictions which are tested and imposed using minimum distance to obtain the restricted parameter vector (ρ, β_Y, β_L, β_U).

Our empirical representation of the production model uses data on registered UK companies. Under UK legislation each registered company is required to provide accounting and other data about their operations to an executive agency of the Department of Trade and Industry know as Companies House. These data are made available in a commercial software package called Financial Analysis Made Easy (FAME), which is produced jointly by Jordans and Bureau Van Dijk (BVD 2003). The production data relate to companies, some of which have plants in a number of different locations. For our analysis it is necessary that the productivity measures

[10] A full description of the system GMM estimator for dynamic panel data is given in Arellano and Bover (1995), Blundell and Bond (2000) and Bond (2002). An application in the context of an ADL(1,1) model is given in Blundell and Bond (2000).

relate to production at one location. It is, however, possible to identify and remove multi-plant firms from the sample because they report more than one trading address.

The FAME data are available for a number of years, although the reporting for individual firms is irregular. We have derived an unbalanced panel of firms over 9 years from 1996 to 2004. Sales (turnover) is used as a proxy for output and we have information on the number of employees and on average wages. We estimate productivity separately for seven industry groups: manufacturing, construction, transport storage and communications, financial intermediation, real estate, business services, and public services.

The FAME data record the full postcode information of each firm in the sample. To construct measures of the agglomeration `experienced' by each firm we use employment data from the Annual Business Inquiry for each of the 11,344 postcode sectors (PCS) in Britain. Our measure of agglomeration captures the accessibility to economic activity at each location as

$$U_i = \frac{E_i}{d_i} + \sum_j \frac{E_j}{d_{ij}}, \qquad \text{(A5)}$$

where E_i is employment in PCS i, d_{ij} is the Euclidean distance between the centroids of PCS i and j, and d_i is an approximation to the internal distance of PCS i.

The model given above allows us to estimate the effect of density, or accessibility, on wages across diverse industries. It does not seek to distinguish localization or urbanization economies or to analyze sources of agglomeration economies. It is designed simply to test the extent to which agglomeration externalities may be heterogeneously distributed within the urban economy. The results are shown in table A1 below.

Table A1: Dynamic reduced form wage equation estimates.

Variable	manufacturing	construction	trans, storage & comm..	financial intermediation	real estate	business services	public services
W (t-1)	0.493 ***	0.473 ***	0.423 ***	0.515 ***	0.446 ***	0.405 ***	0.543 ***
	(0.031)	(0.035)	(0.041)	(0.043)	(0.049)	(0.039)	(0.049)
Y (t)	0.601 ***	0.371 ***	0.294 ***	0.609 ***	0.451 ***	0.405 ***	0.480 ***
	(0.043)	(0.059)	(0.063)	(0.060)	(0.071)	(0.049)	(0.086)
Y (t-1)	- 0.382 ***	- 0.096 ***	- 0.198 ***	- 0.300 ***	- 0.124 ***	- 0.224 ***	- 0.280 ***
	(0.033)	(0.036)	(0.044)	(0.041)	(0.033)	(0.035)	(0.053)
L (t)	- 0.595 ***	- 0.387 ***	- 0.292 ***	- 0.563 ***	- 0.395 ***	- 0.406 ***	- 0.509 ***
	(0.066)	(0.080)	(0.078)	(0.097)	(0.116)	(0.054)	(0.090)
L (t-1)	0.380 ***	0.191 ***	0.200 ***	0.287 ***	0.141	0.223 ***	0.363 ***
	(0.056)	(0.053)	(0.061)	(0.077)	(0.091)	(0.051)	(0.063)
U (t)	- 0.292	0.352	0.036	- 0.712	- 0.165	- 0.322	- 0.672
	(0.198)	(0.684)	(0.385)	(0.474)	(0.951)	(0.376)	(0.651)
U (t-1)	0.343 *	- 0.278	0.098	0.852	0.252	0.472	0.749
	(0.199)	(0.673)	(0.381)	(0.474)	(0.949)	(0.374)	(0.663)
AR1	0.000	0.000	0.000	0.000	0.000	0.000	0.000
AR2	0.963	0.072	0.851	0.055	0.049	0.059	0.600
N	45,087	13,621	9,221	14,063	8,956	11,116	5,147
ρ	0.577 ***	0.451 ***	0.425 ***	0.502 ***	0.387 ***	0.413 ***	0.522 ***
	(0.026)	(0.034)	(0.040)	(0.040)	(0.044)	(0.038)	(0.047)
β_Y	0.699 ***	0.433 ***	0.201 ***	0.603 ***	0.443 ***	0.398 ***	0.593 ***
	(0.037)	(0.057)	(0.056)	(0.056)	(0.071)	(0.049)	(0.070)
β_L	- 0.704 ***	- 0.292 ***	- 0.192 ***	- 0.525 ***	- 0.335 ***	- 0.371 ***	- 0.543 ***
	(0.047)	(0.073)	(0.061)	(0.068)	(0.091)	(0.041)	(0.082)
β_U	0.061 ***	0.153 ***	0.230 ***	0.304 ***	0.208 ***	0.253 ***	0.055
	(0.020)	(0.043)	(0.037)	(0.036)	(0.048)	(0.022)	(0.055)
Comfac	0.000	0.000	0.015	0.148	0.054	0.005	0.115

Notes: Numbers in parentheses are robust standard errors. Estimates are based on two-step GMM with small sample corrections. AR1 and AR2 are the Arrelano and Bond tests for first-order and second-order serial autocorrelation. Comfac gives the significance level of the minimum distance test for the common factor restrictions.

The AR1 and AR2 tests shown in the table are important because the existence of serial autocorrelation invalidates the use of instruments derived according to the system GMM estimator. With the exception of the real estate industry, the test statistics reject the existence of serial autocorrelation in the levels equations. Another important diagnostic statistic is the minimum distance test for the common factor restrictions, which is significant at or above the 10% level in all cases.

Our results on agglomeration economies show positive and significant effects on wages for six of the seven industries shown in the table. We also find substantial variance in the magnitude of this effect across our industry groups. For manufacturing the elasticity is 0.06, which compares well with previous estimates of 0.06 for the US states (Ciccone and Hall 1996) and 0.045 for EU regions (Ciccone 2002). For services, however, we find much stronger associations between density and wages. For financial intermediation we estimate an agglomeration elasticity of 0.30, for business services 0.25, for transport storage and communications 0.23, and for real estate 0.22. Thus, the magnitude of the agglomeration elasticity can take substantially different values for different workers in the urban economy, with those engaged in industries that tend to occupy CBD and central city locations showing the largest agglomeration effects.

Appendix B Analytical detail

First-order conditions for utility-maximization (equation (1))

The first-order conditions with respect to y_{ij}, l_i, and x_i are, respectively:

$$\frac{\partial U_i}{\partial y_{ij}} = \kappa_i ; \frac{\partial U_i}{\partial l_i} = \mu_i ; \kappa_i \left(w_i - \tau_i \right) = \mu_i \left(L_i + t \right)$$

Derivatives of the indirect utility function

For example,

$$\frac{\partial V_1}{\partial \tau_1} = \kappa_1\left(\frac{\partial F_A}{\partial x_1}\frac{\partial x_1}{\partial \tau_1} - x_1 - \tau_1\frac{\partial x_1}{\partial \tau_1}\right) - \mu_1\left((L_1 + t)\frac{\partial x_1}{\partial \tau_1} + N_1\frac{\partial t_1}{\partial x_1}\frac{\partial x_1}{\partial \tau_1}x_1\right)$$

Adding $(w_1 - w_1)\frac{\partial x_1}{\partial \tau_1}$ to the first term and using the first-order condition with respect to *x_1* leads to:

$$\frac{\partial V_1}{\partial \tau_1} = -\kappa_1 x_1 + \left(\kappa_1\left(\frac{\partial F_A}{\partial x_1} - w_1\right) - \mu_1 N_1 t' x_1\right)\frac{\partial x_1}{\partial \tau_1}.$$

Analogously,

$$\frac{\partial V_1}{\partial \tau_2} = -\mu_1 N_2 t' x_1 \frac{\partial x_2}{\partial \tau_2}.$$

References

Arellano Manual and Bover Olympia (1995) Another look at the instrumental variable estimator of error-components models, Journal of Econometrics, 68, 29-52.

Arnott, Richard, 2007, Congestion Tolling with Agglomeration Externalities, *Journal of Urban Economics*, 62, 187-203

Arnott, Richard, Tillman Rave and Ronnie Schöb, 2005, *Alleviating Urban Traffic Congestion*, MIT Press

Blundell Richard W and Bond Stephen R (2000) GMM Estimation with persistent panel data: an application to production functions, *Econometric Reviews*, 19, 32 1-340.

Bond Stephen R (2002) Dynamic panel data models: a guide to micro data methods and practice, *Portuguese Economic Journal*, 1, 141-162.

BVD (2003) *FAME: UK and Irish company information in an instant*, Bureau van Dijk, London.

Ciccone Antonio (2002) Agglomeration effects in Europe*, European Economic Review* 46 (2) 213-227.

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

De Borger, Bruno and Kurt Van Dender, 2003, Transport tax reform, commuting and endogenous values of time, *Journal of Urban Economics*, 53, 510-530

DfT(2005) Values of time and operating costs, London: DfT.

Duranton G and Puga D (2004) 'Microfoundations of urban agglomeration economies' in Henderson JV and Thisse JF (eds) Handbook of Regional and Urban Economics, Volume 4, Amsterdam: Elsevier.

Eberts, R. and D. McMillen (1999) Agglomeration economies and urban public infrastructure, Chapter in HP Cheshire and E S Mills (eds) Handbook of regional and urban economics, Volume III. New York: North Holland

Elhorst, J.P., J. Oosterhaven and W.E. Rom, 2004, *Integral cost-benefit analysis of Maglev technology under market imperfections*. SOM Report 04C22, University of Groningen (forthcoming in Journal of Transportation and Land-Use).

Fujita M and Thisse J (2002) *The economics of agglomeration: Cities, industrial location and regional growth*. Cambridge University Press: Cambridge.

Graham, Daniel J., 2007a, Agglomeration, Productivity and Transport Investment, *Journal of Transport Economics and Policy*, 41, 3, 317-343

Graham, Daniel J., 2007b, *Agglomeration economies and transport investment*, JTRC Discussion Paper 2007-11 (http://www.internationaltransportforum.org/jtrc/DiscussionPapers/jtrcpapers.html)

Mayeres, Inge and Stef Proost, 1997. Optimal tax and public investment rules for congestion type of externalities, *Scandinavian Journal of Economics*, 99, 2, 261-279

Mun, Se-Il and Makoto Yonekawa, 2006, Flex time, traffic congestion and urban productivity, *Journal of Transport Economics and Policy*, 40, 329-358

Parry, Ian W H and Antonio Bento, 2001, Revenue recycling and the welfare effects of road pricing, *Scandinavian Journal of Economics*, 103, 4, 645-671

Rosenthal, Stuart and William Strange (2004) Evidence on the nature and sources of agglomeration economies, Chapter in Henderson JV and Thisse JF (eds) Handbook of Regional and Urban Economics, Volume 4. Amsterdam: Elsevier

Small, Kenneth A. and Jia Yan, 2001, The value of "value pricing" of roads: second-best pricing and product differentiation, *Journal of Urban Economics*, 49, 2, 310-336

Van Dender, Kurt, 2004, Pricing transport networks with fixed residential location, *Regional Science and Urban Economics*, 34, 3, 289-307

Venables, Anthony J. (2007) Evaluating urban transport improvements: cost-benefit analysis in the presence of agglomeration and income taxation, *Journal of Transport Economics and Policy* 41, 173–188.

Verhoef, Erik T., 2000, The implementation of marginal external cost pricing in road transport – Long run vs short run and first-best vs second-best, *Papers in Regional Science*, 79, 3, 307-332