Fourth Symposium in Tax and Economic Growth
"Urban Growth and Finance"
School of Public Policy, University of Calgary
October 9-10, 2013

EFFICIENCY OF RESOURCE ALLOCATION IN A METROPOLITAN AREA: AN URBAN ECONOMIST'S PERSPECTIVE

Richard Arnott*

University of California, Riverside

November 3, 2013

*I would like to thank Juan Carlos Lopez for creative, energetic, able research assistance.

This is the first of a pair of companion papers. The companion paper, "Reflections on Metropolitan Calgary's Spatial Structure", is available as School of Public Policy discussion paper xxxxxx, and is referred to in this paper as "the companion paper."

Efficiency in Resource Allocation in a Metropolitan Area: An Urban Economic Perspective

Modern microeconomic policy analysis is organized around the First Theorem of Welfare Economics[1]. The Theorem lays down a long list of conditions under which "the market" achieves an efficient allocation. Those who favor the free market and oppose government intervention on ideological grounds are apt to overlook the long list of conditions under which the Theorem applies, and claim that the Theorem proves the efficiency of the free market. Those who are leery of markets and well disposed towards government stress that Real World economies come nowhere close to satisfying the long list of conditions, and therefore claim that the Theorem provides the basis for extensive government intervention.

When I was a student, the standard compromise middle ground between these two positions was the theory of classic market failures[2]. The market fails (results in inefficiency) in three main ways -- classic externalities, public goods, and increasing returns to scale in production. Government intervention is potentially justified to deal with all of them. If they are corrected according to the classical prescriptions, then the free market will generate an efficient outcome. Since markets have many virtues that central planning does not, then the ideal economic system (at least from the point of view of efficiency) is organized primarily through markets but with the government intervening to correct the classic failures. Equity, so the argument went, is efficiently dealt with through lump-sum redistribution. Thus, policy to achieve efficiency can be separated from policy to achieve equity.

Since then, there have been two principal assaults on the classic market failure theory of the appropriate role of government in a mixed economy. The first stemmed from the "new, new welfare economics" of the new public economics, which grew out of optimal tax theory, and emphasizes asymmetrical information. One point made was that informational asymmetries preclude efficient redistribution, so that there is a fundamental

---

[1] In order of increasing sophistication, the standard technical textbooks on microeconomic theory are Varian (1978), Varian (1987), and Mas-Colell, Whinston, and Green (1995). Bator (1957) provides a very good calculus-based rather than set-based presentation of the Theorem, and Bator (1958) a good discussion of classic theory of market failure. Salanié (2000) provides an excellent survey of the theory of market failure, including both classic and non-classic market failures.

[2] Economists use the word "distortion" to apply to any source of inefficiency, whether it is due to a classic market failure, a non-classic market failure (such as informational externalities), or a non-market failure.

equity-efficiency tradeoff[3]. Some sacrifice of efficiency is justified to improve equity. This aspect of the new, new welfare economics has been falling out of favor, with many now arguing that income taxation, though distortionary, is the best instrument to achieve equity, and that otherwise efficiency should be pursued[4]. The second assault has come from public choice theory. Just because an ideal government can improve efficiency doesn't mean that an actual government would, since government intervention entails its own waste. Today's compromise middle ground, which I shall stand on, is the theory of classic market failures, taking these cautions into account.

I shall consider each of the three classic market failures, starting with public goods, in the context of public policy for the metropolitan economy. But before doing so, a digression on the theory of the second best is appropriate.

The theory of the first best concerns the optimal allocation of resources in an economy when the only constraints are resource constraints and technological constraints. The theory of the second best concerns the optimal allocation of resources in an economy when there are constraints in addition to resource and technological constraints. To illustrate, consider the following example taken from urban economics. A fixed number of commuters travel from A to B. They have a choice between traveling by car along a congested highway or by bus along an uncongested bus corridor that is separated from the highway. The first best entails pricing both modes at marginal social cost. Suppose, however, that the commuters have voted not to implement congestion tolling on the highway, so that auto travel is priced below marginal social cost, with the result that an inefficiently large share of commuters travel by car. There is then a political constraint in addition to the resource and technological constraints. Second-best policy entails subsidizing bus travel so that the modal share is efficient.

In reality, there are always distortions that have not been efficiently corrected. All should in principle be taken into account in choosing policy. However, there are so many of them, and they are so difficult to measure with any degree of accuracy, that in most contexts economists ignore them, and prescribe policy according to first-best rules. Interestingly, in his work on transportation economics, William Vickrey consistently advocated first-best pricing, even though he wrote one of the original papers in the theory of the second best in the late 1950's. Sometimes however there is a particular distortion

---

[3] The basic idea was originally developed in the context of income taxation. The government can observe an individual's taxable income, but neither how many hours she worked nor her ability. The government would like to redistribute from the able to the less able, but not being able to observe ability must instead tax income. Individuals have an incentive to reduce their number of hours worked -- increase their leisure -- in order to reduce their income, so as to appear less able. Thus, the income tax distorts the labor-leisure decision. The idea was formalized by William Vickrey (1947) and the optimal income tax problem was solved by James Mirrlees (1971) -- the two winners of the 1996 Nobel Prize in Economics.

[4] This point of view is clearly articulated, and its application widely discussed, in Kaplow (2008)

that is so egregious and so quantitatively important that it is considered in policy analysis. Two of the most familiar are industrial pollution and auto congestion. In what follows, I shall follow the received wisdom in presenting the first-best policy rules, except when there is a compelling argument for second-best considerations to be introduced.

Local Public Goods

Almost all the examples of pure public goods that I can think of are global or national in scale. The body of knowledge is a pure global public good[5]. My consuming it does not diminish anyone else's consumption of it. Culture too is a pure global public good. A country's body of laws and institutions is a pure national public good, as is a country's defense system. At a local level, the set of local regulations and local institutions are a pure public good, but all major categories of local government services, which include mass transit, the local road system, local public education, and local crime prevention and law enforcement, are congestible public goods, in which the quality of service degrades with the intensity of use. Public utilities too can be regarded as being congestible public goods.

The first-best theory of congestible public goods/congestible facilities, is well developed, and has clear rules for efficient resource allocation[6]. The first rule is that pricing should be at short-run marginal social cost. The second rule is that capacity should be such that the marginal social benefit of capacity equals the marginal social cost. The marginal social benefit is the reduction in (discounted) user costs, holding fixed the level of utilization at the optimal level, while the marginal social cost is just the marginal expenditure on capacity. The third rule is that, if application of the first two rules results in operation at a deficit, the deficit incurred should be financed out of via lump-sum taxation[7].

- *pricing at short-run marginal social cost*

While the rule is straightforward, its practical application in the urban context entails complications. In my opinion, William Vickrey's 1963 paper, "General and specific financing or urban services" provides the best point of entry into the literature[8]. For specific categories of local public services, he shows what short-run marginal cost pricing

---

[5] It is not however completely non-excludable. Totalitarian regimes can exclude their citizens from access to certain areas of knowledge.

[6] James Buchanan (1965) developed the theory of private congestible goods in his theory of clubs. Herbert Mohring (1976) provides an excellent diagrammatic presentation of the theory in the context of transportation.

[7] This is the appropriate rule when the marginal dollar of general revenue is raised via a lump-sum tax. When it is not, second-best adjustments may need to be made to the optimal pricing and capacity rules.

[8] Vickrey, W., 1963. “General and specific financing of urban services,” in *Public Expenditure Decisions in the Urban Community* edited by H.G. Schaller, Resources for the Future, 62-90.

would entail. For fire protection, for example, each house would be rated according to its fire danger, proximity to other homes, and distance from the fire station, and houses with a higher rating would be charged more. This would provide households with the appropriate incentives to improve the fire safety of their homes, and would also get them to face the added burden they impose on society from living further away from a fire station. My favorite of his many examples is charging the police for the traffic congestion they cause when giving out tickets. That may seem counter-intuitive, indeed a bit wacky, but would give the police the incentive to ticket drivers in ways that would cause less congestion. My purpose in citing these two examples is to illustrate that application of the first rule should be, and indeed is, tempered by practical considerations and common sense. Rating houses according to fire danger might entail greater cost than the incentive benefits from doing so, and charging police for the congestion they cause would be impractical. What I take away from Vickrey's paper is that the ideal should always be short-run marginal cost pricing, but with deviations from that rule being based on common sense.

Public utility pricing too should be at short-run marginal social cost. Public utilities have indeed been employing increasingly sophisticated pricing, but often following the wrong principles. Let me give you a couple of examples from my own experience paying utilities in Riverside, California. The principle of peak-load pricing is used for electricity, with the peak rate being applied during the peak period of the day and the off-peak rate during the rest. That's fine, as far as it goes, but does not go as far as it could or should. For example, it provides me with no incentive to be especially economical during the hottest days of the summer when electricity demand is at its peak. Given current technology, *responsive peak-load electricity pricing*, where this minute's rate is based on this minute's load/capacity ratio, is feasible[9]. Water in Riverside is priced according to a three-rate system, where the marginal rate is based on water consumption per household member. This may be effective in conserving water, but is contrary to the principle of marginal social cost pricing. Water is a necessity, so there is an argument for providing the necessary amount free. But why should I pay more than the cost to society of a marginal gallon of water for my garden, particularly since my nice garden benefits the neighborhood?

Another aspect of pricing at social marginal cost is the use of fees. Fees as a proportion of local public revenue have been rising, at least in the United States. According to the principles outlined above, a move away from the general financing of public services towards their specific financing is good thing if fees are set at or close to marginal social cost. But the actual pattern of fees seems to bear little correlation to the corresponding marginal social costs. On one hand, most of the fees that are actually applied appear to be simple tax (revenue) grabs. It is useful for all dogs to have dog tags, so that they can be traced to their owners if they are running wild or get injured or lost. But it is hard to see how this service can justify Riverside's annual cost of $100. In fact, the main effect of

[9] For a pricing scheme to have the desired effect, consumers must be aware of it. Under responsive pricing, customers would therefore need to be informed of the current aggregate usage.

the excessive dog license fee in Riverside is that pit bull owners (who are disproportionately low socio-economic class) fail to license their dogs. Minor traffic violations in Riverside carry a $500 fine. How this can be justified except as a form of random lump-sum taxation I don't know? On the other hand, there is considerable scope for the justified application of specific fees. The most obvious example is curbside parking in downtown areas. Its social cost derives from the increased traffic congestion it causes. Another is charging homeowners for part of the cost of maintaining the public trees abutting their properties, since they enjoy the private benefits from it in the form of beautification and reduced air conditioning bills.

I have been only a lukewarm supporter of congestion pricing, except in the most congested downtown areas, on the grounds that it would be more hassle than it's worth. Life is too complicated as it is. The same argument applies to many fees. What I am suggesting is not that we subject ourselves to a battery of fees, but rather that we use marginal social cost pricing as the starting point of the policy discussion on specific financing. If we feel that hassle of marginal social cost pricing some service does not justify the efficiency gain, so be it.

- *the degree of cost recovery*

There is an important theorem which states that "A congestible facility of optimal size that prices at short-run marginal social cost and that operates under constant long-run average cost generates exactly enough revenue, in discounted terms, to finance the facility's construction and operating costs".[10] There are also corollaries that relate the degree of cost recovery to how rapidly long-run average cost rises or falls with output.

This theorem has considerable scope for application. Probably the least controversial is to public utilities. The theorem indicates that the proportion of a public utility's costs that come from revenue should depend on the shape of its long-run average cost function, which depends primarily on technology. The deficit should be financed out of general revenue.

More controversial are the theorem's implications for the pricing and privatization of urban freeways. Current estimates are that urban freeways have approximately constant long-run average cost .[11] Thus, if travel on an urban freeway[12] is priced at short-run marginal cost (which translates into setting the toll equal to the marginal congestion externality) and if its capacity is chosen optimally, it will be fully self-financing. This implies that the general taxpayer should not have to cover any of the cost of urban

---

[10] Mohring, H., Harwitz, M. 1962. *Highway Benefits: An Analytical Framework.* Northwestern University Press, Evanston.

[11] See Kraus, M., 1981. Scale economies analysis for urban highway networks. *Journal of Urban Economics* 9, 1-22.; Small, K., Verhoef, E. 2007. *The Economics of Urban Transportation*. Routledge, New York.

[12] A freeway that is priced is no longer free. Perhaps instead of "freeway", I should say "limited access highway".

freeways, since, if priced and operated efficiently, they should cover their costs. In most countries, most freeways were tolled from the day they were opened. Since users have regarded them as a higher-priced but also higher-quality *addition* to the existing road network, which provides free travel, they have been accepted. But tolling a freeway that previously provided travel for free is bound to encounter political opposition. Los Angeles and other jurisdictions have dealt with this in three different ways. The first is "value pricing", which entails making one or more lanes "HOT" (high occupancy and tolling) lanes. HOT lanes have been marketed politically as providing a priced but higher-quality service than the other lanes[13]. The second is to impose tolling on all added capacity, whether new freeways or expansions of existing freeways. The third is to lease out rights of way for private toll roads, with restrictions on the tolls charged[14].

One of the virtues of first-best congestion pricing is that, with constant long-run average costs, it provides the right signals for capacity expansion. Define the economic surplus generated by a congestible facility per unit time as the revenue it generates from first-best pricing minus its amortized capital costs (opportunity cost of the capital plus maintenance cost plus depreciation), operating and maintenance costs, and depreciation If this surplus is spent to expand capacity, either through improvements to the existing facility or through expansion, the facility will expand its capacity at the efficient rate. The intuition is that if demand is higher than anticipated, the facility will be more congested than anticipated, so that first-best congestion pricing will raise more revenue than anticipated, and the excess revenue can and should be allocated to expansion of the facility.

Even more controversial, at least in North America, would be proposals to move in the direction of social marginal cost pricing for mass transit. The reason is that mass transit is used disproportionately by the poor, especially in the United States. One popular view, which was mentioned earlier, is that mass transit travel should not be priced below social marginal cost since income taxation provides the efficient way to redistribute. Counter to this view is the argument that travel by mass transit may supply at least as good a measure of need as documented income.

School voucher programs provide a way of moving towards the marginal cost pricing of education. Give all parents a voucher for each child in school equal to the cost of providing a student with a solid education, and let schools compete for students. Higher-

---

[13] I am surprised that the public has accepted value pricing since its implementation rather obviously increases congestion in the untolled lanes. Perhaps drivers appreciate having the *option* of being able to travel at a higher speed when needed.

[14] The lease contract typically contains clauses stating that the government commit to constructing no directly competing freeways, to ensure that the lessee will receive adequate revenue. The difficulty has been that this commitment is not credible since the government is unlikely to resist pressure to build parallel capacity itself if the private toll road turns out to be heavily congested. Thus, the lessee is faced with the prospect of losing money if demand turns out to be lower than anticipated, and losing money if demand turns out to be higher expected and the government constructs parallel, untolled capacity.

quality schools would charge a higher price, reflecting their higher costs, and children (via their parents' decisions) would sort across schools on the basis of their parents' willingness to pay for education. Limited programs along these lines have been successful in the United States in revitalizing underperforming school districts, but Canadians attach such a high premium to social equity and equality of opportunity that school voucher programs are unlikely to be politically popular in Canada.

The issue then arises as to how metropolitan public facility deficits should be financed. I am going to side step much of this issue, and defer to those paper presenters who are experts on the Canadian fiscal system. One issue is the extent to which the provincial government should redistribute from richer to poorer regions within the province; another is the division of taxing powers between levels of government; yet another is inter-jurisdictional spillover of benefits, which apply with particular force to education. Urban economic theory provides the basis for an unfamiliar answer. Absent the inter-jurisdictional spillover of benefits, public facility deficits should be financed by land taxation[15].

There is however one aspect of the issue that I would like to comment on from the unfamiliar perspective of an urban economist. How should the public infrastructure of suburban expansion be financed? There are five general approaches. The first is financing out of general current revenue; the second is bond financing at the city-wide level; the third is charging suburban developers development impact fees; the fourth is charging the current and future property owners of new suburban areas a special assessment; and the fifth is charging user fees. These five approaches are differentiated along two dimensions, time and space. How should the cost be borne across generations? Should the cost be borne by those who are directly affected by new public infrastructure or by the community at large? The efficiency answer is to apply the benefit principle, since doing so provides the appropriate marginal incentives. Competitive developers will develop at the efficient time and density if they face the full social cost of their development. Prospective new homebuyers will make efficient choices with respect to when to buy and what size of home to buy if they face the full social costs of their choice. And later residents will face the right incentives concerning adding capacity at a later date if they face the social cost of doing so. One way of achieving this goal is to have the developers pay the full cost associated with their development; they in turn will pass on these costs

---

[15] This is an implication of the Henry George Theorem, which states that, in a city of optimal population size and with efficient capacity of its congestible facilities, aggregate land rents equal the deficit incurred from the social marginal cost pricing of services that are produced under increasing returns to scale, which include local public goods and most utilities (e.g., Vickrey, 1977). The broad intuition is that the city is operating at a point of minimum average cost. Under marginal cost pricing, this implies zero aggregate profit, with the deficit incurred from increasing returns to scale activities, the congestible public facilities, exactly equaling the profit derived from the decreasing returns to scale activity -- the production of lots. One imagines that the rent a firm charges for a lot includes the cost of transportation to the lot. Thus, the larger the city population, the more lots are produced, and the higher the marginal production cost -- decreasing returns to scale.

to prospective new homebuyers, etc. Another way of achieving this goal is through an annual local assessment on homes in the neighborhood that is set so equal to that year's benefits. Yet another way is to charge congestion fees for use of the infrastructure.

These are the first-best mechanisms. The mechanism should be adjusted to the extent that the new development generates externalities that are not internalized. For example, if the marginal immigrant to the city imposes a net burden, a possible second-best policy would be to set the development impact fee above the direct cost of the associated infrastructure since doing so would discourage in-migration.

- *preference revelation for local public goods*

One of the central problems facing governments everywhere is how to determine their constituents' preferences for local public goods and services. There is no ideal method. A variety of public good preference revelation mechanisms have been proposed but all are unrealistic in their assumptions concerning the information available to the government and/or residents, and concerning agents' computational sophistication[16]. Local referenda are employed in many jurisdictions in the United States, and generally seem to work well since the issue being voted on is generally clearly stated --- a stated expenditure for a stated purpose with a stated method of finance. For reasons that I do not understand, though they are presumably related to inter-governmental law, referenda are not as common in Canada as in the United States. Information concerning the demand for various public services gained from the study of referenda outcomes in the United States can nonetheless be usefully applied in the Canadian context. In Canada, the pre-eminent preference revelation method is representative democracy. While economists have been consistently at the forefront in developing theories of collective choice (social choice theory, public choice theory, the new political economy, etc.), applied research has been more the domain of political scientists. In any event, while the preference revelation problem is one of the central problems in the efficient management of metropolitan growth, I shall comment on only those aspects of the topic on which I have some expertise, the Tiebout mechanism and congestible public facilities.[17]

The celebrated Tiebout mechanism for preference revelation for local public goods entails "voting with the feet". Different jurisdictions offer different tax/public service packages, and the individual resident chooses to live in that jurisdiction whose tax/public service matches best with her tastes. The Tiebout mechanism has been intensively studied in the context of US metro areas, the larger of which have several hundred different cities and towns within the metro area. In Calgary, however, with its metropolitan government, it would seem to be irrelevant.

---

[16] Hindriks and Myles (2006), an up-to-date intermediate public economics textbook, provides a good introduction to the relevant literature.

[17] Tiebout, C., 1956. “A pure theory of local expenditures,” *Journal of Political Economy* 64, 416-424.

I would however like to return to a topic that I discussed in an earlier subsection --- congestible public facilities. There I looked at the topic from the perspective of cost recovery. Here instead I look at it from the perspective of preference revelation.

There is an important class of public services for which individuals make decisions concerning quantity/frequency of usage. Individuals decide how much electricity to consume and how frequently to use various city streets and highways. Since an individual chooses to use such a service up to the point where the benefit from her last unit of service equals the price, her decisions concerning quantity/frequency of usage provide information about her marginal willingness to pay. This revealed preference can assist in the determination of the appropriate level of the service to provide.

Let us return to the case of a congestible public facility whose provision is characterized by constant long-run average cost. If marginal social cost pricing is employed and if optimal capacity is provided, the revenue generated is just sufficient to cover the cost of constructing, maintaining, and operating the facility. A corollary of the theorem is that if marginal cost pricing is employed, and if less (more) than optimal capacity is provided, then it raises more than enough (does not raise sufficient) revenue to finance capacity. The intuition is simple. If the facility is too small, it will be heavily congested, so that marginal social cost pricing will raise considerable revenue, more than enough to cover the facility's costs. Thus, how much revenue is raised under marginal cost pricing relative to the cost of the facility indicates whether the facility's capacity is inefficiently high or low. This statement of the Self-Financing Theorem is based on a static or steady-state analysis. A dynamic analysis generates rules that are useful in practical application. Consider two polar cases, one in which, once built, the facility's capacity cannot be altered, the other in which, the facility's capacity can be continuously expanded at constant cost. To simplify, ignore operating costs.

In the first case, in which the capacity decision is once and forever, the simple statement of the Theorem does not indicate when the facility should be built or how it should be financed, since obviously it will at that time have generated no revenue. The timing rule is "Build when the social benefit of postponing construction of optimal capacity one period equals the social cost". The social benefit equals the amortized cost of constructing the facility, while the social cost is the cost of not having the facility available for use during the period. Optimal capacity is such that the discounted revenue generated under marginal social cost pricing equals the cost of constructing capacity. In this case, the preference revelation problem is not solved since the decision on capacity has to be made before users have registered their preferences in terms of usage. An example would be the construction of a bridge at the single possible location, when subsequent expansion would be prohibitively costly.

In the second case in which capacity can be continuously expanded, the rule is "Debt finance the facility, using all the surplus that is generated after debt servicing, maintenance, and operating costs have been paid to finance capacity expansion." Note that, under this rule, users in each period pay for the flow cost of the facility during that period, so that capital costs are allocated across generations in proportion to the flow

costs incurred. Furthermore, at each point in times, through their usage of existing capacity, the level of usage (and hence the revenue received from the congestible facility) provides the information needed to determine the optimal rate of capacity expansion Similar rules can be derived for intermediate cases, in which capacity expansion occurs at discrete intervals.

Now consider the class of congestible public facilities that are characterized by decreasing long-run average cost, which include public utilities. Users indicate their preferences for public services through their individual demands. Here however, with social marginal cost pricing and optimal capacity, revenues fall short of costs, and some proportion of costs should be financed out of general revenue. Modified versions of the rules for the constant cost case apply. Suppose, for example, on the basis of engineering studies, that it is ascertained that, under marginal social cost pricing, the efficient cost-recovery rate for a particular congestible public facility is 70%. Then with continuously divisible capacity, the efficient rule for capacity expansion is to take the surplus generated after 70% of the amortized cost of the facility's capacity, plus operating and maintenance costs, have been paid for out of facility revenue to finance capacity expansion.

According to this line of reasoning, determination of the efficient degree of cost recovery of various types of congestible facilities, which depends on the degree of economies of scale, is of central importance.

Externalities

An externality occurs when the action of one economic agent affects the well being (utility in the case of a consumer, profit in the case of a firm) of another agent. Externalities are pervasive. A pecuniary externality operates through prices. A technological externality operates directly rather than through the price system. In the traditional theory of market failure, there are two general rules relating to the efficient treatment of classic externalities.

Rule 1: Pecuniary externalities do not matter.

Rule 2: The inefficiency associated with a technological externality is efficiently treated through internalization of the externality, more specifically by charging the generator of the externality for the cost of the externality he imposes on other economic agents.

Pecuniary externalities create transfers but do not generate inefficiency. For example, suppose that I consume a chocolate bar. Doing so shifts up the aggregate demand for chocolate bars, pushing equilibrium up the supply curve for chocolate bars, and therefore increasing the equilibrium price. This increase in price increases producer surplus and decreases consumer surplus in the same amounts, causing no change in aggregate surplus.

Rule 2 is the standard Pigouvian prescription for the internalization of a (technological) externality. Consider a polluting firm that generates effluent (smoke, hazardous waste, etc.) each kg of which does $6 worth of damage. In the base situation, the firm does not pay for the effluent it generates, and so generates too much of it. If however it is charged an effluent fee of $6 per kg, then it faces the social cost of the effluent it produces, and so produces the efficient amount. The Pigouvian prescription works well when the damage generated by the externality depends only on the generator's actions.

In the metropolitan context, there are six important classes of externalities, congestion externalities, agglomeration externalities, migration externalities, environmental externalities, land use externalities, and fiscal externalities. I shall discuss each in turn.

- *congestion externalities*

I have already discussed congestion externalities in the context of congestible public facilities. Here I discuss them from a somewhat different perspective. Most congestion externalities are *aggregate externalities*, depending only on the total number of people users of the facility (e.g., the density of cars on the road, or the number of passengers on a bus). The congestion created is anonymous and atomistic. In deciding on frequency of use, a user treats the level of congestion of exogenous. For aggregate externalities, the Pigouvian prescription is an efficient remedy, and the familiar analysis is applicable. There are, however, congestion externalities that are not aggregate, and for which the Pigouvian remedy is inefficient. One example that has been quite intensively studied recently is congestion in taking off and landing at airports where there is one or more major carrier that has market power. The major carriers internalize the congestion cost they impose on themselves and so should be charged only for the congestion cost they impose on other carriers.

In the context of metropolitan traffic, the standard Pigouvian prescription is an efficient remedy, and entails the application of a congestion toll. With the application of an efficient congestion toll, drivers have an incentive to make efficient decisions on all margins of choice for which the principle of anonymity applies. Thus, they make efficient choices with respect to modal choice (if the other modes are efficiently priced), trip frequency, trip timing, and route choice. They do not however make efficient choices with respect to driving behavior, since a careless or aggressive driver is charged the same amount as a safe driver, except to the extent that careless or aggressive driving leads to efficient fines or to the efficient increase in insurance rates in the event of accidents.

Urban transportation economists have long advocated congestion tolling as the principal remedy to urban auto traffic congestion. Its implementation has been very slow, however, primarily due to political opposition. Drivers dislike being charged for what previously provided free, they distrust governments, suspecting that tolling is simply another economically unjustified user fee. Furthermore, according to standard theory, the average driver is made worse off unless the congestion toll revenue is spent in ways that benefits her. In addition, since congestion tolling entails paying more in money for a trip

though less in time, it disproportionately benefits those with high values of time, most of whom are rich. The well-known London congestion-pricing scheme was successful primarily because it was voted on by London residents but applied primarily to non-London residents.

Since Calgary essentially has a metropolitan government, since the majority of households live outside the downtown area and commute downtown to work by car, and since Calgarians tend to favor small government, there would likely be considerable opposition to almost any form of congestion pricing. That pretty much rules out implementation of first-best policy, and means that road capacity, mass transit pricing, and mass transit capacity choices must all be second best.

There are however two partial steps towards congestion tolling that are worth exploring in Calgary. Both, which were mentioned before, involve forms of "value pricing" in which tolling part of the road network is introduced as a paid upgrade to the free service provided by regular roads. One is to convert some freeway lanes to HOT lanes (high-occupancy and tolling lanes). There is considerable experience with such lanes in Los Angeles. The other is to permit the private construction and operation of new freeways on government rights-of-way (or perhaps land assembled by the government through eminent domain). Both forms of value pricing have proved to be more popular than was anticipated, perhaps because most people would occasionally exercise the option of paying a supplement for faster travel.

I shall leave more specific discussion of transportation policy in Calgary to the companion paper.

Congestion externalities occur in all types of congestible facilities, which include education and health care. Crime may also be considered a form of congestion externality since the crime rate is typically higher in larger metropolitan areas and hence a function of total (aggregate) population. At a more abstract level, population density may be considered to be a congestion externality, though people have different opinions about whether it is positive or negative[18].

- *agglomeration externalities*

The pattern of economic activity over space is determined by the interaction between centripetal forces, which cause the spatial concentration of economic activity, and centrifugal forces, which cause the spatial dispersion of economic activity. Even in the absence of traffic congestion, the principal centrifugal force is transportation costs, since an increase in city size increases average transportation costs. There are a variety of centripetal forces --- increasing returns to scale in the provision of public services (the extreme case being a pure public good, such as a public monument, whose *per capita*

[18] Almost everyone would like their back door to open up to a bucolic park and their front door to open up to a vibrant neighborhood with lots of street life, and a wide variety of dining and recreational activities.

cost is inversely proportional to population) and utilities, increased variety in products, neighborhoods, cultural activities, etc., and increased productivity. The term agglomeration economies is used to describe the set of centripetal forces, and the term agglomeration externalities to describe aspects of the set of centripetal forces that are not "mediated" by markets. For example, increasing returns to scale *internal* to a firm affect other firms through that firm's influence on product and factor prices, but do not generate uninternalized externalities. In contrast, increasing returns to scale *external* to a firm affects other firms not only through this market channel but also by directly affecting their productivity.

Agglomeration economies must be very important since they offset the significantly higher transportation costs in large compared to small cities, but they have proved very difficult to measure directly. Considerable work has gone into developing the microeconomic theory of agglomeration economies[19], measuring agglomeration economies at the aggregate level, and attempting to infer the relative importance of alternative sources of agglomeration economies from empirical regularities.

I shall not attempt a comprehensive review of the literature[20], but shall rather focus on a few facets that seem particularly relevant to this essay. Economists have described many different possible sources of agglomeration economies. Probably the most familiar is that identified in Adam Smith's statement that "The division of labor is limited by the extent of the market" and illustrated by Smith's example of a pin factory.[21] Paul Krugman identified the home market effect, that the average real (relative to the equilibrium wage rate) consumer price of commodities is lower, the larger the size of a city, since a higher proportion of commodities are produced locally and hence can be delivered to the consumer at lower cost.[22] Alfred Marshall focused on knowledge externalities within industries.[23] Jane Jacobs emphasized cross-industry fertilization of ideas and cross-industry exchange of knowledge.[24] Others have emphasized the importance of having a group of specialized agents meet face to face to make a deal.

Both the theoretical and empirical literatures have focused on external economies of scale in production. Most of the literature has considered the situation in which agglomerative

---

[19] Fujita and Thisse, 2002, being the outstanding contribution.

[20] I recommend three points of entry into the literature: the North-Holland Handbook edited by Jacques Thisse and Vernon Henderson (2004) which provides a broad overview of the theoretical and empirical literatures; Thisse and Fujita's (2002) magisterial overview of the theory; and Fujita, Krugman, and Venables (2001), which considers many extensions of Krugman's core-periphery model.

[21] Smith, A., 1937. *An Inquiry into the Nature and Causes of the Wealth of Nations*. New York: Modern Library.

[22] Krugman, P., 1980. "Scale Economies, Product Differentiation, and the Pattern of Trade," *American Economic Review* 70: 950-959.

[23] Marshall, A. 1890. *Principles of Economics*. MacMillan, London.

[24] Jacobs, J., 1961. *The Death and Life of Great American Cities*. New York: Vintage.

economies of scale enter the individual firm's production function as a multiplicative productivity parameter; for example, for any combination of inputs, a firm at one location might produce, say, 28% more output than a firm at another location. The multiplicative productivity parameter is called the *location potential*. The location potential of a firm in industry i in city j might depend on the value of city j's output, the quantity of output in industry i in city j, a distance-weighted measure of accessibility to other workers in industry i in city j, and so on. Empirical work focuses on estimating the location potential function, and various properties of the estimated location potential function suggest the relative importance of alternative sources of agglomeration economies. Productivity is typically measured by labor productivity (value of output per worker). Some studies examine how labor productivity within a particular industry differs across cities. Other studies examine how the labor productivity within a particular industry differs within a city. Some disaggregate the analysis according to the education level of workers. Earlier studies assumed that agglomeration externalities are specific to an industry, while some more recent studies have admitted cross-industry agglomeration externalities.

Several empirical regularities stand out.

1. The location potential is systematically higher in larger cities.
2. The elasticity of city-industry labor productivity with respect to city-industry employment, measured across cities, differs systematically across industries. Those industries with lower (higher) elasticities locate disproportionately in smaller (larger) cities.
3. Within a city, the elasticity of labor productivity with respect of accessibility to other workers in the same industry differs systematically across industries. Those industries for which this elasticity is higher (lower) locate at more (less) accessible (typically more central) locations. This elasticity is exceptionally high for corporate headquarters, the FIRE industries (finance, insurance, and real estate) and for advertising, and is low for standardized services.
4. The empirical estimates tend to conform to intuition concerning which sources of agglomeration economies are relatively more important in different industries. For example, intuitively the fashion industry tends to be very concentrated spatially because of the importance of being up to date on recent fashion trends and of having access to highly specialized labor available on a job basis.

At the aggregate level, the magnitude of agglomeration externalities can be estimated, although with a considerable margin of error; on a finer spatial scale, they are not measurable with any reasonable degree of accuracy. For these reasons, it is impractical to tax them. As we shall see, this inability to internalize a very important class of externalities makes accommodating metropolitan growth efficiently considerably more difficult that it otherwise would be.

- *migration externalities*

Migration policy is an important aspect of efficient growth. National governments everywhere (and one provincial government in Canada!) decide not only on the level but also the composition of immigration. Though provincial and local governments have

fewer policy instruments at their disposal than do national governments to influence immigration, migration policy is nonetheless important. Does Calgary wish to continue growing rapidly, which entails considerable immigration, or to slow down its growth?

I know that these issues have been extensively and expertly discussed within the local policy community, so shall keep my comments brief.

A larger population affects wages, rents, land and housing prices, and the cost of living in predictable ways. These changes in prices result in some types of households being better off and others worse off. Thus, migration policy entails distributional conflict that is resolved at the ballot box. I shall abstract from these considerations, and suppose instead either that residents are identical or that there is a common goal such as the maximization of surplus.

An immigrant imposes costs on existing residents and also confers benefits on them. The short-run (long-run) costs and benefits are those that occur before (after) the metropolitan area has had a chance to adapt through the addition of infrastructure and new housing. The major short-run costs entail congestion/crowding in both public services and housing, while the major benefits come through agglomeration economies. But there are also "fiscal externalities" -- does the marginal migrant increase government expenditures by more or less than she contributes in taxes? If the additional expense of providing her with services exceeds her tax payment, existing residents bear the burden of the difference. Long-run costs include, in addition, the cost of the infrastructure that is added to accommodate a higher population, which includes both building new infrastructure and expanding existing infrastructure. Long-run benefits too arise primarily from agglomeration economies but may be different from short-run benefits. On one hand, more recent in-migrants are more likely to inject new ideas; on the other, less recent immigrants are likely to be better integrated within the city's production structure.

Much of the earlier literature overlooked the agglomeration benefits of immigrants, which partially accounts for why there used to be such a strong anti-urban bias in rural-urban migration policy in most developing countries.

The efficient rate of immigration is that for which the *social* cost of the marginal in-migrant equals the *social* benefit. Since immigration occurs to the point where the *private* cost of immigration to the marginal migrant equals her *private* benefit, public policy should tax/subsidize and regulate immigration such that the private benefit and cost of in-migration are equalized at the efficient rate of immigration.

In the Calgary context, the most obvious potential source of migration inefficiency is that Alberta residents share Alberta resource rents. An immigrant essentially receives a gift of part of these resource rents, which amounts to an immigration subsidy. These resource rents are most obviously manifest in the Alberta Trust Fund, but also they allow the Alberta to impose the lowest tax rates in Canada, including having no provincial sales tax.

Provincial and local governments have limited policy instruments that permit discrimination between long-term residents and recent immigrants. The most obvious and perhaps the only instruments available are residency requirements for the provision of certain public services. This has been a huge policy issue in the states along the US-Mexico border.

I have discussed migration externalities from the provincial perspective. It should be recognized however that migrants to Alberta are migrants from elsewhere. When a resident migrates from another jurisdiction to Alberta she not only imposes an externality on Albertans but also removes an externality from her former jurisdiction. This point is particularly important with respect to provision of social services to the disadvantaged. If all provinces race to the bottom in raising the residency requirements for the provision of social services, the equilibrium is underprovision of social services to the disadvantaged, which is evident in the United States. The federal government should use the policy instruments at its disposal to avoid this adverse outcome.

- *environmental externalities*

The standard Pigouvian prescription for the treatment of environmental externalities is to charge firms for the social cost of the effluent they produce. This applies to global warming, as well as to more localized externalities. Thus, social responsibility mandates that Alberta tax CO_2 emissions at an appropriate rate.

I know little about Calgary's environmental problems. The most important appear to relate to water usage. Broadly speaking, pricing at marginal social cost is the best way to ensure consumption of water at the socially efficient level. Other types of policies, such as the proper use of cost-benefit analysis, are needed to ensure the efficient transmission and distribution of water. If the government wishes to subsidize agriculture, as most governments do, it should do so more directly, such as allowing postponement of property taxes on agricultural land until its sale for urban development, than through subsidizing agricultural water usage.

- *land use externalities*

Most externalities are dealt with efficiently through choosing taxes and setting their rates so that the generator of the externality pays as closely as possible for the marginal damage done by his economic activity. Land use externalities are an exception. One reason is that most are highly localized, the effects of most extending for a block or less, so that a mosaic of block-specific taxes would be required. Another is that measuring most would be more costly than the benefit gained from the taxing them. Another is that they are intrinsically spatial. An economic agent decides not only what level of pollutants or nuisances to generate but also where to locate. A tax rate on the level of pollutants that is not location specific does not provide appropriate incentives concerning location. A final reason is that most land use externalities are reciprocal in nature, with the magnitude of the damage affected by not only the decisions of the generators of the externalities, but also decisions of the recipients; that is, most land use externalities are Coasean in nature. To take one of the classic Coasean examples, the magnitude of the externality created by a soot-producing factory depends not only on the factory's

activities but also on what land uses choose to locate close to the factory. It is economically inefficient for a laundry that hangs its washing out to dry to be located near the soot generating factory. A laundry's locating near the factor entails "coming to the nuisance".[25]

Throughout most of the 20th century, in North American cities land use externalities were dealt with through "Euclidean" zoning (named after a US court case, *Euclid vs Ambler*, declaring zoning to the constitutional, rather than after the geometer), which entailed the rigid separation of land uses. Land use planning designated separate areas for industrial, commercial, and residential land uses, with further restrictions on each related to structural density, setbacks, coverage ratios, etc. Another type of zoning is hierarchical zoning. Residences can locate anywhere; commercial establishments can locate in industrial zones and commercial zones but not in residential zones; and industrial firms can locate only in industrial areas. As times change and as cities grow, any zoned pattern of land use becomes inefficient. The Canadian economy has moved from heavy industry to light industry and services; technological improvements in goods transportation (such as the development of the intercity highway system and of the assembly line) have caused the efficient location of most heavy industry to move away from the city center to locations near a circumferential highway; technological improvements in auto design and manufacture have encouraged residential suburbanization, while the move towards single-member households has increased the demand for living downtown. The rigid pattern of land uses set down by the City fathers in the late 1920's has been changed by the accumulation of zoning variances and by adjustments to the official land use plan. Nevertheless, especially over the last thirty or forty years, with the revitalization of downtown and the new urbanist movements, there has been a trend away from what is now called Euclidean I zoning towards Euclidean II zoning, which allows and even encourages more mixed land use.

If "all politics is local", then "all land use politics is very local". Landowners will continue to push for zoning variances that make their land more valuable and to be opposed by existing neighboring land uses, especially resident households, who perceive that the proposed zoning variance will compromise their enjoyment of their properties and lower their property values. Furthermore, any large-scale changes in land use regulation will generate large capital gains and losses in property values, and so be vociferously opposed by the losers. Thus, it is prudent to set up a land use policy structure that results in gradual change in the land use system through a zoning variance process that is responsive to local political pressure.

That said, there are aspects of local land use policy that affect the spatial structure of the entire metropolitan area and that should therefore be considered at the metropolitan level. Three effects, I think, are particularly important: the overall level of housing prices, the availability of affordable housing for poorer households, and the implications of residential density regulations on the viability of mass transit.

---

[25] Coase, R., 1960. "The Problem of Social Cost," *Journal of Law and Economics* 3:1-44.

The jurisdictional fragmentation of US metropolitan areas has the benefit of providing households with a larger menu of choices in their tax/public service mixes than in Canadian metropolitan areas. But there have also been costs. One has been the use by suburban jurisdictions of minimum lot size regulation to zone out the poor, or at least those households that would impose a substantial negative fiscal externality on their neighbors[26]. The incentive to employ exclusionary minimum lot size zoning is stronger, the more important is property taxation as a source of local revenue. Minimum lot size zoning has reduced the availability of affordable housing for poorer households in the suburbs, and, by reducing suburban density, reduced the viability of mass transit to suburban areas, both of which have contributed to the ghettoization of poor families in central cities. Another cost has been an increase in spatial segregation by income.

Another aspect of local land use policy that has affected the entire metropolitan area has been the imposition of maximum density regulations in downtown neighborhoods, often in the name of maintaining their character. This policy reduces the "effective" (adjusted for location) supply of housing, which has driven up housing prices and increased average commuting distances.

Some metropolitan areas apply land use restrictions at the metropolitan level that are designed to modify the spatial structure that would be generated by market *cum* local land use regulation. The most obvious are green belts and urban growth boundaries. Their benefits are the preservation of agricultural land and densification, which increases the viability of mass transit to the suburbs. Their primary cost is the increase in housing prices they induce. The empirical evidence suggests that they favor the rich and hurt the poor.

Calgary does not have jurisdictional fragmentation, nor has it applied density regulations in downtown neighborhoods that have substantially reduced residential densities there, not has it applied green belt or urban growth boundary policies, at least with any aggressiveness. That is not to say that its land use at the metropolitan level is efficient. As in all North American cities, urban auto travel has been underpriced for many years[27]. In Calgary, this has led to inefficiently low residential densities and suburban sprawl, and perhaps the rigorous application of a greenbelt or urban growth boundary, or the imposition of minimum density regulations would improve efficiency.

---

[26] The word "class" seems taboo in North American policy discussion. Yet the incentive to zone out the *hoi polloi*, in order that neighbors behave well and that children go to school with other well-behaved children, may be at least as strong as the fiscal motive. O'Sullivan (2009), the leading urban economics textbook, contains very good chapters on zoning and residential segregation.

[27] The underpricing of urban auto travel was less extreme in Canada than in the US, which probably accounts for why most Canadian cities are less sprawled than their US counterparts.

The final aspect of land use policy that I shall discuss is its effect on the location of employment, which I shall also discuss at greater length in the companion paper. The discussion on this point is more speculative than most of the previous discussion since the economic theory of metropolitan land use with non-monocentric cities has been little explored.

In the quarter century following World War II, there was rapid residential decentralization in both Canada and the US, due primarily to the expansion of car ownership[28]. In the subsequent quarter century, the major change in metropolitan spatial structure was the decentralization of employment. About a decade ago, Glaeser and Kahn published a Brookings Institute study that documented the recent history of the decentralization of residences and employment in major US metropolitan areas.[29] They found that in 1990, the median distance of a resident from the CBD is 8 miles, while the median distance of a job from the CBD is 7 miles. Since then both jobs and residences have continued to decentralize, and there has also been a marked increase in "reverse commuting" (commuting from home in the central city to a job in the suburbs). At the same time, in almost all US cities, due to new urbanist advocacy, there has been a push to reduce sprawl[30], increase mass transit modal share, and densify the metropolitan area. The economics literature has focused on the effects of land use policies on urban spatial structure using the monocentric city model. I shall argue in the companion paper that the subsidization of auto travel has likely led to inefficiently centralized urban employment, but what effects land use policies have had on the decentralization of employment I do not know.

- *fiscal externalities*

In previous subsections, I have mentioned two fiscal externalities. The first was the fiscal externality associated with in-migration. The second was the fiscal externality, under property taxation, imposed by a property owner living in homes with below-average property value on households living in homes with above-average property values. In this subsection, I shall simply touch on a third form of fiscal externality that is typically treated as a topic in a sub-field of public economics, fiscal federalism, rather than in urban economics.

Provincial-local fiscal relations are characterized by a bewildering number and variety of provincial-local grants and cost-sharing arrangements, which are the subject of the theory

---

[28] "Flight from blight" was another major cause in US but not Canadian cities. Mieszkowski and Mills (1993) provides a strong discussion of the relative importance of increased auto ownership and flight from blight in residential decentralization in US cities.

[29] Glaeser, E., and M.E. Kahn., 2001. "Decentralized Employment and the Transformation of the American City". *Brookings-Wharton Papers on Urban Affairs* 2: 1-47.

[30] Different professional groups use the term "sprawl" differently, and professional usage of the word differs from colloquial uses. When they use the term sprawl, most economists mean spatial development at *inefficiently* low density.

of intergovernmental grants. When a local government provides a public service, it may provide substantial benefits to residents of other jurisdictions. The primary example is local primary and secondary education[31]. Many primary and secondary students educated in one jurisdiction will migrate to other jurisdictions. Empirical evidence, in the literatures on both the social *vs* private returns to education and on agglomeration, provides strong support that more educated individuals confer greater positive externalities. If there were no provincial-local grants, local governments would provide less than the efficient amount of public education since their residents would incur the full cost of providing the service but would enjoy only a fraction of the benefits; there would be a positive fiscal externality from each jurisdiction to the rest of the world. The efficient arrangement is for the local government to pay for the same fraction of costs as it receives in benefits, which is one aim of provincial-local grant and cost-sharing programs.

How efficient these various programs are in the aggregate is clouded with uncertainty simply because their complexity discourages policy researchers. Whatever their overall efficiency, there are likely substantial efficiency gains to be had by rationalizing and simplifying these arrangements.

There are six main takeaways from this section:

- There are a variety of important externalities in the context of managing metropolitan growth efficiently that potentially merit government intervention.
- To the extent justifiable considering informational constraints and implementation costs, congestion externalities should be dealt with *via* the Pigouvian prescription of pricing them so that economic agents face the full social costs of their actions.
- The principal force encouraging the spatial concentration of economic activity is economies of scale in production. In the metropolitan context, these economies of scale are almost entirely external to the individual firm, and therefore entail positive agglomeration externalities. Unfortunately, since these externalities are largely atmospheric in nature and prohibitively costly to measure at the level of the individual agent, it does not seem feasible to price them. This precludes attainment of the first best, and makes metropolitan transport and land use policy design an exercise in the theory of the second best.
- Immigration should be encouraged when the benefits that the marginal in-migrant confers on existing residents exceed the costs she imposes on them. Migration efficiency is achieved when migration incentives and disincentives are designed to align the private and social benefits of migration. The issue arises concerning whether efficiency in this context means global or local efficiency.
- Most land use externalities are local. Such land use externalities are best dealt with through the local land use regulatory process rather than via pricing policies. Land use externalities and land use regulations have some effects that are metropolitan in

[31] Another context in which these programs are important is central city-suburban fiscal relations. Between 1945 and 1980, when almost all central cities, not only in the US but also in Canada, were in distress, central cities used to argue vociferously that the suburbs should compensate them for the benefits they provide to suburban residents.

nature and need to be dealt with at the metropolitan level. A particularly important one has been the historic underpricing (i.e., pricing below marginal social cost) of metropolitan auto travel, which has resulted in excessive metropolitan expansion at inefficiently low densities.

- The two most important forces determining metropolitan spatial structure are transportation costs, which encourage dispersion, and scale economies in production, which encourage concentration. It is not surprising therefore that the two most important externalities relate to transportation and scale economies in production. The underpricing of the auto congestion externality results in excessive dispersion. The agglomeration externalities that result from scale economies in production being external rather than internal have complex effects on urban spatial structure, which will be discussed in the companion paper.

Economies of Scale

Recall that the First Theorem of Welfare Economics states that competitive equilibrium is efficient under a certain set of conditions, one of which is that production be "convex", which entails diminishing marginal rates of substitution between any pairs of factors in the production of all commodities, as well as constant or diminishing returns to scale in production.

Consider an unregulated economy in which the production of one of the goods is characterized by increasing returns to scale at all levels of production. The more a particular firm in this industry produces, the lower its average cost. Suppose that there is a group of firms in the industry, one of which produces more than all the others. If that firm prices below the average cost of all the other firms in the industry but above it owns average cost, it can force all the other firms out of the market and still make a profit. Thus, increasing returns to scale in an industry generates a natural monopoly. Since average cost exceeds marginal cost when there are economies of scale, if the natural monopolist prices efficiently, at marginal social cost, it will lose money. It will instead choose some form of monopoly pricing, resulting in consumers facing a price that exceeds marginal cost, causing them to consume less of the good than is socially optimal. Now consider an unregulated economy in which production of one of the goods is characterized by a U-shaped average cost curve, with the efficient scale of production about one-third the size of the market. Equilibrium will tend to be characterized by either two or three firms, each of which is a price-setter, and all of which behave strategically towards the others. Whatever the form of the oligopoly equilibrium achieved, it will be inefficient.

The government might choose to intervene in those industries characterized by scale economies. It could solve for the surplus-maximizing level of output in such an industry, allocate the production of that level of output among the firms in the industry so as to minimize total cost, require that each firm price at marginal social cost, and then cover the firms' losses through general revenue; if it is efficient to have only one firm in the industry, the situation would be that of a regulated public utility. The efficient amount of

the good would be produced and it would be priced at the efficient level. Under classical assumptions, this would indeed solve the problems created by economies of scale. But after eighty years of experience with regulated public utilities, we understand that this classic solution is imperfect. The regulated firms know more about the production technology than the regulator, and can exploit their superior information via rent extraction, and furthermore have little incentive to innovate.

Thus, economies of scale in production create problems for the efficient allocation of resources. In the context efficiently accommodating metropolitan growth, economies of scale arise in two contexts, natural monopolies and external economies of scale.

- natural monopolies I: public utilities

In Canada and the United States, "utilities" -- gas, electricity, local telephone service, water, and sewage[32] -- are provided by "semi-local" quasi-governmental agencies. Gas, electricity, and water are produced according to a three-stage production structure: generation, transmission, and (local) distribution. Traditionally, each stage of the production process has been managed by a separate regulatory body, though in recent years the generation stage for some public utilities has been deregulated in those industries in which it was discovered that economies of scale in generation are not strong. There are strong technological economies of scale in transmission[33], which justifies a single company or agency being responsible for transmission over a region. The local public utility companies then buy whatever is being transmitted, and distributes it over the local distribution network, which is also characterized by decreasing long-run average cost. The local rate structure is negotiated between the local public utility company and its utility board, subject to a variety of regulations, which may include a maximum deficit constraint.

The ideal is pricing of the utility at short-run marginal social cost. But, with scale economies in production, pricing at short-run marginal social cost results in operation at a loss. When account is taken that public utilities' losses are financed out of general revenue, which is raised with distortion, and that eliciting information from the public utility company requires providing them with some "informational rent", "second-best" pricing typically involves some form of Ramsey pricing, in which prices are calculated as markups over short-run marginal cost, with the markup rate being inversely proportional to the demand elasticity.

---

[32] Garbage collection is a borderline utility, with only modest economies of scale. In some communities, garbage collection is privatized, though firms and households are required to have their garbage removed for public health reasons.

[33] Consider a pipeline. Flow is proportional to the area of the pipe, while the surface of the pipe is proportional to its radius. Thus, a doubling of the radius of the pipe results in a doubling of the surface of the pipe that needs to be constructed but a quadrupling of the oil flowing through the pipeline.

Optimal distribution capacity is that level of capacity that minimizes the average cost of distributing the optimal quantity. It is also that level of capacity for which the social marginal benefit of expanding capacity equals the social marginal cost.

Efficient capacity policy differs across utilities, differing according to the utilities' specific technologies. Since I know little about the distribution technologies of specific utilities, let me instead first discuss the technology of constructing highway capacity, which is analogous and which I am broadly familiar with.

Traffic engineers define capacity to be the maximum sustainable flow. The most obvious way to expand capacity is to add lanes, but capacity can be expanded in other ways as well, such as providing better banking around corners, smoothing grade variation, improving lighting and traffic signing, providing a road surface that is well suited to local weather conditions, and repaving the road periodically to reduce its roughness. Effective or expected capacity can also be increased by designing the road to make accidents less likely, and by improving the speed at which accidents are dealt with. Thus, even though, at first glance, capacity appears discrete, in fact it is continuously variable.

Textbook analyses typically simplify by assuming that efficient infrastructure design involves only designing infrastructure so that a given level of capacity is obtained at minimum cost, and then determining optimal capacity. But there other aspects to efficient infrastructure design. One that is now receiving attention in the transportation literature is *reliability*, which has various interpretations, one of which is variability in travel time. Another is speed. The cost-minimizing way of designing a road with a given level of capacity may entail slow travel at capacity flow; for a given construction cost, it might be desirable to have a highway with a lower capacity but higher average speed. Another is service quality; driving is more pleasant on smoother road surfaces. Yet another is infrastructure performance under transient rather than steady-state demand conditions. How does the infrastructure respond when the entry flow exceeds capacity? California has introduced "ramp metering" on its freeway entry ramps, which restricts the rate at which cars can enter the freeway. This has markedly reduced the turbulence at entry points that was responsible for much traffic jamming, essentially converting hypercongestion into queuing.

While the engineering detail is specific to the particular utility, the same general economic considerations apply. For example, voltage fluctuations, brownouts, and blackouts can occur in electricity distribution networks when demand exceeds capacity. And there are still some local telephone networks that have not switched to fiber optic technology that experience jammed switches when demand exceeds capacity.

Most cities now face the issue of how efficiently to upgrade their out-of-date and degraded utility networks (for example, lime build-up has reduced the capacity of many water distribution networks). For reasons that I do not understand, these important economic issues have not been taken up by urban economists, and seem to have been overlooked by the profession at large. There is a well-developed body of theory on optimal facility maintenance, periodic upgrading, and replacement, but I do not know of a

literature that applies this body of theory to urban utilities, taking into account the engineering particularities of the different utilities. Nor do I know how cities are budgeting for these costly upgrades.

- natural monopolies II: urban public transportation/mass transit

In an important paper published in 1972, Herbert Mohring identified two intrinsic sources of scale economies in mass transit.[34] The first is *economies of service frequency*. Take a given bus network. Double the flow of passengers and double the frequency (and hence the number) of buses. Per passenger, walking time (from the origin to the boarding bus stop, and from the alighting bus stop to the destination), boarding and alighting time, bus crowding, and operating costs remain unchanged, but since bus frequency doubles, average waiting time halves. The second is *economies of service density*. Hold bus frequency on a given street fixed. Double the flow of passengers and double number of streets serviced by a bus. Per passenger, waiting time, boarding and alighting time, bus crowding, and operating costs remain unchanged, but since service density doubles, average walking time halves.

It turns out that these sources of scale economies are sufficiently important quantitatively that pricing bus travel at short-run marginal social cost and providing optimal capacity results in bus fares covering only about 80% of the operating and amortized capital costs. Thus, even absent second-best considerations, the subsidization of mass transit out of general revenue has a solid justification in theory.

Two other considerations provide support for even higher subsidy rates. The first is the underpricing of urban auto travel. Suppose that there are two ways of getting from A to B, by car and by mass transit, and that the aggregate demand for travel between the two locations is sensitive to price. The underpricing of urban auto travel generates distortion on two margins of choice, trip frequency and travel mode. Too many trips are taken and an excessively high proportion of these trips are taken by car. Now reduce the bus fare below short-run marginal social cost. Doing so reduces the distortion on the modal choice margin but increases the distortion on the trip frequency margin. The empirical magnitudes are such that the second best normally entails setting the bus fare below short-run marginal social cost.

The second consideration providing support for an even higher subsidy rate than that justifiable on the basis of Mohring economies of scale is distributional. In Los Angeles, at any rate, traveling by bus may be a better indicator of need than a low taxable income. If this is the case, subsidizing bus travel may provide a more efficient way of helping the needy than is redistributing through the income tax system.

---

[34] Mohring, H., 1972. "Optimization and Scale Economies in Urban Bus Transportation," *American Economic Review* 62: 591-604.

On the basis of these considerations, Parry and Small argue that current levels of subsidy to mass transit systems are about right.[35]

While the economic principles determining optimal capacity are well established, the analysis of optimal capacity is typically done one mode at a time. Only very recently have metropolitan transportation economists started to analyze the optimal design of entire metropolitan transportation systems. Mohring economies of scale may result in multiple local optima. There are three different reasons why. The first concerns alternative mass transit modes -- bus, LRT, and subway. Take any pair, and suppose for the sake of argument, that their networks coincide and that they are identical in other respects as well, so that they are perfect substitutes in demand. Because of Mohring economies of scale, the cost-minimizing mass transit system entails either one mode or the other, but not both together. The reason is that each mode draws passengers away from the other, undermining the benefits each achieves through a higher volume of traffic. Since the networks of different modes do not coincide, and since they differ in other service characteristics as well, they are in fact not perfect substitutes. Thus, multiple transit modes may be optimal, but multiple local optima are to be expected. The second reason why Mohring economies may result in multiple local optima is the interaction between mass transit and auto travel. To illustrate, suppose that there is a single mass transit mode operating on a fixed network and congestible, auto travel on city streets on a separate but parallel fixed network, that the two modes are perfect substitutes in demand, and that overall trip demand is inelastic. Use of both modes together cannot be optimal. Suppose it were. Then transferring a passenger from auto travel to mass transit travel decreases trip cost for both car drivers and mass transit passengers. Again, when account is taken that the two modes are not perfect substitutes, there are likely to be multiple optima. The third reason to expect multiple local optima is the technology of bus-car interaction, with a bus contributing more to congestion the higher the ratio of cars to buses, and cars contributing more to congestion the higher the ratio of buses to cars.

In recent years, the major transportation issue in California[36] has been the reallocation of the transportation funds away from car travel and towards mass transit travel[37]. The result to date has been worsening freeway congestion (Los Angeles is now the most congested city in North America, according to both Tomtom and the Texas Transportation Institute) but also the City of Los Angeles has experienced a significant percentage increase in transit ridership (though from a low base). Those who favor more spending on mass transit and less on the freeway system not only present the standard new urbanist and green arguments, but also argue that this allocation of the transportation

---

[35] Parry, I. and K. A. Small, 2009. "Should Urban Transit Subsidies Be Reduced?," *American Economic Review* 99: 700-724.

[36] Except perhaps for the proposed high-speed rail line between San Francisco and San Diego, where the issues are only tangentially related to intra-metropolitan transportation.

[37] The reallocation has been largely accidental. The federal gas tax revenues, which are used to finance the freeway system, have been falling; local governments have been upgrading their bus systems; and the City of Los Angeles has been expanding its LRT system.

budget will generate a virtuous cycle. Expansion of the mass transit system, along with a reduction in expenditure on the freeway system will generate densification (which is apparently called intensification in Calgary), which will improve the viability of mass transit. Furthermore, without densification auto traffic congestion would get worse and worse. Those who oppose more spending on mass transit and less on the freeway system argue not only that the automobile metropolis works but also that it is the revealed preference of Southern Californians, who appreciate the cheaper housing, less congested lifestyle, and the greater privacy and convenience it allows. Due to the decentralization of employment that auto travel induces, Los Angeles can expand spatially virtually without limit with only modest increases in commuting times. Furthermore, even though its low density (relative to metropolitan areas of its size) may be partially attributable to the subsidization of auto travel in the past, the densification required to make mass transit viable would be prohibitively expensive even if it were desirable. Both sides of the debate may be partially correct in the sense that each identifies a different local optimum as the goal of transportation policy. Reasonable men may disagree over which of the two local optima is better, though I have my own opinion.

In almost all cities, there is substantial debate over the allocation of the mass transit budget between bus and rail. The vast majority of urban transportation economists favor expansion of the bus system[38] over expansion of the rail system. The main argument is that their technologies are such that bus transportation is considerably cheaper per mile than rail transportation. Another argument, which was made above, is that, due to economies of scale, having one dominant mass transit system is preferable to having a balance, which is supplemented by the argument that inter-modal transfers always present difficulties. Yet another argument is that bus transportation permits more flexibility, easily accommodating uncertain changes in the spatial distribution of trips. I elaborate on these points in the companion paper, questioning whether it is wise for Calgary to extend a light rail system that is strongly oriented towards the city center.

- *agglomeration economies*

I have already discussed agglomeration economies at some length in the section on externalities. The term agglomeration economies is sometimes used in the broader sense to refer to set of forces that induce the spatial concentration of economic activity, and sometimes in the narrower sense of external economies of scale in which productivity increases with spatial concentration. Agglomeration economies generate agglomeration externalities since each extra unit of output produced or each extra unit of labor hired by a particular firm in a metropolitan area increases the productivity of all other firms in the city. Attempting to deal with agglomeration externalities by applying the classical, Pigouvian prescription seems futile since they are atmospheric in nature.

In this subsection I shall address two related questions. First, how do agglomeration economies distort production in a metropolitan area? Second, since it seems futile to

[38] Some cities' mass transit systems are organized around high-speed bus corridors, with feeder buses. Such a system is intermediate between a regular bus system and a rail system.

attempt to internalize agglomeration externalities, what policies might mitigate the distortions that they create?

As a reference point, I start with a classical (Hecksher-Ohlin type) economy with transportation costs in which: i) there are two goods and two industries, and three factors of production, land, skilled and unskilled labor; ii) individuals consume the two goods and land; iii) the economy comprises two equal-sized islands, with the transportation of people and goods being costless on an island but costly between islands; iv) the economy has a fixed supply of both types of labor, with costless migration (though inter-island commuting is costly); and v) the technologies are the same on both islands and exhibit constant returns to scale. There is a unique local optimum, which is also the global optimum, which exhibits symmetry between the islands, and no trade or commuting between the islands Furthermore, the competitive equilibrium is unique and coincides with the optimum. Thus, there is a unique equilibrium that is symmetric, stable, and efficient.

Now modify the model in only one way. Assume that industry 1 is characterized by external economies to scale; in particular, assume that industry 1's location potential on island i is an increasing function of the number of skilled workers employed in industry 1 on island i. The first important point is that, because of the agglomeration externalities in the production of good 1, equilibria and optima will not in general coincide. The second important point is that localized economies of scale in the production of good 1 provides an incentive for all production of good 1 to be located on only one of the islands. Depending on parameter values, the optimal allocation may or may not entail symmetry, and there may be multiple asymmetric local optima. The symmetric allocation is always a local optimum, but it may or may not be stable, depending on the adjustment mechanism, and there may be multiple asymmetric equilibria as well that alternate in stability. While it isn't true that "anything can happen" or that "nothing can be said about the relationship between equilibrium and optimum", the relationship between locally optimal allocations and equilibrium allocations is complex, even in this very simple model. Thus, except in specific cases, one of which I shall discuss in the companion essay, it seems that there are no simple ways to mitigate the efficiency loss associated with agglomeration externalities.

Conclusion

In this essay, I have examined efficient resource allocation in a metropolitan area. The points of departure were the First Theorem of Welfare Economics, which states that, under a stringent set of assumptions, competitive equilibria are efficient, and the classic theory of market failure, which identifies the three classic "distortions" (deviations from the assumptions of the perfectly competitive model), public goods, externalities, and increasing returns to scale, explains how each upsets the efficiency of competitive equilibrium, and derives policies that restore efficiency. I then applied the classic theory of market failure in the context of the allocation of resources at the metropolitan scale and from the perspective of urban economics. The urban economic perspective is not

inconsistent with the public economics perspective but it pays more attention to space, and metropolitan and land use policy, and less to tax policy and intergovernmental fiscal arrangements.

References

Bator, F. M., 1957. "The Simple Analytics of Welfare Maximization," *American Economic Review* 47: 22-59.

Bator, F. M., 1958. "The Anatomy of Market Failure," *The Quarterly Journal of Economics* 72: 351-379.

Buchanan, J. M., 1965 . "An Economic Theory of Clubs," *Economica* 32: 1-14.

Coase, R., 1960. "The Problem of Social Cost," *Journal of Law and Economics* 3: 1-44.

Fujita, M. and J.F. Thisse, 2002. *Economics of Agglomeration: Cities, Industrial Location and Regional Growth*. Cambridge: Cambridge University Press.

Fujita, M., P. Krugman and A. J. Venables, 2001. *The Spatial Economy: Cities, Regions and International Trade*. Cambridge: MIT Press.

Glaeser, E. and M.E. Kahn, 2001. "Decentralized Employment and the Transformation of the American City," *Brookings-Wharton Papers on Urban Affairs* 2: 1-47.

Henderson, J.V. and J.F. Thisse, eds. 2004. *Handbook of Regional and Urban Economics*. Cambridge:Elsevier.

Hindriks, J. and G. Myles 2006. *Intermediate Public Economics*. Cambridge: MIT Press.

Jacobs, J., 1961. *The Death and Life of Great American Cities*. New York: Vintage.

Kaplow, L., 2008. *The Theory of Taxation and Public Economics*. Princeton: Princeton University Press.

Kraus, M., 1981. "Scale Economies Analysis for Urban Highway Networks," *Journal of Urban Economics* 9: 1-22.

Krugman, P., 1980. "Scale Economies, Product Differentiation, and the Pattern of Trade," *American Economic Review* 70: 950-959.

Marshall, A. 1890. *Principles of Economics*. MacMillan, London.

Mas-Colell, A., M. D. Whinston and J. R. Green. *Microeconomic Theory*. New York: Oxford University Press, 1995.

Mieszkowski, P. and E. S. Mills, 1993. "The Cause of Metropolitan Suburbanization," *Journal of Economics Perspectives* 7, 135-147.

Mirrlees, J. A. 1971. "An Exploration in the Theory of Optimum Income Taxation," *The Review of Economic Studies* 38 :175-208.

Mohring, H., 1972. "Optimization and Scale Economies in Urban Bus Transportation," *American Economic Review* 62: 591-604.

Mohring, H., 1976. *Transportation Economics*. Cambridge: Ballinger.

Mohring, H. and M. Harwitz, 1962. *Highway Benefits: An Analytical Framework.* Evanston: Northwestern University Press.

O'Sullivan, A., 2009. *Urban Economics*. New York :McGraw-Hill.

Salanie, B., 2000. *The Microeconomics of Market Failures*. Cambridge: MIT Press.

Parry, I. and K. A. Small, 2009. "Should Urban Transit Subsidies Be Reduced?," *American Economic Review* 99: 700-724.

Small, K. A. and E. T. Verhoef, 2007. *The Economics of Urban Transportation*. New York: Routledge.

Smith, A., 1937. *An Inquiry into the Nature and Causes of the Wealth of Nations*. New York: Modern Library.

Tiebout, C., 1956. "A Pure Theory of Local Expenditures," *Journal of Political Economy* 64, 416-424.

Varian, H., 1987. *Intermediate Microeconomics*. New York: W.W. Norton.

Varian, H., 1978. *Microeconomic Analysis*. New York: W.W. Norton.

Vickrey, W., 1947. *Agenda for Progressive Taxation*. New York: Ronald Press.

Vickrey, W., 1963. "General and Specific Financing of Urban Services," in *Public Expenditure Decisions in the Urban Community* edited by Howard G. Schaller, Resources for the Future, 62-90..

Vickrey, W., 1977. "The City as a Firm." in *The Economics of Public Services* edited by Martin S. Feldstein and Robert P Inman. New York: Palgrave Macmillan.