Mode Boundaries of Automated Metro and Semirapid Rail in Urban Transit

Luigi Moccia*,†, Duncan W. Allen‡, Gilbert Laporte†§¶, and Andrea Spinosa‖

Abstract

Our research question is to what extent, and under what circumstances, full automation in metro lines defines transit mode boundaries with respect to semirapid transit. The modeling approach is based on micro-economic appraisal. Automation, beside changing the investment and operation and maintenance cost profiles of metro lines, can improve some aspects of the user experience. The low marginal cost of frequency possible with automated metro may unlock both users’ benefits via reduced waiting times and crowding, and operator’s savings via shorter trains and smaller platforms and stations. We show how the user’s travel time components are structurally different under several configurations. In particular, we highlight the critical role that different demand and alignment patterns play in the mode comparison.

*Istituto di Calcolo e Reti ad Alte Prestazioni, Consiglio Nazionale delle Ricerche, Via P. Bucci 41C, 87036 Rende (CS), Italy - luigi.moccia@icar.cnr.it
†Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT), Canada
‡IBI Group, 21 Custom House, Boston, US 02110
§HEC Montréal, 3000 chemin de la Côte-Sainte-Catherine, Montréal, Canada H3T 2A7
¶School of management, University of Bath, Bath, UK
‖Università La Sapienza, Dipartimento di Ingegneria Civile, Edile ed Ambientale (DICEA), Roma, Italy
1 Introduction

The mode boundaries as defined in the classic work of Vuchic (2005) contain little overlap between the rapid and semirapid classes of urban transit. We investigate the role of full automation in defining the boundaries of rapid rail, also known as “metro”, versus semirapid rail. Here by “full automation” we refer to the Grade of Automation level four (GoA4) defined by the International Association of Public Transport (UITP, 2019) when no operating staff is required on board. Under the Grade of Automation level three (GoA3) the train is automated but a qualified train operator is still required on board for emergency situations. In 2020 forty cities in the world have at least one GoA3/4 metro line (Figure 1).

Figure 1: Cities with operational GoA3/4 metros in year 2020.

The two urban rail modes that we study typically offer a common set of advantages such as electric traction and high riding comfort. This allows us to concentrate on other aspects that are specific to the comparison between human-operated and automated driving. Guide automation not only changes the investment, operation and maintenance costs of metro lines, but can improve some aspects of the user experience (Wang et al., 2016). We focus our analysis on the user experience aspects where the quantitative assessment methods are more robust, namely the appraisal of the travel
time components related to the users’ access and egress, waiting, and in-vehicle travel time, including a penalization for crowding. We remark that an automated metro can offer additional advantages such as extended service hours, fast response to demand variations, and enhanced safety. Therefore, our analysis is only a partial and conservative answer to our research question. In spite of this limitation, we develop scenarios with a rich mapping between transit technology and design, operator costs, and user’s experience. Guide automation currently requires an alignment with an exclusive right-of-way (RoW) which is not necessary with conventional semirapid rail. An exclusive RoW confers both user advantages, such as a higher travel speed, and disadvantages, such as a slower access to train platforms. An exclusive RoW may occur under several alignment types, for example by elevated or underground sections, which entails further differences in the access and egress times in the same mode category. Alignment differences may also occur in the case of semirapid rail, altering a key performance metric such as the average commercial speed. These alignment differences within and between modes further interact with different types of passenger demand that can be observed in practice. For example, users traveling short distances and accessing the line on foot will place more importance on a fast pedestrian access to closely spaced stations than to the riding time, whereas the opposite will hold for users accessing the line by a motorized feeder mode for a longer trip. These considerations motivated our development of scenarios where the patterns of both the passenger demand and the alignment of the studied modes differ. We show that sweeping generalizations about the prevalence of a mode over another are unsubstantiated, and that specific circumstances, particularly those related to the feasible alignment, play a large role. Nevertheless, we find that automated metro may conveniently extend its coverage to demand levels previously considered exclusive domain of semirapid rail.

The remainder of this paper is organized as follows. Section 2 contains the literature review which motivates our work. Section 3 describes the model which is an adaptation of that of Moccia et al. (2018) updated in Moccia et al. (2020) to rapid transit with full automation. Section 4 discusses plausible scenarios, and Section 5 presents
the results of our computational experiments. Some conclusions and comments on future work are reported in Section 6.

2 Literature review

The modeling approach that we follow is based on micro-economic appraisal of transit, a literature thread that we review in Section 2.1. A critical question often raised against the transit optimization models with fixed demand is their relevance under the prevailing and more general case where demand is elastic. We address this question in Section 2.2. The broader issue of guide automation in passenger transport, some of the models used to assess this technological trend, and the impacts on urban transit are discussed in Section 2.3. Section 2.4 positions our contribution in the reviewed literature.

2.1 Microeconomic appraisal of transit

Jara-Díaz and Gschwender (2003) review the microeconomic appraisal of transit literature and extend it by including the effect of vehicle size on operating costs and the influence of crowding on the value of in-vehicle time. The authors clarify how a broader characterization of user costs significantly increases optimal frequencies, and alters key design variables of a transit system. A possible underestimation bias in crowding measures is discussed by Li and Hensher (2013) who compare crowding levels as measured by some rail operators in Australia with results of passenger surveys. The authors find a significant gap between measured and expressed crowding levels, and posit that this gap may be caused by excessive aggregation of crowding measures with respect to the time and number of stations. Furth and Muller (2006) expand the classic analysis of the additional waiting time caused by service unreliability of Osuna and Newell (1972). The authors introduce the concept of budgeted waiting time: passengers plan their trip with an additional buffer time with respect to the average in order to mitigate the lateness risk. For short headways, the budgeted waiting time is based on
the 95-percentile value of the waiting time distribution, which is strongly dependent on service reliability. For long headways, the budgeted waiting time depends also on the low extremes of the schedule deviation distribution, and must be integrated with other disutilities such as the schedule inconvenience and synchronization cost. These passenger-focused indices of service reliability are lacking in the practice of public transit authorities, as discussed in the international survey of van Oort (2014). The studies of van Oort (2011, 2016) highlight the relevance of service reliability monetization in project appraisal. For example, van Oort (2016) estimates that 2/3 of the total benefits for a tram project in Utrecht, the Netherlands, can be attributed to improvements in service reliability with respect to an existing bus line. Moccia and Laporte (2016) devise a lower convex approximation scheme for an optimization model of a transit line with fixed demand where total cost, the sum of passenger and operator costs, is minimized. This approximation scheme allows to enrich the model of Tirachini et al. (2010) with variable stop spacing and train length over multiple periods while retaining some analytic tractability as in simpler models. Moreover, the approximation scheme facilitates numerical solution methods. Jara-Díaz et al. (2017) tackle the issue of demand imbalance between peak and off-peak periods in the optimization of a transit line. The periods may also differ in duration of the cycle time, average trip length, and vehicle speed due to congestion. In contrast to the previous analyses of Jansson (1984, 1980), Jara-Díaz et al. (2017) find both analytically and numerically a divergence between the bi-period and single-period optimal results. Jara-Díaz et al. (2020) extend the bi-period analysis of Jara-Díaz et al. (2017) to the case where two vehicle sizes may be available: a fleet of smaller vehicles operates in all periods, and a fleet of larger vehicles supports the peak operations. The simultaneous deployment of two vehicle sizes at the peak determines different dwell time at stops, and hence, to avoid different cycle times, a holding strategy must be enforced for the smaller vehicles. The authors find that the two strategies, homogenous and non-homogeneous fleets, yield very similar total costs, although the non-homogeneous fleet with holding is slightly preferable.

Microeconomic appraisal of transit technologies relies on reviews of projects for
capital cost estimates. Levine et al. (2018) review Bus Rapid Transit (BRT) and Light Rail Transit (LRT) projects in the US and present capital cost comparisons based on the type of RoW and service performances. The chosen performance index is the ratio of the transit operating speed versus private car. Thanks to the project cost breakdown into specific elements and the reference to service performances, the authors find smaller capital cost differences between the reviewed technologies with respect to previous estimates. One of their findings, that RoW type is a dominant factor in transit performance, substantiates the classic system analysis of Vuchic (2005). Bruun et al. (2018) explore the concept of productive capacity, another seminal idea of Vuchic (2005), and provide disaggregated cost estimates of transit systems for both capital and O&M costs. As in Levine et al. (2018), capital cost breakdown for RoW, special civil work, and other infrastructure elements define a more complex comparison of transit technologies with respect to simpler approaches.

Moccia et al. (2018) improve the model of Moccia and Laporte (2016) by new aggregate representations of the temporal and spatial variability of demand, and by refinements of the cost functions for both the operator and the passengers. This new model remains solvable by the same solution approach of Moccia and Laporte (2016), a lower convex approximation scheme, and is applied to two scenarios for two semi-rapid modes, namely BRT and LRT. The main focus of Moccia et al. (2018) is on cost assessments of components and design aspects yielding the best performances under realistic conditions for BRT and LRT. By so doing, two common biases are avoided: capital and O&M cost estimates are disaggregated per functional class; and the performance of the offered service remain comparable, avoiding misleading comparisons from the passengers’ point of view. Moccia et al. (2020) present an optimization model for the technology selection and design of a transit line where the demand is spatially disaggregated by segments, and compare this new model with a revised version of Moccia et al. (2018). The new model leads to a revised crowding penalty function of the spatially aggregated model with substantially less crowding underestimation caused by the average bias intrinsic to an aggregated model. By using the approxi-
mation scheme of Moccia and Laporte (2016) a formula for stop spacing is found that
shares similarities with that of Wirasinghe and Ghoneim (1981) derived by a different
method, namely continuous approximation, for a non-uniform many-to-many travel
demand. Moccia et al. (2020) conclude that the revised spatially aggregated model
suffices when the main focus is technology selection.

2.2 Transit design and demand

A relevant issue when comparing transit modes across wide demand levels is the in-
terplay between transit design and captured demand. Daganzo (2012), under very
general assumptions valid not only for transportation but for public services at large,
establishes conditions for the decomposition between system design and control, and
demand forecasting and pricing. In the case of consumer surplus as a metric for the
net user benefit, the studied objective of an elastic demand model is the maximization
of the sum of the consumer surplus and of the net benefit to the rest of society (the
operator’s profit plus the balance of the externalities). The optimality conditions for
the system design and control variables of the former model are proved to be equiva-
 lent to those of the generalized system cost minimization at the optimal demand level.
Moccia et al. (2017) introduce an equivalence in the objective functions of fixed and
elastic demand models for the optimization of a transit line when the elastic demand
can be approximated by a linear function of travel time components and fare. This
finding reinforces the general result of Daganzo (2012) about the equivalence in the
optimality conditions of fixed and elastic demand models for the optimization of pub-
lic services. Moreover, Moccia et al. (2017) confute the claim of Chang and Schonfeld
(1991) about a structural difference in the optimal rules between the elastic and fixed
demand cases—the optimal rules being the same.

2.3 Automation assessments

Wadud et al. (2016) explore the energy and greenhouse gases impacts of future auto-
mated passenger and freight vehicles in the US by the classic decomposition scheme
of Schipper (2002), where the focus is on major factors on activity level, modal share, energy intensity and fuel carbon content. With respect to our line of enquiry, guide automation in urban transit, the work of Wadud et al. (2016) is significant because it highlights that guide automation in passenger vehicles can induce a reduction in transit modal share by a decrease of the generalized cost of a car trip. Fulton (2018) further explores the potential risks of guide automation in terms of increasing auto-dependency. A key factor which may induce alternative transport pathways is the supply cost of fully autonomous transportation services. On this topic, Bösch et al. (2018) provide bottom-up calculations of autonomous vehicles’ supply costs of urban and regional passenger transportation services. Several operational schemes are considered, such as dynamic ride-sharing, taxi, private, and line-based. The cost calculation of car-based services improves with respect to previous literature by including, as suggested by Litman (2015), overhead, parking, maintenance and cleaning—the latest being particularly relevant for shared schemes. For autonomous rail-based services, the supply cost is assumed to be 4.7% less than the current cost of the Swiss passenger train network. This estimate is obtained by subtracting from the total operating train cost the average salary of the driver. Average passenger occupancies, speeds, empty rides, and other operational factors are assumed as given parameters. The study shows that private cars may retain their cost attractiveness under full automation, particularly because of their low variable costs. Moreover, the authors point at other aspects beside costs that can entrench a high modal share of private car, notably the concept of the autonomous car as a private mobility robot to run errands and chauffeur tasks, and the role of car manufactures (on this aspect see also Mattioli et al. (2020), and on the insignificant role of ride-hailing service in changing car ownership patterns, see Diao et al. (2021)). With respect to the supply costs of public transit services, Bösch et al. (2018) hint at a possible role of pooled taxi schemes, although the comprehensive cost accounting highlights the critical role of occupancy. Mass transit may still retain its competitiveness in urban areas and for high demand relationships. Abe (2019) estimates generalized trip costs by autonomous buses and
taxis in several urban contexts in Japan. The main focus is on variations of operating costs with respect to human-driven vehicles with exogenous access and waiting times. Automation benefits may reduce subsidies for buses via higher labor productivity. For taxis, the potential percentage decrease in the operator cost is larger, and thus benefits may extend to users via reduced fares. Feedbacks on modal shares and total distance travelled are not considered. Tirachini and Antoniou (2020) use a total cost model of a transit line to assess scenarios of vehicle guide automation. The model considers one planning period where frequency is optimized and yields vehicle passenger capacity. The variations on capital cost, operational cost, and running speed induced by guide automation with respect to current human driven vehicles are synthetically described by three parameters. The authors acknowledge that there is a significant uncertainty on the future realization of these parameters, and hence investigate scenarios for different socio-economic contexts, namely cities representative of developed and developing countries. Under several scenarios, guide automation increases optimal frequency and reduces optimal vehicle capacity, subsidy, and degree of economies of scale.

Pinto et al. (2020) develop a bilevel optimization-simulation model to assess the possible role of a fleet of shared AVs in a large-scale multimodal transit system. The upper-level model sets the AV fleet size and the frequencies of transit patterns, which are all-stop lines and, for some lines, their short-turn variations. The patterns of some bus lines, those with low demand, can be set to zero frequency, i.e. disabled. By so doing, the upper-level model can reallocate their budget to other transit or to the AV fleet. The service of mass rapid transit lines is ensured by minimum frequency constraints. At the upper-level model, the objective function to be minimized is the sum of passengers’ waiting times and in-vehicle crowding. The transit agency point of view is represented in the upper-level model by two constraints on the capital and operational budgets. The lower-level model decides the mode choice and assigns the traffic flows by agent-based simulation and is based on the previous work of Zhang et al. (2011); Verbas et al. (2015, 2016a,b); Pinto et al. (2018). The bilevel model is solved iteratively until convergence is reached. Congestion effects of the AV fleet are not investi-
gated. Computational experiments are executed on demand data of a Chicago suburb (Evanston), but the entire Chicago transit network is represented to consider in-vehicle congestion. The results hint at a potential trade-off between the service improvement to the AV passengers and the service reduction to the transit users.

Chen and Nie (2017b) study two types of e-hailing service integrated with fixed-route mass transit in an idealized grid network where passengers’ origins and destinations are uniformly and independently distributed, as in Daganzo (2010). In the zone-based type the e-hailing vehicles serve passengers in a zone centered at a transit station. In the line-based type the e-hailing vehicles operate along a fixed-route transit line, but they can deviate from it to pickup and dropoff passengers at other places than transit stations. The zone-based model is an extension of Aldaihiani et al. (2004), and the line-based model derives from Chen and Nie (2017a). The joint optimization of the e-hailing and fixed-route transit services is modeled by continuous approximation, and the numerical solutions are validated by a simulation that includes some realistic operational details not fully captured by the model. Computational experiments show that the line-based e-hailing outperforms the zone-based one in terms of both operator and user costs. The transit system with line-based e-hailing offers a sparser fixed-route network, and hence a longer spacing between stations, but higher frequencies. However, the authors point to a limitation of their set of assumptions, namely that ride-sharing is allowed in the line-based but not in the zone-based e-hailing, as a possible cause of the zone-based under-performance.

Zhang et al. (2019) study a trunk-and-branches transit network with semi- and full-autonomous buses. Semi-autonomous buses operate as conventional buses in the network branches, and cooperate as a platoon in the trunk. This allows labor saving in the trunk operation where only one driver for one platoon is required. Full-autonomous buses operate driverless in the whole network. The continuous approximation model considers bus speed as an exogenous parameter which varies by technology. Stop spacing optimization, demand variations by periods, and in-vehicle crowding are not considered. Although the model allows different demand levels between the trunk
and the branches, the demand is assumed to follow a continuous uniform distribution along the trunk and branches. The results point at the bus speed under automation as a critical issue. Whenever the full-autonomous buses do not reach the average commercial speed of 12 km/h, their value, as measured by the sum of the operator and user costs, is insufficient with respect to conventional transit. Semi-autonomous buses may present trade-offs between users: those traveling in the trunk experience higher waiting times due to platooning. Scenarios of total cost advantage with semi-autonomous buses may occur with low demand in a long trunk and high demand in the branches, i.e., an interurban and not an urban setting.

Fielbaum et al. (2017) review topological representations of urban transit networks and propose a model where by varying few parameters, the spectrum of monocentric, polycentric and dispersed urban forms can be synthetically described. The authors highlight that simpler transit network representations such as the radial and grid structures do not capture characteristics of real cities where there are hierarchies of streets and multiple centers. The proposed network model can yield topological indices closer to those of real cities with respect to the simpler structures. Fielbaum et al. (2016) compare the user and operator costs of four transit line structures, namely direct, exclusive, hub-and-spoke, and feeder-trunk, on the parametric network model of Fielbaum et al. (2017). Their study shows the critical role of the transfer penalty parameter. The significance of the transfer penalty is confirmed in the work of Fielbaum et al. (2018) who refines the line structures of Fielbaum et al. (2016) by using several transit design heuristics. Fielbaum (2020) studies the transit line structures of Fielbaum et al. (2016) with cost parameters under full guide automation. Computational results indicate a potential cost reduction of 10% in the trunk sections, and the relative advantage of more direct line structures with respect to conventional transit.

2.4 Proposed approach

We share with the reviewed literature the concern that an exclusive focus on the novelty of guide automation in passenger transport may obfuscate potential negative ef-
fects of auto-dependency, and contribute to postpone needed investments in urban transit. However, we posit that these potential negative effects may be averted by more comprehensive techno-economic assessments of transit performances: shared, electric, and automated passenger transport is already feasible in cities that can build an automated metro. High-quality urban transit may also be planned where automated metros are unfeasible, for example by conventional semirapid rail.

Within this perspective, we aim to improve the state-of-the-art of techno-economic assessment of urban transit. Our work is different from the reviewed literature in three main ways. First, we detail existing transit technologies, and by so doing the cost uncertainties of automation are substantially reduced. We note that the deployment of full guide automation for cars and buses is highly uncertain. We deal with a specific automatic transit technology, automated metro, where cost and performance uncertainties are reduced by the significant number of its real-world examples. Second, our model endogenously computes operational and strategic characteristics from basic technological parameters. For example, average occupancy is an output of our model and not an input. Similarly, our model computes dwell times, average commercial speed, stop spacing, and other performance measures. Third, we compute both operator and passenger costs with considerably more features than in the previous literature, namely multiple periods, stop spacing, in-vehicle passenger crowding, and variable train length. This total cost approach enables the endogenous computation of the factors mentioned above, and delivers a more holistic view. In particular, endogenous stop spacing allows to explore trade-offs related to the number of stations. However, we recognize that mode choices and the spatial distribution of travel demand require broader frameworks to be assessed.

With respect to previous studies contributed by the first author on transit line optimization (Moccia and Laporte, 2016; Moccia et al., 2017, 2018, 2020) here we provide an extension to rapid transit. This extension induced two advances with respect to previous work. First, a model refinement for the access to and egress from platforms, because this aspect can be significantly different between rapid and semirapid tran-
sit (Section 3). Second, and foremost, an increased detail on the interplay between demand scenarios, transit technologies, and the type of civil construction for the infrastructure (Section 4). We think that this aspect, the relevance of the alignment, is not well recognized when comparing transit modes.

3 Mathematical model

The following model, derived from Moccia et al. (2018, 2020), is adapted to answer the research question of this paper: the comparison between semirapid rail and automated metro lines. These two modes may entail significantly different access and egress times to platforms. Metro lines often exhibit elevated or underground platforms, whereas this feature is less common in light rail lines. When comparing light rail to technologies with similar platform configurations, such as bus rapid transit, the passenger travel time component related to the platform type can be disregarded. This assumption would be misleading for the topic of this paper and we correct it by adding a specific term in the access and egress time computation.

The resulting model is presented in full for the sake of completeness in the remainder of this section which is structured as follows. Section 3.1 and 3.2 introduce the notation, and the main assumptions on the demand, respectively. The optimized variables are defined in Section 3.3. The cycle time is detailed in Section 3.4. The passengers’ time value and the operator cost, which are treated separately and then combined into a total cost, are described in Section 3.5 and Section 3.6, respectively. The side constraints are stated in Section 3.7, and the resulting optimization model is delineated in Section 3.8.

3.1 Notation

We denote as transit unit (TU), see Vuchic (2005), a set of \( n \) physically linked vehicles traveling together. For brevity, we denote the number of vehicles of a TU as the \( TU \) length. Table 1 lists the symbols that express variables. Table 2 summarizes the main
symbols used as parameters, indices, and auxiliary functions. Greek letters are specific to parameters and functions that are dimensionless. Additional symbols are derived as explained below. The subscripts min and max specify bounds of a parameter or of a variable. As will be detailed in the remainder of this Section, we discretize the service time in periods indexed by \( p \). Parameters and variables are thus derived by using this index as subscript.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>( d )</td>
<td>Distance between stops</td>
<td>km</td>
</tr>
<tr>
<td>( f )</td>
<td>Frequency in a generic period</td>
<td></td>
</tr>
</tbody>
</table>
| \( \mathbf{f} \) | Vector of \( p \) frequencies | |}

Table 1: List of symbols and their units of measure used as variables.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f )</td>
<td>Frequency in a generic period</td>
<td></td>
</tr>
<tr>
<td>( \mathbf{f} )</td>
<td>Vector of ( p ) frequencies</td>
<td></td>
</tr>
<tr>
<td>( n )</td>
<td>Number of vehicles per TU, also denoted as TU length, in a generic period</td>
<td>veh</td>
</tr>
<tr>
<td>( \mathbf{n} )</td>
<td>Vector of ( p ) TU lengths</td>
<td>veh</td>
</tr>
</tbody>
</table>

Table 2: List of primary symbols, and units of measure used in the formulae.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \bar{a} )</td>
<td>Average acceleration rate of a TU</td>
<td>m/s²</td>
</tr>
<tr>
<td>( \bar{b} )</td>
<td>Average deceleration rate of a TU</td>
<td>m/s²</td>
</tr>
<tr>
<td>( B )</td>
<td>Deployed fleet of TUs</td>
<td></td>
</tr>
<tr>
<td>( c_{OR} )</td>
<td>Fixed operator cost related to the transit route</td>
<td>$/h</td>
</tr>
<tr>
<td>( c_{Os} )</td>
<td>Fixed operator cost related to a stop</td>
<td>$/h</td>
</tr>
<tr>
<td>( c_{Osv} )</td>
<td>Operator cost related to a stop per extra vehicle</td>
<td>$/veh-h</td>
</tr>
<tr>
<td>( c_{1f} )</td>
<td>Unit operator cost per TU-hour</td>
<td>$/TU-h</td>
</tr>
<tr>
<td>( c_{1v} )</td>
<td>Unit operator cost per vehicle-hour</td>
<td>$/veh-h</td>
</tr>
<tr>
<td>( c_{2v} )</td>
<td>Unit operator cost per veh-km</td>
<td>$/veh-km</td>
</tr>
<tr>
<td>( C_{a} )</td>
<td>Access and egress time value</td>
<td>$/h</td>
</tr>
<tr>
<td>( C_{o} )</td>
<td>Operator cost</td>
<td>$/h</td>
</tr>
<tr>
<td>( C_{u} )</td>
<td>Passengers’ time value</td>
<td>$/h</td>
</tr>
<tr>
<td>( C_{tot} )</td>
<td>Total cost, sum of ( C_{u} ) and ( C_{o} )</td>
<td>$/h</td>
</tr>
<tr>
<td>( c_{vh} )</td>
<td>Coefficient of variation of the headway</td>
<td>-</td>
</tr>
<tr>
<td>( C_{v} )</td>
<td>In-vehicle time value</td>
<td>$/h</td>
</tr>
<tr>
<td>( C_{w} )</td>
<td>Waiting time value</td>
<td>$/h</td>
</tr>
<tr>
<td>( f_{t} )</td>
<td>Threshold frequency for timetable behavior</td>
<td></td>
</tr>
<tr>
<td>( \dot{f} )</td>
<td>Threshold frequency for the high frequency penalty</td>
<td></td>
</tr>
<tr>
<td>( f_{max} )</td>
<td>Upper bound on the maximum frequency</td>
<td></td>
</tr>
<tr>
<td>( H )</td>
<td>Service hours per year</td>
<td>h/year</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$</td>
<td>Capacity of a vehicle</td>
<td>pax/veh</td>
</tr>
<tr>
<td>$K$</td>
<td>Capacity of a TU</td>
<td>pax/TU</td>
</tr>
<tr>
<td>$l$</td>
<td>Average trip length</td>
<td>km</td>
</tr>
<tr>
<td>$L$</td>
<td>Length of the route</td>
<td>km</td>
</tr>
<tr>
<td>$n$</td>
<td>Number of vehicles per TU</td>
<td>veh</td>
</tr>
<tr>
<td>$p$</td>
<td>Index of a period</td>
<td>-</td>
</tr>
<tr>
<td>$\hat{p}$</td>
<td>Number of periods</td>
<td>-</td>
</tr>
<tr>
<td>$q$</td>
<td>Average bidirectional demand at the peak period</td>
<td>pax/h</td>
</tr>
<tr>
<td>$R$</td>
<td>Running time</td>
<td>h</td>
</tr>
<tr>
<td>$s$</td>
<td>Access and egress speed</td>
<td>km/h</td>
</tr>
<tr>
<td>$S$</td>
<td>Commercial speed of the TU</td>
<td>km/h</td>
</tr>
<tr>
<td>$S_{\text{max}}$</td>
<td>Maximum allowed speed of the TU</td>
<td>km/h</td>
</tr>
<tr>
<td>$S_{\text{run}}$</td>
<td>Running speed of the TU excluding stop service</td>
<td>km/h</td>
</tr>
<tr>
<td>$t_a$</td>
<td>Average access and egress time of a user</td>
<td>h</td>
</tr>
<tr>
<td>$T_a$</td>
<td>Time loss caused by acceleration and deceleration phases</td>
<td>h</td>
</tr>
<tr>
<td>$t_b$</td>
<td>Boarding and alighting time per user and vehicle</td>
<td>s/pax-veh</td>
</tr>
<tr>
<td>$T_b$</td>
<td>Boarding and alighting time per user and TU</td>
<td>s/pax-TU</td>
</tr>
<tr>
<td>$t_c$</td>
<td>Commercial cycle time</td>
<td>h</td>
</tr>
<tr>
<td>$t_{oc}$</td>
<td>Operating cycle time</td>
<td>h</td>
</tr>
<tr>
<td>$t_d$</td>
<td>Time loss caused by opening and closing of doors</td>
<td>s</td>
</tr>
<tr>
<td>$t_e$</td>
<td>Average dwell time at stops</td>
<td>s</td>
</tr>
<tr>
<td>$t_{eb}$</td>
<td>Fixed stop clearance time</td>
<td>s</td>
</tr>
<tr>
<td>$t_{ev}$</td>
<td>Stop clearance time for an extra vehicle length</td>
<td>s/veh</td>
</tr>
<tr>
<td>$t_g$</td>
<td>Access and egress time to underground platform for a user</td>
<td>s</td>
</tr>
<tr>
<td>$T_i$</td>
<td>Time loss caused by acceleration, deceleration, and door operations</td>
<td>h</td>
</tr>
<tr>
<td>$t_{if}$</td>
<td>Fixed component of the terminal time</td>
<td>s</td>
</tr>
<tr>
<td>$t_{iv}$</td>
<td>Component of the terminal time variable with the TU length</td>
<td>s</td>
</tr>
<tr>
<td>TU</td>
<td>Transit unit</td>
<td>-</td>
</tr>
<tr>
<td>$t_u$</td>
<td>Average time lost at intersections per unit distance</td>
<td>min/km</td>
</tr>
<tr>
<td>$t_v$</td>
<td>Average in-vehicle time of a user</td>
<td>h</td>
</tr>
<tr>
<td>$t_w$</td>
<td>Average waiting time of a user</td>
<td>h</td>
</tr>
<tr>
<td>$V$</td>
<td>Unit value of the in-vehicle time</td>
<td>$$/pax-h$$</td>
</tr>
<tr>
<td>$w$</td>
<td>Waiting time at a stop when $f &lt; f_t$</td>
<td>min</td>
</tr>
<tr>
<td>$y$</td>
<td>One-stage technical life</td>
<td>year</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>Fraction of the hourly demand in the most loaded segment of the line</td>
<td>-</td>
</tr>
<tr>
<td>$\beta$</td>
<td>Multiplicative factor of the operating cycle time</td>
<td>-</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>Ratio of the period demand to the peak demand</td>
<td>-</td>
</tr>
<tr>
<td>$\delta$</td>
<td>Crowding penalty function wrt the instantaneous occupancy rate</td>
<td>-</td>
</tr>
<tr>
<td>$\Delta$</td>
<td>Crowding penalty function wrt the average occupancy rate</td>
<td>-</td>
</tr>
<tr>
<td>$\epsilon$</td>
<td>Rate of the average waiting time to the headway</td>
<td>-</td>
</tr>
<tr>
<td>$\zeta$</td>
<td>Spare capacity factor for the fleet</td>
<td>-</td>
</tr>
<tr>
<td>$\eta$</td>
<td>Fraction of the longest dwell time in the maximum frequency formula</td>
<td>-</td>
</tr>
</tbody>
</table>
### Table 2 – Continued from previous page

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>θ</td>
<td>Vehicle occupancy rate</td>
<td>-</td>
</tr>
<tr>
<td>ε</td>
<td>Discount rate</td>
<td>-</td>
</tr>
<tr>
<td>λ</td>
<td>Ratio of the average trip length to the length of the route</td>
<td>-</td>
</tr>
<tr>
<td>Λ</td>
<td>Ratio of the land capital cost to the route and stations construction cost</td>
<td>-</td>
</tr>
<tr>
<td>μ</td>
<td>Discount factor of the waiting time under timetable behavior</td>
<td>-</td>
</tr>
<tr>
<td>ν</td>
<td>Spare capacity factor for the TU</td>
<td>-</td>
</tr>
<tr>
<td>Ξ</td>
<td>Ratio of the residual value to the initial value of a capital component</td>
<td>-</td>
</tr>
<tr>
<td>ρ</td>
<td>Slope of the linear part of the crowding penalty function δ</td>
<td>-</td>
</tr>
<tr>
<td>τ</td>
<td>Ratio of the maximum to the average period demand</td>
<td>-</td>
</tr>
<tr>
<td>νₐ</td>
<td>Multiplicative factor of V for the access and egress time</td>
<td>-</td>
</tr>
<tr>
<td>νₐₑ</td>
<td>Multiplicative factor of V for the waiting time</td>
<td>-</td>
</tr>
<tr>
<td>φ</td>
<td>Ratio of the maximum to the average vehicle occupancy rate</td>
<td>-</td>
</tr>
<tr>
<td>ψ</td>
<td>Ratio of the maximum to the average dwell time</td>
<td>-</td>
</tr>
<tr>
<td>χ</td>
<td>Ratio of the period hours to the total service hours</td>
<td>-</td>
</tr>
<tr>
<td>ω₁₂</td>
<td>Parameters of the high frequency penalty</td>
<td>-</td>
</tr>
</tbody>
</table>

#### 3.2 Assumptions on the demand

We assume that a bidirectional transit line operates on a route, where by “route” we designate the physical alignment and the infrastructure, and by “line” we refer to the service operated. The route length is \( L \) and the line length is \( 2L \), i.e. a complete round-trip. We first describe how the demand is temporally distributed, and then we introduce the assumptions on the spatial aspects. The transit line serves passengers for \( \hat{p} \) periods in the operating year. The average bidirectional hourly demand (boardings) \( q_p \) in a period is equal to \( q \gamma_p \), where \( p \) is the period index, \( \gamma_p \) is a positive parameter not exceeding one, and \( q \) is the average hourly demand during the peak period. Without loss of generality we index the periods in non-increasing order of demand, the peak period is indicated by \( p = 1 \) and hence \( \gamma_1 = 1 \). For each period the maximum demand is \( q_p \tau_p \), where \( \tau_p \) is a parameter equal to or larger than one. The ratio of a period’s hours to the total service hours is denoted by \( \chi_p \), and \( \sum_p \chi_p = 1 \). These demand parameters must reflect the total service hours in a year, indicated by \( H \).

Spatially, we describe the demand by two parameters. We indicate by \( \alpha \) the fraction
of the period demand at the most loaded segment, where by "segment" we refer to a 
continuous part of the line, and thus a segment implies a direction. This parameter 
is related to what is usually called "bottleneck capacity", measured in passengers per 
hour and direction (pphd). The parameter \( \lambda \) controls the amount of travel served, and 
is the ratio of the average trip length to the route length. Thus, the average length of 
passenger trip, \( l \), is equal to \( \lambda L \). In general, these two spatial parameters, \( \alpha \) and \( \lambda \), 
may vary by period. However, there is a lack of real-world data on these aspects and 
 obtaining yearly averages is already challenging. This is why we assume that these 
two parameters are constant, but we acknowledge that this is a limitation.

3.3 Variables

The stop spacing is an optimized continuous variable and is denoted by \( d \). We denote 
by \( n_p \) the number of vehicles of a TU in period \( p \), which is an integer variable that 
we optimize. The frequency \( f \) is expressed as the number of TUs per hour. There 
is a frequency for each period, namely \( f_p \), which is also a continuous variable that is 
optimized.

3.4 Cycle time

We distinguish between operating and commercial cycle times. The operating cycle time 
is the sum of the running time between stations, including acceleration and decelera-
tion, of the time lost at intersections, and of the dwell time for boarding and alighting. 
The commercial cycle time is determined as a function of the operating cycle time plus 
some time components related to providing of variability in running times and opera-
tions at terminals.

More formally, we model the operating cycle time as the sum of four terms de-
scribed in the following. We assume that a TU accelerates up to and decelerates down 
from a maximum allowable speed \( S_{\text{max}} \). For a mode without an exclusive right-of-
way, the TU loses an average \( t_u \) minutes per km. This time loss occurs mainly at inter-
sections and diminishes with higher investments in runningway improvements and
traffic signal priority (TSP) systems. For a mode with exclusive right-of-way, such as automated metro, $t_u$ is set to zero. In general, the average speed excluding user service at stops, $S_{run}$, is

$$S_{run} = \frac{1}{\frac{1}{S_{max}} + \frac{t_u}{60}} \quad S_{max}[\text{km/h}], t_u[\text{min/km}]. \quad (1)$$

The resulting running time, denoted $R$, is equal to $2L/S_{run}$ and is the first term of the operating cycle time. We assume that on average a TU leaves a stop accelerating up to $S_{run}$, travels at this speed, and then decelerates to halt at the next stop. Let $\bar{a}$ and $\bar{b}$ be the average acceleration and deceleration rates of a TU. The incremental time loss caused by the acceleration and deceleration phases is denoted by $T_a$, and is equal to

$$T_a = \frac{S_{run}}{25920} \left( \frac{1}{\bar{a}} + \frac{1}{\bar{b}} \right) \quad S_{run}[\text{km/h}], \bar{a}, \bar{b}[\text{m/s}^2], \quad (2)$$

(see e.g. Vuchic and Newell, 1968). We add to the standing time a fixed component $t_d$, which accounts for the opening and closing of the doors, and we denote by $T_i$ the lost time for acceleration, deceleration, and door opening and closing:

$$T_i = T_a + \frac{t_d}{3600} \quad T_a[\text{h}], t_d[\text{s}]. \quad (3)$$

In a round-trip, the number of times when the phenomena counted by $T_i$ occur is equal to $2L/d$, henceforth the second term of the operating cycle time is $2LT_i/d$. The third term of the operating cycle time expresses the load-dependent dwell time which is related to $T_{br}$ the boarding and alighting time per user of a TU, and the number of passengers using a TU, given by $q/f$. The boarding and alighting time of a TU depends on the number $n$ of vehicles per TU and their door configuration. We assume the boarding and alighting service time per user of a vehicle to be equal to $t_b$, hence $T_{br} = t_b/n$. We introduce a fourth term of the operating cycle time accounting for extra delays at intersections, links, and stops under high frequencies, where for high frequencies we mean those exceeding a threshold frequency $\hat{f}$, for example 25 TU/h. When this frequency
is reached, the design TSP may underperform, and the interactions between signals and stops, as well as the disturbances induced by trespassing may become significant.

This term depends on the ratio of the frequency to this threshold frequency raised to the power $\omega_2$, and multiplied by a coefficient $\omega_1$ to the base operating cycle time loss at intersections, which is equal to $2t_uL$. This term was used in Moccia et al. (2018) and derived from simulations of operations in reserved lanes for the VIVA semirapid transit in York Region, Ontario, Canada (Allen, 2016). The exponential formulation was adapted from the historic US Bureau of Public Roads formula expressed in terms of a volume to capacity ratio. The intuition behind this formulation is that, as for general vehicular traffic, the effects may be negligible for a considerable range, then may rapidly increase with volume once the effects become noticeable, e.g. as the likelihood of vehicles queuing behind one another grows.

The operating cycle time $t_{oc,p}$ for each period is

$$
t_{oc,p}(f_p, d, n_p) = R + \frac{2L}{d} T_i + \frac{t_b}{3600} \frac{q_p}{n_p f_p} + \omega_1 \frac{2t_u L}{60} \left( \frac{f_p}{f} \right)^{\omega_2}
$$

$$
t_b[s-pax/veh], q_p[pax/h], T_i, R[h], t_u[mi/min/km], d, L[km],
$$

$$
n_p[veh/TU], f_p, f[TU/h].
$$

The commercial cycle time is obtained by adding the following three terms. The first term is obtained by multiplying the operating cycle time by $\beta$, a parameter larger than one. The parameter $\beta$ accounts for the schedule time recovery at the terminals in one cycle, often referred to as "padding" time. The sum of the padding times in one day is assumed also to be sufficient for change of the TU lengths between intra-day periods, if any. The second term is the fixed terminal time $t_{tf}$ for minimum crew rest and securing vehicle keys, if any. The third term is the terminal time proportional to $t_{tv}$ and variable with the TU length for the crew walking distance between the TU ends, if any, and safety checks.
The commercial cycle time, in the following referred to as cycle time for brevity, is

\[
t_{c,p}(f_p, d, n_p) = \beta t_{oc,p}(f_p, d, n_p) + \frac{t_{ff} + t_{tv}n_p}{3600}
\]

\[
t_{oc,p}[h], n_p[\text{veh/TU}], t_{ff}[s], t_{tv}[s/\text{veh}].
\]  

(5)

### 3.5 Passengers’ time value

The passenger time value is a monetized value of time composed of three parts: access and egress, waiting, and in-vehicle time values. The in-vehicle time value under un-crowded conditions is denoted by \( V \), and in the following this is also indicated as the base-time value. Multiplicative factors larger than one of the base-time value account for the additional discomfort usually encountered by passengers in the parts of a trip outside a transit vehicle.

Users access to and egress from the nearest stop at speed \( s \) at the surface level where the average distance is \( d/4 \) at the origin and at the destination. Hence, the average total access and egress length is \( d/2 \) at the surface level. For a mode with underground or elevated stations, a fixed access and egress time of the platforms, \( t_g \) is considered. The average access and egress time \( t_a \) of a user is

\[
t_a(d) = \frac{d}{2s} + \frac{t_g}{3600} \quad d[\text{km}], s[\text{km/h}], t_g[s].
\]  

(6)

The value of one unit of access and egress time is the product of the base-time value and a parameter \( \nu_a \). The average access and egress value \( C_a \) is

\[
C_a(d) = \nu_a V \left( \frac{d}{2s} + \frac{t_g}{3600} \right) q \sum_p \gamma_p \chi_p.
\]  

(7)

The waiting time depends on the frequency and we distinguish between low and high frequencies. In the case of high frequencies, users arrive at stops at a constant rate and the average waiting time is \( \epsilon/f \). Values of \( \epsilon \) strictly larger than 1/2 can model cases where the headways have a large variance, for example by using the correction of the
waiting time formula of Osuna and Newell (1972)

\[ \epsilon = \frac{1}{2} + c_{vh}^2, \]  

(8)

where \( c_{vh} \) is the coefficient of variation of the headways. In general, this formulation would have a dependency of the coefficient of variation on the frequency, \( c_{vh}(f) \), which would introduce additional complexity. In the following we assume that \( c_{vh} \) is a fixed parameter for the technology and the range of frequencies studied, and consequently \( \epsilon \) is given. We detail in Section 4.4 the choice of this approach for the technologies studied in this paper.

In the case of low frequencies, users follow timetables and arrive at stops \( w \) minutes before the expected time of service. The waiting time at the stop saved by this behavior still represents a disutility for the user who may have to redefine his or her schedule. This disutility, often referred to as schedule delay, is discounted by a factor \( \mu \) less than one, for example \( \mu = 1/3 \), with respect to the disutility of waiting at the stop. The threshold frequency for these two behavior regimes is defined by \( f_l \), for example six TU per hour, which results in a headway of 10 minutes. The average waiting time \( t_w \) of a user is

\[
t_w(f) = \begin{cases} 
\frac{w}{60} + \frac{\epsilon}{f} & \text{if } f < f_l \\
\frac{\epsilon}{f} & \text{if } f \geq f_l, \ f \in [\text{TU/h}, w \in \text{min}]
\end{cases}.
\]  

(9)

The average value of waiting \( c_w(f) \) borne by \( q \) users at the frequency \( f \) is

\[
c_w(f) = \upsilon_w V t_w(f) q \quad V[\$/\text{pax-h}], t_w[\text{h}], q[\text{pax/h}],
\]  

(10)

where \( \upsilon_w \) is the multiplicative factor for the waiting time with respect to the base value of time. The average value of waiting for the \( \hat{p} \) periods, \( C_w \), is

\[
C_w(f) = \sum_p c_w(f_p) \gamma_p \chi_p \quad c_w[\$/\text{h}],
\]  

(11)

where \( f \) is the vector of \( \hat{p} \) frequencies.
The average in-vehicle time $t_v$ of a user is modeled as a fraction of the operating cycle time $t_{oc}$. This fraction is equal to the ratio of the average trip length $l = \lambda L$ to the total distance $2L$ covered by a TU in a cycle, and hence

$$t_v = \frac{\lambda}{2} t_{oc} [\text{h}].$$

The value of the in-vehicle travel time is multiplied by the crowding penalty function $\Delta$ introduced in Moccia et al. (2020) and explained in the following. A function $\delta$ expresses the crowding disutility with respect to the instantaneous vehicle occupancy rate $\theta$, and the penalty function $\Delta$ aims at reducing underestimation of crowding when it is synthetically applied to the average of the vehicle occupancy rate. We assume that the function $\delta$ is piecewise linear as follows: $\delta = 1$, i.e., there is no penalty up to a vehicle occupancy rate of $\theta_{min}$, and for larger values of $\theta$ the penalty increases linearly with a slope value $\rho$. The threshold $\theta_{min}$ for no penalty can be selected according to the seat configuration and the prevailing type of trip. For a long commuter trip the passenger can experience discomfort when she has no other choice than taking a seat adjacent to an already occupied seat, i.e., when circa half of the seats are occupied. For a short urban trip, the discomfort can arise from standing, which may occur when the seats are nearly full. We note that seat configuration, i.e., the number of seats per square meter, may change, and this is why we express the crowding disutility in terms of $\theta$, the fraction of the maximum capacity.

Figure 2 provides an example of this penalty function with $\theta_{min} = 0.3$, a value appropriate for urban transit where one-third of the vehicle capacity is usually offered as seats, and standing is common. Moccia et al. (2017, 2018) further discuss the relevant literature that supports this approach for in-vehicle passenger crowding.

Formally, this disutility function is (for notational compactness in the following two
Figure 2: Example of the instantaneous crowding disutility function with $\rho = 1$ and $\theta_{\text{min}} = 0.3$.

The crowding penalty function $\Delta$ is computed on the average occupancy rate $\bar{\theta}$. At each period, the average occupancy rate $\bar{\theta}$ depends on the frequency and on the TU length. The average occupancy rate is

$$\bar{\theta}(f, n) = \frac{lk}{2Lknf},$$

where $k$ is the passenger capacity of a vehicle, and hence the capacity $K$ of a multi-unit TU is equal to $k \times n$. The penalty function $\Delta$ is reported in Table 3, where the parameter $\phi$ is such that $\phi \bar{\theta}$ is the maximum value of the vehicle occupancy rate. We denote by
Table 3: Formula of the crowding penalty.

\[
\Delta(\bar{\theta}) = \begin{cases} 
1 + \rho \left( \frac{\phi^2 \phi \bar{\theta} - \theta_{\min}}{2(\phi - 1)} + \frac{(\phi \bar{\theta} - \theta_{\min})^3}{6\theta^2(\phi - 1)} - \phi \frac{(\phi \bar{\theta} - \theta_{\min})^2}{2\theta(\phi - 1)} - \theta_{\min} \frac{(\phi \bar{\theta})^2 - \theta^2_{\min}}{4\theta^2(\phi - 1)} \right) & \text{if } \bar{\theta}(2 - \phi) < \theta_{\min} < \phi \bar{\theta}, \phi \leq 2 \\
1 + \rho \left( \phi^2 \bar{\theta} + \frac{4\bar{\theta}(\phi - 1)^2}{3} - 2\phi \bar{\theta}(\phi - 1) - \theta_{\min} \right) & \text{if } \theta_{\min} \leq \bar{\theta}(2 - \phi), \phi \leq 2 \\
1 + \rho \left( 2(\phi \bar{\theta} - \theta_{\min}) + \frac{2(\phi \bar{\theta} - \theta_{\min})^3}{3(\phi \bar{\theta})^2} - \frac{2(\phi \bar{\theta} - \theta_{\min})^2}{\phi \bar{\theta}} - \theta_{\min} \frac{(\phi \bar{\theta})^2 - \theta^2_{\min}}{(\phi \bar{\theta})^2} \right) & \text{if } \theta_{\min} < \phi \bar{\theta}, \phi > 2 \\
1 & \text{if } \theta_{\min} \geq \phi \bar{\theta}
\end{cases}
\]
\( \Delta p \) the resulting \( \hat{p} \) penalty functions.

The value of in-vehicle time, \( C_v \), is

\[
C_v(f, d, n) = V \lambda^q \sum_p \chi_p \gamma_p \Delta_p(\hat{\theta}_p(f_p, n_p)) t_{oc,p}(f_p, d, n_p),
\]

(15)

where \( n \) is the vector of \( \hat{p} \) TU lengths. The total passengers’ time value \( C_u \) is then

\[
C_u = C_a + C_w + C_v \quad C_a, C_w, C_v \quad [\$/h].
\]

(16)

### 3.6 Operator cost

The operator cost consists of six components. The first component is the construction and maintenance of the route and is denoted by \( c_0r \). The second and third components are related to the construction and maintenance costs of the stations. For each station these costs can be decomposed into a fixed part \( c_{0s} \) and a variable part \( c_{0sv} \), which depends on the TU length at peak hours. There are \( 2 + 2L/d \) one-way stops, including the terminal. The second component is related to the terminal, and the third to the other stops which depends on the stop spacings. The fourth component depends on the fleet size and reflects vehicle capital and administrative costs. Let \( c_{1v} \) be the unit operator cost per vehicle-hour which accounts for the capital and administrative costs. The deployed fleet size \( B_1 \) of TUs is the product of frequency and cycle time at the peak period: \( B_1 = f_1 t_{c,1} \). The vehicle fleet size is equal to \( \zeta n_1 B_1 \), where \( \zeta > 1 \) accounts for operation and maintenance (O&M) spares. The fifth component expresses the crew costs, if any, and depends on \( c_{1t} \), the unit operator cost per TU-hour. The sixth component accounts for running costs such as energy, tires or wheelsets, lubricants, etc. Let \( c_{2v} \) be the unit operator cost per vehicle-km. The amount of vehicle-km is the product of the commercial speed \( S \) and the fleet size in a period. The commercial speed is obtained by dividing the total length \( 2L \) by the cycle time. Thus, the amount of vehicle-km is
\[ S \times nB = 2L/t_c \times nft_c = 2Lnf. \]  

The operator cost \( C_o \) is then

\[
C_o(f, d, n) = c_{0r} + 2(c_{0s} + c_{0sv}(n_1 - 1)) + (c_{0s} + c_{0sv}(n_1 - 1)) \frac{2L}{d} + c_{1v} \zeta n_1 f_1 t_{c,1}(f_1, d, n_1) + c_{1t} \sum_p \chi_p f_p t_{c,p}(f_p, d, n_p) + 2c_{2v} L \sum_p \chi_p n_p f_p. \tag{17}
\]

### 3.7 Side constraints

The frequency is constrained to lie between \( f_{\text{min}} \) and \( f_{\text{max}} \). The value \( f_{\text{min}} \) must account for capacity as follows. Let \( \alpha q \tau \) be the largest load served by the line in a generic period, where \( \alpha \leq 1 \), and \( \nu \) be a spare capacity design factor. For example, a value of \( \nu \) smaller than one accounts for random demand fluctuations and represents a safety margin, whereas \( \nu \) larger than one allows crush loading. However, at low demand levels this \( f_{\text{min}} \) computation may yield values that are considered as unacceptable by decision makers and users. Under these circumstances \( f_{\text{min}} \) can be set by a “policy headway” rationale, i.e., there is a minimum guaranteed frequency \( f_{\text{pol}} \). Thus, \( f_{\text{min}} \) is

\[
f_{\text{min}} = \max \left( f_{\text{pol}}, \frac{\alpha q \tau}{\nu k_n} \right) \quad f_{\text{pol}}[\text{TU/h}], \quad q[\text{pax/h}], \quad k[\text{pax/veh}], \quad n[\text{veh/TU}]. \tag{18}
\]

The \( f_{\text{max}} \) is defined according to general principles from the Transit Capacity and Quality of Service Manual (TCQSM, 2013). We set an upper bound \( \tilde{f}_{\text{max}} \) on the value of \( f_{\text{max}} \) to reflect rail vehicle operation with safe separation. TCQSM (2013) effectively defines \( f_{\text{max}} \) as the maximum number of transit vehicles that can pass a given location in a given time period, and is inversely proportional to a minimum achievable headway under which acceptable service can be maintained. The minimum achievable headway depends on the longest dwell time, where by dwell time we consider the time at a service station with wheels stopped. We assume that the longest dwell time is \( \psi \) times the average dwell time, where \( \psi \) is a parameter larger than one. The average dwell
time $t_e$ is a function of frequency, stop spacing, and TU length:

$$
t_e(f, d, n) = t_d + t_b \frac{qd}{2n f L} t_d[s], t_b[s-pax/veh], q[pax/h],
$$

$$
d, L[km], n[veh/TU], f[TU/h].
$$

(19)

The $f_{max}$ formula is

$$
f_{max}(f, d, n) = \min\left(\frac{3600}{t_{e0} + t_{ev}(n - 1) + \eta \psi t_e}, f_{max}[TU/h],
$$

$$
t_{e0}, t_{ev}, t_e[s],
$$

(20)

where the dependency of $f_{max}$ on $f, d,$ and $n$ occurs through $t_e,$ because the achievable minimum headway is limited by the longest dwell time $\psi t_e.$ The values of $t_{e0},$ the fixed stop clearance time, $t_{ev},$ the stop clearance time for an extra TU length, and $\eta,$ the fraction of the longest dwell time, are set in accordance with the vehicle characteristics we define for each scenario, and with the assumed station and runningway characteristics for which capital cost estimates are established.

Finally, we impose side constraints on the stop spacing. Reaching the speed $S_{max}$ requires a stop spacing larger than a threshold value $d_{min},$ which depends on acceleration and deceleration rates

$$
d_{min} = \frac{S_{max}^2}{25920} \left(\frac{1}{\bar{a}} + \frac{1}{\bar{b}}\right) S_{max}[km/h], \bar{a}, \bar{b}[m/s^2],
$$

(21)

(see e.g. Vuchic and Newell, 1968). Values of $d$ less than this minimum distance will not be allowed. We observe that $d_{min}$ will be different for types of vehicles and speed limits, and is a constraint that is rarely binding, i.e., the solutions that we obtain present stop spacings larger than $d_{min}. An upper bound $d_{max}$ is also defined so as to keep results in a range representative of urban semirapid or rapid transit rather than regional systems. In general, this upper bound may be required by other policy criteria.
3.8 Optimization model

The total cost $C_{tot}$, sum of passengers’ time value and operator cost, is a function of frequencies, stop spacing, and TU lengths. The model follows:

$$\text{minimize } C_{tot}(f, d, n)$$ (22)

subject to

$$d_{\text{min}} \leq d \leq d_{\text{max}}$$ (23)

$$\max\left(f_{\text{pol},p}, \frac{\alpha q_p T_p}{\nu n_p k}\right) \leq f_p \leq f_{\text{max},p}(f_p, d, n_p), \forall p \in \{1, \ldots, \hat{p}\}$$ (24)

$$n_{\text{min}} \leq n_p \leq n_{\text{max}}, n_p \in \mathbb{N}, \forall p \in \{1, \ldots, \hat{p}\}$$ (25)

$$n_p f_p l_{c,p}(f_p, d, n_p) \leq n_1 f_1 l_{c,1}(f_1, d, n_1), \forall p \in \{2, \ldots, \hat{p}\}.$$ (26)

Constraints (23) set minimum and maximum values for the stop spacing. Constraints (24) enforce minimum and maximum values for the frequencies. Constraints (25) specify the feasible range of TU lengths, and constraint (26) ensures that the maximum fleet is deployed at peak times, where for “maximum deployed fleet” we refer to the number of vehicles needed for the scheduled service. Without this constraint non-peak periods could call for a larger fleet that is available. This could happen because the fixed costs of the fleet, capital amortization and administrative expenses, are computed in the objective function by the variables of the peak period.

The model is solved by an updated version of the algorithm presented in Moccia and Laporte (2016) by constructing a separable lower convex envelope of the objective function. The algorithm is implemented in Python 3.7 with the L-BFGS-B solver, a quasi-Newton code for bound-constrained optimization, see Zhu et al. (1997).
4 Scenarios

We consider two urban transit modes, namely semirapid rail (SR) and automated metro (AM). By “mode” we refer to a relatively large set of specific transit implementations that we identify as “technologies”. Semirapid rail requires a right-of-way with a partial separation from other traffic (Vuchic, 2005; Vuchic et al., 2012). Automated metro is completely grade-separated from vehicular or pedestrian crossings. We exemplify the spectrum of these two modes by six specific implementations: four for the semirapid rail, and two for the automated metro. These technologies are based on an assessment of the literature (ITA, 2004; Vuchic, 2005; Vuchic et al., 2012; Casello et al., 2014; Bruun et al., 2018), and on IBI Group’s internal database of transit projects. The four semirapid rail technologies are synthesized in Section 4.1. The two automated metro technologies are described in Section 4.2. The technologies are each embedded in two scenarios motivated by opposite demand patterns that can be found in practice. These two demand patterns are presented in Section 4.3. Finally, Section 4.4 details the full set of techno-economic parameters of the six technologies in the two scenarios.

4.1 Semirapid rail

We exemplify the spectrum of semirapid rail by four specific implementations that we label as SR1, SR2, SR3, and SR4.

The runningway of SR1 consists of a dedicated arterial RoW obtained by a simple change in use of an existing lane. A photo example of the SR1 runningway is shown in Figure 3. We assume an average of six intersections per km, four of which are signalized. The traffic signal priority is assumed to be present in the relatively modest form prevalent in North America. Stations are part of the curb or sidewalk environment, with only modest passenger shelters and amenities, but allow level boarding, see Figure 4 for an example of a SR1 station. Fares are collected on board and no fare vending occurs at stations. Vehicles are single-unit low-floor trams of length 34 m. The two terminals have a loop that is not part of the service route length and vehicles are
in a single-ended configuration. The maximum authorized speed is assumed to be 50 km/h. The pictogram of Figure 5 summarizes the main SR1 features.

The runningway of SR2 consists of an exclusive at grade right-of-way where traffic signal priority is provided, Figure 6. Stations may offer multi-modal transfer and allow off-vehicle fare collection, level boarding, and other passenger amenities as detailed in Moccia et al. (2018), see Figure 7 for an example. Vehicles may operate as multiple units with a number of vehicles ranging from one to four, and the maximum authorized speed is assumed to be 75 km/h. These SR2 features are depicted in Figure 8.

We define SR3 as SR2 but with an alignment that allows for faster operations. The SR3 runningway has two sections with a maximum speed of 90 km/h: an elevated section and a at-grade section with barrier- or gate-protected highway crossings. The former represents 15% of the SR3 runningway, and the latter 60%. The remaining part of the SR3 runningway is at-grade with TSP as in the SR2 case. We assume a weighted maximum speed of SR3 equal to 86 km/h. Examples of an elevated runningway and
a station for SR3 are reported in Figures 9 and 10. A pictogram of this technology is illustrated in Figure 11.

The runningway of SR4 is specified to represent local conditions where a surface alignment is severely constrained by existing buildings and geography. In these cases the alternatives on the drawing table are either the simplest alignment as in SR1, a dedicated lane in an existing road, or an alignment with a significant underground section. We posit for SR4 an alignment mix with the three following sections: 60% underground, 15% elevated, and 25% at-grade as in SR2. The underground section allows a maximum speed of 90 km/h, and the weighted maximum speed of SR4 is equal to 86 km/h. Examples of SR4 projects are shown in Figures 12 and 13, and a pictogram of this technology is illustrated in Figure 14.

4.2 Automated metro

Grade of Automation 4 (GoA4) has become state-of-the-practice for metros or rapid transit, therefore it is appropriate to use automated metro (AM) as a benchmark for the
rapid functional class of public transport. AMs must be completely grade-separated from vehicular or pedestrian crossings, which in urban contexts often means the alignment and stations need to be elevated or underground. AMs may be built in an at-grade configuration where pedestrians and highway vehicle crossings are made by underpasses or overpass bridges; this means that an at-grade AM station may be more massive and complex than an at-grade SR station. AMs are assumed to have platform edge doors in stations to increase passenger safety; this is the prevalent practice for new metros. Based on a survey of AMs already in service, the technology used as a baseline is the conventional duorail Bombardier "Innovia 300" metro vehicle as op-
erated by the Riyadh Metro. It was assumed that trains could be as long as twelve four-axle vehicles or as short as a single vehicle; this spans the capacity range between two-car trains of small AMs to long conventional metro trains in cities like New York, so as to allow for continuous treatment of a wide range of passenger traffic densities. Maximum speed capability is assumed to be 90 km/h, a capability the vehicle manufacturer offers. The difference between the two AMs in this paper is not one of technology or performance, but rather of the mix of at-grade, elevated, and underground sections, which can depend on the particulars of costs and RoW availability in a city.
The choice of expensive elevated construction, or even more expensive underground construction, reflects both local and planning considerations. Considering the large number of AM systems already in service, two cases, labeled AM1 and AM2, are selected to represent the range of values observed for this mode. AM1 represents the “leanest” mix of construction types likely to be found in a rapid transit network: 60 percent at-grade and 40 percent elevated. AM2 represents the “most expensive” likely mix of 40 percent elevated and 60 percent underground. Stations are assumed to have the same mix of construction type as the runningway. Photographic examples of these two construction types for runningways and stations are provided in Figures 15–16 and 18–19. The two pictograms of these AM variants are shown in Figures 17 and 20, for AM1 and AM2, respectively. In Section 4.4 we provide capital and O&M estimates for these two mixes of AM construction types.

Figure 15: An example of elevated AM runningway and vehicle (Rennes, La Poterie, ©Bernard Chatreau).

Figure 16: An example of an elevated AM station (Rennes, Pontchaillou, ©Bernard Chatreau).

Figure 17: AM1 pictogram.
4.3 Demand patterns

To capture some of the wide variations in demand characteristics among cities, we consider two distinct demand distributions along the transit line. These two demand patterns induce two scenarios, in the following labeled as A and B.

Scenario A exemplifies the case of a radial line that extends between a city center or central business district and a point towards the edge of the city or in the suburbs. This type of transit line often exhibits a “commuter” type of passenger flows, that can be captured by the following three characteristics: a strong flow imbalance between the two directions, a significant passenger concentration in a section of the line, and a relatively high average trip distance. Moreover, in a commuter line there is a significant share of passengers who access the transit line by motorized, private and public, modes. Consequently, facilities such as bus berths and car parkings at the stations are more substantial than in the case of a prevalent pedestrian access.
Scenario B represents the case of a diametrical line that connects two outlying points via the city center. In this pattern the passenger flows are more evenly distributed between the two directions, the most loaded section registers a smaller share of the total demand, and the average trip distance is shorter. Stations are simpler because of the prevalent pedestrian access. We note that the demand patterns induce different scenarios not only because of the demand-specific parameters, but also because of the different capital and O&M costs of stations as detailed in Section 4.4.

4.4 Parameters

This section reports the techno-economic parameters used in the computational experiments. Monetary figures are expressed in US dollars for the year 2012. Capital amortization per hour of service is computed for infrastructure and rolling stock as

\[
\frac{P(1 - \Xi)\iota}{H(1 - (1 + \iota)^{-y})},
\]

(27)

where \(P\) is the purchase price, \(H\) is the number of service hours in a year, \(\iota\) is the discount rate, \(\Xi\) is the fraction of the residual value, and \(y\) is the one-stage technical life. The one-stage technical life is lower than a typical service lifetime because it expresses the equivalent years including the cost of a mid-life rebuild at the prevalent discount rate. Land capital cost is estimated as a fraction \(\Lambda\) of the construction cost of the route and stations. Land capital cost is annualized by multiplying it by the discount rate, i.e., assuming an infinite service life.

We list the full set of parameters in Tables 4–10 organized as follows. Table 4 reports the parameters that are constant among the scenarios and technologies. These parameters pertain to the passengers’ values such as the monetization of the travel time components and crowding, to the system configuration such as the route length, to the general economic assumptions such as the discount rate, etc. More in detail, the in-vehicle base VoT is derived from US-DOT (2011), and the access and waiting VoT multipliers, 1.25 and 1.5, respectively, are as in Tirachini et al. (2010) and Moccia and
The rate of the average waiting time to the headway $\epsilon$ is conservatively set equal to $1/2$ for all technologies because we study highly reliable modes in newly developed alignments and we do not allow for congested operations in our technical assumptions. A techno-economic assessment of transit operations under congested conditions, as is often the case for example with buses or trams in mixed traffic, is out of the scope of this analysis. The upper bound for the maximum frequency that we set, 40 TU/h, and the refinement of the maximum achievable frequency, equation (20), depending on the dwell time, equation (19), corroborates the assumption that the studied technologies perform reliably.

Table 5 lists the technical parameters that are specific to a mode such as acceleration and deceleration rates, and operational coefficients pertaining to the frequency management. Table 6 details the techno-economic parameters that are specific for each of the six technologies and are constant between the scenarios. These parameters pertain to the capital and O&M costs of the route and rolling stock, to the terminal and per stop operational aspects, to the passenger access and egress, and to the service speed computation. Table 7 lists the parameters that differentiate Scenarios A and B. Scenario A entails a higher passenger access speed, 12 km/h, with respect to that of Scenario B, 4 km/h. This is consistent with the assumption that in Scenario A we study a prevailing multimodal access whereas in Scenario B there is a prevailing pedestrian access. Scenario A deals with a longer average trip than that of Scenario B, and the passenger concentration at the critical segment of the line is as well larger in Scenario A than in B. Table 8 and Table 9 provide the economic parameters of the six technologies for Scenario A and B, respectively. These two tables detail the differences in terms of capital costs of the stations between the two demand scenarios, with stations in Scenario A requiring significantly higher construction costs than those of Scenario B because of the more complex multimodal features.

Table 10 reports the period-related parameters. We consider four periods, and the parameters derive from the same temporal distribution of the demand used in Moccia et al. (2018), the turnstile data of the Boston subway network, the difference being that
we now discretize the service time by four periods instead of three as in the previous paper. This allows automation to show its strength in terms of adaptability to demand variations. An illustration of the scaled demand curve duration, its discretization in four periods, and the period parameters is provided in Figure 21.

Table 4: Parameters related to the users and to the transit system that are common to all technologies.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Unit</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit value of in-vehicle time</td>
<td>$V$</td>
<td>$$/h$</td>
<td>13.37</td>
</tr>
<tr>
<td>Multiplicative factor for the access time</td>
<td>$\nu_a$</td>
<td>-</td>
<td>1.25</td>
</tr>
<tr>
<td>Multiplicative factor for the waiting time</td>
<td>$\nu_w$</td>
<td>-</td>
<td>1.50</td>
</tr>
<tr>
<td>Average occupancy rate up to $\delta = 1$</td>
<td>$\theta_{min}$</td>
<td>-</td>
<td>0.3</td>
</tr>
<tr>
<td>Slope of the linear part of $\delta$</td>
<td>$\rho$</td>
<td>-</td>
<td>1.0</td>
</tr>
<tr>
<td>Rate of the average waiting time to the headway</td>
<td>$\phi$</td>
<td>-</td>
<td>0.5</td>
</tr>
<tr>
<td>Discount factor of the waiting time under timetable behavior</td>
<td>$\mu$</td>
<td>-</td>
<td>0.3</td>
</tr>
<tr>
<td>Threshold frequency for timetable behavior</td>
<td>$f_t$</td>
<td>TU/h</td>
<td>6.0</td>
</tr>
<tr>
<td>Waiting time at a stop when $f &lt; f_t$</td>
<td>$w$</td>
<td>min</td>
<td>5</td>
</tr>
<tr>
<td>Route length</td>
<td>$L$</td>
<td>km</td>
<td>15.0</td>
</tr>
<tr>
<td>Number of service hours per year</td>
<td>$H$</td>
<td>h/year</td>
<td>9940</td>
</tr>
<tr>
<td>Spare capacity factor for the TU</td>
<td>$\nu$</td>
<td>-</td>
<td>0.95</td>
</tr>
<tr>
<td>Spare capacity factor for the fleet</td>
<td>$\zeta$</td>
<td>-</td>
<td>1.20</td>
</tr>
<tr>
<td>Discount rate</td>
<td>$\iota$</td>
<td>-</td>
<td>0.03</td>
</tr>
<tr>
<td>Ratio of the residual value to the initial value of the rolling stock</td>
<td>$\Xi$</td>
<td>-</td>
<td>0.05</td>
</tr>
<tr>
<td>Ratio of the residual value to the initial value of the infrastructure</td>
<td>$\Xi$</td>
<td>-</td>
<td>0.00</td>
</tr>
<tr>
<td>Ratio of the land capital cost to the route and stations construction cost</td>
<td>$\Lambda$</td>
<td>-</td>
<td>0.1</td>
</tr>
<tr>
<td>Ratio of the maximum to the average dwell time</td>
<td>$\psi$</td>
<td>-</td>
<td>1.550</td>
</tr>
<tr>
<td>Upper bound of the stop spacing</td>
<td>$d_{max}$</td>
<td>km</td>
<td>2.5</td>
</tr>
<tr>
<td>One-stage infrastructure technical life (route and stations)</td>
<td>$y$</td>
<td>year</td>
<td>40</td>
</tr>
<tr>
<td>One-stage vehicle technical life</td>
<td>$y$</td>
<td>year</td>
<td>25</td>
</tr>
<tr>
<td>Multiplicative factor of the operating cycle time</td>
<td>$\beta$</td>
<td>-</td>
<td>1.07</td>
</tr>
<tr>
<td>Upper bound on the maximum frequency</td>
<td>$f_{max}$</td>
<td>TU/h</td>
<td>40</td>
</tr>
<tr>
<td>Fraction of the longest dwell time in the maximum frequency formula</td>
<td>$\eta$</td>
<td>-</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Table 5: Mode-specific parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Unit</th>
<th>SR</th>
<th>AM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit operator cost per TU-hour</td>
<td>$c_{1t}$</td>
<td>$$/TU-h$</td>
<td>60</td>
<td>10</td>
</tr>
<tr>
<td>Average acceleration rate</td>
<td>$a$</td>
<td>m/s$^2$</td>
<td>1.15</td>
<td>1.20</td>
</tr>
<tr>
<td>Average deceleration rate</td>
<td>$b$</td>
<td>m/s$^2$</td>
<td>1.00</td>
<td>1.15</td>
</tr>
<tr>
<td>Threshold frequency for the high frequency penalty</td>
<td>$f$</td>
<td>TU/h</td>
<td>25</td>
<td>-</td>
</tr>
<tr>
<td>Fixed time lost for a stop (doors and other fixed times)</td>
<td>$t_d$</td>
<td>s</td>
<td>9.0</td>
<td>5.0</td>
</tr>
<tr>
<td>Fixed stop clearance time</td>
<td>$t_{e0}$</td>
<td>s</td>
<td>57</td>
<td>32</td>
</tr>
<tr>
<td>Exponent of the high frequency penalty</td>
<td>$\omega_2$</td>
<td>-</td>
<td>1.40</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 6: Technology-specific parameters in both scenarios.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Unit</th>
<th>SR1</th>
<th>SR2</th>
<th>SR3</th>
<th>SR4</th>
<th>AM1</th>
<th>AM2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Route capital cost</td>
<td>-</td>
<td>m$/km</td>
<td>11.63</td>
<td>15.58</td>
<td>18.27</td>
<td>42.1</td>
<td>24.15</td>
<td>46.58</td>
</tr>
<tr>
<td>Route maintenance cost</td>
<td>-</td>
<td>m$/km-year</td>
<td>0.03</td>
<td>0.12</td>
<td>0.15</td>
<td>0.39</td>
<td>0.23</td>
<td>0.43</td>
</tr>
<tr>
<td>Stop maintenance cost, one-way</td>
<td>-</td>
<td>$/stop-year</td>
<td>18827</td>
<td>35236</td>
<td>40521</td>
<td>114517</td>
<td>113612</td>
<td>140003</td>
</tr>
<tr>
<td>Incremental stop maint. cost per additional vehicle, one-way</td>
<td>-</td>
<td>$/stop-veh-year</td>
<td>-</td>
<td>24184</td>
<td>29998</td>
<td>40508</td>
<td>17423</td>
<td>22979</td>
</tr>
<tr>
<td>Vehicle length</td>
<td>-</td>
<td>m/veh</td>
<td>34</td>
<td>28</td>
<td>28</td>
<td>28</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>Vehicle capital cost</td>
<td>-</td>
<td>m$/veh</td>
<td>3.2</td>
<td>2.9</td>
<td>2.9</td>
<td>2.9</td>
<td>2.6</td>
<td>2.6</td>
</tr>
<tr>
<td>Vehicle capacity</td>
<td>k</td>
<td>pax/veh</td>
<td>227</td>
<td>191</td>
<td>191</td>
<td>118</td>
<td>118</td>
<td>118</td>
</tr>
<tr>
<td>Unit operator cost per vehicle-km</td>
<td>c_{2o}</td>
<td>$/veh-km</td>
<td>3.51</td>
<td>1.96</td>
<td>1.96</td>
<td>1.96</td>
<td>1.11</td>
<td>1.11</td>
</tr>
<tr>
<td>Vehicle administrative cost</td>
<td>c_{1vb}</td>
<td>$/veh-year</td>
<td>56129</td>
<td>54908</td>
<td>54908</td>
<td>54908</td>
<td>35306</td>
<td>35306</td>
</tr>
<tr>
<td>Time lost at intersections per average km</td>
<td>t_u</td>
<td>min/km</td>
<td>0.99</td>
<td>0.62</td>
<td>0.15</td>
<td>0.15</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Vehicle boarding and alighting time per user</td>
<td>t_b</td>
<td>s/veh/pax</td>
<td>1.5</td>
<td>1.4</td>
<td>1.4</td>
<td>1.4</td>
<td>2.1</td>
<td>2.1</td>
</tr>
<tr>
<td>Stop clearance time for an extra vehicle length</td>
<td>t_{ev}</td>
<td>s/veh</td>
<td>-</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>1.2</td>
<td>1.2</td>
</tr>
<tr>
<td>Access and egress extra-time to platform for a user</td>
<td>t_g</td>
<td>s</td>
<td>0</td>
<td>0</td>
<td>22</td>
<td>182</td>
<td>89</td>
<td>218</td>
</tr>
<tr>
<td>Fixed terminal time</td>
<td>t_{tf}</td>
<td>s</td>
<td>360</td>
<td>210</td>
<td>210</td>
<td>210</td>
<td>110</td>
<td>110</td>
</tr>
<tr>
<td>Variable terminal time</td>
<td>t_{tv}</td>
<td>s</td>
<td>-</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Parameter of the high frequency penalty</td>
<td>\omega_1</td>
<td></td>
<td>0.135</td>
<td>0.135</td>
<td>0.034</td>
<td>0.034</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Minimum number of vehicles per transit unit</td>
<td>n_{min}</td>
<td>veh/TU</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Maximum number of vehicles per transit unit</td>
<td>n_{max}</td>
<td>veh/TU</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Maximum allowed speed</td>
<td>S_{max}</td>
<td>km/h</td>
<td>50</td>
<td>75</td>
<td>86</td>
<td>86</td>
<td>90</td>
<td>90</td>
</tr>
</tbody>
</table>
### Table 7: Scenario-specific parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Unit</th>
<th>Scenario A</th>
<th>Scenario B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access and egress speed</td>
<td>$s$</td>
<td>km/h</td>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>Fraction of demand in the most loaded segment of the line</td>
<td>$\alpha$</td>
<td>-</td>
<td>0.45</td>
<td>0.25</td>
</tr>
<tr>
<td>Ratio of the average trip length to the length of the route</td>
<td>$\lambda$</td>
<td>-</td>
<td>0.40</td>
<td>0.25</td>
</tr>
</tbody>
</table>

### Table 8: Scenario A, technology-specific parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Unit</th>
<th>SR1</th>
<th>SR2</th>
<th>SR3</th>
<th>SR4</th>
<th>AM1</th>
<th>AM2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stop capital cost, one-way</td>
<td>-</td>
<td>m$/stop</td>
<td>0.36</td>
<td>1.42</td>
<td>1.64</td>
<td>4.28</td>
<td>2.83</td>
<td>4.49</td>
</tr>
<tr>
<td>Incremental stop capital cost per additional vehicle, one-way</td>
<td>-</td>
<td>m$/stop-veh</td>
<td>0.73</td>
<td>0.78</td>
<td>2.14</td>
<td>0.76</td>
<td>1.51</td>
<td></td>
</tr>
</tbody>
</table>

### Table 9: Scenario B, technology-specific parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Unit</th>
<th>SR1</th>
<th>SR2</th>
<th>SR3</th>
<th>SR4</th>
<th>AM1</th>
<th>AM2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stop capital cost, one-way</td>
<td>-</td>
<td>m$/stop</td>
<td>0.36</td>
<td>1.00</td>
<td>1.15</td>
<td>3.00</td>
<td>1.98</td>
<td>3.15</td>
</tr>
<tr>
<td>Incremental stop capital cost per additional vehicle, one-way</td>
<td>-</td>
<td>m$/stop-veh</td>
<td>0.51</td>
<td>0.55</td>
<td>1.50</td>
<td>0.53</td>
<td>1.06</td>
<td></td>
</tr>
</tbody>
</table>

### Table 10: Period-related parameters common to all scenarios.

<table>
<thead>
<tr>
<th>Ratio</th>
<th>Symbol</th>
<th>$p = 1$</th>
<th>$p = 2$</th>
<th>$p = 3$</th>
<th>$p = 4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Period average demand to average peak demand</td>
<td>$\gamma_p$</td>
<td>1.00000000</td>
<td>0.61421007</td>
<td>0.59064449</td>
<td>0.13704000</td>
</tr>
<tr>
<td>Maximum to average peak period demand</td>
<td>$\gamma_p$</td>
<td>1.27952633</td>
<td>1.35324947</td>
<td>1.21027921</td>
<td>2.40126798</td>
</tr>
<tr>
<td>Period hours to total service hours</td>
<td>$\chi_p$</td>
<td>0.08009259</td>
<td>0.13043981</td>
<td>0.20509259</td>
<td>0.58402778</td>
</tr>
</tbody>
</table>
5 Results

We now report results of the optimization model of Section 3 applied to the six technologies and two scenarios presented in Section 4. The two scenarios differ in the average journey length, and to abstract from this we present results according to a passenger travel density (PTD) index which expresses the amount of traveled distance by passengers per unit of route length. This index is defined as

$$\text{PTD} = Hq\lambda \sum_{p} x_p \gamma_p,$$

and its unit of measure is pax-km/year-km. We note that the “km” in the PTD unit “pax-km/year-km” is not redundant because the “km” in the numerator refers to the unitary distance traveled by a passenger, and the “km” in the denominator expresses the unitary route length.
For Scenario A the algorithm is executed on 118 peak hour demand levels uniformly spaced between 2000 and 60000 pax/h, extrema included, and for scenario B the studied peak demand range is 2000–120000 pax/h. These ranges of peak hour demands and the other parameters of the two scenarios yield PTDs between $\sim 10^6$ and $\sim 6 \times 10^7$ pax-km/year-km. For the two technologies with underground sections, SR4 and AM2, we have omitted in the charts the results below $3 \times 10^6$ PTD where these technologies are implausible and their construction intensive infrastructure yields high costs that would render some figures difficult to read.

5.1 Scenario A

A commuter demand in Scenario A induces relatively large stop spacings, Figure 22, which decline with increasing PTDs. This contributes to best-in-class average commercial speeds (ACSs) for the six technologies, Figure 23, which neatly differentiate between the AM mode and the semirapid technologies with SR3 and SR4 clustering at similar ACSs. Figure 24 illustrates the average operator costs. The supply of SR1, the less capital intensive technology, is cheaper up to a moderate travel density, $\sim 5 \times 10^6$ PTD, SR2 and SR3 hold up to $\sim 10^7$ PTD from where AM1 yields comparable costs to those of SR2 and SR3. SR4 and AM2, which share a similar capital intensive construction, present similar supply costs. However, the passenger time values, Figure 25, are stratified, with AM1 offering the best values. SR3 approaches AM1 at PTDs larger than $\sim 5 \times 10^6$. AM2, SR2 and SR4 tend to yield similar values, and SR1 registers significantly worse performances induced by the slower commercial speeds. We note that the slower commercial speed of SR2 with respect to AM2 is almost balanced in terms of smaller access costs, because of the smaller stop spacing and the easier access to platform of semirapid rail versus rapid rail with a significant fraction of underground stations. Combining operator and user costs yields the total cost depicted in Figure 26. We observe that for a commuter demand the two slower semirapid technologies, SR1 and SR2, offer a competitive advantage versus a lean automated metro, AM1, only at low travel densities up to $\sim 3 \times 10^6$ PTD. The fastest semirapid technology without an
underground section, SR3, matches the AM1 performances at high PTDs and yields the best overall performances below $\sim 10^7$ PTD. The two technologies with underground sections, SR4 and AM2, offer similar values, with a slight AM2 advantage above $\sim 10^7$ PTD.

To clarify why the reported performance indices may present "jumps" in their curves, we comment on the discrete variables that we use for all technologies with the exception of SR1: the TU length, i.e., the number of vehicles composing a train. In the computational results, the AM2 TU lengths that we obtain are very similar to those of AM1. The SR3 and SR4 optimal TU lengths are slightly different from those of SR2, but the general pattern is the same. Therefore, the TU lengths and related indices for AM2, SR3, and SR4 are in the following omitted for brevity. We illustrate in Figures 27 and 28 the optimal TU lengths for AM1 and SR2, respectively. Comparing AM1 and SR2 TU lengths, we observe that there is a much larger TU length diversity among periods with automated metro than with semirapid rail. Figures 29 and 30 report the deployed fleets per period of AM1 and SR2, respectively. We note that automated metro tends to have a larger variation in the size of deployed fleet between periods than semirapid rail, which almost always uses the same fleet size at the two busiest periods. We recall from Table 10 that the two busiest periods account for $\sim 21\%$ of the operating time in our experiments. The impact of the discreteness of the TU length variables is noticeable on the frequencies. Figures 31 and 32 report the frequencies per period of AM1 and SR2, respectively. In these two charts we use a linear scale for the horizontal axes, and the upper horizontal axis expresses the "bottleneck capacity", i.e., the capacity offered at the most loaded segment of the line at the peak hour. The charts depict both the optimal frequencies and, by dotted lines, the minimum frequencies as per equation (18). By so doing, we can show when the constraint on the minimum frequency is binding. The analysis of the frequencies allows us to draw five main observations. First, both AM1 and SR2 at the peak period offer the minimum frequency required by the bottleneck capacity constraint, i.e., the frequency at the peak period, which determines the fleet size, does not offer extra-capacity to further reduce the user cost. Second, at
non-peak periods both AM1 and SR2 supply frequencies larger than those strictly required by the bottleneck capacity constraint. The low variable cost of rail-based transit, see Figure 33, allows for these user benefits. Third, AM1 consistently offer larger frequencies than those of SR2, particularly at off-peak periods. Fourth, AM1 in periods two and three is constrained by the maximum frequency, 40 TU/h. This result and the fact, Figure 29, that AM1 does not fully deploy the fleet at off-peak periods, indicate that it would be optimal, in a total cost perspective, to strive for higher frequencies at these periods. Fifth, the optimal frequency of SR2 at the peak period is always lower than the upper bound on the frequency. This result indicates that the constraint on the maximum frequency derived from the maximum dwell time is binding for SR2. On the last point, the maximum dwell time, we observe that this value is also related to the number of stations, which is inversely related to their construction costs. This fact can be inferred from Figure 34 that shows the maximum dwell time at the peak period for all technologies. The worsening of the maximum dwell time at higher station costs lowers the maximum attainable frequency for semirapid rail. This effect can be appreciated in Figure 35 where we report the maximum and minimum frequencies at the peak period for all technologies.

To illustrate how remarkably different may be the performance offered by semirapid rail and automated metro, also when the average user cost is similar, we depict in Figures 36 and 37 the breakdowns of the user costs of AM2 and SR2, respectively. Semirapid rail provides a lower access cost with respect to automated metro, but at the expense of higher waiting and in-vehicle travel costs. A similar conclusion can be drawn by comparing the breakdowns of the user costs of AM1 and SR3, which we report in Figures 38 and 39, respectively.

We further observe that automated metros and the two fastest semirapid rail, SR3 and SR4, may offer lower crowding, i.e., lower average passengers per unit of vehicle length up to $\sim 9 \times 10^6$ PTD, see Figure 40.
5.2 Scenario B

Scenario B deals with a demand pattern of short trips and a prevalence of pedestrian access to the stations. Therefore, the access time plays a more crucial role than in the previous scenario. Figure 41 depicts the stop spacings, which, as expected, are significantly lower than those of Scenario A. Average commercial speeds are also lower, Figure 42, but they still separate between the AM mode and the four semirapid rail technologies with SR3 and SR4 converging at high PTDs. The operator costs are larger than in the previous scenario, Figure 43, because this demand pattern is more costly to serve, and SR1 holds its convenience up to a large demand density, $\sim 10^7$ PTD, beyond which AM1, SR2, and SR3 become the cheapest options. SR4 and AM2 present similar costs. Passenger time values are stratified, with AM1 and SR3 still the most convenient options, followed by SR2, AM2, SR4 and SR1, Figure 44. The average total costs show that SR1 has a limited convenience only at the lowest travel densities up to $\sim 10^6$ PTD, SR2 break-even with AM1 at $\sim 6 \times 10^6$ PTD, but in the range between $\sim 3 \times 10^6$ and $\sim 4 \times 10^7$ PTD differences between these technologies and the overall best performer SR3 are small. As in Scenario A, AM2 and SR4 are less convenient of the other technologies and show similar values, Figure 45.

We now provide some results on the optimal TU lengths, deployed fleets, and frequencies. In Scenario B computational results of frequencies and derived indices show a slight numerical instability because of multiple near-optimal local minima. However, the relative smoothness of the objective function, the total cost, reported in Figure 45, allows us to be confident in the overall validity of the numerical results. Analogously to our presentation in the previous section, we focus for the sake of brevity on the results of AM1 and SR2. For these two technologies, respectively, Figures 46 and 47 illustrate the optimal TU lengths, Figure 48 and 49 depict the deployed fleets, and Figures 50 and 51 present the frequencies. We observe that four patterns noted for Scenario A hold as well for Scenario B. First, AM1 presents more diversity among periods in the TU lengths and deployed fleets than SR2. Second, AM1 offers larger frequencies than SR2. Third, AM1 in periods two and three is constrained by the maximum fre-
frequency. Fourth, the optimal frequency of SR2 at the peak period is always lower than the upper bound on the frequency because of the binding constraint on the frequency derived from the maximum dwell time, see Figures 52 and 53.

However, in Scenario B the bottleneck capacity constraint is not always binding at the peak period, and some extra-capacity may be offered there to further lower the user cost. In fact, for all technologies, the variable supply cost is relatively low and decreasing with demand, see Figure 54.

Figures 55 and 56 report the breakdowns of the user costs of AM1 and SR3, respectively, showing also under Scenario B different structures of comparable average user costs of the two studied modes. Automated metros and to a certain extent SR3 and SR4 exhibit lower crowding than the two slower semirapid rail technologies, Figure 57.

Figure 22: Scenario A, stop distance.
Figure 23: Scenario A, average commercial speed.

Figure 24: Scenario A, average operator cost.
Figure 25: Scenario A, average passenger time value.

Figure 26: Scenario A, average total cost.
Figure 27: Scenario A, Transit Unit length per period for technology AM1.

Figure 28: Scenario A, Transit Unit length per period for technology SR2.
Figure 29: Scenario A, deployed fleet per period for technology AM1.

Figure 30: Scenario A, deployed fleet per period for technology SR2.
Figure 31: Scenario A, optimal and minimum frequencies per period for technology AM1.

Figure 32: Scenario A, optimal and minimum frequencies per period for technology SR2.
Figure 33: Scenario A, Operation and maintenance cost, excluding annuity of the fleet, per unit of service supplied.

Figure 34: Scenario A, maximum dwell time at the peak period.
Figure 35: Scenario A, minimum and maximum frequencies at peak period.

Figure 36: Scenario A, disaggregation of the average passenger time value in its components for technology AM2.
Figure 37: Scenario A, disaggregation of the average passenger time value in its components for technology SR2.

Figure 38: Scenario A, disaggregation of the average passenger time value in its components for technology AM1.
Figure 39: Scenario A, disaggregation of the average passenger time value in its components for technology SR3.

Figure 40: Scenario A, average passenger occupancy per unit of vehicle length.
Figure 41: Scenario B, stop distance.

Figure 42: Scenario B, average commercial speed.
Figure 43: Scenario B, average operator cost.

Figure 44: Scenario B, average passenger time value.
Figure 45: Scenario B, average total cost.

![Graph showing average total cost vs. PTD (pax-km/year-km) for different technologies.]

Figure 46: Scenario B, Transit Unit length per period for technology AM1.

![Graph showing Transit Unit length per period vs. PTD (pax-km/year-km) for AM1.]
Figure 47: Scenario B, Transit Unit length per period for technology SR2.

Figure 48: Scenario B, deployed fleet per period for technology AM1
Figure 49: Scenario B, deployed fleet per period for technology SR2

Figure 50: Scenario B, optimal and minimum frequencies per period for technology AM1.
Figure 51: Scenario B, optimal and minimum frequencies per period for technology SR2.

Figure 52: Scenario B, maximum dwell time at the peak period.
Figure 53: Scenario B, minimum and maximum frequencies at peak period.

Figure 54: Scenario B, Operation and maintenance cost, excluding annuity of the fleet, per unit of service supplied.
Figure 55: Scenario B, disaggregation of the average passenger time value in its components for technology AM1.

Figure 56: Scenario B, disaggregation of the average passenger time value in its components for technology SR3.
6 Conclusions and future work

We have presented a techno-economic assessment of conventional and automated urban rail technologies. The spectrum of relevant cases has been sampled by selecting opposite but representative passenger demand and transit alignment patterns. For the passenger demand we have individuated two patterns: local, with prevalent pedestrian access and egress for short trips; and commuter, with prevalent motorized access and egress for long trips. For the transit line alignment we have selected patterns that differ in the amount of civil construction for each of the two studied modes. In the case of semirapid rail we have studied an alignment with minimal construction in a pre-existing arterial road, and three newly developed alignments that can provide higher operating speeds and less interference with transversal traffic. Automated rail, because of the exclusive right-of-way that it requires, belongs to a civil construction class that is considerably more capital intensive than conventional semirapid rail. Nevertheless, it is still possible — and insightful — to differentiate between a lean alignment with a mix of at-grade and elevated sections, and a rich one, with substantial underground
sections. These two alignments differ both in the capital and O&M requirements for the
civil structures, the route and the stations, and in the passengers’ experience, mainly
because of the faster access and egress in the lean case. Results of the model show that
these differentiations are highly relevant, and without them a broad generalization
across transit modes would be misleading.

The demand and the alignment patterns synergistically determine the relative and
absolute weights of access and egress, and in-vehicle travel time components in the
user cost computation. Under a local demand pattern, technologies with low aver-
age operating speed but which can offer stations with low capital and O&M costs, and
fast access and egress, may hold their convenience at higher travel densities than in the
case of a commuter pattern. A more capital intensive alignment is particularly relevant
in the case of conventional semirapid rail, but only to a point. SR1 has a limited scope
and only at low travel densities under a local demand pattern. SR2 excels in broader
demand ranges, particularly under a commuter demand pattern. SR3 yields the best
performances matching those of the automated metros, but SR4 is never competitive
with the other semirapid technologies. Similarly in the case of automated metro, the
more capital intensive alignment with underground sections fares poorly with respect
to the lean case. The average operational speeds are similar in the two cases, but the
lower access and egress time, and the lower capital and O&M costs of the lean align-
ment make it predominant.

In synthesis, the relevance for the scientific literature and for the public transport
practice is as follows. First, we show the relevance of the demand pattern to be served
and of the type of feasible alignments. Second, automated metro considerably expands
the range of cases where it is competitive with conventional semirapid rail if, and only
if, an alignment that is relatively lean in terms of construction is feasible. Third, under
most circumstances, the fastest technologies, automated metros, SR3, and SR4, achieve
their best results at a lower average occupancy than slower technologies. We remark
the relevance of this latter point, that it is only conservatively captured in our model, in
a world where a pandemic risk is not anymore — unfortunately — a remote possibility.
Future work may proceed along three directions. First, the model could be applied to other technological trends such as the electrification of bus rapid transit. Electrified BRT is now feasible with a wide range of technologies encompassing batteries, overhead contact systems, and hybrids between these two options. Battery advantages such as route flexibility and disadvantages such as recharging times and reduced payload should be evaluated versus an overhead contact system, which has opposite pros and cons. Second, the model could generate marginal supply cost curves to be used for pricing studies. Third, dataset of results over several demand scenarios and technologies could be the basis of decision support systems for the screening of transit projects.

Acknowledgements

Luigi Moccia was partly supported by CNR (Italy) under project “Sequence/Stream Analytics and Process Intelligence”. Gilbert Laporte was funded by the Canadian Natural Sciences and Engineering Research Council under grant 2015-06189. These supports are gratefully acknowledged. Luigi Moccia and Duncan W. Allen thank Eric C. Bruun for fruitful discussions on transit planning and operations. We thank the Editor and the referees for their valuable comments.

References


sess the impact of a suburban first-mile shared autonomous vehicle fleet service on transit demand. In *Transportation Research Board Annual Meeting*, Washington, D.C.


Zhang, K., Mahmassani, H. S., and Vovsha, P. (2011). Integrated nested logit mode choice and dynamic network micro-assignment model platform to support conges-
