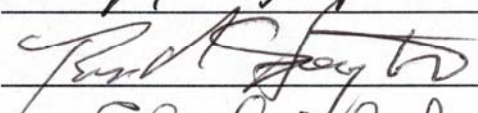IMPACTS OF TRAFFIC CONGESTION ON REGIONAL PRODUCTION
EFFICIENCY: CASES OF U.S. URBAN AREA

by

Junyang Yuan
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Public Policy

Committee:
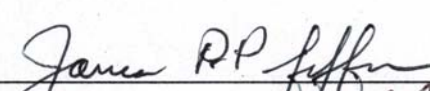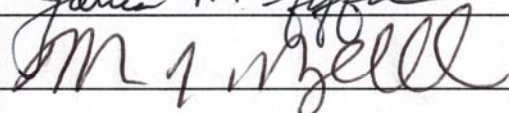
_____   Kenneth J. Button, Chair

_____   Roger R. Stough

_____   Edmund Zolnik

_____   Jean-Claude Thill, External Reader

_____   James P. Pfiffner, Program Director

_____   Mark J. Rozell, Dean

Date: _December 19, 2013_____   Spring Semester 2014
                                  George Mason University
                                  Fairfax, VA

Impacts of Traffic Congestion on Regional Production Efficiency:
Cases of U.S. Urban Area

A Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy at George Mason University

by

Junyang Yuan
Master of Management
Tsinghua University, 2006
Bachelor of Engineering
Tsinghua University, 2003

Director: Kenneth J. Button, Professor
Department of Public Policy

Spring Semester 2014
George Mason University
Fairfax, VA

DEDICATION

This is dedicated to my beloved family, for their unconditional love and support, especially to the memory of my mother. I hope I make her proud.

ACKNOWLEDGEMENTS

I would like to thank many people who have helped me on the path to make this happen.

I am deeply indebted to my advisor and mentor, Dr. Kenneth Button, for his patient and invaluable guidance as well as continuous encouragement throughout my dissertation preparation and writing in past several years. I would also like to sincerely thank my committee members, Dr. Roger Stough, who helped me so much in my graduate studies at George Mason University, and Dr Edmund Zolnik, for their efforts and time in helping me improve this work. Without their insightful comments, it is impossible for me to complete my dissertation. My appreciation is also extended to Dr. Kingsley Haynes and my external reader, Dr. Jean-Claude Thill at University of North Carolina, who have provided considerable efforts at various stages during this procedure. Additionally, I would like to thank Dr. Jonathan Gifford and Dr. Emilia Istrate, whose suggestions are really valuable for me towards the completion of this work.

I also really appreciate all faculty members in School of Public Policy, for their intelligence and support offered to me in my graduate studies. In addition, my special thanks are given to Beth Eck and Shannon Williams Hettler, whose warmhearted and considerate assistance kept me organized.

I am grateful to all my friends and colleagues at George Mason University, for their innovative ideas and comments on my research, as well as their precious friendship.

Last but not least, none of this would have occurred without the love and care of my family. My deepest gratitude goes to my father, who always supports and encourages me through this endeavor, and my dear wife, Li Ji, who always stands by my side in both good and bad time.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

IMPACTS OF TRAFFIC CONGESTION ON REGIONAL PRODUCTION
EFFICIENCY: CASES OF U.S. URBAN AREA

Junyang Yuan, Ph.D.

George Mason University, 2014

Dissertation Director: Kenneth J. Button

This dissertation aims to provide another way to evaluate traffic congestion's impacts on

regional economy besides of traditional method of calculating congestion costs in terms

of traffic delay and wasted fuel which has encountered considerable critiques on

debatable definition and measurement of both benchmark speed and value of time. To

additionally measure traffic congestion's indirect and long-term influences, an

econometric approach is applied in this study. Since traffic congestion should be

considered as a factor that may affect the efficiency of production procedure rather than a

direct input in production function, a two-step approach is implemented here. Initially, a

Translog production function model with three inputs is applied to calculate the Total

Factor Productivity (TFP) growth. Sequentially, regression analysis is conducted to detect

traffic congestion's impacts on TFP growth. To comprehensively investigate congestion's

influences, a stochastic frontier analysis is further introduced to decompose the TFP

growth into technical change, scale efficiency change and technical efficiency change.

The relationship between traffic congestion and each component of TFP growth is then

probed into. To verify the results from the parametric analysis, a non-parametric analysis

is also applied in this dissertation. The Malmquist productivity index as well as its components is calculated using the Data Envelopment Analysis (DEA), and then each of them is regressed based on the same information applied in the parametric analysis. Results derived from both methodologies are compared at the end. The data covers 31 large and very large American urban areas, boundaries of which are defined by the author using ArcGIS software, during the period from 1990 to 2009. Influenced by the conversion from SIC to NAICS in later 1990s, in order to keep data's consistency, the research periods is divided into two segments: 1990-2000 and 2001-2009, which also provides an opportunity to make comparisons between these two periods. After considering various econometric and statistical issues, such as stationarity, spatiality, multi-collinearity, heteroscedasticity, auto-correlation, cross-sectional dependence, and endogeneity, a SYS-GMM (System General Moment Methodology) is implemented and corresponding results show that traffic congestion has significantly negative impacts on TFP growth during 1990 and 2000, while this impact becomes positive and still significant in the successive period. Moreover, congestion didn't affect technical change significantly in both periods. Its impact on technical efficiency change became trivial after 2000, though it negatively influenced this component in the previous period. For scale efficiency change, traffic congestion seems like a positive contribution continuously. Both parametric and non-parametric analysis provide similar results, though still tiny differences exist. Three possible explanations are provided correspondingly: (1) adaptation and adjustment to congestion in a long run; (2) "Hidden-behind" factors of traffic congestion; and (3) Redistribution effects in urban areas. This

dissertation also shed light on various ways in mitigating urban traffic congestion presently, and emphasized the pros and cons of implementing traffic congestion pricing which is economically welcomed, but politically objected (in general) in policy implication.

CHAPTER 1   INTORDUCTION

1.1 BACKGROUND

Growth of urbanization is stable in past thirty years (see Figure 1-1), and this

trends may remain in following several decades. Reported by the *2011 Revision*

published by the World Bank[1], on average over 50% of the global population resides in

urban areas, and it's expected that this ratio will grow up to 67% in 2050. Particularly, in

more developed regions, more than 77% of population lived in urban areas in 2011, and

this ratio may climb up to 99% by 2050. During the same period, the ratio of urbanized

population in less developed regions may increase to 64% from less than a half.

One direct result caused by urbanization is the higher population density in urban

areas, and inevitably urban residents' lives are influenced in various aspects when people

have more and more friends or competitors (in most time), one of which is commuting.

Hence, growing urbanization is usually accompanied with traffic congestion that gives

rise to many complaints about great inconvenience in daily lives and high costs in

business operations. So far, traffic congestion has become an inevitable conundrum in

urban areas, whether in Washington DC, New York, Los Angles, Beijing, Frankfurt, or

Seoul. This issue has attracted extensive and in-depth attention from governors,

economists, and engineers to investigate and analyze its causes, types, measurements, and

---

[1] United Nations. World Urbanization Prospects The 2011 Revision Highlights. Retrieved from
http://esa.un.org/unup/pdf/WUP2011_Highlights.pdf

possible solutions. One of these concerns is about the magnitude of congestion costs to

regional economy.
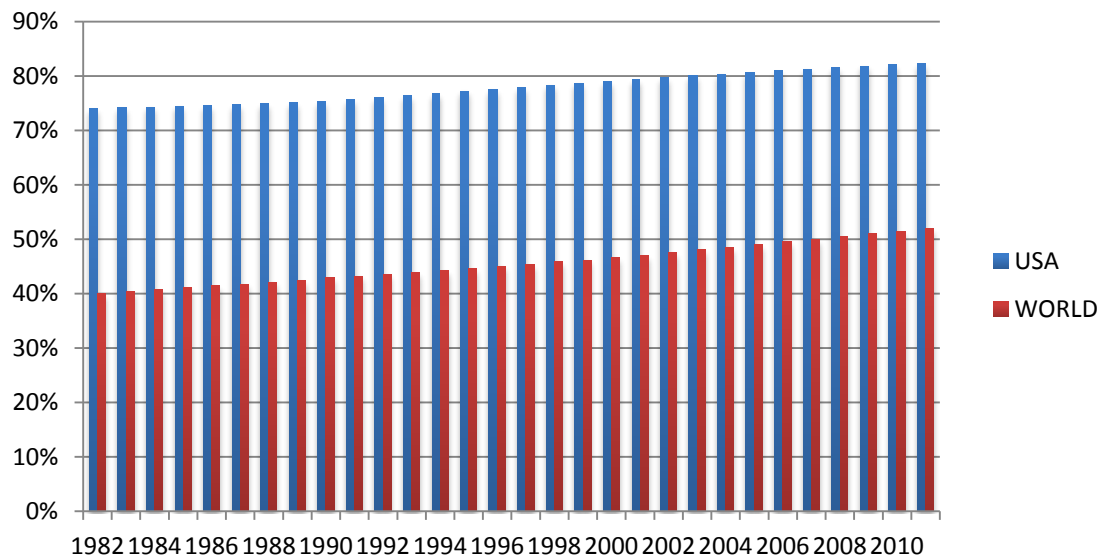
Figure 1-1 Growing Urbanization in U.S. and the World
Data Source: World Bank Group.

The estimation of congestion cost is important for local governments, since it

helps planners better understand the real 'harm' of traffic congestion, allowing them to

implement appropriate measures to deal with this problem, especially in conditions of

currently severe limited budgets. However, there has not been an undisputed method to

quantify congestion costs. At present, a common approach is to calculate the most

obvious costs of traffic congestion in terms of time delay and fuel waste. The product of

total delay hours and the time value plus the product of the amount of additional fuel

consumed and the unit fuel price provides a straightforward measurement of congestion

costs. Using this method, the latest *Annual Urban Mobility Report 2012* issued by the

Texas Transportation Institute (TTI) reported that the overall American congestion costs

in 2011 was $121 billion, including 5.5 billion hours of delayed time and 2.9 billion

gallons of wasted fuel.[2] Figure 1-2 displays the trend of national congestion costs

(measured in 2011 dollars), total delay (in hours) and fuel wasted (in gallons) between

1982 and 2011. Obviously, the total traffic congestion costs increased continuously

between 1982 and 2006, and then its growth slowed down in the latest economic

recession with continuously dropping down and touching the bottom in 2008. In

succession, it increased again with the slow recovery of economy, and at present traffic

congestion returns to the level of 2004, while keeping almost stagnant in recent years.

---

[2] The delay cost is an estimate of the value of lost time in passenger vehicles and the increased operating costs of commercial vehicles in congestion. The average cost of time is $16.30/hour, and the commercial vehicle operating cost is $88.12/hour (including both truck travel time cost and operating cost, but excluding diesel cost). The passenger vehicle fuel cost is calculated by multiplying additional fuel consumption with average gas price in each state. The additional fuel consumption formula is also shown in the report. The benchmark condition is the free-flow travel condition. Values for gasoline and diesel are reported separately by American Automobile Association (AAA).
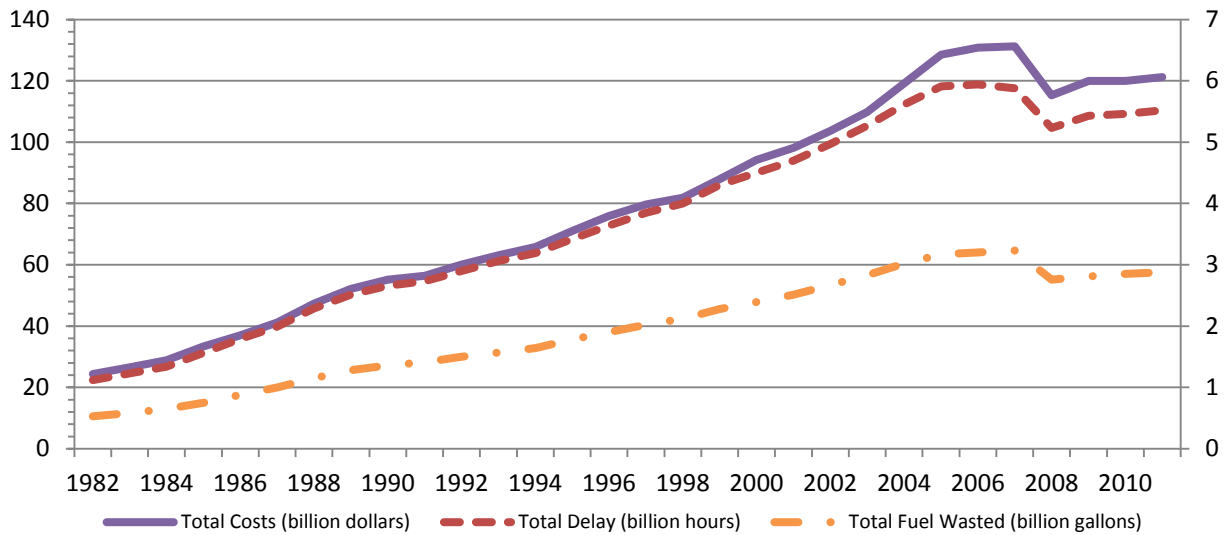
Figure 1-2 National Congestion Trend From 1982 To 2011
Data Source: Annual Urban Mobility Report 2012. Texas A&M Transportation Institute.

Other scholars also apply similar methodologies to measure congestion costs.

Delucchi (1998) estimates external costs of U.S. road congestion, including time delay

(containing both monetary and non-monetary costs)[3] and increased fuel consumption,

totaled from $55.9 to 223.6 billion in 2011 dollars.[4] Winston and Langer (2006) find that

congestion costs to the nation's motorists in 72 largest urbanized areas of the U.S. in

1996, amounted to roughly $35.8 billion (in 2011 dollars), including both time delay and

additional fuel consumption. Douglass Lee of the Volpe National Transportation Systems

---

[3] Monetary cost items can be traded in real markets and hence valued directly in dollars, such as foregone paid work, while non-monetary cost times are those travel delays with unpaid activities.
[4] The low and high bounds just provide the range where the congestion costs may be located, and don't have exact possibilities. The author references various investigations and set up a broad range to cover different parameters produced by those researchers.

Center provides two distinct estimates: one is $141.5 billion (in 2011 dollars) of total

delays from traffic congestion in the U.S.; and the other one is only $15.7 billion (in 2011

dollars) of the corresponding economic loss. (Roth, 2006) The lower end is based on

drivers' willingness to pay[5] for increased traffic speed, rather than the time value

estimated for calculating the upper end. Lee points out that drivers would be prepared to

accept much of the delay, rather than paying for its elimination.[6] Nevertheless, the

willingness to pay may be not equal to real payment, since the bias between expectation

and implementation always exists. In addition, drivers normally have various preferences

for travel speed and these preferences may change under different conditions, which

makes it impossible to take a comprehensive survey to quantify each road user's delay

cost precisely.

1.2 MOTIVATION

Significant differences in cost estimates arouse suspicion of their accuracy. The

basic formula to calculate time delay cost is $C = (T_a - T_f) \cdot V_t$, where, C is the time

delay cost, T_a is the actual travel time and T_f represents the travel time under free-flow

---

[5] The willingness-to-pay is a partial equilibrium, based on the assumption of no income effect on the commodity under consideration. For individual, it makes sense, since the interviewee doesn't consider the disposable income after the payment for more convenient commuting. Albeit, when considered in the whole society, the considerable amount of payment for road using will reduce the consumption level in other aspects, and thus may produce income effects that will further influence his/her willingness-to-pay in transportation. As a result, the survey willingness-to-pay doesn't really make a lot of sense. (Hayashi, 2012)

[6] In reality, many commuters would like to move slowly in clogged traffic flow or leave early (late) in the morning (afternoon) from home (office) to avoid congestion rather than to pay for driving on less congested lanes during rush hours as long as they are not in urgency. One reason may be from people's traditional recognition: roads are public goods that should not be charged. In addition, for many commuters they can not be rewarded even if they could save time in traffic as long as they arrive at office on time. In other words, scholar's estimate of time value usually exceeds commuter's actual willingness to pay.

condition. V_t denotes the value of time. However, this methodology has received considerable critics in spite of its popularity.

First, there isn't an identical criterion to define the free-flow speed. Federal Highway Administration sets the 85th percentile speed in the previous three months during weekday off peak times (9am-4pm, 7pm-10pm) and weekend/holiday times (6am-10pm), not exceeding posted speed limits or 60 mph where the posted speed is unknown, as the free-flow speed. Texas Transportation Institute defines free flow as the traffic speeds in light traffic hours (e.g., 10 p.m. to 5 a.m.), with an upper limit of 65 mph on freeways and no limit on arterial streets. Various definitions are not the only factor resulting in complexity. The posted speed limit may be altered on the same road. For example, there are at least three speed limits, 55, 60 and 70 mph, on different I-66 segments located in Northern Virginia. In addition, the speed limit is also changeable for the same road segment. In 2010, three-fifths of the miles on I-66 increase the speed limit from 65 to 70 mph. In 2013, the Ohio Department of Transportation also plans to increase speed limits on rural interstates from 65 to 70 mph and allows some two-lane highways to be changed from 55 mph speed limits to 60 mph. Accordingly, the amount of traffic delay will change once the benchmark is altered.[7]

Second, there also exist considerable debates on quantifying the value of time. A simplified approach that uses the average or prevailing wage rate (Hartgen, 2007) is

---

[7] Besides above comments, traffic congestion is also regarded as both a physical and a relative phenomenon by the OECD/ECMT report (ECMT, 2007). The former means that vehicles impede each other's movement, while the latter relates to user's expectation compared with road system performance, e.g., at which speed drivers will feel congested or feel 'free'. Some drivers prefer to driving at least 70 mph on a highway, while some others may think 50 mph has been fast enough. In this case, free-flow speed is a kind of subjective recognition.

broadly questionable, since the economic loss for the time delay varies based on trip purposes. For those non-work related trips, such as shopping, recreation and driving kids to school, travelers wouldn't be paid any money even if they could save time from avoiding the traffic jam. For work related trips, employees may not be paid extra money for their early arrivals, while late arrivals might result in severe results. Thus, it's too arbitrary to quantify time cost using the average wage. Other scholars develop this idea and try to provide more accurate estimation. Small and Yan (2001) recommend a 50% of the wage as the value of time for work related trips. Gwilliam (1997) suggests that for work (paid) trips, the recommended time value is 133% average wage, and for other personal (unpaid) travels, adults assume a 30% household hourly income as the value of time, while kids also assume a 15% household hourly income. Obviously, significant differences among quantifying value time undoubtedly result in distinct estimates.

Third, in essence, free-flow condition is not a reasonable benchmark. Bradford (2009) argues that "there is no realistic, hypothetical state of the world in which we would experience perfect, free-flow traffic everywhere", because it would be impractical to build enough roads complying with continuously growing traffic demand, and temporary free-flow speeds on roads would eventually entice more drivers. Hence, there is not such a Utopian case that the economically optimal level of congestion is zero in modern urban area (Downs, 2004). Correspondingly, it seems unreasonable that every deviation from the zero-delay ideal can be described as a cost. In fact, congestion could only be mitigated rather than eliminated, as inevitable incidents always occur. About 50% of traffic congestion can be classified as non-recurring congestion that is impossible to

remove in any situation. The benchmark of free-flow condition is too optimistic to be true. Cortright (2010) criticizes this idea and comments that Texas Transportation Institute overestimates true congestion costs by about 300%.

Table 1-1 Sources of Congestion and Their Contributions

Category	Percentage of Total Congestion
Bottlenecks	50%
Traffic accidents	25%
Work zones	15%
Bad Weather	10%
Poor signal timing	5%

Source: American Highway Users Alliance, *Unclogging America's Arteries,* 1999–2004.

Last, congestion costs measured using the above approach only quantify their direct impacts on time delay and wasted fuel, while indirect impacts on regional economy are not concerned at all. Traffic congestion may reduce the mobility of resources and even accessibility to products and services, and thus might further affect regional economic performance. Environmental costs are not considered, either. Non-measurable indirect costs enhance the difficulty to evaluate congestion's impacts comprehensively.

1.3 ESEARCH QUESTIONS

Inaccurate and incomplete estimates of congestion' impacts may mislead governments' decision making in solving traffic congestion. Overestimation of congestion costs may lead to unnecessary expenditure, while underestimation would induce the insufficient supply of transportation service. Both aftereffects result in inefficient resource allocation, which may negatively contribute to regional economic performance. Under the contemporary condition of limited budget encountered by governments, this issue seems more important for appropriate resource distribution. In

addition, whether traffic congestion significantly harm regional economy is still

debatable. Taylor (2002) argues that traffic congestion is just an unfortunate consequence

of prosperity and a drag on otherwise high levels of accessibility, not a cause of economic

decline and urban decay. In reality, more developed urban areas are usually associated

with severer congestion. How to mitigate traffic congestion effectively and efficiently is

another important issue. Cost-effective methods are crucial, but also complex. Hence, this

thesis mainly investigates the following two questions.

    1. Is traffic congestion a significant factor that negatively influences regional

economic efficiency?

    2. Which measures are preferred in transportation policy to alleviate traffic

congestion effectively and efficiently?

1.4 RESEARCH CONTEXT

This research is conducted at the urbanized area level across the United States. In

general, congestion is severe in areas with high population density and highly developed

business activities. The excessive traffic demand relative to limited transportation

capacity is the direct cause of traffic congestion. Hence, urban areas are more appropriate

research subjects, rather than the nation or states. Econometric models will be applied to

test the relationship between traffic congestion and regional economic efficiency. To

validate the result, a non-parametric method (Data Envelopment Analysis) will be

applied.

1.5 THESIS STRUCTURE

Following the introduction in this chapter, Chapter II reviews relevant studies to further demonstrate the literature gaps. In succession, Chapter III details the statistical methods used in regression as well as variables and data source. Chapter IV discusses traffic congestion's impacts using the parametric analysis, while Chapter V focuses on the non-parametric method to validate previous results. Chapter VI introduces and discusses various experiences of alleviating traffic congestion as well as their advantages and limitations. Chapter VII presents a summary of the research, concluded with key findings and its unique contribution. Discussions on future research are also included.

CHAPTER 2   LITERATURE REVIEW

To some extent, congestion reflects the performance of local transportation system. To better understand its impacts, the tie between transportation and regional economy is established firstly. This review discusses transportation's role in regional economic development in terms of its influences in firms and labor market, relevant empirical analysis and conclusions as well as methodologies. In sequence, the viewpoint is shifted to traffic congestion. Studies on its definition, measurement and impacts on economic efficiency are remarked.

2.1 TRANSPORTATION AND REGIONAL ECONOMY

2.1.1 Theoretical Analysis

An integrated economy is composed of interactive activities in production, distribution, and consumption (Say, 1836). Transportation plays a critical role in connecting production, commodity circulation, and social economic turnover. In an economic system, an efficient transportation system helps enhance the circulation and allocation of resources, and thus stimulates economic growth.

Transportation determines movements of physical capital (raw materials, intermediary products, equipment, and products), and thus affects firms' behaviors in location selection and logistic operation. Based on industrial location theory, firms in early periods had to locate themselves where the sum of transportation costs of raw materials and final product and other costs was a minimum, since the former one occupied a great share of the total cost. (Weber, 1909/1929) Later, development of transportation technology significantly reduced transportation costs as well as its

proportion in total costs. Correspondingly, firms adjusted their strategies in location

selection, and benefited greatly owing to less constraint in geography.

The dropping transportation cost doesn't imply its less importance. The further

development of social division of labor makes the evolvement of logistics still important

in current firms' operation. In 1960s, the management guru Peter Drucker stated that

logistics was the last great unexplored continent of business. (cited in Allen, 1997, p.

109) In 1980s, the concept of supply chain management (SCM) was developed to

illustrate the process from extraction of the raw materials to scheduling of final assembly,

including the movement of work-in-process inventory. (Smock, 2003) In SCM's Seven

Rights of Fulfillment[8], right time, right place and right cost are directly related to the

transportation system which performance may determine delay costs, inventory costs,

reliability costs, just-in-time processing costs, and so on. (Shirely and Winston, 2004) At

present, more and more retailers abandon conventional in-store inventory to save spaces

for cutting rents, and their daily operations heavily rely on timely delivery of shipment.

Any unpunctual delivery may result in potential economic loss.

Besides additional logistics costs, companies' operations would also be affected

by transportation system's performance in product/service market and labor market.

Transportation efficiency determines the range in which products/services could be

delivered to or obtained by consumers within a reasonable period of time. Greater

mobility is beneficial for enlarging the market size, so that companies may sell more

---

[8] Seven rights: 1) The right product; 2) To the right customer; 3) At the right time; 4) At the right place; 5) In the right condition; 6) In the right quantity; and 7) At the right cost.

products and encounter more business opportunities. Meanwhile, the distance in which employees could commute to work each day is also determined by transportation condition. The more efficient the transportation system, the longer the distance from residential places to work could be achieved within the same time. (Goodbody Economic Consultants, 2003) As a result, firms have better accessibility for labors who reside far away, which implies that more potential qualified employees may match job positions.

In Rodrigue's book (2013), impacts of transportation on economic growth are discussed in five aspects: (1) Networks, enabling new or existing interactions between economic entities; (2) Performance, improving existing passenger's and freight's movements in time and cost attributes; (3) Reliability, improving punctuality as well as reducing loss and damage; (4) Market size, accessing to a wider market size so as to improve economies of scales in production, distribution and consumption; and (5) Productivity, growing through the access to a larger and more diverse base of inputs, such as raw materials, parts, energy and labor, and broader markets for diverse outputs -- intermediary and finished goods. Rodrigue comments that an efficient transportation system could conduce to geographical specialization, increased competition, large scale production and increase land value, and positively contribute to economic growth after consideration of some disadvantages, such as air quality, noise, and land take.

In consequence, an efficient transportation system is deemed as a catalyst for regional economy, as it intensifies connections among producers and consumers, and induces more business opportunities. In addition, by exploiting geographical comparative advantages as well as developing economies of scale and scope, a more efficient division

of production is achieved. Hence, production efficiency may be enhanced with better

resources' allocation owing to greater mobility and accessibility.

2.1.2 Empirical Analysis

In past three decades, economic impacts of transportation have been investigated

richly. The following table lists brief information about these studies in terms of various

aspects, including types of economic benefits, research units, methodologies, and

measurements of transportation infrastructure.

Table 2-1 Impacts of Transportation Infrastructure on Economic Development

Study	Scope	Model	Type	Transportation Measure	Results [*]
Albala 2004	Chile and Mexico regions 1950–2000	gap approach using a Leontief production function	regional output	infrastructure capital stock (transportation, communication, general purpose)	no effect (in Chile) + (in Mexico)
Aschauer 1990b	48 U.S. states 1960-1985	production function	per capita income	vehicle density, highway capacity, and pavement quality	+
Aschauer 1989	United States 1949-1985	production function	productivity of private Sector	transportation, water, sewer, gas and electricity	+0.24 +0.39 (aggregated public capital stock)
Berndt and Hansson 1992	Sweden 1960-1988	variable cost function dual to production function	private sector costs	transportation, water and sewer, electricity	-
Boarnet 1996	55 California counties 1969-1988	production function	private output	highway infrastructure in a region plus some fraction of highway infrastructure in neighboring	+ in own counties - in neighboring counties
Bosca, et al. 2002	Spain, regions 1980–1993	generalized Leontief cost function	private sector costs	infrastructure capital stock (transportation, communication, general purpose	- 0.08
Bruinsma, Rienstra, and Rietveld 1997	regions in Netherlands	a reference region approach, a regional labor market approach, a survey among entrepreneurs	employment growth; firm growth	one major new highway	no effect (employment growth) + (firm growth)

Study	Scope	Model	Type	Transportation Measure	Results [*]
Cohen & Morrison Paul 2004	48 US, states, manufacturing 1982–1996	generalized Leontief cost function with spatial spillover index	manufacturing costs	public highway stock	infrastructure investment reduces own costs and increases cost reducing effect of adjacent states
Conrad and Seitz 1994	3 sectors, West Germany 1961-1988	dual cost function model	sector output and costs	proxy for transportation Infrastructure including traffic and energy	- + (total factor productivity)
Coughlin, Terza and Arromdee 1991	48 U.S. states 1981-1983	conditional logit model	foreign direct investment in manufacturing	highway miles per square mile of state land	+
Crihfield and Panggabean 1995	277 Standard Metropolitan Statistical Areas 1960-1977	production function	output	highway miles	+, but marginal contribution is no more than other forms of investments.
Dalenberg and Partridge 1995	28 metropolitan areas 1966-1981	general translog function	employment	highway spending/per income	-
Demetriades and Mamuneas 2000	12 OECD countries, 1972–1991	quadratic cost function	national output	public infrastructure capital stock	+ 0.36 (UK) to +2.06 (Norway) long-run rates are much higher than short-run rates but declining over time
Duffy and Eberts 1991	28 metropolitans, 1980-1984	two-equation system model	per capita income	transportation, water and sewer, public hospitals	+ 0.094

Study	Scope	Model	Type	Transportation Measure	Results [*]
Fernald 1999	9 industry groups United States 1953-1989	production function	industry productivity, industry output	road stock	+1.4 (before 1973) +0.4 (after 1973) congestion becomes empirically important after 1973
Garcia and McGuire 1992	48 U.S. States 1970-1982	production function	gross state product	highway miles per square mile	+ 0.045
Garcia, McGuire, and Porter 1996	48 U.S. States 1970-1983	production function	gross state product	highway capital	+ 0.127 no effect on growth rate
Gkritza et al. 2008	Indiana	simultaneous equation model	employment, income, output	characteristics of 117 highway projects	+
Harmatuck 1996	Untied States 1949-1985	transfer function model	gross national product	Non-military public investment	+ 0.03
Holleyman 1996	369 four-digit SIC industries 1969-1986	translog cost function model	manufacturing costs	highway capital stock	
Holtz-Eakin and Schwartz 1995	48 U.S. States 1971-1986	new classical growth model	productivity growth	highways, water and sewer, gas and electricity	no effect
Jones 1990	48 U.S. states 1964-1984	disequilibrium-adjustment model	employment, income, investment	highway spending per capita	+ (some periods) - (other periods)
Lynde and Richmond 1993	Untied Kingdom 1966-1990 (quarters)	dual function of production technology and cost function	labor productivity growth rate in manufacturing sector	nonresidential public capital	+ (prior to 1979) - (after 1979)

17

Study	Scope	Model	Type	Transportation Measure	Results [*]
Moomaw, Mullen, and Williams 1995	48 U.S. states 1970, 1980, 1986	production function	gross state product	highway capital stock	+ 0.001~0.027 public capital's impacts are confined by other regional elements
Morrison and Schwartz 1996a	6 New England states, 1970-1987	flexible variable cost function and incorporated with short-run fixity of both private and capital stock	manufacturing costs	highways, water and sewer	-
Morrison and Schwartz 1996b	48 U.S. states 1970-1987	generalized Leontief variable cost function model	manufacturing costs	highways, water and sewer capital	- 0.10
Munnell 1990a	Untied States 1949-1987	production function	productivity of private Sector	transportation, water and sewer, gas and electricity	+0.21~0.39 +0.31~0.37(total nonmilitary public capital)
Munnell 1990b	48 U.S. States 1970-1986	production function	gross state product	highway capital stock	+0.06 (+0.07 in Northeast and +0.36 in South)
Nadiri and Mamuneas 1994	12 two-digit U.S. manufacturing industries, 1955-1986	cost function	industry costs and labor demand	public financed infrastructure	- 0.11 to – 0.21 (costs) - (labor demand)
RESI 1998	9 industries, Maryland, 1982-1996	cost function	industry costs, output	highway investment	- 0.05 (costs) + 0.06 (output)
Seitz 1993	31 manufacturing industries, West Germany 1970-1989	generalized cost function using duality theory	productivity of private capital	length of motorway network and capital stock of total road network	+

Study	Scope	Model	Type	Transportation Measure	Results [*]
Seung and Kraybill 2001	Ohio, 1990	dynamic general equilibrium model	gross state output	public capital stock and investment	public investment benefits the output growth, while the magnitude depends on the public capital stock
Stephan 2003	11 West German regions, 1970-1996	production function	output of manufacturing sector	infrastructure capital (transportation and communications)	+ 0.38 (first differences) + 0.65 (log levels)
Tatom 1993	United States 1949-1990	Granger causality test	Productivity of private sector	public sector capital	no effect (either the growth rate of public capital stock or level of public sector investment)

*: + implies positive and significant impacts; - denoted significant but negative impacts; no effect represent statistical insignificance. The number behind the sign indicates the elasticity of dependent variable to transportation variable.

2.2 RESEARCH UNITS

Units of research analysis include counties, metropolitans, states, and nations. Studies at national level are more likely to achieve positive and larger coefficients (Aschauer, 1989; Munnell, 1990a; Demetriades and Mamuneas 2000; Stephan 2003), compared with studies using state, MSA (Metropolitan Statistical Area), or county level data. One criticism refers to the lack of specification in researches at macro level that omits regional unique features (individual heterogeneity), such as residential travel pattern, industry structure, and urban pattern, which leads to the aggregation bias[9] in statistics (Elliott et al., 2008; Pesaran, et al., 1989). On the other hand, a county may be too small to cover integrated effects of a local transportation system, except those large ones located in mid-west and west, as highly developed transportation network connects various districts and influences regional economy in a broader range. For example, an interstate highway serve multi-states, and a local transportation infrastructure may serve several counties or even the whole urban area. Definition of research subjects should be based on specific research questions.

2.3 MEASUREMENT OF TRANSPORTATION INFRASTRUCTURE

How to measure transportation infrastructure also affects the explanatory power of results. There are three typical types of measurement: total non-military (residential) public capital stock, aggregated core infrastructure capital stock (transportation, communication, water, sewer, gas and electricity), and pure transportation infrastructure

---

[9] Aggregation of data may reduce the amount of variance in the data that would both eliminate important information and falsely inflate the value of R^2 yielded by the regression analysis (Walker and Catrambone, 1993).

measurement, such as highway density, highway/rod capital stock, and specific

transportation projects. Because of aggregation bias, neither non-military public capital

nor aggregated core infrastructure capital could accurately measure transportation

specifically. Introduction of other types of public capital makes it difficult to differentiate

transportation's influence from aggregated impacts. As a result, either monetary or non-

monetary highway/streets capital proxy may provide more pertinent and useful

information. Therein, monetary measurement is preferred, for it covers more information

of each project (construction, maintenance, capital outlay, etc.) and more importantly, is

comparable. Though physical measures could avoid problems inherent in the estimation

of monetary capital stock (e.g. data's unavailability under some conditions or

questionable quality of estimate based on the perpetual inventory technique), they still

have limitation in application. The comparability among various types of transportation

infrastructure is poor, and it is impossible to make simple mathematic operation on

physical measures belonging to different categories.

2.4 METHODOLOGIES

Four well-known techniques applied in learning transportation infrastructures'

impacts are econometric approach, cost-benefit analysis (CBA), input-output model, and

computational general equilibrium (CGE) model, with respective advantages and

limitations that determine when, where and how they should be used.

2.4.1 Econometric Analysis

Econometric analysis is the most frequently applied approach in this field. Based

on various economic theories, such as neo-classical economic theory and endogenous

growth theory, integrated with statistical methods, econometric analysis explores the relationship and/or causality between transportation and regional economic performance, and further provides theoretical support for government decision makers. Since the evolvement of neo-classic economics (e.g. Solow Growth Model), mathematic models and production functions have been integrated to analyze economic growth. The generalized new-classical economic production function model is expressed as: $Y = f(X_1, X_2, \ldots, X_n)$, where Y measures the output, $X_i(i \in (1, 2, \ldots, n))$ represent n inputs, and $f(\cdot)$ is the production function with positive but diminishing marginal products. One of the most widely used specific forms is the constant elasticity log-linear specification, also known as the Cobb-Douglas production function, which can be written generally in the form as (Aschauer, 1989): $Y = \tilde{A} K^\alpha L^\beta P^\gamma$, where Y is the output. K, L and P indicate the real stock of private capital input, labor input and real stock of public capital input, and α, β and γ are output elasticities of labor, private capital and public capital, separately. $\tilde{A}$ is the Hicks-neutral level of technology.

However, the production function omits input prices which could affect factor utilization, and the minimal structure imposed on the data is only one explanation for the wide range of estimates (Nadiri and Mamuneas, 1996). Besides, the production function is treated with the assumption that the management of firms acts to maximize economic profits under the condition of perfect competition, i.e. all firms perform efficiently. Whereas, in reality, the market is not perfectly competitive, and X-inefficiency always exists, violating the assumption of production function and resulting in biased results.

The cost or profit function approach offers the possibility to trace the effects of infrastructure investment on a firm's cost-minimizing or profit-maximizing behavior, such as adjustment of combination of inputs (i.e. employment, materials and private capital), in producing a proposed amount of output for a given level of technology[10]. Prices are deemed as the only exogenous variables since they are market determined beyond the immediate control of the firm (under the assumption of free market). The stock of transportation capital is considered a fixed and free input that influences production technology, because the changes in transportation cost may affect firms' behaviors to achieve their cost-minimizing strategy by adjusting their production as well as requirements for inputs. The general cost function model takes a following form: $C = C(w, r, z, t, Q, P)$ in which w, r, and z are the prices of labor, private capital, and intermediate inputs separately, t (time) represents a proxy for technical change, Q is the output, and P indicates the stock of transportation infrastructure capital within a jurisdiction. The cost function is derived by minimizing the firm's production cost: $C(Q) = wL(Q) + rK(Q) + zM(Q)$, subject to its production function: $Q = f(L, K, M, t, P)$, where L, K and M are labor, private capital and intermediate inputs separately. Normally, the Shephard's lemma, which states that the optimal input demand equation (L^*, K^* and M^*) can be obtained by partially differentiating the cost function

---

[10] In cost function, the output is restricted to be fixed, while in revenue function, the inputs are fixed. (Coeli, et. al, 2005)

with respect to the price of each input in production function, is applied to derive the

conditional input demand functions.[11] (Jiwattanakulpaisarn, 2008)

The adjustment effects of transportation infrastructure on the input demand can be

estimated by differentiating the optimal demand function with respect to P. Take

$\partial K^*/\partial P$ as an example, if $\partial K^*/\partial P > 0$, it indicates the private capital and transportation

capital are complements; if $\partial K^*/\partial P < 0$, they are substitutes, whereas transportation

infrastructure is neutral with respect to private capital if $\partial K^*/\partial P = 0$. For empirical

implementation, a Cobb-Douglas specification could be used in the production function

side, but to avoid the restriction of a unitary elasticity of substitution among inputs, more

flexible functional forms are applied, such as the translog form [12] (Seitz and Licht, 1995)

or generalized Leontief functions [13] (Seitz, 1993; Cohen and Paul, 2004).

Though the cost function approach could accommodate more factors in

estimation, especially price variables, it is still deficient. Deno (1988) points out that

input demand functions are the conditional demand for one input, holding other inputs

constant. Since the improved transportation infrastructure could reduce the cost of firms,

which may lead to an expansion of output, the cost function approach is not capable of

capturing the mechanism by which transportation infrastructure may have indirect effects

---

[11] $L^* = \frac{\partial C(w,r,z,t,Q,P)}{\partial w} = L(w,r,z,t,Q,P)$; $K^* = \frac{\partial C(w,r,z,t,Q,P)}{\partial r} = K(w,r,z,t,Q,P)$; $M^* = \frac{\partial C(w,r,z,t,Q,P)}{\partial z} = M(w,r,z,t,Q,P)$

[12] Translog (transcendental logarithmic) production function is a generalization form of Cobb-Douglas Production function with the following form: $lnY = \alpha + \sum_i \beta_i lnX_i + \sum_i \sum_j \delta_{ij}(lnX_i)(lnX_j)$, where X_i, X_j are inputs and Y is the output. (Griffin, et. al 1987)

[13] Leontief production function, also called fixed proportions production function, assumes that the inputs will be used in fixed (technologically pre-determined) proportions, as there isn't substitutability between inputs. It has the form of $Y = \min(\beta_1 X_1, \beta_2 X_2, ..., \beta_n X_n)$ $\beta \gg 0$. $\beta_i, i \in (1, ..., n)$ is technologically determined constants. A further generalized Leontief production function could be represents as the following equation: $Y = \sum_i \sum_j \delta_{ij} X_i^{1/2} X_j^{1/2}$ (Griffin, et. al 1987)

on the demand for private inputs through the output expansion effect. Hence, this method may underestimate transportation infrastructure's real effects. (Jiwattanakulpaisarn, 2008)

To relax the assumptions of either fixed output or inputs, a profit function is estimated with the form of $\max[pQ - (wL + rK + zM)]$, subject to the production function $Q = f(L, K, M, t, P)$, and can be further expressed as: $\Pi = \pi(p, w, r, z, t, P)$, where p is the price of output. According to the Hotelling's lemma[14], the first-order conditions result in unconditional demand functions for labor, private capital and other intermediate private inputs. Since the price of output is introduced in the equation, the estimation could capture the output expansion effect caused by the change of transportation infrastructure, though profit function model is still a partial equilibrium.

In summary, both the production function and the cost/profit function models make it possible to investigate whether transportation influences economic performance in terms of output, cost or profit. The requirement of price variables in cost or profit function models makes them more frequently applied in firm, sector or industry research at micro level. Comparatively, at macro or meso level, measurement of wage, or intermediate good value may induce aggregation bias, since each industry possesses its own features. Accordingly, production function model has greater applicability at macro level research. Moreover, in production function model, transportation infrastructure

---

[14] A counterpart of Shephard's lemma in cost function, could be express as: $Y(p) = \frac{\partial \Pi(p)}{\partial p}$, where $Y(p)$ is a firm's net supply function in terms of a certain good's price (p).

capital could be used as an input, while in cost or profit function models, it seems

impractical to define the price of transportation service.

Econometric analysis integrates statistics and economics. Though economic

theories and statistical methods often require rigid assumptions that are prone to induce

debates on results' reliability, econometric approach still has overwhelming advantages in

this field since they can handle both short-run and long-run problems[15].

2.4.2 Input-Output (I-O) Analysis

Input-output model[16], following the pioneering work of Leontief (1953), is one of

the earliest methods of empirically and quantitatively modeling the structure of economic

interdependence of different sectors/industries within an economic system. The I-O

model is based on *Quesnay's Tableau Économique* (Economic Table) (1759/1792) which

is developed in full generality by Walras's general equilibrium theory (1874/1954), and

then Leontief further makes the model applicable in the real economy. In I-O model,

every economic activity is assigned to both production and consumption sectors. Hence,

---

[15] In the long-run economic impact evaluation, a dynamic economic model is preferred to guarantee the estimation accuracy (Pleeter, 1980).

[16] Strictly speaking, Leontief's Input-Output Model may be not fit for regional economic research in terms of transportation infrastructure. In theory, Leontief investigates how inputs (raw materials, labor and intermediate goods) change given additional demand for outputs, rather than what results will be achieved if extra inputs are provided, which is more likely a statement of Keynesian. The focus lies on how to differentiate government procurement from government investment, though both of them belong to the category of government expenditure. It seems that government procurement determines the demand of final outputs, while government investment is approximate to some kind of input in economic system. Leontief (1944) discusses how the national level of employment would be affected by the cessation of war purchases of planes, guns, tanks and ships. These purchases are government procurement. Actually, in many regional economic studies, public capital stock is often considered as input rather than output, so is the transportation capital stock. Under this condition, Leontief's I-O analysis may be not a good alternative theoretically. However, I-O model has been developed with the idea of Keynes, and thus it's worthwhile to reviewing it here.

industries are inter-dependent, i.e., industries use products of other industries as inputs to produce its own products that will be used as inputs of other industries.

The I-O model helps us track the flow of products measured in monetary value between sectors in a given economic system (Leontief, 1987). It has three components: 1). Transactions table: contains basic data on the flows of goods and services among suppliers and demanders. Normally, various industries are regarded as intermediate suppliers and purchasers, while households are considered as primary suppliers and final purchasers. It shows the monetary flow of goods and services in a local economy. 2). Direct requirements table: shows the proportion of inputs from different intermediate suppliers required to produce one unit output for each intermediate purchaser. 3). Total requirements table. This is achieved following several-round calculation based on above two tables. The multipliers are drawn from this table, and tell us how much of an increase in each sector's input to expect as a result of each additional amount of final demand.

Three impacts are measured in the I-O analysis: 1) Direct demand effects – the value of the immediate changes in an assigned industry, directly affected by change in final demand; 2) Indirect effects – the value of inputs purchased by the backward-linked industries in additional rounds of spending resulting from direct demand changes, caused by inter-industry exchanges; and 3) Induced effects – the impacts on all local industries caused by the expenditures of new household income generated by the direct and indirect effects. (Leontief, 1987)

Above features of I-O model make it advantageous in following aspects: first, it has a more detailed classification and description of economic activities than econometric

models; and second, it is possible to track each industry's feedback for a given change in the output[17]. However, an economic system is too complex to estimate the change quantitatively without setting up some restrictive assumptions: a) constant returns to scale – an industry's production function is linear; b) no supply constraints – an industry has unlimited access to inputs that are also perfectly elastic, and its output is only determined by demand; c) fixed commodity input structure --no substitution of inputs; d) homogenous sector output – if an industry can produce multiple outputs, it will not increase the output of one product without proportionately increasing the output of all its other products; e) homogenous industry technology – an industry uses the same technology to produce all products, and different firms producing the same product use the same process; and f) perfect market structure – there are no unused or underused local resources, and excess capacity in firms and labor is not recognized.

These assumptions limit the application of multipliers and I-O analysis' feasibility in long-term cases. In a long period, magnitudes of multipliers calculated based on existing relationships within the local economy, become less certain if completely new types of economic activities are introduced into the region. The automatic adjustment in a market during a long term would also cause multipliers' overestimation of the impact of a change. In conclusion, I-O analysis depicts a static environment, while changing

---

[17] The Bureau of Economic Analysis has developed IMPLAN (Impact Analysis for Planning) and RIMSII (the Regional Input-Output Multipliers II) to facilitate the application of I-O model in analyzing regional economic development.

technology, substitution of input factors, interregional and intraregional trade patterns, and prices would induce the inaccuracy of multipliers.[18]

Theoretic assumptions limit the applicability of I-O analysis in a long-term dynamic economic system. In reality, this method is also very costly to implement, because all industries' input and output data need to be collected. Highly labor-division in modern society enhances the interdependence of various sectors, and makes the establishment of transaction table extremely complex, let alone successive data collection work. Then the estimate of multipliers would be another challenge. In brief, both theoretical and practical deficiencies limit its application in investigating transportation's long term impacts on regional economic growth.

2.4.3 Computable General Equilibrium (CGE) Models

CGE models, creatively developed by Johansen (1960), are a class of economic models used to estimate how an economic system reacts to changes in policy, technology, and some other external factors. Similar to I-O model, the core of CGE models also relies on the general equilibrium theory, while CGE is mainly contributed by successive Arrow-Debreu's achievements (1954).

Though there isn't a unified definition of CGE models, the basic CGE model has following features: 1) it complies with the neo-classical economic theory, assuming cost-minimizing/profit-maximizing behavior by producers, average-cost pricing, and household demands based on optimizing behavior (i.e. utility-maximization); and 2) the

---

[18] Long-term development of I-O analysis has relaxed several assumptions that exist in basic versions, such as fixed coefficients of inputs, no role of price, and static. Moreover, I-O analysis has not been a mainly model of production, and is applied in many fields (i.e. personal income, racial/ethnic impacts, etc.). However, the extended I-O model would not be discussed in this paper.

market is competitive and clearing (i.e. the interactions among multiple agents in the system are regulated by prices). A typical CGE model contains four basic components: producers, consumers (households), government and foreign trade.

a) Production: in CGE models, producers in each industry attempt to maximize their profits in given production technology and constrained resources[19] (raw materials, intermediated goods, electricity, fuel…). Normally there are two types of production function: a constant elasticity of substitution (CES) function for value added, and a Leontief function for intermediate goods.[20]

b) Consumption: households will maximize their utilities through optimal combination of products and services (including investment and entertainment) under the constraint of their income obtained from production and transfer.

c) Governmental activities: governments collect taxes and tariff, while they are also consumers in the market. They invest in public affairs, and provide subsidies or other financial transfer. Policies are usually introduced as exogenous variables in CGE models.

d) Foreign trade: the constant elasticity of transformation (CET) model is normally used to describe the procedure in which products/services are distributed between domestic market and exports in order to maximize total profits.[21]

---

[19] In the U.S., many firms actually don't focus on profit, such as non-profit organization, and government-sponsored enterprises.

[20] Leontief production function: $X_j = Leontief(Z_{1j}, …, Z_{ij}, …, Z_{nj}, PI_j, OCT_j)$, where X denotes product or service, Z_{ij} represents the ith intermediate good required to produce the product/service j, PI is the primary input (raw materials, energy, etc…), and OCT denotes other cost needed.
CES production function: $Z_{ij} = CES(Z_{ij}^D, Z_{ij}^M)$; $PI_j = CES(K_j, L_j)$; in which D represents domestic, while M denotes imported. K is capital, and L is labor.

[21] Exports in this model only refer to non-military ones. Another option is to use the Armington Model (Lloyd and Zhang, 2006) to describe the procedure in which the combination of domestic products and imported products is optimized to minimize the total cost.

In CGE models, economic activities in each segment should reach its own equilibrium. For example, products/services should be balanced both in quantity and value in commodity market. In labor market, labor will not encounter institutional barriers in migration (i.e. full employment). In capital market, total investments evenly match total savings. For government, budget deficit should be equal to the difference between expenditure and revenue. Household should also obey the rule, i.e., residential savings must equalize the difference between expenditure and income. For foreign trade, trade surplus represented by the inflow of foreign capital should equate the trade deficit represented by the outflow of domestic capital.

Finally, a group of equations are integrated to achieve the solution. Based on the closure rule (Sen, 1963), at least one market's equilibrium should be dropped from the equation group. Several important closures associated with specific economic theories or 'schools' have been developed, such as the neo-classical closure, the Keynesian closure, and the Pigou closure[22] (Thissen, 1998). Except that the neo-classical closure is fundamental to Walraisan CGE models, other closures have tried to describe a more actual economy (macro CGE models), considering unemployment and imperfect competition. Different closures will result in distinct results even using the same CGE model. Hence, incorrect selection of closure may lead to misplaced conclusions – advocating inappropriate policy recommendations.

---

[22] Other closures include the Johansen closure, the Kaldorian closure (neo-Keynesian closure), the Kaleckian closure, the loanable funds closure,

To run the simulation of CGE models, a social accounting matrix (SAM)[23] representing flows of all economic transactions that take place within an economy is required. The SAM captures the transactions and transfers between all economic agents in the system, each cell of which records the payment for a transaction. The SAM provides the benchmark database of CGE model, and normally is used as the calibration technique to estimate parameters (Thissen, 1998). In form, SAM seems like a counterpart of the transaction table in I-O analysis, with a much greater emphasis on institution accounts, such as investment, international trade, tax and transfer.

Compared with I-O analysis, CGE models regard price endogenous and the market's demand and supply are influenced by price. Furthermore, it contains all economic information in an economy, much closer to reality. However, as the other side of a 'sword', CGE models require a huge amount of efforts in implementation. The data in SAM that bridges the detailed linkages between one sector and another one are very complex and overwhelmingly difficult to obtain. The accuracy of parameters is not guaranteed no matter through subject expert judgment or objective estimates based on various calibration techniques. In addition, similar to I-O models, CGE is not an effective option for a long-term analysis, either. As a result, the implementation of CGE models is severely limited by its high cost and strong proneness to inaccuracy. In brief, CGE is ideal but not very practical for macro and long-term economic research.

---

[23] The first SAM was created in UK in 1962 (Stone and Brown, 1962) and then was built as a matrix representation of the National Account (Pyatt and Round, 1977).

2.4.4 Cost-Benefit Analysis

Cost-benefit analysis (CBA) is a systematic process to calculate and compare benefits and costs of a project. It has a broad application for decision making in both business organizations and governments, since it not only determines the feasibility of an option, but also illustrates whether this decision is better than other alternatives.

Boardman et al. (2006) summarizes nine steps to run a generic CBA, including establishment of alternative projects, determination of stakeholders, collection of all cost and benefit elements, prediction outcome of cost and benefits over the duration of the project, monetization of all elements, application of discount rate, calculation of the net present value (NPV), validation of the system (sensitivity test)[24] and final recommendation. By this token, CBA could assess the desirability of projects in both long-term[25] and wide-range[26] views, i.e., it implies the enumeration and evaluation of all relevant attribute performances (costs and benefits) (Prest and Turvey, 1965). Sen (2000) clarifies three fundamental principles of CBA: explicit valuation, broadly consequential evaluation, and additive accounting. Correspondingly, there are also three structural requirements to implement a perfect CBA: assumed completeness, full knowledge or probabilistic understanding, and non-iterative and non-parametric valuations. Failure to fulfill these requirements may result in considerable challenges/debates on its application in reality as following. (Preset and Turkey, 1965; Baram, 1980; and Sen, 2000)

---

[24] Sensitivity test is a systematic method for examining how the outcome of BCA changes with variations n inputs, assumptions, or measurements of costs and benefits.
[25] Investigates repercussions in the further, the nearer and the future.
[26] Considers side-effects of various aspects on a plenty of entities, such as people, industries and regions.

a). Inadequate identification of costs and benefits, especially associated externalities and spillover effects. For example, a built metro station may not only improve the traffic mobility, but also increase the housing value in adjacent areas.

b). Difficulty in quantifying intangible attributes. Many social factors, such as happiness and satisfaction, are impossible to be quantified, let alone to be monetized.

c). Dynamic depreciation rate for valuing future benefits and costs in present analysis. For convenience, people usually set a fixed depreciation rate in a project's life time. However, it usually changes based on the real market condition.

d) Improper distribution of costs and benefits on various entities. The determination of proportions of costs/benefits for each stakeholder in a project is usually subjective, and it is also changeable since each entity's preference may not be invariable all the time. Besides, private investors consider more on short-term cost/benefit, while governments or social scholars may be prone to focus on long-term welfare.

e) Future uncertainty. Nature and humane minds are impossible to be forecasted precisely, so occurrence of disasters or the changes in people's minds would both overset pervious evaluation.

f) Tendentious judgment based on self-interest and/or other analytical temptations. Decision makers always have subject preferences that may influence the objectiveness of evaluation.

All above deficiencies limit the application of CBA approach. Benchmarking projects are usually referenced, but marked differences in function and in skill levels of team members increased the difficulty of executing CBA on even similar projects. The

subjectivity and orientation in CBA further damages its accuracy.[27] Chichilnisky (1997)

points out that CBA can be dangerous if taken literally on large issues and large

timescale. As a result, CBA is more micro, practical and project-oriented, and more often

applied on specific projects rather than in a macro or meso economic research.

2.5 OVERVIEW OF TRAFFIC CONGESTION

2.5.1 Congestion Definition

According to Federal Highway Administration (FHWA)'s report, "[traffic]

congestion usually relates to an excess of vehicles on a portion of roadway at a particular

time resulting in speeds slower -- sometimes much slower-- than normal or "free flow"

speeds." [28] Though congestion is relatively easy to recognize and reminds us with such

words as "clog", "impede" and "excessive fullness", it doesn't have a single, broadly

accepted definition (OECD/ECMT, 2007). In fact, congestion is both a physical

phenomenon that vehicles impede each other's movement and a psychological one

related to the manner in which users' expectation on traffic speed is not satisfied. One

more sophisticated definition might be "the impedance vehicles impose on each other,

due to the speed-flow relationship, in conditions where the use of a transport system

approaches its capacity".[29]

The basis to evaluate congestion costs or effectiveness of congestion relief

method is to measure it quantitatively. In this case, traffic congestion is normally defined

---

[27] Some judgments may be biased owing to either deficiency in knowledge or inclination to certain stakeholders on purpose.

[28] Traffic Congestion and Reliability Trends and Advanced Strategies for Congestion Mitigation. Cambridge Systematic, Inc., and Texas Transportation Institute. Sep. 2005. p. 2-1.

[29] Dargay, J.M. & Goodwin, P.B. (1999) Traffic congestion in Europe. Round Table 110, European Conference of Ministers of Transport, OECD, Paris, 155-200, p.160.

"physically", such as "travel time or delay in excess of that normally incurred under light or free-flow travel conditions" (INCOG, 2001). In Minnesota, traffic flow with speed below 45 mph on arterial roads for any length of time in any direction between 6:00 a.m. and 9:00 a.m. or 2:00 p.m. and 7:00 p.m. is defined congested. In Michigan, a congested traffic condition has a volume/capacity ratio greater than one. In California, when the average speed on a freeway drops below 35 mph for 15 minutes or more on a typical weekday, the freeways is regarded congested.[30]

2.5.2 Measurement of Traffic Congestion

Various indicators have been developed through measuring travel time, travel speed, traffic volume, level of transportation service (LOS)[31], traffic signal cycle failure[32] and so on.

The Travel Rate Index (TRI) calculates the additional travel time over free-flow conditions when there exists congestion.[33] The result indicates the reciprocal of the ratio of peak-flow travel speeds to free-flow travel speeds. The higher the TRI, the slower the travel speed in peak hours, i.e., the severer the traffic congestion. It can be used to calculate other indicators such as annual hours of delay and the portion of travel under

---

[30] Bertini, R. L. (2005). *You are the traffic jam: an examination of congestion measures.* In Transportation Research Board Annual meeting ed. Portland State University. Retrieved from ww.its.pdx.edu/pdf/congestion_trb.pdf on October, 12, 2012.

[31] Typical LOS measures include volume/capacity, density, delay, number of stops, and so on, to define a scale from A (free flow) to F (forced or breakdown flow).

[32] One has to wait through more than one cycle to clear the queue

[33] Four steps to calculate TRI: (1) estimate peak period vehicle mileage; (2) assign each road segments to a five congestion levels (extreme, severe, heavy, moderate, and free-flow) which have different standards of average daily traffic per lane and average vehicle speed based on road type (highway or arterial); (3) calculate vehicle travel delay, based on the difference between average and free-flow travel speeds on each segment, time vehicle mileages on that segment; and (4) calculate average passenger-speed for each road section based on vehicle occupancy.

congested conditions. However, TRI only reflects recurrent congestion, while a

significant portion of delay is caused by incident/accidents. The Travel Time Index (TTI)

scales both recurrent and non-recurrent delays. TTI is calculated based on information

from the Smart Dust Network which can provide real traffic speed data for specific times

and locations recorded in billions of discrete reports from GPS-enabled probe vehicles.

Thus, TTI is more accurate than TRI and also used to calculate bottleneck factors, such as

the number of congested hours for a particular intersection or link.

There are also many other congestion indicators representing different

perspectives and assumptions, some of which only consider impacts on motorists and are

not fit for evaluating benefits from congestion reduction through mode shifts or more

accessible land use, such as roadway LOS, TRI, and CRM (Congested Road Miles)

(Litman and Doherty, 2009). In addition, indicators measuring impact on per capita

rather than that on per vehicle are better for estimating total congestion costs, like Annual

Delay Per Capita, Fuel Per Capita, and Congestion Burden Index (CBI)[34].

Reliability indicators reflect another important facet of congestion impacts, traffic

reliability. Popular reliability indicators include the 90th or 95th percentile travel time[35],

the buffer index[36], the planning time index[37], and the frequency that congestion exceeds

some threshold.

---

[34] CBI is the travel rate index multiplied by the proportion of commuters subject to congestion by driving to work.

[35] Reflects the longest travel time during a ten or twenty day period.

[36] Reflects the extra time which has to be added by travelers to their travel schedule to ensure on-time arrival. It is calculated as the difference between the 95th percentile and average travel time, divided by the average travel time.

[37] Reflects the total travel time required to provide adequate buffer time, including both recurrent and unexpected delay. It compares the near-worst travel time to a travel time in light or free-flow traffic.

Bertini (2005) points out that many of above options (e.g., traffic counts, the volume/capacity ratio and LOS) are usually derived from simple and limited data, are extrapolated over large segments of the road network, and do not consider the impacts on different types of users. These poor measurements may lead to irrational policy decisions. Latest development which allows for more robust data collection that could reflect actual system performance should substitute those old ones. Furthermore, the benchmark of free-flow speed, an ideal but unrealistic one, is suggested to be replaced by median speeds or a set of benchmark values (e.g., percentage of maximum legal speed or different speed bands) to provide more reasonable estimates. (ECMT, 2007)

2.5.3   Traffic Congestion and Regional Economy
As a kind of public capital, transportation infrastructure constitutes an element in the macroeconomic production function and its stock may enter the production function directly, as a third input. However, different from private capital which is used to purchase equipment, hire employers, and invest in other necessary inputs in production, transportation infrastructure is more likely a platform on which other inputs could be used to produce. In this case, the performance of transportation system may determine the production efficiency and then affects production output indirectly. Thus, traffic congestion that measures the fluidity of economic resources may not only result in wasted time and fuel accompanied with unreliability of shipment, but also further produces economic multiplier effects where the price of commodities, goods or services may increase with deterioration of transport situation. Furthermore, congestion's indirect

impacts may lead to any change in industry layout, people's residential choice, and other business and residential behaviors closely related with transportation.

Boarnet (1997) examines the relationship between highway congestion, labor productivity and output using data of California counties over the period from 1977 to 1988. The evidence suggests that congestion reduction on existent stock is significant to productivity, while effects of expanding the street and highway stock are suspicious. Fernald (1999) emphasizes the effects of road infrastructure and congestion on output using a dataset of cross-section U.S. industries between 1953 and 1989. It's found that vehicle-intensive industries benefit more from roads-building, but roads construction and expansion are not always productive at the margin. In addition, congestion does not appear statistically significant to negatively influence output before 1973, but becomes a significant factor thereafter. Hymel (2009) explores the link between traffic congestion and employment growth in large U.S. metropolitan areas with a cross-sectional analysis. Results illustrate that the initial congestion level has a significant negative effect on aggregated employment growth rate in following years under controlling initial employment, colleges, and crime level.

Boarnet, Fernald and Hymel obtain similar results through their researches: traffic congestion has significant and negative impacts on output or employment, but have different focuses. Boarnet uses county level data of California, while Hymel chooses 85 largest metropolitan statistical areas (MSA) of United States. Fernald's study is based on the nation level with disaggregated industrial data. Each of above studies has its own limitations.

Firstly, to avoid aggregation bias, the match between areas where primary regional economic activities occur and areas where traffic congestion really takes effect is crucial. Traffic congestion is prevalent and severe in core cities or counterparts with dense business and human activities, and might not be a significant problem in rural areas. A nation or state level analysis eliminates variations among regions and may result in great bias in estimation. It also fails to provide any valuable information to local governments that may be more interested in congestion issues. On the other side, a county level analysis seems too microcosmic to cover congestion's integrated impacts. Congestion occurring in certain segments of arterial roads or highways normally influences upstream traffic flow and other parts of transportation network. If the congestion issue is analyzed in separated sections (counties) instead in the entire system (urban area), the estimation bias may occur. To be sure, in Boarnet's research, there exists an exception. In California, some counties are large enough to cover whole regional transportation systems, such as Riverside, San Diego, and Los Angles. These counties could be regarded as qualified research units, while some other counties, including Orange, San Francisco and its adjacent ones, are not large enough and need to be combined together to compose urban areas or MSAs. Compared with nation, state or county, MSA seems a better option, albeit a MSA is still not the optimal choice, because it includes not only urban areas but also satellite cities plus intervening rural land socio-economically connected to the urban core city. Thus, a better research subject is the urban area.

Secondly, how to measure traffic congestion is another pivotal point. Fernald

models congestion as a function of the total miles driven by trucks, automobiles, and

other motor vehicles. However, the magnitude of total driven miles actually represents

traffic demand rather than traffic congestion, though given slowly increased traffic supply

(limited transportation infrastructure), excessive traffic demand is a direct factor inducing

congestion. The rapid development of transportation management system (e.g. intelligent

transportation system[38]) and more reasonable design in transportation infrastructure and

urban layout have improved transportation capacity remarkably even without large-scale

road construction and expansion. As a result, increment in total driven miles could not

definitely result in traffic congestion with equivalent increment, i.e., this proxy may

overestimate the congestion level. Boarnet references the capacity adequacy variable

which is the ratio of the highway's rated capacity divided by peak hour travel flow.[39] It is

also debatable. The application of ITSs helps increase highway's rated capacity which is

set fixed in Boarnet's model. The selection of the peak hour traffic flow also seems a

little arbitrary and subjective. Besides, most states in the United States lack similar

monitoring systems to collect data in all primary road segments, and thus this variable is

only applicable in CA cases. Hymel introduces the travel delay per capita produced by

Texas Transportation Institution in his research. This index measures the average hours

---

[38] The U.S. Department of Transportation sees the ITSs as encompassing a broad range of wireless and wire line communication-based information and electronics technologies. These technologies are integrated into the transportation infrastructures and vehicles to relieve congestion, improve safety and enhance social benefits and productivity.

[39] For rural road segment, the peak hour traffic flow is the flow during the 30th highest volume hour during the year, and for urban road segment, peak hour flow is the volume during the 200th highest volume hour for each year.

of travel delay experienced by a resident of an urbanized area because of recurring congestion and incidents. It evaluates the average delay experienced by per resident in a given area instead of those who really drive in the peak period. Hence, this average index may underestimate congestion level, for a part of population who use public transit rather than cars for rush-hour commute as well as those who work at home are still counted in the denominator. In addition, it doesn't consider the effect of urban size. In small urban areas, the average commuting distance is shorter, so is the commute time. Severe congestion may not lead to a heavy delay. While in large urban areas, even lighter congestion may result in a greater delay index.

CHAPTER 3   METHODOLOGY

In Chapter II, four methodologies in transportation research field have been introduced and discussed. One principle is matching research methods to research questions. Transportation system is an integrated network instead of isolated infrastructure. Thus, it is better to evaluate its performance in macro- or meso- environment. Meanwhile, both direct & indirect and short-term & long-term impacts of transportation infrastructure should be considered. Moreover, planning, public acceptance and construction of transportation infrastructure in the U.S. are usually time consuming, and its induced effects in regional economy also need a long time to take effect. Consequently, characteristics of traffic congestion, including locality, network, sociality and chronicity, determine that the classical econometric model is preferred here.

In most studies, traffic congestion is considered as a direct input in the model with the dependent variable of regional output or employment (Boarnet 1997; Hymel 2009). However, as criticized by Miller and Upadhyay (2002), the approach that treats all possible determinants beyond basic factors of production as inputs may be conceptually inaccurate, since many of them may have only indirect effects on output, and these additional determinants just affect the efficiency of utilizing real inputs, which consequently influences the total factor productivity. In this study, traffic congestion is not considered an engagement in production procedure, and a coherent two-step analytical procedure is proposed. In the first stage, regional economic efficiencies are

defined and quantified. Then, the impact of traffic congestion on pre-calculated

efficiency is probed.

To validate the robustness of results, two distinct methods are applied: Total

Factor Productivity (parametric analysis) and Malmquist Productivity Index[40] (non-

parametric analysis). Detailed discussion on these two methodologies will be expatiated

in following two chapters, and in this section the procedure of implementing regression

analysis will be focused[41].

3.1 MODEL ESTABLISHMENT

As the dataset covers both time series and cross sections, an inevitable apprehension

is the spurious relationship.[42](Engle and Granger, 1987) Granger and Newbold (1974)

comment that spurious regressions tend to lead investigators to commit a high frequency

of Type I errors[43]. Hence, it's frequently recommended to do the stationary test in advance.

If both explained variables and explanatory variables are stationary, estimates based on the

original model are relatively precise. Otherwise, it's necessary to take some measures to

obtain stationary data, one of which is to take the first difference of variables at the sacrifice

of removing some useful long-term information of original model and making the model

difficult to explain since it changes the original model form by investigating the

---

[40] It's a bilateral index used to compare the production efficiency of two economies. (Caves et al., 1982)
[41] Indeed, the parametric analysis in the first step also uses the regression methodology. Therefore, they have the same procedure to deal with the data.
[42] In some cases, dependent and independent variables may not have direct causal connection, yet it may be incorrectly inferred that they do, due to either coincidence or the presence of an unseen factor, which results in the phenomenon of 'spurious regression', commonly represented by very high coefficients of determination (R^2), prevalent significant results for both coefficients (t-test) and the regression equation (F-test), or severe auto-correlation of residuals Durbin and Watson (1950).
[43] In statistics, a Type I error implies the rejection of a potentially true null hypothesis, caused by underestimating p-value in significance test.

relationship between first-difference value of variables rather than their level values based on economic theories. Another option is to take the co-integration test and then select a proper model type based on the test result[44].

The unit root[45] test is the most popular one used to examine data's stationarity. In early period, unit root tests were developed specifically for time series data, and recently more and more methods have been further developed for panel data, such as LLC[46](Levin et al., 2002), IPS(Im et al., 2003), Breitung (Breitung, 2000), ADF-Fisher and PP-Fisher (Maddala and Wu, 1999). According to Ramirez (2007), these investigators have shown that panel unit root tests are more powerful (less likely to commit a Type II error[47]) than their counterparts applied to individual time series because the information in the time series is enhanced by that contained in the cross-section data. Furthermore, in contrast to individual time-series unit root tests which have complicated limiting distributions, panel unit root tests lead to statistics with a normal distribution in the limit (Baltagi, 2001).

In this study, two tests are applied: LLC test assuming a common (identical) unit root process across the relevant cross-sections[48], and ADF-Fisher test assuming different

---

[44] Either Vector Error Correction Model (VECM) under the condition of cointegration or Vector Auto-Regression Model (VAR) for non-cointegration. Model details are discussed later.

[45] A linear stochastic process has a unit root if 1 is a root of the process's characteristic equation.

[46] LLC test permits various intercepts and time trends, heteroscedasticity and higher-order autocorrelation, and is fit for the panel data with medium size (both T and N are no more than 250).

[47] In statistics, a Type II error is the failure to reject a potentially false null hypothesis, caused by overestimating p-value in the significance test.

[48] The LLC test employs a null hypothesis of a unit root using the following basic Augmented Dickey Fuller (ADF) specification: $\Delta y_{it} = \alpha y_{it-1} + \sum \beta_{ij} \Delta y_{it-j} + X_{it} \delta + \varepsilon_{it}$, where y_{it} refers to the pooled variable, X_{it} represent exogenous variables in the model, and ε_{it} denotes the error terms that are assumed to be mutually independent disturbances. It's assumed that $\alpha = \rho - 1$ is identical across all cross-sections, but the lag order for the difference terms across all sectors is allowed to vary.

45

unit roots processes across the relevant cross-sections[49]. If and only if in both tests, the hypotheses that there are common/various unit roots across cross-sections are rejected, it could be concluded that the variable has a stationary process. In addition, for each unit root test, if and only if results from all three modes (with both intercept and trend, with only intercept, and with none)[50] show that hypotheses couldn't be rejected, could we consider that the variable is non-stationary. Another issue needed to concern is cross-section dependence which violates the assumption that the error terms are independent across cross-sections (Chang, 2002; Pesaran, 2007). Following the suggestion from Levin, Lin, and Chu(2002), the cross-sectional averages are subtracted from the series, and then the unit root test is applied for modified series to mitigate the impact of cross-sectional dependence.[51]

For non-stationary economic variables, the co-integration analysis developed by Granger (1981), and Engle & Granger (1987), which builds on error correction models, provides a powerful and widely adopted framework for studying long-run as well as short-run relations.[52] Variables are co-integrated if there exists a long-term equilibrium among them. Popular co-integration tests for time-series data include EG test (Engle and Granger,

---

[49] The ADF-Fisher test estimates a separate ADF regression for each cross-section to allow for individual unit root process, which implies $\alpha_i = \rho_i - 1$ may vary across cross-sections.

[50] The test starts with the ADF regression including both intercept and trend in the model, then the model changes including only intercept, and finally the model doesn't contain any of those two terms.

[51] Before undertaking such an approach, a test should be done to ascertain the existence of cross-section dependence. In the context of large T and small N, the LM test statistic proposed by Breusch and Pagan (1979) can be used, while in panel data models with small T and large N, STATA command XTCSD implement two semi-parametric tests proposed by Friedman (1937) and Frees(2004), as well as the parametric testing procedure proposed by Pesaran (2004). (see STATA help document)

[52] Stock H. J. (1997). Cointegration, long-run comovements, and long-horizon forecasting. In D. M. Kreps and K. F. Wallis (Eds.). *Advances in economics and econometrics: theory and applications: Seventh world congress volume 3*. Cambridge: Cambridge University Press.

1987), JJ test (Johansen and Juselius, 1990), and S test (Shin, 1994)[53], while more attention

have been paid to panel data (Kao, 1999; Pedroni, 1996, 1997, 1999; Westerlund, 2005).

Cointegration demonstrates that the original model could be regressed without concerning

about spurious problem[54]. Meanwhile, a $VECM$ (vector correction model) needs to be

established to consider short-run fluctuations[55]. In brief, no matter variables are stationary

or not (non-stationary variables must be co-integrated), the original model could obtain

satisfactory estimators[56], with the exception that non-stationary co-integrated variables

requires a $VECM$ investigating short-run non-equilibrium for complements.

In advance of establishing detailed models, we detect the multi-collinearity[57]

among explanatory variables in each regression. Especially in the first-stage, the

application of the translog production function model makes multi-collinearity not

ignorable, as squared terms and interaction terms are included. For cross-sectional data, the

tolerance inflation factor (VIF) is usually used to determine whether collinearity is

---

[53] The EG test is a multivariate generalization of the DF test with the null hypothesis of no-cointegration, while the S test addresses the cointegration hypothesis directly. Both EG and S tests are two-step, single-equation, and residual-based tests. On the other hand, the JJ likelihood ratio trace test is a system method based on vector autoregression (VAR), and is not a single step method either. This test is seriously biased toward spuriously detecting cointegration in the case of low-order VAR model or small samples ($n < 100$). (Konya 2004).

[54] Existence of cointegration presents that there is long-run steady equilibrium among variables, and the regression residual is stationary. As a result, regressive estimators based on the original model are relatively precise.

[55] According to the Granger representation theorem (Engle and Granger, 1987), if Y and X are co-integrated, their short-run non-equilibrium relation could be expressed by a $VECM$: $\Delta Y_t = lagged(\Delta Y, \Delta X) - \lambda \mu_{t-1} + \varepsilon_t$, where μ_{t-1} is the non-equilibrium error term and λ is the short-run adjustment parameter. $\mu_{t-1} = Y_{t-1} - \alpha_0 - \alpha_1 X_{t-1}$, if the long-run equilibrium exists in the model: $Y_t = \alpha_0 + \alpha_1 X_t$.

[56] In both cases, regression residuals are stationary.

[57] Though existence of multi-collinearity is not a violation of the OLS basic assumptions, it violates the assumption that X matrix is full ranked, making variance of the model as well as variances of coefficients inflated and inducing unreliable inference. Basic assumptions include: a) $E(\varepsilon_i) = 0$; b) $var(\varepsilon_i) = \sigma^2$; c) $cov(\varepsilon_i, \varepsilon_j) = 0$; d) $cov(\varepsilon_i, x_i) = 0$; and e) $\varepsilon_i \sim N(0, \sigma^2)$.

significant enough to be concerned. Though VIF is not applicable for the panel data because of the data structure, through observing several typical symptoms[58], we can still conclude its existence. In general, to cope with this problem, methods in two fields are frequently used. One refers to dealing with raw variables, such as selection of proper variables based on stepwise methodology or re-combination of highly related variables using Principle Component Analysis (PCA). The other one implicates the statistical model, such as ridge regression, a biased estimation with lower variances of estimates. Here, an unabridged translog production function should be followed, since each term in this model is meaningful, and unbiased estimators of coefficient are crucial for subsequent calculation. Thus, the PCA is used.

A common panel data model is described as:

$$Y_{it} = \sum_{j=1}^{k} X_{jit} \beta_j + \mu_i + \delta_t + \varepsilon_{it} \tag{1}$$

where X_j are observed explanatory variables, μ_i and δ_t are known as unobserved effect. The subscript i and t refer to the unit of observation and time period, severally. ε_{it} is a disturbance term assumed to satisfy the usual regression assumptions. μ_i refers to impacts that only vary among units but unchangeable with time, while δ_t are those only varying with time but keep constant over all units. Introduction of μ_i and δ_t enhances the complexity in regression analysis, because they may be correlated with any of the X_j variables that leads to a violation of OLS assumptions, which may result in unobserved

---

[58] Including: a) small changes in the data, such as removing several independent variables from the model, produce wide swings in the parameter estimates, including signs' conversion and significances' change; b) coefficients may have 'wrong' signs or implausible magnitude, compared with theoretical background; and c) coefficients may have very high standard errors and low significance levels though they are jointly significant and R^2 for the regression is quite high.

heterogeneity bias. To solve this issue, several statistical techniques are developed for panel data model specifically, two main approaches of which are known as fixed-effect regression and random-effect regression. There are three versions of the fixed-effect approach: within-group estimator (FE), first-difference estimator (FD), and least square dummy variable estimator ($LSDV$), first two of which eliminate the unobserved effects while the last one uses dummy variables to represent them.

Both FE and FD approaches tackle unobserved heterogeneity bias by removing μ_i in respective regression. However, there is a price for everything. Any explanatory variable that remains constant for each individual over time will drop out of the model. Compared with FE, the cost of applying FD seems larger, because of two factors: (1) it could not cope with time-fixed effects δ_t. The unobserved term $\delta_t - \delta_{t-1}$ still remains in the regression equation; and (2) it may result in autocorrelation of the disturbance term, as a negative moving average autocorrelation is induced between ($\varepsilon_{it} - \varepsilon_{it-1}$) and ($\varepsilon_{it-1} - \varepsilon_{it-2}$). Meanwhile, the unobserved term $\delta_t - \delta_{t-1}$ may also encounter autocorrelation problem. In contrast, using the $LSDV$ approach in which unobserved characteristics of individuals are represented by dummy variables, impacts of unobserved fixed effects are measured as intercepts. This method has the advantage to keep even time-invariant explanatory variables in the model without worrying losing information. Albeit, defining dummy variables for all observations exhausts the degrees of freedom, especially in the case of a large N.[59]

---

[59] There are $N \cdot T - k - N$ degrees of freedom if the panel is balanced, the same as that in FE approach.

Different from the fixed-effect regression which estimates parameters under the premise of controlling individual unobserved effects, the random-effect regression has no attempt to realize individual unobserved effects at all. Compared with fixed-effect regression, random-effect regression has three additional assumptions: (1) individual effects are treated as being drawn randomly from a given distribution[60]; (2) individual effects are uncorrelated with explanatory variables[61]; and (3) individual effects are independent with original disturbance term. Consequently, all unobserved effects could be summed into the disturbance term. The complexity in this regression refers to the combined disturbance term that has both autocorrelation because of existence of μ_i in each period and cross-section correlation owing to the inclusion of δ_t in each cross section. In this case, although OLS estimator remains unbiased and consistent, it is inefficient and the OLS stand errors are computed incorrectly[62]. The generalized least square (GLS) method or even a more advanced approach, feasible generalize least square ($FGLS$) method could achieve unbiased, consistent and efficient estimates[63].

Compared with fixed-effect model, random-effect model has two significant advantages: a). it could tackle with time-invariant explanatory variables; and b). it keeps more degrees of freedom in calculation. Nonetheless, it must satisfy rigid assumptions. Otherwise, random-effect regression may produce not only biased but also inconsistent estimates. Adversely, in any case, estimates based on fixed-effect regression will be both

---

[60] $\mu_i \sim IID\left(0, \sigma_\mu^2\right)$ and $\delta_t \sim IID\left(0, \sigma_\delta^2\right)$.

[61] $Cov(\mu_i, x_{it}) = 0$ and $Cov(\delta_t, x_{it}) = 0$

[62] $E(\eta_{it}) = E(\mu_i) + E(\delta_t) + E(\varepsilon_{it}) = 0$, but $\sigma_\eta^2 = \sigma_{\mu+\delta+\varepsilon}^2 = \sigma_\mu^2 + \sigma_\delta^2 + \sigma_\varepsilon^2$, since all interaction terms are equal to 0 (assumption of independence).

[63] The different between GLS and $FGLS$ is whether σ_η^2 is pre-given (GLS) or pre-unknown ($FGLS$).

unbiased and consistent. Therefore, which model to be selected should depend on detailed requirements of the research and corresponding statistical characteristics.

Simultaneously incorporating above theoretical and practical considerations into model choice seems a daunting task. Researchers often rely on the Hausman (1978) specification test to detect if explanatory variables are orthogonal to unit effects. If there is no correlation between independent variables and those effects, then the difference between estimates of β in FE model and that in RE model should be statistically insignificant.[64] If the null hypothesis of orthogonality is rejected, the RE regression should be rejected in favor of the FE regression. However, traditional Hausman test assumes that μ_i, δ_t and $\varepsilon_{i,t}$ are independent identically distributed (iid). Under the existence of heterogeneity, the test may provide incorrect results, such as the negative value of H. Here, clustered robust variances are employed to obtain more precise results from a modified Hausman test (Chen, 2010).

After model establishment, more statistical tests need to be done as the criteria to select proper statistical methods in order to obtain relatively accurate estimates. For panel data, primary concerns are heteroscedasticity, autocorrelation and cross-sectional dependence, each of which is detected.[65]

---

[64] The Hausman test statistic H measures the difference between the two estimates:
$H = (\hat{\beta}_{FE} - \hat{\beta}_{RE})'[Var(\hat{\beta}_{FE}) - Var(\hat{\beta}_{RE})]^{-1}(\hat{\beta}_{FE} - \hat{\beta}_{RE}) \sim \chi^2_K$

[65] In this study, the following procedure is followed: 1). Heteroscedasticity test: Heteroscedasticity violates the assumption of homoscedasticity in OLS: $Var(e_{it}) = \sigma^2$. It doesn't result in biased parameter estimates, but may lead to higher variance estimate and cause the falseness of significance tests. 2). Autocorrelation test: Autocorrelation violates the assumption of OLS that error terms are uncorrelated: $Cov(e_{i\,t}, e_{i\,t-s}) = 0$. Though it doesn't bias the OLS coefficient estimates, it tends to underestimate (overestimate) the standard errors when the autocorrelation of errors at low lags are positive (negative). And 3). Cross-section dependence test: Cross-section dependence is a unique feature of panel data that implies there may exist correlation among individuals, and violates the unit independence in OLS: $Cov(e_{it}, e_{jt}) = 0$. It may lead to

In mathematics, different statistical approaches fit for distinct regression formulas. STATA provides various commands to tackle with these regressions under specific restrictions.

In general, primary sponsors of transportation infrastructure, especially roads and highway, are governments. The project plans and sizes are usually based on either the forecast for future development or the necessity to solve contemporary problem. Corresponding expenditures are directly affected by government budgets which are determined by government revenue, a reflection of total economic output. Thus, economic development may also contribute to transportation investment. The same issue may exist between employment/private capital and output. In consequence, endogneity[66] could not be omitted. Here the Granger causality test[67] is implemented to determine whether the dependent variable is the Granger cause of each independent variable, and then a Hausman test[68] is applied to judge if the model with those independent variables Granger caused by the dependent variable is endogenous or not. If the test can't reject the hypothesis that the model is exogenous, the endogeneity issue should be concerned.

---

incorrect variance estimates. Cross-section dependence can arise due to spatial or spillover effects, or could be due to unobserved common factors.

[66] Endogeneity means the error term is correlated with one or more independent variables in the model, i.e., $Cov(\varepsilon_i, x_i) \neq 0$. The reason may be either the existence of omitted explanatory variables correlated with existent variables in the model, or simultaneity that the dependent variable also determines independent variables. OLS estimate with endogeneity may lead to biased and inconsistent results.

[67] The Granger causality test is developed be Granger (1969) for determining whether a stationary time-series vector (x) is useful in forecasting another stationary time-series vector (y). 'Useful' means x is able to increase the accuracy of the prediction of y with respect to a forecast, considering only past values of y. It can demonstrate whether the simultaneity between dependent variable and independent variables exists.

[68] The Hausman test is a generalized test that could detect the statistical significance of differences $(\beta_2 - \beta_1)$ between estimates resulted from two different models using a statistic H asymptotically subject to χ^2 distribution with the number of degrees of freedom equal to the rand of matrix $Var(\beta_1) - Var(\beta_2)$. If the null hypothesis is rejected, one or both of the estimators is inconsistent.

Two ways are usually used to cope with the endogeneity issue: the two-stage least square ($2SLS$) method and the general moment method (GMM), both of which need to define instrument variables (IVs). In this study, it's difficult to choose appropriate instrumental variables not only because of the rigorous statistical requirements[69] but also owing to economic significance. For example, when using PCA approach based on the translog production function model to calculate coefficients of each input, each principle component incorporates all explanatory variables simultaneously, and such an integration makes it more difficult to find qualified instrument variables that can explain all information synchronously, are also uncorrelated with error terms, and are definable using economic theory simultaneously. In practice, the time lags of endogenous variables are often used as instrument variables, as there is a good theory under which the explanatory variables from one period ago can't be caused by a current explained variable (i.e. the past cannot be caused by the future), although this method has its intrinsic shortcomings[70]. Following Baum et al. (2003), lags and first-difference terms of endogenous variables are considered as instrument variables here.

It is arbitrary to use instrumental variables recklessly before testing their qualifications. In GMM, the over-identification test is a common one for detecting

---

[69] The selection of instrument variables has to obey two principles: 1). instrumental variables z should be highly correlated with endogenous explanatory variable x: $Cov(x, z) \neq 0$; and 2) instrumental variables should be exogenous in the model, i.e., they should be uncorrelated with error terms: $Cov(z, \varepsilon) = 0$.
[70] The use of lags of endogenous variables as IVs decrease the operating sample size as the number of instruments rises. Moreover, lagged variables may fail to be proper IVs, either because assumptions of zero correlation are not robust and fail due to a more complex pattern of serial correlation than the econometric assumes, or because these lagged variables are not correlated with the variables they are instrumenting. However, for macro-economic research, these two problems seem not so severe owing to the existence of social inertia. (McFadden,1994) Retrieved from http://emlab.berkeley.edu/~mcfadden/e240b_f01/ch4.pdf on October, 16, 2012.

exogeneity. In addition, Hansen(1982) creates the J- test[71] to verify the orthogonality conditions between IVs and error terms[72]. Another principal concern is whether IVs are highly correlated with endogenous regressors. If the correlation between them is really poor, then these IVs are called weak instrumental variables that may result in very large asymptotic bias in estimates. Stock and Yogo (2005) propose testing for weak instruments using F- statistics with the assumption of i.i.d. errors, and the Cragg-Donald F can be robustified in the absence of i.i.d. errors by using the Kleibergen-Paap statistics (Baum, 2009). All these tests could be done using corresponding commands in STATA.

Compared with the least square approach (including[73]: OLS, GLS, WLS, 2SLS, etc.) and the maximum likelihood estimation (MLE), GMM has advantages for dealing with large sample data without knowing variable distributions.[74]GMM could also handle the heteroscedasticity and autocorrelation in the model through defining corresponding weighting matrix in calculation. Thus, GMM always provides more efficient estimates than

---

[71] The other popular over-identification test is the Sargan test (Sargan, 1958). As Andersen and Sorensen (1996) document, the proliferation of IVs vitiates the J- test, while Sargan test is not influenced by the size of IVs as it does not depend on an estimate of the optimal weighting matrix. However, it requires homoscedastic errors for consistency, while J-test is fit for heteroscedastic errors. Hence, there is no precise guidance on relatively safe number of instruments. Generally, Sargan test is in a sense too big (too difficult to pass), yet J is too small (too difficult to reject) (Roodman, 2009). Given that in this research the dataset has the form of short panel (N > T), the number of IVs shouldn't be very large while the heteroscedasticity problem can not be ignored. The J- test is preferred to the Sargan test. Retrieved from http://www.uio.no/studier/emner/sv/oekonomi/ECON5103/v10/undervisningsmateriale/PDAppl_16.pdf on October, 16, 2012.

[72] If the null hypothesis that all IVs are exogenous is rejected, the results may be even worse than just taking the OLS, because not only does the endogenity problem still exist, but also the freedom of model is reduced.

[73] GLS: generalized least square. WLS: weighted least square.

[74] The premise of using MLE is knowing the variable distribution that's normally defined subjectively. Incorrect definition of distribution may lead to biased estimates. GMM estimators are always consistent, no matter for large sample or small sample. However, only the sample size is large, are estimates asymptotically efficient.

other methods if above tests on IVs are passed.[75] Therefore, GMM is preferred in this study to solve the endogeneity issue.

Two types of GMM are frequently used: $DIF - GMM$ (first-difference GMM) and $SYS - GMM$ (system GMM). $DIF - GMM$ was proposed by Holtz-Eakin et al. (1988), and later developed further by Arellano and Bond (1991). Taking the first difference of the Equation 1, the bias caused by omitted variables could be removed through dropping those unobservable/un-measurable individual factors, and the following model is obtained:

$$\Delta y_{it} = \beta \cdot \Delta x_{it} + \Delta \varepsilon_{it} \tag{2}$$

Arellano and Bond use lagged values of the endogenous regressors as IVs in the first-differenced model, since they are pre-determined and not correlated with the first-differenced error term. However, $DIF - GMM$ estimator is calculated using partial information because impacts of those time-invariant variables are removed. Moreover, lagged levels are weak instruments for first-differences if the instrumented variables are close to a random walk[76]. In such a case, the Arellano and Bond estimator has been found to have poor finite sample properties, in terms of bias and imprecision. To overcome the disadvantages of $DIF - GMM$, Blundell and Bond (1998) develop the $SYS - GMM$ estimator through constructing a system of equations, constituting the first difference equation and the original level equation.

---

[75] $2SLS$ is just a particular case of GMM, with error terms being both homogeneous and un-autocorrelated. For an over-identified equation, $IV - GMM$ cluster-robust estimates will be more efficient than $2SLS$ estimates.

[76] In mathematics, a random walk is a time-series data that consists of a succession of random steps. A random walk model could be described as the equation: $y_t = \alpha + y_{t-1} + \varepsilon_t$, where ε_t is the white noise with the property: $E(\varepsilon_t) = 0$ and $Var(\varepsilon_t) = \sigma^2$, and α is a constant drift.

$$\begin{cases} y_{it} = \beta x_{it} + \mu_i + \varepsilon_{it} \\ \Delta y_{it} = \beta \cdot \Delta x_{it} + \Delta \varepsilon_{it} \end{cases} \tag{3}$$

Lagged values of endogenous variables, $x_{j,t-k}$, are used as IVs of $\Delta x_{j,t}$ in first-differenced equation. While in the level equation, variables in levels, $x_{m,t}$, are instrumented with suitable lags of their own first differences, $\Delta x_{m,t-n}$. Compared with $DIF - GMM$, $SYS - GMM$ improves by dramatically reducing bias in finite sample and significantly gaining precision from exploiting additional moment conditions, in cases where the autoregressive parameter is only weakly identified from the first-differenced equations (Blundell and Bond, 1998). Thus, in this empirical work, the $SYS - GMM$ approach is employed.

3.2 VARIABLES AND DATA SOURCE

In following two chapters, the same dataset is examined with different methods. Thus, all variables with corresponding procedure and data source to obtain them will be introduced here. A three-input production function, using regional gross domestic production as the output and private capital stock, labor capital stock as well as highway and street capital stock as inputs is used to calculate production efficiency. [77] The introduction of highway and streets capital is after serious and cautious consideration. First, in macro economy, this item should be introduced since it also contributes to output directly. Second, different from healthcare, sewage and other similar public capital, transportation

---

[77] There are also some researchers who introduce energy and materials (besides labor and capital) in the production model (Hudson and Jorgenson, 1974; Berndt and Wood, 1979; Lindenberger and Kummel, 2002). However, the residential energy consumption data is not available at urban level or county level. Besides, energy production and consumption are integrated with capital and labor. So it's not considered in this study.

plays a more direct and positive role in improving economy.[78] Third, compared with public transit (including bus, subway and rail), highway and streets are more likely private goods,[79] as not only is this market full of competition but also it receive tiny subsidy[80]. Considering that the production function model is developed at the firm level with neo-classical economic theory, when it's used in the field of aggregated economy, it may be better to include those inputs with market characteristics. As a result, the three-input production function model has advantages in both theories and practical feasibility.

In succession, impacts of traffic congestion on economic efficiency will be investigated. In parametric analysis, economic efficiency is represented by the Total Factor Productivity (TFP), while in non-parametric analysis, it's denoted with the Malmquist Productivity Index. Plenty of studies have analyzed sources of the TFP growth, some of which are referenced as following.

3.2.1 Variable Selection

The impact of innovation on productivity growth through investment in research and development (R&D) has been well defined in literature (Griliches, 1980; Coe and Helpman, 1995 among others). R&D investment stimulates the innovation of new

---

[78] Some kinds of public capital are mainly invested for social welfare rather than for economic development. More investments in healthcare may occupy limited resources that might have been used in other fields which may be more effective and efficient for economic growth at least in short run.

[79] Public transit is more likely regarded as a charitable industry, because the objective is not for profit but social welfare. The price is controlled without enough competition. Moreover, even if some organizations are managed by private, such as WMATA and AmTrack, they still receive a considerable amount of subsidies from government. Another important component in America transportation, airline industry, is included in the category of private capital.

[80] Randal OToole, Are Highway Subsidized? November 2006. http://www.thefreemanonline.org/features/are-highways-subsidized/, visited on May, 16, 2013.

technology, and enhances firms' ability to assimilate and exploit existing information to improve efficiency (Cohen and Levinthan, 1989).

Appropriate human capital is another important factor that boosts productivity (Schultz, 1962; Lucas, 1988; Romer, 1990). It is considered useful to increase the ability of adapting and implementing existing technology or creating new techniques. Its positive and significant impacts on TFP growth are also testified in a lot of studies (Coe et al. 1997; Miller and Upadhyay, 2000 among others). In spite of the existence of strategic complementarities between human capital and R&D (Acemoglu and Zillibotti, 2001), to avoid omitted variables bias, following Bronzini and Piselli's study (2009), human capital is still kept as an additional factor affecting the regional TFP growth.

Another determinant is the agglomeration effect. The theoretical link between agglomeration and productivity is well established (Bannister and Stolp, 1995; Glaeser and Gottlieb, 2009 among others). Firms located in agglomerated areas tend to operate more efficiently, for such firms can take advantage of external economies (e.g. scales economies, localization economics, and urbanization economies) stemming from the spatial proximity to a larger market, a thicker and more varied labor pool, the existence of more entrepreneurial talent and technical expertise, more general knowledge and personal contacts, and more convenient transportation connection. A number of empirical studies have shown that the agglomeration effect may contribute positively to productivity growth (mostly to labor productivity and capital productivity) (Ciccone and Hall, 1996; Cingano and Schivardi, 2004 among others).

Besides, industry structure may also imposes some impacts on total factor productivity growth. The financial intermediary sector alters the path of economic progress by lowering capital flow costs with consequent ramifications for capital allocation, and has been proven exerting a large, positive impact on TFP growth (Beck et al. 2000). Another focus is on information and communication sector, which totally alters the nature of modern economy through rapid diffusion of personal computers, the Internet, mobile technology, and broadband networks (Seo and Lee, 2006). It drastically reduces the cost of information transfer. In other words, this sector ameliorates the efficiency of knowledge creation, utilization, and distribution in knowledge-based economies. It establishes a network that facilitates communication between firms, while also helping to streamline their production processes and lowers transaction costs (Schreyer, 2000).

Furthermore, the TFP growth in each region is influenced by both regional specific characteristics and national economic situation. Hence, a macroeconomic variable is introduced in the model to represent national foreign-trade, openness and so on. For each region, its initial condition also determines corresponding TFP growth (Nehru and Dhareshwar, 1994). In fact, these two control variables refer to part of individual fixed-effects and time fixed-effects, but they are observed and could be introduced in the model as explanatory variables.

3.2.2 Variable Description and Measurement
Following contents describe all variables used in this study and their measurements as well as necessary procedures to deal with raw data.

1. Output: BEA and Census don't provide ready-to-use output data for each urban

59

area. The solution here is to apply BEA's local area personal income as a proxy to calculate each county's GDP and aggregate them to urban level.

$$Q_{i,j}(t) = \left[\frac{PI_{i,j}(t)}{PI_j(t)}\right] Q_j(t)$$

$$Q_s(t) = \sum_{j=1}^{m} \sum_{i=1}^{n} Q_{i,j}(t) \tag{4}$$

where, PI is the personal income, and Q denotes GDP. Subscripts i, j and s corresponds to county, state, and urban area respectively. m and n are the number of states and counties belonging to each urban area.

2. Labor capital stock: because of the infeasibility of data on hours of work in each urban area, the employment in each county is aggregated to achieve the urban level labor capital.

3. Private Capital Stock: The real net non-residential stock of fixed assets (structures, equipment and software) at the end of previous year is used to represent the private capital stock. The Bureau of Economic Analysis only provides capital stock estimates for the nation, so the urban level data needs an approximation. Following Garofalo and Yamarik (2002), the national capital stock estimates is apportioned to each urban area using annual private industry earning as a proxy[81]. For each two-digit BEA industry, the national fixed asset is apportioned by the ratio of this industry's earning within each urban area to its total earning within the whole nation. As a result, the estimate of each

---

[81] Garofalo and Yamarik don't provide theoretical support for the reason to use industry earnings as the proxy. However, they did some tests and proved that this approach has produced real information in estimates of the private capital stock, rather than a simple replication of national relationship between industry income and capital (significant variations among capital stock in different regions, and low explanation power in a simple regression between capital and industry income).

urban area's private capital stock is then the sum of all industries' estimates in this area.

$$K_{i,j}(t) = \left[\frac{E_{i,j}(t)}{E_i(t)}\right] FA_i(t)$$

$$K_s(t) = \sum_{j=1}^{m} \sum_{i=1}^{n} K_{i,j}(t) \tag{5}$$

where E represents the industry earning, FA is fixed asset, and K denotes private capital stock. Subscripts i, j and s indicate industry, county, and urban area severally. m is the number of counties in urban area s, while n refers to the number of industries defined for calculation.

As BEA converted the SIC (Standard Industrial Classification) code to NAICS (North American Industry Classification System) code after 1997[82], BEA provides two reports for fixed assets: the *Fixed Reproducible Tangible Wealth in the United States*[83] series reports, and the *Fixed Assets and Consumer Durable Goods*[84] series reports. The former one covers the period before 1997 using the SIC code to classify industry, while the latter one applies both SIC and NAICS codes in different periods (SIC between 1998 and 2000, NAICS after 2001). Since available data from BEA in both codes don't match available code conversion rules that's only feasible at lower industry level, it is

---

[82] For SIC-code industry, there are nine non-farming industries included in the dataset: 1). farming; 2). agricultural services, forestry, and fishing; 3). mining; 4). construction; 5). manufacturing; 6). transportation and public utilities; 7). wholesale trade; 8). retail trade; 9). finance, insurance, and real estate; 10). services. For NAICS-code industry, this number increases to twenty : 1). farming; 2). forestry, fishing and related activities; 3). mining; 4). utilities; 5). construction; 6). manufacturing; 7). wholesale trade; 8). retail trade; 9). transportation and warehouse; 10). information; 11). finance and insurance; 12). real estate and rental and leasing; 13). professional, scientific, and technical services; 14). management of companies and enterprises; 15). administrative and waste management services; 16). educational services; 17). health care and social assistance; 18). arts, entertainment, and recreation; 19). accommodation and food services; 20). other services, except government.

[83] Musgrave, J. C. (1998). Fixed reproducible tangible wealth in the United States: Revised estimates for 1995-97 and summary estimates for 1925-97. *Survey of Current Business*, 78(9), 36-73.

[84] Bennett, J., Glenn, M., & Wasshausen, D. (2011). fixed assets and consumer durable goods for 1997-2010. *Bureau of Economic Analysis.*

impossible to unify the measurement of some variables across different periods, vice versa, the research period is divided into two parts: before 2000 (based on SIC), and after 2000 (based on NAICS). Missing data marked with 'D' in BEA dataset are those not shown to avoid disclosure of confidential information, but the estimates for such item have been included in the totals. These data could be obtained by extracting other items from sub-total or total number. For those which can not be obtained through extraction, the interpolation method as well as other simple estimates will be applied based on various conditions.

As the data by industry is not always available at county level, following procedure is applied: 1). For missing values with available ones in adjacent (backward and forward) years: a. if the industry doesn't have any affiliation (e.g., construction, retail trade, etc.), the missing value is the average of adjacent two values. b. If the industry has branches (e.g. manufacturing, transportation, etc.), and primary sub-industries' data are available through all years, the average proportion of this majority part in other years could be used as the reference to estimate the missing year's value. 2) For missing values at both ends, the 1-b method is applied if possible. Otherwise, the missing value can be calculated using the total value extracting other industries' values. If more than one industry has missing values in the same year, and none of them could be estimated using above ways, then all of these industries are regarded as an integrity with values equal to the total value subtracting other industries' aggregation.

One shortcoming of this method lies that different county may have various industries with missing value, and this procedure may cause inconsistency in estimation.

However, all industries with missing values belong to the minority part in each area. Their value only covers a very small proportion of total amount (normally less than 5%). Hence, the estimate procedure is still acceptable.

4. Highway Capital Stock: The BEA also provides the fixed asset data of highway and streets which belongs to the category of government fixed assets, but still at nation level. The Census of Government (COG) also provides individual government's [85] finance statistics which contains expenditure on highway. However, the perpetual inventory method is also required to obtain part of the highway capital stock[86], and the capital stock in benchmark year is difficult to calculate. In addition, this data is provided every five years, and the latest available data is 2007[87]. Any value during the interval also needs estimate. Therefore, a similar method as used in estimating private capital stock is applied to distribute national highway/streets capital stock to each county. Albeit, for this public capital, physical characteristics are used as proxy, since monetary information for local highway/streets is not available. Though physical measurement is usually not accurate to compare, considering that only specific transportation infrastructure is measured, it's still acceptable. The stock of highway and street capital can be conducted as shown below:

$$H_j(t) = \left[\frac{LM_j(t)}{LM(t)}\right] H(t)$$

---

[85] It includes municipalities, counties, townships and any other forms of local governments.

[86] In each county, different agencies (federal, state and local government) may own and maintain distinct highway/streets. COG data only provides information on each-level government's expenditure on highway, while for each county, the total highway/street expenditure is actually a combination of different governments' expenditures. As a result, the expenditure from federal and state government still needs to be decomposed to each county and then the capital stock stills needs to be calculated.

[87] The latest 2012 data is advertised but yet published in Census website when this dissertation was finished.

$$H_s(t) = \frac{L_s}{\sum_j^m L_j} \sum_j^m H_j(t) \tag{6}$$

where H indicates the highway and streets capital stock. LM represents lane miles[88], and L is road length[89]. Subscripts s and j imply urban area and state respectively. m is the number of states intervened with the urban area. Both lane miles and road length data could be obtained from annual *Highway Statistics* report, published by FHWA (Federal Highway Administration).[90]

5. Human capital: the percentage of persons who are 25 years and over with bachelor's or higher degree is used as the index of human capital (Berry and Glaeser, 2005)[91]. The Census Bureau provides this index for every ten years (1980, 1990, 2005~2009 in our case). The missing value is estimated using the interpolation method, assuming a constant growth rate in 1980s, 1990s, and 2000s, separately.

6. Congestion Index: the National Transportation Statistics[92] provide three types of congestion indices: a). amount of delay or wasted fuel caused by congestion, totally or individually; b). Travel Time Index; and c). Annual Roadway Congestion Index. The measurement on amount, no matter time delay or wasted fuel, could not represent the

---

[88] Roads with different functions have various lanes. The capital layout and maintenance expenditure on a 4-lane road should be higher than a 2-lane road with the same length, though the ratio between them may not be two. Using lane miles instead of length should be more accurate.

[89] At MSA level, data of lane miles are not available. To calculate the ratio, road length is used.

[90] An option to estimate road length is using the ArcGIS software to measure the length of each road segment and sum them up together. The 2011 TIGER/Line road shapefiles published by Census Bureau provides the data source. However, there is only one year data, so the ratio is constant during all periods. Though relying on some previous data collection work, I find that the ratio changes so slightly that could be regarded constant to some extent. I prefer using the data from FHWA, as its data are time variant.

[91] A better measurement of human capital should be based on subjects of degrees, since various fields may contribute differently to economic growth, such as the distinction between science and art.

[92] Research and Innovative Technology Administration, Bureau of Transportation Statistics. Retrieved from http://www.bts.gov/publications/national_transportation_statistics/#chapter_1 on October, 18, 2012.

congestion degree accurately, since it is influenced by the size of the transportation system. The roadway congestion index is a measure of vehicle travel density on major roadways. It is a direct comparison of miles traveled with the miles of road available to travel on. The shortcoming of this index is that it does not take into account many factors that influence congestion, such as accidents or incidents on road segments or bad weathers. The travel time index (TTI) is the ratio of actual travel time in peak hours to the supposed travel time in free-flow conditions for the same trip, measuring the amount of additional travel time required due to heavy traffic, roadway incidents or accidents, or any other factor that dispute traffic. Though its value is still influenced by enactment of free-flow condition, this ratio is comparable among different areas. Hence, *TTI* is used to measure the congestion level here.

7. Research and Development: two methods could be used to estimate the R&D level in each urban area: the ratio of annual R&D investment to the total GDP in each urban area (Huang and Liu, 1994), and the number of utility patents per 10,000 people in the same area (Pakes and Griliches, 1980). The former one measures resources devoted into R&D, while the latter one partially indicates the R&D capability, as not all innovations will apply for patents, and only a few of them will be granted as patents as well. The amount of R&D fund can be obtained from the *Annual Research and Development in Industry* series report, published by NSF (National Science Foundation). However, this data is only available at state level. A rough estimate is to use the regional employment as the proxy to apportion the state level data to county level and then sum up the county data to corresponding urban area. Albeit, the R&D investment is distributed unevenly in

companies with various sizes (employment). It's inconvincible to simply use employment as a proxy. A better choice is to use the employment by size of company or by industry as the proxy. NSF issues the R&D investment data categorized by size of company and by industry in each state. Meanwhile, employment data by size of company are available in Census Bureau, and employment by industry at county level could be obtained from BEA. Nevertheless, data on total (company, Federal, and other) industrial R&D fund by industry and company size at state level only covers the period after 2003. In period from 1995 to 2002, it only includes top 10 states. Therefore, to keep data's consistency, the R&D investment index is omitted and the number of utility patents per 10,000 people is employed as the R&D index. The patent data at county level is available in the database of the United States Patent and Trademark Office (USPTO) covering the periods 1990-1999, and 2006-2010. The data between 2000 and 2005 can be achieved in the Institute for Strategic and Competitiveness at Harvard Business School. [93]

8. Agglomeration Index: There are two concepts of agglomeration, the spatial agglomeration for an industry and the sectorial agglomeration for an area (Ceapraz, 2008). If most plants of an industry are located in a certain area with a few dispersed in other areas, the industry has a high spatial agglomeration (concentration) level. If most industries gather in a certain area, while other areas have a few industries, it could be regarded that this area has a higher sectorial agglomeration level. Indeed, the sectorial

---

[93] On their website, only MSA-level patent data between 1998 and 2009 is obtainable, while I already have urban-level patent data from 1990 to 1999, and from 2006 to 2009. To avoid redundant estimates that may influence data's accuracy and keep data's consistency, the MSA-level patent data between 2001 and 2009 is used to substitute corresponding urban-level patent data in this research. Based on my calculation, the number of utility patents in our targeted urban area occupies more than 85% of that in corresponding MSAs.

agglomeration is synonymous with the industrial specialization. A great number of

indices are provided to measure an industry's spatial agglomeration, such as Gini

coefficient and Herfindahl index.[94] Since the research unit is urban area, the sectorial

agglomeration is concerned rather than the spatial agglomeration. A simple

transformation of spatial agglomeration indices could calculate the sectorial

agglomeration A low specialized area implies that more various industries are

concentrated there, which are beneficial for interaction among industries and knowledge

spillover, especially in current highly divided economic system. Because the sectorial

agglomeration level among different regions must be comparable in the model, the

Herfindahl index is calculated using employment in each industry.

$$H_i = \sum_j s_{i,j}{}^2$$
$$s_{i,j} = \frac{E_{i,j}}{\sum_j E_{i,j}} = \frac{E_{i,j}}{E_j} \qquad (7)$$

where $s_{i,j}$ is the share of the employment E in the industry j of the region i in the total

employment E_j of the region i. However, it may not sufficiently indicate how flourishing

the region's economic activities are. Thus, the density of establishment of each industry

in each area is also introduced. The number of establishment data is available at county

level since 1986 in County Business Pattern database of U.S. Census Bureau. An

integration is necessary to obtain the urban-level data.

9) Industry structure: as discussed in the above section, in this dissertation, only

communication industry and finance industry are introduced in the model to control the

impacts of change in economic structure on economic efficiency. The revenue share of

---

[94] Others include EG-index, MS-index, D-index, K-density and so on.

each industry[95] in each area's total industry revenue is used to represent the industry's relative importance.

10) Macro-economic institution and policies. Since all urban areas in this study are located in contiguous 48 states of the United States, they are influenced by the same economic system and federal policies, though each urban area has its own different features owing to respective political system (Democracy or Republican). The national GDP is used to represent macro-economic situation that covers foreign trade, market openness, political stability and so on.

11) Initial economic condition: the per capita income in the first year of each period is used to denote the initial economic condition of every urban area.[96]

3.3 RESEARCH SUBJECTS AND PERIODS

To better estimate congestion's impacts on regional economic efficiency, analysis is conducted at the urban area level. There are not ready-to-use urban data for most variables, so we have to aggregate data of counties which are defined as components of corresponding urban area. Counties located totally within an urban area and those which more than 1/5 area is contained in the urban area[97] or those in which 1/5 of urban area is contained are defined as qualified components of an urban area. The ArcGis software is applied to complete this task. Take the New York urban areas as an example. In the

---

[95] In both SIC category and NAICS category, information sector refers to traditional publishing, issuing, and printing (books and newspapers), which is different from modern digital information concept. In NAICS category, there are two communication related sectors: telecommunication and internet service providers, both of which are integrated to denote the communication sector.
[96] Employment is another option, but income is measured with monetary value which covers more information, in both quantity and quality.
[97] Meanwhile, no other urban area overlaps the same county.

following map, shaded areas with orange color are counties which are considered as parts

of the urban area with blue color. The larger one with net-grid refers to the New York

MSA. Obviously, NY-MSA is larger than corresponding NY-urban area.
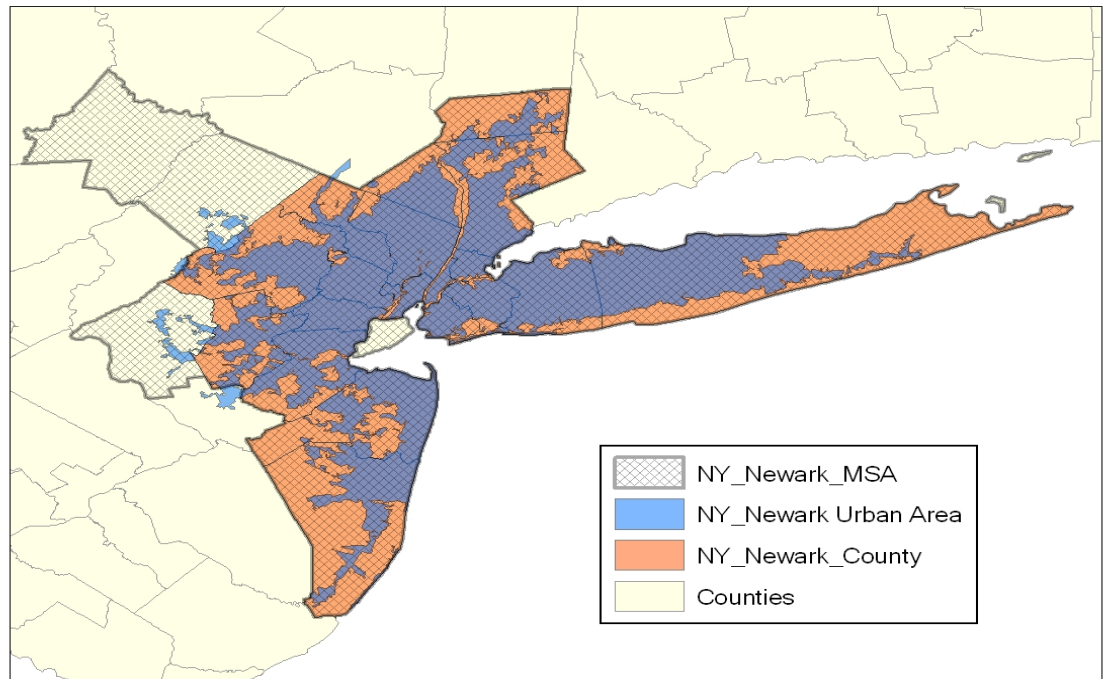
Figure 3-1 Map of Defined NY Urban Area

Only urban areas located in 48 American contingent states except those in West are

considered. The primary reason is that a considerable amount of land in western states is

owned by federal government. Since federally owned land can't be traded freely in market,

in the West, firms could not locate themselves completely free. This fact may affect the

distribution layout of enterprises in a relatively broad area, which is decisive in calculation

69

of the agglomeration index. To avoid potential bias, only urban areas located in other

regions are selected.[98] Generally, the larger the urban area, the higher the congestion level.

Thus, large and very large urban areas defined by FHWA in total of 31 are covered.

Figure 3-2 Federal Land as a Percentage of Total State Land Area
Data source: U.S. General Services Administration, Federal Real Property Profile 2004, excludes trust properties.

3.4 TIME FRAME

Traffic congestion not only slows down vehicle speeds on roads, but also affects

adjacent residential and business activities. Increased transportation cost may result in

---

[98] Though federal government still owns land in these states, the percentage is so small that could be ignored.

changes in activities of firms and individuals in a long run, such as relocation, alternation

of working schedule, and so on. Limited by the data availability of several primary

variables, the paper covers the period between 1990 and 2009. Furthermore, several

primary variables, including private capital stock, agglomeration index and industry

share, are calculated based on detailed industry categories which definitions change with

the conversion from SIC to NAICS in late 1990s. Correspondingly, their values are not

comparable in different periods. To guarantee the consistency of variables, the research

era is divided into: 1990~2000 and 2001~2009. On one hand, it avoids the estimation

bias due to data's inconsistency. On the other hand, this also makes it possible to

compare traffic congestion's impacts in various periods.

CHAPTER 4   IMPACTS OF TRAFFIC CONGESTION ON REGIONAL ECONOMIC EFFICIENCY: PARAMETRIC ANALYSIS

In this section, parametric analysis is executed to calculate the economic efficiency in each urban area, and successively a further investigation is conducted to analyze how traffic congestion influences regional economic efficiency.

4.1 MEASUREMENT OF TFP GROWTH

The total factor productivity (TFP) growth is used as the index of economic efficiency. Different from labor productivity (output per hour/labor) or capital productivity (output per unit of value of fixed production assets), the measure of TFP growth treats all inputs systematically. In the production function context, the TFP growth can be defined as the portion of output growth not explained by the amount of inputs growth.[99] As such, its level is determined by how efficiently and intensely the inputs are utilized in production. Economists also recognize that TFP growth acts as a determinant factor in long-term growth process (Solow, 1956, among others).

The TFP growth could be measured using the conventional Divisia index, $\dot{TFP} = \dot{Y} - \sum_j S_j \dot{X}_j$, where Y is output, X_j is a vector of inputs ($j = 1, 2, \dots \dots, J$), and S_j is the share of input X_j in the total cost. $S_j = w_j X_j / C$, w_j being the price of input X_j, C is the observed cost, and a dot over a variable denotes its rate of change. Solow (1957) develops this concept, applying it in a Cobb-Douglas production function, and gets

---

[99] These non-input factors include technological progress, economies of scale, capacity utilization, market efficiency and others that can make the use of inputs more efficient or effective and enable higher output using the same quantities of inputs.

$T\dot{F}P = \dot{A}/A$, where A denotes the Hicks-neutral technology. Based on Solow's assumptions, the TFP growth should be equal to the technical change (progress). Ohta (1974) and Denny et al. (1981) have argued that only in a single-product case, with constant return to scale and cost efficiency, can TFP growth equalize the technical change. With non-constant returns to scale or imperfect input market, the techical change only contributes partially to the TFP growth (Kumbhakar, 2000).

Solow residuals require continuous data, while in reality data are always discrete[100]. Jorgenson and Griliches (1967) introduce the Tornqvist index correspondingly, and Diewert (1976) proves that if the production function has the form of transcendental logarithm, its Tornqvist index is an exact index.[101] Since the price data is not always available, instead of calculating input shares, scholars usually estimate these parameters using production function models. On the premise of market equilibrium, the input share should be equal to the output elasticity with respect to the corresponding input.[102]

In this chapter, a three-input production function, $Y_{i,t} = F(L_{i,t}, K_{i,t}, G_{i,t})$, is applied to calculate the TFP growth. Y measures the quantity of output, L denotes the input of labor services, K is the input of private capital services, and G represents the input of highway and streets capital. Two subscripts i and t denote region and time, respectively.

---

[100] Economic data are always recorded annually, quarterly, monthly or even daily, but strictly speaking, they are still discrete and couldn't be used in calculus.

[101] Based on Jorgenson and Griliches' method, the average TFP growth in adjacent two years could be measured using following equations: $T\dot{F}P = \frac{1}{2}\ln\frac{Y_t}{Y_{t-1}} - \sum_j \overline{S}_J \ln\frac{X_t}{X_{t-1}}$, $\overline{S}_J = \frac{1}{2}[S_{j,t} + S_{j,t-1}]$.

[102] Necessary conditions for a producer equilibrium in each market are given by equalities between the value shares and elasticities of output with respect to the corresponding inputs, $S_j = \partial \ln Y / \partial \ln X_j$. (Christensen et al. 1981)

Compared with the widely utilized Cobb-Douglas production function, the translog production function (Christensen, Jorgenson, and Lau, 1973) has less constraints on C-D's rigid premises. The elasticity of substitution between input factors is permitted to vary and the linear relationship between the output and inputs (in logarithm terms), which are taken into account in $C - D$ production function, is passed to a non-linear one. Assumptions of Hicks-neutral technological change[103] and the constant return of scale in $C - D$ production function are also revoked. In conclusion, the translog production function is more flexible in form, and hence more approximated to the real economy. In effect, Antras (2004) concludes that the U.S. economy is not well described by a $C - D$ aggregate production function after dealing with the nonsphericality of the disturbances, the endogeneity of the regressors, and the nonstationarity of the series involved in the estimation.

The translog form may generally be viewed as a second order Taylor approximation to an unknown aggregate production function (Hoff, 2002). In the case of three inputs, the trasnslog production function with time index, which means the production function itself is allowed to shift over time to account for technological change (Jorgenson and Nishimizu, 1978), is as following:

$$
\begin{aligned}
\ln Y_{i,t} = {} & \alpha_0 + \alpha_L \ln L_{i,t} + \alpha_K \ln K_{i,t} + \alpha_G \ln G_{i,t} + \lambda T + \beta_{LK} \ln L_{i,t} \ln K_{i,t} \\
& + \beta_{LG} \ln L_{i,t} \ln G_{i,t} + \beta_{KG} \ln K_{i,t} \ln G_{i,t} + 0.5\beta_{LL} (\ln L_{i,t})^2 \\
& + 0.5\beta_{KK} (\ln K_{i,t})^2 + 0.5\beta_{GG} (\ln G_{i,t})^2 \\
& + \gamma_{LT} T \ln L_{i,t} + \gamma_{KT} T \ln K_{i,t} + \gamma_{GT} T \ln G_{i,t} + 0.5\gamma_{TT} T^2
\end{aligned}
$$

(8)

---

[103] A technological innovation is called Hicks neutral, following the context of Hicks (1932), if a change in technology does not change the ratio of one input's marginal product to another input's marginal product for a given ratio between these two inputs quantities.

where α, β and γ are unknown parameters to be estimated. T denotes the time index, which could be either the real year or the distance from the benchmark year. The parameters γ_{LT}, γ_{KT} and γ_{GT} measure biases of the technological change. If one of them is positive, the corresponding value share increases with technology, and it is called an input-using bias. If it is negative, the value share decreases with technology, and technical change is input-saving. Finally, if it is zero, the value share is independent of technology, i.e., a neutral technical change.[104] The following output elasticities with respect to each input could be achieved:[105]

$$\varepsilon_{K_{i,t}} = \frac{\partial \ln Y_{i,t}}{\partial \ln K_{i,t}} = \alpha_K + \beta_{LK} \ln L_{i,t} + \beta_{KG} \ln G_{i,t} + \beta_{KK} \ln K_{i,t} + \gamma_{KT}T \tag{9}$$

$$\varepsilon_{L_{i,t}} = \frac{\partial \ln Y_{i,t}}{\partial \ln L_{i,t}} = \alpha_L + \beta_{LK} \ln K_{i,t} + \beta_{LG} \ln G_{i,t} + \beta_{LL} \ln L_{i,t} + \gamma_{LT}T \tag{10}$$

$$\varepsilon_{G_{i,t}} = \frac{\partial \ln Y_{i,t}}{\partial \ln G_{i,t}} = \alpha_G + \beta_{LG} \ln L_{i,t} + \beta_{KG} \ln K_{i,t} + \beta_{GG} \ln G_{i,t} + \gamma_{GT}T \tag{11}$$

Finally, the TFP growth equation could be expressed as:

$$T\dot{F}P = \dot{Y}_{i,t} - \varepsilon_{L_{i,t}}\dot{L}_{i,t} - \varepsilon_{K_{i,t}}\dot{K}_{i,t} - \varepsilon_{G_{i,t}}\dot{G}_{i,t} \tag{12}$$

4.2 STEPS FOR ANALYSIS

4.2.1 Principle Component Analysis

The translog production function has an intrinsic problem of multi-collinearity, so the principle component analysis (PCA) is used. The Table 4-1 displays the pairwise correlation matrix among all independent variables in the production function. The matrix

---

[104] Jorgenson and Weldeau. Econometrics: Econometric modeling of producer behavior. 2000. MIT Press
[105] According to elasticity definition, if $Q = F(X)$, then $\varphi_X = (\Delta Q/Q)/(\Delta X/X)$. Let $\ln X = u$ and $\ln Q = v$, we can get $X = e^u, Q = e^v$, so $e^v = F(e^u)$. Taking derivatives of both sides of this equation with respect to u and $\frac{d(e^u)}{du} = e^u \frac{dF(e^u)}{d(e^u)}$, so $\frac{dv}{du} = e^u \frac{dF(e^u)}{d(e^u)\cdot e^v}$, $\therefore \frac{d \ln Q}{d \ln X} = X \cdot \frac{dQ}{QdX} = \frac{dQ}{Q}/\frac{dX}{X}$, with discrete data, $\frac{dQ}{Q}/\frac{dX}{X} = \frac{\Delta Q}{Q}/\frac{\Delta X}{X}$, therefore, we can get $\frac{d \ln Q}{d \ln X} \equiv \frac{\partial \ln Q}{\partial \ln X} = \varepsilon_X$.

shows that there are two groups within which variables are highly correlated (above 0.9):

variables with the time factor and those without that. Correlation coefficients between

variables across two groups are really low (less than 0.3). Following Cavatassi et al.

(2004), I pool the data for all years in each research period and estimate the principle

components over the combined data. The obtained weight is then applied to the variable

for each round in the dataset, and thus the PCA value for each unit in every year is

obtained. A principle assumption is that the impact of included variables during each

period has been averaged. However, this method doesn't change the relative importance

of each variable in determining the index, and any instance change in importance of

variables are essentially averaged over all years. Moreover, this approach helps to

facilitate components' consistency and comparability over time.

An important step of PCA is to reasonably assess the number of components,

without either under-extraction or over-extraction of them. The Parallel Analysis (PA)

has been proven consistently accurate in determining the threshold for significant

components, variable loadings, and analytical statistics (Franklin et al., 1995). It suggests

that in each period, two principle components, each of which explains more than 96% of

variances of all covered variables, should be selected (shown in Figure 4-1 and Table 4-

2). Thus, two components have involved most information of raw data.

Each principle component could be calculated based on formula: $PCA_i =$

$\sum_{i=1}^{k} \theta_i X_i$, where X denotes explanatory variables in translog production function model,

and k is the number of them. θ refers to the coefficient of each variable in PCA. Then a

model with two independent variables is established.

$$lgdp_{i,t} = \beta_1 PCA_{1,i,t} + \beta_2 PCA_{2,i,t} + \mu_i + \delta_t + \varepsilon_{i,t} \tag{13}$$

where $lgdp = \log(GDP)$. β_1 and β_2 are parameters to be estimated.

4.2.2 Spatial Analysis

Since research units in this paper are urban areas that have geographic features, it is worthwhile to detect whether spatial associations exists according to Tobler's First Law: "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). Two geographical entities are said related if there is a positive or negative correlation between these entities, and spatial association does not necessarily imply causality (Miller, 2004). Either a casual relationship, such as spillover effects, or other hidden variables omitted in the model that may result in measurement errors, could lead to spatial autocorrelation. However, it seems dubious to determine the existence of spatial association without taking any specific test, even though most urban areas are not adjacent to others and the distance between them varies a lot, from tens of miles[106] (Washington D.C. to Baltimore) to thousands of miles (New York to St. Antonio). Hence, two tests are implemented here: one for individual variable and the other for the integrated model.

The first test applies the Moran's I index[107], which measures the spatial autocorrelation based on both feature locations and feature values simultaneously..

---

[106] The distance is calculated between the center of an urban area and its counterpart in another one. The center point is located in the form of longitude and latitude provided in Google map.

[107] Another similar and popular index is Geary's C index: $C = \frac{(n-1)\sum_i^n \sum_j^n w_{ij}(y_i-y_j)^2}{2\sum_i^n \sum_j^n w_{ij}\sum_j^n(y_i-\bar{y})^2}$ (Geary, 1954). The difference between Moran's I and Geary's C is that the former is a more global measurement and sensitive to extreme value of variables, whereas Geary's C is more sensitive to differences in small neighborhoods. Generally, they result in similar conclusions, but Moran's I is preferred in most cases since it is consistently more powerful than Geary's C (Cliff and Ord, 1975).

$$I = \frac{n}{\sum_i^n \sum_j^n w_{ij}} \cdot \frac{\sum_i^n \sum_j^n w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_i^n (y_i - \bar{y})^2} \qquad (14)$$

where w_{ij} is the weight matrix, and here $w_{ij} = \frac{1}{d_{ij}}$ since urban areas are distributed

sparsely[108] and the reciprocal of distance reflects the nearness of geographical entities.

d_{ij} is the Euclidean distance calculated based on longitude and latitude of the center of

each urban area, i.e. the straight-line segment connecting two locations[109]. y denotes the

value of each variable in this research. Moran's I has a value between -1 and 1. If $I > 0$,

it implies positive correlation; if $I < 0$, it denotes negative correlation; and if $I = 0$, it

represents non-correlation. Moreover, this index needs to be calculated year by year, as

it's a static index, and all variables are tested with the Moran's I index[110] (See Table 4-3).

---

[108] For concentrated areas, the weight matrix of values $(0, 1)$ is more frequently used. The value of 1 implies that two entities share a boarder (the Rook contiguity) or a vertex (the Bishop contiguity), and otherwise it's 0. (LeSage, 1999)

[109] In past decades, more and more scholars recognized that transportation and communication technologies have shrunk the world in an incredible degree, thus the nearness could be measured through considering both space and time (Couclelis and Getis, 2000; Janelle, 1995). However, this measure is much more complex, and the Tolber First Law is still proven valid even in current era (Couclelis, 2000; Graham and Marvin, 1996).

[110] In general, only the dependent variable is tested since it has included information of independent variables. However, this case assumes that all independent variables have spatial impacts in both the same magnitude and the same sign. Only considering the dependent variable may result in aggregated bias: some independent variables have positive spatial correlation while some others have negative correlation, and their aggregations may produce insignificant conclusions.

Table 4-1 Correlation Among Independent Variables

	lpc	lemp	lhic	lpc2	lemp2	lhic2	lpcemp	lpchic	lemphic	t	t2	ltpc	ltemp	lthic
lpc	1													
lemp	0.910	1												
lhic	0.910	0.846	1											
lpc2	0.999	0.908	0.912	1										
lemp2	0.927	0.998	0.865	0.926	1									
lhic2	0.917	0.851	0.999	0.921	0.871	1								
lpcemp	0.984	0.968	0.906	0.984	0.978	0.914	1							
lpchic	0.976	0.896	0.977	0.979	0.916	0.981	0.967	1						
lemphic	0.950	0.947	0.971	0.952	0.959	0.974	0.973	0.983	1					
T	0.0580	0.0229	0.124	0.0564	0.0232	0.120	0.0423	0.0910	0.0810	1				
t2	0.0555	0.0213	0.114	0.0541	0.0217	0.110	0.0403	0.0846	0.0744	0.975	1			
Ltpc	0.194	0.148	0.246	0.193	0.150	0.243	0.177	0.224	0.210	0.988	0.964	1		
Ltemp	0.163	0.138	0.220	0.161	0.139	0.217	0.154	0.194	0.190	0.992	0.967	0.998	1	
Lthic	0.188	0.145	0.264	0.187	0.148	0.261	0.173	0.230	0.219	0.987	0.963	0.998	0.996	1

Table 4-2 Results of PCA in Both Periods

Component	1990-2000			2001-2009		
	Eigenvalue	Proportion	Cumulative	Eigenvalue	Proportion	Cumulative
1	8.934	0.638	0.638	8.789	0.628	0.628
2	4.561	0.326	0.964	4.709	0.336	0.964
3	0.280	0.0200	0.984	0.309	0.0221	0.986
4	0.176	0.0126	0.997	0.146	0.0104	0.997
5	0.0409	0.00290	0.999	0.0398	0.00280	1.000
6	0.00439	0.000300	1.000	0.00353	0.000300	1.000
⋮						
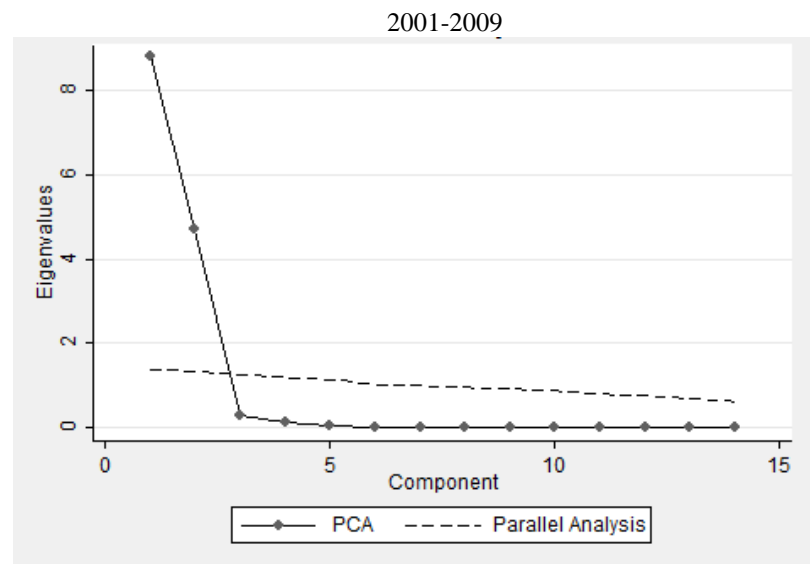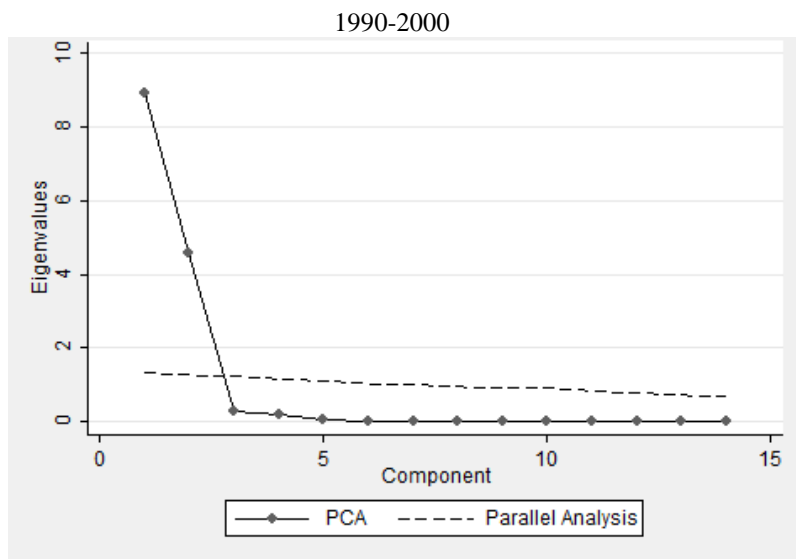
Figure 4-1 Parallel Analysis of PCA

Table 4-3 Results of Moran's I Index of Each Variable

year	Igdp	PCA_1	PCA_2
1990	0.045*	0.016	0.017
1991	0.041*	0.016	0.017
1992	0.037	0.006	0.008
1993	0.034	0.003	0.005
1994	0.028	0.001	0.003
1995	0.024	0	0.002
1996	0.02	-0.002	-0.001
1997	0.018	-0.003	-0.002
1998	0.015	-0.005	-0.004
1999	0.015	-0.006	-0.005
2000	0.014	-0.007	-0.006
2001	0.020	-0.013	-0.011
2002	0.021	-0.010	-0.009
2003	0.022	-0.007	-0.006
2004	0.022	-0.008	-0.007
2005	0.022	-0.010	-0.009
2006	0.021	-0.010	-0.008
2007	0.023	-0.010	-0.008
2008	0.027	-0.008	-0.005
2009	0.033	-0.008	-0.004

*: significant at 10% level.

The results show that absolute values of all three variables' Moran's I indices are even no more than 0.05 during the whole period, and correspondingly all of them are not statistically significant at 5% level.

In succession, whether the model is spatially dependent is detected using two models: spatial error and spatial lag. The spatial error model incorporates spatial effects through error term as: $y = \beta x + \varepsilon$ and $\varepsilon = \lambda W \varepsilon + \xi$, where ε is the vector of error terms, spatially weighted using the weight matrix W. λ is the spatial error coefficient, and ξ is a

vector of uncorrelated error terms. The spatial lag model incorporates spatial dependence

by adding a spatially lagged dependent variable on the right-hand side of the regression

equation[111]: $y = \rho W y + \beta x + \varepsilon$, where ρ is the spatial coefficient, and x is a matrix of

observations on the explanatory variables. Both spatial error and spatial lag effects may

result in incorrect variance estimates in OLS if they are not considered. The difference

between them lies that spatial error effect doesn't result in biased coefficient estimates,

while spatial lag effect does. Using STATA, significances of both effects are detected

(See Table 4-4).

Table 4-4 Results of Spatial Lag Effect Tests

Year	Spatial error	Spatial lag	Year	Spatial error	Spatial lag
1990	0.428	2.3	2000	0.05	1.819
1991	0.169	1.96	2001	0.181	0.4
1992	0.087	2.961	2002	0.135	0.167
1993	0.217	3.249	2003	0.136	0.048
1994	0.131	3.604	2004	0.241	0.021
1995	0.029	3.393	2005	0.281	0.028
1996	0.002	3.398	2006	0.189	0.043
1997	0.002	2.792	2007	0.34	0.016
1998	0.005	2.555	2008	0.871	0.492
1999	0.032	1.876	2009	1.121	0.035

For both model types, their robust Lagrange multiplier values demonstrate that

there isn't significant spatial association in terms of either spatial error or spatial lag. In

conclusion, spatial elements could be omitted in the model establishment and estimates

based on the original model are reliable.

---

[111] A more complex form is to add spatially lagged independent variables on the right-hand side of the regression equation. For simplicity, it's not discussed in this paper.

4.2.3 Stationary Test

It is shown that all three variables could be regarded stationary after controlling intercept, trend or neither of them in both LLC and ADF-Fisher tests, based on results displayed in Table 4-5 and 4-6. Thus, no further co-integration test is required and the original model could be regressed without concerning spurious regression.

4.2.4 Hausman Test

Two Hausman tests are applied: the most frequently used one with the assumption of homoscedasticity across sections, and the other one robust to heteroskedasticity as a comparison, as shown in Tables 4-7. For panel data, it's arbitrary to determine the absence of heteroskedasticity, so the results of the second Hasuman test are used as criteria. In phase I (1990-2000), two Hausman tests provide distinct results, while in phase II, their results are similar. In general, a more robust estimate is preferred. Thus, in phase I, a random-effect model is used, while in phase II, a fixed-effect model is implemented.

4.2.5 Statistical Tests for Panel Data

In both models across all periods, heteroskedasticity, auto-correlation and cross-section dependence are necessary to consider in following regressions based on corresponding test results from Table 4-7 to Table 4-10.

Table 4-5 Hausman Test Results

Ordinary Hausman Test				Hausman Test Robust to Heteroskedasticity			
Phase I		Phase II		Phase I		Phase II	
κ^2	p-value	κ^2	p-value	κ^2	p-value	κ^2	p-value
15.99	0.0003	34.08	0.0000	1.671	0.4337	17.32	0.0002

Table 4-6 Levin-Lin-Chu unit-root test Results

H0: Panels contain unit roots   H1: Panels are stationary

LGDP				PCA_1				PCA_2			
1990-2000		2001-2009		1990-2000		2001-2009		1990-2000		2001-2009	
Statistic	p-value	Statistic	p-value	Statistic	p-value	Statistic	p-value	Statistic	p-value	Statistic	p-value
-2.1418	0.0161	-6.9115	0.0000	-9.8116	0.0000	-7.8036	0.0000	-4.1196	0.0000	-18.0082	0.0000

Table 4-7 Augmented Dickey-Fuller Fisher unit-root test Results

Ho: All panels contain unit roots   H1: At least one panel is stationary

	LGDP				PCA_1				PCA_2			
	1990-2000		2001-2009		1990-2000		2001-2009		1990-2000		2001-2009	
	Statistic	p-value	Statistic	p-value	Statistic	p-value	Statistic	p-value	Statistic	p-value	Statistic	p-value
P	319.5	0.0000	113.7	0.0001	349.4	0.0000	225.0	0.0000	457.3	0.0000	181.4	0.0000
Z	-3.391	0.0003	-3.735	0.0001	-8.439	0.0000	-3.074	0.0011	-11.17	0.0000	-2.305	0.0106
L*	-10.89	0.0000	-3.682	0.0002	-15.16	0.0000	-7.091	0.0000	-20.52	0.0000	-5.262	0.0000
Pm	23.13	0.0000	4.643	0.0000	25.81	0.0000	14.64	0.0000	35.50	0.0000	10.73	0.0000

P: Inverse chi-squared   Z: Inverse normal   L*: Inverse logit t   Pm: Modified inverse chi-squared

Table 4-8 Heteroskedasticity Test Results

H0: no groupwise heteroskedasticity

Phase I		Phase II	
Likely Ratio κ^2	*p-value*	*κ^2 **	*p-value*
346.99	0.0000	6670.02	0.0000

Table 4-9 Autocorrelation Test Results

H0: no first-order autocorrelation

Phase I		Phase II	
ALM (lambda=0)	*p-value*	*F(1, 30)**	*p-value*
17.86	0.0000	37.153	0.0000

Table 4-10 Cross-Section Dependence Test Results

H0: no cross-section dependence

Phase I		Phase II	
Friedman's statistic	*p-value*	*Pesaran's statistic**	*p-value*
62.293	0.0005	16.081	0.0000

*: The estimators used for corresponding tests in FE model are different from their counterparts in RE model.

4.2.6 Endogeneity Test

The Eviews software is used to run the panel Granger Causality test, and very low probabilities (0.03 and 2E-32) (Table 4-11) demonstrate that both principle components should be considered as potential endogenous variables in the model. Then the STATA software is used with the endogeneity Hausman Test (Table 4-12) to detect whether the regression model is endogenous. Results illustrate that the null hypothesis is rejected so that there exist systematic differences between coefficients resulted from the model including endogenous variables and that including exogenous variables. Hence, it's preferred to apply GMM. In order to achieve relatively precise estimates, the $SYS-$

GMM is employed, and corresponding Arellano-Bond test and Hansen test also pass

(Table 4-13).

Table 4-11 Pairwise Granger Causality Tests
Sample: 1990 2000   Lags: 2

Null Hypothesis:	F-Statistic	Prob.
LGDP does not Granger Cause PCA_1	3.554	0.0299
LGDP does not Granger Cause PCA_2	96.72	2.E-32

Table 4-12 Endogeneity Hausman Test *
H0: difference in coefficients not systematic

Phase I		Phase II	
κ^2	p-value	κ^2	p-value
93.11	0.0000	40.87	0.0000

*: For FE model, STATA also provides another exogeneity test, the Davidson-MacKinnon statistic, and the values is $F(2,182) = 3.197$ and P-value = .0432.

4.3 FIRST STAGE REGRESSION
Using $SYS - GMM$, coefficients of two principle components are estimated.

Table 4-13 Results of *SYS-GMM*

Dependent Variable: Lgdp	Phase I	Phase II
PCA_1	.2717*	.2663*
PCA_2	-.0676*	-.0774*
constant	10.86*	11.34*
Wald chi2	234.57*	541.67*
Arellano-Bond test for AR(1) in first differences	1.94*	-1.44
Arellano-Bond test for AR(2) in first differences	0.77	0.21
Hansen test of overid. restrictions	29.38	29.77

*: significant at 5% level.

The Wald χ^2 test shows that both models are significant. Results of Arellano-

Bond test and Hansen test also satisfy requirements of using $SYS - GMM$. As a result,

coefficient γ of each independent variable is obtained. Integrated with expressions of

principle components, the final equation could be expressed as:

$$lgdp = \alpha_0 + \gamma \cdot PCA = \alpha_0 + \sum_{i=1}^{2} \gamma_i \sum_{j=1}^{k} \theta_j x_j \qquad (15)$$

The coefficient of each real variable in translog production function is calculated, and output elasticities with respect to each input variable could be obtained. Successively, the TFP growth for each individual unit is calculated following the Equation 8.

4.4 SECOND STAGE REGRESSION

It is turn to detect the relationship between traffic congestion and the TFP growth, and the corresponding multiple regression is established as following:

$$T\dot{F}P = \gamma Z + \mu_i + \delta_t + \varepsilon_{i,t} \qquad (16)$$

where Z denotes the set of all determinant factors. A particular issue in this stage is that the data source of agglomeration index uses SIC code before 1997 (including 1997) and after 1997 the NAICS code is applied, which doesn't match data sources of other variables. Thus, two models are established in phase I: one includes all variables excluding the agglomeration index for the period 1990-1997; and the other one covers all variables for the period 1990- 2000.

Test results show that all variables are stationary, and there isn't severe pairwise correlation among them[112]. The result of Hausman test shows that a RE model is preferred, but the Breusch and Pagan Lagrangian multiplier test[113] detects that the OLS should be used[114]. Granger causality test also demonstrates that $T\dot{F}P$ is not the Granger cause of any independent variable. Hence, the endogeneity issue is omitted in this model.

---

[112] All pairwise correlation coefficients are below 0.7 and VIF test as well as Collin test doesn't show any multi-collinearity problem.

[113] The LM-BP test helps us decide between a random effects regression and a simple OLS regression. The null hypothesis is that variances across entities is 0, i.e. no significant differences across units (no panel effect).

[114] When the linear OLS regression is used, the heteroscedasticity needs to be tested to make sure the regression estimators are robust. The White Test is used for homoskedasticity against unrestricted forms of heteroskedasticity.

Table 4-14 Regression Results with TFP Growth as Dependent Variables

TFP growth	*Phase I*		*Phase II*
	1990-1997	1990-2000	2001-2009
Traffic Congestion	-.204**	-.137 **	.146**
Human Capital	.148	.108*	-.148**
R&D	.0005	.001	-.0002
Communication	.138	.030	.100
Finance	-.403*	-.172	-.055
Density	1.33 e-05	7.72 e-06	1.36 e-05
Sectorial Agglomeration	-.226		-.048
National GDP	.221**	.087**	-.013
Initial Per capita Income	.292	.126	-.324
F	3.08**	2.89**	2.94**

*: significant at 10% level; **: significant at 5% or lower level.

Table 4-14 represents that during the period between 1990 and 2000, in both

models each parameter's coefficient has the same sign (positive or negative direction)

and the same significance level. The national GDP is significant, demonstrating the

positive contribution of macro-economic condition to TFP growth. Traffic congestion

always has significant and negative impacts on the TFP growth, which complies with

conventional theoretical analysis that traffic congestion damages the efficiency of social

and business activities and further negatively influences the total factor productivity.

However, during the period between 2001 and 2009, traffic congestion becomes a

significant contributor to the TFP growth. This finding stirs me to dig deeper for the

truth. Dumbaugh (2012) detects whether vehicle delay had a negative effect on urban

economies. His result also demonstrates that being clogged in traffic leads people to be

more productive. Integrating his advice with my understanding, I summarize following

factors that may explain above results to some extent: adaptability for congestion, "hidden-behind" factors, and alteration in urban layout.

Generally, it's recognized that traffic congestion could result in shrinkage of either product market or labor market and indirectly reduce the scale efficiency of business activities. However, the real world is dynamic , especially in a long run. Increased delay and unreliable trips push both individuals and firms to adjust their behaviors once additional costs caused by congestion couldn't be endured. Employees might relocate themselves to areas adjacent to offices on the cost of greater payment for rent or mortgage, which are usually compensated by saved time and fuel as well as some other indirect benefits. (Levinson and Kumar, 1994) As a result, firms may not lose talents because of congestion. In addition, highly developed areas always attract various talents from everywhere and companies may still hire enough qualified employees[115].

For firms, their product/service market size is actually determined by the potential number of consumers. Hence, traffic congestion's impacts ascribe to whether it really impedes firms' potential consumers. One phenomenon deserving attention is that urban centers or central business districts are usually not the most congested areas, since normally many transportation alternatives are available there. People may walk, ride bicycles, take subways or buses rather than drive by themselves in these areas. It is arterial roads or primary highways connecting above areas with surrounding residential areas that encounter the severest congestion during rush hours. As a result, businesses

---

[115] Unemployment rate is at least 2.1% by all education attainments in 2012 based on Bureaus of Labor Statistics data. Though it's an average number that does not consider uneven distribution in various areas, it still illustrates that on average the labor market is at saturated situation.

located in urban centers still have high accessibility for their customers. In addition, if reasonable schedules are made, it's very possible to avoid rush hours and not necessary to suffer severe travel delay and additional congestion costs for firms.

Meanwhile, various industries have difference reliance on transportation. For example, manufacturing's market sizes may not shrink due to traffic congestion, as long as its products could be delivered to buyers prior to contracts' deadline. Some other industries, especially certain services, may not be influenced by congestion either, unless they have to provide emergent services during rush hours. Appointments could be made on weekends when traffic congestion is not very severe. Moreover, with rapid development of telecommunication and internet, some industries could expand their markets as widely as they can, including financing and communication. For these industries, traffic congestion's influence in their daily operation seems trivial.

Indeed, traffic congestion is not an element that could be purposely created (different from private capital investment or infrastructure expenditure), but a phenomenon actually caused by various factors. Those 'hidden behind' factors may affect TFP growth indirectly, some positively (e.g. excessive traffic demand reflected in more vehicle-miles travelled) and some negatively (e.g. deficient traffic supply, un-reasonable urban layout and poor traffic management system). Their different influences in TFP growth may counteract each other, and finally congestion's impact on TFP growth is just the net influence of above factors. Even though the congestion level in a region keeps the same, the efficiency of transportation system may vary greatly. With advanced technologies, reasonable working schedules and other proper measures, current

transportation system has accommodated much more traffic demand than before. According to the Bureau of Transportation Statistics, the average Travel Time Index in American urban Areas increased by 3.5% from 1990 to 2009, while the annual vehicle mileage travelled in all urban areas grew by 54.7%. If we just consider the period between 1990 and 2007 after considering the economic recession which occurred in 2007, the average TTI in urban areas increased by 7%, still much lower than the growth of VMT which is 56.7%. If we focus on one specific case, the New-York urban area, the increment of VMT between 1990 and 2009 is 31.7% (35.9% between 1990 and 2007), much larger than its TTI growth of 7.3% (14.7%) during the same period. Therefore, from the viewpoint of transportation engineering, the transportation system's performance has been improved greatly, even though many people are still unsatisfied on contemporary traffic congestion.

Conventional concept that traffic congestion is always bad may root in poor pavement, outdated traffic management system and deficient urban planning. However, rapid technology development and better urban planning have made it possible to fully utilize the capacity of current transportation system, though it still desires improvement. Correspondingly, the congestion might not be as bad as supposed to be under this condition. Garrett (2004) ever argues not only some level of urban traffic congestion inevitable, but also desirable: "A roadway with miles of bumper-to-bumper traffic is clearly an overused resource, but a roadway with no congestion at all is an underused resource. Thus there exists some optimal level of congestion." Morre (2009) also comments that "[S]ome amount of highway congestion is an unavoidable and desirable

byproduct of good planning and efficient investment…. Heavy use means that people value the places and activities that highways take them to and, by extension, that they value the highways." In consequence, the degree of traffic congestion might not be a good criteria to measure transportation system's efficiency any more. To some extent, this kind of congestion could be named 'quasi-optimal congestion', since the optimal level could yet be achieved without the implementation of road pricing policy[116]. In summary, we should focus on optimizing the congestion based on benefit and costs of dealing with this phenomenon rather than eliminating it arbitrarily.

Traffic congestion's impacts on business behaviors are also demonstrated in FHWA report, *Traffic Congestion and Reliability (2005)*, one of which is that "businesses tend to stagnate or move out of areas with high operating costs and limited markets, while they locate and expand in areas with lower operating costs and broader market connections."[117] Current solutions for deteriorate traffic congestion also include establishment of satellite cities outside urban center. The 1991-2010 Beijing Municipality Master Plan proposed 14 satellite cities to be developed as independent economic and cultural centers. On July 2nd, 2012, the South Korea formally opened the Sejong Special Autonomous City, a special administrative district located 75 miles away from Seoul, to redistribute ministries and national agencies in order to relieve capital's traffic pressure. Moscow also has a similar plan to move part of its administrative institutions away from

---

[116] It's debatable to regard highway and streets as pure public good. In theory, externalities of traffic congestion make it necessary to apply pricing scheme as a regulative mechanism. Without introducing road pricing strategy, congestion could not reach its optimal level

[117] The report also mentions that "[t]he magnitude of these changes varies by industries, based on how strongly the industry's total operating costs is affected by transportation factors."

the urban center to re-adjust distribution of resources and alleviate over crowded

pressure. No matter spontaneous move of businesses or administrative planning in urban

layout, peripheral areas of urban centers may gather more resources in this process. Since

these areas are normally less developed previously, their growth rate may be much higher

than existent urban centers. Hence, when the whole urban area is considered as an entity,

redistribution of businesses and following relocation of resources, especially to peripheral

areas, may enhance the average productivity growth in the whole area.

   All above factors take time to be functional, and that may explain why traffic

congestion has negative impact before 2000, but behaves positively in the last decade.

4.5 DECOMPOSITION OF TOTAL FACTOR PRODUCTIVITY

Controversial effects of traffic congestion on TFP growth enhance my interest to

investigate further on congestion's influences in components of TFP growth.

In Solow's (1957) study, technical change is regarded as the unique source of TFP

growth. However, it is a theoretical conclusion under rigid constraints of neutral technical

progress, perfect market and efficient production. In reality, constant returns to scale in

production function are not guaranteed. The phenomena of scale economies prove the

existence of variable returns to scale. Besides, not all firms could produce efficiently due

to weak management ability, market imperfectness and so on. As a result, technical

change can't equalize the TFP growth under most conditions. Several other factors may

also contribute to TFP growth, such as scale and efficiency elements.

Time delay and traffic unreliability caused by congestion (especially the latter)

may influence a firm's optimal usage of resources, i.e. labor and capital inputs, since it

has to assume additional costs, either expected or unexpected. Hence, congestion may affect the productive efficiency which covers various fields, such as production scale, technique efficiency and allocation of resources. It may not definitely encumber technical progress, at least in a short term. To investigate this assumption empirically, a stochastic frontier approach (SFA) is applied to decompose the TFP growth following Kumbahkar's (2000) method.

SFA is an evolvement of traditional production function models which obey the rule that the production procedure has been efficient enough and just analyzes the relation between inputs and output while overlooking the inefficiency issue (Mishra, 2010). Strictly speaking, the formulation of production function assumes that the engineering and managerial problems of technical efficiency have already been addressed and solved, so that the analysis can focus on the problems of allocative efficiency[118] (Mishra, 2010). Thus, a production function should be defined as a relationship between the maximal technically feasible output and inputs needed to produce that output (Shephard, 1970). However, in reality, not all firms could produce efficiently in the same way, owing to various factors, including outdated technology and deficient management. The stochastic

---

[118] Allocative efficiency implies that producers only produce those types of goods and services that are most desirable in the society and also in high demand. It's allocatively efficient where the marginal benefit is equal to the marginal cost, and at this point the social surplus is maximized with no deadweight loss. Free market under perfect competition generally are allocatively efficient, yet are not for the cases of monopoly, externalities, and public goods (for market failure), or price controls (for government failure in addition to taxation). In theory, allocative inefficiency results in utilization of inputs in the wrong proportions, given input prices (Schmidt and Lovell, 1979). However, accumulated empirical evidences have suggested that the problem of allocative efficiency is trivial, and improvement in 'nonallocative efficiency' is an important aspect of the process of growth (Leibenstein, 1966). This type of 'nonallocative efficiency' is called as X-effeiency by Leibenstein, and it is similar to the concept of technical efficiency, yet not the same (Leibenstein, 1977).

frontier model relaxes assumptions of perfect competition and optimal production which are set in neo-classical production function models, and it's more close to the real market.

The frontier analysis dated back to 1950s when Koopmans (1951) measured a firm's performance with respect to a best practice frontier. Farrell (1957) extended Koopmans's work and suggested measuring inefficiency as the observed deviation from a frontier isoquant. Following studies focus on the best way to define the frontier of the production possibility set, and two methods are frequently used: a) parametric methods, based on the econometric estimation of the frontier; and b) non-parametric methods based on linear programming techniques (e.g. data envelopment analysis DEA). In this chapter, the parametric approach -- a stochastic frontier analysis, is implemented based on the translog production function model. Compared with DEA, the frontier decided using SFA is stochastic with a probability distribution, which means the estimated inefficiency constituting two parts: a) the stochastic part; and b) the real production inefficiency equal to the gap between the performance of observed individual and the optimum frontier, while DEA couldn't differentiate these two parts.

The stochastic frontier model could be defined as:

$$Y_{it} = F(X_{it}, \ t) \exp(v_{it} - u_{it})$$

(17)

where $F(\cdot)$ is the production frontier[119] and $u_{it} \geq 0$ measures the technical inefficiency

and alters over time. v_{it} is the idiosyncratic and stochastic part of frontier and

$v_{it} \sim N[0, \sigma_v^2]$. Taking the natural logarithm of both sides of above equation, we can get:

$$\ln Y_{it} = \ln F(X_{it}, \ t) + v_{it} - u_{it}$$

(18)

Totally differentiating above equation with respect to time, we obtain the

following formula[120]:

$$\dot{Y} = \frac{d \ln F(X,t)}{dt} - \frac{du}{dt}$$

(19)

Totally differentiating the natural logarithm of $F(\cdot)$ with respect to time t:[121]

$$\frac{d \ln F(X,t)}{dt} = \frac{\partial \ln F(X,t)}{\partial T} + \frac{\partial \ln F(X,t)}{\partial X} \cdot \frac{dX}{dt} = TC + \sum_j \varepsilon_j \dot{X}_j$$

(20)

The first term on the right-hand side of equation (20) measures the change in the

frontier output caused by technical change TC, and the second one measures the output's

increment or reduction caused by the change in input use. Combining equation (19) and

(20), the equation (20) can be rewritten as:

$$\dot{Y} = TC + \sum_j \varepsilon_j \dot{X}_j - \frac{du}{dt}$$

(21)

---

[119] A production frontier represents a point at which an economy could most efficiently produces goods or services, therefore, allocating its resources in the best way.
[120] v_{it} is white noise and its differentiation with respect to t is zero.
[121] Subscripts i and t are omitted for simplicity.

Hence, the overall productivity change is not only affected by technical change and changes in input use, but also influenced by the change in technical inefficiency (TE). For a given level of inputs, if exogenous technical change shifts the production frontier upward (downward), then TC is positive (negative). TE improves (deteriorates) over time, if du/dt is negative (positive). $-du/dt$ can be interpreted as the rate at which an inefficient producer catches up to the production frontier (Kim and Han, 2001, p 271). Using the Divisia index of TFP growth and equation (21), we can get

$$\dot{\text{TFP}} = \dot{Y} - \sum_j S_j \dot{X}_j = TC + \sum_j \varepsilon_j \dot{X}_j - \frac{du}{dt} - \sum_j S_j \dot{X}_j$$

$$= TC - \frac{du}{dt} + \left(\sum_j \varepsilon_j - \sum_j S_j\right) \dot{X}_j$$

$$= TC - \frac{du}{dt} + (RTS - 1)\sum_j \lambda_j \dot{X}_j + \sum_j (\lambda_j - S_j) \dot{X}_j \qquad (22)$$

where $RTS = \sum_j \varepsilon_j$ denotes the measure of returns to scale, and $\lambda_j = \varepsilon_j / RTS$. The last component on the right-hand side in Equation 22 measures inefficiency in resource allocation resulting from deviations of input prices from the value of their marginal product.

Based on above equation, TFP growth could be decomposed into technical change (TC), technical efficiency (TE) change, scale components and the allocative efficiency (AE) change. The technical change measures the increment or reduction of the maximum output that can be produced from a given level of inputs resulted from technical advance or regress by a shift in production frontier. Technical efficiency change is the change in a firm's ability to achieve maximum output given its set of inputs reflected by how close it is to the production frontier. Scale efficiency change represents

the change in the degree to which a firm is optimizing the scale of its operation

demonstrated by a firm's movement along the production function curvature. (Sena,

2003) Allocative efficiency change reflects the change in a firm's ability to select a level

of inputs so as to ensure that the input price ratios equal to the ratios of the corresponding

marginal products. The benchmark, production frontier, is determined by the production

function model. In this research, because of the lack of price data, the AE change is

omitted, and only three components are considered. To avoid debates on abundant

assumptions, the translog production function is still used.

Taking the partial derivative of both sides of Equation 21 with respect to T, we

can obtain the technical change/progress:

$$TC = \varepsilon_T = \frac{\partial \ln Q_{i,t}}{\partial T} = \lambda + \gamma_{LT} \ln L_{i,t} + \gamma_{KT} \ln K_{i,t} + \gamma_{GT} \ln G_{i,t} + \gamma_{TT} T \qquad (23)$$

where ε_T is defined as the growth of output with respect to time, holding capital input,

labor input, and highway/streets input constant.

4.5.1 Regression Results
The following table displays the result of regression on technical change with the

same group of determinants.

Table 4-15   Technical Change as Dependent Variables

TC	Phase I		Phase II
	1990-1997	1990-2000	2001-2009
Traffic Congestion	.0003	.00004	-.00002
Human Capital	.0018 **	.0033**	.00004
R&D	.0001 **	.00003**	2.26e-06**
Communication	.0047**	.0039**	.0005**
Finance	.0033**	.0017**	-.0001
Density	-2.73e-07**	-3.52e-07**	-1.85e-08**
Sectorial Agglomeration	.0039**		.0001**
National GDP	-.0019**	-.0013**	-.0615**

Initial Per Capita Income	.0291**	.0311**	.0013**
F	470093**	57504**	6.51e+09**

*: significant at 10% level;
**: significant at 5% or lower level.

In this case, traffic congestion has no significant impact on technical change in all models over two periods, which complies with previous discussion that congestion may not influence the progress of techniques. In the information era, physical trips may not play an important role in knowledge spillover and innovation any more. Other factors may be more decisive in this procedure. All signs of significant variables keep consistent except the human capital, so the results are robust in the main. R&D, communication, agglomeration, and initial per capita income contribute to technical change in both periods, while human capital and financial service were only significant in Phase I. The result basically accords with previous studies. Density is negatively significant, reflecting that urban areas have been over-crowded and result in negative impacts. The significantly negative coefficient of national GDP illustrates that the macro economic situation in America, including economic structure and policy, may also impede the technical progress, arousing the concern on the macro policy.

Table 4-16 Scale Efficiency Change as Dependent Variables

SC	Phase I		Phase II
	1990-1997	1990-2000	2001-2009
Traffic Congestion	.0153*	.0135**	.0130**
Human Capital	-.0072	-.0032	-.0223**
R&D	.0001	.00004	-.00005
Communication	.0364**	.0351**	.1227**
Finance	.0245	.0118	-.0164**
Density	-1.14e-06	-8.52e-07**	1.25e-06
Sectorial Agglomeration	.0075		-.0059
National GDP	.0028	.0006	.0087**
Initial Per capita Income	.0096	.0242	.1227

F	63.7**	114.4**	2778.4**

*: significant at 10% level;
**: significant at 5% or lower level.

For scale efficiency change, communication seems a constant contributor in all models during the whole period, demonstrating that it effectively reduces costs in production, operation and management of firms, and prompts them to approach correspondingly optimal scale levels. Traffic congestion has no significant impact during 1990 and 1997, while it has significant and positively influence after 2000. On one side, firms' product market size may not really shrink owing to congestion, so the distance between their current scale level and optimal scale level may not be affected by congestion, either. On the other side, the change of urban layout that leads to relocation of resources may optimize economic scales. Previously over-crowded resources may have damaged the production efficiency, while redistribution of businesses may lead to better allocation of resources, which is beneficial for economic growth. Moreover, impacts of congestion's 'hidden factors' may still also exert positive impacts after inter-counteraction. More economic activities reflected by higher VMT would improve regional scale economy, even though some other factors may negative influence scale efficiency.

Table 4-17 Technique Efficiency Change as Dependent Variables

TE growth	Phase I		Phase II
	1990-1997	1990-2000	2001-2009
Traffic Congestion	-.0085**	-.0024	.0002
Human Capital	.0055**	.0046**	.0081**
R&D	-.00009**	-8.17e-06	.00003**
Communication	-.0583**	-.0258**	.0008
Finance	-.0169**	-.0068**	.0006

100

Density	8.10e-07**	7.09e-07**	5.38e-08
Sectorial Agglomeration	.0159**		-.0089*
National GDP	.0003*	-.0012**	.0027**
Initial Per capita Income	.0115	.0089**	-.0085**
F	3859.7**	5.96e+11**	3917.2**

*: significant at 10% level;
**: significant at 5% or lower level.

When using technical efficiency change as the dependent variable, coefficients vary markedly. Except the human capital, which plays both significant and positive roles in two periods, other variables have inconsistent coefficients either in significances or in signs. Traffic congestion seems not a determinant factor after 2000, but it harmed technical efficiency growth between 1990 and 1997 while considering the sectorial agglomeration simultaneously. The adaptation of both individuals and firms to congestion may explain why its impacts on technical efficiency change became insignificant in the past decade. Development of technologies, especially in instant communication, also makes it possible that commuters could make business calls and check business emails even seated in their stop-and-go vehicles. As a result, traffic delay may not definitely result in efficiency reduction. For firms, rescheduling shipment delivery and corresponding adaptation in operation and management may also alleviate congestion's negative impacts, which makes the statistical results insignificant possibly.

4.6 SUMMARY

Through rigid quantitative analysis, traffic congestion's impacts on TFP growth as well as its components are investigated thoroughly in this study. Actually, traffic congestion's impacts are much more complicated than what's expected based on traditional theories. The adaptation to congestion during a long-term as well as the

development of technologies counteracts its theoretical negative impacts on labor and product market for firms. Moreover, relocation of firms and emergence of new business districts (e.g. edge cities), directly or indirectly caused by traffic congestion, produce more growth opportunities. In these emerging areas, the TFP growth is normally higher than those mature urban centers. As a result, the average TFP growth in the whole urban area may also increase.

Hidden-behind factors that influence the congestion level indirectly affect TFP growth as well. Rapid improvement in transportation technologies, traffic management and urban planning makes it possible to utilize transportation resources effectively, and gradually leads to a 'quasi-optimal' congestion level, at which marginal congestion cost is approximate to marginal benefit of additional traffic in the lack of consideration of congestion's externalities. That could explain why traffic congestion results in distinct impacts in the past decade compared with its counterparts during the last ten years in the 20th century. Besides, traffic congestion has no significant impacts on technical change, since it doesn't really slow down the transfer of knowledge and skills owing to the development of communication at least in a relatively short term. Its impacts on technical efficiency change also faded with the time passing, and its equivalents on scale efficiency change are positive and significant. That might be explained by alteration in firms and labor's locations caused by congestion as well as the trade-off among congestion's hidden-behind factors. In conclusion, traffic congestion may not really damage local economic efficiency in U.S. large urban areas.

CHAPTER 5   IMPACTS OF TRAFFIC CONGESTION ON REGIONAL ECONOMIC EFFICIENCY: NON-PARAMETRIC ANALYSIS

In the previous chapter, the productivity change denoted by TFP growth is measured as the variation over time of the firm's distance from the frontier and is decomposed into changes in technology, technical efficiency, and scale efficiency, employing a stochastic frontier analysis (SFA) based on the translog production function model. However, the parametric approach requires a function form for the production technology and a distributional assumption for the inefficiency component that has often been criticized on the ground. No matter using simple Cobb-Douglas production function, relatively complex trans-log production function, or more generalized CES (Constant Elasticity of Substitution) (Arrow et al. 1961) and VES (Variable Elasticity of Substitution) (Revankar, 1971) production functions, each model has its own specific assumptions. Violation of these assumptions directly results in poor performances. In addition, each production function corresponds to a specific form in which inputs are combined to produce outputs[122]. Thus, the selection of production function model leads to a series of implications with respect to the shape of the implied isoquants and values of elasticities of factor demand and factor substitution (Greene, 2008). As a result, estimate results may vary greatly using different production functions.

---

[122] The simplest function form should be the linear production function: $Q = \alpha_1 X_1 + \cdots + \alpha_n X_n$.
A classic exponential function form is the Cobb-Douglas production function: $Q = AK^\alpha L^\beta$.
Another special function form is the Neumann-Leontieff production function: $Q = \min(\alpha_1 X_1, \dots, \alpha_n X_n)$.

The distribution of the inefficient term μ is another major focus. In theory, four types of distribution are introduced: half-normal, exponential, truncated normal, and gamma. Albeit, in practice, it's impossible to determine which distribution type the μ obeys, and people usually subjectively select distribution types. Plenty of empirical evidences prove that sample mean efficiencies are sensitive to the chosen distribution (Badunenko and Stephan, 2007). Though Greene (1990) comments that estimated sample mean efficiencies are similar no matter which distribution type is assumed, and the rank correlation coefficients between the pairs of efficiency estimates for all sample observations are high enough to suggest using relatively simple distribution (e.g. half-normal or exponential), it's still arbitrary to conclude that the distribution of μ will not influence the estimated result at all.

In conclusion, although SFA has a well-established structure and could explicitly show statistical significances between inputs and output, the complexity and limitations in application still provide the space for the emergence and development of non-parametric analysis. Waived requirements on modeling structure as well as relaxed assumptions prompt its empirical applications.

5.1 DATA ENVELOPMENT ANALYSIS

The concept of data envelopment analysis (DEA) was built on ideas of Farrell (1957). Farrell suggests that one could usefully analyze technical efficiency in terms of realized deviations from an idealized frontier isoquant. (Greene, 2008) In parametric analysis, this concept naturally falls into an econometric approach in which the inefficiency is identified with disturbances included in a regression model, while in non-

parametric analysis, the *DEA* proposed by Charnes, et al. (1978) is applied to scale the

inefficiency level. The kernel of *DEA* lies in using the linear programming to wrap a

quasi-convex hull around the data, which is in essence the Farrell's efficient unit

isoquant. Technically, this method searches for points with the lowest inputs for any

given output or those with the highest outputs for any given inputs, and connects these

targeted points to form the efficiency frontier.
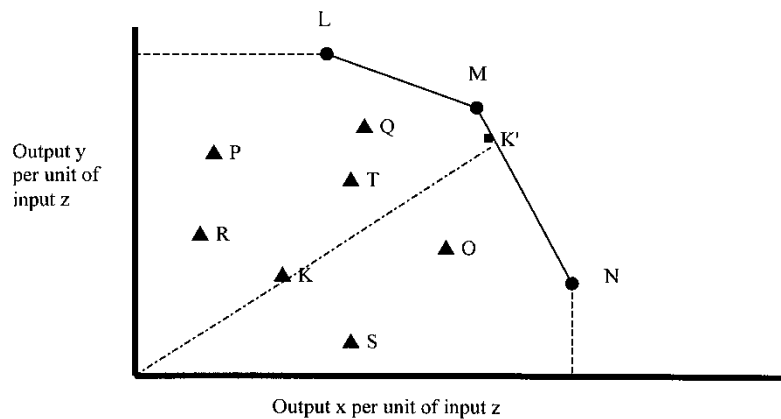
Figure 5-1 Example of DEA
Source: http://www.emeraldinsight.com/content_images/fig/0320170501001.png

Obviously, points on the frontier compose the efficient subset (L, M and N), and

those not on the frontier (e.g. P, Q and T) are deemed as inefficient producers. A

numerical coefficient is given to each producer, defining its relative efficiency. *DEA*

differs fundamentally from the econometric approach in its interpretation of the frontier

generating mechanism, but is analogous in its philosophical underpinnings. In other

words, DEA is just one method that could be summarized into the catalogue of

deterministic frontier analysis, which is unfortunately an disadvantage compared with SFA.

In sum, DEA is a mathematical programming approach to assess the relative efficiency[123] for a group of homogenous decision making units (DMUs) with multiple inputs and outputs. An efficient DMU, the benchmark, is the observed one that lies on the 'frontier' of the production possibility set. The efficiency score of a DMU in the presence of multiple input and output factors is defined as $Efficiency = \frac{weighted\ sum\ of\ outputs}{weighted\ sum\ of\ inputs}$.

Assuming that there are n DMUs, each with m inputs and s outputs, the relative efficiency score of a DMU p is obtained by the following model (Charnes et al. 1978):

$$\max \frac{\sum_{k=1}^{s} v_k y_{k,p}}{\sum_{j=1}^{m} u_j x_{j,p}}$$
$$s.t \ \frac{\sum_{k=1}^{s} v_k y_{k,i}}{\sum_{j=1}^{m} u_j x_{j,i}} \leq 1 \quad \forall i \tag{24}$$
$$v_k, u_j \geq 0 \quad \forall\ k, j$$

where $k = 1, 2, \ldots, m$, $j = 1, 2, \ldots, s$, and $i = 1, 2, \ldots, n$. y_{ki} denotes the amount of output k produced by DMU i, and x_{ji} is the amount of input j utilized by DMU i. v_k and u_j are weights given to output k and input j, separately.

Since the nonlinear programming formula is an extended form of an ordinary fractional programming problem, it could be converted to a linear programming (LP) equivalents, which will provide the optimal v_k combination for each DMU, complying

---

[123] The efficiency here refers to technical efficiency that might be characterized as either 1) the feasible increase in outputs for a given set of inputs or 2) the feasible reduction of inputs for a given set of outputs when there is no waste.

with subjective conditions.[124] According to the dual theory of linear programming

problem, its dual linear programming can be described as follows:

$$\min_{\theta\lambda} \theta$$
$$s.t \ \sum_{i=1}^{n} \lambda_i x_{ji} - \theta x_{jp} \leq 0 \quad \forall j$$
$$\sum_{i=1}^{n} \lambda_i y_{ki} - y_{kp} \geq 0 \quad \forall k \tag{25}$$
$$\lambda_i \geq 0 \quad \forall i$$

where θ represents the efficiency score, and λ is a vector of dual variables. A DMU is

considered to be efficient if θ is equal to 1[125], and a score of less than 1 implies its

inefficiency[126]. The lower the score, the less efficient the DMU. For every inefficient

DMU, DEA identifies a set of corresponding efficient units that can be used as

benchmarks which could be achieved based on above equations. If a composited DMU, a

linear combination of efficient units, can be identified, this virtual DMU will utilize less

input than that of the corresponding real DMU to produce the same amount of outputs.

The model demonstrate a classical form of DEA model, the input-oriented CCR

model[127] which incorporates the assumption of constant return to scale in production[128].

Färe et al. (1983), Banker et al. (1984), and Byrnes et al. (1984) extend Charnes'

---

[124] The linear programming has the following form:

$$\max_{u,v} \sum_{k=1}^{s} v_k y_{kp}$$
$$s.t \ \sum_{j=1}^{m} u_j x_{jp} = 1$$
$$\sum_{k=1}^{s} v_k y_{ki} - \sum_{j=1}^{m} u_j x_{ji} \leq 0 \quad \forall i$$
$$v_k, u_j \geq 0 \quad \forall k, j$$

[125] Strictly, only $\theta = 1$, does it represent that DMU is weak efficient. If and only if $\theta = 1$, $\sum_{i=1}^{n} \lambda_i x_{ji} = \theta x_{jp}$, and $\sum_{i=1}^{n} \lambda_i y_{ki} = y_{kp}$, could a DMU be considered efficient.

[126] The objective seeks the minimum θ that reduces the input vector x_{jp} to θx_{jp}, while remaining the output at frontier line (efficient level).

[127] Another model is the output-oriented CCR model that focuses on how to maximize outputs given inputs.

[128] In effect, Banker et al. (2004) comments that the CCR model simultaneously evaluates the technical and returns-to-scale performances, i.e. it evaluates scale and purely technical inefficiencies at the same time. Thus, its result is the same as the technical efficiency estimates under the condition of constant returns to scale (without considering scale effect).

approach by relaxing this limitation and allowing variable returns to scale in production. They design the BCC DEA model through including an additional convexity constraint ($\sum_{i=1}^{n} \lambda_i = 1$) in the CCR model. This convexity constraint takes into account the possibility that the average efficiency level at the most efficient scale size may not be attainable for other scale sizes at which a particular DMU may operate (Banker et al., 1984). Hence, the efficiency score conducted in BCC DEA model measures the pure technical efficiency of a DMU at the given scale of operation. Since the CCR DEA model measures the overall technical efficiency, the ratio between two measures of efficiency in CCR and BCC DEA models is a measure of scale efficiency.[129] In addition to CCR and BBC DEA models, many additive forms have also been proposed, such as performance-based clustering methods for identifying appropriate benchmarks[130] (Doyle & Green, 1994) and super-efficiency DEA model[131] (Andersen and Petersen, 1993). In this research, the calculation of Mamlquist productivity index in the following step only employs conventional CCR and BBC DEA models.

In conclusion, DEA is a methodology directed to frontiers rather than central tendencies. Instead of trying to fit a regression plane through the center of data as in the statistical regression, DEA 'floats' a piecewise linear surface to rest on top of the observations. (Cooper et al. 2011) Hence, DEA doesn't require explicitly formulated assumptions and variations with various types of production function. Furthermore, it has

---

[129] Dalmau-Matarrodona and Puig-Junory, 1998, pp. 459.

[130] One difficulty in identifying benchmark while using DEA is that the composite DMU that dominates the inefficient DMU may not exist in reality (Talluri, 2000). .

[131] In the CCR model, it's impossible to rank or differentiate efficient DMUs. The core idea of super-efficiency DEA is to exclude the DMU under evaluation from the reference set, and it can take any value greater than or equal to 1.

a significant advantage of handling multiple inputs and outputs simultaneously. Even though, DEA also has several disadvantages in following aspects: 1). The efficiency score could be very sensitive to extreme points, especially when data may be contaminated by measurement error (Timmer, 1971); 2). It's impossible to detect the relationship between inputs and outputs, including both magnitudes and significances; and 3). The number of efficient DMUs on the frontier tends to increase with the number of inputs and outputs variables (Berg 2010). In our analysis, raw data are from authorized organizations, including Census and BEA, and have been reasonably processed to mitigate measurement error as much as possible. Additionally, only three inputs and one output will be handled, and parametric features have been discussed in the previous chapter. Thus, *DEA*'s deficiencies may not result in significant problems in our analysis.

5.2 MALMQUIST PRODUCTIVITY INDEX

As a counterpart of the TFP growth in the parametric analysis, the Malmquist productivity index is prevalent in the non-parametric analysis to measure the production efficiency, and could be calculated using the *DEA* methodology.

As early as in 1953, Malmquist introduced the input distance function and developed a standardized consumption quantity index as the ratio of a pair of input distance functions[132]. Then, the Malmquist's consumption index was converted into an input quantity index in the context of production analysis. In succession, an analogous output quantity index was introduced by Shephard (1970). However, Malmquist's

---

[132] The Malmquist index measures the amount by which one consumption bundle need to be radially scaled in order to generate the same utility level provided by several base consumption bundle. (Grifell-Tatje and Lovell, 1995, pp. 169).

quantity index wasn't applied to measure the productivity change until Caves et al. (1982) contributed their influential work to establish the relationship between the Malmquist index and the Tornqvist (1936) index. Given the assumption of translog production technologies with equal second-order parameters, the geometric mean of two Malmquist input (output) quantity indices is proved to be equal to a Tornqvist input (output) index. In addition, the geometric mean of two Malmquist input-based (output-based) productivity indices is equal to the Tornqvist productivity index, corrected by a scale factor in the case of variable returns to scale.

There are two definitions of Malmquist productivity index. The first one is partially oriented, being based either on ratios of output distance functions or on ratios of input distance function. Caves et al. (1982) argues that output based productivity indexes treat productivity differences as differences in maximum output conditional on a given level of inputs, and input based productivity indexes treat productivity differences as differences in minimum input requirements conditional on a given level of outputs. In both cases, data from one period are compared with counterparts in the adjacent period, in either an output-enhancing or an input-saving direction. An alternative definition is proposed by Diewert (1992) based on the ratio of a Malmquist output quantity index divided by a Malmquist input quantity index. Accordingly, this index includes both output and input distance functions. Bjurek (1996) further popularizes this idea as an empirical index, often referred as the Malmquist total factor productivity index. Rather than defining the productivity as input oriented or output oriented, the Malmquist total factor productivity is simultaneously oriented, and the new definition also solves an

110

inherent problem in traditional definition that the index may not be bounded for all units

under variable returns to scale non-parametric technologies. Färe et al. (1996) prove that

these two productivity indexes are equal if, and only if, technologies exhibit inverse

homogeneity. However, though these restrictions are unlikely to hold empirically, the two

indices generate little different measures of productivity change (Färe et al., 1996; and

Balk, 1993).

5.2.1 The Malmquist Output Based Productivity, and Total Factor Productivity[133]

Based on the theoretical contribution of Farrell's (1957) and the DEA method

developed by Caves et al. (1982), input and output based efficiency can be measured for

comparisons over time. Assume a production unit producing a vector of outputs, $y = (y_1, y_2, ..., y_m)$, with a vector of inputs, $x = (x_1, x_2, ..., x_n)$, observed at different time

periods $t = 1, 2, ..., T$. The production technology F models the transformation of inputs

x into outputs y: $F = \{(x, y) : x \ can \ produce \ y\}$. The output efficiency is defined as the

proportional change in all observed outputs quantities given inputs quantities, compared

to the outputs quantities produced based on the same inputs quantities using the frontier

technology. Utilizing the output distance function, we can define the output based

efficiency as $D_O^t(y^t, x^t) \equiv \max_\delta \{\delta : (\delta y^t, x^t) \in F^t\}$, where O denotes the output, and D

measures the relative distance between a production unit and its corresponding

benchmark unit or unit combination with frontier technology. In particular, note that

---

[133] The only differences between input-oriented and output-oriented productivity are their directions and objectives (maximum or minimum). They share very similar forms in equations. For simplicity, only the output-oriented productivity index is discussed here.

$D_0^t(y^t, x^t) = 1$ if and only if the production unit is on the boundary of frontier, when production is technically efficient. $D_0^t(y^t, x^t)$ is the within-period distance, defined using period t inputs and outputs as well as period t production technology. Adjacent-period distance functions that use data from one period and technology from an adjacent period, are defined as $D_0^t(y^{t+1}, x^{t+1}) \equiv \min_\delta \left\{ \delta : (\frac{1}{\delta} y^{t+1}, x^{t+1}) \in F^t \right\}$ and $D_0^{t+1}(y^t, x^t) \equiv \min_\delta \left\{ \delta : (\frac{1}{\delta} y^t, x^t) \in F^{t+1} \right\}$.

$D_0^t(y^{t+1}, x^{t+1})$ measures the maximal outputs produced to make (y^{t+1}, x^{t+1}) feasible in relation to the technology at period t. Similarly, $D_0^{t+1}(y^t, x^t)$ measures the maximal outputs obtained to make (y^t, x^t) feasible in relation to the technology at period $t + 1$. As a result, the period t output-based CCD Malmquist productivity index is defined as:

$$M_0^t(x^t, y^t, x^{t+1}, y^{t+1}) = D_0^t(y^{t+1}, x^{t+1})/D_0^t(y^t, x^t) \qquad (26)$$

This index measures the productivity change between periods t and $t + 1$, from the perspective of period t technology. Likewise, the period $t + 1$ output-based CCD Malmquist productivity index with the technology at period $t + 1$ as the reference has the following form:

$$M_0^{t+1}(x^t, y^t, x^{t+1}, y^{t+1}) = D_0^{t+1}(y^{t+1}, x^{t+1})/D_0^{t+1}(y^t, x^t) \qquad (27)$$

Either $M_0^t(x^t, y^t, x^{t+1}, y^{t+1})$ or $M_0^{t+1}(x^t, y^t, x^{t+1}, y^{t+1})$ could be larger, equal or smaller than 1, as productivity may grow, stagnate or decline between periods t and $t + 1$. Because the two indices are not necessarily equal, in order to avoid the selection of an

112

arbitrary benchmark, normally people define the output-based CCD Malmquaist productivity change index as the geometric mean of the two, as follows:

$$M_O(x^t, y^t, x^{t+1}, y^{t+1}) = \left[\left(\frac{D_O^t(y^{t+1}, x^{t+1})}{D_O^t(y^t, x^t)} \right) \left(\frac{D_O^{t+1}(y^{t+1}, x^{t+1})}{D_O^{t+1}(y^t, x^t)} \right) \right]^{1/2} \tag{28}$$

$M_O(x^t, y^t, x^{t+1}, y^{t+1})$ is defined relative to a best practice technology satisfying variable returns to scale, and it could measure the productivity growth, stagnation or decline, net of the effect of scale economies. Grifell-Tatje and Lovell (1995) criticize its inaccuracy in the presence of non-constant returns to scale technology. Indeed, the calculated values of $M_O(x^t, y^t, x^{t+1}, y^{t+1})$ are biased downward in the region of increasing returns to scale, and biased upward in the region of decreasing returns to scale. Färe et al. (1994) (FGNZ) argue that a Malmquist productivity index must be based on distance functions defined on a bench technology characterized by constant returns to scale in order to provide an accurate measure of productivity change.[134] Such a Malmquist productivity index can be expressed as:

$$M_{OC}^t(x^t, y^t, x^{t+1}, y^{t+1}) = D_{OC}^t(y^{t+1}, x^{t+1})/D_{OC}^t(y^t, x^t) \tag{29}$$

where subscript C represent the constant returns to scale. Others without C correspond to the assumption of variable returns to scale.

In addition to Fare and his colleagues' contribution, Grifell-Tatje and Lovell (1999) also developed an output-based period t generalized Malmquist productivity

---

[134] In addition, the constant returns to scale technology is preferred so as to avoid the feasibility problem that may arise while estimating mixed-period distance functions given variable returns to scale (Mahlberg and Sahoo, 2011).

index which is expressed as the product of a period t CCD Malmquist productivity index and a period t Malmquist scale index with the following form: [135]

$$GM_O^t(x^t, y^t, x^{t+1}, y^{t+1}) = M_O^t(x^t, y^t, x^{t+1}, y^{t+1}) \cdot E_O^t(y^t, x^t, x^{t+1}) \tag{30}$$

where E_O^t denotes a period t output-based Malmquist scale index, defined as:

$$E_O^t(y^t, x^t, x^{t+1}) = S_O^t(x^{t+1}, y^t)/S_O^t(x^t, y^t), \text{ with } S_O^t(x^t, y^t) = D_O^t(y^t, x^t)/D_{OC}^t(y^t, x^t)$$

and $S_O^t(x^{t+1}, y^t) = D_O^t(y^t, x^{t+1})/D_{OC}^t(y^t, x^{t+1})$.

Besides conventional CCD Malmquist productivity index, updated FGNZ index and the generalized version from Grifell-Tatje and Lovell, Bjurek (1996) introduces the definition of Malmquist total factor productivity following Diewet's contribution work based on the Fisher quantity indices(Fisher, 1992)[136]. According to Moorsteen (1961), the real output quantity index could be defined as $Q_O^s(y^t, y^{t+1}, x^s) =$ $D_O^s(y^{t+1}, x^s)/D_O^s(y^t, x^s)$, where s refers to either period t or $t+1$, and Q denotes quantity. It measures the change in observed output quantities between adjacent two periods t and $t+1$.[137] The output quantity index presents a ratio of an output efficiency measure for a combination of outputs and inputs in different periods to the corresponding

---

[135] Grifell-Tatje and Lovell (1999) also demonstrate that the generalized Malmquist productivity index does provide an accurate measure of productivity change in the presence of scale economies with one input and one output sample. Furthermore, under the market behavior and technology conditions maintained by Caves et al. (1982), the generalized Malmquist productivity index is proved to be equal to a Tornqvist productivity index, with any number of outputs and inputs.

[136] Diewert (1992) presents strong economic justifications for the use of the Fisher ideal input and output quantity indices (the geometric mean of the Laspeyres and Paasche quantity indices) that are analogous to the economic justifications for the use of translog input and output indices. He also obtains a strong economic justification for the use of the Fisher productivity index that is presented as the ration between the Fisher output quantity index and the Fisher input quantity index.

[137] When $s = t$, $D_O^t(y^{t+1}, x^t)$ measures the maximal proportional change in outputs required to make (y^{t+1}, x^t) feasible in relation to the technology at period t. When $s = t + 1$, $D_O^{t+1}(y^t, x^{t+1})$ measures the maximal proportional change in outputs required to make (y^t, x^{t+1}) feasible in relation to the technology at period $t + 1$. Either $D_O^t(y^t, x^t)$ or $D_O^{t+1}(y^{t+1}, x^{t+1})$ is a is a standard output efficiency measure for a unit observed at either period t or $t + 1$.

standard output efficiency measure for a combination of outputs and inputs in the same periods.[138] Similarly, the real input quantity index has a similar form, $Q_I^s(y^s, x^{t+1}, x^t) = D_I^s(y^s, x^{t+1})/D_O^s(y^s, x^t)$. [139]

As a result, the Malmquist total factor productivity is defined as the ratio between an output quantity index and an input quantity index, as follows:

$$MTFP^s = \frac{Q_O^s(y^t, y^{t+1}, x^s)}{Q_I^s(y^s, x^{t+1}, x^t)} = \frac{D_O^s(y^{t+1}, x^s)/D_O^s(y^t, x^s)}{D_I^s(y^s, x^{t+1})/D_O^s(y^s, x^t)} \tag{31}$$

Compared to output-oriented Malmquist productivity indices, Bjurek's Malmquist total factor productivity is not restricted in directions. It also relaxes the assumption on the scale effect. However, in spited of its theoretical advantages, in practice it has not been developed well enough to handle real cases owing to the complexity of its concept. Based on the study conducted by Grifell-Tatje and Lovell (1999), except the CCD Malmquist productivity index, the $M_{OC}(x^t, y^t, x^{t+1}, y^{t+1})$, $GM_O(x^t, y^t, x^{t+1}, y^{t+1})$ (the geometric mean) and $MTFP(y^t, y^{t+1}, x^t, x^{t+1})$ (the geometric mean) all could provide accurate estimates on productivity growth.

5.2.2 Decomposition of the Malmquist Productivity Indices

Similar to the TFP growth, the Malmquist productivity index could also be decomposed into several parts. The period t CCD Malmquist productivity index decomposes as:

$$M_O^t(x^t, y^t, x^{t+1}, y^{t+1}) = \frac{D_O^t(y^{t+1}, x^{t+1})}{D_O^{t+1}(y^{t+1}, x^{t+1})} \frac{D_O^{t+1}(y^{t+1}, x^{t+1})}{D_O^t(y^t, x^t)} \tag{32}$$

---

[138] If this index is greater than unity, more outputs are produced at period $t + 1$ than at period t, for given technology level and inputs at period s.

[139] If this index is less than unity, there are less inputs required in the production at period $t + 1$ than at period t, for given technology and output quantities at period s.

The first item on the right side is an index of technical change along a ray through period $t + 1$ data, and the second item is an index of technical efficiency change, both calculated relative to variable returns to scale reference technologies.

The period t FGNZ Malmquist productivity index includes effect of returns to scale, and could be decomposed as:

$$M_{OC}^t(x^t, y^t, x^{t+1}, y^{t+1}) = \frac{D_{OC}^t(y^{t+1}, x^{t+1})}{D_{OC}^t(y^t, x^t)} =$$

$$\frac{D_{OC}^t(y^{t+1}, x^{t+1})}{D_{OC}^{t+1}(y^{t+1}, x^{t+1})} \frac{D_O^{t+1}(y^{t+1}, x^{t+1})}{D_O^t(y^t, x^t)} \frac{D_{OC}^{t+1}(y^{t+1}, x^{t+1})/D_O^{t+1}(y^{t+1}, x^{t+1})}{D_{OC}^t(y^t, x^t)/D_O^t(y^t, x^t)} \tag{33}$$

Similar to the decomposition of CCD Malmquist index, FGNZ decomposition also has a technical efficiency change term and a technical change term. However, the technical change term is calculated relative to constant returns to scale reference technologies using period $t + 1$ data. In addition, it has a scale effect that captures the impact of change in scale efficiency from period t to $t + 1$.

The FGNZ Malmquist productivity index could be further expressed as:

$$M_{OC}(x^t, y^t, x^{t+1}, y^{t+1})$$

$$= \frac{D_O^{t+1}(y^{t+1}, x^{t+1})}{D_O^t(y^t, x^t)} \cdot \left[\frac{D_{OC}^t(y^{t+1}, x^{t+1})}{D_{OC}^{t+1}(y^{t+1}, x^{t+1})} \cdot \frac{D_{OC}^t(y^t, x^t)}{D_{OC}^{t+1}(y^t, x^t)}\right]^{1/2} \cdot \frac{D_{OC}^{t+1}(y^{t+1}, x^{t+1})/D_O^{t+1}(y^{t+1}, x^{t+1})}{D_{OC}^t(y^t, x^t)/D_O^t(y^t, x^t)}$$

$$= TE\Delta(x^t, y^t, x^{t+1}, y^{t+1}) \cdot T\Delta_C(x^t, y^t, x^{t+1}, y^{t+1}) \cdot SE\Delta(x^t, y^t, x^{t+1}, y^{t+1}) \tag{34}$$

where $TE\Delta(x^t, y^t, x^{t+1}, y^{t+1})$ measures the technical efficiency change on the best practice technologies, $T\Delta_C(x^t, y^t, x^{t+1}, y^{t+1})$ measures the geometric mean of the magnitudes of technical change along rays through (x^{t+1}, y^{t+1}) and (x^t, y^t), and $SE\Delta(x^t, y^t, x^{t+1}, y^{t+1})$ measures the scale efficiency change from period t to period $t +$

1. Though there are still some critics on FGNZ Malmquist index decomposition and some other forms have been developed and analyzed, such as the Generalized version, and Ray-Desli (1997) version (see Appendix for details), the FGNZ decomposition is the most common way in application.

In sum, the FGNZ Index could be decomposed into technical change, technical efficiency change, and scale efficiency change. In calculation, DEA is used with four basic distance functions, $D^t(y^t, x^t)$, $D^t(y^{t+1}, x^{t+1})$, $D^{t+1}(y^t, x^t)$ and $D^{t+1}(y^{t+1}, x^{t+1})$, with benchmark technologies satisfying either constant returns to scale (CCD model) or variable returns to scale (BBC model). In this paper, the DEAP software is used[140], which decomposes Malmquist Productivity Index based on the form of FGNZ.

5.3 RESULTS AND ANALYSIS

One of DEA's advantages is no limitation on dimensions of either inputs or outputs, so it's not necessary to take logarithm of each variable to mitigate variances as manipulation in the parametric analysis, or any other procedure to deal with the raw data. However, it's still debatable that whether inputs and outputs should be either all physical measurements or all monetary ones. At present, most scholars haven't paid attention to this issue, and often use mixed measurements of inputs and outputs with DEA. It may be fine if the research is limited in relatively small area where the variation in currency value could be neglected.

When DMUs belong to different regions, states or nations, currency values may vary significantly among these DMUs. For example, a firm located in region A may need

---

[140] The STATA software is also applied as a comparison, and they are almost the same.

less money to rent the land if the real estate price is not as high as that in region B. However, it may still need the same number of employees to run the business if the firm keeps the same size as its counterpart's in region B. As a result, it may produce outputs with same amount of goods but with less inputs in monetary value (the same amount of inputs in physical value) compared with its counterpart in region B. In this case, DEA may offer misleading results (the firm in Region A is more efficient than its counterpart in Region B in production).

Hence, utilizing inputs and outputs with the same features (e.g. physical or monetary) may lead to more accurate analysis. In this paper, except the labor capital, other inputs and the output are all measured with their monetary values. To solve the potential problem, the wage and salary as a proxy of labor capital is used to provide a comparison. In advance to do the regression analysis, a simple t-test is deployed to test whether there is significant difference between two DEA results with distinct types of inputs. Consequently, both Malmquist Productivity Index and its technical change component are significantly different from their counterparts, while there are not significant differences in either scale efficiency or technical efficiency changes in all phases[141]. Finally, each urban area has a list of annual efficiency scores which are in range between 0 and 1, representing values of dependent variables, Malmquist productivity index as well as its components, which will be regressed against all independent variables following Equation 16 listed in Chapter 4.

---

[141] It's interesting that in Phase I, the mean value of technical and TFP changes calculated with the number of labors is smaller than its counterpart estimated with the monetary value of wage and salary, while in Phase II, this result is adverse.

Tables 5-1 – 5-4 display regression results using various dependent variables in different periods and provide the comparison between coefficients resulted from two datasets: one is calculated using the number of employment as an input, while the other one is based on the salary and wage as the equivalent input.

When the Malmquist productivity index is used as the dependent variable (Table 5-1), though results differ in magnitude when we use different types of input measures, most coefficients keep constant in both signs and significances. In Phase I (1990-2000), when all independent variables are included in the regression, no matter which type of input is used, only the national GDP significantly influence the TFP growth positively. Traffic congestion doesn't play a significant role, but in both modes, all coefficients of traffic congestion have negative signs. In Phase II (2001-2009), results in two models are highly similar except R&D and Initial Per Capita Income. Traffic congestion becomes a positive and significant factor for Malmquist Productivity Index, though they are significant at different levels (one at 5% and the other one at 10%). Human capital still has negative coefficients but becomes significant. Another major difference lies on the National GDP which is also significant but with negative coefficients. The consistency of most results in both DEA and parametric analysis proves their robustness to some extent, further enhancing the reliability of my study. As discussed in Chapter 4, the gradual adaptation of both firms and individuals to traffic congestion and indirectly induced change in urban layout may explain the alteration of congestion's role in influencing productivity growth at a whole.

For technical change (Table 5-2), in Phase I, two regressions provide relatively different results. National GDP contributes positively and significantly in both regressions, while Human Capital is only significant at 10% level in one regression but still keeps the same sign in another one. Traffic congestion is insignificant and its sign also varies in two regressions, albeit the sign keeps the same in both regressions with different modes (including agglomeration index or not). In Phase II, results seem more distinctively, although most variables' signs keep constant. Traffic congestion becomes significant and positive when all monetary measurement inputs are used in DEA. This result is a little different from that in the parametric analysis, but it's still similar since in most cases traffic congestion is insignificant. In addition, it also complies with results of regressions on other dependent variables. A possible explanation may be that some hidden-behind factors of traffic congestion prompt technical change and their effects exceed negative ones caused by other determinant factors. Meanwhile, the change of urban layout may also contribute to the average level of technical change in the whole urban area.

Table 5-1 Regression Results of Malmquist Productivity Index

Malmquist Productivity Index	Employment		Salary and Wage		Employment	Salary and Wage
	1990-1997	1990-2000	1990-1997	1990-2000	2001-2009	2001-2009
Traffic Congestion	-.0522	-.0378	-.0487	-.0221	.0876**	.0445*
Human Capital	-.0104	.0156	-.0282	-.0018	-.0643**	-.0430*
R&D	.0008	.0008	.0003	-.0003	.00002	-.0001
Communication	-.0326	-.0792	.0756	-.0511	.0802	.1032
Finance	.0131	.0250	-.0246	-.0001	-.0499	-.0491*
Density	-2.45e-06	-5.03e-07	-3.50e-06	-5.26e-06**	2.41e-06	-1.82e-06
Sectorial Agglomeration	-.0871		-.0534		-.1969**	-.0430
National GDP	.1117**	.0341**	.04500**	-.0079	-.1267**	-.2019**
Initial Per Capita Income	-.0058	-.0259	.1128	.1066	-.0771	.0470
F	4.20**	2.13*	5.74**	5.50**	8.23**	13.93**

*: significant at 10% level;   **: significant at 5% or lower level.

121

Table 5-2 Regression Results of Technical Change

Tech change	Employment		Salary and Wage		Employment	Salary and Wage
	1990-1997	1990-2000	1990-1997	1990-2000	2001-2009	2001-2009
Traffic Congestion	-.0168	-.0095	.0157	.0176	.0398	.3566**
Human Capital	-.0271	-.0178	-.0890*	-.0445**	-.0004	1.266**
R&D	.0006	.0004	.0009	-.0002	-.0002	-.0064**
Communication	-.0591	-.0857	.1569	.1128*	-.0019	-.5631
Finance	.0407	.0258	.1009	-.0247	-.0083	.4482
Density	-1.05e-06	-1.83e-06	-3.05e-06	-3.84e-06	-3.59e-06	-.00001
Sectorial Agglomeration	-.0163		.0589		-.0619	.4139
National GDP	.0772**	.0579**	.1764**	.0492**	-.0293*	-.5245**
Initial Per Capita Income	.0415	.0586	-.0278	.1151*	-.0008	-5.902
F	3.64**	4.82**	7.98**	19.04**	4.03**	46.62**

*: significant at 10% level;   **: significant at 5% or lower level.

Table 5-3 Regression Results of Pure Technical Efficiency Change

Pure efficiency change	Employment		Salary and Wage		Employment	Salary and Wage
	1990-1997	1990-2000	1990-1997	1990-2000	2001-2009	2001-2009
Traffic Congestion	-.1352**	-.0757	-.0976**	-.0541	.0795**	.0358
Human Capital	.1066*	.0669	.0501	.0259	-.1392**	-.0787**
R&D	-.0009	.0001	-.0014	.0001	.0010**	.0006
Communication	-.1627	.0181	-.2522	-.0466	.3477**	.3082**
Finance	-.1923	-.1020	-.0508	-.0422	-.1150**	-.1282**
Density	6.83e-06	6.09e-06	-5.41e-07	2.27e-06	6.97e-06*	3.50e-06
Sectorial Agglomeration	-.0502		.0833		-.0529	-.0559
National GDP	.0188	-.0399**	-.0621**	-.0389**	.0332	.1033**
Initial Per Capita Income	.1877	.1058	.0952	.0104	.0876	.2423**
F	0.80	1.61	2.55**	2.10**	2.91**	1.37

*: significant at 10% level;   **: significant at 5% or lower level.

122

Table 5-4 Regression Results of Scale Efficiency Change

Scale efficiency change	Employment		Salary and Wage		Employment	Salary and Wage
	1990-1997	1990-2000	1990-1997	1990-2000	2001-2009	2001-2009
Traffic Congestion	.1228**	.0589	-.1164*	.0080	-.0357	-.0850**
Human Capital	-.1109*	-.0447	-1.332	.0213	.0872*	.0797**
R&D	.0009	.0003	.0036	-.0003	-.0009*	-.0002
Communication	.1400	-.0215	-2.750**	-.1394	-.3077	-.2051
Finance	.2241	.1247	-.1501	.0636	.0913	.0723
Density	-.00001	-6.38e-06	-.00004	-4.32e-06	-1.15e-06	-3.69e-07
Sectorial Agglomeration	.0133		-.1224		-.0703	.1075
National GDP	.0046	.0134	.1709*	-.0191	-.1449	-.0666
Initial Per Capita Income	-.2855	-.2187	11.46**	-.0117	-.2033	-.0551
F	0.80	0.76	149.9**	1.29	2.29**	1.44

*: significant at 10% level;   **: significant at 5% or lower level.

When the technical efficiency change is considered (Table 5-3), in Phase I, estimators are similar in both regressions. Traffic congestion has negative and significant impacts on the technical efficiency change between 1990 and 1997, the same as that in the parametric analysis. When the agglomeration index is removed from the model, traffic congestion becomes insignificant in period 1990-2000, but still with negative signs. In Phase II, traffic congestion becomes a determinant contributor to the technical efficiency change when employment is used as the input in DEA. In another case, traffic congestion becomes insignificant but has a positive sign. Compared with results in the parametric analysis, results are highly approximate though different in magnitude. In the past decade, congestion's influence in technical efficiency change alters from significantly negative to positive (significant or insignificant depending on how the data is processed), which also demonstrates the change in impacts of traffic congestion

For the last component, the scale efficiency change (Table 5-4), there exist great differences among various regressions. In the category of using the number of employment as an input, results in Phase I show that traffic congestion contributes to the scale efficiency change between period 1990 and 1997, the same as the achievement in the parametric analysis. However, in other cases, traffic congestion becomes insignificant, which is different from results in Chapter 4. When salary and wage are used in DEA, traffic congestion shows significantly negative at 10% level. In Phase II, it is even significant at 5% level, but still negative, controversial to results in the parametric analysis. The robustness of the regression seems weak under this condition. In addition, low F-value also demonstrates that the reliability might be mitigated by the fact that the

model may not fit for a linear relationship very well. In this case, results derived from parametric analysis seem more convincing.

5.4 SUMMARY

In conclusion, results of non-parametric analysis are approximate to those obtained from parametric analysis, illustrating the robustness of our models and analysis. In addition, coefficient estimates based on two different DEA models are similar but not the same. In the regression using results from the DEA with unified measurement of variables, more variables become significant and the model's F-value is relatively larger than that in the other case, which may prove the advantage of applying unified measurements of inputs in DEA. However, compared with parametric analysis, non-parametric analysis is prone to perform poorly in regression, i.e. several models are insignificant at all demonstrating the uncertainty of linear relationship between dependent variable and independent variables. One possible reason may be that variables in the parametric analysis are either directly obtained economic variables or combinations of them, while non-parametric analysis measures the distance and calculate the ratio using linear programing technology that is more mathematically than economically, and may cut off indirect link among variables.

Even though, results from both parametric analysis and non-parametric analysis are proven consistent in most cases, which enhances the confidence of estimated results. As discussed before, traffic congestion's influence in regional economic efficiency differs from the common sense held by most people. In most cases, after 2000, traffic congestion's influence becomes either insignificantly or significantly positive rather than

significantly negative before 2000. Results from non-parametric analysis further

strengthen the reliability of conclusions obtained from parametric analysis.

CHAPTER 6   POLICY IMPLICATION

None of this is to suggest that there is no benefit in having our transportation system operate efficiently. But automobile congestion, vehicle delay, and their proxy, level-of-service, are not measures of system efficiency. Nor are they measures of economic vitality. They are nothing more or less than measures of how convenient it is to drive an automobile. -----Eric Dumbaugh (2012)

Results in both parametric and non-parametric analysis demonstrate a truth that may defy the common sense that traffic congestion seriously harms economic efficiency. Contrarily, in the past decade, traffic congestion's impacts become insignificant and even contribute to regional productivity growth, so do its several components. This result may ascribe to a relatively macro research area in which redistribution or spillover effects indirectly caused by congestion, and some hidden-behind factors that dominate congestion's influence. Another reasonable explanation refers to the adaptation of individuals and firms to traffic congestion. Not only corresponding behaviors, but also plenty of schemes owing to technical development have been executed to minimize congestion's negative impacts, and results based on our analysis may prove their success to some extent.

Nevertheless, it's very important to recognize that the change in traffic congestion's impacts is associated with plentiful efforts devoted to solve this issue, because before 2000 traffic congestion still had negative and significant influences. Dumbaugh (2012) comments that automobile congestion and vehicle delay are not measures of system efficiency. Nor are they measures of economic vitality. Though this comment may not be correct completely (at least not correct before 2000), it

still prompts our further consideration on congestion. Our results show that in the past decade congestion's impacts on regional economic efficiency have altered. To some extent, traffic congestion may not measure system efficiency any more. Eliminating traffic congestion to achieve free-flow condition seems uneconomical and infeasible in practice, especially in flourishing urban areas. By contrary, if a system could achieve its optimal level of congestion at which traffic speed and traffic volume could achieve a perfect balance, and meanwhile social cost could equal to social benefit, the transportation system could be regarded as efficient even though congestion still exists. Therefore, it seems more important to implement reasonable measures to guarantee the fluidity of traffic flow rather than simply mitigate congestion level based on its quantitative measurement, i.e., the 'quality' of congestion is more critical than its 'quantity'.

6.1 MEASURES TO MITIGATE TRAFFIC CONGESTION

In essence, traffic congestion is a result of excessive traffic demand relative to limited traffic supply, i.e. road capacity. In practice, it could be either temporary or permanent, either recurrent or occasional, and either in partial segment or in the whole transportation system. Different methods are fit for distinct types of congestion.

6.1.1 Measures for Non-Recurrent Congestion

Non-recurrent congestion is generally caused by spontaneous, unplanned occurrences, e.g. traffic accidents and incidents, emergency maintenance, and severe weather conditions. Except for the last one, other factors usually result in partial-segment congestion. In addition, most non-recurrent traffic congestions are hardly predictable,

which affects traffic reliability directly. Traffic unreliability is much worse than predictable traffic delay, for which commuters can adjust their travel behaviors correspondingly to guarantee they could still keep their schedules, while unreliable traffic conditions could either make people easily to miss their appointments or affect businesses' operations owing to unexpected late delivery of shipments.

Nowadays, advanced technologies have been widely applied to cope with though not eliminate non-recurrent traffic congestion, a typical example of which is the Intelligent Transportation System (ITS)[142]. ITS is a systematic project that contains various applications, some of which are designed for dealing with non-recurrent traffic congestion, such as emergency vehicle notification systems and collision avoidance systems. The former one could shorten the response time once an accident/incident occurs, while the latter one is helpful in avoiding accidents. Table 6-1 displays some real cases in which incident and accident management is applied as well as their effectiveness.

---

[142] The U.S. Department of Transportation sees the ITS as encompassing a broad range of wireless and wire line communication-based information and electronics technologies.

Table 6-1 Real Cases of Traffic Management Scheme

Location	Program	Results
San Antonio, TX	TransGuide ITS System	In the first year, the incident response times are reduced by 20%.
Maryland, MD	CHART (Coordinate Highway Action Response Team)	In 2009, the average durations for clearing an incident with and without the assistance of CHART were, respectively, about 28.4 minutes and 41.1 minutes. CHART contributed to a reduction in blockage duration of about 31 %.
Hudson Valley, NY	HELP (Highway Emergency Local Patrol)	The average clearance time was approximately 35.5 minutes with HELP, compared to 42.5 minutes average for accidents that occurred on weekday evenings, and a 50.3 minute average for those on weekends. The average response time in weekdays with HELP, is approximately 8 minutes, compared to 20 minutes on weekends, and 12 minutes on weekday evenings.
Oregon	IR (Incident Reponse)	The duration of delay-causing incidents has dropped by 14% to 31%. The average delay per incident has dropped from 36% to 66% on different road segments
New York	IIMS (Integrated Incident Management System)	It can reduce roadway damage incident duration (5%-37% in three case studies) and reduce incident verification and communications times (0-92% in three cases).
Philadelphia, PA	Traffic and Incident Management System .	It has reduced freeway closure time by 55%.
Seattle and Tacoma, WS	Service Patrols	On average, assistance took less than 5 minutes to arrive at the scene compared with previous 10 minutes. The reduction in incident response time ranged from 44 to 77 %.
San Francisco Bay Area, CA	FSP (Freeway Service Patrol)	The average time savings were 16.5 minutes for the FSP assisted breakdowns and 12.6 minutes for accidents based on the field observation.

ITS can also help drivers get real-time traffic information through their mobile phones or other devices (e.g. GPS), so commuters could change routes to avoid being jammed somewhere[143]. In addition, variable speed limit is another strategy to avoid accidents through dynamically adjusting speed limits according to real-time traffic

---

[143] Drivers can even know available parking information and thus reduce the 'mobile congestion' in seeking available parking spaces especially those along road sides.

conditions. This strategy is used to reduce vehicles speed in order to avoid potential collisions in downstream road segments[144].

In sum, because non-recurrent congestion is hard to predict, relevant strategies have to be able to cope with a dynamic situation. Though non-recurrent congestion is annoying, it normally influences road segments rather than the whole system, except that caused by severe weather, and it usually lasts a relatively short time compared with recurrent congestion[145]. In brief, non-recurrent traffic congestion is difficult to forecast but not too difficult to deal with.

6.1.2 Measures for Recurrent Congestion

Recurrent traffic congestion is the predictable delay which happens regularly or periodically. A primary factor that induces recurrent congestion is the high traffic volumes during the same daily periods (peak commute hours or holiday events) and at peak locations (e.g. urban business centers, intersections, interchanges, toll plaza areas, and major long-term construction zones). Therein, rush-hour congestion is the greatest headache in terms of two dimensions: 1). it lasts a relatively long term, usually two or three hours in the morning and the same amount in the afternoon; and 2) it always occurs in certain time periods in a day because of similar working schedules for a majority of people and businesses. In peak hours, the giant volume of traffic flow exceeds the road capacity and causes traffic jam in the whole road network. That's why this issue is

---

[144] Usually, these road segments either have existent accidents, or have slow speed traffic flow caused by various reasons. VSL strategies are used for improving safety rather than alleviating congestion directly.
[145] Strictly, the duration of the delay caused by incidents is determined by temporal traffic condition, severity of incidents and the efficiency of clearance.

difficult to cope with. An intuitive impression is either to increase road supply or reduce

traffic demand, but in practice it becomes much more complex than in theory.

On the supply side, the essence is to augment the capacity of road network. The

term 'capacity' should be considered in two dimensions: static and dynamic. Static

capacity refers to the width of road (a proxy of number of lanes), and the number of roads

in the network[146]. The larger the capacity, the more traffic volumes the system could

accommodate. However, its improvement highly relies on new infrastructure

constructions, which has been proven only effective in a short term after completion of

projects and faded away with the time passing (Downs, 1962). Induced traffic demand is

a primary explanation for this law, since expansion of roads temporarily results in

reduced congestion level that attracts more people to drive. Finally, congestion reappears

even at a higher level than before, because it's impossible to expand the whole network

simultaneously and proportionately, which leads to growing imbalance of capacities to

hold traffic in various road segments. In consequence, cities that spend the most on road

building end up with the worst congestion[147]. Beijing, the capital of P.R. China, has the

---

[146] In a broad sense, other transportation alternative should also be considered as complements of cars. Public transit, such as subway, light rail and bus, plays a very important role in transportation of large cities over the world. However, it has two primary limitations: 1). the service range is limited. Generally, the further to the urban center, the lower the utilization ratio; and 2). huge costs. Construction costs are considerable for subway and light rail, while operation costs are numerous for buses. In United States, the relatively low residential density greatly damages public transit's effectiveness with the poor ridership (even in DC, LA and NY) compared with that in Tokyo, Soule, Beijing, and some European cities. The layout of urban area as well as American people's living culture makes it extremely difficult to popularize public transit everywhere. In addition, almost all public transit projects survive on government subsidies that at present are disliked by majority. Thus, even though they have the necessity to exist, especially for social welfare, they are not the most efficient and sustainable way to mitigate traffic congestion in the United States. As a result, it's not discussed in the context of this chapter.
[147] Sierra Club. New roads are not the answer: Avoiding traffic congestion through transportation choices. Retrieved from http://www.sierraclub.org/sprawl/induced.pdf on May 16, 2013.

widest road (8 lanes in double directions) located in urban center, while it also assumes

worst traffic congestion over the nation. Meanwhile, building new roads cost too much,

especially in already highly dense areas. Expenditure on land appropriation, construction

and maintenance will be a huge burden for U.S. governments suffering considerable

deficits. Expanding roads is almost a mission impossible in urban centers for there is no

more spare space for new constructions. In some peripheral areas, land appropriation is

also difficult, as those lands either belong to private properties or are public reserved

lands. For most of its route through Northern Virginia, US 29 is constructed to possess at

least two lanes in each direction, while the segment passing through the Manassas

National Battlefield Park has only single lane in each direction for approximately three

miles, causing severe traffic congestion in morning rush hours. [148]

Dynamic capacity implies the fluidity of traffic flow. It can be influenced by both

physical and non-physical elements of a transportation system. Road pavement is

classified in the physical category. On-schedule maintenance could increase vehicles'

average speed since drivers don't need to decelerate or change the lane in case of

encountering uneven pavements for safety. Non-physical factors involve reversible

lanes, ramp control, signal coordination, dynamic traffic light, E-Z pass and so on, most

of which could be categorized in the field of ITS. Their principle is to maximize the

traffic flow in road network through appropriately intervening vehicles' movement or

temporarily enlarging the road capacity in a certain direction during particular periods. In

---

[148] The author ever experienced a 45-min driving through this 3-mile segment in the weekday morning during rush hours without encountering any incident or accident.

practice, plenty of relevant schemes have been implemented and receive significant

effects through before-and-after studies.

Table 6-2 Real Cases of Traffic Supply Scheme

Type*	Location	Results
SC	Temecula, CA	Yield citywide improvements in travel time saving by 14% and corridor speeds increment by 17%, as well as reductions in stops by 29%.
SC	Phoenix, AZ	An increase in average speed of almost 10 mph on McKellips Road. A 6 % increase in the average speed, and a 4.3 % less in vehicle stops.
SC	Los Angeles, CA	The average reductions reported in some projects were: 7.4% in travel time, 16.5% in delay, and 17% in stops. For all projects, there are a 11.4% travel time reduction, 24.9% delay reduction, and a 27% stops reduction.
SC	Greenwood Village, CO	13% reduction in travel time and a 17% improvement in travel speed system wide.
SC	Richmond , VA	In the central business district area, travel time decreased by 9 to 14 %, total delay decreased by 14 to 30 %, and stops decreased by 28 to 39 %.
SC	Gloucester, VA	The corridor travel times were improved by 30 to 34 % over the non-coordinated system, while stopped delays on non-coordinated approaches were increased about 14 %.
SC	Syracuse, NY	Reductions in travel time on five main arterials ranged from 1.2% to 34.2 % during the AM peak, -2.7% to 35.1 % during the mid-day period, and -13.9% to 31.2 % during the PM peak.
SC	Oakland, MI	Travel time decreased by 8.6 % in the morning peak direction of travel and 7 % in the evening peak direction of travel. Off peak and non-peak direction travel times were also improved, decreasing 6.6% to 31.8 %.
SC	Tucson, AZ	Reduced travel times at several intersection by 7.9 % and delay by 17.9 %.
ITS	New Jersey Turnpike, NJ	Deployment of E-Z Pass reduced delay for all vehicles at toll plazas by 85 %.
RL	Dearborn, MI	On average, the time required to traverse the reversible segment dropped an average of 16.5% in both peak. The average speeds within the segment increased by an average of 21.6% in both rush hours.
RL	Atlanta, GA	Morning travel times in the major-flow direction decreased by 25% and by 5% in the minor-flow direction. During the evening peak period, travel time reductions were reduced by 24% for flows in the heavier directions and 3.5% in their lighter directions.
RL	Tampa, FL	Provided motorists a trip time of 10 min or less for their morning and afternoon commute into and out of Tampa. Before-speeds of less than 15 mi/h in the peak hours rose to free-flow speeds of about 60 mi/h.
RM	Seattle, WA	The travel time in a 6.9-mile site decreased by 43% in the first two years.

Type*	Location	Results
RM	Milwaukee, WI	Speeds increased by 13% in the segment between Capitol Drive and Burleigh Street, by 10% between North Avenue and Wisconsin Avenue, and by 6% between Bluemound Road and Greenfield Avenue.
RM	Chicago, IL	Individual motorists saved up to 5 minutes in traversing the 3.6-mile section.
RM	Portland, OR	On the northbound I-5, the afternoon peak hour average speed improved from 16 mph to about 40 mph. On average, 39% reduction in travel time, and 60% increase in speed.
RM	Twin Cities, MN	In the absence of metering, there was an increase of freeway point-to-point travel time of 22 % (2.5 minutes) during the peak period on the tested corridor segments, a 7% decreased in speeds. When meters on turned on, the travel time improvement on various road segments is from 4% to 32%.
RM	Other cases in North America	Detroit, MI: 8% increase in speed. Austin, TX: 60% increase in speed. Long Island, NY: 9% increase in speed.

*: SC – Signal Coordination; RL – Reversible Lanes; RM – Ramp Meters

In summary, maximizing dynamic capacity of existent road network seems more feasible than just building or expanding roads. With the rapid development of technologies, less expensive instrument and more advanced management strategies would be applied broadly. Nevertheless, it may still lead to induced traffic demand that may offset previous efforts. That's the reason why it's worthwhile to pay attention to measures on the traffic demand side.

On the demand side, the essence is to constrain overfull traffic volume during peak periods. One strategy is to smooth the uneven distribution curve of traffic flow in a whole day through affecting commuters' travel schedules between rush hours and non-rush hours[149]. Another one is to reduce the traffic demand, using technological, administrative and/or economic measures. The spreads of telecommunication technologies have substituted a certain amount of motorized trips, by providing the possibility that people

---

[149] The limitation of such a strategy is that it only fits for those commuters with flexible working schedules in several industries.

could work at home while communicating with others as well as shopping through internet instead of making physical trips. This strategy is very cost-effective, while it's not a catholicon for traffic congestion yet[150].

Administrative measures are defined as organizational or legal agreements and institutions that encourage or discourage certain activities so as to achieve the objective of alleviating traffic congestion, such as trip reduction ordinance (e.g. odd-even license plate car registration plan and limitation on plate issuance), and HOV (High Occupancy Vehicle) lanes. The greatest deficiency of administrative measures lies in its economic inefficiency. For example, in order to restrict the growth rate of vehicles in Beijing which has suffered the severest congestion in mainland China, people are required to draw lots to obtain a certificate of entitlement issued with limited number annually by the government. Such a policy results in a situation that people who urgently need to purchase a car may fail to obtain the right, let alone it can't move vehicles away from congested streets. The odd-even license plate car registration plan has been proven effective in during 2008 Olympics in Beijing and 2010 Asian Games in Guangzhou, China. Albeit, this strategy even further separates the permit to drive from the demand to drive and results in inefficient allocation of resources. In addition, it could only be implemented temporarily for specific events, otherwise some households may purchase two car, one with odd-number plate and the other one with even-number plate, to cope with the ridiculous policy. Comparatively, HOV lane strategy is the most popular way to

---

[150] Most significant obstacles are posed by industrial, institutional and societal barriers. In some industries, work has to be done on sites. Some businesses prefer face-to-face contacts, and lots of employers are accustomed of supervising their employees in office. In addition, a considerable amount of people really enjoy the feeling of shopping in malls instead of that with keyboard and mouse.

135

mitigate congestion, and encourages travelers to carpool or choose other alternatives, and

reduce the number of vehicles on roads to maintain relatively high traffic speed.

Table 6-3 Real Cases of HOV Lane in the U.S.

Location	Results
CA	The average travel time saving is 1.7 min in the 1171 lane-mile whole HOV system.
Los Angles, CA	Travel times of 37.7 min and 54.3 min along the HOV and GP (General Purpose) lanes, a travel time saving of 16.6 min in 27.5 mile route on I-210 W.
Twin Citie, CA	The estimated current HOV speed is about 56 to 62 mph, while the general-purpose lanes travel at about 32 to 38 mph during the peak periods.
Seattle, WA	HOV time savings in Seattle vary between 30 seconds to 1 minute per mile
Hampton Roads, VA	Four HOV stations always keep speeds near 65 mph, but vehicles in GP lanes have speeds varying from 49 to 65 mph. I-64 HOV provides travel time savings from 0.8 to 2 min in 9.75 miles and a more reliable trip to HOV lane users.

Though HOV scheme is more acceptable than other administrative schemes, it

still has shortcomings. Except the similar inefficiency problem, in the United States, low

residential density outside urban centers makes it more difficult and higher cost to form

carpooling. Moreover, most HOV lanes are not established additionally, but converted

from existent general purpose lanes, such as I-66 segments inbound I-495 belts in DC

area. These HOV lanes occupy already limited and congested road space and force more

vehicles running on other roads where the congestion may further deteriorate.

Economic measures seem more efficient no matter in theory or in practice. Its

principle is based on the economic rationalist assumption: commuters make decisions

only relying on their estimation of private costs and benefits, and they just select the

option with the highest benefit-cost ratio[151]. Simply speaking, economic measures are

---

[151] This assumption is really theoretical, because of its infeasibility at individual level. One difficulty is the incompleteness of information before their departure, such as road and weather condition. Another one is

financial incentives and disincentives to result in commuters' alteration in transportation modes and/or travel periods. Popular economic measures include road pricing, parking pricing and transportation subsidies for public transit[152]. Here, road pricing is the main point we want to discuss[153].

6.2 THEORETICAL BACKGROUND OF ROAD PRICING

Road pricing strategy could stimulate commuters to choose the way that is the most beneficial through redistributing road resources among different travelers with distinct preferences. Vickrey (1963) and Schreiber & Clemmer (1982) summarize its four potential impacts: (a) travelers switch to alternative transportation modes; (b) they adjust their schedules to avoid trips in peak hours when road pricing is levied; (c) drivers may change routes to keep away from charging segments; and (d) people relocate jobs or homes to shorten the commuting distance.[154] Moreover, Schreiber and Clemmer also state that road pricing could stimulate more carpooling, since the lower trip cost assumed by each passenger could counteract the inconvenience of forming car pools, including efforts to find colleagues, sacrifice of individual freedom, and unwillingness of social

---

how to quantify traffic costs by individual, including time delay and wasted fuel if driving, or inconvenience if taking public transit. Generally, commuters just make decisions based on their preferences and feeling. Commercial drives may consider more on this issue.

[152] Subsidies are not efficient in economics, since it's kind of transferred payment mainly used for improving social welfare rather than efficiency. In this segment, economic measures only refer to marketing behaviors instead for governmental activities.

[153] Though the parking pricing scheme also belongs to purely economic issue, it still has some characteristics that limit its effects. It could only influences people whose destinations are urban centers, but not commuter just driving through. In addition, park-and-ride policy is usually associated with public transit program. Its primary purpose is to encourage more people take public transit, and the valuation of its effectiveness should be combined with the subsidy in transit.

[154] The third one might be the most common phenomenon when road pricing strategy is only applied in specific road segments or zones. People who don't want to pay are more jammed on other routes, until they realize that they may assume higher traffic costs and decide to pay for less traffic. The fourth one is not so common, but more and more people begin to consider congestion issue before they locate themselves.

interactions sometimes. These potential impacts sound appealing, and the legitimacy to charge commuters has also been solved by economists.

The theoretical background of road pricing focuses on two dimensions, the property of roads and the externality of traffic congestion. According to Samuelson (1954), public goods should be non-excludable and non-rivalrous[155], but "specific roads and road systems can be and have been… club goods[156], private goods, or common goods, depending upon the institutional environment in which the roads are provided." (Benson, 2006, pp.245). Common goods are non-excludable but normally scarce and subject to rivalry in consumption. As a result of overuse, congestion deteriorates the quality for all users. Since no user is fully liable for the individual cost, there exists negative externality. Most existent roads and streets in congestion comply with above description, so it's concluded that roads should be considered as common goods rather than purely public goods[157]. As a result, it makes sense that roads users could obtain better road service with additional charge.

---

[155] Rivalry: a good's use or consumption by one diminishes its availability for use or consumption by another. Excludability: a good's use excludes another to use it.

[156] In Buchanan (1965)'s classic work, *An Economic Theory of Clubs*, club goods are defined as goods purchased by a club, a voluntarily-formed close-knit group of individuals who have mutually beneficial interactions, and consumed by its all members. Club goods are excludable but non-rivalrous, until reaching a point where congestion happens. If there is congestion with growing number of members, one member's consumption of this good may diminish, though not completely, other members' benefits. Therefore, the optimal population size of a club has to be defined such that the marginal gain from an additional member is equal to the marginal cost of congestion. Roads which are built in gated residential communities are a kind of club goods (Benson, 2006). To some extent, a club good is not a purely public good, though Buchanan believes that the optimal sharing group of a club good "is more than one person or family but smaller than an infinitely large number" (i.e., the whole public) (Buchanan, 1965, pp. 2).

[157] Without congestion, there isn't rivalry among drivers, since each vehicle could run without disturbing others or being disturbed by others. In this case, roads are more likely to be seen as purely public goods.

In addition, externalities of traffic congestion further explain why and how much road users on congested roads should be charged. An additional vehicle that inserts into already congested traffic flow will slow down everyone a little bit and the marginal cost of accommodating this extra vehicle becomes a little higher than before. To achieve economic efficiency, commuters should pay for externalities caused by themselves besides fuel cost and depreciation of vehicles. Otherwise, commuters always underestimate the total trip costs associated with their journeys, which easily results in overuse of roads and thus leads to social efficiency loss.

Pigou (1920) is the first one who introduced the idea of a congestion toll[158]. With a Pigouvian tax imposed, travelers have to reevaluate their trip costs, and some may abandon their original travel plans if updated costs exceed benefits. Finally the number of vehicles on roads during rush hours will be reduced to the level Q_S from the previous

---

[158] Another precursor in this field is Knight (1924) who introduced the idea of solving traffic congestion using Coase approach. Its fundament is to privatize roads and use free market to deal with externality problem, which could maximize social welfare through an efficient allocation of resources. The premise is clearly defining property rights of roads at the beginning, under which no transaction cost exists and the free market mechanism would achieve the optimal level of congestion without government intervention. However, this assumption is unrealistic, since most roads are owned by governments and it will be extremely difficult to privatize roads no matter in techniques, politics or economy. Moreover, limited by geography, roads could not be established freely and there will never be a totally competitive market. "A private road without alternatives close by will likely exploit its locational monopoly characteristics, threatening a diminution of society's welfare" (Timothy, 1998). Even though, Coase's idea still stirs some innovative designs, a typical example of which is the tradable permit system (Goddard, 1997). The total supply of permits is set to achieve specified target for congestion reduction, and then they are distributed to everyone. People could sell or buy permits based on their real need, and the price is determined by the market. In Goddard's proposal, government plays an important role in determining the number of permits, fining vehicles on roads without permits, and enforcing compliance for this policy. However, such a policy proposal requires so many technical details to consider (how many permits to issue? are permits valid during peak hours or the whole day day), let alone political issues (who could get permits from government initially), that it's still just a concept so far.

level Q_C, and the marginal social cost is equal to the marginal social/private benefit with maximum social efficiency (a + b + f + c + g). (Button and Verhoef, 1998)
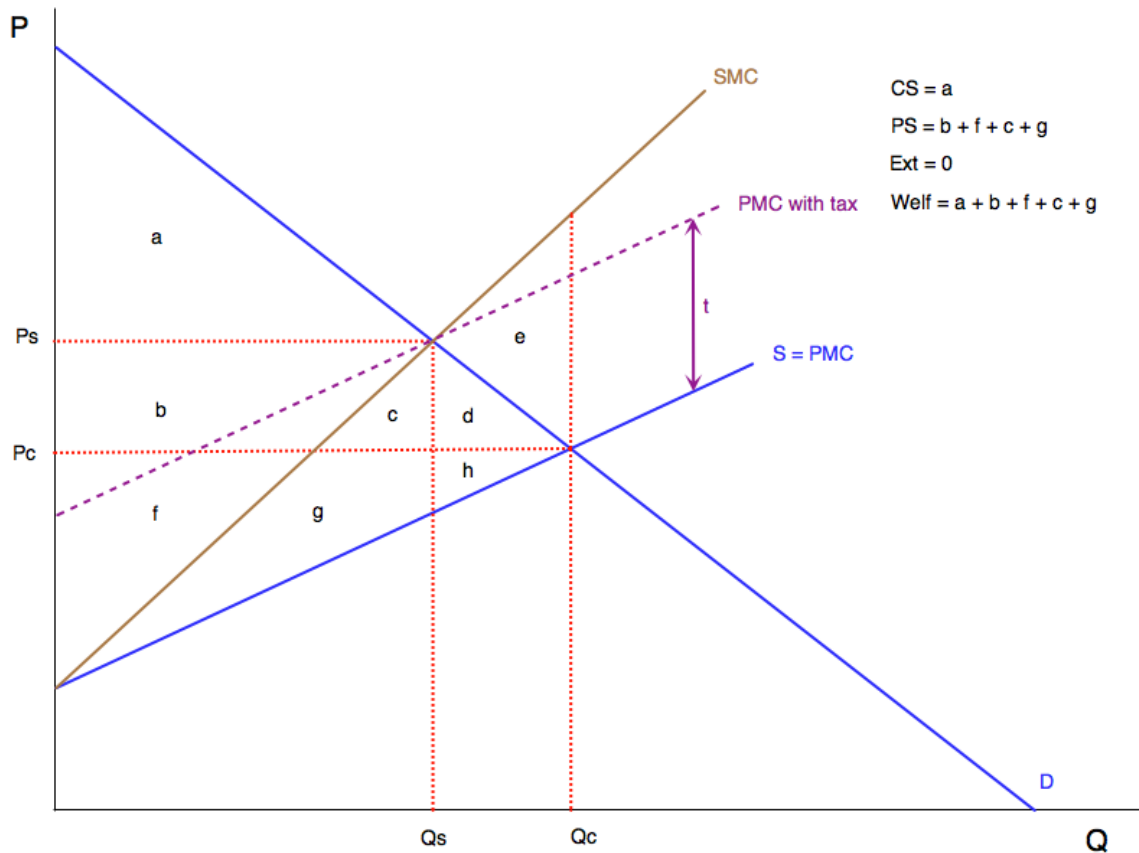

Figure 6-1 Effects of A Pigouvian Tax

Scholars have developed various congestion pricing schemes. Since Walters (1961) introduced a basic model under assumptions of homogenous traffic with identical vehicles, uniform speed and densities along the road, independence of time, and the same cost for each vehicle, more real conditions have been gradually considered in models,

including time elements in Vickrey's bottleneck model (1969), heterogeneity in both travelers and vehicles (Lindsey and Verhoef, 2000)[159], elastic trip demand (Arnott et al., 1993), multiple departure time and routes (Mahmassani and Herman, 1984), non-queuing situations (Henderson, 1974), and general-equilibrium model[160] (Parry and Bento, 2001).

Above models outline the first-best price set to match the marginal cost produced by each traveler. In theory, a first-best congestion price could produce maximum economic benefit, while it has to obey rigid assumptions corresponding to an ideal condition, under which the price is dynamic with various dimensions affecting the actual marginal external costs of each trip, such as mileage, time length, schedule, routes and vehicles used (Verhoef et al. 1996). Besides, the price should be charged in the whole road network rather than several road segments, and all other transportation alternatives should also comply with market rules. That's why a second-best congestion pricing scheme is much easier to accept and implement. Instead of a continuously time-varying toll, a uniform or step tolls seem more practicable because of simplicity and lower management costs, even though they are not fully efficient. Relatively, step tolls yield much greater efficiency gains than uniform tolls because they reduce queuing by altering departure times (Arnott et al. 1990). More schemes, including system with priced and non-priced roads (Marchand, 1968), HOV exemptions (Chu, 1999), elastic demand (Emmerink et. al 1996), the choice on public transit (Tabuchi, 1993), parking charge

---

[159] Travelers have different congestion tolerances, values of time, trip preferences, desired speed and so on, while vehicles differ in their occupied road space, weight and acceleration capabilities, the number of people they can carry, and visual obstruction they impose on other vehicles drivers.
[160] This type of model considers broader fields, such as congestion on un-priced routes, accident and pollution externalities, suboptimal transit pricing, and gasoline taxes as well as impacts on the distribution of income and well-being.

(Glazer and Niskanen, 1992), and vehicle types (Dafermos, 1973), have also been developed and discussed. Furthermore, to involve the situation of incidents/accidents in model, the stochastic methodology is also introduced (Emmerink et al. 1998).

6.3 EMPIRICAL IMPLEMENTATION OF ROAD PRICING

No matter how complicated above models are, the practical application of congestion charge is the ultimate objective. An ideal price strategy (first-best pricing) advocated by Vickrey (1963) possesses following characteristics: (1) is as close as possible to the marginal social cost of each trip; (2) varies smoothly over time to avoid forming other peaks just before a scheduled start or just after a scheduled end; (3) takes into account the impact of a single trip on other traffic from the time when the trip is made until it leaves the congestion period; (4) charged on the basis of the trip length and congestion condition of the trip; (5) charged on the ex post actual impact of experienced trip during congestion; and (6) is universal for all vehicles without exception. Such an ideal scheme provides a perfect template for charging congestion price, but it's too theoretical to implement in practice. Comparatively, the second-best pricing strategy is more practicable. At present, two congestion charge systems are applied in the world.

(1) Toll cordons. Vehicles are charged at specific points in order to travel within a specified area, defined by a boundary. So far, it has been implemented in Singapore, London, and Stockholm[161].

---

[161] In some other cities, such as Oslo, Bergen, Trondheim, Rome and Santiago, the cordon-based pricing scheme has also been applied. However, their primary purpose is not for congestion reduction, but revenue generation (in Norwegian cities), historical district preserving (in Rome), or air pollution reduction (in Santiago).

As the pioneer in cordon-toll, Singapore updates its system from early Area Licensing Scheme (ALS) (1975-1998) to subsequent Electronic Road Pricing System (EPR) (1998-ongoing), and benefits significantly. The initial drop in traffic entering the restricted zone (RZ) was 44%, and the average speed inside the zone in the AM peak hours increased by more than 20%. On most congested streets, traffic speed rocketed to 30 kilometers per hour (kph) from previous 15-18 kph. In addition, on inbound radials leading to the RZ, there was also a 10% increase in traffic speed. After 1998, this scheme was still proven effective. Within the RZ, average speeds were further improved to 40-45 kph from previous 30-35 kph with effectual traffic volume control and restriction.

The central London is another well-known case. During the first few months of this program in 2003, automobile traffic demand declined about 20%. Between 2003 and 2006, average traffic speed during charging days increase by 37% and peak hour delay dropped by 30%[162]. A similar scheme is also implemented in central Stockholm with a time-varying tax levied on most vehicles entering and exiting the center area. In the seven-month trial period in 2006, 22% reduction of traffic was achieved in inner city[163]. Then this scheme was formally approved by the parliament in 2007, and its performance was satisfactory: traffic volume was down by 18% and waiting time to enter the city

---

[162] After 2006, because of a lot of construction projects for 2012 London Olympics being commenced, traffic congestion level was deteriorated even below the level in 2002. (Central London Congestion Charging: Impacts monitoring Sixth Annual Report) Retrieved from http://www.tfl.gov.uk/assets/downloads/sixth-annual-impacts-monitoring-report-2008-07.pdf on May, 20, 2013.
[163] Stein A. (August 7, 2006). Stockholm's successful experiment with congestion pricing. Terrapass. Retrieved from http://www.Terrapass.Com/Blog/Posts/Congestion-Tax on December 19, 2012.

centre during peak hours was reduced by 50%.[164] Similar to London, public transport has seen significant improvement in ridership.

(2) Toll lanes (also called HOT lanes since the fee could be waived for high occupancy vehicles that accommodate enough passengers[165]). This scheme is more popular than toll zones because of its smaller scale and simplicity to implement. Usually, HOT lanes are parallel to general purpose lanes in order to provide more options to commuters. Plenty of case studies have shown that HOT lanes could effectively maintain traffic speed at free-flow level, and reduce travel time significantly.

Table 6-4 Real Cases of HOT Lane in the U.S

Routes[166]	Location	Results
SR 91 Variable-Toll Express Lane[*]	Between Orange and Riverside , CA 4 added 10-mile lanes	During the peak hour, the delay reduction was an estimated 12-13 minutes. The average speed of vehicles in the express lanes over 60 mph (on-and-off peak hours), while the average speed in the freeway lanes averages less than 15 mph, with typical speeds averaging 1-5 mph during peak hours.
SR 167 Express Lane[**]	Between Renton and Auburn, WA. 10.76 miles northbound and 7.69 miles southbound, single lane	The HOT lanes maintain average traffic speeds of 45 mph or more during peak-hours at least 95 percent of the time. The Northbound HOT lanes result in the maximum time savings in HOT compared with general purpose lane is 8.5 minutes. The Southbound HOT lane provided weekday drivers with an average savings of 5 minutes during the peak-hour.
I-394 MnPass [**]	Minneapolis, MN 11-mile single lane	A 20 mph increase in their speed, and those in the general purpose lanes will see a slight increase in speed (up to 15%).
I-15 Express Lanes [**]	San Diego, CA 8-mile two reversible lanes	Free-flow travel conditions were maintained at nearly all times. In the worst-case scenario, I-15 Express lane users can save up to 20 minutes per trip.

[164] Peach J. (August 23, 2011). The success of Stockholm's congestion pricing solution. Cars and Roads, Europe, Featured, Transport. Retrieved from http://thisbigcity.net/the-success-of-stockholms-congestion-pricing-solution/ on April 16, 2013.

[165] Normally three and above passengers are required, varying in different cases.

[166] At present, more and more toll lanes have been established or converted from previous HOV lanes, such as I-35 W express lanes in MN, I-495 express lanes in DC area, and I-25 express lanes in Denver. Limited by space, only a few studies are introduced in the context.

I-10 Express Lanes [**]	Houston, TX 13 mile single lane	Average speed on general-purpose lanes was 25 mph, while average speed on the HOT was 59 mph (over 17-minute time saving).
US 290 QuickRide [**]	Houston, TX 13.5 miles 1 reversible lane	Relative travel time savings were 11 minutes.

[*]: new added lanes
[**]: converted from existing HOV lanes[167]

Many HOT lanes were converted from existing HOV lanes, so vehicles with enough passengers could waive the fee (normally HOV 3+) and only those with less passengers need to pay. On SR-91 toll lanes, 50% discounts are provided for HOV 3+ as incentives for carpooling. Considering that previous HOV lanes normally have HOV 2+ threshold, converted HOT lanes may push some vehicles with only two passengers to drive on GP lanes, and thus induce more traffic on those free lanes. Therefore, in some cases, HOT strategy may result in more congestion on GP lanes during peak periods, such as I-25/US 36 express lanes in Denver. Newly added toll lanes are preferred since they expand existing roads and provide additional capacity, and as a result, traffic speed on GP lanes may also grow as some vehicles could shift to new HOT lanes. Moreover, shift of traffic flow from GP lanes to HOT lanes could also ameliorate traffic condition on GP lanes, such as in I-394 MnPass case: SOV has the option to drive on these HOT lanes now with paying for a fee.

Both toll cordons and toll lanes accept the strategy of variable price bases on time, though it's rather a discrete price than a continuous one. In both AM and PM peak hours, the price rises to summit and drops or even become null in non-rush hours. A long HOT

---

[167] Transportation administration departments increase the threshold for waiving fee, and vehicles with less passengers have the right to pay for running on these lanes.

with several exits may also accept a distance-based pricing scheme that is also time-based simultaneously. The newly established I-495 express lane in DC area charges vehicles with different prices when they select different entries and exits. A more advanced technology is the congestion-metering system with the charge directly related to the prevailing level of congestion on the road network. It varies with the change of real-time congestion situation and is approximate to a dynamic and 'first-best' solution in theory, even though it may not capture the precise marginal cost. On I-15 express lanes, the Fastrack program charged commuters a dynamic per-trip fee which altered based on time of day and traffic flow, and there was a noticeable volume increase by 48% during the entire three-year monitoring period since 1996 Fall. (Small and Gomez-Ibanez, 1998)

Emmerink (1998) illustrates factors which may influence attractiveness of serving congestion pricing schemes in the following table, and the toll cordon system with time-dependent charges seems the most appealing option.

Table 6-5 Assessment of Various Types of Congestion Pricing

Type	Costs	Impact on Congestion	User Friendliness [b]	Side Effects [c]	Implementation [d]	Overall [e]
Cordon_F	+[a]	-	+	-	+	+
Cordon_T	+	+	+	-	+	++
Distance-based_F	-	-	+	-	-	-
Distane-based_T	-	+	+	+	-	+
Time-based_F	--	+	-	+	--	-
Time-based_T	--	+	--	+	--	-
Congestion-metering	--	+[f]	--	+	--	-

F: fixed charges
T: time-of-day dependent charges
[a] The impacts are measured at an ordinary scale ranging from most favorable (++) to most unfavorable (--) with + and – as intermediate level.
[b] Reflecting the predictability for the congestion price. It might be related to users' behavioral responses to current congestion level, knowing the costs of mobility during congested periods.
[c] Reflecting shift in land-use patterns.
[d] Reflecting difficulties associated with putting the theory into practice.
[e] Overall assessment taking all relevant factors into account.

Theoretically, this scheme can lead to optimal traffic volume. However, owing to the low predictability of the congestion price, the impact might be much lower in practice.

However, although congestion pricing is almost impeccable in theory and existing programs have shown its effectiveness in congestion mitigation, many proposals were still abandoned or at least postponed (sometimes indefinitely), including toll-cordon plans for Hong Kong, Cambridge, Randstad and New York. (Small and Gomez-Ibanez, 1998).

One important reason is the "potential Pareto improvement" (Emmerink, 1998). Even though a first-best price could be implemented to maximize social welfare, it's not true that all parties would be better off. Before appropriate redistribution of toll revenues, everybody except government appears to be worse off. Another reason roots in political opposition that public regard congestion price in most cases nothing but a kind of additional tax (Harrington, 2001). People's concern on fairness of using roads is also not ignorable, since they believe that roads are public goods, and all drivers should have equal access to roads, regardless of income and status. Moreover, whether the traffic off the road network owing to congestion price should be regarded as the least essential in social terms is still questionable. In other words, the ability to pay is not synonymous to the importance of the trip. Some business owners also worry that congestion pricing may push some customers to other free areas with lower mobility costs, though this conclusion

is debatable[168]. Last but not the least[169], the amount of charge in real cases is still roughly

determined, since it's impossible to accurately monetize each commuter's marginal cost

covering time values as well as other external costs from congestion.

Though more and more successful projects have demonstrated mitigated pressure

from the opposition side, at present, there are still two concerns focusing on the

redistribution of revenues and double taxing issue. Hence, revenue allocation and fee-

waived scheme are key determinants of the acceptability of congestion pricing by public.

A consensus is returning part of toll revenue to low-income drivers and those living in

toll cordons through travel allowance and tax reductions or subsidizing them directly with

discounted price. (Schreiber and Clemmer 1982; Small, 1992) The left part should be

used to improve transportation services throughout the area including funding new

highways and improving public transit[170]. Even if there is a comprehensive package of

using toll revenues, some critics still doubt government's monopoly power in charging

and using the money. Undoubtedly, government has incentives to raise revenues at the

expense of road users, and shift the money on other public projects, such as health care,

---

[168] In fact, no matter in central London or in Singapore, retailers located in tolled zones haven't suffered lost customers. In adverse, they have better performance. Retailers have seen a 6% increase in business in centre London. (Small and Gomez-Ibanez, 1998; Litman, 2011) One possible reason is that lower traffic flow lead to faster mobility and higher productivity. More convenient public transit service also counteracts reduced vehicles.

[169] Another concern about the privacy issue is caused by electronic road pricing, e.g. the tracking of individual car trips by authorities. However, it seems not so important compared with other issues, since many pricing schemes as well as new technologies could alleviate commuters' worry on privacy.

[170] According to Verhoef et al. (1997)'s survey, in road users' preferences of ways to distribute revenues, the 1st option is the investment in new roads, and investment in public transport (new or better services) and subsidies for public transport (lower fares) rand 3rd and 4th. Indeed, general equilibrium studies of congestion pricing cautiously treat revenue allocation schemes designed solely to improve public acceptability, because these schemes may lead to efficiency loss which may even outweigh the initial improvements (Lindsey and Verhoef, 2000).

education and welfare. In this case, governments have the possibility to overcharge drivers and undersupply necessary road infrastructures and services. (Newbery, 1990)

Above oppositions and suspicions on congestion pricing don't imply a desperate fate of this strategy, but prompt scholars and officials to scheme this strategy more carefully. Jones (1998) proposes four priorities for a feasible congestion pricing scheme: (1) provide conditional free access to the road network in the toll zone, and limit toll charges to certain hours in weekdays[171]; (2) minimize road pricing's impacts on shopkeepers and businesses in order to mitigate the opposition from this sensitive and powerful group; (3) charge discriminative prices based on different vehicle types[172]; and (4) provide reliable and real-time information to drivers to make sure they know what and how much they are paying for[173]. May (1992) also introduces five approaches to achieve public and political support for congestion pricing scheme.

1. Policy-led approach. Congestion pricing should be included in a package of a broad policy that is easier to be approved. Singapore's initial ALS was projected as a part of an overall package of measures to improve transport situations.

2. Technology-led approach. Advanced technologies should be applied. EZ-Pass technology effectively eliminates queue waiting phenomenon at traditional toll plazas and avoids induced congestion at the entrance/exit of toll lanes or boundary of toll zones.

3. Revenue generating approach. The policy which objective is to raise fund for transportation infrastructures' maintenance or construction is often easier to gain public

---

[171] In London, residents in the central area benefit a 90% discount of congestion charge.
[172] In all cordon cases mentioned above, some specified service vehicles are exempted from congestion charging, such as firefighter vehicles and school buses.
[173] Current information technology has made this possible and affordable.

support. Toll rings in Norway are designed primarily to generate revenue at the beginning, and untended congestion mitigation effects were observed with the execution of toll rings in these cities. (Small and Gomez-Ibanez, 1998)

4. Make use of demonstration projects. The case in Stockholm has shown a successful demo. After a seven-month trial program that achieved remarkable congestion relief in 2006, the congestion tax scheme was approved by the parliament and became a permanent policy in 2007.

5. Analysis-led approach. This approach requires a careful and objective assessment made to be resolved, and an analytical program designed to answer them.

Many strategies have been provided to facilitate approval and implementation of road pricing policy, while the core is to how to use the revenue from congestion fee. As King et al. mentioned (2007, p.115), "[c]ongestion pricing will be politically viable when it has well-organized winners who see massive gains, and these massive gains are to be found in the toll revenue." Goodwin (1997, p.2) also comments, "discussion of road pricing without explicit attention to the use of revenue stream is inherently unlikely to be able to command a consensus in its support."

In theory, both Goodwin (1989) and Small (1992) have offered proposals to utilize congestion tolls in ways designed to maximize political support. Goodwin intends to create constituencies who would benefit from toll revenue, while Small attempts to prevent opposition from the group who may lose because of such a policy. Goodwin's solution is to distribute toll revenues in a manner that obtains the wide possible group of supporters, which a third of revenue put toward road improvements, a third toward public

transport, and another third toward the general fund of local government. Such a "Rule of Three" is designed to create potential political beneficiaries and to compensate commuters who pay the tolls. Objected to Goodwin's proposal that devotes too much money on roads and public transit, Small creates his three-way distribution of revenues: one third to travelers, one third to reduction in general transportation-funded taxes, and the last third toward new transport services. In practice, both "Rule of Three" may not be executed strictly. Take London as an example, of the net revenue of $222 million in 2008, 82% went for bus improvement, 9% for roads and bridges, and the remaining 9% for road safety, pedestrian and cycling facilities, borough plans, and environmental improvements.[174] But obviously, most revenues are used in transportation related items.

Rapid development of private-public partnership (PPP) in these years solves another tough problem of shortage in funding for prompting congestion pricing projects. Compared with other transportation projects, congestion pricing scheme provides more incentives to private companies since they could get revenues by charging users. Take the I-495 HOT lanes as an example. This $1.4 billion project is financed by the professional service leader, Fluor Corporation, and the international toll road developer and manager, Transurban. After its completion, Transurban operates and maintains the program, while VDOT owns the facility.

In addition, with the application of advanced technologies, costs of congestion pricing project have been diminished to an acceptable level. Button and Vega (2007)

---

[174] Komanoff, C. (February 15,2013) . Lessons from London after 10 Years of the congestion charge. STREETSBLOG. Retrieved from http://www.streetsblog.org/2013/02/15/lessons-from-london-after-10-years-of-the-congestion-charge/ on August 6, 2013.

compare costs of various international toll cordons and find that average operating cost per user of the road/network was tiny, lower than $0.2 in most cases[175]. Even though capital costs occupy a large proportion in the whole cost, the considerable amount of revenues could also offset the costs and makes these projects economically attractive.

6.4 SUMMARY

This chapter introduces non-recurring congestion and recurring congestion with corresponding measures to mitigate congestion level from both supply and demand sides. Currently, in an already mature transportation network, large-scale construction or expansion of roads seems impractical owing to limited space and huge costs. ITS is a more feasible and cost-effective option to maximize road network capacity based on existent physical infrastructure. On the demand side, administrative schemes seem effective but at the cost of loss in efficiency. Economic measures except subsidies on public transit, such as road pricing, have been proven legitimate in theory and effective in practice. However, political obstacles slow down its popularization. Even though, "the continual rise in car ownership and the reluctance of society to continue to construct infrastructure to meet the resulting unconstrained traffic flow has …[made] authorities …[to adopt] quasi-economic approaches to traffic management and , while not moving to a free market, have begun to use Road Pricing as a device to allocate scarce road space to those who gain most economically from its use." (Button and Vega, 2007, p.284)

---

[175] The highest operating cost is in London, above $4 per user.

Transportation network is a systematic project, so is the congestion alleviation. Single scheme always encounters unintended situation or induces unexpected results. Various solutions should be considered as a package to deal with this complex issue. Road pricing scheme should be considered as a mainstream strategy in the future because of its comprehensive advantages. Similarly, governments also have to double check every transportation policy's potential impacts before implementing it. In the latest China's National Holiday, Chinese government issued a policy that was supposed to encourage trips during the long-holiday period (7 days): fees were waived on all expressways in most provinces for vehicles with less than 7 seats. However, though travelers could apparently waive a lot of highway charge[176] when they took a long-journey, more and more people began to complain this policy: severe congestion everywhere. Obviously, the free-access policy induced huge traffic demand that exceeded the road network capacity, by shifting considerable trips from crowded rails or expensive airlines to free expressways. In addition, to charge unqualified vehicles, except those vehicles installed with electronic devices, other vehicles had to take paper permits at entrance and submit them at exit (indeed nothing was paid for most qualified cars) and that made the congestion even worse[177]. Obviously, this policy seems more like an administrative subsidy, and the effects were totally unexpected and inappreciative even by commuters[178].

---

[176] In China, most highways are expressways that charge travelers. And relative to the average income level, the charge is considerable. The national highway system is outdated with limited lanes and poor pavement.

[177] Electronic device and monitoring system may solve this issue, but the system is not so advanced yet.

[178] There was news that because of the congestion on expressway during that holiday, a pregnant woman had an abortion because she couldn't arrive at hospital on time.

CHAPTER 7   CONCLUSION

Traffic congestion has become an inevitable trouble for modern cities, and congestion costs are an important indicator to measure congestion's negative impacts. However, deficiencies in measurements result in estimate's inaccuracy and unreliability. Econometric analysis provides another way to evaluate congestion's relatively comprehensive effects, including long-term and indirect impacts. Albeit, theoretical drawbacks of introducing congestion as an input in production function model also induce debates. Hence, this thesis discusses whether traffic congestion has significant impacts on regional economic efficiency using parametric and non-parametric analysis, with a dataset covering 31 domestic very large and large urban areas in the United States during 1990 and 2009. And the conversion from SIC code to NAICS code forces us to divide the research period into two segments: 1990-2000 and 2001-2009.

7.1   RESEARCH ACHIEVEMENTS AND ANALYSIS

In parametric analysis, the total factor productivity growth and its components, technical change, scale efficiency change and technical efficiency change are calculated based on the translog production function model using stochastic frontier analysis. In non-parametric analysis, the DEA approach is applied. In this study, DEA is used twice: one with all monetary inputs (salary and wage as the proxy of labor capital, private capital stock and highway capital stock) and the other one with mixed measurement inputs (number of employments as the proxy of labor capital). Then these dependent variables are regressed against travel time index and other control variables. In the period

1990-2000, whether introducing the agglomeration index in the model further leads to two models which are applied for periods 1990-1997 and 1990-2000, severally. In all regressions, multi-collinearity, heteroskedasticity, auto-correlation, cross-sectional dependence and endogeneity in statistics are tested and solved correspondingly

Results show that in parametric analysis, during the first period, traffic congestion has significant and negative impacts on total factor productivity growth no matter which model is applied. However, it has various influences in TFP's components. It harms technical efficiency change, contributes to scale efficiency change, and has no significant influence in technical change. After 2000, results display somewhat different phenomena. Traffic congestion's impacts on total factor productivity growth become positive and its influence in technical efficiency change isn't significant any longer. For technical change and scale efficiency change, traffic congestion still plays the same role. Counterpart in non-parametric analysis match most above results though some differences still exist.

Table 7-1 Summarized Coefficients of Travel Time Index in Both Analysis *

Period	TFP Growth			TC			TE			SE		
	P	DEA_1	DEA_2	P	DEA_1	DEA_2	P	DEA_1	DEA_2	P	DEA_1	DEA_2
1990~1997	-.203	~***	~	~	~	~	-.009	-.135	-.098	.015	.123	-.116
1990~2000**	-.137	~	~	~	~	~	~	~	~	.014	~	~
2001~2009	.146	.088	0.045	~	~	0.357	~	.079	~	.013	~	-.085

*: 'P' denotes parametric analysis. DEA_1 implies DEA with mixed-measurement inputs. DEA_2 is the DEA with monetary inputs.

**: In this period, the agglomeration index is not included in the model because of data's consistency.

***: symbol '~' denotes insignificant at 10% level.

All numbers shown in Table 28 are significant coefficients of travel time index in various models. Except the obvious difference in the category of scale efficiency when monetary inputs are used in DEA, other regressions provide approximate results, which proves the robustness of our analysis. The consistency in both analysis may overthrow the conventional concept on traffic congestion to some extent. Possible reasons can be summarized as:

1.   People adapt to traffic congestion gradually. Commuters may make carpools to drive on HOV lanes or take public transit to save time during rush hours. They may also leave earlier in the morning or later in the afternoon to avoid rush hour traffic. Firms may change delivery schedules to make sure the punctuality of shipment. Flexible working schedules make it possible for commuter to avoid congestion in peak hours. Advanced technologies also provide travelers the possibility that they could deal with some business issues even seating in their slow-moving vehicles. Indeed, as long as the unreliability of trips doesn't deteriorate too much, long-time waiting in the queue may not really reduce productivity significantly, because production rarely occurs outside working schedules. That could explain why in some cases, traffic congestion is insignificant.

2.   Hidden-behind factors of traffic congestion. The reason why traffic congestion should be rejected in the production function model as an input is that congestion could only reflect the result of trade-off between traffic demand and traffic supply. High traffic demand normally indicates more flourishing economy owing to more economic and residential activities. Low traffic supply implies poor road infrastructure and traffic management. Hence, some elements, such as VMT (vehicle mileage travelled), may

157

prompt productivity growth, since more business and individual movement may be beneficial for economic communication. Others, such as poor maintenance and management of transportation system, may harm productivity growth. Therefore, congestion's impacts on economic efficiency are actually aggregated results induced by various factors. In the last decade, improvement in traffic management and advanced technology has facilitated transportation systems to accommodate more travelers even though congestion is still severe. The "essence" of congestion has altered.

      3.   Redistribution effects in urban areas. "When the streets become congested and driving inconveniently, people move to more accessible areas, rebuild at higher densities, travel short distances, and shift travel modes" (Dumbaugh, 2012). Some businesses move to less congested peripheral locations (with convenient accessibility) from congested centers, which improves local productivity from its original low level. As a result, the productivity growth in the whole unban area may encounter an improvement, so does the technical efficiency changes. This process takes time, and that's why in the lasted decade, congestion's impacts became positive from previous negative one except for the scale efficiency change.

      In brief, it's necessary to reconsider traffic congestion's role in the modern economy. As Taylor (2002) mentioned, traffic congestion is the evidence of social and economic vitality: empty streets and roads are signs of failure. Though it doesn't imply that congestion should be warmly welcomed, it still implies that traffic congestion might not be a nightmare in people's urban lives, at least in terms of economic development. It's obligatory to recognize the formation of traffic congestion. Unreasonable urban

layout, poor transportation infrastructure, outdated traffic management system and high traffic demand jointly result in traffic congestion. Except the traffic demand, former three elements may result in low efficiency. With development of advanced technologies, such as ITS, and better maintenance of transportation infrastructure as well as more efforts devoted in traffic management, current transportation system has improved remarkably. However, more and more vehicles on roads still induce higher congestion level. In this case, congestion may not measure the efficiency of transportation system any more at least on the supply side. In other words, the capacity of transportation system has been exerted furthest.

7.2 POLICY IMPLICATION

Even though traffic congestion may not damage productivity growth significantly in the past decade, it still has caused great inconvenience in our daily lives. It may influence people's emotions with anxiety and impatience, which are hardly to define and let alone to quantify these psychological factors' indirect impacts. It will also result in environmental problems which are also difficult to estimate. More importantly, most transportation systems have potentials to improve further. Hence, a comprehensive discussion on various policies dealing with congestion is also provided in this study.

Different types of congestion are discussed with corresponding strategies on both supply and demand sides. Roads construction and expansion are expensive and usually infeasible especially in urban centers. In addition, it could results in induced traffic

which counteracts its effects in long-term.[179] Administrative measures, including governmental control on vehicle registration, permits to drive and subsidies to public transit, are usually effective but not efficient. Technology implementation, such as ITS, is preferred, since these schemes are normally cost-effective, maximizing road network capacity without too much expenditure[180] and effectively reducing the possibility of accidents/incidents as well as shortening response and duration time. Based on economic theories, a transportation system without charging the externalities of congestion couldn't achieve the maximum benefits as well as optimal efficiency. As a result, economic measures, especially road pricing strategy, have been proven effective in reducing vehicle trips in certain areas or routes during peak hours, though so far they are still more likely a second-best approach. In addition, road pricing scheme could reduce unnecessary trips during peak hours. "A fine toll not only determines an optimal time pattern of trips conditional on a given demand, but also the optimal set of users". [181]

Though greatly impeded by public and political oppositions, more and more successful projects have proven congestion pricing's effectiveness. It's predictable that governments become prone to take this scheme to deal with severe traffic congestion, especially under the condition of limited budget. A package of road pricing policy that considers interests of various groups seems easier to receive public support. Meanwhile, trial projects are necessary as demos to persuade public and politicians. Furthermore, it's

[179] One benefit is to accommodate more vehicles in the system, which it may also induce severer congestion.
[180] Compared with expenditure in road expansion or construction.
[181] Lindsey, C. R., & Verhoef, E. T. (2000*). Traffic congestion and congestion pricing* (No. 00-101/3). Tinbergen Institute. pp. 8.

critical for governments to have the idea of systematic plan in designing policies to mitigate traffic congestion, since congestion is a systematic issue rather than an isolated one. Expanding bottle-neck road segments, applying ITS, providing better public transportation service, and executing road pricing strategy are all crucial components in traffic congestion mitigation. Overlooking any part may result in limited effects or more encumber.

7.3   AREAS OF FUTURE RESEARCH

Though we have considered as many factors as we can to enhance the reliability of our models, it's still far from perfectness. Limited by the data's availability and consistency, a relatively long-term period has to be divided into two shorter ones that may lead to many potential problems in statistical regression, even though we have the chance to make comparisons between two periods. For some variables, only second-best options are available. For example, human capital has to be interpolated because of the time interval of the Census survey, and a more reasonable one considering detailed subjects is infeasible. R&D has to be measured using patent rather than the direct investment in this field. In addition, the sample size is not big enough. Medium and small urban areas are not included purposely, because we concern more on those congested ones that mainly are large and very large areas. Nevertheless, if more urban areas could be introduced in the dataset with dummy variable defining its size, it may be useful to include more information in the model.

In the future work, the sample size could be enlarged. More control variables, if possible, should also be introduced to establish better models. Another concern is on the

DEA results. Since this non-parametric analysis calculates the distance from the frontier line through one-round calculation, existence of outliers may result in biased estimates. Thus, the bootstrapping Malmquist index is developed through generating a bunch of sample data and calculating the average value of repeated estimators. Such an approach could provide confidence intervals for the Malmquist index and reduce estimation bias. Meanwhile, industry rather than the urban area should be used as the research objective. Owing to various levels of dependence on transportation network, different industries may perform distinctively facing traffic congestion. For local governments, they may also concern more on their core industries, and such a research could provide more customized guidance on dealing with traffic congestion. In addition, case studies on road pricing will be paid more efforts. Detailed internal mechanism, operation modes, and both short- and long-run effects are our focus in order to deeply and comprehensively understand this economic approach. We look forward to cooperating with transportation departments to design practical plans or trial projects of toll roads/cordons together before long.

APPENDIX

Ray and Desli (1997) argue that $T\Delta_C(x^t, y^t, x^{t+1}, y^{t+1})$ has little do to with the magnitude of a shift in the best practice technology[182], so it may overstate or understate the magnitude of technical change on the best practice technologies. Furthermore, $SE\Delta(x^t, y^t, x^{t+1}, y^{t+1})$ is based on quantity vectors and technology from periods t to $t + 1$, hence there seems to be some double counting of technical changes. As a result, Ray and Desli offer an alternative decomposition of $M_{OC}(x^t, y^t, x^{t+1}, y^{t+1})$ with following form:

$$RD\_M_{OC}(x^t, y^t, x^{t+1}, y^{t+1})$$
$$= TE\Delta(x^t, y^t, x^{t+1}, y^{t+1}) \cdot T\Delta(x^t, y^t, x^{t+1}, y^{t+1}) \cdot S\Delta(x^t, y^t, x^{t+1}, y^{t+1})$$

where

$$TE\Delta(x^t, y^t, x^{t+1}, y^{t+1}) = \frac{D_O^{t+1}(y^{t+1}, x^{t+1})}{D_O^t(y^t, x^t)}$$

$$T\Delta(x^t, y^t, x^{t+1}, y^{t+1}) = \left[\frac{D_O^t(y^{t+1}, x^{t+1})}{D_O^{t+1}(y^{t+1}, x^{t+1})} \cdot \frac{D_O^t(y^t, x^t)}{D_O^{t+1}(y^t, x^t)}\right]^{1/2}$$

and the 'scale change factor'

$$S\Delta(x^t, y^t, x^{t+1}, y^{t+1}) = \left[\frac{SE^t(y^{t+1}, x^{t+1})}{SE^t(y^t, x^t)} \cdot \frac{SE^{t+1}(y^{t+1}, x^{t+1})}{SE^{t+1}(y^t, x^t)}\right]^{1/2}$$

and $SE^t(y^{t+1}, x^{t+1}) = D_{OC}^t(y^{t+1}, x^{t+1})/D_O^t(y^{t+1}, x^{t+1})$.

Compared with FGNZ decomposition, the Ray and Desli efficiency change term is the same. Their technical change term differs in that it is defined on the best practice

---

[182] Best practice technology refers to the technology satisfying variable returns to scale.

technologies, and the scale change factor is the geometric mean of a pair of scale

efficiency ratios which are measured on periods t and $t + 1$ technologies separately. So it

refers to the quantity vectors but not to the technologies. Grifell-Tatje and Lovell (1999)

also decompose their generalized Malmquist productivity index[183], and they obtain

almost the same components as that of Ray and Desli, with the only difference in scale

change factor[184]. Fare et al. (1997) criticize the Ray and Desli's decomposition, because

the scale change factor cannot measure scale efficiency change since each component

uses only a single period technology. In addition, they state that they think of technical

change as change in maximal average product as the response to the suspicion that the

benchmark technology satisfies constant returns to scale instead of variable returns to

scale.

---

[183] $GM_O^t(x^t, y^t, x^{t+1}, y^{t+1}) = \frac{D_O^t(y^{t+1}, x^{t+1})}{D_O^{t+1}(y^{t+1}, x^{t+1})} \cdot \frac{D_O^{t+1}(y^{t+1}, x^{t+1})}{D_O^t(y^t, x^t)} \cdot \frac{D_{OC}^t(y^t, x^{t+1})/D_O^t(y^t, x^{t+1})}{D_{OC}^t(y^t, x^t)/D_O^t(y^t, x^t)}$

$GM_O(x^t, y^t, x^{t+1}, y^{t+1}) = TE\Delta(y^t, y^{t+1}, x^t, x^{t+1}) \cdot T\Delta(y^t, y^{t+1}, x^t, x^{t+1}) \cdot S'\Delta(y^t, y^{t+1}, x^t, x^{t+1})$

[184] $S'\Delta_O(y^t, y^{t+1}, x^t, x^{t+1}) = \left[\frac{SE^t(y^t, x^{t+1})}{SE^t(y^t, x^t)} \cdot \frac{SE^{t+1}(y^t, x^{t+1})}{SE^{t+1}(y^t, x^t)}\right]^{1/2}$

REFERENCES

Acemoglu, D., & Zilibotti, F. (2001). Productivity Differences. *The Quarterly Journal of Economics*, 116(2), 563–606.

Allen, B. (1997). The logistics revolution and transportation. *Annals of the American Academy of Political and Social Science*, 553, 106-116.

Andersen, P., & Petersen, N.C. (1993). A procedure for ranking efficient units in data envelopment analysis. *Management Science*, 39(10), 1261-1265.

Andersen, T. G., & Sorensen, B. E. (1996). GMM estimation of a stochastic volatility model: A Monte Carlo study. *Journal of Business & Economic Statistics*, 14(3), 328-352.

Antras, P. (2004). Is the U.S. aggregate production function Cobb-Douglas? New estimates of the elasticity of substitution. *Contributions to Macroeconomics*, 4(1).

Arellano, M., & Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The Review of Economic Studies*, 58(2), 277-297.

Arnott, R., de Palma, A., & Lindsey, R. (1990). Departure time and route choice for the morning commute. *Transportation Research Part B: Methodological*, 24(3), 209-228.

Arnott, R., De Palma, A., & Lindsey, R. (1993). A structural model of peak-period congestion: A traffic bottleneck with elastic demand. *The American Economic Review*, 83(1), 161-179.

Arrow, K. J., & Debreu, G. (1954). Existence of an equilibrium for a competitive economy. *Econometrica*, 22 (3): 265–290.

Arrow, K. J., Chenery, H. B., Minhas, B. S., & Solow, R. M. (1961). Capital-labor substitution and economic efficiency. *The Review of Economics and Statistics*, 43(3), 225-250.

Aschauer, D. A. (1989). Is public expenditure productive? *Journal of Monetary Economics*, 23(2), 177-200.

Badunenko, O., & Stephan, A. (2007). Lecture 1: Introduction to basic concepts. In *Infratrain*. Retrieved October 6, 2012, from http://www.ersa.org/IMG/pdf/Infra_lecture1.pdf

Balk, B. M. (1993). Malmquist productivity indexes and Fisher ideal indexes: Comment. *The Economic Journal*, 103(418), 680-682.

Baltagi, B. (2001). *Econometric analysis of panel data* (2nd ed.). New York: John Wiley & Sons, Ltd..

Banker, R. D., Cooper, W. W., Seiford, L. M., Thrall, R. M., & Zhu, J. (2004). Returns to scale in different DEA models. *European Journal of Operational Research*, 154(2), 345-362.

Bannister, G. J., & Stolp, C. (1995). Regional concentration and efficiency in Mexican manufacturing. *European Journal of Operational Research*, 80(3), 672-690.

Banum, C. F. (2009). *Instrumental variables and panel data methods in economics and finance*. Retrieved October 1st, 2013 from http://fmwww.bc.edu/GStat/docs/StataIV.pdf

Baram, M. S. (1980). Cost-benefit analysis: **A**n inadequate basis for health, safety, and environmental regulatory decision making. *Ecology Law Quarterly*, 8, 473-532.

Baum, C. F., Schaffer, M. E., & Stillman, S. (2003). Instrumental variables and GMM: estimation and testing. *Stata Journal*, 3(1), 1-31.

Beck, T., Levine, R., & Loayza, N. (2000). Finance and the sources of growth. *Journal of Financial Economics*, 58, 261-300.

Benson, B. L. (2006). The rise and fall of private roads in England. In G. J. Roth (Eds.), *Street smart: competition, entrepreneurship and the future of roads* (pp. 245-276). New Brunswick, NJ: Transaction.

Berg, S. (2010). *Water utility benchmarking: measurement, methodologies, and performance incentives*. International Water Association (IWA) Publishing.

Berndt, E.R., & Wood, D.O. (1979). Engineering and econometric interpretations of energy-capital complementarity. *American Economic Review*, 69, 342-354.

Berry, C. R., & Glaeser, E. L. (2005). The divergence of human capital levels across cities. *Papers in Regional Science*, 84(3), 407-444.

Bertini, R. L. (2005). You are the traffic jam: An examination of congestion measures. In *Transportation Research Board Annual Meeting*. Portland State University. Retrieved October 21, 2011, from http://www.its.pdx.edu/pdf/congestion_trb.pdf

Bjurek, H. (1996). The Malmquist total factor productivity index. *The Scandinavian Journal of Economics*, 98(2), 303-313.

Blundell, R., & Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics*, 87(1), 115-143.

Boardman, A., Greenberg, D., Vining, A., & Weimer, D. (2006). *Cost-benefit analysis: concepts and practice.* Upper Saddle River, NJ: Prentice Hall.

Boardman, N. E. (2006). *Cost-benefit analysis, concepts and practice* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Boarnet, M. G. (1997). Infrastructure services and the productivity of public capital: The case of streets and highways. *National Tax Journal*, 50(1), 39-57.

Bradford, C. (2009, July 14). TTI Hyperbole. *Austin Contrarian*. Retrieved September 30, 2013, from http://www.austincontrarian.com/austincontrarian/2009/07/tti-hyperbole.html

Breitung, J. (2000). The local power of some unit root tests for panel data. In B. H. Baltagi (ed.), *Advances in Econometrics* (Vol. 15, 161–178).

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47(5), 1287-1294.

Bronzini, R., & Piselli, P. (2009). Determinants of long-run regional productivity with geographical spillovers: The role of R&D, human capital, and public infrastructure. *Regional Science and Urban Economics*, 39(2), 187-199.

Buchanan, J. M. (1965). An economic theory of clubs. *Economica*, 32(125), 1-14.

Button, K. J., & Weyman-Jones, T. G. (1994). X-efficiency and technical efficiency. *Public Choice*, 80(1), 83-104.

Button, K., & Vega, H. (2007). The costs of setting up & operating electronic road pricing in cities. *Traffic Engineering & Control*, 48(6), 280-284.

Byrnes, P., Färe, R., & Grosskopf, S. (1984). Measuring productive efficiency: an application to Illinois strip mines. *Management Science*, 30(6), 671-681.

Cavatassi, R., Davis, B., & Lipper, L. (2004). *Estimating poverty over time and space: construction of a time-variant poverty index for Costa Rica* (Working paper No. 04-21). Rome: Food and Agriculture Organization. Retrieved August 6, 2012, from ftp://ftp.fao.org/docrep/fao/007/ae403e/ae403e00.pdf

Caves, D. W., Christensen, L. R. & Diewert, W. E. (1982). The economic theory of index numbers and the measurement of input, output, and productivity. *Econometrica*, 50 (6), 1393-1414.

Ceapraz, I. L. (2008). The concepts of specialisation and spatial concentration and the process of economic integration: Theoretical relevance and statistical measures. The case of Romania's regions. *Romanian Journal of Regional Science*, 2(1), 68-93.

Chang, Y. (2002). Nonlinear IV unit root tests in panels with cross sectional dependency. *Journal of Econometrics*, 110(2), 261–292.

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision-making units. *European Journal of Operational Research,* 2(6), 429–444.

Chen, Q. (2010). *Advanced econometrics with STATA*. Beijing, China: Higher Education Press.

Chichilnisky, G. (1997). The costs and benefits of benefit-cost analysis. *Environment and Development Economics*, 2(02), 195-221.

Christensen, L. R., Cummings, D., & Jorgenson, D. W. (1981). Relative productivity levels, 1947–1973. *European Economic Review*, 16 (1), 61–94.

Christensen, L. R., Jorgenson, D. W., & Lau, L. J. (1973). Transcendental logarithmic production frontiers. *The Review of Economics and Statistics*, 55(1), 28-45.

Chu, X. (1999). Alternative congestion pricing schedules. *Regional Science and Urban Economics*, 29(6), 697-722.

Ciccone, A., & Hall, R. E. (1996). Productivity and the density of economic activity. *American Economic Review*, 86(1), 54-70.

Cingano, F., & Schivardi, F. (2004). Identifying the sources of local productivity growth. *Journal of the European Economic association*, 2(4), 720-744.

Cliff, A. D., & Ord, J. K. (1975). The choice of a test for spatial autocorrelation. In J. C. Davies and M. J. McCullagh (Eds.), *Display and Analysis of Spatial Data* (pp. 54-77). London: John Wiley & Sons.

Coe, D. T., & Helpman, E. (1995). International R&D spillovers. *European Economic Review*, 39(5), 859-887.

Coe, D.T., Helpman, E., & Hoffmaister, A.W. (1997). North-south R&D spillovers. *The Economic Journal*, 107(440), 134–149.

Coeli, T. J., Rao, D. P., O'Donnell, C. J., & Battese, G. E. (2005). *An introduction to efficiency and productivity analysis*. U.S., Springer Science + Business Media, Inc..

Cohen, J. P., & Paul, C. J. M. (2004). Public infrastructure investment, interstate spatial spillovers, and manufacturing costs. *Review of Economics and Statistics*, 86(2), 551-560.

Cohen, W. M., & Levinthan, D. A. (1989). Innovation and learning: the two faces of the R&D. *The Economic Journal*, 99(397), 569-596.

Cooper, W. W., Seiford, L. M., & Zhu, J. (2011). *Data envelopment analysis: History, models, and interpretations*. In W. W. Cooper et al. (Eds.), *Handbook on data envelopment analysis* (2nd ed., pp. 1-39). New York, NY: Springer.

Cortright, J. (2010). Driven apart: How sprawl is lengthening our commutes and why misleading mobility measures are making things worse. *CEOs for Cities*. Retrieved June 12, 2013, from http://www.ceosforcities.org/pagefiles/DrivenApartXSFINAL.pdf

Couclelis, H. (2000). From sustainable transportation to sustainable accessibility: Can we afford a new tragedy of the commons? In D. G. Janelle and D. C. Hodge (Eds.), *Information, place and cyberspace: Issues in accessibility* (pp. 341–356). Berlin: Springer.

Couclelis, H., & Getis A. (2000). Conceptualizing and measuring accessibility within physical and virtual spaces. In D. G. Janelle and D. C. Hodge (Eds.), *Information, place and cyberspace: Issues in accessibility* (pp. 15–20). Berlin: Springer.

Dafermos, S. C. (1973). Toll patterns for multiclass-user transportation networks. *Transportation Science*, 7(3), 211-223.

Dalmau-Atarrodona, E., & Puig-Junoy, J. (1998). Market structure and hospital efficiency: evaluating potential effects of deregulation in a national health service. *Review of Industrial Organization*, 13(4), 447-466.

Delucchi, M. (1998). *The annualized cost of motor-vehicle use in the U.S., 1990-1991: Summary of theory, data, methods, and results* (UCD-ITS-RR-96-3(1)). Institute of Transportation Studies, University of California, Davis.

Demetriades, P.O., & Mamuneas T.P. (2000), Intertemporal output and employment effects of public infrastructure capital: Evidence from 12 OECD economies. *Economic Journal*, 110(465), 687–712.

Denny, M., Fuss, M., & Waverman, L. (1981). Substitution possibilities for energy: Evidence from U.S. and Canadian manufacturing industries. In E. Berndt and B. Field (Eds.), *Modeling and measuring natural resource substitution* (pp. 230-257). Cambridge, MA: MIT Press.

Deno, K. T. (1988). The effect of public capital on US manufacturing activity: 1970 to 1978. *Southern Economic Journal*, 55(2), 400-411.

Diewert, W. E. (1976). Exact and superlative index numbers. *Journal of Econometrics*, 4(2), 115-145.

Diewert, W. E. (1992). Fisher ideal output, input, and productivity indexes revisited. *Journal of Productivity Analysis*, 3, 211-248.

Diewert, W. E. (1992). Fisher ideal output, input, and productivity indexes revisited. *Journal of Productivity Analysis*, 3(3), 211-248.

Downs, A. (1962). The law of peak-hour expressway congestion. *Traffic Quarterly*, 16(3), 393-409.

Downs, A. (2004). *Still stuck in traffic*. Washington DC: Brookings Institution Press.

Doyle, J., & Green, R. (1994). Efficiency and cross-efficiency in DEA: Derivations, meanings and uses. *Journal of the Operational Research Society*, 45(5), 567-578.

Dumbaugh, E. (June 1, 2012). Rethinking the economics of traffic congestion. In *The Atlantic Cities Place Maters*. Retrieved October 1st, 2013 from http://www.theatlanticcities.com/commute/2012/06/defense-congestion/2118/

Durbin, J., & Watson, G.S. (1950). Testing for serial correlation in least-squares regression, I. *Biometrika*, 37(3/4), 409-428.

Elliott, G., Komunjer, I., & Timmermann, A. (2008). Biases in macroeconomic forecasts: Irrationality or asymmetric loss? *Journal of the European Economic Association*, *6*(1), 122-157.

Emmerink, R. H. (1998). Road-Pricing: An Introduction. In *Information and Pricing in Road Transportation* (pp. 35-57). Springer Berlin Heidelberg.

Emmerink, R. H., Verhoef, E. T., Nijkamp, P., & Rietveld, P. (1996). Information provision in road transport with elastic demand: A welfare economic approach. *Journal of Transport Economics and Policy*, 30(20), 117-136.

Emmerink, R. H., Verhoef, E. T., Nijkamp, P., & Rietveld, P. (1998). Information effects in transport with stochastic capacity and uncertainty costs. *International Economic Review*, 39(1), 89-110.

Engle, R.F. & Granger, C.W.J. (1987). Cointegration and error correction: Representation, estimation, and testing. *Econometrica*, 55, 251-276.

Färe, R., Grosskopf, S., & Brännlund, R. (1996). *Intertemporal production frontiers: with dynamic DEA*. Boston: Kluwer Academic.

Färe, R., Grosskopf, S., & Lovell, C. K. (1983). The structure of technical efficiency. *The Scandinavian Journal of Economics*, 85(2), 181-190.

Färe, R., Grosskopf, S., Norris, M., & Zhang, Z. (1994). Productivity growth, technical progress, and efficiency change in industrialized countries. *The American Economic Review*, 84(1), 66-83.

Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society. Series A (General)*, 120(3), 253-290.

Fernald, J. G. (1999). Roads to prosperity? Assessing the link between public capital and productivity. *American Economic Review*, 89(3), 619-638.

Franklin, S. B., Gibson, D. J., Robertson, P. A., Pohlmann, J. T., & Fralish, J. S. (1995). Parallel analysis: a method for determining significant principal components. *Journal of Vegetation Science*, 6(1), 99-106.

Frees, E. W. (2004). *Longitudinal and panel data: analysis and applications in the social sciences*. Cambridge, UK: Cambridge University Press.

Frees, E.W. (1995). Assessing cross-sectional correlations in panel data. *Journal of Econometrics*, 64, 393-414.

Frees, E.W. (2004). *Longitudinal and panel data: analysis and applications in the social sciences*. Cambridge, UK: Cambridge University Press.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32, 675-701.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200), 675-701.

Garofalo, G. A., & Yamarik, S. (2002). Regional convergence: Evidence from a new state-by-state capital stock series. *Review of Economics and Statistics*, 84(2), 316-323.

Garrett, T. A. (2004). *Light-Rail Transit in America Light-Rail Transit in America*. Federal Reserve Bank of St. Louis. Retrieved April 10, 2012, from http://www.stlouisfed.org/community_development/assets/pdf/light_rail.pdf

Geary, R. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician*, 5, 115-145.

Glaeser, E. L., & Gottlieb, J. D. (2009). *The wealth of cities: Agglomeration economies and spatial equilibrium in the United States* (Working paper, No. w14806). National Bureau of Economic Research. Retrieved August 21, 2012, from http://www.nber.org/papers/w14806.pdf?new_window=1

Glazer, A., & Niskanen, E. (1992). Parking fees and congestion. *Regional Science and Urban Economics*, 22(1), 123-132.

Goddard, H. C. (1997). Using tradeable permits to achieve sustainability in the world's large cities: Policy design issues and efficiency conditions for controlling vehicle emissions, congestion and urban decentralization with an application to Mexico City. *Environmental and Resource Economics*, 10(1), 63-99.

Goodbody Economic Consultants (2003). *Transport and Regional Development*. Ballsbridge, Dublin. Retrieved July, 24, 2012, from www.cati.org.pl/download/reg_dev/transport_and_regional_development.pdf

Goodwin, P. (1989). The rule of three: a possible solution to the political problem of competing objectives for road pricing. *Traffic Engineering and Control*, 30(10), 495-497.

Goodwin, P. (October 23, 1997). Solving congestion: Inaugural lecture for the professorship of transport policy. Retrieved November 12, 2012, from http://discovery.ucl.ac.uk/1244/1/2004_22.pdf

Goodwin, P. B. (1989). The "Rule of Three": A possible solution to the political problem of competing objectives for road pricing. *Traffic Engineering & Control*, 30(10), 495-497.

Graham, S., & Marvin, S. (1996). *Telecommunications and the city: Electronic spaces, urban places*. New York, NY: Routledge.

Granger, C. W. (1981). Some properties of time series data and their use in econometric model specification. *Journal of Econometrics*, 16(1), 121-130.

Granger, C. W. J. & Newbold, P. (1974). Spurious regression in econometrics. *Journal of Econometrics*, 2, 111-120.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3), 424-438.

Greene, W. H. (1990). A gamma-distributed stochastic frontier model. *Journal of Econometrics*, 46, 141-164.

Greene, W. H. (2000). *Econometric analysis* (4th edition). Upper Saddle River, NJ: Prentice-Hall.

Greene, W. H. (2008). The econometric approach to efficiency analysis. In H.O. Fried, C. A. K. Lovell, & S. S. Schmdit (Eds.), T*he measurement of productive efficiency and productivity growth* (92-250). Oxford, NY: Oxford University Press.

Grifell-Tatjé, E., & Lovell, C. K. (1995). A note on the Malmquist productivity index. *Economics Letters*, 47(2), 169-175.

Grifell-Tatjé, E., & Lovell, C. K. (1999). A generalized Malmquist productivity index. *Top*, 7(1), 81-101.

Griffin, R. C., Montgomery, J. M., & Rister, M. E. (1987). Selecting functional form in production function analysis. *Western Journal of Agricultural Economics*, 12(2), 216-227.

Griliches, Z. (1980). Returns to research and development expenditures in the private sector. In J.W. Kendrick & B. Vaccara (Eds.), *New developments in productivity measurement* (pp. 419-454). Chicago: University of Chicago Press.

Gwilliam, K. M. (1997). *The value of time in economic evaluation of transport projects: Lessons from recent research* (Working Paper: Transport NO.OT-5). Washington DC: World Bank.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4), 1029-1054.

Harrington, W., Krupnick, A. J., & Alberini, A. (2001). Overcoming public aversion to congestion pricing. *Transportation Research Part A: Policy and Practice*, 35(2), 87-105.

Hartgen, D. T. (2007). Traffic congestion in North Carolina: Status, prospects and solutions. *John Locke Foundation*. Retrieved May 16, 2012 from http://www.ncleg.net/documentsites/committees/21stCenturyTransportation/Prioritization-Best%20Practices-Efficiency/References/TrafficCongestion.pdf

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, 46, 1251-1271.

Hayashi, T. (2012). Smallness of a commodity and partial equilibrium analysis. *Journal of Economic Theory*, 148(1), 279-305.

Henderson, J. V. (1974). Road congestion: a reconsideration of pricing theory. *Journal of Urban Economics*, 1(3), 346-365.

Hoff, A. (2002). *The translog approximation of the constant elasticity of substitution production function with more than two input variables* (Working Paper 14/2002), Danish Research Institute of Food Economics.

Holtz-Eakin, D., Newey, W., & Rosen, H. S. (1988). Estimating vector autoregressions with panel data. *Econometrica*, 56(6), 1371-1395.

Huang, C. J., & Liu, J. T. (1994). Estimation of A Non-Neutral Stochastic Frontier Production Function. *Journal of Productivity Analysis*, 5(2), 171-180.

Hudson, E.A. & Jorgenson, D.W. (1974). U.S. Energy Policy and Economic Growth, 1975-2000. *The Bell Journal of Economics and Management Science*, 5, 461-514.

Hymel, K. (2009). Does Traffic Congestion Reduce Employment Growth? *Journal of Urban Economic,* 65(2), 127-135.

Im, K.S, Pesaran, M. H., & Shin, Y. (2003). Testing for Unit Roots in Heterogeneous Panels. *Journal of Econometrics*, 115(1), 53-74.

Indian Nation Council of Governments (INCOG) (2001). *Congestion management system.* Tulsa, Oklahoma.

Janelle, D. G. (1995). Metropolitan expansion, telecommuting and transportation. In S. Hanson (ed.), *The geography of urban transportation* (pp. 407-434). New York: Guilford.

Jiwattanakulpaisarn, P. (2008). *The impact of transport infrastructure investment on regional employment: an empirical investigation* (Doctoral dissertation, Department of Civil and Environmental Engineering, Imperial College London).

Johansen, L. (1960)**.** *A multisectoral study of economic growth*. Amsterdam: North-Holland.

Johansen, S. & Juselius, K. (1990). Maximum likelihood estimation and inference on cointegration - with applications to the demand for money. *Oxford Bulletin of Economics and Statistics*, 52, 169-210.

Jones, P. (1998). Urban road pricing: public acceptability and barriers to implementation. In: K. Button & E. Verhoef (Eds.) *Road pricing, traffic congestion and the environment* (pp. 263 - 284). Cheltenham, UK: Edward Elgar Publishing

Jorgenson, D. W., & Griliches, Z. (1967). The explanation of productivity change. *The Review of Economic Studies*, 34(3), 249-283.

Jorgenson, D. W., & Nishimizu, M. (1978). US and Japanese economic growth, 1952-1974: An international comparison. *The Economic Journal*, 88(352), 707-726.

Kao C. (1999). Spurious regression and residual-based tests for cointegration in panel data. *Journal of Econometrics*, 90, 1-44.

Key, N., McBride, W., & Mosheim, R. (2008). Decomposition of Total Factor Productivity Change in the U.S. Hog Industry. *Journal of Agricultural and Applied Economics*, 40, 137-149.

Kim, S., & Han, G. (2001). A decomposition of total factor productivity growth in Korean manufacturing industries: A stochastic frontier approach. *Journal of Productivity Analysis*, 16(3), 269-281.

King, D., Manville, M., & Shoup, D. (2007). The political calculus of congestion pricing. *Transport Policy*, 14(2), 111-123.

Knight, F. H. (1924). Some fallacies in the interpretation of social cost. *The Quarterly Journal of Economics*, 38(4), 582-606.

Konya, L. (2004). Unit-root, cointegration and Granger causality test results for export and growth in OECD countries. *International Journal of Applied Econometrics and Quantitative Studies*, 1(2), 67-94.

Koopmans, T. C. (1951). Efficient allocation of resources. *Econometrica: Journal of the Econometric Society*, 19 (4), 455-465.

Kumbhakar, S. (2000). Estimation and decomposition of productivity change when production is not efficient: A panel data approach. *Econometric Reviews*, 19, 425-460.

Leibenstein, H. (1977). X-efficiency, technical efficiency, and incomplete information use: A comment. *Economic Development and Cultural Change*, 25(2), 311-316.

Leibenstein, H. (1966). Allocative efficiency vs. X-efficiency. *American Economic Review*, 56 (3), 392–415.

Leontief, W. (1944). Output, employment, consumption, and investment. *The Quarterly Journal of Economics*, 58(2), 290-314.

Leontief, W. (1953). Domestic production and foreign trade: The American capital position re-examined. *Proceedings of the American philosophical Society*, 97(4), 332-349.

Leontief, W. (1987). Input-output analysis. *The New Palgrave: A Dictionary of Economics*, 2, 860-864.

LeSage, J. P. (1999). *The theory and practice of spatial econometrics* (Working paper 28). Toledo, Ohio: University of Toledo. Retrieved September 12, 2012, from http://raptor1.bizlab.mtsu.edu/s-drive/TEFF/sbook.pdf

Levin, A., Lin, C.F. & Chu, C.S.J. (2002). Unit root tests in panel data: Asymptotic and finite-sample properties. *Journal of Econometrics*, 108(1), 1-24.

Levinson, D., & Kumar, A. (1994) The rational locator: Why travel times have remained stable. *Journal of the American Planning Association*, 60(3), 319-332.

Lindenberger, D. & Kummel, R. (2002). Energy-dependent production functions and the optimization model 'PRISE' of price-induced sectoral evolution. *International Journal of Applied Thermodynamics*, 5(3), 101-107.

Lindsey, C. R., & Verhoef, E. T. (2000). *Traffic congestion and congestion pricing* (Working paper, No. 00-101/3). Tinbergen Institute.

Lindsey, C. R., & Verhoef, E. T. (2000). *Traffic congestion and congestion pricing* (Working paper, No. 00-101/3). Tinbergen Institute. Retrieved November 1, 2012, from http://papers.tinbergen.nl/00101.pdf

Litman, T. (November 24, 2011). *London congestion pricing: Implications for other cities*. Victoria transport policy institute. Retrieved October 23, 2012, from http://www.vtpi.org/london.pdf

Lloyd, P., & Zhang, X. G. (2006). The Armington Model. *Social Science Research Network.* Retrieved July 6, 2012, from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=883565

Lovell, C. K. (2003). The decomposition of Malmquist productivity indexes. *Journal of Productivity Analysis*, 20(3), 437-458.

Lucas, R. E.(1988). On the mechanics of economic development. *Journal of Monetary Economics*, 22(1), 3-42.

Maddala, G. S and Wu, S. (1999). A comparative study of unit root tests with panel data and new simple test. *Oxford Bulletin of Economics and Statistics*, 61(S1), 631-652.

Mahlberg, B., & Sahoo, B. K. (2011). Radial and non-radial decompositions of Luenberger productivity indicator with an illustrative application. *International Journal of Production Economics*, 131(2), 721-726.

Mahmassani, H., & Herman, R. (1984). Dynamic user equilibrium departure time and route choice on idealized traffic arterials. *Transportation Science*, 18(4), 362-384.

Marchand, M. (1968). A note on optimal tolls in an imperfect environment. *Econometrica*, 36(3/4), 575-581.

May, A. D. (1992). Road pricing: An international perspective. *Transportation*, 19(4), 313-333.

Miller J. H. (2004). Tobler's first law and spatial analysis. *Annals of the Association of American Geographers*, 94(2), 284–289.

Miller, S. M., & Upadhyay, M. P. (2000). The effects of openness, trade orientation, and human capital on total factor productivity. *Journal of Development Economics*, 63(2), 399-423.

Miller, S. M., & Upadhyay, M. P. (2002). Total factor productivity and the convergence hypothesis. *Journal of Macroeconomics*, 24(2), 267-286.

Mishra, S. (2010). A brief history of production functions. *IUP Journal of Managerial Economics*, 8(4), 6–34.

Moore, T. (2009). *Congestion, VMT, and Public Policy*. Retrieved May 2, 2012, from http://www.mdot.maryland.gov/Office_of_Planning_and_Capital_Programming/Moving_Maryland/Moving_Maryland/UMD_Background_Paper.pdf.

Moorsteen, R. H. (1961). On measuring productive potential and relative efficiency. *The Quarterly Journal of Economics*, 75(3), 451-467.

Munnell, A. H. (1990a). Why has productivity growth declined? *New England Economic Review*, January/February, 3-22.

Munnell, A. H., (1990b). How does public infrastructure affect regional economic performance? *New England Economic Review*, September/October, 11-32.

Nadiri, M. I., & Mamuneas, T. P. (1994). The effects of public infrastructure and R&D capital on the cost structure and performance of U.S. manufacturing industries. *The Review of Economics and Statistics*, 76(1), 22-37.

Nehru V. & Dhareshwar A. (1994). *New estimates of total factor productivity growth for developing and industrial countries* (No. 1313). Washington DC: The World Bank.

Newbery, D. M. (1990). *Pricing and congestion: economic principles relevant to pricing roads*. UK: Oxford.

OECD/ECMT Transportation Research Center (2007). *Managing urban traffic congestion*. Retrieved January 23, 2012, from http://www.internationaltransportforum.org/Pub/pdf/07Congestion.pdf.

Ohta, M. (1974). A note on the duality between production and cost functions: Returns to scale and the rate of technological progress. *Economic Studies Quarterly,* 25(3), 63–65.

Pakes, A., & Griliches, Z. (1980). Patents and R&D at the firm level: A first report. *Economics Letters*, 5(4), 377-381.

Parry, I. W., & Bento, A. (2001). Revenue recycling and the welfare effects of road pricing. *The Scandinavian Journal of Economics*, 103(4), 645-671.

Pedroni, P. (1996). *Fully modified OLS for heterogeneous cointegrated panels and the case of purchasing power parity* (Working Paper, No. 96–20). Indiana University.

Pedroni, P. (1997). *Panel cointegration: Asymptotics and finite sample properties of pooled time series tests with an application to the PPP hypothesis* (Working paper, No. 95–013), Indiana University.

Pedroni, P. (1999). Critical values for cointegration tests in heterogeneous panels with multiple regressors. *Oxford Bulletin of Economics & Statistics*, 61, 653-670.

Pesaran, M. H. (2004). *General diagnostic tests for cross section dependence in panels*. (Working Papers 0435), University of Cambridge.

Pesaran, M. H. (2004). *General diagnostic tests for cross section dependence in panels* (CESIFO Working Paper NO. 1229). Retrieved September 20, 2012, from http://www.cesifo.de/pls/guestci/download/CESifo+Working+Papers+2004/CESifo+Working+Papers+July+2004/cesifo1_wp1229.pdf

Pesaran, M. H. (2007). A simple panel unit root test in the presence of cross section dependence. *Journal of Applied Econometrics*, 27(2), 265–312.

Pesaran, M. H., Pierse, R. G., & Kumar, M. S. (1989). Econometric analysis of aggregation in the context of linear prediction models. *Econometrica*, 57(4), 861-888.

Pigou, A. (1932). *The economics of welfare, 1920*. London: McMillan & Co.

Pleeter, S. (1980). *Economic impact analysis*. Dordrecht: Martinus Nijhoff.

Prest, A. R., & Turvey, R. (1965). Cost-benefit analysis: **A** survey. *The Economic Journal*, 75(300), 683-735.

Pyatt, G. & Round, J. I. (1977). Social accounting matrices for development planning. *Review of Income and Wealth*, 23(4), 339-364.

Pyatt, G., & Round, J. I. (1977). Social accounting matrices for development planning. *Review of Income and Wealth*, 23(4), 339-364.

Quesnay, F. (1972). *Quesnay's Tableau économiqu* (R. L. Meek, Trans.), (M. Kuczynski & R. L. Meek, Eds.), London: Macmillan. (Originally work published in 1759).

Ramirez, M. D. (2007). A panel unit root and panel cointegration test of the complementarity hypothesis in the Mexican case: 1960–2001. *Atlantic Economic Journal*, 35(3), 343-356.

Ray, S. C., & Desli, E. (1997). Productivity growth, technical progress, and efficiency change in industrialized countries: comment. *The American Economic Review*, 87(5), 1033-1039.

Revankar, S. N. (1971). A class of variable elasticity of substitution production functions. *Econometrica*, 39(1), 61-71.

Rodrigue, J.P. (2013). *The geography of transport systems* (3rd ed.). New York: Routledge.

Romer, P. M. (1990). Endogenous technological change. *Journal of Political Economy*, 98(5), S71-S102.

Roodman, D. M. (2009). A note on the theme of too many instruments. *Oxford Bulletin of Economics and Statistics*, 71(1), 135–158.

Roth, G. J. (Ed.). (2006). *Street Smart: Competition, Entrepreneurship, and the Future of Roads*. Oakland, CA: Transaction Publishers.

Samuelson, P. A. (1954). The pure theory of public expenditure. *The Review of Economics and Statistics*, 36(4), 387-389.

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica*, 26(3), 393-415.

Say, J. B. (1836). *A treatise on political economy: Or the production, distribution, and consumption of wealth*. Philadelphia, PA: Grigg & Elliot.

Schmidt, P., & Lovell, K. C.A. (1979). Estimating technical and allocative inefficiency relative to stochastic production and cost frontiers. *Journal of econometrics*, 9(3), 343-366.

Schreiber, A. F. & Clemmer, R. B. (1982). *Economics of urban problems* (3rd ed.). Boston: Houghton Mifflin.

Schreyer, P. (2000). *The contribution of information and communication technology to output growth: A study of the G7 countries* (STI Working Paper, 2000/2). Paris: Organisation for Economic Co-operation and Development.

Schultz, T. (1962). Reflections on investment in man. *Journal of Political Economy*, 70(5), 1-8.

Seitz, H. (1993). A dual economic analysis of the benefits of the public road networks. *The Annals of Regional Science*, *27*(3), 223-239.

Seitz, H., & Licht, G. (1995). The impact of public infrastructure capital on regional manufacturing production cost. *Regional Studies*, *29*(3), 231-240.

Sen, A. (1963). Neo-classical and neo-Keynesian theories of distribution**.** *Economic Record*, 39(85), 53-64.

Sen, A. (2000). The discipline of cost-benefit analysis. *The Journal of Legal Studies*, 29(S2), 931-952.

Sena, V. (2003). The frontier approach to the measurement of productivity and technical efficiency. *Economic Issues-Stoke on Trent*, 8(2), 71-98.

Seo, J. H. & Lee, Y. S. (2006). Contribution of information and communication technology to total factor productivity and externalities effects. *Information Technology for Development*, 12(2), 159-173.

Shephard, R.W. (1970). *Theory of cost and production functions*. Princeton, NJ: Princeton University Press.

Shin, Y. (1994). A residual-based test of the null of cointegration against the alternative of no cointegration. *Econometric Theory*, 10, 91-115..

Shirley, C., & Winston, C. (2004). Firm inventory behavior and the returns from highway infrastructure investments. *Journal of Urban Economics*, 55(2), 398-415.

Small, K. (1992). Using the revenues from congestion pricing. *Transportation*, 19(4), 359-382.

Small, K. A. (1992). Using the revenues from congestion pricing. T*ransportation*, 19(4), 359-381.

Small, K. A., & Yan, J. (2001). The value of "Value Pricing" of roads: Second-best pricing and product differentiation. *Journal of Urban Economics*, 49, 310-336.

Small,K.,& Gómez-Ibáñez, J. (1998). Road pricing for congestion management: The transition from theory to policy. In K. Button and E.Verhoef (Eds.), *Road pricing,*

traffic congestion, and the environment: issues of efficiency and social feasibility (pp. 263-284). Cheltenham, UK: Edward Elgar Publishing .

Smock, D. (2003). Supply chain management: What is it? *Purchasing*, 132(13), 45.

Solow, R. (1956). A contribution to the theory of economic growth. *Quarterly Journal of Economics*, 70(1), 65-94.

Solow, R. M. (1957). Technical change and the aggregate production function. *Review of Economics and Statistics*, 39(3), 312-320.

Stephan, A. (2003). Assessing the contribution of public capital to private production: Evidence from the German manufacturing sector. *International Review of Applied Economics*, 17(4), 399–418.

Stock, J. H., & Yogo, M. (2005). Testing for weak instruments in linear IV regression. In: J.H. Stock & D.W.K. Andrews (Eds.). *Identification and inference for econometric models: a festschrift in honor of Thomas Rothenberg* (pp. 80-108). Cambridge: Cambridge University Press.

Stone, J. R., & Brown, A. (1962). *A computable model of economic growth: A programme for growt*h. Cambridge: Chapman & Hill.

Stone, R. & Brown, A. (1962). *A computable model for economic growth*. Cambridge, UK: Cambridge Growth Project.

Tabuchi, T. (1993). Bottleneck congestion and modal split. *Journal of Urban Economics*, 34(3), 414-431.

Talluri, S. (2000). Data envelopment analysis: models and extensions. *Decision Line*, 31(3), 8-11

Taylor, B. D. (2002). Rethinking traffic congestion. *Access*, 8, 8-16.

Terry M. (2009). *Congestion, VMT, and public policy: Discussion paper for a transportation symposium.* Retrieved August 30, 2012, from http://www.mdot.maryland.gov/Office_of_Planning_and_Capital_Programming/Moving_Maryland/Moving_Maryland/UMD_Background_Paper.pdf

Thissen, M. (1998). *A classification of empirical CGE modelling*. Groningen, Netherlands: University of Groningen.

Timmer, C. P. (1971). Using a probabilistic frontier function to measure technical efficiency. *The Journal of Political Economy*, 79, 776-794.

Timothy, D. H. (1998), Congestion pricing and road investment. In K. J. Button & E. T. Verhoef (Eds.). R*oad pricing, traffic congestion and the environment* (pp. 39-78). Cheltenham, England : Edward Elgar Publishing Ltd..

Tobler W., (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(2), 234-240.

Törnqvist, L. (1936). The Bank of Finland's consumption price index. *Bank of Finland Monthly Bulletin*, 10(1936), 1-8.

Verhoef, E. T., Nijkamp, P., & Rietveld, P. (1997). The social feasibility of road pricing: a case study for the Randstad area. *Journal of Transport Economics and Policy*, 31(3), 255-276.

Verhoef, E., Nijkamp, P., & Rietveld, P. (1996). Second-best congestion pricing: the case of an untolled alternative. *Journal of Urban Economics*, 40(3), 279-302.

Vickrey, W. S. (1963). Pricing in urban and suburban transport. *The American Economic Review*, *53(*2), 452-465.

Vickrey, W. S. (1969). Congestion theory and transport investment. *The American Economic Review*, 59(2), 251-260.

Walker, N., & Catrambone, R. (1993). Aggregation bias and the use of regression in evaluating models of human performance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 35(3), 397-411.

Walras, L. (1954). *Elements of pure economics* (W. Jaffe, Trans.). Homewood, IL: Richard D. Irwin, Inc.. (Originally work published in 1874).

Walters, A. A. (1961). The theory and measurement of private and social cost of highway congestion. *Econometrica: Journal of the Econometric Society*, 29(4), 676-699.

Weber, A. (1929). *Alfred Weber's theory of location of industries* (C. J. Friedrich, Trans.). Chicago: University of Chicago Press. (Originally work published in 1909).

Westerlund, J. (2005). New simple tests for panel cointegration. *Econometric Reviews*, 24(3), 297-316.

Winston, C., & Langer, A. (2006). The effect of government highway spending on road users' congestion costs. *Journal of Urban Economics*, 60(3), 463-483.

CURRICULUM VITAE

Junyang Yuan received his Bachelor of Engineering from Tsinghua University, China in 2003, and her Master of Management from Tsinghua University in 2006. His research interests and experiences mainly focus on transportation and urban development, particularly in transportation planning and policy, highway and aviation research, and regional economic development. While attending the graduate school, he also worked at the National Association of Counties.

.