Chapter 1
Introduction

1.1. Urban congestion

Urbanization is perhaps the milestone of the last century’s huge economic growth, emerging from it and simultaneously causing it. The current trends show an uninterrupted process and foreshadow the total triumph of the city: recent forecasts estimate that 70% of the world population -which raises to 86% for OECD countries- will be living in urban areas by 2050. In the same time horizon, global urban land cover is projected to increase from 603,000 km$^2$ in 2000 to over 3 million km$^2$ in 2050 (Angel et al., 2011), hosting 6.5 billion people. Undoubtedly, such a global transformation will intensify the existing significant challenges faced by modern cities: environmental degradation, air pollution and the efficient provision of public goods and services.

This dissertation focuses on one of these major challenges. Urban congestion, i.e. the overcrowding of urban roads and suburban highways by private vehicles during peak-hours, is a post-war phenomenon underlied by a series of decisive factors: the constant revolution in the automobile industry, the abundance of cheap (undertaxed in many countries) oil and the immense investment in highway construction. These factors gradually dwarfed the private cost of automobile use, creating what is widely known as car-dependent societies. Urban congestion is by no means an innocent byproduct of this process. Maddison and Pearce (1996) estimated the on-peak marginal external cost of it in the range of £0.16-0.37 (in 1990 prices) per vehicle kilometer. Schrank et al. (2012) computed that, for the US, urban road congestion in 2011 caused a total of 5.5 billion hours of delay. Rough calculations reveal that only the time costs of congestion sum up to approximately 0.9% of GDP.$^1$ The same report estimates that in 2011 the value of wasted fuel (due to congestion) summed up to $121 billion, or an additional 0.78% of the country’s GDP. On top of these numbers, someone should also consider the external social cost of pollution (health and other indirect costs) generated from the combustion of the 2.9 billion gallons of fuel wasted due to congestion in the US; these costs are associated with the risks traffic-related pollutants, e.g. particulate matter (PM 2.5 and PM 10), impose on human health, which may be substantial (Lelieveld et al., 2015).

Several practical remedies to urban congestion have been proposed (the enumeration that follows below is by no means exhaustive) and practiced during the previous decades. Road capacity expansion (widening the existing road lanes) and network augmentation (adding new links to an existing network) have been the main policy responses, providing temporary relief but generating long-term incentives for car ownership and wider use of private vehicles. The outcome can be the so-called Downs-Thomson paradox: more congestion, more private road

---

$^1$ Using the average household income reported by U.S. Census Bureau, and assuming 240 working days per year, eight working hours per day and a mean value of time at 70% of hourly income.
infrastructure to be maintained costly in the long-run, and larger public transport deficits to be covered through distortionary taxation.\(^2\) Recently, similar remedies that approach the issue from the supply side suggested that a more intelligent transport system, in which traffic flows are managed more efficiently through various technological breakthroughs, may be key in combating urban congestion. But these technological breakthroughs will most probably not be widely available in the very near future, and it is difficult to predict how drastic they are going to be in providing a viable solution on their own. Like capacity investments, these technologies may create incentives for private vehicle use that wipe out a large part of the initial benefits. Furthermore, like all supply side interventions, they are going to require substantial investments, to be funded by existing highly-distortionary taxes in the economy.

In many cities (mainly in Asia), public transport has managed to generate a modal split that keeps traffic congestion in reasonable levels. In other parts of the world (especially in North American cities), a series of policies (density regulations, zoning) persistently incentivized the dispersion of residences from jobs, giving rise to the phenomenon of urban sprawl: low-density, fragmented expansion of the suburban fabric at the detriment of countryside. Urban sprawl hampers the efforts to curb congestion, because it renders the provision of public transportation in acceptable levels of service financially disadvantageous, both in terms of capacity (frequency) but also in terms of spatial coverage. Capacity costs increase due to the necessary increments in bus and rail kilometers needed to maintain an acceptable level of service. Coverage costs increase because as cities expand radially, accessibility to the pre-existing public transport network falls. Without costly network expansions, the modal split is inelastic in capacity increments. Downs (2004) claims that, even if transit capacity in the US quadrupled morning peak-hour transit would not exceed 11% of trips and private vehicle trips would not fall by more than 9%. Therefore, for these parts of the world, the study of urban congestion cannot be detached from the study of the tremendous momentum urban sprawl currently possesses.

Peak-hour congestion pricing has received increasing attention during the last decades as a viable alternative to many of the above practices shown to have limited success in curbing the problem of traffic congestion in urban areas. Unlike the above measures, congestion pricing increases the private cost of car use; it is therefore not subject to the feedback effects capacity expansion suffers from and can be more effective in reducing congestion. Furthermore, congestion pricing generates revenues in contrast to supply-side approaches that require funds to be accomplished. This is an important factor: in an era in which many economies undergo fiscal consolidation, revenue from externality taxation can contribute in balancing public budgets, or in financing public spending that would otherwise derail these budgets.

To some extent, the comparative advantages of congestion pricing have been mirrored by an observable policy trend: a growing number of cities have adopted or consider adopting the scheme. Cordon tolls have been introduced in Norway (Oslo, Trondheim), Sweden (Stockholm, Göteborg) and Italy (Milan) and have been shown to be effective in reducing peak-hour congestion. In some cases, follow-up studies provide support to the theoretical findings by

\(^2\) Mogridge (1997) offers a review of the literature on the paradox.
showing significant average welfare gains from these measures, generating enthusiasm in part of the scientific community and policy makers. Some do not share this enthusiasm, focusing on the relatively low political acceptability such measures have enjoyed so far. No matter where the truth stands, it is the obligation of researchers to explore the properties of congestion taxes in depth and add useful insights in the public discussion of this timely topic. The following dissertation contributes towards this direction.

1.2. Background and research questions

Under certain theoretical conditions, road pricing can maximize economic efficiency by reducing traffic to its optimal level. The fundamental prerequisite is knowledge of the spatiotemporal variation of the externality generation and propagation mechanisms. At any given time point, vehicles in different locations of a network cause different external costs, because traffic conditions are different: congestion is not linear in flows and the average socioeconomic characteristics of the drivers composing the flow in each location are different. Similarly, vehicles at any given location of a network generate different external costs at different time points. If the externality propagation mechanism (i.e. what every user \( x \) causes to each user \( y \) in the network) and the valuation of the externality (i.e. what each user \( y \) is willing to pay to avoid the external effect of \( x \) on her) are known with certainty, an optimal road pricing scheme (that varies over space and time) can be designed. Furthermore, in order for such a first-best scheme to be feasible, the costs of acquiring such information and implementing a respective pricing scheme must be reasonable.

In addition to the above technological and informational requirements, the implementation of a first-best scheme would also require political acceptability of a measure that charges every link in the urban network, leaving no alternative route untaxed. Combined, these requirements limit severely the practical applications of first-best (i.e. Pigouvian) road taxes. Until today, no city has implemented a scheme that may be considered -in efficiency terms- to be sufficiently close to the first-best. Singapore is perhaps an exception, being still far-off but substantially closer to that benchmark.

The practical limitations of Pigouvian road pricing motivated a voluminous literature that explores the properties of various second-best schemes that leave parts of a road network untaxed. Some contributions in this partial taxation stream derive rules for optimal road pricing in a generic static network. For instance, Verhoef (2002a; 2002b) offers a general analytical solution for the second-best problem where not all links of a congested network can be charged; an algorithm based on this analytical solution is then tested on a medium size network. Also, Van Dender (2004) shows that constraints in network pricing can cause the optimal toll to deviate in a complex way from the marginal external cost of congestion. The cordon toll, a particular instance of partial taxation, received special attention due to its practical relevance. Some contributions suggested that it is highly efficient, i.e. it captures a surprisingly large fraction of
the first-best theoretical gains in various spatial configurations, especially in monocentric settings (Verhoef, 2005).

The recent advances in double dividend theory (e.g. Parry and Bento, 2001) have highlighted another reason for which second-best settings may emerge. This regards the existence of at least one distortionary tax in the economy (even inside the transport system, e.g. public transport subsidies) that interacts with the externality tax on the road, with the interaction taking the form of a common base the two taxes partially share. Perhaps the most illustrative example of such a distortionary tax is the tax on labor income. To a large extent, urban tolls and labor income taxes share the same base (labor supply) as the majority of peak hour trips are commuting trips. If modal substitution is impossible (i.e. the elasticity of substitution between car and public transport is zero) and labor supply is elastic only at the extensive margin (i.e. labor adjusts only through the number of working days, but not through the duration of a working day) the two tax bases overlap completely: increasing the toll or the labor tax by \( x \) monetary units triggers an identical behavioral response, as both taxes reduce the private benefit of labor supply (and thus of commuting trip) equally. As the modal elasticity of substitution increases and labor becomes adjustable in the intensive margin (i.e. by adjusting the duration of the working day), the two tax bases detach from each other and the tax interaction effect fades out.

In the presence of a relevant distortionary tax, road taxation affects welfare through two channels: it offers externality abatement and alters the burden of the tax in the distorted market. In such a setting, Pigouvian taxation is only a quasi first-best because it is suboptimal to begin with. In fact, as illustrated by Parry and Bento (2001), the quasi first-best toll under tax base overlap can be welfare decreasing, indicating that the optimal level of the tax lies below the marginal external cost. In some cases, traffic in the no-toll equilibrium lies below its welfare maximizing level, calling for a negative optimal toll. Double dividend literature suggests that the introduction of the externality tax should be accompanied by a reduction in the pre-existing distortionary tax. That is, the revenue from the former tax should finance cuts in the latter tax in the form of a revenue-neutral tax swap.

Double dividend introduces a series of policy-relevant questions that are fundamental in the taxation of urban congestion and other road externalities. Can urban tolls provide welfare gains without a sophisticated revenue-recycling program? If yes, are the welfare gains substantial? What is the role of the labor and property tax in optimal congestion pricing? Is knowledge over marginal external costs on the road a sufficient condition to design welfare-increasing, revenue-neutral tax swaps? In a period that is considered by many as a turning point in congestion management, answers to these questions are timely because they increase our understanding over the economic efficiency of road pricing. In an era in which the idea of road pricing begins gaining momentum, these answers will help in understanding the extent to which policy agendas that were previously confined to provide separate solutions by separate authorities at different administrative levels (urban planning principles, transportation planners,
local and federal governments) will have to merge or be restructured in order to provide vertically efficient solutions to congestion externalities.

This thesis attempts to provide answers to the above research questions with a series of contributions lying at the intersection of three streams of literature, which are summarized in Section 1.3: congestion pricing, urban economics and the theory of double dividend. All research questions are addressed with the use of general equilibrium modeling, which constitutes the common methodological denominator of the entire dissertation.

Figure 1.1. Schematic outline of the thesis

Chapter 2 examines the design of congestion taxes in a monocentric city model with pre-existing labor taxes, repositioning some of the key results provided earlier by the non-spatial model of Parry and Bento (2001). The latter show that a high elasticity of labor supply may render the quasi first-best scheme (i.e. marginal external cost pricing) welfare-decreasing (see Section 1.3). However, the lack of a spatial dimension prevents the investigation of questions involving a differentiation of road taxes over space. By adding space to the analysis, Chapter 2 facilitates the evaluation of second-best congestion pricing schemes such as the cordon toll (Mun et al. 2003; 2005) or a flat kilometer tax (Sullivan, 1983) in the presence of a uniform-in-space labor tax. One of the key findings is that, under lump-sum revenue recycling, a cordon toll may generate welfare gains even when the quasi first-best scheme causes substantial welfare losses. The finding has substantial practical applications, since it highlights the conditions under which an archetype second-best scheme becomes superior to the Pigouvian scheme, whose informational requirements and implementation costs are very high. The same chapter shows
that, when road toll revenue is used to finance labor tax cuts, 35% of the optimal road tax does not reflect marginal external congestion costs, but rather functions as a Ramsey-Mirrlees tax, i.e. an efficiency enhancing mechanism allowing for an indirect spatial differentiation of the labor tax. This result adds a quite different motivation to road pricing, since welfare gains can be produced even in absence of congestion.

Chapter 3 investigates second-best congestion pricing in a monocentric city characterized by distortionary regulatory mechanisms in the housing and land markets (such as building height restrictions and zoning) as well as distortionary property taxation. Both interventions have in common that they affect the welfare gains that externality pricing in the primary market of interest (road transport) could bring. But an important difference is that, whereas tax-induced distortions invoke deviations from the Pigouvian tax rule (i.e. the optimal tax deviates from the marginal external cost), such deviation is not efficiency enhancing in the case of quantitative restrictions (command and control regulations). The findings highlight the conditions under which road demand management can be detached from urban planning and public finance decisions. This detachment bears high policy relevance, since the authorities that control the respective policy instruments are horizontally and vertically scattered in the governmental structure. More important, the policy implications are general: Chapter 3 shows that the question of whether generic externality taxation must be adjusted does not so much depend on whether distortionary policies in related markets exist, but much more on the type of these distortions.

Finally, the chapter provides an extensive sensitivity analysis of the Pigouvian toll welfare gains across cities with identical household preferences and road technologies but different building height limits. This analysis explores the extent to which gains from congestion charges for cities without any housing regulation (as commonly assumed in the literature) can be extrapolated to cities with restrictive land-use regulations.

The second part of the dissertation regards analyses of polycentric structures. Chapter 4 provides the general structure of a General Equilibrium model of Land Use and Transport (GELUT), which is one of the few consistent options in modeling polycentricity; GELUT models are juxtaposed vis-à-vis prior approaches, such as Land Use and Transport Interaction (LUTI) models, that have received severe criticism for the ad hoc way prices and quantities are determined (see also Section 1.3.3.3). The chapter focuses on the complex and large-scale nature of GELUT models and highlights why their accurate and fast solution by numerical methods is an important issue. It provides the specification of two candidate solution approaches and the detailed anatomy of the respective algorithms. The first approach is based on an all-in-one algorithm that solves for the economic (land-use) and stochastic user equilibrium simultaneously. The second is based on an iterative-shortcut approach, in which the algorithm (i) iterates between the two equilibria and (ii) reduces drastically the computational burden associated with the economic equilibrium. Comparative analysis of the algorithms shows how and why the latter approach can provide from significant to enormous time savings. The analysis involves decomposition and inspection of the separate components of each algorithm and illustrates how time savings evolve as a function of the model’s network resolution (i.e. the number of zones and
The final part of the chapter identifies the conditions under which the two formulations are fully equivalent, i.e. their respective solution algorithms reach the same equilibrium. Detailed design patterns, schematic depictions and pseudocode disengaged from any programming language are provided throughout the chapter.

Chapter 5 investigates the interaction between congestion tolls and distortionary labor taxation in the context of a GELUT model. The proposed model has a clear geographical reference. It is tailored to Randstad conglomeration, a polycentric area comprising the Netherlands' four largest cities (Amsterdam, Rotterdam, the Hague and Utrecht). The region is of considerable economic significance; while it covers only 20% of the country's land area, at least 40% of the population resides there, and half of the national income is generated within its boundaries. Despite being a prosperous region, it has experienced lower productivity growth compared to other regions in the Netherlands and Europe for a series of years (annually 1.7% over the period 1995-2005). It is characterized by large commuting flows between zones and severe congestion during the peak hours. The territorial review by OECD (2007) places heavy congestion and the incoherence of public transport system as the most important drivers of this sluggish growth. Roughly 80% of the traffic jams in the Netherlands in 2005 occurred in Randstad. The model is calibrated to fit a series of stylized facts characterizing the behavior of the average household (expenditures shares, allocation of time, etc.) and the characteristics of the region: the general spatial lay-out and network, the population and employment share of each zone, the average commuting speed of modes, modal split, and the relative land rents, housing prices, wages and floor-to-area ratios.

The effects of various pricing schemes (cordon tolls, Pigouvian taxes and the optimal toll) accompanied by two distinct revenue recycling programs (lump-sum transfers and labor tax cuts) are explored with the above setting. In line with more stylized models, a Pigouvian road tax with revenue returned lump-sum is shown to generate considerable welfare losses. Surprisingly, the Pigouvian toll is also welfare decreasing even when the road tax revenue is used to finance labor tax cuts. By computing the optimal road tax for each road link, it is shown that the latter may not only lie far below its Pigouvian level: it may also be negative in a large part of the network. Therefore, optimizing the tax system from an environmental point of view and returning the revenue optimally may not guarantee welfare gains, since an environmentally optimal tax system may still incorporate strong negative tax interactions (reflected in the distance between optimal taxes and their Pigouvian levels). In fact, these interactions may be strong enough to offset the beneficial effects from the abatement of the externality and the recycling of the revenue. This is confirmed by approximating the Pigouvian, tax-interaction and revenue recycling effect for each link of the network.

Extensive sensitivity analyses are employed to investigate the efficiency properties of the optimal tax scheme, the full network Pigouvian toll and a system of cordon tolls around the largest destination (employment) nodes of the network. The latter policy leaves most of the links untaxed and is found to produce welfare gains, which are nevertheless shown (in a broad parameter range) to be persistently small. This is attributed to the combined impact of the
network configuration (polycentricity) and the magnitude of labor market distortion. As the labor tax gets smaller, the distortion fades away and the welfare gains of the cordon toll system recover, attaining values that are in alignment with those predicted by earlier studies in cordon tolls under polycentricity.

Chapters 2-5 assume the existence of a (federal) government capable of imposing a different level of a road tax on each part of the network and returning the revenue either lump-sum or in the form of a distortionary tax cut. Despite the above setting can potentially lead to an (economically) efficient allocation \( i.e. \) when the correct level of a tax is combined with the correct revenue-recycling program, the implementation of tax reforms introducing urban road charges has met strong resistance. One of the possible drivers behind the low acceptability is the spatial profile of such reforms: decisions over the details of the pricing scheme (type, charges) and the use of the generated revenue are usually centralized, leaving the impression that local welfare is decreased after the introduction of these policies.

A peripheral focus of this dissertation concerns the fiscal externalities of horizontal toll competition. Chapter 6 provides numerical computations for the social cost of a decentralized, \( i.e. \) locally-autonomous, toll system in the GELUT model for Randstad used in Chapter 5. The extension pertains the introduction of local governments that are assigned specific, non-overlapping dominions. These authorities are granted the permission to i) set toll levels in their dominions and ii) redistribute toll revenues according to their objectives. To simplify things while keeping the context in line with that of Chapter 5, the focus is limited on games resulting from a dichotomy of the Randstad area in two large fiscal dominions. Furthermore, the number of instruments controlled by each local government is reduced to just one: a flat kilometer tax on all roads falling into its dominion. Two different local government objectives are considered: local utility maximization and local revenue maximization. In each case, the welfare of the resulting unique Nash equilibrium is compared to the social planner solution, which maximizes the average utility in the two regions. The findings show that toll competition in terms of local welfare may reduce overall utility, but it could favor the large area whose links are used widely (and asymmetrically) by residents of the smaller region. When the objective shifts to revenue maximization, both regions are found to end up with much higher toll levels and are worse-off compared to the no-toll equilibrium. These decentralized settings are juxtaposed against two centralized cases, in which the kilometer tax is set by a social planner aiming to maximize the average utility of the two regions. In these social planner settings, the kilometer tax can (in the latter case) or cannot (in the former) be differentiated across the two regions. In both cases, the resulting allocation is found to be Pareto-preferred to the no-congestion toll equilibrium by both regions. The chapter closes with an investigation of a scheme that combines centrally-set toll levels aiming to maximize average utility combined with local revenue recycling.
1.3. Building blocks

The thesis builds on three wide-and to some extent overlapping-streams of literature: the voluminous literature of congestion pricing, the multiple contributions in the theory of double dividend and the various modeling tools offered by urban economic theory. The amount of existing work in these fields is large enough to render a complete review of each stream impossible. Rather, what is attempted is a careful placement of the dissertation’s contributions vis-à-vis some fundamental work in each field.

1.3.1. Congestion pricing

The optimal taxation of externalities arising from any arbitrary equilibrium allocation of production and consumption activities is a complex topic that has received generous attention through many strains of economic theory. It dates back to the classic work of Pigou (1920) and builds further with the well-known contributions of Coase (1960), Baumol (1972) and Sandmo (1975). The classic works by Walters (1961) and Vickrey (1963) focused entirely on the specific externalities generated by road use.

Figure 1.2. Left: The fundamental flow diagram (lower panel) in a one-link setting. Right: Pigouvian road toll and its effect on congestion. (source: Parry, 2008)

In general, the externalities induced by road use resemble those produced by the consumption or production of any generic “dirty” commodity. Road use gives rise to pollution, congestion, accidents and other negative externalities. A plethora of features and peculiarities, however, establish road pricing as a special form of environmental taxation whose optimal design requires specific methods. Unlike most of the conventional goods that can be easily classified as final or intermediate, roads are facilities whose use is complementary to generic production processes and consumption activities. Production requires labor hours and freight transport, which translates into derived demand for road facilities by commuters and freight transport agents. Consumption and leisure activities require shopping and other types of trips,
thus generating demand for road use. Furthermore, road facilities may be complements or substitutes to each other, depending on the state of their users. Finally, the generation and propagation mechanisms of road externalities (i.e. the way congestion and traffic-related pollution are generated and who they affect across space and time) are unique and constitute an autonomous field of study.

The literature of urban congestion pricing is voluminous and a thorough review can be a cumbersome task. However, a brief recap of the basic flow congestion model and a quick charting of the advances and frontiers that have resulted as responses to its shortcomings are necessary at the outset of this dissertation. To some extent, the recap attempted is based on Parry (2008); the interested reader is referred to any of the multiple comprehensive reviews of the basic models used for pricing urban congestion and the respective insights from each. The absolute departure point contemplates the fundamental relations of traffic congestion shown in Figure 1.2, as they result in an isolate road segment of homogenous capacity (without bottlenecks) used by drivers with homogenous value of time. These are the relations between speed, flow and density.

The upper left quadrant of the left panel in Figure 1.2 displays the relation between average speed and density, with the latter being the inverse of the average distance that separates vehicles in the road segment. The speed-density relation is monotonically decreasing because vehicles slow down as the road gets more dense (and therefore the distance between them decreases) to avoid accidents. When density reaches the maximum capacity of the road, speed falls to zero; when the road is empty, a free-flow speed is feasible. The upper right quadrant shows the non-monotonic speed-flow relation. Because flow is the product of speed and density, every feasible level of it can be achieved in two qualitatively different states: in a normal phase (the upper part of the speed-flow curve), in which flow is realized with relatively high speed and relatively low density, and in a phase of hypercongestion, in which flow is realized with high density and low speed. Flow is maximized for a unique density below the segment's maximum capacity. At that density level there is still space for vehicles to enter the link; but as density increases further (i.e. in hypercongestion) both flow and speed fall. Hypercongestion is not treated explicitly in the context of the dissertation; the interested reader is referred to Small and Verhoef (2007) for a detailed description of its mechanics and relevant ways to model it.

The resulting backward bending flow curve represents the average private cost of road use, reflecting the technological constraints embodied in the speed-density relation. To determine the (partial) unregulated equilibrium level of volume, someone needs to consider preferences as well. The downward sloping curve in the right panel of Figure 1.2 represents the marginal private benefit, (i.e. the marginal willingness to pay) of an extra trip in the road segment under consideration. When flow equals $V^0$, marginal private benefits are equal to average private costs. The equilibrium that occurs is partial, in the sense that there are no distinct relations that relate

---

3 Consider, for instance, the serial network characterized by the sequence of nodes $a$, $b$ and $c$. For a user that has decided to live in node $a$ and work in node $c$, links $a \rightarrow b$ and $b \rightarrow c$ are complements; however, in a different setting where the user is facing a choice between the alternative to live in node $a$ and work in node $b$ versus the alternative to live in node $b$ and work in node $c$, the above two links become substitutes.
the drivers’ demand for using the road link to other consumption or production activities in a structural (i.e. consistent to budget and time constraints) way.

At volume $V^0$, like at any level of traffic, an additional trip increases density and therefore reduces the speed of all other vehicles in the road segment, imposing on them a marginal external cost. Because higher density translates into a larger number of drivers incurring the delay, the marginal external cost (i.e. the disparity between the marginal social cost and the average private cost) increases with flow. The optimal level of traffic ($V^*$) is determined by the equality of the marginal private benefit with the marginal social cost of traffic flow. To bring down traffic to this level, a regulator has to introduce a road toll equal to $\tau^*$, i.e. the textbook Pigouvian remedy to the problem. The careful reader can notice that the right panel of Figure 1.2 is drawn in a way that excludes multiple equilibria. In fact, other (hypercongested) equilibria may be possible, but their dynamic stability is subject of a fierce debate.

The model for flow congestion is the appropriate framework to address the problem in a time-averaged manner. In many situations, however, congestion is not expressed in the form of lower speed due to the general level of vehicle density in a specific link. Rather, it takes the form of density reaching the maximum capacity in a specific road segment. In the bottleneck model, this translates into the formation of a vertical (and thus spaceless) queue in which drivers are served in a first-in-first-out (FIFO) manner. The reader should note that the bottleneck model is not used explicitly in any of the chapters included in the thesis. The use of a bottleneck model in conjunction with a double-dividend setting, however, is one of the future challenges discussed in the conclusions (Chapter 7).

1.3.2. Double dividend

The theory of double-dividend focuses on the occurrence of failures, most often tax-induced distortions in markets that interact with the externalities (environmental, congestion, etc.) to be regulated. When such a market does not operate efficiently, there is a divergence between the marginal willingness to pay for the commodity and the corresponding marginal social benefit from its consumption. This divergence mirrors a negative tax-interaction effect: the efficiency loss in the distorted market caused by a marginal increase of the environmental tax. A negative tax interaction effect could be strong enough to outweigh the Pigouvian effect of an environmental tax, i.e. the welfare benefit from a marginal reduction of externality. In that case, the optimal environmental tax does not only lie below its Pigouvian level: it is essentially negative. This fundamental issue has received only limited attention in the literature of transport economics, although similar questions have spawned several contributions in the literature of environmental economics (see, for example, Bovenberg and De Mooij, 1994; Parry, 1995; Goulder, 1995a; Bovenberg and Goulder, 1996; Parry and Bento, 2000).

An exception is the well-known contribution by Parry and Bento (2001) that highlights the possibility of a welfare-decreasing quasi-first best Pigouvian toll in a single-link setting, in

---

4 See Arnott et al. (1993) for an intuitive description of basic model and the properties of the optimal toll.
which the departure and destination locations are exogenous. With a pre-existing distortionary tax on labor income, the traffic level in a no-toll equilibrium may lie below its optimal level despite the generalized price of a commuting trip falls short of its marginal social cost. Therefore, a policy intervention that introduces road tolls but leaves the labor tax unaffected (for example because the revenues are returned lump-sum) may be expected to produce a decline in social welfare rather than the increase that is hoped for. The conclusion is that, in order to increase welfare, road toll revenues must be used to reduce the distortionary tax.

The double-dividend stream of literature sheds additional light on the interplay between an externality-correcting tax (e.g. a road toll) and a pre-existing distortionary tax (e.g. a tax on labor) by investigating revenue-neutral and marginal swaps between them. Apart from the Pigouvian and tax interaction effects (defined above), these swaps produce a revenue recycling effect, i.e. a welfare gain in the initially distorted (labor) market, due to the reduction of this tax. This reduction is financed by the additional revenue the marginal increase of the externality tax (on road use) generates. If the revenue recycling effect dominates the tax interaction effect, a so-called strong double dividend emerges, and this is a case for setting the externality tax above its Pigouvian level. However, most of the earlier double-dividend literature suggests that this will not occur (Bovenberg and Mooij, 1994; Parry, 1995; Goulder, 1995b; Bovenberg and Goulder, 1996).

1.3.3. Space

The incorporation of congestion- and double-dividend-related settings in explicit spatial contexts is the main source of insights in this dissertation. What follows below is a brief review of the modeling toolbox offered by modern urban economic theory.

1.3.3.1. Monocentric city models

Early contributions in the field of urban economics attempted to explain city size, population density and land rents in a monocentric city (Alonso, 1964; Mills, 1967).\(^5\) In these settings land-use is simply determined by a split between a developed area representing the city and undeveloped agricultural land. In turn, this split is determined by the relative accessibility of each location and the exogenous opportunity cost of converting agricultural land to residential areas. The well-known extension by Muth (1969) allowed for spatial variation in structural density (floor-to-area ratio) without affecting the mechanism through which land-use is determined. Other important extensions include Brueckner (1977) and Wheaton (2004), who incorporate mixed land-use in the model by allowing jobs to be dispersed, rather than concentrated in a central business district.\(^6\)

---

\(^5\) See Fujita (1989) for a rigorous exposition of monocentric city models.

\(^6\) McDonald (2009) calibrated the version of Wheaton (2004) for Los Angeles (using data from 1958); he shows that, with job locations fixed, measures such as peak-hour congestion pricing can significantly reduce the amount of residential redistribution needed to achieve economic efficiency.
Monocentric models have been used to compute the monetized welfare losses from land-use regulations and restrictions. For instance, Bertaud and Brueckner (2005) provide illustrative calculations of the social cost of building height restrictions.\textsuperscript{7} Calibrating the model to fit stylized facts, the authors conclude that this cost may account for 2.2\% of income. Because the model disregards congestion and pollution externalities, as well as the negative externalities associated with high structural density in central areas, the above social cost collapses to the additional transport costs.\textsuperscript{8} The latter costs are generated directly by the building height restriction, as city size and therefore commuting distance increases in the long run (see below on the issue of time and adjustment).

Monocentric city models have also been used to calculate the welfare gains from road price schemes that target congestion and pollution externalities. For instance, McConell and Straszheim (1982) formulate a model that incorporates both pollution and congestion externalities generated from vehicle use. Verhoef (2005) provides numerical simulations for the (relative) efficiency of two archetype road pricing schemes, a cordon toll and a flat kilometer tax. It is shown that, for plausible parameter values, the two policies provide a promising reduction in the number of vehicle kilometers (15-17\%) and capture a large part of the gains from the optimal tax scheme.

The basic version of the monocentric city model is not equipped to explain fragmentation and leapfrog development observed in many urban areas, \textit{i.e.} the phenomenon of land parcels being left undeveloped while others, seemingly less accessible, are built up. One way of accounting for scattered development is to assign an amenity value to public open space, so that individuals may be willing to trade off additional commuting costs associated with living further away from the city center for proximity to open space near their residence (for instance Turner, 2005). Overall, the monocentric city model is a conceptual device that provides important theoretical insights.\textsuperscript{9} However, further policy-oriented insights regarding polycentric areas are difficult to derive using a system of monocentric cities.\textsuperscript{10}

\textbf{1.3.3.2. General equilibrium models of land-use and transport}

Recently, the above difficulties motivated the development of a class of models that partition space into a set of zones across which economic activity (\textit{e.g.} employment, residential, structural density) can display any pattern (see Anas and Kim, 1996; Anas and Xu, 1999; Tscharaktschiew and Hirte, 2010; Anas and Hiramatsu, 2012, 2013). General equilibrium models of land-use and transport (hereafter, GELUT models) offer the ideal platform to perform welfare analysis at the

\textsuperscript{7} See also Arnott and McKinnon (1977) on the same issue.

\textsuperscript{8} The additional negative externalities are associated with reduced view, less exposure to sunlight and reduced access to open spaces. Brueckner \textit{et al.} (1999) develop a theory of endogenous amenities based on income disparities.

\textsuperscript{9} Several studies have tested the empirical validity of the monocentric city model. For an example see Brueckner and Fansler (1983) who find that that income, population and agricultural rent are statistically significant determinants of urban land area.

\textsuperscript{10} The interested reader is referred to the contribution by Lucas and Rossi-Hansberg (2002).
urban scale. These models are closely related to Regional Computable General Equilibrium models (see for instance Bröcker, 1998).

Space partition in GELUT models is crucial for several reasons. First, it facilitates the construction of models in which the urban landscape morphology and the road network, both playing a key role in the generation of congestion and environmental externalities, are represented in a more realistic way. Second, data on economic activity and environmental externalities are organized in a discrete way: job, population and structural densities, emissions and pollution concentration (air quality) are always available at some zonal aggregation. Therefore, these models can be calibrated easier to fit these data. Third, GELUT models do not focus on the city size per se, as the entire surface (e.g. a functional urban area) under consideration is exogenous. Therefore, fundamental concepts such as urban sprawl, which appear to be unidimensional (i.e. represented by population density only), are represented in a much more elaborate way: the GELUT framework is capable of reproducing vacant land, leapfrog development, fragmentation, and other observable variables synthesizing the urban structure.

Furthermore, GELUT models (like all general equilibrium frameworks) possess specific advantages due to the fact that their entire structure represents a consistent, closed economic system, in which no value can be generated or destroyed through an ad hoc mechanism. Furthermore, all prices are set through mechanisms that are consistent with each other and compatible to the rules governing the behavior of the model’s agents. Because prices and quantities are codetermined consistently, GELUT models provide the ideal framework to calculate the total welfare effect of environmentally-related taxes and regulations in double dividend settings emerging from the interactions of these interventions with other pre-existing taxation.

General equilibrium models have been criticized, among others, for their limited capacity to provide forecasts over the variables of interest. The infinite speed of adjustment, inherent in every equilibrium model, seems to be a critical driver behind this weakness. Two other classes of models that possess diametrically opposite properties, i.e. higher forecasting power but lower consistency regarding welfare analysis, are briefly discussed in the next section. In the future extensions of this work (see conclusions) it is argued that many elements from the following section can be injected into a GELUT model to increase its forecasting power.

1.3.3.3. Other approaches

Land-use and transport interaction models (hereafter, LUTI) provide an alternative way to perform forecasts and welfare analyses in spatial settings. These models usually lack micro-foundations, i.e. explicit agent behavior and interactions defined in different markets. Instead,

---

11 In this case the most interesting variables include those describing the urban pattern and the environmental indicators.
12 Here adjustment refers to endogenous variables adjusting to an equilibrium in a given time point. This is to be distinguished from the dynamic adjustment of endogenous variables (e.g. towards a steady-state) across time points, which is a property of any dynamic general equilibrium model.
they model directly market equilibria with *ad hoc* adjustment processes in land-use and the transport system. For instance, some applications introduce pseudo-adjustment processes in land and housing markets by imposing an exogenous time horizon (e.g. five years). This horizon is roughly in alignment with empirical observations, but a LUTI model does not have the intrinsic capacity to explain it or reproduce it. Despite this, the *ad hoc* elements or restrictions introduced to replicate the observed correlation of the variables of interest (e.g. household and job density) allow the urban form to be represented elaborately. For instance, *ad hoc* mechanisms may include exogenous chains of demand (e.g. land demand determines demand for floor space; floor space allocation determines economic activity and so on). For this reason, LUTI models’ forecasting power is usually larger than this of fully micro-founded approaches. Because of these *ad hoc* mechanisms, LUTI models are very diverse with respect to the treatment of time, land, space, transport networks and services. Often, they are characterized by lack of documentation, which sometimes renders their deciphering a challenging task.13

Microsimulation models include dynamic disequilibrium (adjustment-oriented) agent-based (or activity-based) models. These models differ from LUTI approaches because they are micro-founded to a larger extent and incorporate adjustment time explicitly. Furthermore, microsimulation models’ resolution boils down to the parcel level. This facilitates the incorporation of expectations and mechanisms such as bargaining and auctions. These mechanisms (together with the general incentive structure) underlie phenomena such as vacant land and fragmentation that are key to the determination of transport-related externalities at the city level. The most characteristic example of an urban microsimulation model is UrbanSim (Waddell, 1998a; 1998b; 1998c; Waddell et al., 1998; Waddell, 2002), which comes in numerous versions and applications.14 A more elaborate discussion of the various differences between the approaches discussed in this section and GELUT models is offered in the introduction of Chapter 4.

---

13 Some well-known LUTI models include the Integrated Transportation and Land Use Package (Putman, 1983; 1991), MEPLAN (Echenique, 1969) and more recently TIGRIS XL (Zondag and de Jong, 2011; Zondag et al., 2015). The interested reader is referred to Hunt et al. (2005) for an elaborate discussion of these models.

14 For historical reasons someone we also refer to the Detroit prototype of the NBER urban simulation model (Ingram et al., 1972).