Congestion and Metropolitan Development

Alex Anas
Professor of Economics
State University of New York at Buffalo
Amherst, New York, USA

Twelfth Annual International Conference "Economy and Geography"
The Leontief Center
St. Petersburg, Russia

February 14-16, 2013

Written March 26, 2013[1]

1. Urban travel and the congestion problem

All over the world as people move to urban areas and as they get wealthier in cities they buy cars and – other things being equal – the vast majority of people prefer private mobility to public transportation. In rapidly growing Chinese and other Asian cities car ownership is increasing so fast that roads and highway capacity must be steadily expanded and public transit systems must be built as well.

In the United States where almost all people who would move to cities have already done so, road capacity is no longer an issue in most places and the thinking has shifted to better managing existing capacity rather than building more. But trips are increasing even in the United States because of several reasons. First, there is still a net movement of people from the central cities to the suburbs, and as people suburbanize they must rely more on driving, since in American suburbs public transit is scarce and walking impractical. Second, even though US car ownership is not increasing rapidly, people are making more trips especially of a purpose not involving going to work.

Traffic congestion is getting worse in almost all major world cities. The Texas Transportation Institute which tracks congestion on a regular basis has consistently documented an increase in urban congestion in American cities (TTI, 2012). And according to the International Association of Public Transportation, the urban road travel

[1] This article has been written based on a power point presentation prepared for the conference. I am thankful for the support of grant 142934 from the University of California, of which I am acting as the scientific director.

speed in Beijing, one of the world's most congested cities was only 11 km/hour on average in 2001. (IAPT, 2001)

The congestion problem implies serious costs both private and social. The private cost of congestion is that time is allocated to travel can be used to do other more productive or more enjoyable activities that are work or leisure-related. So if people can travel faster there would be benefits in terms of higher production or a higher quality standard of living. But the more important cost of congestion is its social cost which arises from the fact that congestion is a negative externality. Congestion occurs when the number of cars sharing a road exceeds some measure of the road's capacity so that adding a car delays the other cars using the road. This cost of the delay each car imposes on other cars remains un-priced. That is, when I join the traffic stream I do take my own delay into account and so pay for it but cannot take into account the delay I cause on others since I am not required to pay for it. The delays cars cause on one another make up the social cost of congestion. In economics, a negative externality is characterized by the social marginal cost exceeding the private average cost. This means that for congestion to occur at a socially optimal level, a traveler who causes congestion should be charged a tax (commonly known as a congestion toll) that extracts from that traveler the value of the delays he is causing.

What a congestion toll would do is to raise the cost of travel to the point where the gap between social marginal cost and private average cost is eliminated which would mean that the negative externality is internalized and that there is no more a social loss from urban travel. Under the congestion toll there would be less travel and it is optimal that this be the case. But optimal congestion is not zero congestion, so there would still be congestion.

The congestion toll is a simple idea based on the work of British economist Arthur Pigou (see Pigou, 1932) who was the first to examine the welfare implications of externalities. His ideas were proposed and refined for the traffic congestion problem by William Vickrey (1969) and others. But, despite all the theoretical work, the Pigouvian congestion toll is also a difficult concept to implement in the real world. One reason for this is that congestion differs greatly by time of day and by particular pieces of a road. It needs to be measured carefully and levied mile-by-mile, hour-by-hour. Another reason is

that delays occur mostly in time units like minutes but congestion tolls must be levied in monetary units like dollars. This requires that the concept of the value of time be used to convert delays from time to monetary units. This seems simple to do, but consider that when I get on the road I am delaying a variety of others sharing the road with me. Some of these people have high values of time, while others have low values of time. For example, I may be delaying a doctor who is racing to save a life and a young unemployed brat who is simply aimlessly joy driving. What should be my congestion toll? Based on the correct theory, I should pay a very high sum indeed for delaying the doctor and close to nothing for delaying the youngster. My congestion toll cannot be calculated precisely unless the doctor's and the youngster's values of time can be known.

The above example may sound terribly discouraging. But economists have shown that congestion tolls can be welfare improving even if they cannot be calculated perfectly accurately as long as we are willing to limit our thinking to average situations. Making people pay more – but not pay too much – will reduce the congestion externality and will improve welfare though not to the optimal level. Especially when congestion is very high it is attractive to think about making people pay more without risking the overcharging. It appears that in most practical situations we are pretty confident that we would not overcharge for the congestion externality.

Another difficulty with congestion pricing is that of applying it road-by-road and minute-by-minute (or hour-by-hour). With GPS technology even such a scheme seems within reach but in all real cases where congestion tolling has been implemented – Norwegian cities, Singapore, London, Stockholm, Milan – it has been done in a limited way by making cars pay when they cross a cordon line enclosing a highly congested central urban area or when traveling inside a pre-designated highly congested area. All these schemes are considered successes and seem to have found the acquiescence if not the outright support of citizens. And it is reasonable to expect that there will be more applications in the future. Congestion pricing has been considered for Manhattan and less formally for Chicago but rejected in both cases. This is not surprising because American cities, even Manhattan, are not very congested compared to European and Asian megacities although the level of congestion seems to be rising in most places.

2. Travel time and urban size in the U.S.

Since the main focus of this article is how congestion affects urban development and the internal organization of cities, it is useful to start with a cross section of US metropolitan areas (the 50 largest) and ask just how long it takes for people to commute to work. To be sure, total travel time from home to the workplace includes both congested and uncongested travel and so measuring total travel time does not directly say anything about the social cost of congestion. However, there is a great deal of evidence that people consider total travel time in their location decisions and also in their decisions to allocate time among leisure, work and travel. So total travel time is an important variable to study if we want to understand how congestion affects urban development which we will do in the next section.

Now turning to Figure 1 we see that it shows a regression line. The horizontal axis is the logarithm of the size of the metropolitan area measured as the number of workers older than 16 years of age. The vertical axis is the logarithm of the one-way travel time in minutes from home to work. The sample points (many of which are shown as dots) are the 50 largest US metropolitan areas. The data is all from the decennial Census which surveys large numbers of individual commuters no matter what their mode of commuting. The point in time is the year 2000 (results are very similar for 1990). The slope of the regression line is almost exactly 0.1. This means that if we consider a metropolitan area twice as large in workers as another, the larger of the two will have only 10% higher travel time on average.

To be more precise, Table 1 summarizes the essence of the finding by extracting the numbers for Louisville, Pittsburg, Houston, Chicago and New York. Going from one of these metropolitan areas to the next population is almost exactly doubled and average travel time increases by about 10% in each step. To put it more coarsely, New York is 16 times as big as Louisville but has only 50% longer travel commute times on average. Most people, economist and graduate students included, find these results surprising. There is an ingrained expectation that for a metropolitan area twice as big, the average travel time should be increasing by much more than 10%! A common guess is 50% increase in commute time if the metropolitan size is doubled. If that were true, the average one-way commute time in New York would be 115 minutes, and it would take

the average New Yorker almost 4 hours each day to get done with commuting! And – if that were true – I dare say with a great deal of confidence – that New York would not be New York as we know it but quite different.

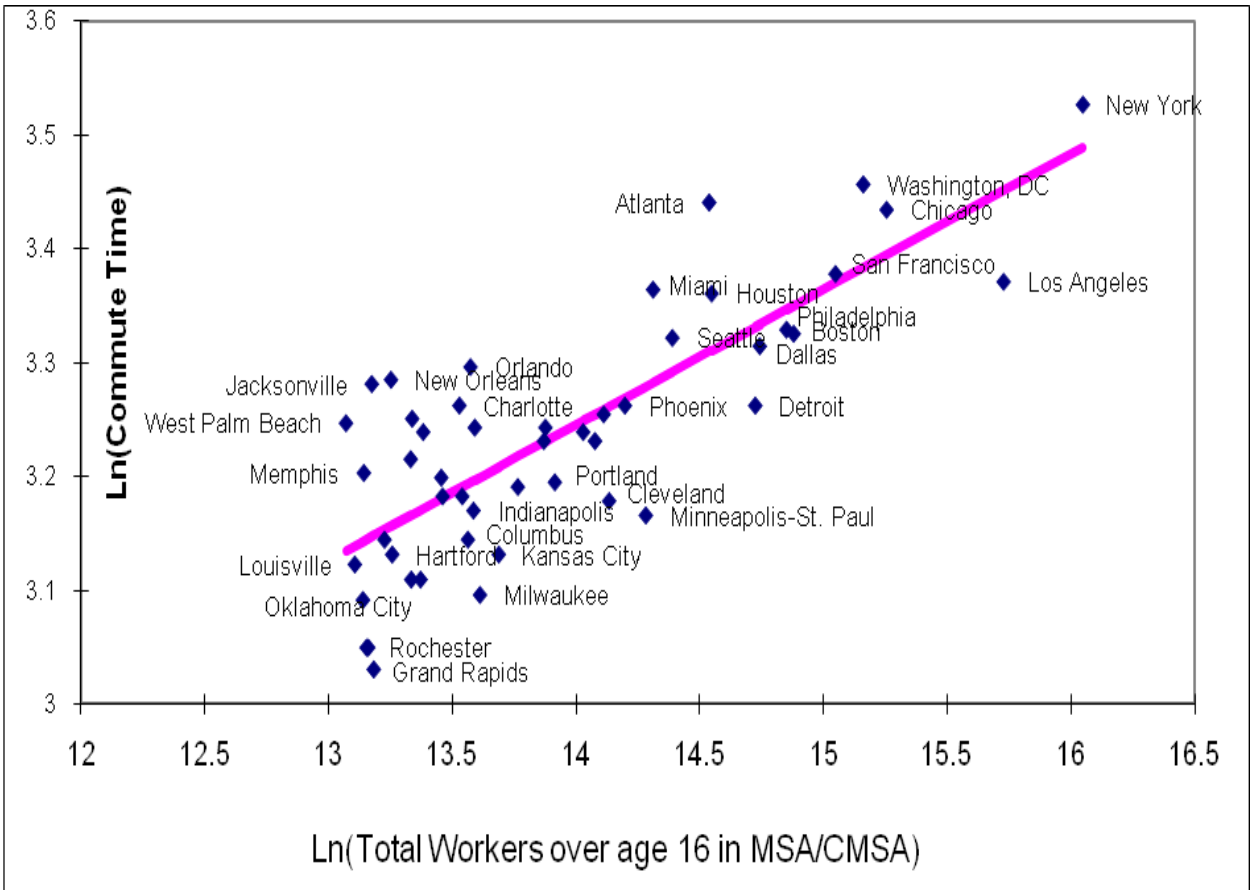


Figure 1
(*Source*: Anas, 2011)

Doubling population increases commute time by 10%

URBAN AREA	WORKERS	AVERAGE COMMUTE
LOUISVILLE	0.5 million	22.7 minutes
PITTSBURG	1.0 million	25.5 minutes
HOUSTON	2.0 million	28.8 minutes
CHICAGO	4.0 million	31.0 minutes
NEW YORK	8.0 million	34.0 minutes

- *New York has 16 times more workers than Louisville but only 50% higher commute time*

Table 1

There are some other findings hidden in the regression line. One is that the higher is the percentage of workers who commute by public transportation the longer is the travel time. Although this effect is statistically significant it is not extremely so. This is as expected because commuting by public transit is cheap but slower. It requires accessing stations and waiting both at the beginning and at the end of the trip. The share of commuting by transit increases with size reaching the mid to high 30% range in New York. We will discuss the role of public transportation in more depth later. As well, congested travel increases as metropolitan size increases, something we know from the studies of the Texas Transportation Institute (TTI, 2012). Importantly, metropolitan size is not independently determined. As congestion rises adjustments occur that determine both the arrangement of economic activity internally within a metropolitan area and at the same time also determine the total size of the metropolitan area measured in population, workers or land area.

Some other things about the regression are not important in this presentation. Note, for one, that the variance around the regression line is larger for smaller than it is for

larger metropolitan areas. This may be a function of the fact that there are simply more small metropolitan areas than there are larger ones. Or it may reflect something more subtle. But it will not concern me here.

Somewhat more interesting is whether the regression tells us meaningful things about particular metropolitan areas. I claim that it does. Note for example, that Atlanta is highly congested for its size, lying much above the regression line, while Los Angeles is not as highly congested for its size as we would expect, lying somewhat below the regression line. I think I know why. In the case of Atlanta, it had been a congested city with much pollution and it had a hard time getting federal funds for road building because it exceeded pollution standards. It did, however, build a transit system which lengthens travel times. So I believe Atlanta's higher than expected travel times stem from a lack of road capacity. In the case of Los Angeles the opposite is true. L.A. is the poster child of the automobile oriented city with only 2% of metropolitan trips by public transit and plenty of road capacity. I will say more on this later also. There is no claim here that any of these findings are also true for countries outside the US. In fact I do not believe that the results would be very similar.

3. How do metropolitan areas adjust to congestion?

To understand how a metropolitan area would adjust to congestion, I draw on several additional observations about US metropolitan areas. One of these is shown in Table 2. It shows that the largest share of commutes occurs from suburban homes to suburban workplaces. The share is higher in the USA than in Canada. Suburb to central city commuting which was dominant in the 1950s has now decreased to a share of a bit over 20%. For suburb to suburb commuting to have flourished in the last five or six decades a lot of jobs must have moved to the suburbs alongside with households which, we know has been true.

US & Canadian Commuting Patterns

		United States	Canada
Residence	**Workplace**	**2000 Census (%)**	**2001 Census(%)**
Central city	*Central city*	**27.5**	**46.1**
Central city	*Suburb*	**8.9**	**7.5**
Suburb	*Central city*	**20.2**	**16.2**
Suburb	*Suburb*	**43.4**	**30.2**
		100.0	**100.0**

Table 2

Figures 2 and 3 highlight another aspect of suburbanization using a particular area– where I also happen to reside and work – the Buffalo-Niagara Falls metropolitan area. This is a midsize metropolitan area of a bit under one million people and it has very low congestion. Figure 2 shows that while the population has remained essentially flat between 1950 and 2000, the urbanized land area has increased more than fourfold! Figure 3 shows this urbanized land expansion in real geographic space where the red color denotes the new urbanized land added since 1950.

This problem of urban expansion that is rapid relative to population growth is known as the urban sprawl problem. American urban planners have been very concerned about the extent of urban sprawl. By some measures, while population in urban areas has increased by about 1% per year, urbanized land area as a whole has expanded by 2.5% per year. To an economist this is not necessarily very surprising since agricultural land in the United States is rather cheap (by some measures about $1/6^{th}$ as expensive as in Western Europe), incomes have been high and congestion has been low. So people

spread out buying big houses in the suburbs and get closer to suburban amenities and schools. In doing so they have distanced themselves from the crime, blight and generally lower quality schools of the central cities. Since 1950 the average house size per household in the US has increased considerably.

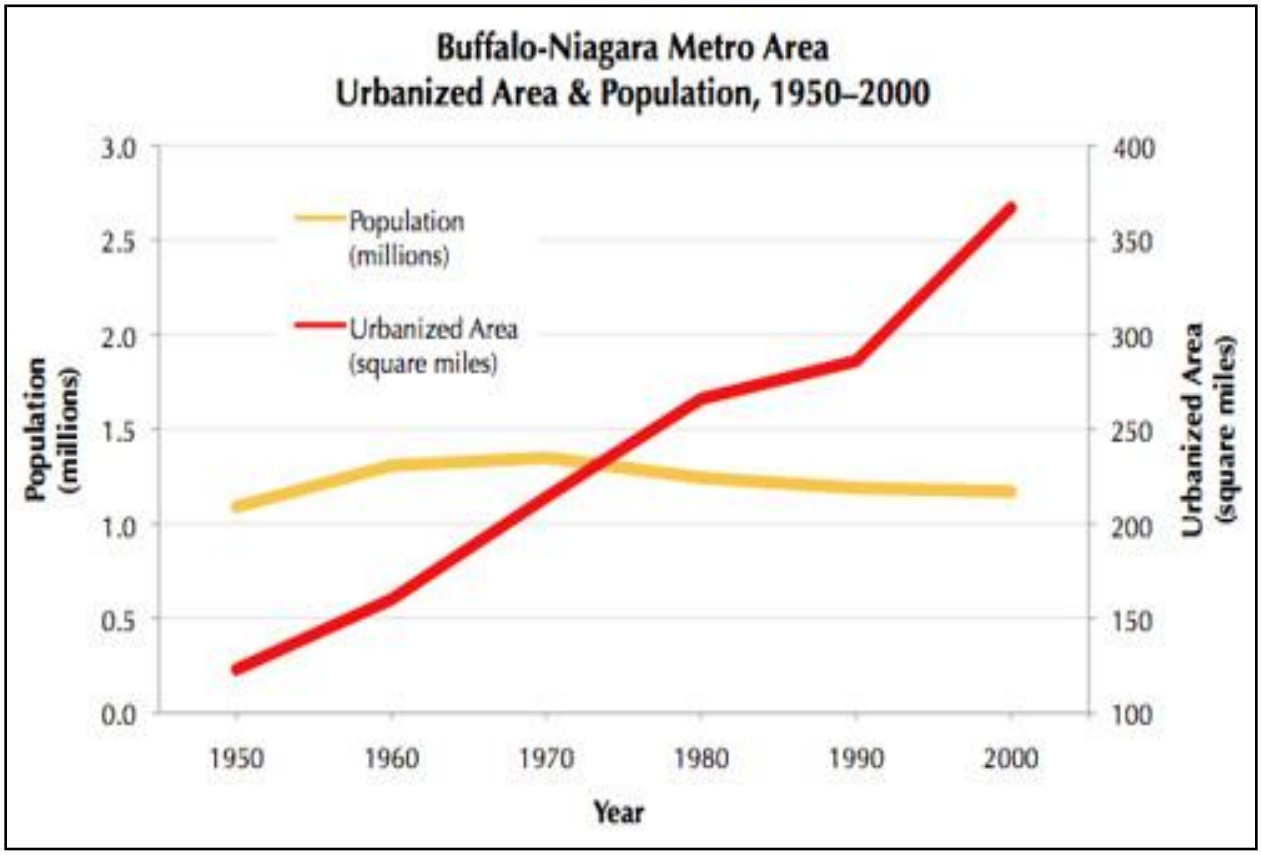
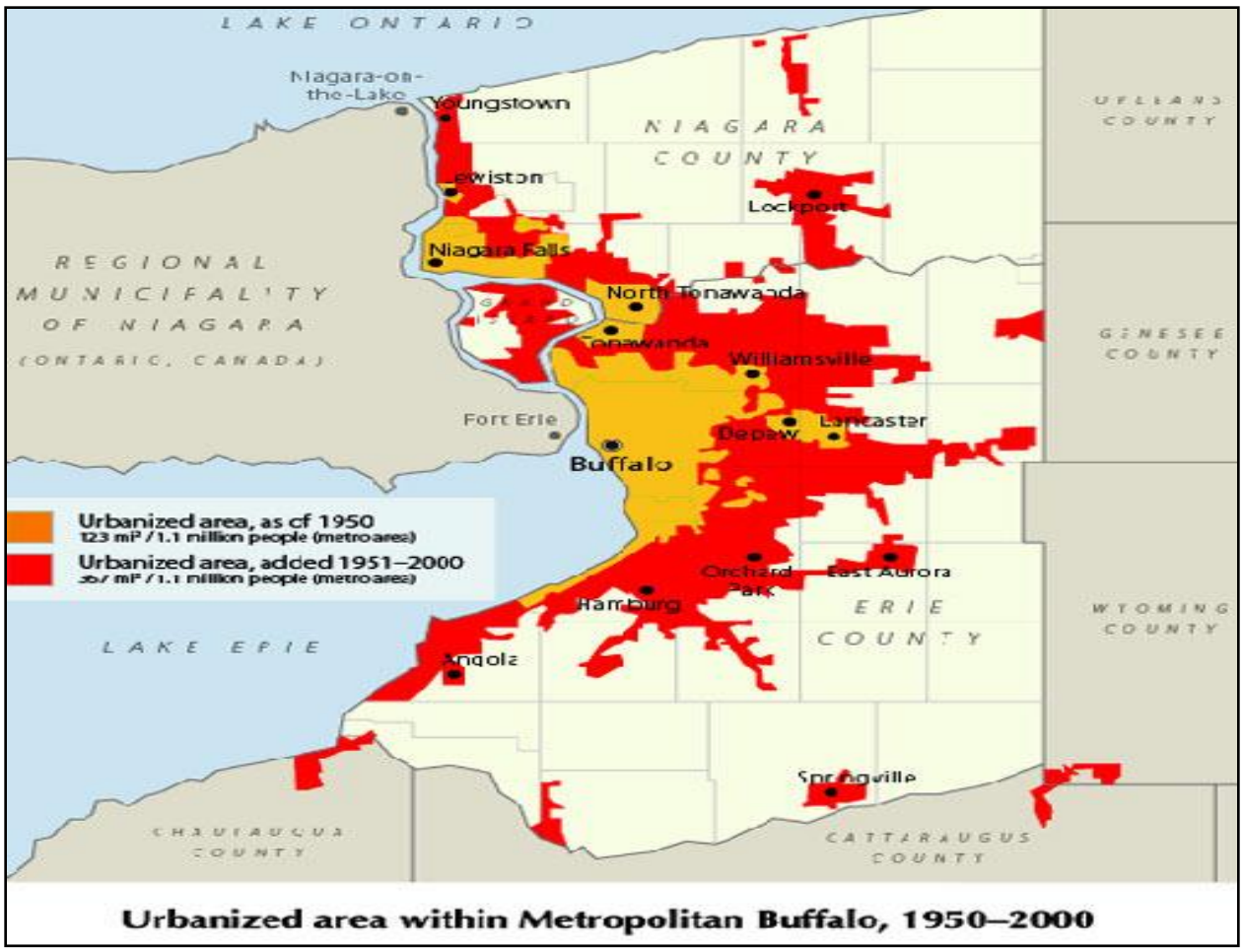


Figure 2
(Source: *Joe the Planner*)



Urbanized area within Metropolitan Buffalo, 1950–2000

Figure 3
(Source: *Joe the Planner***)**

These facts about job and residential suburbanization, travel to work and urban sprawl may be juxtaposed against the work of theoretical urban economics since its birth in the early 1960s. Virtually all of the early theory was developed by assuming a very simple and highly counterfactual urban form, that of the monocentric city (Alonso, 1964). In this model it is assumed that all jobs in the urban area are pinned to the center of a city and cannot change location. So households must spread themselves around the center to which they must commute daily. This situation is supposed to describe the North American city as it was in the late 1800s and early 1900s. But soon after that time it ceased to be monocentric as jobs increasingly moved to the suburbs.

A large number of articles have been written by very smart people since the 1970s analyzing the adjustment to traffic congestion in a monocentric city. Indeed, it is rather

easy to explain what happens. Suppose that I am located in a monocentric city that does not have public transportation. Say that my residence is 5 miles away from the downtown where all the jobs are and that I commute there daily. As more people join my city, roads get more congested. So over time it takes me longer and longer every month or year to commute. How can I make my own adjustments to offset this effect of congestion on me? Since each mile that I travel on is getting more congested and takes more time to travel, I have only one choice. I must reduce the total distance I travel on so that even though each mile is more congested and takes longer to travel, I keep my total travel time tolerable by traveling over fewer miles. So, perhaps, I move my place of residence from 5 miles from the center, to 3 miles from the center. I have shortened my trip by 2 miles. By moving closer to the center, I am now renting housing that is more expensive per square meter so I may also choose a smaller house so that my rent bill is not excessive. Rent is higher because there is less land available near the center since most cities are circular in shape or fractions of a circle and so land is more expensive nearer the center. By moving closer to the center, I am trading a higher rent for lower travel time. I would not make the move unless it was overall beneficial to me. Therefore the travel time saved must be more valuable to me than the higher rent per square meter paid for the smaller house.

In a monocentric city in which public transportation is not available there is no other margin of adjustment except the distance to the center. Indeed using sophisticated mathematics and microeceonomics, urban economists have shown by repeated analyses that as population increases in monocentric cities, the cities expand but housing density, rents and congestion per mile increase and more so near the centers of those cities. The conclusion is that population increase causes more congestion but denser urban areas.

I consider myself among the luckier urban economists because –after working a bit on the monocentric city – I developed models in which both jobs and households can relocate when congestion increases and other things change (see Anas, 2011 for a summary). The best way to see how jobs may relocate is to relax the assumption of the monocentric city that jobs must remain in the center no matter what happens. Consider a business located in the center. Over time the city grows in population but does not add more highway capacity. This causes the existing roads to become more congested and the workers of the business take longer to commute to the center. One effect will be that the

aggregate labor supplied to the center will be reduced. This will cause the market wage that must be paid by a business located in the center to increase. The business can stay put in the center and pay the higher wages and many businesses will continue to do that. But the business can also consider moving closer to its workers in order to reduce their travel distances and thus also their travel time offsetting the effect of the higher congestion per mile. Because the supply of labor to a location outside the center would not be reduced as much as the supply of labor to the center (since most congestion delays occur near the center), the business that relocates out of the center would be paying a lower wage relative to the business that stayed in the center. But there are other benefits as well. The business locating out of the center will be locating at a lower rent area and so can buy or rent larger facilities with more space per worker thus making each worker more productive.

Businesses moving out of the city centers may experience some costs as well. Being in the centers has advantages known as agglomeration effects that were noted by Alfred Marshall (Marshall, 1890; Fujita and Thisse, 2002). Being closer to many other businesses means better access to information and other non-pecuniary benefits. These are diminished by moving out of the center. But the dominant trend has been for businesses to disperse to the suburbs. Therefore, the pull of the agglomeration benefits of staying in the centers has been gradually overcome by the pull of the suburban locations to offer lower rents and better access to labor at lower productivity-adjusted wages.

The above discussion ignored the role of public transportation by assuming that none is available. But its presence would cause the above adjustments to be modified substantially. In most places in the world public transit lines, especially rail and subway, are focused to serving the city centers. The city centers are normally the biggest job concentrations and have enough economies of scale in trip-ends so as to make public transit lines serving them economically viable. In a city in which the center is served by public transportation an increase in road congestion could be accommodated by a traveler, not by relocating closer to the center but by switching to rail. While riding the train may often take longer (access and egress from stations are time consuming), road congestion is largely avoided and riding the train has some other benefits such as being able to read something during the ride. Because of public transportation, congestion's

effect on the decrease of labor supply is moderated and businesses located in the center may be less willing to move out and into the suburbs since market wages in the center would not increase as much as they would if public transportation were not available.

4. Studies with CGE models: Chicago, Paris, Los Angeles

The ideas explained above were confirmed by using computable general equilibrium (CGE) models to study real metropolitan areas. The RELU-TRAN (Regional Economy, Land Use and Transportation) model is the most complete CGE model of a metropolitan area. It draws on my previous work with partial urban modeling that dates back to the early 1980s. My various papers can be found on my website.[2] So far I have applied this model to Chicago and Paris and beginning this summer we will be applying it to the Los Angeles metropolitan area. I now offer a brief summary of what has been learned about congestion and metropolitan development from the Chicago and Paris applications and what more we expect to learn by applying the model to Los Angeles. Table 3 shows how different these three metropolitan areas are in terms of public transit use and employment decentralization.

	Public transit share in commuting	**Employment dispersion**
Chicago, MSA	13%	About 30% of jobs in the 4 largest job centers
Ile-de-France	50%	About 50% of jobs in the City of Paris and 10 surrounding growth poles
Los Angeles, MSA	2%	About 30% of jobs in the 30 largest job centers

Table 3

Note that Greater Paris (Ile-de-France) is highly centralized compared to US cities, and commuting relies very heavily on public transit. At the same time, the region exhibits high employment centralization with about 50% of the jobs in the central core area which includes the suburban growth poles and the City of Paris. By comparison,

[2] http://sites.google.com/site/alexanashomepage/

Chicago has only about 30% of the jobs in the four largest job centers and all except the downtown are suburban job centers. The downtown includes about 10% of the jobs. Los Angeles as a metropolitan area has very low transit share and 30% of the jobs are in 30 highly dispersed job centers. The downtown includes only about 4% of the jobs.

In the case of Chicago simulations with RELU-TRAN were made to see how the metropolitan area would evolve in the period 2000 to 2030 as population and total employment increased by about 24% (according to projections). These simulations were run for two scenarios. In one of these, road capacity is not changed at all (no new roads are built), while in the other road capacity is increased somewhat according to planners' projections. Consider the first scenario. As population increases but no new roads are built congestion increases on virtually all miles of road. This means that virtually all trips would get longer. Figure 4 shows the effects of the population increase on aggregate new construction by single family housing (SF), multiple family housing (MF), commercial and industrial floor space. Most notably it also shows that to make room for these new buildings, the available vacant and developable land decreases by about 14% during the 30 years.

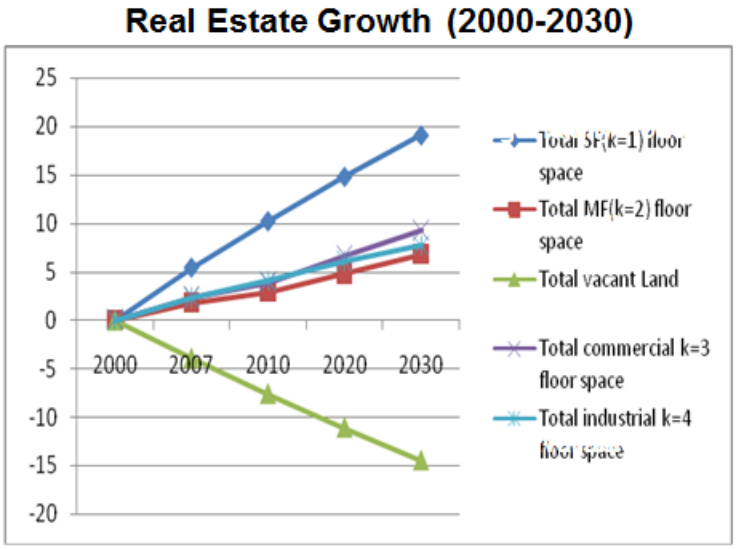


Figure 4

These results paint a picture of urban sprawl. Both jobs and residences spread out to new outlying areas. Figures 5 and 6 are very instructive. Figure 5 shows what happens to aggregate and per-capita VMT (vehicle miles traveled by cars) when no new

roads are built, while Figure 6 shows the same things when new road capacity is added according to projections. From Figure 5 note that the blue line which measures aggregate VMT is rising by about 13% from the year 2000. The reason there is a rise in VMT is simply that more people are added and since many of them drive aggregate VMT increases. Economists refer to an increase in an aggregate that comes about from the expansion of a market as the "extensive margin". But note that aggregate VMT is rising by only about 13% when aggregate population and jobs rise by 24% (mentioned earlier). As a result, the VMT per capita is actually decreasing by about 8-9%. A small part of this is explained by the fact that some of the additional population uses public transit for most (but not all) of its trips. But this is not a large fraction since public transit in Chicago has a share of trips amounting to only 12-13% of all trips, and much lower in the suburbs where almost all the new construction goes. So why does per-capita VMT decrease by as much as it does? The reason is that the new jobs and population suburbanize apace and thus the average distance between residence and workplace or residence and shopping place is reduced. This is the response to rising congestion that was explained in the previous section. Jobs and residences move closer to each other to offset as much as possible the effect of the higher congestion on each mile. Economists refer to the reduction in quantity that occurs at the micro level as the "intensive margin". So we see in these simulations that while the effect of congestion on travel time in the extensive margin is positive, the effect of congestion in the intensive margin is negative. The two in fact balance out in such a way that the travel time per trip remains very stable, increasing only very slightly over the 30 year period.

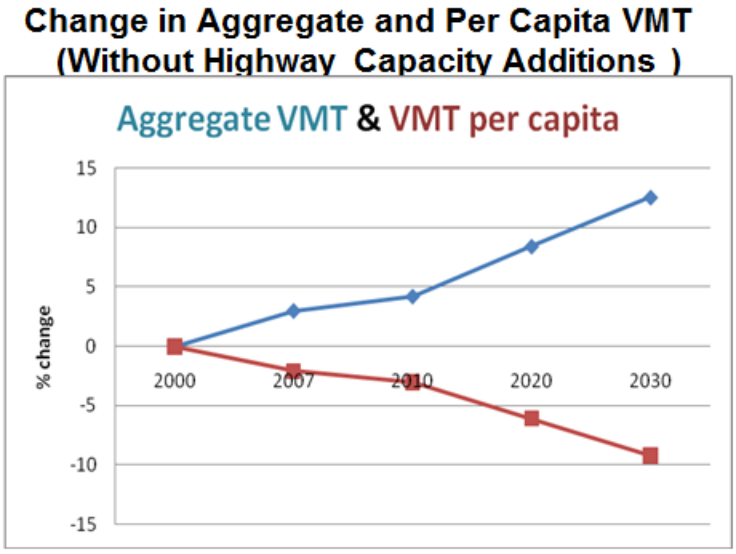


Figure 5

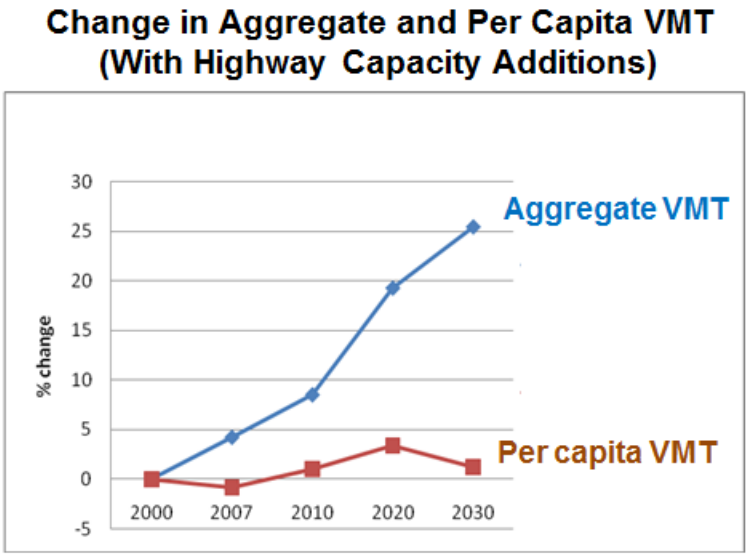


Figure 6

Figure 6 show the effect of expanding road capacity. If road capacity is expanded, then − not surprisingly − aggregate VMT rises more and per capita VMT decreases less. In this case, the intensive margin effect (people and jobs move closer to each other) and the extensive margin effect (more people drive) balance each other so that per capita VMT is essentially flat over the 30-year period. The effect of more congestion on personal travel time is offset by residences and jobs relocating closer to each other on average.

In the case of Paris, the objective was to use the model to evaluate the effects of investments in public transit planned for 2025 and 2035. These investments are aimed at

connecting the 10 growth Poles surrounding Paris to each other and to Paris. Unlike the traditional radially oriented public transit projects, these rail and subway investments are meant to make easier the peripheral circulation around the City of Paris. The City of Paris itself is essentially locked out of future redevelopment. As it is now there is very little empty land inside Paris and increasing the city's capacity to accommodate more jobs and residences would come at the expense of much congestion and taller buildings that would ruin the skyline. There is a vacancy rate of about 8% which means that some more job and population growth can be accommodated without increasing aggregate floor space. These assumptions and features of Paris were treated by the model.

The expectations from the planned public transit investments is that sprawl will be contained and that public transit ridership will increase and, more importantly, that as much as possible of the new job growth will concentrate in the 10 growth poles surrounding Paris. The results supported these expectations and are illustrated in Figures 7-9. The City of Paris is the white area in the middle. The improved accessibility does concentrate jobs in the growth poles and concentrates residences less so. Rents per square meter of floor space increase and more so in the growth poles than elsewhere. Rents also generally increase in the City of Paris (but this is not shown in Figure 9).

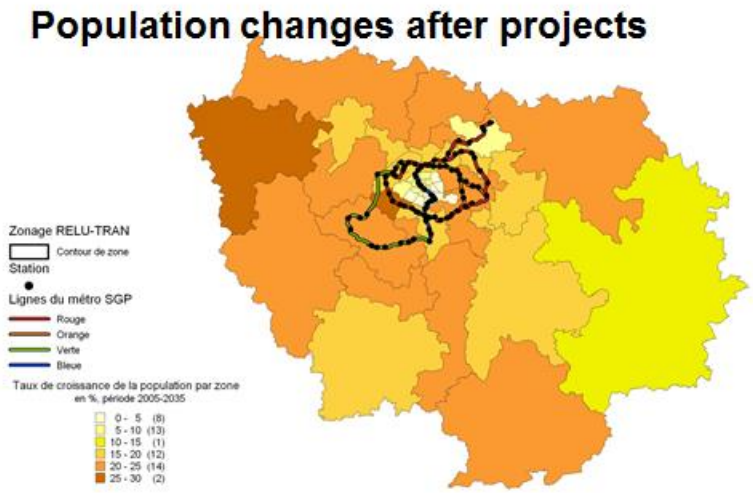


Figure 7

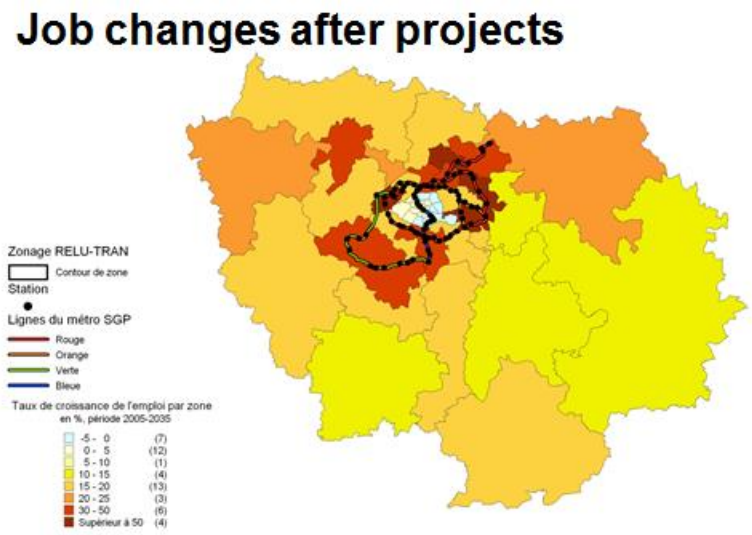


Figure 8

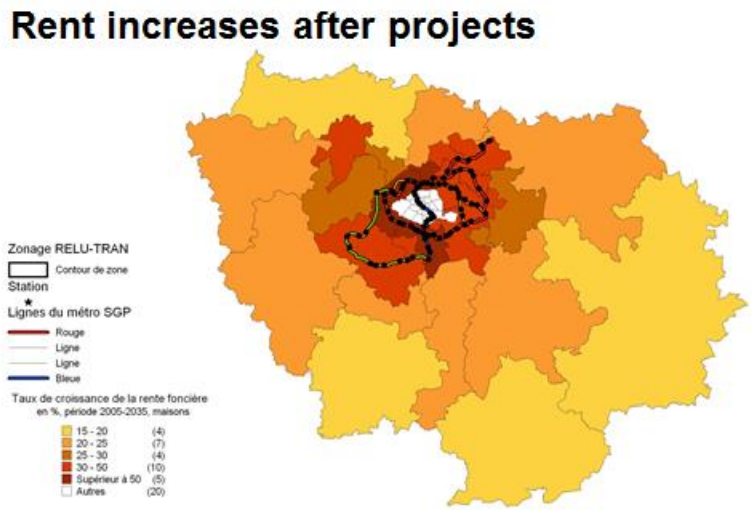


Figure 9

The Los Angeles version of the model has just been calibrated and we plan to start simulations sometime in the summer. As shown is Table 3, jobs in Los Angeles are very dispersed and public transit very unavailable. The 30 subcenters that contain the 30% of the jobs are shown in Figure 10. In this situation, and since the model cannot yet be simulated, we may conjecture about results we are expecting. The basic intuition is supported by what we learned in the Chicago simulations. Since almost all travel relies on road transportation, an increase in congestion that could come about by population growth would result in several possible responses. One is that new jobs would form new

additional job subcenters, the other that new jobs would concentrate in the existing subcenters. The basic feature of these changes would be to reduce the distance between residence and job on average. This then would be like what we observed in the case of Chicago but much more pronounced since the public transit margin is not sufficiently available.

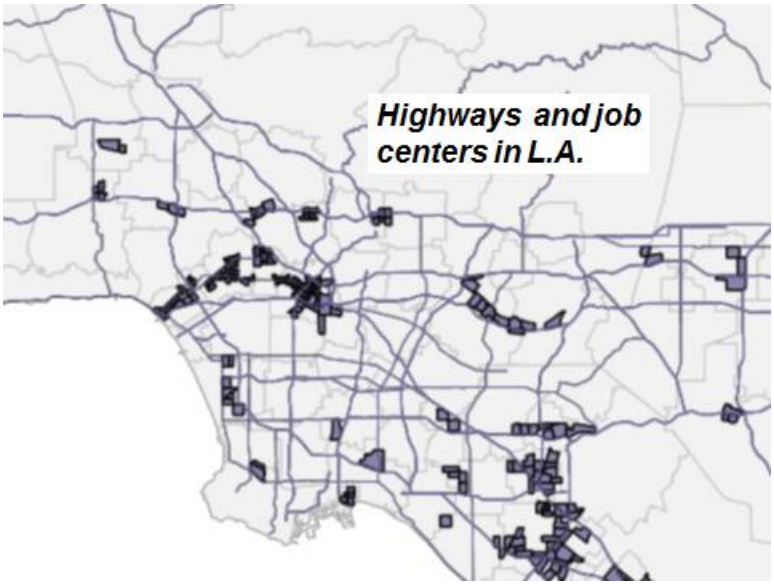


Figure 10

REFERENCES

Alonso, William, 1964. *Location and Land Use***,** Harvard University Press, Cambridge, Mass.

Anas, Alex, 2011. *Discovering the efficiency of urban sprawl*, Chapter 6 in the *handbook of Urban Economics and Planning*, Kieran Donaghy, nancy brooks and Gerrit Knaap, co-editors, Oxford University Press.

Fujita, Masahisa and Jacques Francois Thisse, 2002. *Economics of Agglomeration: Cities, Industrial Location and Regional Growth*, Cambridge University Press.

International Association of Public Transport, 2001. *Millenium Cities Database for Sustainable Transport*, Brussels.

Joe the Planner (website). Musings on Town Planning, Politics, Economics and Citizenship (with a Buffalo Perspective). http://joeplanner.blogspot.com/

Marshall, Alfred, 1890. *Principles of Economics*, MacMillan, London (1920, 8th edition).

Texas A&M Transportation Institute, 2012. *Annual urban Mobility Report*, http://mobility.tamu.edu/ums/

Pigou, Arthur, 1932. *The Economics of Welfare*, MacMillan, London (4th edition)

Vickrey, William, 1969. "Congestion theory and transport investment" *The American Economic Review* Vol. 59, No. 2, Papers and Proceedings of the Eighty-first Annual Meeting of the American Economic Association (May, 1969), pp. 251-260