Transportation Research Part A 94 (2016) 411–430

Contents lists available at ScienceDirect

Transportation Research Part A

journal homepage: www.elsevier.com/locate/tra

Time-dependent congestion pricing system for large networks: Integrating departure time choice, dynamic traffic assignment and regional travel surveys in the Greater Toronto Area

Aya Aboudina Ph.D. [a,b,*], Hossam Abdelgawad Ph.D., P.Eng. [a,b,*], Baher Abdulhai Ph.D, P.Eng. [a], Khandker Nurul Habib Ph.D, P.Eng [a]

[a] *Department of Civil Engineering, University of Toronto, M5S 1A4, Canada*
[b] *Cairo University, Faculty of Engineering, 12631 Giza, Egypt*

ARTICLE INFO

Article history:
Received 5 January 2016
Received in revised form 29 August 2016
Accepted 4 October 2016

Keywords:
Congestion-pricing
Dynamic traffic assignment simulation
Departure time choice
Large-scale network
Schedule-delay cost

ABSTRACT

Congestion pricing is one of the widely contemplated methods to manage traffic congestion. The purpose of congestion pricing is to manage traffic demand generation and supply allocation by charging fees (i.e., tolling) for the use of certain roads in order to distribute traffic demand more evenly over time and space. This study presents a framework for large-scale variable congestion pricing policy determination and evaluation. The proposed framework integrates departure time choice and route choice models within a regional dynamic traffic assignment (DTA) simulation environment. The framework addresses the impact of tolling on: (1) road traffic congestion (supply side), and (2) travelers' choice dimensions including departure time and route choices (demand side). The framework is applied to a simulation-based case study of tolling a major freeway in Toronto while capturing the regional effects across the Greater Toronto Area (GTA). The models are developed and calibrated using regional household travel survey data that reflect the heterogeneity of travelers' attributes. The DTA model is calibrated using actual traffic counts from the Ontario Ministry of Transportation and the City of Toronto. The case study examined two tolling scenarios: flat and variable tolling. The results indicate that: (1) more benefits are attained from variable pricing, that mirrors temporal congestion patterns, due to departure time rescheduling as opposed to predominantly re-routing only in the case of flat tolling, (2) widespread spatial and temporal re-distributions of traffic demand are observed across the regional network in response to tolling a significant, yet relatively short, expressway serving Downtown Toronto, and (3) flat tolling causes major and counterproductive rerouting patterns during peak hours, which was observed to block access to the tolled facility itself.

1. Introduction and background

As traffic congestion levels soar to unprecedented levels in dense urban areas, and governments are challenged to meet the demand for transportation and mobility; congestion pricing is becoming one of the widely contemplated methods to combat congestion (Washbrook et al., 2006).

* Corresponding authors at: Department of Civil Engineering, University of Toronto, M5S 1A4, Canada.
E-mail addresses: aya.aboudina@mail.utoronto.ca (A. Aboudina), hossam.abdelgawad@alumni.utoronto.ca (H. Abdelgawad), baher.abdulhai@utoronto.ca (B. Abdulhai), khandker.nurulhabib@utoronto.ca (K.N. Habib).

http://dx.doi.org/10.1016/j.tra.2016.10.005
0965-8564/© 2016 Elsevier Ltd. All rights reserved.

The "tragedy of the commons" concept has been established longer than a century ago as mentioned by Hardin (1968). A famous example is when herders are given free access to open grassland for their cows to graze, cows tend to overgraze and deplete their source of sustenance to the detriment of everyone. The parallel to the tragedy of the commons in traffic could not be more direct. While transportation authority and society at large would like to "optimize" travel and minimize overall cost of travel, travelers act very differently. Travelers act independently and rationally, based on their self-interest, i.e., minimizing their direct cost while not paying attention to the societal cost and the detriment to others. Consequently, the purpose of congestion pricing is to manage traffic demand generation and supply allocation to ensure a more rational use of roadway networks. This is accomplished by charging fees for the use of certain roads in order to reduce traffic demand or distribute it more evenly over time (away from the peak period) and space (away from overly congested facilities).

Numerous studies have investigated the potential of congestion pricing schemes in reducing the vehicular demand subject to travel and behavioral characteristics. While fully enumerating all congestion pricing studies is beyond the scope of this paper, the following section briefly reviews what is highly relevant to our scope:

In a study conducted by Washbrook et al. (2006) at University Drive (Burnaby, British Columbia), single-occupant vehicle (SOV) commuters completed a discrete choice experiment in which they chose between driving alone, carpooling or taking a hypothetical express bus service when choices varied in terms of time and cost attributes. The results of this study indicate that a potential increase in drive alone costs brings greater reductions in SOV demand than an increase in SOV travel time or improvements in the times and costs of alternatives (i.e., carpooling and bus express service). Another study conducted by Duranton and Turner (2011) at the University of Toronto assessed the potential of congestion pricing against capacity expansions and extensions to public transit as policies to combat traffic congestion. The study concludes that vehicle kilometers traveled (VKT) is quite responsive to price as opposed to transit or capacity expansions. Moreover, Sasic and Habib (2013) showed that the recommended strategy to lighten peak period demand while maintaining transit mode share in the Greater Toronto and Hamilton Area (GTHA) requires imposing a toll (around $1) for all auto trips in addition to a 30% flat peak transit fare hike. Furthermore, their results suggest that such a pricing policy would have a larger effect on shifting travel demand over time than any other policies not including a road toll.

Tolling studies in the literature range from applying a flat or simple pricing structure, e.g., Lightstone (2011) and Sasic and Habib (2013), on a small or sometimes hypothetical network, e.g., Gragera and Sauri (2012) and Guo and Yang (2012), to a network-wide pricing scheme, e.g., Verhoef (2002) and Morgul and Ozbay (2010). Finkleman et al. (2011) studied the acceptability and impacts of HOT lanes in the GTA through a stated preference survey of more than 250 drivers, under various trip conditions and for various traveler characteristics. Other efforts, e.g., Nikolic et al. (2015), studied dynamic tolling of HOV lanes on specific corridors in a micro-simulation environment; in which the network-effect and routing options affected by tolling were not considered. Mahmassani et al. (2005), Lu et al. (2006, 2008), Lu and Mahmassani (2008), and Lu and Mahmassani (2011) developed a multi-criterion route and departure time user equilibrium model for use with dynamic traffic assignment applications to networks with variable toll pricing. The model considers heterogeneous users with different values of time, values of (early or late) schedule delay, and preferred arrival time (PAT) in their choice of departure times and paths characterized by travel time, out-of-pocket cost, and schedule delay cost. Furthermore, the model was applied to an actual relatively small network (180 nodes, 445 links, and 13 zones) through a simulation-based algorithm. The authors, however, acknowledge that their algorithm suffers from computational limitations in a large network setting.

All these studies contribute considerably to the state-of-the-art and state-of-the-practice in congestion pricing; nevertheless, the literature has some or a combination of the following limitations:

- scarce case studies on large-scale realistic regional networks/models (as opposed to hypothetical small networks);
- hypothetical tolling scenarios that lack methodological/practical basis; and
- disregard of travelers' individual responses to pricing (e.g., choice of departure time, choice of mode, and choice of route). Additionally, the limited number of studies that considered some of those responses ignored the drivers' personal and socioeconomic attributes affecting the decision made in response to pricing, perhaps due to lack of large scale travel surveys.

In light of the aforementioned gaps, this study is motivated to develop a robust framework for the methodological derivation and evaluation of variable congestion pricing policies to manage peak period travel demand, while explicitly capturing departure time and route choices in a large-scale dynamic traffic simulation environment. The study, through rich travel survey data available in the Greater Toronto Area (GTA), considers the drivers' heterogeneity in their values of (early or late) schedule delay and desired arrival times. Moreover, drivers' personal and socio-economic attributes – affecting the choice of departure times – are taken into account besides the trip-related travel time, out-of-pocket cost, and schedule delay cost. The DTA model is calibrated using actual traffic counts from the Ontario Ministry of Transportation and the City of Toronto. The framework addresses the impact of tolling on: (1) road traffic congestion (supply side), and (2) travelers' choice dimensions including departure time and route choices (demand side). Mode choice responses to tolling are beyond the focus of this study and will be considered in future work. The framework is applied to a simulation-based case study of tolling a major freeway in Toronto (the Gardiner Expressway) while capturing the regional effects across the GTA, in Ontario, Canada. The case study examined two tolling scenarios: flat and variable tolling.

2. Modeling framework

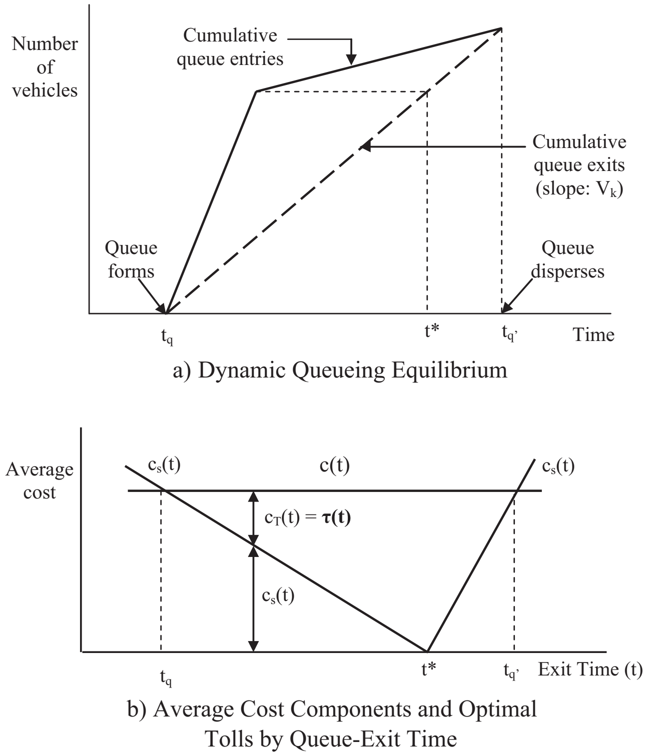
This section presents a framework for evaluation of variable congestion pricing policies as a method of spatial and temporal traffic management. The proposed framework integrates departure time choice and route choice within a large-scale dynamic traffic simulation environment.

The framework is based on four key pillars: (1) the bottleneck model for dynamic congestion pricing which is the theoretical basis of the variable tolling structure adapted in this study; (2) an econometric (behavioral) model of departure time choice that is built and calibrated using regional house-hold travel survey data which capture the heterogeneity of travelers' personal and socioeconomic attributes; (3) a dynamic traffic assignment simulation platform that is used to assess the impact of various pricing options on routing and congestion patterns; and (4) finally, the integration and implementation of the above into a single framework that incorporates variable tolling while looping between the departure time choice layer and the DTA layer until departure time choices and route choices reach equilibrium, and the impact of tolling on system performance is assessed. The above key pillars of the approach are described next.

2.1. Theoretical basis: the bottleneck model for dynamic congestion pricing

Dynamic models consider that congestion peaks over time then subsides. Therefore, there is a congestion delay component that peaks with congestion that the travelers experience. Dynamic models assume that travelers have a desired arrival time; deviations from which imply early or late schedule delays. Travelers who must arrive on time during the peak periods encounter the longest delay; i.e., there is a trade-off between avoiding congestion delay and arriving too early or too late.

The basic Bottleneck Model is the most widely used conceptual model of dynamic congestion pricing (Small and Verhoef, 2007). It involves a single "bottleneck" and assumes that travelers are homogeneous and have the same desired arrival time, t∗. Moreover, the model assumes that for arrival rates of vehicles not exceeding the bottleneck capacity and in absence of a queue, the bottleneck's outflow is equal to its inflow and as a result no congestion (delay) occurs. When a queue exists, vehicles exit the queue at a constant rate, which is the same as the bottleneck capacity V_k. Fig. 1a illustrates the un-priced equilibrium condition of this model (i.e., equilibrium in the absence of tolling) and Fig. 1b shows the two components of the total cost $c(t)$ in the un-priced equilibrium condition, namely, travel delay cost $c_T(t)$ and schedule delay cost $c_s(t)$ (early and late arrival costs). As noticed in the figure, the schedule delay cost is assumed to be a piecewise linear function in this model. The summation of the two costs (i.e., the total cost) is constant in the un-priced equilibrium, as illustrated in the figure.



a) Dynamic Queueing Equilibrium

b) Average Cost Components and Optimal Tolls by Queue-Exit Time

Fig. 1. Equilibrium in the basic bottleneck model (Small and Verhoef, 2007).

Note that the total number of travelers that enters the system ultimately exits the system after being queued for a while. The optimal toll in this case attempts to ‘‘flatten” the peak in order to spread the demand evenly over the same time period. In this case, the price is set such that the inflow equals road capacity, which in turn equals the outflow. The optimal tolled-equilibrium exhibits the same pattern of exits from the bottleneck as the un-priced equilibrium, but it has a different pattern of entries. Pricing affects the pattern of entries with a triangular toll schedule, with two linear segments, that replicates the pattern of travel delay costs in the un-priced equilibrium. This toll is shown in Fig. 1b as $\tau(t)$. It results in the same pattern of schedule-delay cost as in the un-priced equilibrium, but it produces zero travel delay cost (i.e., no travel delays exist in the optimal case). Instead of queueing-delay, travelers trade off the amount of toll to be paid versus schedule delay such that a traveler who arrives right on time t$*$ pays the highest toll. The resulting tolled-equilibrium queue-entry pattern therefore satisfies an entry rate equal to the capacity V_k, i.e., queue entry rate equals the queue exit rate in Fig. 1a.

The toll structure introduced in the current study is motivated by the above theoretical bottleneck pricing theory; where key benefits arise from rescheduling of departure times from the trip origin (temporal distribution). The optimum toll $\tau(t)$ in the Bottleneck Model varies continuously over time, as illustrated in Fig. 1b. It is however impractical to change the toll every second as suggested by the model. ‘Step tolls’ are the closest approximation of this ideal situation in practice; different toll values are set at discrete time intervals and the toll is constant within each interval.

Although the bottleneck model provides the core concept, it is limited to the case of a single bottleneck, where the departure time choice is the only choice travelers have to respond to pricing. In large urban networks, there is a myriad of origin-destination pairs, trip lengths, travelers’ schedules, routing options and travel behavior that vary across the population. Therefore, our pricing framework extends the conceptual triangular pricing structure suggested by the bottleneck model to the more complex and general case of a large urban network. Our framework uses econometric departure time choice modeling based on regional travel surveys in conjunction with dynamic traffic assignment to capture departure time choice and routing dynamics in response to tolling.

2.2. The econometric model for departure time choice

In order to capture users’ individual responses to pricing, this study uses a discrete-choice module to capture the departure time choice dynamics in response to tolling. The discrete choice module considers drivers’ socio-economic attributes and the network level-of-service attributes. This study extends a discrete choice model developed by Sasic and Habib (2013) at the University of Toronto that describes departure time choice in the Greater Toronto and Hamilton Area (GTHA). The model is extended to incorporate a schedule delay cost component for realistic modeling of morning peak travel behavior. The model was developed and calibrated in the original study and retrofitted in this study using the Transportation Tomorrow travel Surveys (TTS) of 2006 and (the latest) 2011 respectively (DMG, 2015). The developed departure time choice model belongs to the Generalized Extreme Value (GEV) class of models for discrete choice applications that make use of random utility maximization theory, where each agent (traveler) is assumed to choose an alternative that maximizes its random utility. The random utility for any alternative is defined as a systematic and a random component (where the joint density of all random components is distributed according to the extreme value distribution).

Two types of scale parameters are introduced in this model. These are root scale parameter and nest scale parameter of a particular choice set. Moreover, the modeling framework uses a scale parameterization approach. This approach captures heteroskedasticity in departure time choices. It also captures heterogeneity in users’ departure time choice responses to variations in trip-related attributes (e.g., travel time and cost) at each choice interval. Further details on the choice set structure, the utility function variables, the model parameters’ adjustment process, and the model calibration results are presented in the simulation-based case study on the GTA region later in this paper.

2.3. The mesoscopic dynamic traffic assignment model

Congestion pricing is typically sought in congested large urban areas, where congestion spreads over space for long peak hours. Therefore, to dynamically control traffic in large-scale congested networks, three systems are needed concurrently: (1) a prescriptive decision-setting/control tool (e.g., a demand or supply control policy such as congestion pricing or ramp metering), (2) a descriptive econometric departure time choice model as discussed above, and (3) a descriptive dynamic traffic assignment model that captures route choice dynamics and the evolution of traffic congestion resulting from travelers seeking the least-generalized-cost routes to their destinations. A large scale dynamic traffic assignment simulation model is, hence, required for practical congestion pricing policy derivation and application; a model that can realistically capture the route choice dynamics network-wide (over time and space) resulting from fixed or variable tolls along key corridors. It is noteworthy that these tolls would in turn affect travelers’ departure time choice; therefore the need to integrate both the route and departure time choices within the same framework.

For that purpose, and to capture system-wide effects of tolling in large urban areas, a mesoscopic dynamic traffic assignment (DTA) model is used in this study. In general, mesoscopic models simulate the movement of vehicles in the transportation network in groups according to the fundamental diagrams of traffic theory. These models offer a compromise between microscopic and macroscopic models; unlike macroscopic models, they model individual vehicles, and unlike microscopic models, they are less computationally demanding and hence are more suited for modeling large networks (Abdelgawad and Abdulhai, 2009).

More details related to the demand patterns, which are inputs to the mesoscopic simulation, the key traffic assignment control parameters, and the simulation calibration results will be discussed within a case study on the GTA simulation network in the following sections.

2.4. The integrated variable congestion pricing framework

Fig. 2 shows the integrated variable congestion pricing framework. The ultimate goal of this framework is to provide a tool for variable congestion pricing policy derivation and evaluation, while taking into account the route choice and departure time choice dimensions in large-scale regional networks. The system works in the following order:

- ***Input Data:*** The system first takes as input the network topology, anticipated demand and user demographics to form a hybrid dynamic traffic assignment and travel behavior model. Moreover, a nonlinear version of the price structure of the bottleneck model (i.e., step tolls rather than a continuous toll structure) is to be provided as input to the system for the facility of interest in the network (e.g., link, road, corridor or area). The nonlinear triangular price structure rises from zero to a maximum then falls back to zero when congestion diminishes, as shown in the "Dynamic Toll Schedule" module in Fig. 2. It is important to mention that the framework is intended to test different tolling scenarios; e.g., HOT lanes, congested highway sections, and cordon tolls. Toll values can be discretized per time (up to a toll value per-minute) and space (up to a toll value per-link).
- ***Run DTA Simulation Model:*** The DTA simulation model takes the network topology, toll structure (if any, which is added to travel costs), anticipated demand; and performs iterative dynamic user-equilibrium traffic assignment; resulting in OD travel times, updated network conditions, and routing options given the inputs received.
- ***Apply Discrete Choice Model:*** The discrete choice model takes as input the toll structure, the heterogeneous personal and socio-economic attributes of impacted drivers (for which the discrete-choice model is applied), and the average OD travel times and costs calculated across the network from the most recent DTA simulation run. The output of the discrete-choice model, in turn, represents the *new* temporal demand patterns (with modified start times) due to tolling.
- ***Integrate Departure Time and Route Choices:*** The equilibrium in drivers' behavioral responses to variable pricing policies is sought by iteratively and sequentially simulating the changes in route choice and departure time choice in response to tolling through the DTA simulator and the discrete-choice model, respectively. At the end of each iteration, the discrete-choice module estimates the impact of the input toll schedule given the most recent network conditions (travel times and costs) on travelers' individual departure time choices. The updated choices are then fed back into the dynamic traffic assignment simulator, which, in turn, produces the new network conditions and so on until certain convergence criterion is met.

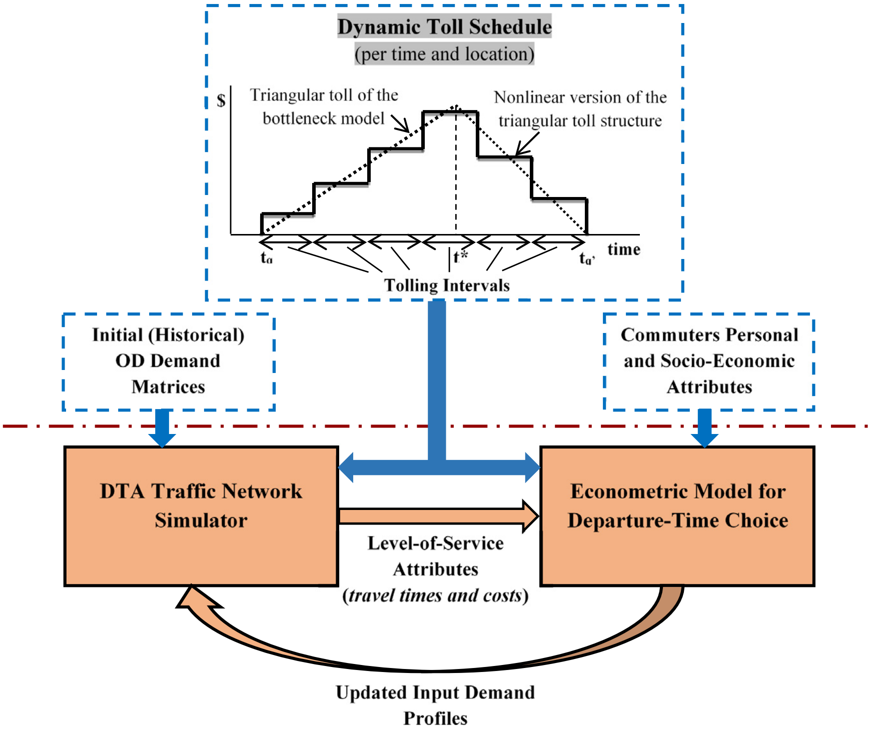

Dynamic Toll Schedule
(per time and location)
$
Triangular toll of the bottleneck model
Nonlinear version of the triangular toll structure
t*
tα
tα'
time
Tolling Intervals
Initial (Historical) OD Demand Matrices
Commuters Personal and Socio-Economic Attributes
DTA Traffic Network Simulator
Econometric Model for Departure-Time Choice
Level-of-Service Attributes
(travel times and costs)
Updated Input Demand Profiles

Fig. 2. Framework for variable congestion pricing evaluation.

As described in this section, two levels of equilibrium are sought in the implementation of the proposed framework. The first one (*inner* iterative loop) is the dynamic user equilibrium within the traffic assignment simulation model. The convergence criteria used for traffic assignment is referred to as the Relative GAP (RG); it is a measure of how close the current assignment solution is to the User Equilibrium (UE) network assignment (Chiu et al., 2008). The traffic assignment iterations terminate when the RG drops below certain pre-specified convergence threshold or when a pre-specified maximum number of iterations is reached. The second one (*outer* iterative loop) is the equilibrium in the departure time choice model output in response to changes in the traffic network travel times and costs after tolling, as shown in Fig. 2. The outer loop terminates when travelers cease to change their departure time interval, i.e., when the maximum (absolute) relative difference in the total number of vehicles at any departure time interval drops below a pre-specified convergence threshold, as clarified in the following formula:

$$\max_i abs\left(\frac{C_i - P_i}{P_i}\right) \leqslant \propto, \quad i = 1, 2, \ldots, n$$

where n is the number of departure time choice intervals, C_i is the number of drivers who chose to depart during interval i in the current iteration, P_i is the number of drivers who chose to depart during interval i in the previous iteration, and $\propto$ is the convergence threshold that represents the maximum acceptable error in the departure time choice model output (compared to the output of the previous iteration). This criterion could be difficult to satisfy if very few drivers depart during some intervals. However, this is not usually the case in large-scale studies featuring a large population of drivers and a time period that encompasses widely-traveled times.

The above is further detailed via an application of the proposed framework to a simulation-based case study of the GTA, as discussed in the next section.

3. Model calibration and application: case study on the GTA

Traffic congestion is reaching a crisis level in larger cities and metropolises in Canada and worldwide. The Greater Toronto and Hamilton Area (GTHA) in Ontario, Canada, is a vivid example in terms of widespread congestion on all modes, particularly roads. Toronto is one of the ‘top ten’ most congested North American cities (TomTom International BV, 2014). In 2006, the annual cost of congestion to commuters in the GTA was $3.3 billion. Looking ahead to 2031, this cost is expected to rise to $7.8 billion (GTTA, 2008).

Different levels of government in Canada are contemplating congestion pricing options to alleviate traffic congestion problems. The Ministry of Transportation Ontario (MTO) is actively evaluating High Occupancy Toll (HOT) lane options, (Nikolic et al., 2015). In 2013, Metrolinx (an agency of the government of Ontario) released its investment strategy in which it recommended the implementation of HOT lanes as a potential source of fund for transit expansion in the region.

Together these factors strengthen the need to analyze, test, and deploy various traffic control policies (such as the one proposed herein) in order to tackle the alarming congestion problems in the GTA region. This region involves widespread activities, heterogeneous travel behavior, different values of time among diverse drivers, multiple routing options, as well as many satellite cities; which makes it an ideal case study on which to test the proposed framework for variable congestion pricing. It should be emphasized that the GTHA region contains the GTA and that the GTA comprises a large majority of the GTHA. Therefore, the use of data (e.g., the TTS survey datasets) and a model (e.g., the departure time choice model) for the GTHA should be suitable for application to the GTA.

This section presents the details of the first implementation of the proposed framework to a key corridor (the Gardiner Expressway) in the GTA and the resulting impact on the full region. The section starts with a brief description of the data used in this study, followed by a detailed explanation of the modeling process of the GTA network (in terms of network geometry, travel demand, and key simulation parameters calibration and validation process). The last, and main, part of this section corresponds to the departure time discrete choice model; its formulation, variables, the parameters adjustment process, and the empirical model validation results.

3.1. Demand data sources

The travel demand related data used in this study (as input to the traffic network simulation model and the discrete-choice model, as shown in Fig. 2) is extracted from the latest 2011 Transportation Tomorrow Survey (TTS), (DMG, 2015). TTS is a household based travel demand survey that is conducted in the Greater Toronto and Hamilton Area (GTHA) every five years. The survey provides detailed information on trips made on a typical weekday by all individuals in the selected households. Information collected in the survey includes household related attributes (e.g., the number of people and the number of vehicles available for personal use), person related attributes (e.g., their age, driver licence availability, and work/school location), and trip related attributes (e.g., origin, destination, purpose, start time, and type of transportation used). Five percent of the GTHA households are contacted by telephone and all trips made by residents eleven years of age or older on a specific weekday are recorded. Expansion factors are used to expand the collected data to represent the total population of the survey area in the year of the survey. The expansion factors are determined based on geographical areas and verified based on Canada Census data that are used as the control total for calculating the expansion factors.

3.2. GTA dynamic traffic assignment simulation model

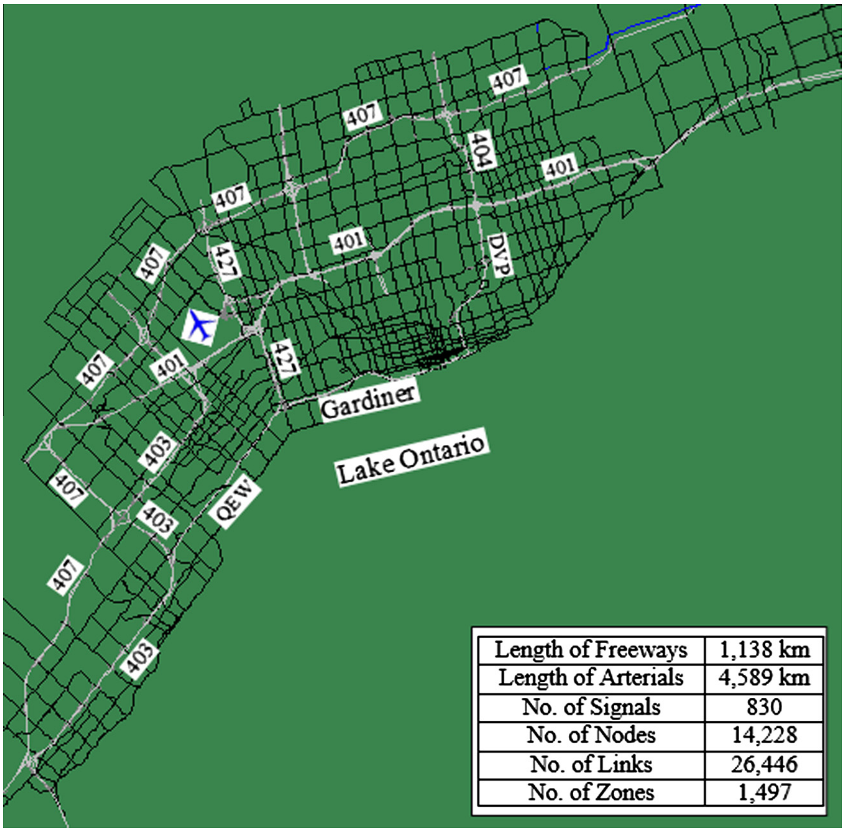
3.2.1. GTA simulation model specification - network geometry

The simulation model of the GTA network, used in this study, incorporates all highways, major arterials, on-and-off ramps, as well as traffic signal information at the major signalized intersections in the GTA. The DTA platform used is DynusT (Chiu et al., 2008). As shown in Fig. 3, the network was built to cover all major arterials and freeways within the GTA to capture all routing options in response to tolling scenarios. The network covers 1497 traffic analysis zones (TAZs), as defined by the most recent zoning system (DMG, 2015), 1138 km of freeways, 4589 km of arterials, and 830 traffic signals.

3.2.2. GTA simulation model specification - travel demand

As mentioned in the data description section, the time-dependent OD matrices used as input for the GTA simulation model were extracted from the 2011 TTS data survey, after applying the reported expansion factors to cover the total demand in the survey area. The demand extracted included all auto modes (SOV, HOV, taxi passenger, and motorcycles), and morning trips from 6:00 to 10:30 am generated every 15 min. The majority of home-based work trips in the GTA - on which we focus in this study - were observed to occur during this time interval (Sasic and Habib, 2013). Additionally, the background demand (i.e., trips that pass through the GTA network but start and/or end outside of it) was added - at the proper time intervals - to the GTA demand. A demand shifting procedure was conducted to capture the time elapsed until those background trips reached the boundaries of GTA network, then added to the demand from those boundary zones. Moreover, loop detector counts across multiple highways in the GTA were used to refine the OD matrix. Fig. 4 shows the total demand at each 15 min time interval during the am period, before and after adding the background demand.

Despite their unquestionable impact on traffic conditions, truck demand and transit on-street units (e.g., buses and street cars) are not included in the input demand considered in this study. This is primarily due to the absence of their relevant data in the TTS survey from which the input demand was extracted. Additionally, the DTA simulation software used does not include transit assignment model to simulate/assign transit units in the network. However, the absence of trucks and transit units in the model was compensated for by adjusting the demand of some OD's during the model calibration process. This was applied to OD's feeding corridors where loop detector readings exceeded simulated traffic (probably due to shortage in the input demand). It is also important to emphasize that this study focuses on the am peak period of auto traffic during which truck demand is relatively low (Roorda et al., 2010).



Length of Freeways	1,138 km
Length of Arterials	4,589 km
No. of Signals	830
No. of Nodes	14,228
No. of Links	26,446
No. of Zones	1,497

Fig. 3. Snapshot of the GTA DynusT simulation model.

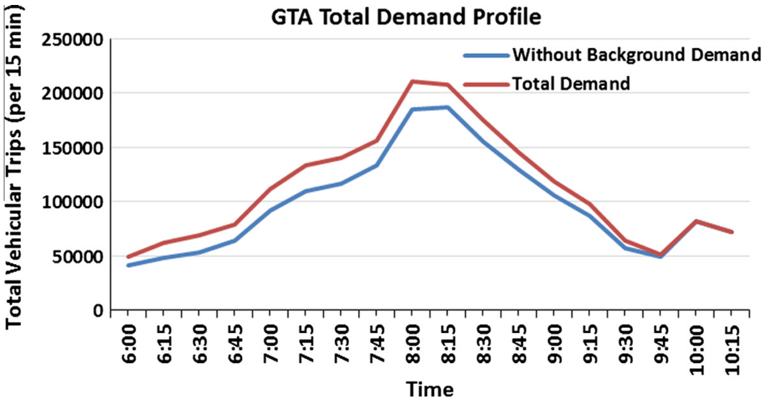


Fig. 4. GTA total demand profile (Kamel et al., 2015).

In this study, two modes of releasing traffic demand (i.e., generating vehicles) into the simulation model are utilized: (1) typical OD demand matrix, and (2) vehicle-by-vehicle input with detailed start-time and path information. Initially, the vehicles in the network are simulated from the input time-dependent OD matrices, extracted from travel survey data, over the simulation horizon. After a DynusT run from OD demand matrix mode converges to UE, information on the simulated vehicles (e.g., vehicle ID, start-time, and origin and destination zones) is listed in output files. DynusT can use this vehicle-by-vehicle output information as input for alternative scenarios.

The advantage of using the second input mode is to have an apples-to-apples comparison between the base-case scenario and a variety of other scenarios with the same input vehicles. In other words, the vehicle-by-vehicle input mode removes the possibility of variability in simulation results stemming from different vehicle input. As described, the second mode is based on the completion and output of a DynusT run from the OD demand matrix mode. Accordingly, in order to apply the departure time choice model to capture the impact of variable tolling on the start-times of specific vehicles in the network, the base-case network is re-simulated (after a complete run with OD demand mode) with the imposed tolling scenarios using the detailed vehicle-by-vehicle input mode. The total demand estimated from the TTS for the 4.5 h period is 1.8 million vehicles. Although the core demand was estimated for 4.5 h, the simulation was conducted for 6 h period to capture the shoulders of rush hour.

3.2.3. GTA network calibration

The parameters adjusted in the calibration process of the GTA DynusT simulation network include the traffic flow model parameters, the freeway bias factor (that controls traveler's perception bias towards freeway travel time), and the demand values among certain OD pairs at specific time intervals. The simulated hourly traffic volumes at 177 locations over highways 400, 401, 403, 404, QEW, the Gardiner expressway, and the Lakeshore Blvd were compared against real data collected from loop detectors (Kamel et al., 2015).

The GEH statistic (named after Geoffrey E. Havers who invented it in the 1970s), widely used in calibrating traffic simulation models, was used by Kamel et al. (2015) as an evaluation criterion for the simulated volumes in the GTA model. Its value reflects the difference between the observed and the simulated volumes. The GEH statistic is computed as follows:

$$GEH = \sqrt{\frac{2(V-C)^2}{(\mathrm{V}+\mathrm{C})}}$$

where V is the model simulated hourly volume at a location and C is the actual hourly count at the same location. The average GEH of the whole model is 9.75, as shown in Fig. 5. This value of 9.75 falls in the cautiously acceptable range of the calibration targets developed by Wisconsin DOT, as summarized in Table 1.

The best attained GEH of 9.75 is interpreted and accepted with two factors in mind: (1) the sheer size of the regional network and (2) the large number of loop detector stations and the inevitable variability in the quality of the loop detector data used in the calibration process.

The traffic assignment software used in this study allows for only single-user class with single value of time (VOT). According to Lu et al. (2006), considering multiclass traffic assignment (i.e., considering heterogeneity in VOT in route choice) is generally challenging in large-scale simulation models due to computational-efficiency and solution storing-space issues. The VOT used in the simulation model for the GTA is $15/h, according to Habib and Weiss (2014).

As mentioned in Section 2.4, the convergence criteria used for the traffic assignment model is referred to as the Relative GAP (RG). The RG measures the relative difference between the experienced and the shortest total path travel times for each OD and simulation time interval (typically 5 min) combination. It reflects how close the assignment solution (at each iteration) is to the target User Equilibrium (UE) network assignment. The detailed formulas of the RG are provided in (Chiu et al., 2008). Fig. 6 illustrates the evolution of the RG over a 20 iterations run of the GTA network simulation model.

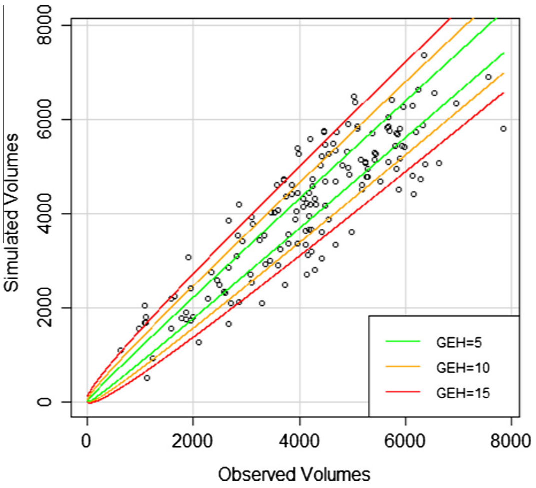


Fig. 5. Scatter plot of the observed and simulated hourly volumes (Kamel et al., 2015).

Table 1
GEH calibration targets (www.wisdot.info/microsimulation).

GEH less than 5	Acceptable fit, probably OK
GEH between 5 and 10	Caution: possible model error or bad data
GEH greater than 10	Warning: high probability of modeling error or bad data

The GTA network contains thousands of links and nodes, and millions of vehicles, making it one of the largest mesoscopic dynamic traffic simulation models built in the region. A number of challenges were faced while building, calibrating, running, and processing the output of the GTA simulation platform, as will be summarized in this section. This is mostly due to the size of the network, the volume of data, the variety of data sources, the veracity and value of the data used to build and calibrate the model, and the volume of the output data produced during the simulation. On an i7 Machine with 16 GB of RAM, the DTA run-time of the GTA simulation model until convergence (using 20 iterations, as illustrated in Fig. 6) is around 7.5 h. As clarified in the description of the integrated variable congestion pricing framework, the GTA simulation model is run several times in sequence with the departure time choice model until convergence is reached.

3.3. Econometric model for departure time choice in the GTA

Several approaches can be followed to simulate drivers' departure time along with route changes within a traffic simulation environment. The most simple, yet naïve and non-realistic, approach is to induce random perturbation of trips start-

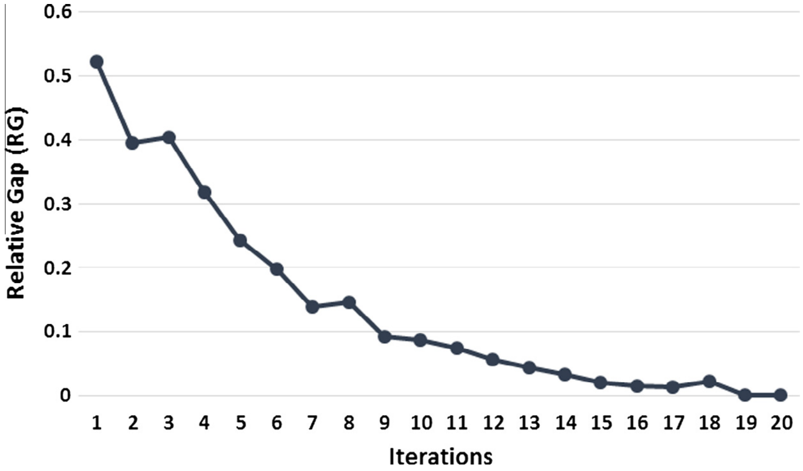


Fig. 6. GTA DTA simulation model convergence.

times throughout the simulation, based on certain pre-set probability, as in (Balmer et al., 2008). This approach is easy to implement and not computationally demanding. However, the stochastic mutations might bring unrealistic start-times (e.g., work trip starting at 2:00 am). Additionally, the changes in start-time are not directly affected by policies introduced (like time-dependent congestion pricing).

Another approach, followed in (Lu et al., 2006), involves joint departure time and route-choice algorithms - implemented iteratively until equilibrium - based on a set of trip attributes that include travel time, out-of-pocket cost, and schedule-delay cost. The model developed in this study was applied to an actual relatively small network (180 nodes, 445 links, and 13 zones) through a simulation-based algorithm. This approach has the advantage of realistically Modeling the joint nature of both departure time and route choices within a simulation environment. However, it cannot be handled in a large network setting within the limits of practical computational capabilities. Additionally, it does not consider the impact of driver related attributes (e.g., personal and socio-economic characteristics) on the choice making process.

A third approach, followed in this study, is through integrating an econometric behavioral departure time choice model (that considers both trip and driver attributes) into a large-scale traffic assignment simulation model. It provides a computationally tractable tool to estimate departure time and route choice responses to traffic management policies that affect travel times and costs, in a large-scale setting. The problem with this approach is the underlying assumption that departure time and route choices are made sequentially (rather than jointly). However, this is compensated for in this study by iterating and feeding back between departure time and route choices until both choices reach equilibrium.

This study extends a departure time choice model in the GTHA, developed by Sasic and Habib (2013), to incorporate a schedule delay cost component for realistic modeling of morning peak travel behavior. The developed model is a Heteroskedastic Generalized Extreme Value (Het-GEV) model that further enhances the Choice Set Generation Logit (GenL) captivity component developed by Swait (2001). The Het-GEV model explicitly captures the correlation between adjacent choice alternatives (by allowing choice alternatives to appear in multiple clusters) while the GenL form captures the captivity of decision makers to specific choice alternatives due to schedule constraints.

This section describes the details of the econometric model used in this study to model the departure time choice in the proposed variable congestion pricing framework. The section starts with an introduction to the model choice set formulation, followed by a discussion of the original variables used in the utility functions as well as the extensions and assumptions done to incorporate schedule delay and toll cost components in the model variables. Lastly, the re-calibration process of some model parameters and the final validation results are presented.

3.3.1. Model formulation

The datasets from the 2006 TTS survey (DMG, 2015) were used for the empirical model of departure time choices of home-based commuting (home to work or school) trips in the Greater Toronto and Hamilton Area (GTHA) (Sasic and Habib, 2013). The datasets from the latest 2011 TTS survey are used in this study to retrofit the 2006 model for 2011 conditions, as will be explained later in the paper. In this model, departure time is represented as nine half-hour intervals that span the morning peak, when the majority of home-based work trips occur. For compatibility with the departure time choice model, the variable-tolling intervals used (for step tolls) are the same nine intervals used in the model. The choice framework is shown in Fig. 7. This framework resembles the decision making process where an individual chooses his/her departure time within a specific range (portion) of the day. In this framework, the probability that an individual chooses to depart from home to work during some interval is defined as the weighted sum of the probability of choosing this time interval over the one preceding it and the probability of choosing this time interval over the one following it. Moreover, the probability of

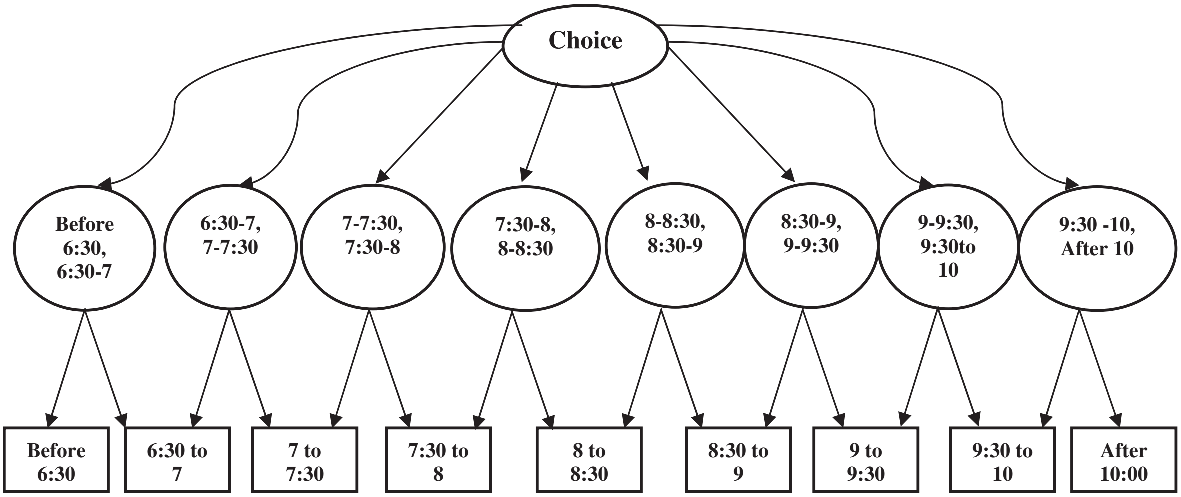


Fig. 7. Departure time choice framework in the Het-GEV model (Sasic and Habib, 2013).

choosing some departure time interval is affected by the explanatory variables, as well as the root and nest scale parameters (μ_R and μ_c) that explain additional choice heterogeneity. In particular, the probability of choosing certain alternative j, P_j, is calculated as follows:

$$P_j = \sum_{c=1}^{8}((P_j|c) * Q_c), \quad j = 1, 2, \ldots 9,$$

where $P_j|c$ is the conditional probability of alternative j in the choice set c and Q_c is the probability of the choice set c. Q_c is calculated based on the following formula:

$$Q_c = \frac{\exp(\mu_R I_c)}{\sum_{c=1}^{8}\exp(\mu_R I_c)},$$

where I_c is the inclusive value of a particular choice set c. I_c is calculated as follows:

$$I_c = \frac{1}{\mu_c}\ln\left(\sum_{k=1}^{K}\exp(\mu_c V_k)\right),$$

where K is the total number of alternatives in the choice set c. The conditional probability of any alternative j in a particular choice set c is calculated according to the following formula:

$$P_j|c = \frac{\exp(\mu_c V_j)}{\sum_{k=1}^{K}\exp(\mu_c V_k)}.$$

3.3.2. Model variables

The model has two types of explanatory variables in the systematic utility functions and the root and nest scale parameters: commuters' personal and socio-economic attributes such as work duration, occupation category (general office, manufacturing, or professional), gender, job status (full-or-part- time), and age category; and transportation level-of-service (LOS) attributes corresponding to alternative departure time segments such as travel time, travel distance, and travel cost. It should be mentioned that the departure time choice model is only applied to commuting trips (representing the majority of morning trips) for which the original model was estimated. Hence, route choice is assumed to be the only choice non-commuting trips have to respond to pricing; it is modeled through the DTA simulator. We believe that this assumption should not create much bias in the overall measured effect because only a fraction of travelers typically respond to a toll or other shock by changing departure time. A lack of response from non-commuters could be compensated by a more-than-proportional response from commuters so that the overall response is similar to a case in which all travelers are flexible.

Preparing the commuters' attributes required: (1) extracting the records of the original and background commuting trips – considered in this study – from the TTS datasets; and (2) determining if certain trip in the model is commuting and properly extracting its attributes (from the database prepared in the first steps) based on its OD and start-time interval. Whereas, preparing the LOS attributes involves processing the detailed path and time trajectories of millions of vehicles, stored in large output files of the simulation model. The records processed for each vehicle contain its OD, start-time, travel time, links traversed, and time spent on each link along the trip. The travel distance of each commuting trip is calculated by summing the lengths of links traversed during that trip and the travel cost is calculated via multiplying the travel distance by the average cost of auto use per unit distance. The value used for the latter parameter is 0.1534 $/km; as was calculated in Miller et al. (2015) based on average gas and other car-related operations and maintenance costs in the GTA.

It is important to note that the model above does not include an explicit variable for the toll cost as the TTS survey dataset contains no toll information to assist in the coefficient estimation of such parameter. For the sake of variable pricing policy testing in this study, the imposed tolls are added to the travel cost variable. The coefficient of the inserted toll variable (in the utility of each departure time choice) is set such that the ratio between the coefficients of travel time and toll variables is compatible with the average VOT used in the DTA simulation model of $15/hr.

It should be noted that forecasting the impact of hypothetical transportation demand management strategies based on revealed preference (RP) model parameters might underestimate the impact of these policies (Habib et al., 2013). In other words, using the auto cost parameter might not be ideally suited for tolls. This is due to the fact that drivers – to some extent – may not be very elastic to increases in travel time and basic costs (maintenance, fuel, etc.); however, they may react more clearly to changes in parking cost and road charges (i.e., out-of-pocket money), as it is something they can avoid. Nevertheless, adding the toll cost to the travel cost variable is expected to give approximate estimation of drivers' behavioral responses to variable pricing. More realistic Modeling of commuters' responses to pricing in the GTA might be achieved by re-estimating the departure time choice model based on stated preference (SP) data surveys incorporating toll information, in addition to the existing revealed preference information in the TTS surveys, which is beyond the scope of this study and can be done in future work.

Although the schedule delay (early or late arrival) cost is intuitively an important factor contributing to the departure time choice for morning commuting trips (having specific desired arrival time), this variable is absent from this model since the work/school start times (i.e., desired arrival times) of commuting trips are not reported in the TTS survey. The schedule delay cost is, however, crucial to attain the anticipated departure time rescheduling effects of tolling in accordance with the bottleneck model triangular pricing structure adapted in this study. Without schedule delay cost, the model would erroneously exaggerate shifting commuting trips to outside the toll period. In other words, schedule delay cost is what keeps commuters "anchored" to their desired arrival times. Accordingly, this variable is added to the model. The detailed formula used in this study for schedule delay as well as the determination process of its parameters are presented in the next subsection.

3.3.3. Empirical model

The departure time choice model considers the following: (1) alternative specific constants, (2) coefficients of variables defining systematic utility functions, (3) coefficients of variables defining root scale parameters, and (4) coefficients of variables defining nest scale parameters. As a result, the model has 74 parameters. As mentioned before, the empirical model was originally estimated based on the datasets from the 2006 TTS survey. The alternative specific constants (ASCs) were hence updated to be consistent with the 2011 dataset, according to the following rule (Train, 2003):

$$ASC_{i_{New}} = ASC_{i_{Original}} + \ln\left(\frac{A_i}{S_i}\right), \quad i = 1, 2, 3, \ldots 9$$

where A_i is the share of decision-makers in the 2011 population who chose departure time interval i; whereas S_i is the share of decision-makers in the 2006 population who chose alternative i. Table 2 shows the ASCs before and after adjustment; the updated constants are proportional to the corresponding shares of drivers in the 2011 dataset - at each time interval - yet carrying the behavioral information involved in the originally estimated constants.

The schedule delay cost, c_s, used in this study takes the following formula (Small, 1982):

$$c_s = \begin{cases} \beta(t_d - t - T(t)) & \text{if } t + T(t) \leqslant t_d \quad (\textit{Early Arrival Cost}) \\ \gamma(t + T(t) - t_d) & \text{if } t + T(t) > t_d \quad (\textit{Late Arrival Cost}) \end{cases}$$

where β and γ are the shadow prices of early and late arrival delays, respectively. t is the trip start time, $T(t)$ is the travel time, and t_d is the desired arrival time. As mentioned before, the commuters' desired arrival time info is not reported in the TTS survey; only actual arrival time (shifted from desired time by an unrevealed schedule delay component) is reported. Accordingly, the desired arrival time (t_d) is randomly generated – in this study – for each vehicle in the network following a 'Log-Normal' distribution. i.e., $\ln(t_d)$ is assumed to have a 'Normal' distribution with parameters μ (mean) and σ (standard deviation). The Log-Normal distribution is suitable for random variables inherently positive. Additionally, it has a quasi-bell shape that enforces ascending probabilities for values (i.e., desired arrival times) close to the mean, and vice versa. Accordingly, it is believed to produce more realistic distribution of simulated desired arrival times than following a uniform distribution.

Several values were tested for the mean and standard deviation of this distribution; 8:30 am (i.e., minute 150 counting from 6:00 am) was ultimately selected as the mean desired arrival time (i.e., $\mu = \ln(150)$) and $\sigma = 0.05$ – measured in ln(*minute*) – was set as the standard deviation. The selected parameters were found to bring adequate validation results among other tested values, when the integrated – departure time and traffic assignment – testbed is applied in the base-case. More specifically, they resulted in the closest output distribution of simulated departure (hence arrival) times for commuting trips as those obtained in the GTA base-case traffic assignment simulation results (without applying the departure time choice model), as will demonstrated later. Furthermore, the selected parameters entail the best relationship between travel time and schedule-delay cost values; such that the minimum schedule-delay costs are observed at the same time-interval where the maximum travel time delays are experienced, and vice versa, as suggested by the bottleneck model (Fig. 1b).

According to Small (1982), the early and late arrival delays are perceived differently by commuters and hence have different coefficients (i.e., shadow prices) in the schedule delay cost function with a ratio of 1–4, respectively. This ratio was further modified in this study, during model validation, to be 1–2 which better fits the GTA data. As mentioned before, the departure time choice model uses a scale parameterization approach where the root and nest scale parameters do not take constant values; rather, they vary according to trip and driver attributes. These parameters are eventually multiplied by the coefficients of different variables in the model, including the integrated schedule delay cost variable. Accordingly, the scale parametrization approach implicitly captures heterogeneity in drivers' values of (early or late) schedule delay.

After the schedule delay cost is added to the model, the coefficients of travel time in the utility functions needed to be recalibrated. The calibrated parameters were determined using a factorial design procedure, (Cheng, 2013), and are reported in Table 3. The objective of this procedure was to determine the set of parameters that minimize the absolute error between the observed and the estimated values at all time intervals; for the following measurements:

Table 2
Original and new Alternative Specific Constants (ASCs).

Original ASCs	0	−0.4508	−0.2099	0.1803	0.3659	0.1143	0.007	−0.3665	1.3054
New ASCs	1.0010	1.0426	1.4983	2.0899	2.3680	2.2478	2.0469	1.3484	1.7053

Table 3
Original and adjusted coefficients of the travel time variable.

Original Time Coeffs	0	−0.0107	−0.0087	−0.0149	−0.0196	−0.03	−0.0332	−0.0182	0
New Time Coeffs	−0.015	−0.0187	−0.0167	−0.0249	−0.0196	−0.015	−0.0102	−0.0082	−0.005

- number of commuters who chose to depart at each time interval;
- average resulting travel time per km (calculated by averaging the travel time of each commuter divided by the distance traveled in km, over all commuters departing at each time interval); and
- average travel distance traveled.

The simulation process of commuters' departure time selection process in the system proposed involves looping over all commuting vehicles in the model. The personal attributes of each commuter are linked to the corresponding trip LOS attributes generated – at all departure time intervals – by the DTA simulation model of the GTA. The schedule-delay and toll costs are then calculated for that commuter at all time intervals. The extracted and calculated variables are hence plugged into the model formulas to obtain the probability of choosing each departure time interval. The commuter departure time choice is determined using a Roulette Wheel selection approach, in which choices with higher probabilities have higher chances to get selected (Back, 1996). The trip new start-time is then calculated by adding or subtracting multiples of 30 min (depending on the departure time interval chosen) to its original start-time set in the DTA simulation run under original TTS demand. After all commuting trips are processed, their start-times are updated in the input demand file of the DTA simulation model.

In Fig. 8, the number of commuting trips started at each half-hour interval and their corresponding average travel time per km are compared among two simulation runs, whose measurements are referred to in the figure as "original demand" and "modified demand". The total number of commuting trips in the GTA model – for which departure time choice model is applied and plots in Fig. 8 are reported – is around 1,270,000 trips (out of a total of 1.8 million trips in the model), as mentioned before. The measurements under "original demand" are obtained from the output of a GTA DTA simulation run in base-case conditions (i.e., without tolling) using the original demand extracted from TTS survey data, without applying the departure time choice model. Whereas, those under "modified demand" are obtained from applying the retrofitted/re-calibrated departure time choice model iteratively with the GTA DTA simulation model under base-case conditions. The patterns shown in the figure indicate the best attainable correspondence, in the absolute values and the overall trends, between the 'original' and 'modified' demand related measurements after performing all model retrofitting/calibration steps.

The 9 departure time intervals used in this model (6–6:30, 6:30–7 ... 10–10:30) were assigned the numeric indices 0, 1 ... 8. For each vehicle in the simulation model, the difference between its observed (original) departure time interval index and its estimated one was calculated at the end of the iterative simulations. The value of the difference lies between −8 and 8. Intuitively, the higher the percentage of vehicles with a zero difference (when estimated and observed departure time intervals coincide) the better. The chart in Fig. 9 shows the percentage of vehicles whose difference lies in each index difference group when applying the calibrated discrete choice model iteratively with the DTA simulation model in the base case (i.e., without tolling). It is clear that the estimated departure time choice of more than 80% of the commuters lies within 3 intervals (before or after) from the original (half-hour) choice, which we believe is acceptable; given the continuous nature of the departure time and the boundary value problems that may result from time discretization. The findings from

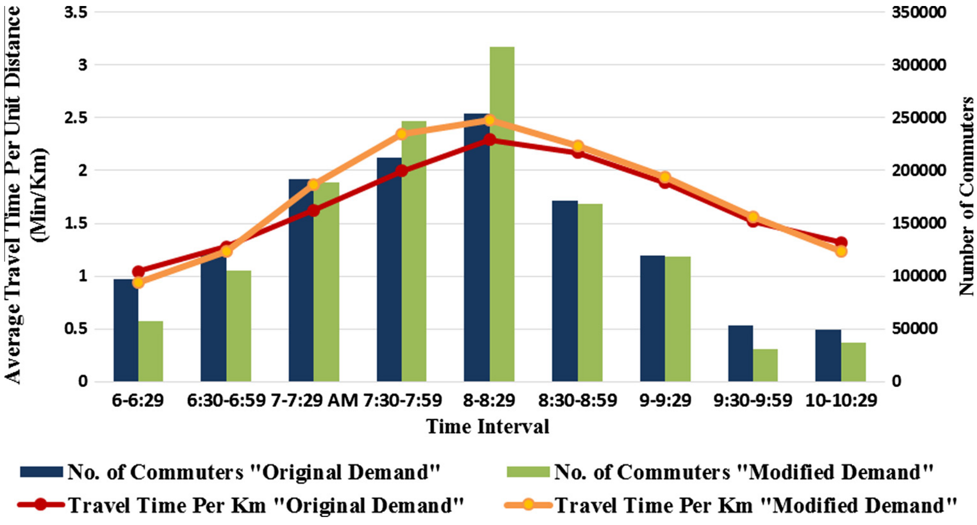


Fig. 8. Comparisons between observed and estimated simulation measurements.

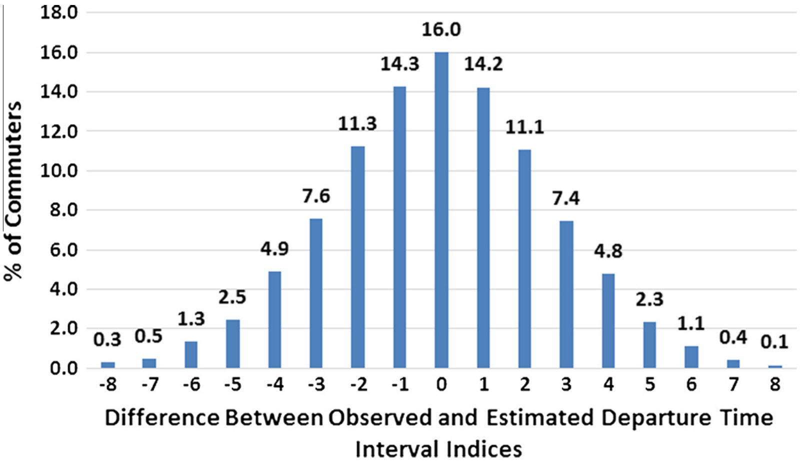


Fig. 9. Percentage of commuters vs. indices difference.

Figs. 8 and 9 demonstrate the performance of the calibrated framework when applied to the GTA in the base case without tolling.

The discrete choice model has 74 statistically significant parameters, among which only 18 needed to be adjusted for the validation of the model outputs. The adjustment was necessary for the following reasons: (1) to update the model to be consistent with the 2011 TTS dataset being used, and (2) to adapt with the added schedule component cost. It should be noted that we retrofitted the 2006 model to the target year of 2011 for three reasons: (1) the 2006 model was recently developed, i.e., existed, and repeating the estimation for 2011 was out of our scope, (2) updating a 2006 model using 2011 dataset is a way of using two repeated cross-sectional datasets in a pseudo-panel data formation where 2006 data are used for estimation and 2011 data for validation, and (3) estimating a departure time choice model that captures toll cost and schedule delay cost directly was not possible as neither the 2006 nor the 2011 TTS data contained the necessary information, i.e., retrofitting was unavoidable. The retrofitting process performed, however, shouldn't affect the robustness of the original model formulation given its relatively large number of parameters and statistically significant explanatory variables, as well as the parameterized root and nested scale parameters.

The integration of the described discrete choice model into the proposed congestion pricing framework is important to assess the differential impact of pricing scenarios on the departure time choice of distinct drivers, based on their personal and socio-economic attributes used in the model. The departure time choice model considers users' heterogeneity in values of (early or late) schedule delay and desired arrival time. At the DTA level, however, the traffic assignment software allows for only single-user class with single value of time. Considering multiclass traffic assignment (i.e., considering the effect of heterogeneous VOT on route choice) is not a simple modification of the software and is hence deferred to future work.

As illustrated in Section 2.4, the DTA network simulation model and the departure time choice model run sequentially and iteratively until convergence in the departure time model output (i.e., drivers' start-time rescheduling responses to tolling) is reached. This is achieved when the maximum absolute relative difference in the total share of vehicles at any departure time interval drops below a pre-specified convergence threshold, denoted as $\propto$. It should be noted that the randomness inherent in the nature of the probabilistic discrete departure time choice process might cause some variation/difference in the discrete choice model output, even when the model is applied repeatedly under identical inputs. Therefore, the value of $\propto$ should be higher than the upper limit of those potential differences. Observing the model output - across different runs - when applied on the GTA morning commuting trips (around 1,270,000), under identical inputs, it was found a suitable value for $\propto$ to be used in this application is 0.1. According to the convergence criteria specified, it takes the integrated framework around 3 iterations of the outer loop to converge in the GTA simulation-based case study. This is a relatively fast convergence in terms of the number of iterations required for such a large-scale application. On an i7 Machine with 16 GB of RAM, the run-time of the integrated departure time and DTA simulation models until convergence, under each tolling scenario being tested, is around 18 h.

The departure time choice model integration process was accompanied by many challenges. For example, preparing the driver-related data required by the model entailed time-consuming efforts to process the raw 2011 TTS survey datasets and properly extract the attributes linked to each (original or background) commuter identified in the GTA model. Moreover, calculating the network-related attributes required by the model involves processing vehicles' records stored in massive output files, produced by the traffic assignment simulation model. The calculation process is obviously time and computationally demanding. Moreover, it is repeated iteratively – post the termination of each GTA traffic assignment simulation run – to provide the departure time choice model with the updated network attributes based on which the model estimates the new demand profiles to be fed back to the traffic simulation model, and so on until convergence. As a result, this process represents the second major factor, after the DTA simulation model run-time, causing the running time of the full (integrated) system to be long (around 18 h).

In the following section, the application of the overall variable congestion pricing framework is illustrated using tolling scenarios in the GTA.

4. Tolling scenario evaluation

4.1. Tolled route (Gardiner Expressway- GE)

The implemented framework is intended to test different tolling scenarios; e.g., single or multiple freeways, urban corridors, HOT lanes, a sub-network and cordon tolls. As a first implementation, the route selected in this study to be tolled is the Gardiner Expressway (GE). The GE, as shown in Fig. 3, is the main artery running through Downtown Toronto, the core and economic hub of the GTA and arguably Canada. The expressway is 18 km long between Highway 427 and the Don Valley Parkway (DVP). It is six-to-ten lanes wide in varying locations. In addition to the fact that the GE suffers from extended periods of congestion, there is an ongoing debate on whether to tear it down, to toll it and use the revenue for its maintenance, or to apply other hybrid proposals to improve its operation. Hence, the GE was our first choice to test the proposed variable congestion pricing framework. It is important, however, to emphasize that although the pricing strategy is applied only to this main artery within the heart of Toronto, the impact of doing so is regional, as it draws demand from across the GTA. Therefore, the simulations and analyses are conducted on the entire GTA network, due to the inter-connectivity and multiple routing options existing in this network and to capture regional effects.

4.2. Toll structure

The toll considered is distance-based and its value is entered in $/km. Different toll values are set at different time intervals, according to a triangular structure (as shown in Fig. 2). The study period is focused on the morning period from 6:00 to 10:30 am when the majority of commuting trips in the GTA occurs, and significant traffic utilizes the GE to downtown Toronto. The variable-tolling intervals used are the nine half-hour intervals shown in Fig. 7, for compatibility with the departure time choice model.

4.2.1. Determination of morning peak period duration and variable toll structure

The number of trips during 6:00–10:30 am morning period on the GE corridor (i.e., the Gardiner Expressway and its parallel arterials) is approximately 90,000. In order to attain the departure time scheduling benefits of variable tolling, the toll pattern should replicate the queueing-delay pattern during the peak period, as suggested by the bottleneck model (presented in Section 2.1). For that purpose, the toll structure determination process starts with determining the morning peak period start and end times (on the GE corridor) as well as the pattern of excess travel time (i.e., queueing-delay) during that period.

As shown in Fig. 1a, the peak period is considered to start when the inflow exceeds the available route capacity, resulting in traffic queues and increased travel times that build-up to a maximum when the inflow starts decreasing below capacity. The peak does not end at this point of time; rather, it ends when all travelers who entered the system (from the beginning of the peak period) ultimately exit after being queued for a while.

According to this definition, and based on the demand information and the base-case simulation results of the trips made on the GE corridor, the peak period start and end times were found to be 7:00 am and 9:30 am, respectively. Consequently, no toll is imposed before 7:00 am or after 9:30 am in the variable pricing scenarios tested in this study. Additionally, and according to the bottleneck model optimum triangular pricing rules, the toll pattern selected replicates the pattern of queueing-delays on the corridor in the un-priced equilibrium, shown in Fig. 10, which - as mentioned - was found to exist

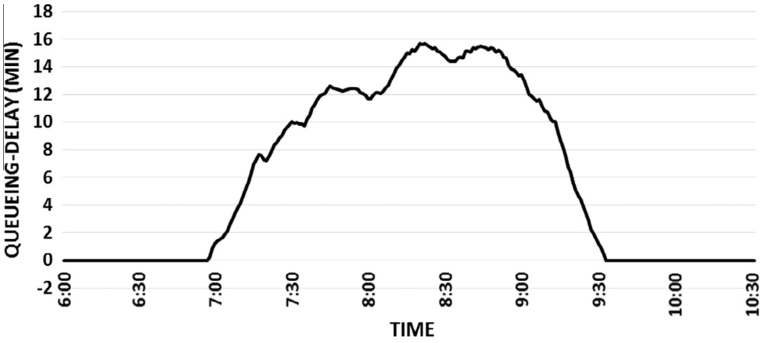


Fig. 10. Average (base-case) queueing-delay on the GE corridor.

between 7:00 am and 9:30 am. The queueing-delay, at any time instant, was calculated as the average excess travel time, at this time instant, over the travel time experienced just before the peak started.

As mentioned earlier, the continuous toll pattern of the bottleneck model is approximated in this study through step tolls in which distinct toll values are imposed on half-hour tolling intervals. The first and last tolling intervals are identified as the half-hour intervals during which the peak period of that facility starts and ends, respectively. The variable toll pattern replicates the estimated base-case queueing-delay pattern in order to attain the desired rescheduling benefits of variable tolling. Therefore, the variable toll is assigned a zero value during early and late intervals having zero queueing-delay; whereas, it is assigned the maximum value during the interval having the largest average queueing-delay.

As a reference, a toll value of 0.15 $/km should be set per interval for every 1 min/km average 'queueing-delay per km' experienced in the base-case during that interval. This value was found – among multiple values tested – to create moderate route shifts to parallel (non-tolled) arterials, under the average VOT used in the simulation model. Accordingly, the toll value at each tolling interval ($i \in \{1, 2, \ldots, 9\}$) - in the variable toll structure - is calculated by multiplying 0.15 by the average 'queueing-delay per km' estimated on the GE during that interval. The latter value represents the average of queueing-delay values (plotted in Fig. 10) estimated during the designated interval divided by the GE length in km.

4.2.2. Tolling scenarios

In order to study the effectiveness of the proposed framework in variable congestion-pricing policy evaluation, two tolling scenarios are investigated: (1) variable tolling structure that replicates the queueing-delay pattern, as described above, and (2) flat tolling across all time intervals; its value was set by taking the average of the time-dependent non-zero toll values of the first tolling scenario, for a fair comparison between two tolling structures having the same 'average' order of magnitude, as shown in Fig. 11.

4.3. Results and conclusions

4.3.1. Network-wide analysis

Total travel times network-wide dropped from 595,346 h - in the base-case - to 585,510 h as a result of variable tolling; i.e., 9836 h (1.7%) were saved network-wide. Whereas, flat tolling resulted in an increase in total travel times to 598,659 h; i.e., base-case travel times network-wide increased by 3313 h (0.6%). Fig. 12 shows the major routing decision points for traffic approaching Toronto. The results are summarized in the form of percentage difference of overall traffic flow during the period from 6:00 am to the end of the tolling period (in each case) along the key corridors between the base case, the flat tolling and the variable tolling scenarios. Examining the results indicates the following:

4.3.1.1. Variable tolling.

- Overall, the variable toll resulted in mild routing changes across the GTA when compared to the flat tolling scenario; −1% at QEW, +5% at Highway 401, and −7% at DVP.
- At the GE, only 5% divergence was observed at the bifurcation to Lake Shore; resulting in maximizing the efficiency of the downstream sections of the GE.

4.3.1.2. Flat tolling.

- Overall, the flat toll resulted in more pronounced re-routing patterns across the GTA compared to variable tolling; showing −2% at QEW, +5% at Highway 401, and −8% at DVP. Flat tolling is less conducive to departure time changes as all periods have the same toll, and therefore its impact is predominantly on re-routing.

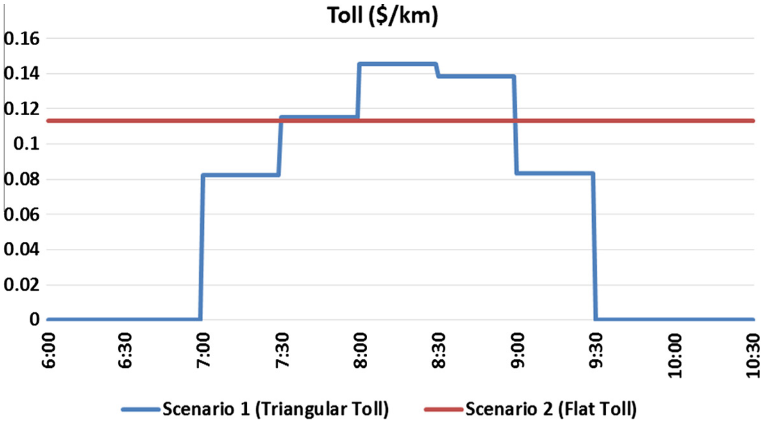


Fig. 11. Tolling scenarios 1 and 2 for the Gardiner expressway.

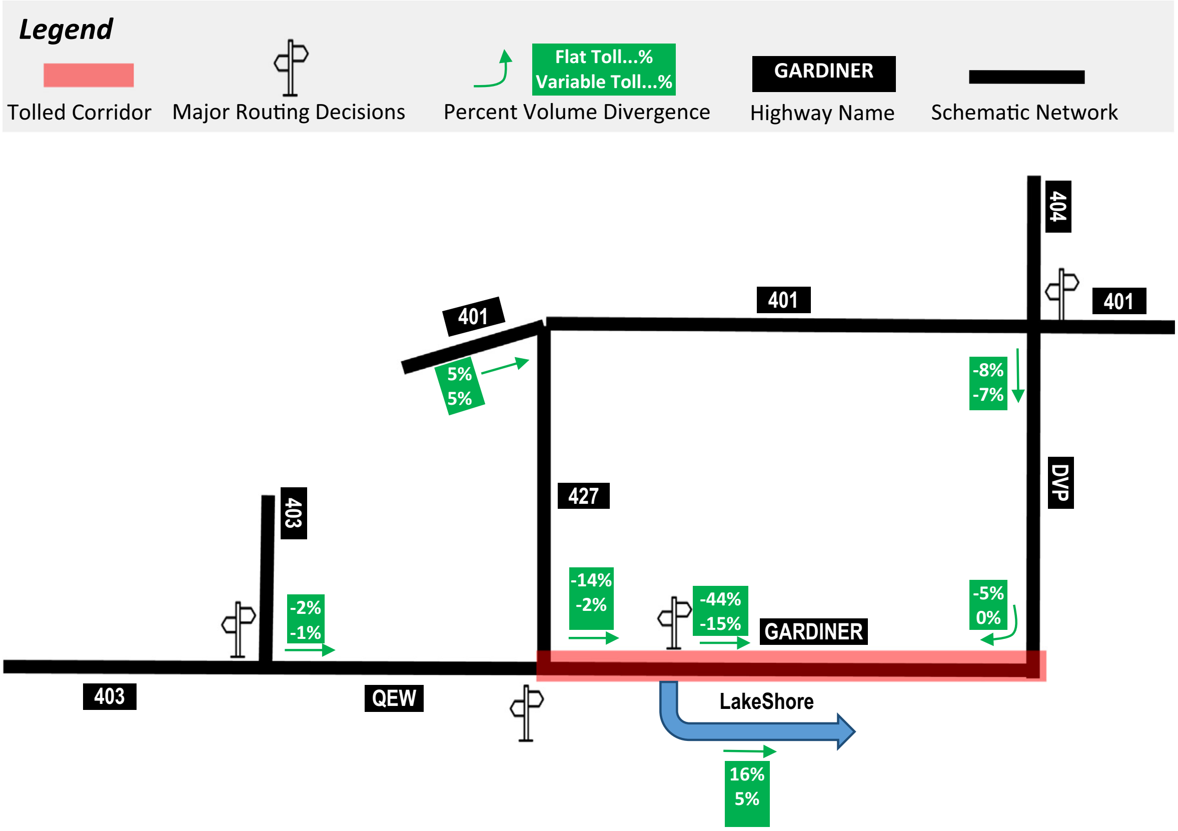


Fig. 12. Major routing decision points for GE Corridor traffic.

- On the GE, significant divergence (re-routing) was observed at the bifurcation to Lake Shore; resulting in shockwave and congestion upstream of this bifurcation. This congestion resulted in – interestingly – less flow on the GE downstream from the off-ramp to Lake Shore, i.e., underutilizing the GE by as much as 44%. This observation was confirmed by the low speed values (20–28 km/h) along the sections of the GE upstream of the off ramp.

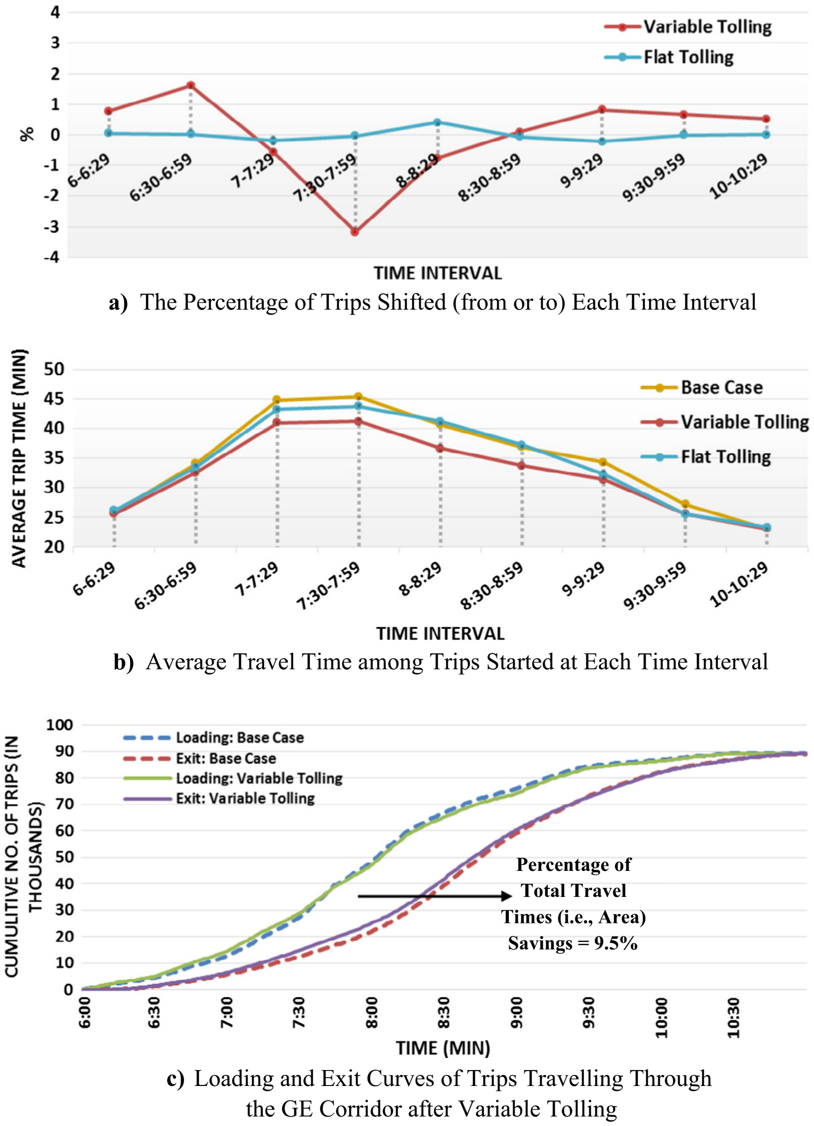
4.3.2. Trip-based analysis

Fig. 13 shows: (a) the changes in departure time choices, (b) travel times, and (c) the patterns of entry and exits from the network for the original 90,000 trips that traveled through the GE corridor (i.e., the GE and its parallel arterials) in the morning period, under different tolling scenarios. This analysis involves all the trips that are affected by tolling the GE, including:

- trips passing through the tolled route;
- trips diverting from the tolled route to other parallel arterials after tolling (e.g., Lake Shore Blvd); and
- trips on the parallel arterials that might be affected by those shifting their route to avoid the tolled route.

4.3.2.1. Variable tolling. As clear from Fig. 13a, variable tolling induced shifting approximately 5% of the peak hour traffic passing through the corridor (from 7:30 am to 8:30 am) to earlier and later time intervals. As a result, lower travel times are observed at all time intervals after variable tolling, as shown in Fig. 13b. Further, the variable pricing scenario resulted in 9.5% savings in the total travel times of the trips that traveled through the corridor (at all time intervals), relative to the base case as shown in Fig. 13c. In Fig. 13c; the total area between the loading and exit curves of the trips that traveled through the corridor (which represents the total travel times spent on the network by those trips) shrunk by 9.5%. The benefits come from rescheduling of departure times from the trip origin, in addition to the route shift impacts of tolling. Moreover, this figure shows that – unlike in the simple Bottleneck Model – variable tolling on real-world road networks affects not only the cumulative loading curve but also the cumulative exit curve.

4.3.2.2. Flat tolling. Flat tolls create no incentive for drivers to avoid relatively congested periods by changing their departure times across the tolled periods, as they have the same toll. This is noticed in Fig. 13a. This scenario outperforms the base case by only 2% net savings in the total travel times compared to 9.5% in the variable tolling case. The benefits under flat tolling come solely from the route shift impacts of tolling. However, as clear in Fig. 13b, this gain is realized more at early and late



a) The Percentage of Trips Shifted (from or to) Each Time Interval

b) Average Travel Time among Trips Started at Each Time Interval

c) Loading and Exit Curves of Trips Travelling Through the GE Corridor after Variable Tolling

Fig. 13. Analysis of the trips that traveled through the GE Corridor at different tolling scenarios.

intervals while some deterioration in travel times is observed at peak time intervals (8–9 am). An additional explanation for these findings will be given in the next section.

4.3.3. Tolled route-based analysis

Fig. 14 shows the average travel times on the tolled route (the GE), eastbound direction, from highway 427 to the DVP. The times are reported at each time interval for different tolling scenarios.

4.3.3.1. Variable tolling. As noticed in Fig. 14, variable tolling entails noticeable decrease in travel times on the tolled route; especially at the middle congested time intervals. The maximum saving observed is 7 min (out of 27 min), i.e., around 25%, at the 8:00–8:30 am time interval.

4.3.3.2. Flat tolling. Flat tolling results in improvements in travel times at early and late intervals. However, it causes significant increase in travel times on the tolled route from 8:30 to 9:30 am, as clearly shown in Fig. 14; which agrees with the findings from the trip-based analysis. The deterioration occurs due to the excessive demand at peak hours that didn't shift to other time intervals due to absence of any incentive to do so (i.e., no toll variation over time). This demand tries to exit the tolled route (the GE) to the immediate parallel arterials (Lake Shore Blvd) and is limited by off-ramp and arterial capacity constraints. Therefore, it creates congestion upstream that blocks the tolled route itself at peak hours, which is very

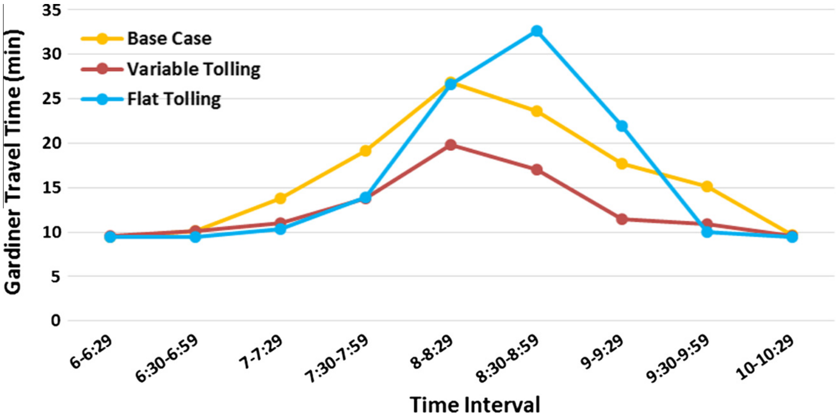


Fig. 14. Average travel time on the Gardiner Expressway Eastbound (from 427 to DVP).

counterproductive. In fact, this observation demonstrates how flat tolling on real-world road networks (in which congestion propagates in the form of spillbacks, shockwaves, etc.) can have appreciably different effects than has been suggested by studies of single links or toy networks.

5. Conclusions and future work

As presented in this paper, congestion pricing is widely viewed among economists and practitioners as one of the promising control tools to tackle traffic congestion. A significant amount of research has been conducted thus far in this area. However, theoretically and or methodologically sound studies are often applied to small or even hypothetical networks, i.e., case studies on large-scale urban network models are scarce. Additionally, the tolling scenarios applied in most practically oriented studies lack methodological justification. Furthermore, the users' individual responses to pricing (e.g., departure time and route choices) were usually disregarded and, if considered, the impact of personal and socio-economic attributes on their choices was often neglected. In this study, a framework for variable congestion pricing policy evaluation has been presented with detailed implementation information on a simulation-based case study in the GTA, in Ontario, Canada. The framework considers the heterogeneity in drivers' values of (early or late) schedule delay and desired arrival time; it involves a discrete-choice model for departure time choice that has been extended in this study to incorporate a schedule delay cost component (besides the existing personal, socio-economic, and trip related attributes) for realistic modeling of morning peak travel behavior. The framework is intended to be general and applicable to a variety of tolling scenarios (e.g., congested highway sections, HOT lanes, cordon tolls, etc.). As a first implementation, it has been utilized in this study to analyze the impact of different tolling scenarios on the Gardiner Expressway (GE), a key freeway passing through downtown Toronto. Impacts are assessed at the regional level, trip level as well as tolled-route level. The results affirm the effectiveness of the integrated variable pricing framework in analyzing the effect of variable tolling in a large-scale simulation application. Moreover, the results obtained demonstrate how congestion pricing on real-world road networks can have different effects than has been suggested by studies of single links or toy networks. For example, unlike in the simple Bottleneck Model, variable tolling affects not only the cumulative loading curve but also the cumulative exit curve. Another example is that imposing a flat toll on a link can actually increase travel time on the link because of spillback.

It can be concluded from the analysis of different tolling scenarios presented in this study (on network, trip, and tolled-route basis) that:

1. In a large-scale interconnected network (like the GTA) where long-distance trips have diverse routing options, tolling a relatively short, yet major, highway like the GE creates temporal and spatial traffic changes network-wide that go beyond the tolling interval and the tolled route. This confirms the necessity of conducting the simulations on a regional scale for policy determination and assessment.
2. More benefits are gained from departure time re-scheduling due to variable pricing, compared to just re-routing as in flat tolling. This emphasizes the importance of the integrated discrete-choice module to the proposed variable congestion pricing framework, to provide a realistic modeling of users' individual departure time responses to variable pricing policies.
3. Pricing that induces re-routing only (and no departure time re-scheduling), or excessive re-routing due to, for instance, over pricing, can send traffic to off ramps to parallel routes so aggressively that it blocks the off ramp and backs up onto the main freeway, limiting access to the priced road itself, which is not only counterproductive but also nullifies the very purpose of pricing itself. This emphasizes the importance of variable pricing to mirror congestion patterns over time, which is the methodological basis (adapted from the bottleneck model) of the proposed variable tolling framework.

The next step in this research is to integrate an optimization module to the variable congestion pricing framework to fine-tune the variable toll structures obtained based on the bottleneck model pricing rules, in order to consider the network-wide dynamics and the possible route shifts to parallel arterials that were absent in the bottleneck model. This should eventually produce the optimal spatio-temporal toll structure (i.e., toll value on each tolled link for every time interval) resulting in the optimal schedule of entering the system and spatial distribution of traffic across the network that would minimize the total travel delay and maximize infrastructure utilization. Additionally, mode and destination choice are potential responses to tolling that will be added to the framework in future work. The framework presented in this study can also be extended by: including transit and trucks demand; developing/integrating the details of transit networks in the GTA and a transit assignment module to the DTA simulation model; considering multi-class traffic assignment though heterogeneous (rather than single) VOT assumption in the route choice models which requires a modification of the DTA simulation software used or using other DTA software that allows for multi-class assignment; and investigating drivers' perception and behavioral responses towards variable tolling policies in the afternoon/evening peak period.

References

Abdelgawad, H., Abdulhai, B., 2009. Optimal spatio-temporal evacuation demand management: methodology and case study in Toronto. In: Proceedings of the 88th Annual Meeting of the Transportation Research Board, Washington, DC.

Back, T., 1996. Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms. Oxford University Press, p. 1996.

Balmer, M., Meister, K., Rieser, M., Nagel, K., Axhausen, K.W., 2008. Agent-based simulation of travel demand: structure and computational performance of MATSim-T. Vortrag. In: 2nd TRB Conference on Innovations in Travel Modeling, Portland, June 2008.

Cheng, C.-S., 2013. Theory of Factorial Design: Single- and Multi-Stratum Experiments. CRC Press LIc.

Chiu, Y.-C., Nava, E., Zheng, H., Bustillos, B., 2008. DynusT User's Manual. <http://wiki.dynust.net/doku.php>.

Costs of Road Congestion in the Greater Toronto and Hamilton Area: Impact and Cost Benefit Analysis of the Metrolinx Draft Regional Transportation Plan. Final Report, Greater Toronto Transportation Authority (GTTA), 2008.

DMG, 2015. Transportation Tomorrow Survey: Design and Conduct of The Survey. Data Management Group, University of Toronto, Joint Program in Transportation. <http://www.dmg.utoronto.ca/reports/ttsreports.html> (accessed in July 2015).

Duranton, G., Turner, M.A., 2011. The fundamental law of road congestion: evidence from US cities. Am. Econ. Rev. 101 (6), 2616–2652.

Finkleman, J., Casello, J., Fu, L., 2011. Empirical evidence from the Greater Toronto Area on the acceptability and impacts of HOT lanes. Transp. Policy 18, 814–824.

Gragera, A., Sauri, S., 2012. Effects of time-varying toll pattern on social welfare: case of metropolitan area of Barcelona, Spain. In: Transportation Research Board Annual Meeting 2012, Paper #12-4723.

Guo, X., Yang, H., 2012. Pareto-improving congestion pricing and revenue refunding with elastic demand. In: Transportation Research Board Annual Meeting 2012, Paper #12-6650.

Habib, K.M., Sasic, A., Weis, C., Axhausen, K., 2013. Investigating the nonlinear relationship between transportation system performance and daily activity-travel scheduling behaviour. Transp. Res. Part A 49, 342–357.

Habib, K.M., Weiss, A., 2014. Evolution of latent modal captivity and mode choice patterns for commuting trips: a longitudinal analysis using repeated cross-sectional datasets. Transp. Res. Part A 66, 39–51.

Hardin, G., 1968. The tragedy of the commons. Science 162, 1243–1248.

Kamel, I.R., Abdelgawad, H., Abdulhai, B., 2015. Transportation big data simulation platform for the Greater Toronto Area (GTA). In: The EAI International Conference on Big Data and Analytics for Smart Cities, 2015.

Lightstone, Adrian, 2011. Congestion Charging in the City of Toronto: Distance based Road Pricing on the Don Valley Parkway and Gardiner Expressway M. Sc. Thesis. Royal Institute of Technology, Stockholm, Sweden.

Lu, C.-C., Zhou, X., Mahmassani, H.S., 2006. Variable toll pricing and heterogeneous users. Transp. Res. Rec. 1964, 19–26.

Lu, C.-C., Mahmassani, H.S., 2008. Modeling user responses to pricing. Transp. Res. Rec. 2085, 124–135.

Lu, C.-C., Mahmassani, H.S., 2011. Modeling heterogeneous network user route and departure time responses to dynamic pricing. Transp. Res. Part C 19, 320–337.

Lu, C.-C., Mahmassani, H.S., Zhou, X., 2008. A bi-criterion dynamic user equilibrium traffic assignment model and solution algorithm for evaluating dynamic road pricing strategies. Transp. Res. Part C 16, 371–389.

Mahmassani, H.S., Zhou, X., Lu, C.-C., 2005. Toll pricing and heterogeneous users: approximation algorithms for finding bi-criterion time-dependent efficient paths in large-scale traffic networks. Transp. Res. Rec. 1923, 28–36.

Miller, E.J., Vaughan, J., King, D., Austin, M., 2015. Implementation of a "Next Generation" activity-based travel demand model: the Toronto case. In: Paper Prepared for Presentation at the Travel Demand Modelling and Traffic Simulation Session of the 2015 Conference of the Transportation Association of Canada, Charlottetown, PEI.

Morgul, E.F., Ozbay, K., 2010. Simulation Based Evaluation of Dynamic Congestion Pricing M.Sc. Thesis. State University of New Jersey.

Nikolic, G., Pringle, R., Jacob, C., Mendonca, N., Bekkers, M., Torday, A., Rinelli, P., 2015. On-line dynamic pricing of HOT lanes based on corridor simulation of short-term future traffic conditions. In: Transportation Research Board Annual Meeting 2015.

Roorda, M.J., Hain, M., Amirjamshidi, G., Cavalcante, R., Abdulhai, B., Woudsma, C., 2010. Exclusive truck facilities in Toronto, Ontario, Canada: analysis of truck and automobile demand. Transp. Res. Rec. 2168, 114–128.

Sasic, A., Habib, K.M., 2013. Modelling departure time choices by a Heteroskedastic Generalized Logit (Het-GenL) model: an investigation on home-based commuting trips in the Greater Toronto and Hamilton Area (GTHA). Transp. Res. Part A 50, 15–32 (M.Sc.).

Small, K.A., 1982. The scheduling of consumer activities: work trips. Am. Econ. Rev. 72, 467–479.

Small, K.A., Verhoef, E.T., 2007. The Economics of Urban Transportation. Routledge, England.

Swait, J., 2001. Choice set generation within the generalized extreme value family of discrete choice models. Transp. Res. Part B 35 (7), 643–666.

TomTom North America Congestion Index, 2014. TomTom International BV.

Train, K., 2003. Discrete Choice Methods with Simulation. Cambridge University Press.

Verhoef, E.T., 2002. Second-best congestion pricing in general networks, heuristic algorithms for finding second-best optimal toll levels and toll points. Transp. Res. Part B 36, 707–729.

Washbrook, K., Haider, W., Jaccard, M., 2006. Estimating commuter mode choice: a discrete choice analysis of the impact of road pricing and parking charges. Transportation 33, 621–639.