C. Barnhart and G. Laporte (Eds.), *Handbook in OR & MS, Vol. 14*

DOI: 10.1016/S0927-0507(06)14002-5

Chapter 2
Public Transit

Guy Desaulniers

Department of Applied Mathematics and Industrial Engineering, École Polytechnique and GERAD, Montréal, Québec, H3T 2A7 Canada
E-mail: Guy.Desaulniers@gerad.ca

Mark D. Hickman

Department of Civil Engineering and Engineering Mechanics, The University of Arizona, Tucson, AZ 85721, USA
E-mail: mhickman@engr.arizona.edu

1 Introduction

For several decades now, operations research has been successful for solving a wide variety of optimization problems in public transit. Several commercial software systems based on operations research techniques have been designed and used by the transit agencies to help them plan and run their operations. Operations researchers have been attracted by the public transit problems because of their size and complexity. Indeed, some of them are huge in practice. For instance, the New York City Transit Authority employs more than 12,000 drivers to operate approximately 4500 buses that serve over 240 bus routes. Furthermore, these problems are complex because they involve passengers, buses, and drivers that are subject to individual preferences and constraints, and interact with each other according to a set of prescribed relationships.

The main goal of most transit agencies is to offer to the population a service of good quality that allows passengers to travel easily at a low fare. The agencies thus have a social mission which aims at reducing pollution and traffic congestion, as well as increasing the mobility of the population. In most cases, the goal is usually not to make profits, as is the case for almost all other transportation organizations such as airlines, railroads, and trucking companies. They are, however, subject to budgetary restrictions that force them to manage expensive resources such as buses, drivers, maintenance facilities, and bus depots as efficiently as possible. Briefly stated, the global problem faced by the agencies consists of determining how to offer a good-quality service to the passengers while maintaining reasonable asset and operating costs.

Addressed as a whole, this global problem is not tractable. Hence, it is divided into a set of subproblems that are usually solved sequentially at various stages of the planning process (strategic, tactical, and operational), and even during operations (real-time control). Strategic planning problems con-

cern long-term decisions such as the design of the transit routes and networks. Most of these problems fall within the category of network design problems and require solving passenger assignment problems as subproblems or for evaluation purposes. These strategic problems aim at maximizing service quality under budgetary restrictions. Tactical planning problems concern the decisions related to the service offered to the public, namely the frequencies of service along the routes and the timetables. These problems are usually solved on a seasonal basis, with occasional updates. These problems also focus on the quality of service. Operational planning problems relate to how the operations should be conducted to offer the proposed service at minimum cost. They include a wide variety of problems such as vehicle scheduling, driver scheduling, bus parking and dispatching in garages, and maintenance scheduling. These problems are solved at various intervals that range from once per month for driver scheduling to once per day for bus parking and dispatching. In contrast to the objectives of the previous problems, the objective for the operational planning problems is clearly one of minimizing total cost. Finally, real-time control problems manage perturbations to the plan using several control strategies. These problems are solved in real time during operations and aim at minimizing passenger inconvenience. Usually, minimal or no operating costs are involved since they consider minor perturbations to the scheduled service.

The goal of this chapter is to review state-of-the-art models and approaches for solving these public transit problems. This review is not exhaustive as it mostly covers the recent contributions that have been applied or have the potential to be applied from our viewpoint. Readers interested on earlier works are referred to the survey paper by Odoni et al. (1994), as well as the series of books arising from the Computer-Aided Transit Scheduling conferences that have contributed tremendously to the practice and growth of operations research in public transit. These books are listed at the beginning of the references, i.e. Wren (1981), Rousseau (1985), Daduna and Wren (1988), Desrochers and Rousseau (1992), Daduna et al. (1995), Wilson (1999), Voss and Daduna (2001), and Hickman et al. (in press).

In certain cases, computational results are briefly reported to give an idea of the problem sizes that can be solved. However, these results are not intended to compare the different approaches. In fact, they can hardly be compared because they most of the times have been obtained using different computers or on different datasets that did not necessarily exhibit the same characteristics.

2 Strategic planning

At the strategic level, transit planning is concerned with the design of transit routes and networks. This involves designing a network of routes to meet passenger demand. Since the demand is based in large part on the network design, the network design problem relies heavily on methods to determine

passengers' route choices (or "assignment" to routes) serving their origins and destinations. This section includes a description of the transit network design problem and a discussion of research in passenger assignment.

2.1 Network design

The public transit network design problem is somewhat more complicated than the traditional network design problem. In addition to determining what links to include in the network, the transit network design includes assembling these links into fixed routes, and determining the frequency of service on each route. The result of the network design, then, should include a set of routes and their frequencies. Most commonly, the problem is formulated on a graph with nodes, links, and (subsequently) routes. Let $G = (N, A)$ be a graph with N, the set of nodes, and A, the set of links, and R represents the set of routes. Nodes represent intersections (e.g., road intersections), but can also represent zone centroids where a geographic zone is represented by a single point (the centroid). A link between nodes represents a particular mode of transport between nodes, and a route represents a sequence of nodes and links of a single mode.

As input to this problem, the formulation typically assumes that there is an existing origin–destination (O–D) matrix, covering the demand between a set of nodes or zones, either on a daily basis or for a specific period within the day. Alternately, it is reasonable (although complicated) to assume that demand is endogenous and determined as an equilibrium problem, in which the flows are a function of the network design. With the origin–destination flows and an assignment of these flows to routes, the set of routes and their frequencies must be determined.

One of the challenges of network design is in the specification of the objective function. Most commonly, the objective is to minimizing the total travel time or the generalized cost of travel. The generalized cost may be found by applying different weights in the objective function to the different components of travel time such as walking (or access) time, initial waiting time, in-vehicle time, transfer time, and egress time separately. Some formulations also include the number of transfers as a component in the generalized cost.

In addition, the costs to the transit operator may also be considered, either explicitly in the objective function or through a constraint on the total budget (or operating profit or loss). Such costs can include the operating cost, given as a function of the route length (in distance or time) and frequency, and the fixed cost of a bus fleet and/or infrastructure along the route network (for rail transit networks). If operator costs are included, a composite objective may be formulated, or the problem can be specified as a multiobjective programming problem.

In addition to this traditional formulation of the objective, several other constraints often enter into the problem in practice; these include: (1) ensuring adequate coverage in the network to provide access to specific nodes or zones

in the service area; (2) ensuring minimum frequencies of service to specific nodes or links in the network; and (3) any other design considerations such as the availability of infrastructure or right-of-way for routes.

In practice, the network design problem incorporates each of these objectives and constraints through a more interactive formulation of the problem, where sample routes may be constructed by computer methods, but are ultimately selected in conjunction with manual review by network designers and planners. In this review, we discuss some of the more significant methods that rely heavily on mathematical programming techniques.

The network design problem is known to be NP-hard (Magnanti and Wong, 1984). As a result, approaches to the problem rely on heuristic techniques to solve problems of reasonable size. Much of the initial work decomposed the problem into two stages: in the first, the set of routes is constructed; in the second, the set of frequencies for these routes are determined. This includes the work of Lampkin and Saalmans (1967) and Silman et al. (1974). In the first stage, heuristic methods are used to construct "skeleton" routes, and these skeleton routes are expanded to cover the full set of nodes in the network. Once the routes are defined, the frequencies are determined by minimizing the total passenger travel time, calculated as the sum of the O–D demand D_{ij} multiplied by the travel time T_{ij}, subject to a constraint on the total fleet size (as a budget constraint):

$$\text{minimize} \quad \sum_{i \in N} \sum_{j \in N} D_{ij} T_{ij}(\mathbf{f}) \tag{1}$$

subject to:

$$\sum_{r \in R} \lceil RT_r f_r \rceil \leqslant \text{total fleet size}. \tag{2}$$

In this formulation, RT_r is the round-trip time on route r and f_r is the frequency on route r. The travel time T_{ij} includes the expected waiting time, as a function of frequencies $\mathbf{f}$ of routes serving the origin i and any transfer node k on the shortest path serving the O–D pair i, j. Lampkin and Saalmans (1967) used a random gradient-based search procedure to determine the final frequency values. In Silman et al. (1974), a penalty is added to the objective for the estimated number of standees on the bus; this penalty is given as a piecewise differentiable function of the route frequency. Their approach uses a gradient projection method to minimize the total travel time.

Dubois et al. (1979) decomposed the network design problem into three subproblems. The first involves determining the links in the street network on which to operate the service; the second determines the routes themselves; and the third determines the optimal frequencies on each route. In the first step, a traditional network design problem is formulated, where the objective is given in (1), using only in-vehicle times as the travel times, subject to a budget constraint on the cost of operating on a street, and binary decision variables indicating whether a street segment is in the final solution. A heuristic is used

to solve this problem, beginning with an initial spanning tree to minimize the total travel time and adding links to minimize this total. A simple all-or-nothing assignment of the O–D flow on the shortest paths is used to estimate the objective function. In the second step, a maximal set of routes is generated from the street network. This set of routes is then subject to heuristic rules to determine the final route structure: (1) routes with heavy transfer flow are joined; (2) route segments are deleted where the demand is effectively served by other routes; and (3) routes that overlap are joined. In the third step, the optimal route frequencies are found using a gradient-based search heuristic, similar to Lampkin and Saalmans (1967). Waiting times are explicitly considered in this last step.

A more formal presentation of the transit network design problem, from a mathematical programming approach, is given by Hasselström (1981). Hasselström proposed a two-stage process of network design in which routes and frequencies are determined simultaneously. In the first stage, an initial route network is generated; in the second stage, this route network is refined and a detailed evaluation of the routes and the passenger assignment is performed. In addition to solving the route and frequency problem simultaneously, the advantage of Hasselström's method is in its implementation and application for realistic network sizes.

In the first stage, Hasselström's (1981) formulation includes a direct demand function, allowing the demand to be determined endogenously. The form of the direct demand model is based on a traditional gravity model with parameter β, where all terms not dependent on the route structure or the frequencies are rolled into a constant term K_{ij} for each O–D pair i, j. Remaining elements of the generalized cost are given by C_{ij}, which is a function of the set of frequencies $\mathbf{f}$. The frequency of each route r is denoted f_r. The objective function, maximizing consumer surplus, is equivalent to maximizing the number of passengers with this demand function. As constraints, Hasselström (1981) considered a budget constraint $\overline{C}$ that includes a cost per vehicle on each route c_r. Also, there is a required minimum frequency of service for a zone $s \in S$, given as Δ_s. If a route r serves zone s, this is indicated with a binary parameter δ_{rs}. Finally, the frequencies must be included in a feasible set $\mathcal{F}$ (e.g., nonnegative integers).

The network design problem in the first stage is then formulated as follows (in this formulation, the notation is simplified to deal with a single transit mode; multiple modes may also be considered in the network):

$$\text{maximize} \quad \sum_{i \in N} \sum_{j \in N} D_{ij} \tag{3}$$

subject to:

$$D_{ij} = K_{ij} \mathrm{e}^{-\beta C_{ij}(\mathbf{f})}, \quad \forall i, j \in N, \tag{4}$$

$$\sum_{r \in R} c_r f_r \leqslant \overline{C}, \tag{5}$$

$$\sum_{r \in R} f_r \delta_{rs} \geqslant \Delta_s, \quad \forall s \in S, \tag{6}$$

$$f_r \in \mathcal{F}, \quad \forall r \in R. \tag{7}$$

In (4), the generalized cost term C_{ij}, a function of the frequencies $\mathbf{f}$, includes waiting and transfer times for the O–D pair, as determined in the passenger assignment. The demand is therefore given as a function of the service frequencies.

To solve this model in the first stage, an initial network is generated by enumerating all possible routes serving a pair of terminals. A set of heuristic rules is then used to prune clearly inferior routes from this set. Then, the final routes and frequencies are constructed by solving a mathematical program: one method uses a linear program to maximize the passenger flow; a second method uses a convex nonlinear program to maximize the consumer surplus. In this first stage, the assignment uses all common routes (see Section 2.2.1) to determine waiting and transfer times. The decision variables in both formulations are the frequencies of each route; routes with very low frequencies can be pruned from the solution space.

In the second stage, a detailed assignment is performed, and the routes are refined. Passenger assignment on existing routes is performed with the heuristic of Andreasson (1977) (discussed in Section 2.2.1). With this new assignment, several route refinements are considered. First, optimization of the connection of route segments at route intersections is proposed; this problem is formulated as a maximum weighted matching problem of combining route segments at the point of intersection. Also, a nonlinear program is formulated for re-optimization of frequencies, using the vehicle fleet size constraint. This optimization is solved by Lagrangian relaxation.

Hasselström (1981) reports on a case study with 50 local bus routes, 10 tram routes, and express bus service. Since this time, the methodology has been developed as commercial software, and hence is clearly able to solve realistic problem sizes.

In the work of Ceder and Wilson (1986), two different mathematical formulations of the bus network design problem are suggested. The first formulation considers the passenger objective of minimizing excess travel time upon boarding, expressed as the sum of "excess" travel time (larger than the shortest travel time with a direct route) plus the transfer time (if any), summed across all O–D pairs. This objective is minimized subject to constraints on the maximum O–D travel time (as a percentage above the shortest path), lower and upper bounds on the route length (expressed in units of running time), and a constraint on the maximum number of routes. A second formulation adds the passenger waiting time and vehicle operating and capital costs to the objective function; it also includes constraints on the minimum frequency for each route and a constraint on the maximum fleet size. To generate a large set of feasible routes, Ceder and Wilson (1986) proposed a heuristic in which each designated terminal node is processed separately. A (topological) breadth-first

search is conducted, in which potential routes that do not meet the constraints on the maximum travel time are eliminated. In addition, the total "excess" passenger hours are also calculated for the route. This set of feasible routes can then be used for further screening and passenger assignment.

The mathematical formulation of Ceder and Wilson (1986) was extended more recently by Israeli (1992) and related papers (Israeli and Ceder, 1989, 1995; Ceder and Israeli, 1998). In this research, the transit network design problem is formulated as a multiobjective programming problem, with two objectives: the total passenger cost (Z_1) and the operator fleet size (Z_2). The total passenger costs Z_1 includes the in-vehicle passenger hours spent between the origin and destination PH_{ij}, the waiting and transfer time spent traveling from the origin to the destination WH_{ij}, and the empty seat-hours on a route r denoted EH_r. These three terms are weighted (weights a_1, a_2, and a_3) in the objective. Formally, the objectives are described as

$$\text{minimize} \quad Z_1 = a_1 \sum_{i \in N} \sum_{j \in N} PH_{ij} + a_2 \sum_{i \in N} \sum_{j \in N} WH_{ij} + a_3 \sum_{r \in R} EH_r, \quad (8)$$

$$\text{minimize} \quad Z_2 = \text{fleet size}. \quad (9)$$

Constraints in the formulation include the passenger assignment from a fixed demand matrix, and minimum frequencies on each route. This problem is solved with the following heuristic:

1. The full set of feasible routes is enumerated, in a manner similar to Ceder and Wilson (1986).
2. Additional direct routes are added to the network between O–D pairs with high demand, where the origin and destination nodes are not terminals. Second, the number of transfers required for each O–D pair for the given route structure is calculated, up to a maximum (e.g., 1 or 2 transfers).
3. A minimal set of routes is obtained through a heuristic procedure. This problem is set up as a set covering problem with the full set of feasible routes. Each column is defined as a route or a feasible combination of routes meeting the maximum number of transfers; the rows are O–D pairs. The objective minimizes the deviation from shortest paths, while maintaining constraints on connectivity for each O–D pair (i.e., reachable within the maximum number of allowable transfers).
4. The assignment of flow to paths and the frequencies on each path are determined iteratively. The frequencies are determined based on the peak load segment on each route, and these in turn are used in calculating the waiting and transfer times in the assignment. The assignment procedure is loosely based on that of Marguier and Ceder (1984), described in Section 2.2.1. With this information, Z_1 is calculated.
5. The minimum fleet size Z_2 is determined using the method of Stern and Ceder (1983).

6. New routes are considered to explore other points in the solution space for the two objectives. In this procedure, a column generation technique is used to avoid re-evaluating previously accepted route sets. These new route sets are re-evaluated by repeating steps 3–5.
7. The route sets in the efficient frontier are evaluated and presented to the decision-maker.

The example problem and solution in Israeli and Ceder (1995) is a problem with 8 nodes and 14 links (based on a similar problem from Ceder and Wilson (1986)). It is unknown how the solution method would perform on larger, more realistic networks.

Another mathematical programming approach to the transit network design problem was proposed by van Nes et al. (1988). In this approach, the routes and frequencies are determined simultaneously. The objective function maximizes the number of direct trips (i.e., trips without transfers) served in the network, for a given fleet size. A direct demand model is proposed to estimate the origin–destination trips by public transit; trips are proportional to the attraction of the origin zone i, O_i, and the destination zone j, D_j, and is an exponential function of the cost, similar to that of Hasselström (1981). The objective function is formulated as

$$\text{maximize} \quad \sum_{i\in N}\sum_{j\in N} aO_iD_j\mathrm{e}^{-\beta C_{ij}(\mathbf{f})}. \tag{10}$$

The generalized cost term C_{ij} is defined as an explicit function of the frequencies of the optimal subset of routes serving the passenger's origin i, R_i^* (Chriqui and Robillard, 1975):

$$C_{ij}(\mathbf{f}) = K_{ij} + \frac{60\alpha}{\sum_{r\in R_i^*} f_r} + c. \tag{11}$$

In this equation, α and c are constants. This equation assumes, for direct service, a constant K_{ij} for access time, egress time, and in-vehicle travel time, and adds a term for the waiting time as a function of the frequencies on acceptable routes.

A variety of constraints are used in this formulation. For these, consider a set $M = \{1, \ldots, m\}$ of vehicle types, with N_m the total number of vehicles available of type m. Operating a vehicle of type m incurs a cost factor k_m. A total of N_r vehicles are assigned to route r, with the indicator b_{mr} equal to 1 if vehicle type m is assigned to route r. With these variables, the constraints include: a budget constraint of $\overline{C}$, the vehicle availability N_m, the set of feasible frequencies $\mathcal{F}$, and the allocation of buses among routes based on the frequency and the round-trip time on the route RT_r. Mathematically, these constraints are formulated as:

$$\sum_{m\in M} k_m \sum_{r\in R} N_r b_{mr} \leqslant \overline{C}, \tag{12}$$

$$\sum_{r \in R} N_r b_r \leqslant N_m, \quad \forall m \in M, \tag{13}$$

$$f_r \in \mathcal{F}, \quad \forall r \in R, \tag{14}$$

$$N_r - 1 < f_r \frac{RT_r}{60} \leqslant N_r, \quad \forall r \in R. \tag{15}$$

The solution technique adopted by van Nes et al. (1988) is a heuristic in which all frequencies on proposed routes are set to 0. Each route is evaluated with respect to its potential to improve "efficiency", defined as the ratio of passengers added by the direct service to the additional cost of increasing the frequency, evaluated in (12)–(15). The route with the highest efficiency is selected and the frequency on that route is increased, until the budget and the available vehicles are consumed. It is shown that this heuristic is similar to evaluating the Kuhn–Tucker conditions for the problem when the budget constraint is included in the objective with a Lagrange multiplier. The Lagrange multiplier is identical to this "efficiency" measure, and these should be approximately equal across routes in the final solution.

van Nes et al. (1988) report testing this solution technique on a network from the Netherlands with 182 nodes and 115 zones, for a network of 8 routes. The paper also reports that the modeling system is capable of solving instances up to 250 nodes, 150 zones, and 750 possible routes.

More recent work has also included a number of metaheuristic methods. Baaj and Mahmassani (1995) and related work (Baaj and Mahmassani, 1990, 1992) decompose the network design problem into three elements: a route generation step, in which routes and frequencies are constructed; a network analysis procedure, defining measures of effectiveness at the network-, route-, and stop-level; and a route improvement algorithm to improve the route design. The heuristic proposed by Baaj and Mahmassani (1995) begins by generating additional skeleton routes connecting the highest O–D pairs in the demand matrix with direct service. With these skeletons, a set of possible node selection and insertions strategies are used to generate full routes. Then, passenger assignment follows the method of Han and Wilson (1982) (see Section 3.1) and determines the frequency and number of buses on each route according to a pre-specified maximum loading factor. Once this assignment is completed, the network evaluation tool is used to identify the number of trips satisfied by direct, one-transfer, and two-transfer trips, and the total waiting time, in-vehicle travel time, and transfer time in the network. The route improvement procedures then are called to improve the route structure through heuristics that: (1) prune off low ridership routes and/or route segments and joining these with other routes; and (2) consider improvements to routes by splitting routes into two parts or by exchanging route legs between routes at points of intersection.

A recent study by Fan and Machemehl (2004) examined simulated annealing, tabu search, genetic algorithms, local search, and random search techniques to solve the network design problem. As with other previous methods,

all the techniques begin with a set of skeleton routes. These metaheuristics are used to generate additional routes; the output is run through a network evaluation tool. Subsequent iterations between the network analysis tool and the metaheuristics are used to improve the quality of the solution.

Other authors have investigated the use of genetic algorithms for the transit network design problem; these include (among many others) recent works by Pattnaik et al. (1998), Bielli et al. (2002), Tom and Mohan (2003), and Verma and Dinghra (2005). These methods involve two steps. First, all feasible routes are generated, typically in a method similar to that proposed by Ceder and Wilson (1986). A set of routes are coded into the genetic algorithm as a string, containing a certain number of routes. For this technique, a fixed number of routes are required by the algorithm, although the number of potential routes can vary so as to determine the optimal number of routes. These strings are then evaluated using the assignment and network evaluation techniques of Baaj and Mahmassani (1995), and consistent with the methods of genetic algorithms, the pool of strings to be evaluated is evolved to a new population, and the process iterates. The size of networks in the genetic algorithm approach can be somewhat larger than with the analytic methods; Bielli et al. (2002) report solving an instance of 1134 nodes, 3016 arcs, 459 stops, and 22 routes. Tom and Mohan (2003) report solving an instance with 1332 nodes (bus stops) and 4076 arcs.

2.2 Passenger assignment

One of the critical issues in strategic planning is determining the demand on each route and other measures of service consumption. Most of the more common strategic measures of performance, from the perspective of the passenger, relate to the amount of time and money spent traveling in the network; i.e., elements of the passenger's path from the origin to destination. Hence, the passenger assignment is critical to determining system performance.

The passenger assignment problem can be defined as follows. Given an origin-to-destination flow, what are the flows on paths through the transit network, taken by the passengers? In formulating this problem, the passenger's objective is assumed to be minimizing travel time or generalized cost. The travel time may consist of some or all of the following variables, with perhaps different weights: the time to access a stop, the waiting time in the stop, the in-vehicle time, the transfer time, the number of transfers, egress time, and any monetary cost. The passenger then faces the task of selecting a route or set of routes that may be able to get from the origin to the destination with the minimum time or generalized cost. In the literature, this problem is addressed within the network design problem (Section 2.1), as part of the task of determining frequencies on routes (Section 3.1), and also as a unique problem itself.

In contrast to the simplicity of the problem definition, there are a number of aspects of the problem that have led to several different research approaches.

One important concept in passenger assignment is the determination of the "minimum cost" path. Important elements in defining the attributes of cost include:

1. The characterization of time-dependence and stochastic attributes in the minimum cost path.
2. The characterization of a solution as: (1) a *single path*, including only a route or combination of routes; (2) a path that can include a set of *common lines*, including cases where multiple routes may overlap on some part of the shortest path; or (3) a *strategy*, allowing passengers to choose their own boarding rules as they travel from origin to destination.
3. The effect of capacity and crowding in the transit network.

In the characterization of the minimum cost path, a traditional approach assumes that deterministic values can be used for travel times on each link. Traditional shortest path techniques have been easily modified to solve these problems. More recent approaches have included stochastic and time-dependent features of the travel time: passenger arrivals, vehicle arrivals, and travel times may be stochastic. As might be expected, these stochastic processes will greatly affect the path assignment approach (Nuzzulo, 2003). There is considerable evidence that passenger arrivals appear to be Poisson for higher-frequency (lower-headway) service, with headways up to 10–15 minutes. In these cases, the assumption of Poisson passenger arrivals appears to be common. However, at longer headways (lower frequencies), some fraction of passengers may actually time their arrivals with the schedule, which may again significantly complicate the analysis (Turnquist, 1978; Bowman and Turnquist, 1981). Moreover, the treatment of vehicle arrivals may be considered deterministic (according to schedule or at the given headway) or stochastic. If vehicle arrivals are assumed to be Poisson, many of the calculations in the path assignment simplify considerably.

The element of time-dependence, relating to the transit schedules, can also affect the modeling approach. In its simplest case, with perfect adherence to the schedule, the choice of a "minimum cost" path decomposes into a time-dependent shortest path problem. This is usually well solved using variants of existing shortest path techniques; see, for example, Tong and Richardson (1984). However, when some combination of both time-dependence and stochastic travel times are introduced, the problem is not so well behaved. As was shown by Hall (1986), the problem of finding a stochastic and time-dependent shortest path suffers from the fact that subpaths do not necessarily concatenate; instead, a possibly exponential number of paths must be evaluated to ensure an optimal path is found. When vehicle arrivals are random but somewhat correlated with a schedule, the assignment becomes significantly more complicated (Hickman and Bernstein, 1997).

A second complication for transit networks is that there may not be a single route or set of routes which has the minimum cost. This may occur in cases where multiple transit routes may overlap on some part of the origin–

destination path. This problem is commonly referred to as the *common lines* problem, in which a passenger may take one of many routes for at least part of the path from the origin to the destination. The more general case of multiple origin–destination paths has led to the term *strategies* (Spiess and Florian, 1989), reflecting possible boarding rules the passenger may use in traveling from an origin to a destination. In a graph-theoretic model, the subnetwork of eligible paths from the origin to the destination in a strategy is characterized as a *hyperpath* (Nguyen and Pallottino, 1988).

The final characterization that may be made is based on the treatment of capacity and crowding (de Cea and Fernández, 1996, 2000). Much of the early literature in the passenger assignment assumed that vehicle capacity was not typically exceeded, and as a result, capacity and crowding effects could safely be ignored. This allowed certain simplifications of the problem, although this is clearly not applicable in all circumstances. Rather, if passenger volumes are assumed to run close to or over the capacity of a route, it might be expected that passengers may not be able to board the first vehicle to arrive. Hence, waiting time and transfer times may be directly affected by the volume of passengers on the route, creating congestion effects. This congestion affects the problem formulation and solution techniques. As a result, the discussion that follows in Sections 2.2.1 and 2.2.2 is decomposed into the passenger assignment under "uncongested" conditions and "congested" conditions, respectively.

2.2.1 Uncongested assignment

The earliest methods of transit assignment used variants of well-known shortest path algorithms; examples include Dial (1967), Lampkin and Saalmans (1967), le Clercq (1972), Silman et al. (1974), and Last and Leak (1976). In these cases, the full demand of each O–D pair is assigned to a shortest path. The variations from existing shortest path methods are based on two exceptions: waiting times and common lines in the network. These two issues are intertwined. A general formulation of the waiting time, as a function of the frequency, suggests that waiting time is related to the inverse of the frequency of routes serving the passenger. If R_i is the set of routes serving the stop i that also serve the passenger's destination or an intermediate (transfer) node, then the expected waiting time is given as

$$E[WT] = \frac{\alpha}{\sum_{r \in R_i} f_r}, \tag{16}$$

where α is a parameter, such that $\alpha = 1$ for Poisson vehicle arrivals, $\alpha \approx 0.5$ with deterministic arrivals. With the expression in (16), the shortest path problem can be solved by creating additional links in the network representing the corresponding waiting time. Moreover, if more than one route serves a node i and also serves a given intermediate node or destination node for an O–D pair, the assignment is made to each route on the basis of the frequency share. That

is, the fraction of passengers served by route r', $P_{r'}$, is given as

$$P_{r'} = \frac{f_{r'}}{\sum_{r \in R_i} f_r}. \tag{17}$$

This formulation assumes that the passenger takes the first bus to arrive at a stop, among all routes serving that stop.

Chriqui and Robillard (1975) provided a more rigorous treatment of the problem when "common lines" serve some part of an O–D path. Specifically, it may not be to the passenger's advantage to choose the first among all routes serving the pair of stops. The most notable case is where one or more of the routes has a shorter travel time to the destination than the others. In this case, it may be advantageous not to board a bus on a slow route, if it is the first to arrive. Chriqui and Robillard (1975) formulated this problem as follows. Assume the passenger will choose a subset of routes, and will board the first route of this subset to arrive at the origin stop. One may also assume that the travel time to the destination after boarding is constant for each route, but may vary across the set of routes serving the stop. Finally, we assume that the passenger desires to minimize the sum of waiting time and time after boarding. Then, the selection of routes becomes a hyperbolic programming problem of selecting routes to include in this subset. Practically, this problem can be solved to optimality by enumerating the possible route subsets. The result is an optimal route subset R_i^* that can be used to define the waiting time at the node and frequency shares for each route in the subset, using the subset R_i^* in the summations of (16) and (17). Chriqui and Robillard (1975) also derived expressions for waiting times and frequency shares for both uniform- and exponentially-distributed bus arrival times. These results were extended by Marguier and Ceder (1984) and Israeli and Ceder (1996), in which routes can be grouped into slow routes and fast routes.

A related heuristic approach was suggested by Andreasson (1977), in which a route is considered in the desirable subset if the travel time upon boarding a bus on that route is less than or equal to the waiting time plus the travel time upon boarding of the minimum time route. Waiting times and route shares are then determined based on this route set. Andreasson's method was also extended by Jansson and Ridderstolpe (1992), in which they present an iterative heuristic for calculating waiting times and route proportions for transit networks with multiple routes and modes between an O–D pair. Jansson and Ridderstolpe (1992) showed that the functions (16) and (17) can create poor approximations of the waiting time and route shares under deterministic headways; instead, these functions depend heavily on the exact timetable.

The work of Spiess (1983) and Spiess and Florian (1989) introduced the concept of passenger *strategies*. In their formulation, the passenger is assumed to minimize the sum of waiting time and on-board time, where these may vary based on the passenger boarding rules. They formulate this assignment as one to minimize the total travel time in the network, taken as the product of arc

flows and their associated travel times, plus the total waiting time in the network. This waiting time depends on the routes in each strategy. For a given node i, let A_i^+ be the set of arcs leaving i and A_i^- be the set of arcs entering i. Let v_a be the flow on arc a, t_a be the travel time on arc a, and f_a is the frequency of service on arc a. Also, ω_i is the total waiting time experienced by passengers boarding at node i. The demand generated at node i is g_i, and V_i (a parameter) is the total flow entering node i. The formulation of the passenger assignment problem is given as an integer linear program, in which the decision variables x_a are binary variables indicating if an arc a is in the strategy. The relaxation of this problem is

$$\text{minimize} \quad \sum_{a \in A} t_a v_a + \sum_{i \in N} \omega_i \tag{18}$$

subject to:

$$\sum_{a \in A_i^+} v_a - \sum_{a \in A_i^-} v_a = g_i, \quad \forall i \in N, \tag{19}$$

$$\omega_i = \frac{V_i}{\sum_{a \in A_i^+} f_a x_a}, \quad \forall i \in N, \tag{20}$$

$$v_a \leqslant f_a w_i, \quad \forall a \in A_i^+, \forall i \in N, \tag{21}$$

$$v_a \geqslant 0, \quad \forall a \in A, \tag{22}$$

$$0 \leqslant x_a \leqslant 1, \quad \forall a \in A. \tag{23}$$

The dual of this linear program has the form of a shortest path problem, resulting in an optimal label-setting algorithm for its solution.

A similar formulation of the assignment problem is presented by de Cea and Fernández (1989), with the use of nonlinear equality constraints for (21). The resulting nonlinear optimization problem is solved by incorporating the nonlinear constraints into the objective function. The solution technique can then be decomposed into three parts: the selection of common lines (a hyperbolic programming problem) for each O–D pair; the assignment of the O–D volumes to links representing common routes in the hyperpath; and the assignment of common route flows to specific routes by frequency share.

More recent study of passenger assignment has focused on the assignment of passengers to specific scheduled vehicle trips. That is, the assignment identifies a particular vehicle trip to which a passenger is assigned. In this formulation, it is necessary to have a time-dependent origin–destination matrix $D_{ij}(t)$. An extensive discussion of various approaches to schedule-based transit assignment, both for uncongested and congested transit networks, is found in the recent volume edited by Wilson and Nuzzolo (2004).

In Tong and Wong (1999), the time-dependent shortest path technique of Tong and Richardson (1984) is used to assign trips to the network. Additional complications are added using stochastic weights on the various components

of travel time (walk time, waiting time, and transfer time), relative to in-vehicle time.

A combination of stochastic and time-dependent travel time attributes are included in Hickman and Bernstein (1997). This model characterizes the following passenger behavior: upon arriving at a stop, the passenger waits until a bus arrives, and then determines whether or not to board the bus based on the time spent waiting and the set of additional bus arrivals expected in the future. The formulation of this problem essentially requires full enumeration of all paths from the origin to the destination, accompanied by the derivation of the probability distributions of travel times on all paths, at the time the boarding decision is made. Such "clever" passenger behavior can be used to simulate passenger behavior under real-time information, as illustrated by Hickman and Wilson (1995). A similar framework for passenger boarding strategies with information at the stop was also presented by Gentile et al. (2005).

2.2.2 Congested assignment

The area of congested transit assignment has evolved in examining the effects of capacity limitations on passenger path assignment. With vehicle capacities, it may be the case that demand is sufficiently high that a passenger desiring to board a given vehicle may be unable to. This results in higher waiting times when in such crowded conditions. This has led to formal specification of *equilibrium* transit assignment in which the delay that each passenger imposes on other passengers is explicitly included in the model. In general, the effect of crowding and vehicle capacity is incorporated in the modeling through the impact of additional waiting and/or transfer time caused by passengers being unable to board a desired vehicle because it is full. It may also be included as additional "discomfort" experienced by passengers while on board the vehicle. However, in these models it is important to note that the effect of congestion is clearly not symmetrical by arc flows.

Mathematically, the most common approach to include capacity and crowding has been to formulate the waiting time as an increasing function of the volume on a particular line, both in terms of the passengers on-board and the passengers waiting to board at a stop. A graph-theoretic structure for transit equilibrium assignment was developed by Nguyen and Pallottino (1988). These authors proposed the graph concept of a *hyperpath*, defined as an acyclic, directed subgraph of routes connecting the passenger's origin–destination pair. Such a hyperpath may be defined using a particular passenger *strategy* (Spiess and Florian, 1989). Passengers are assumed to travel on the *shortest hyperpaths* connecting their origin and destination.

When equilibrium conditions occur, the problem formulation and solution methods from traditional traffic assignment can be used (refer to Chapter 10 in this volume). Nguyen and Pallottino (1988) suggested using traditional traffic assignment techniques, with the adaptation of these techniques to calculate shortest hyperpaths (or strategies) rather than shortest paths, when calculating a direction for improvement in the assignment.

In de Cea and Fernández (1993), the waiting time is considered to be a function of both the passengers desiring to board at the given stop and those passengers traveling through the stop on board the route. Mathematically, the passenger travel time is calculated as a power function of the *conflicting volume* divided by the capacity to approximate the cost of congestion. The conflicting volume is the sum of the boarding volume and the passengers on board. With congestion on the waiting or boarding arcs, a common approach is to define an *effective frequency* f', determined as the average frequency observed by the passenger, assuming he/she may be denied boarding on the first vehicle in his/her strategy. This effective frequency, calculated as the reciprocal of the expected waiting time for a given route, can be sensitive to the level of crowding upon boarding. In turn, the effective frequency f' can be used in the more traditional waiting time and frequency-based assignment approaches found in (16) and the proportional assignment in (17); for a discussion of these issues, see Bouzaïene-Ayari et al. (2001).

de Cea and Fernández (1993) determined the congestion cost using a linear function of the conflicting volume divided by the capacity. The model uses a variational inequality formulation with nonlinear constraints, with assignment to routes based on the effective frequencies. If the assignment to routes is made on the basis of actual route frequencies, rather than the effective frequencies, the problem has linear constraints. In either case, this problem can be solved using diagonalization, making it susceptible to traditional traffic assignment algorithms.

Wu and Florian (1993) and Wu et al. (1994) extended this work to include a formal strategy formulation of the congested assignment problem. The problem is formulated as an asymmetric network equilibrium problem, but with a variable transformation to solve in the space of hyperpath flows rather than route flows. The solution method uses a symmetric linearization (similar to diagonalization), called the linearized Jacobi method, for solving the resulting variational inequality problem.

Moreover, Cominetti and Correa (2001) extended the models of Wu et al. (1994) to consider an *effective frequency* model of transit assignment, considering possible congestion on the boarding links. The principle finding in this work is the definition of some necessary and sufficient conditions that an equilibrium transit assignment has been reached, both in terms of arc flows and in terms of the strategy. A straightforward cost function is established as an objective. The proposed algorithm performs shortest-hyperpath assignment on the inner loop, while using the method of successive averages to update the flows after each iteration.

Lam et al. (1999) presented a stochastic user equilibrium model for passenger assignment. This model assumes a simple bottleneck model for congestion on a link; i.e., the additional delay is linear with the ratio of the conflicting volume to the capacity. A multinomial logit model is used for the selection of paths. The problem is formulated as a nonlinear programming problem with linear constraints. It is shown that conditions on the Lagrange multipliers in the

Kuhn–Tucker conditions can be specified such that the route capacities are not exceeded. Rather, when route capacities are reached, the bottleneck delays are proportional to the Lagrange multipliers. With this observation, the problem is solved using existing solution techniques for the stochastic user equilibrium problem in traffic assignment (see Chapter 10).

The work of Lam et al. (2002) extended this model to a more disaggregate model of route operations. In a more detailed model of stop operations, the total route travel time is affected by the delay in boarding and alighting at the stop. The frequency on the route, in turn, is determined by the number of vehicles divided by the round-trip travel time. In this case, the waiting time is a function of the frequency, but the frequency itself is a function of the delay caused by crowding. As a result, the assignment is then a fixed point in both the space of frequency and boarding and alighting volumes. This fixed-point problem forms an outer iteration on the assignment and is solved using the method of successive averages.

A more general network model has been presented in the work of Lo et al. (2003). This model accommodates nonlinear fare structures and transfers through the use of a *state-augmented multimodal* (SAM) network. In this network representation, the node itself is augmented based on the opportunities to transfer at the node, the number of transfers made by the passenger in the network, and explicit representation of direct links in the network to represent nonlinear costs. This network is then applied in a stochastic user equilibrium assignment, using a logit model. This framework has also been extended to a nested logit structure by Lo et al. (2004).

Nielsen (2000) considered a stochastic user equilibrium model that is based on congestion both for waiting times as well as in-vehicle discomfort. Both measures are functions of the flow on the associated link and the on-board capacity of the vehicle. The stochastic user equilibrium is based on the probit model for the selection of paths, among common lines (line aggregation). Existing methods for probit-based traffic assignment are used to solve the passenger assignment (see Chapter 10). An application using a nested logit model for the stochastic user equilibrium assignment on a large regional transit network is presented in Nielsen (2004).

As with the uncongested assignment, recent work has been examining assignment to specific vehicle trips in the schedule. A transit equilibrium assignment model that explicitly considers schedules was formulated by Nguyen et al. (2001). In their model, the capacity constraints on boarding and transfer links are hard constraints. The additional delay is a function of the *available capacity* of the vehicle as it arrives. The passenger assignment to routes is based on the passengers' desired arrival times at the destination; a penalty term (schedule delay cost) is added to the passenger cost for arriving at the destination at a time other than the desired arrival time. This problem is set up as a nonmonotonic variational inequality problem and solved using simplicial decomposition, in the form of a column generation technique that generates new extreme points in the arc flow solution space.

A separate line of thinking has been developed by Nuzzulo et al. (2001). In these models, the transit passenger is assumed to make a discrete choice of route and trip, based on the attributes of the trip as well as the desired arrival time at the destination. The discrete choice model uses for the utility function typical variables related to travel times, transfers, and related variables. The choice of run is based on trading off these terms in both a within-day assignment and a day-to-day assignment based on learning, which is accomplished through an exponential filter of run attributes. Congestion is included in the model using a measure of delay in the passenger boarding and in-vehicle time.

Also in the area of trip-based assignment, Poon et al. (2004) have extended the work of Tong and Wong (1999) to congested assignment. In this work, a time-dependent origin–destination demand is given, and the assignment is made in a dynamic network based on specific vehicle trips. Congestion is considered through queuing delay at the stop or station as passengers prepare to board, and through the available space on the transit vehicle when it arrives at a stop. The assignment in a single iteration follows Tong and Richardson (1984) using the latest network travel times, and is performed by moving passengers incrementally in time through the network. The final assignment is solved iteratively by the method of successive averages, with the queuing delay and vehicle loading being updated after each iteration.

Finally, a multiagent approach to transit assignment has been developed by Wahba and Shalaby (2005). In this approach, passengers are represented as agents in the transit network. The passenger behavior is described in terms of their route, stop, and departure time choice, based on a desired arrival time at the destination. This behavior is simulated in the network on a given day, and reinforcement learning is used to represent day-to-day adaptation of passengers to their experiences in the network. This is repeated for many days to achieve a final transit network assignment.

3 Tactical planning

In tactical planning, we are concerned with intermediate steps in the planning process in which the frequencies of routes are constructed and the service schedule is determined. We include these two parts together in the tactical level because they are predominantly oriented toward structuring and improving service to the passenger. In this context, Section 3.1 describes the selection of frequencies on the set of transit routes, and Section 3.2 describes methods to construct timetables.

3.1 Frequency setting

The process of determining frequencies for transit routes has already been introduced in the process of network design (Section 2.1). While a set of frequencies are a necessary product of the network design, it is also true that

a transit agency will evaluate and determine frequencies on routes more often than this. Variations in passenger demand patterns and smaller changes in route design may precipitate a need to adjust frequencies. In this section, we begin with the problem of frequency setting for typical routes, and then we briefly describe methods of determining frequencies for some other types of routes that are commonly used.

The problem of setting frequencies can be approached from several different ways. The primary goal is to select frequencies that maximize the passenger service, which can be defined in a number of different ways, subject to a number of possible constraints. These include constraints on the overall fleet size (which is assumed fixed for this process), a constraint that capacity on a route must be sufficient for the demand, and any policy constraints on minimum desirable frequencies. Other input data include the round-trip and required layover time on each route. With this information, a transit agency will choose an allocation of the fleet to particular routes; this allocation will then directly indicate the frequency of service on each route. Finally, these frequencies may be specified by time of day and day of week.

The most common practical approach is to design frequencies to meet the maximum passenger demands without exceeding the capacity, or without exceeding some threshold value of bus utilization, the ratio of demand to capacity (Ceder, 1984). In cases where this produces unreasonably low frequencies, minimum frequencies (or maximum headways) are also commonly applied. More rigorous mathematical programming models have been developed and are outlined below, but these have rarely been applied because of their complexity.

Early work on this problem focused on determining frequencies with common route structures. Scheele (1980) formulated the problem of determining route frequencies as a nonlinear program, based on minimizing the total generalized passenger travel time, with decision variables being the frequency of each route. Simultaneously, this model solves for the flow on each O–D path (the passenger assignment problem). An O–D path is defined strictly as a sequence of route segments. The formulation includes constraints that the demand cannot exceed the available capacity on a route, flow conservation in the assignment, and fleet size constraints. An entropy constraint is also included to distribute trips in the transit network and to ensure accessibility between all origins and destinations. An iterative solution methodology is proposed in which the set of frequencies is fixed, and the O–D and path flows are determined through a Lagrangian function. From the Lagrangian, a descent direction for the frequencies is determined, and the frequencies are updated. The new frequencies are used to iterate on the assignment, until the frequencies converge.

Similarly, Han and Wilson (1982) formulated the problem of solving for frequencies on each route as an allocation of vehicles among routes in a network. The problem is formulated as one of solving for frequencies, with the constraints being the passenger assignment to individual links, the capacity of each

route, and the total fleet size. In contrast to Scheele (1980), however, the objective is to minimize the maximum "occupancy level" at the maximum load point for each route in the network. To solve this problem, Han and Wilson (1982) proposed a two-stage heuristic as follows. In the first stage, a base allocation is achieved that guarantees that all routes have sufficient frequency so that all passengers are served, but there is 100% utilization on at least one route segment for each route. In this first stage, the passenger assignment, O–D pairs are decomposed into those with only a single path (so-called "captive" flow) and those with multipath ("variable" flow) assignment. For the multipath assignment, a simple frequency-share model is adopted for both direct paths and transfer paths. In the base allocation, the captive flow is assigned in the network, and a lower bound on the frequency for each route is determined based on setting the frequency equal to the maximum link flow divided by the vehicle capacity. Second, an iterative procedure is used in which the "variable" flow is assigned and the frequencies are updated to equal the maximum link flow on each route divide by the vehicle capacity. This process iterates until the "variable" flow on each route segment converges. In the second stage, any remaining vehicles in the fleet are allocated to routes to reduce the utilization uniformly. In the example in the paper, this is achieved by increasing the frequency of all routes directly in proportion to the remaining vehicles in the fleet.

Furth and Wilson (1982) presented a model to determine route headways that maximize the consumer surplus (measured in terms of waiting time) plus the total ridership, as a function of the headway. In this formulation, the demand is a function of the headway, making the total ridership and waiting time dependent on the headway. The formulation includes constraints on the total subsidy, the total fleet size, and maximum headway values (as a policy device). The problem is solved through an algorithm using the Kuhn–Tucker conditions on a relaxation where the maximum headway and fleet size constraints are relaxed. Violations of the maximum headway constraint are projected back to the maximum headway, with associated reductions in the available subsidy. Violations of the fleet size constraint are accommodated with a new set of Kuhn–Tucker conditions where this constraint is binding. The algorithm iterates through these conditions until all routes have similar multipliers. The result is an optimal allocation of buses to routes.

A more complex model, minimizing passenger waiting cost, operating cost, and the cost of vehicle "crowding" was introduced by Koutsopoulos et al. (1985). In their formulation, demand is assumed to be fixed, and constraints on the available subsidy, the maximum fleet size, and available capacity on each route. This is formulated as a nonlinear program, which under certain simplifying assumptions is formulated and solved as a linear program.

Recent work by Gao et al. (2004) has explored a bi-level model for determining line frequencies and the corresponding network assignment. The upper level solves for the optimal frequency of each route, minimizing the total passenger cost. This solution is then iterated with the lower-level passenger

assignment, which is based on the congested assignment model of de Cea and Fernández (1993). The assignment is solved using a diagonalization approach.

Additional work has been done in consideration of special scheduling cases, particularly in high-demand corridors. These include short-turning, zone scheduling (or express services), and deadheading. In short-turning, some vehicles on a route will serve only one segment of the route before returning to the terminal. The objective is to reduce the total number of vehicles serving a route, while still meeting passenger demand and/or minimum levels of passenger service. Furth (1987) presented a model to consider short-turn design in which the objective is to minimize the total fleet size serving a route, with constraints that the load cannot exceed capacity at any point on either the full route or on the short-turn segment. For a given short-turn segment, the problem is cast separately for different multiples of the full route: a 1:1 strategy implies one vehicle on the full route for every one on the short-turn segment; a 1:2 implies one on the full route for every two on the short-turn route, etc. This problem is formulated and solved as a linear program with two decision variables: the frequency on the full route; and the relative offset of the dispatch times on the short-turn route versus the full route. The offset is used to balance the loads on the full route and the short-turn route, depending on the loading pattern over the common route segment. From the continuous linear program solution, a simple rounding technique can be used to find offsets at the nearest minute. Additional deadheading and interlining options are considered in a heuristic technique to reduce the total fleet requirement.

The problem description by Ceder (1989) considered a variety of possible short-turn segments on a route. As an input, a full route schedule is assumed. With this information, Ceder (1989) presented a method to determine which trip segments in the schedule could be eliminated by including short-turn trips. This uses a heuristic to minimize the maximum headway for a route segment when the short-turn trip is introduced. Once the short-turn trips are scheduled, a new estimate of the fleet size is found using the technique of Stern and Ceder (1983), based on the creation of deadheading trips and interlining of trips. Since it works with an existing (initial) route schedule, the technique results in both the definition on the short-turn segments as well as the schedule of the short-turn trips.

A more recent investigation by Site and Filippi (1998) posed the short-turning problem as one of determining the short-turn segment, the types of vehicles to operate on both the full route and the short-turn segment, and the frequency of service on both the full route and the short-turn segment. The objective function is to maximize the net benefits, given as the passengers' consumer surplus less the net subsidy (total costs minus fare revenues) for the operator. Costs for the operator include both capital and operating costs. In this model, passenger demand is endogenously determined as a function of the selected frequency. The model includes constraints to ensure demand does not exceed capacity, and a constraint on the maximum available subsidy. The problem is formulated as a nonlinear program. Site and Filippi (1998) decomposed

the full problem into smaller subproblems; each subproblem is solved for the optimal frequencies, for a given short-turn segment and set of vehicle types. These subproblems are solved heuristically using random search methods, and the global solution is found as the subproblem that maximizes the objective of net benefits.

The zone scheduling (or express service) problem was introduced by Jordan and Turnquist (1979) as a strategy to improve service reliability on bus routes. In the zone scheduling concept, a bus serves only selected segments of a route as it travels from one terminus to the other. In their simplified route model, Jordan and Turnquist (1979) assume that all passengers are destined for a single terminus. Their decision variables are the number of zones, the first stop in each zone, and the number of buses allocated to serve each zone. The problem is formulated as one of minimizing the passenger utility, comprising the expected value and variance of the waiting time and on-board travel time for all passengers. The problem is formulated using a dynamic programming recursion, in which the stages are the number of zones and the state variables are the combination of beginning stop and the number of buses allocated to that zone. The means and variances of waiting times and on-board running times are formulated using a stochastic model of transit service calibrated from data in Chicago.

The work of Furth (1986) extended the model of Jordan and Turnquist (1979) for several additional cases: (1) zone scheduling where additional stops outside the zone can be served for alighting (inbound) and for boarding (outbound); (2) zone scheduling where the zone boundaries are asymmetric, depending on the direction of the trip during any given period; and (3) zone scheduling for branching corridors. Again, dynamic programming is used to solve these cases of the zone scheduling problem.

Finally, Furth (1985) presented a model for deadheading, in which the desirable headway in the off-peak direction is higher than that in the peak direction. In providing service in the off-peak direction, vehicles not in service are deadhead back to the initial terminal in the peak direction. This strategy, while offering less service in the off-peak direction, may allow sufficient time savings for fleet size reduction. The problem of determining the minimum fleet size, for given (fixed) maximum headways in both the peak and off-peak direction, is found by solving a maximum flow problem on a time-space network. When the headway constraints are changed to inequalities (i.e., with only a maximum headway), it is possible that shorter headways could yield fleet size reductions, due to the scheduling requirements for these trips. Furth (1985) presented an algorithm to solve for the minimum number of vehicles, using the ratio of the number of peak trips per outbound trip in service. Related techniques are used to solve for the optimal headways in the peak and off-peak directions for a given fleet size, for two cases: (1) minimizing the total passenger waiting time; and (2) minimizing a combination of passenger waiting time and operating costs.

3.2 Timetabling

Timetabling is the process of converting the desired frequency of service on each fixed route into a schedule. The inputs to this process include the route structure, including running times between major timepoints, the frequency of service, and any necessary layover times at terminals or schedule slack (extra time built into the schedule) at the major timepoints on the route. The result is a set of trips and the scheduled times at the terminals and major timepoints on the route.

In most treatments of transit planning methods, timetabling is included within operational planning, since it occurs frequently, with every service adjustment (e.g., every 3–6 months). Also, it is from the timetables that vehicle and crew schedules are constructed. In this chapter, it is included as an element of tactical planning in the sense that it involves determining the schedule to maximize passenger service. This is in contrast to the vehicle and crew scheduling problems that are typically associated with operations planning, which are intended to minimize transit operating costs.

In many cases, the timetabling is relatively straightforward (Ceder, 1986), primarily working on the assumption that headways are constant and that demand is relatively uniform over the time period of interest. Some clock time is specified for the first vehicle trip of each period, and vehicle departures from a terminal or a maximum load point are set as multiples of the desired headway. Estimated running times between timepoints are used to determine the schedule, and layover times are used to schedule return trips. The simplicity of this task may explain in part why it has received relatively little attention from researchers. In this section, some methods for timetabling are described. However, a major complication in timetabling occurs when schedules are intended to be coordinated at a transfer stop or terminal; methods to create timetables under these circumstances are also presented in this section.

Initial work into the timetabling problem for time-dependent passenger arrivals was suggested by Newell (1971), Salzborn (1972), and Hurdle (1973a, 1973b). These works formulated the timetable problem for a single route with the objective of minimizing passenger waiting time. These results suggest that the optimal rate at which vehicles are dispatched is proportional to the square root of the passenger arrival rate, but with the constraint that vehicle capacity cannot be exceeded. These models were extended by Sheffi and Sugiyama (1982) to consider multiple origin–destination pairs (complicating the derivation of the capacity constraint) and to the case of boarding-dependent dwell times.

The work of Wirasinghe and Liu (1995) considers the problem of determining optimal schedule slack times at intermediate timepoints on a route. Given the running time distributions on the route and a schedule-based holding policy, the research gives a model for determining the slack time by minimizing the sum of the passenger waiting time, the passenger schedule delay, and the

operating cost. The problem is solved using first-order conditions on the objective function.

Perhaps the most pressing challenge in timetabling is the synchronization of vehicle timetables so that transfers within the network are well timed. Specifically, one would like to time the arrival of a vehicle on one route with that on another route so that passengers transferring between routes can make the connection with the minimum waiting time. Much of the early work on this problem focused on methods for synchronization at a single timepoint; more recent methods have used heuristics for larger network problems. However, the combinatorial nature of the problem indicates that it is NP-hard, and the computational issues of exact solutions are still vexing. This limitation is important to note, in that only the more recent approaches have considered more practical problems.

The work of Salzborn (1980) considers two related problems. The first problem is determining the feasibility of scheduling a single transfer route through a series of transfer locations ("interchanges"). Inputs to this analysis include the running times on the transfer route, the slack time built into the transfer route and all connecting services ("feeder routes") at these interchanges, and the minimum time required for passengers to make the transfer. In this analysis, the feasibility of the schedule for the transfer route depends on the ability of the schedule to meet the necessary time windows at each interchange. For the second problem, Salzborn (1980) considers the scheduling of the feeder routes at the interchange, and derives conditions under which a feasible feeder route schedules can be constructed. In this case, the feasible scheduling of the feeder routes requires that the headway on these routes be a multiple of the headway on the transfer route. In addition, the scheduling of the feeder routes requires that departures and arrivals at the interchange be balanced (the total number of departures equals the total number of arrivals) over one headway on the feeder routes, so that time slots at the interchange can be scheduled.

Hall (1985) considers the more specific problem of scheduling at the interchange when the feeder route may be delayed. Under an assumed exponential distribution of this delay, Hall (1985) derives equations for the optimal slack time, based on the objective of minimizing passenger delay. In this case, "slack time" is defined as the time between the scheduled arrival on the feeder route and the scheduled departure on the transfer route. A similar approach was used by Knoppers and Muller (1995) to characterize the optimal slack time on the transfer route, with a normal distribution of arrival times on the feeder route. Knoppers and Muller (1995) also investigate the possible reductions in average waiting time if a holding policy is used: a vehicle on the transfer route is held until the feeder bus arrives. In this case, the optimal slack time can be reduced, with corresponding reductions in the average waiting time. However, the work does not examine the implications of additional waiting at downstream stops from the holding strategy.

There has also been a number of more analytic studies to optimize both the slack time and the headways of connecting routes at a transfer point. Purely

analytic optimization models for a single transfer point have been developed by a number of researchers, including Lee and Schonfeld (1991) and Chien and Schonfeld (1998). More recently, these analytic models have been extended to cover multiple transfer points and multiple modes. Heuristics for this problem, to solve for the slack times and any common headways among routes, have been proposed by Chowdhury and Chien (2002) and Ting and Schonfeld (2005). Both of these studies use a combination of operator costs, from the added slack time, and user costs, including waiting time, transfer time, and in-vehicle time.

A more rigorous treatment of the transfer synchronization problem was presented originally by Klemt and Stemme (1988) for a completely deterministic problem, and later by Bookbinder and Désilets (1992) for the case of random delay. The synchronization problem is defined as one to determine the ideal "offsets" for the schedule for each route r in the set R. Here, an *offset* t_r for a given route r is defined as the minutes after some given time at which the first departure from the terminal is scheduled. If the headway on route r is h_r, the possible offsets are assumed to be in the set of integers $T_r = \{0, 1, \ldots, h_r - 1\}$. Suppose the set of transfer opportunities between routes is given as the set $K = \{1, 2, \ldots, k\}$, with the set A_{ij} describing the complementary set of pairs of routes i and j representing transfer opportunity k (i.e., at the intersection of routes i and j). The utility of a given transfer opportunity is given as $D_k(t_i, t_j)$, and the number of passengers making this transfer is given as n_k. Then, the optimization problem to determine the optimal offsets is given as

$$\text{minimize} \quad \sum_{i \in R} \sum_{j \in R} \sum_{k \in A_{ij}} n_k D_k(t_i, t_j) \tag{24}$$

$$\text{subject to:} \quad t_i \in T_i, \quad \forall i \in R. \tag{25}$$

To solve this model, Bookbinder and Désilets (1992) use a heuristic developed by Rapp and Gehner (1976), in which the offset of each route is determined iteratively. In Bookbinder and Désilets (1992), several utility functions $D_k(t_i, t_j)$ based on the passenger waiting time are evaluated, using a simulation model to generate random transfer arrivals. These are evaluated initially for a single transfer connection and also two small networks with multiple transfer connections.

Ceder and Tal (1999) and Ceder et al. (2001) introduced a model to maximize the number of synchronized connections between routes. The objective is simply to maximize the number of potential connections that can be made. In this formulation, the decision variables include the offset of the initial trip in the schedule (as before), and the headway of each vehicle trip is permitted to vary from some minimum to a maximum headway. This creates more flexibility in the construction of schedules, as each vehicle on the route may have a different headway. The problem is formulated as a mixed integer linear program, and solved using a heuristic that processes the nodes sequentially, using some selection criteria. For each node, the headways of all routes at the node are

matched if possible, and offsets of all connecting routes are set so as to create a simultaneous arrival of all routes at the given node. The heuristic proceeds through the set of interchanges until all vehicle departure times are set.

4 Operational planning

Given a set of timetabled trips to be operated, a set of available resources (buses and drivers) and their distribution among the transit agency depots, the operational planning phase aims at constructing vehicle and crew schedules that minimize total costs while respecting all operational constraints and work regulations. This phase also includes planning the assignment of the buses to parking slots in garages and their dispatch to bus schedules, as well as establishing bus maintenance schedules.

As mentioned in the Introduction, these various problems are usually solved sequentially. Figure 1 illustrates the usual solution sequence and the possible types of feedback. The frequency setting and timetabling tactical problems define the main input for the operational planning phase, namely the timetable. After this, the vehicle scheduling problem is solved first. This stage is crucial to assess whether the proposed timetable can be operated with the available fleet of vehicles and how costly it will be. When the vehicle scheduling problem is infeasible or its operating cost is excessive, a revision of the timetable and possibly the route frequencies is performed in order to facilitate vehicle scheduling. Once a vehicle schedule is determined, driver duty scheduling, which consists of constructing anonymous work days for the drivers, is performed to ensure a complete coverage of the vehicle schedule at a reasonable cost. When this problem is not feasible or too costly, the vehicle schedule must

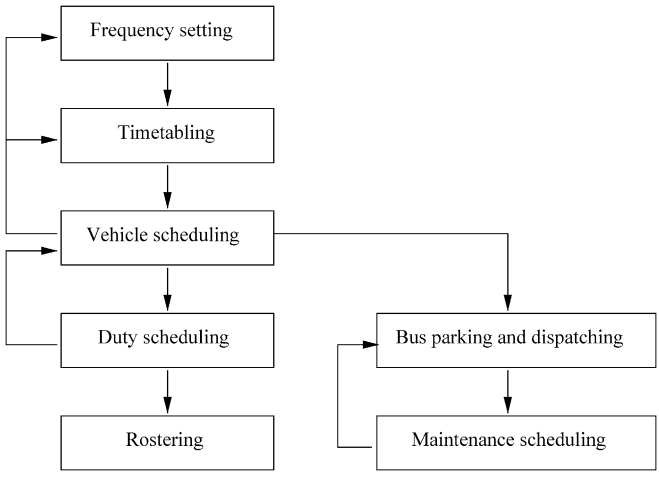

Fig. 1. Relationships between the tactical and operational problems.

be updated to ease the construction of the driver duties. Once the driver duties are known, a rostering problem is solved to establish the personalized driver schedules over a given time horizon (e.g., monthly). This stage rarely requires one to revise the previous decisions since part-time drivers are usually available to provide additional flexibility in the rosters. In parallel to the driver scheduling problems, the bus parking and dispatching problem as well as the bus maintenance scheduling problem are tackled once the vehicle schedule is known. These two problems are more or less solved on a daily basis since they are heavily impacted by the perturbations of the planned vehicle schedule. These two problems are also quite dependent.

Looking at Figure 1, one can see that feedback may frequently occur during the operational planning process. This shows that there is an opportunity for integrating some of these steps. In this regard, Section 4.3 discusses the integration of vehicle and duty scheduling.

4.1 Vehicle scheduling

Vehicle scheduling plays an important role in the management of a public transit agency since it is the first planning step where the primary focus is put on minimizing costs, namely, the acquisition, and operational costs of the buses. Previous steps put a large emphasis on passenger service, which is fixed for vehicle scheduling. Indeed, at this stage, the service offered to the customers is completely determined by the fixed trip timetable. In general, the same daily timetable applies for the weekdays, while a different timetable is defined for days on the weekend. These timetables are usually valid for a certain period of time (a season). Thus vehicle scheduling at a planning level needs to be performed once per timetable and season. It should be noted that in large cities buses operate 24 hours per day, which makes it more difficult to define a daily problem. However, given the low volume of activities during night-time, a 24-hour timetable is split into a day timetable and a night timetable in practice.

The vehicle scheduling problem faced by the public transit agencies corresponds to the single-depot vehicle scheduling problem (SDVSP) when the agency operates its fleet of buses out of a single depot, and to the multidepot vehicle scheduling problem (MDVSP) when several depots are used or when several vehicle types are available. This problem can be stated as follows. Let $\mathcal{T} = \{1, 2, \ldots, n\}$ be a set of n timetabled trips where trip $i \in \mathcal{T}$ starts at time s_i and ends at time e_i. These trips are qualified as *active* since passengers travel along them. Denote by τ_{ij} the travel time (possibly including some layover time) between the end location of trip i and the start location of trip j. We assume that this travel time is the same for all vehicles. Two trips i and j are said to be *compatible* if and only if they can be covered consecutively by the same vehicle (j immediately follows i), that is, if and only if $e_i + \tau_{ij} \leqslant s_j$. The traveling between two such trips is called a *deadhead trip* since there are no passengers on board.

Let $K = \{n+1, n+2, \ldots, n+m\}$ be the set of m depots housing the buses that must be assigned to cover the active trips. Depot $k \in K$ manages v^k identical buses which must start and end their schedule at this depot. A bus leaving a depot to reach the start location of an active trip is said to be performing a *pull-out trip*, while it performs a *pull-in trip* when it returns to the depot from the end location of an active trip. A *feasible schedule* for a bus housed in depot k is composed of a pull-out trip starting at k, a sequence of active trips separated by deadhead trips, and a *pull-in trip* ending at k. Consecutive active trips must be pairwise compatible.

The cost structure is as follows. A cost is incurred each time that a vehicle performs a deadhead, pull-out or pull-in trip. This cost is denoted by c_{ij} for the deadhead trip connecting trips i to j, by c_{kj} for the pull-out trip linking depot k to the start location of trip j, and c_{ik} for the pull-in trip returning to depot k from the end location of trip i. Note that the active trips bear no cost since they represent a fixed amount for any feasible solution. Note also that vehicle fixed costs can be added to the pull-out or the pull-in trip costs. The cost of a schedule is simply the sum of the costs of the trips it contains.

The MDVSP can be defined as the problem of finding a set of feasible vehicle schedules such that each active trip $i \in \mathcal{T}$ is covered by exactly one schedule, at most v^k schedules are defined for each depot $k \in K$, and the sum of the schedule costs is minimized. The SDVSP simply corresponds to the case where $|K| = 1$. In certain versions of the MDVSP additional constraints are considered. For instance, there may exist trip–depot compatibility constraints which restrict the set of depots that can provide a vehicle to perform an active trip, especially when the depots are associated with different vehicle types (see Costa et al., 1995; Löbel, 1998). A soft version of these constraints, when they take the form of preferences rather than strict constraints, can also be handled by defining depot-dependent deadhead costs. Another example of additional constraints consists of imposing a maximum duration or length to every vehicle schedule (Freling and Paixão, 1995). Such a constraint may be needed in a extra-urban context where the potential for driver exchanges is restricted or when fueling considerations must be taken into account.

It should be noted that, in the context of public transit, a deadhead trip that involves a long waiting time before the start of the next active trip is often replaced by a pull-in trip, an idle period at the depot, and a pull-out trip. The vehicle schedules are then seen as sequences of *vehicle blocks*, where each block consists of a sequence of trips that starts and ends at the same depot without returning to it in the middle of the sequence.

The SDVSP arises for small to medium-size transit agencies that rely on a single depot. It can also appear as a subproblem of the MDVSP. The SDVSP is solvable in polynomial time. In fact, it can be modeled as a minimum-cost network flow problem. It has also been formulated as a linear assignment problem, a transportation problem, a quasiassignment problem, and a matching problem. Surveys on the SDVSP and its extensions can be found in Daduna and Paixão (1995) and Desrosiers et al. (1995). Recently, Freling et al. (2001b)

proposed an efficient auction algorithm for solving the quasiassignment formulation of the SDVSP. Based on this algorithm, they also developed a two-phase approach where blocks are built first and combined afterwards, and a core-oriented approach that starts with a network containing a subset of the arcs (the core) and adjusts it iteratively according to the reduced cost of the arcs not considered. The two-phase approach is only valid under certain cost assumptions. Computational experiments showed that these approaches outperform the algorithms previously exposed in the literature.

The MDVSP is common in medium-size transit agencies, and inevitable in larger ones. As proposed by Ribeiro and Soumis (1994), it can be modeled using an integer multicommodity flow formulation as follows. Associate with each depot $k \in K$ a directed graph $G^k = (N^k, A^k)$, where N^k and A^k denote its sets of nodes and arcs, respectively. The node set is defined by $N^k = \mathcal{T} \cup \{k\}$. The arc set is given by $A^k = C \cup (\{k\} \times \mathcal{T}) \cup (\mathcal{T} \times \{k\})$, where C is a subset of $\mathcal{T} \times \mathcal{T}$ that contains an arc $(i, j) \in \mathcal{T} \times \mathcal{T}$ if and only if trips i and j are compatible. Then, define a binary variable X_{ij}^k for each $k \in K$ and each arc $(i, j) \in A^k$ that indicates the flow (0 or 1) of buses originating from depot k on the arc (i, j).

Using this notation Ribeiro and Soumis (1994) formulated the MDVSP as follows:

$$\text{minimize} \quad \sum_{k \in K} \sum_{(i,j) \in A^k} c_{ij} X_{ij}^k \tag{26}$$

subject to:

$$\sum_{k \in K} \sum_{i:(i,j) \in A^k} X_{ij}^k = 1, \quad \forall j \in \mathcal{T}, \tag{27}$$

$$\sum_{j \in \mathcal{T}} X_{k,j}^k \leqslant v^k, \quad \forall k \in K, \tag{28}$$

$$\sum_{i:(i,j) \in A^k} X_{ij}^k - \sum_{i:(j,i) \in A^k} X_{ji}^k = 0, \quad \forall k \in K, j \in \mathcal{T} \cup \{k\}, \tag{29}$$

$$X_{ij}^k \in \{0, 1\}, \quad \forall k \in K, (i, j) \in A^k. \tag{30}$$

The objective function (26) aims at minimizing total costs. Constraints (27) ensure that each active trip is covered exactly once, while constraints (28) limit the number of buses that can be used from each depot. Flow conservation and binary constraints are given by (29) and (30), respectively.

The MDVSP has been studied for more than twenty-five years. Given that it is an NP-hard problem when $m \geqslant 2$ (Bertossi et al., 1987), the early work on this problem focused on heuristic algorithms (for reviews on these methods, see Dell'Amico et al., 1993; Odoni et al., 1994). Since the end of the 1980s, several exact algorithms have been proposed in the literature: Carpaneto et al. (1989), Ribeiro and Soumis (1994), Forbes et al. (1994), Bianco et al. (1994),

Löbel (1998), Mesquita and Paixão (1999), Hadjar et al. (2006), and Kliewer et al. (2006). The results in the last three papers clearly show that real-world large-scale instances can be solved efficiently. In the following we present the methodologies introduced in these three papers.

The approach proposed by Löbel (1998) consists of solving the linear relaxation of the multicommodity network flow model (26)–(30) using a column generation method directly on this formulation; that is, the generated variables are the X_{ij}^k. Before starting the column generation process, a heuristic procedure is used to find a feasible solution. The positive-valued variables of this solution are added to the initial restricted master problem in order to speed up the solution process. Then, at each column generation iteration, the restricted master problem is solved by the dual simplex algorithm and the columns are generated based on a so-called Lagrangian pricing strategy (discussed below) and the standard reduced cost criterion. When the progress in the objective value becomes too small, only the standard reduced cost criterion is used and the restricted master problem is re-optimized by the primal simplex algorithm. Using this methodology, the optimal linear relaxation solution found by Löbel (1998) was already integer for most of the instances treated. When this was not the case, a simple rounding procedure was used to derive an integer solution that was often proved to be optimal.

Lagrangian pricing is a strategy which allows the simultaneous generation of negative reduced cost variables and nonnegative reduced cost variables that complement well the former set of variables. In this way, the column generation process does not require additional iterations to identify these complementary variables as is often the case with the traditional pricing strategy. Given a Lagrangian relaxation of the linear relaxation of the model (26)–(30), the Lagrangian pricing proposed by Löbel (1998) consists of solving the Lagrangian subproblem for a given set of multipliers, namely the values of the dual variables associated with constraints (27)–(29) in the current restricted master problem. All the variables taking a positive value in the solution of this subproblem are then candidates that can be added to the restricted master problem, even if their reduced costs are nonnegative.

Löbel (1998) considered at each iteration two Lagrangian relaxations. In the first, the trip covering constraints (27) are relaxed in the objective function to obtain m independent minimum-cost flow problems that are easily solvable. For the second Lagrangian relaxation, the redundant covering constraints

$$\sum_{k \in K} \sum_{j:(i,j) \in A^k} X_{ij}^k = 1, \quad \forall i \in \mathcal{T}, \tag{31}$$

are added to the formulation (26)–(30) before relaxing constraint sets (28) and (29). The resulting Lagrangian subproblem is also a minimum-cost flow problem that can be solved by inspection. Its solution can however produce bus schedules containing deadhead trips assigned to different depots.

Using this column generation approach, Löbel (1998) reports solving real-world instances from German public transit companies involving up to 49 de-

pots and 24,906 trips. It should be noted however that trip–depot compatibility constraints were considered, yielding a maximum average of 4 depots per trip among these large instances.

Recently, Hadjar et al. (2006) developed a branch-and-bound approach for the MDVSP that combines column generation, variable fixing, and cutting planes. As introduced in Ribeiro and Soumis (1994), traditional column generation is used to compute a lower bound at each node of the branch-and-bound search tree. This column generation process is executed on a set partitioning type reformulation of the MDVSP that can be derived from model (26)–(30) by applying the Dantzig–Wolfe decomposition principle (Dantzig and Wolfe, 1960). In contrast to Löbel (1998) approach, columns in this set partitioning model correspond to vehicle schedules. They are generated by solving shortest path problems.

The variable fixing strategy used by Hadjar et al. (2006) is similar to the one developed by Bianco et al. (1994) and consists of fixing to zero the variables X_{ij}^k that satisfy the following criterion. To simplify notation, we rewrite model (26)–(30) as $\min\{cx \mid Ax = b, x \in \mathbb{Z}_+^\eta\}$, where $x = (x_i)_{i=1}^\eta$ is a vector of η $(= \sum_{k \in K} |A^k|)$ variables, $\mathbb{Z}_+$ is the set of nonnegative integers, and the equality $Ax = b$ is, in fact, an inequality $(Ax \leqslant b)$ for constraint set (28). Denoting by $\bar{x}$ a feasible solution to this problem and by $\bar{\pi}$ a feasible solution to the dual of its linear relaxation, a variable x_i can be set to zero if its reduced cost is greater than or equal to $c\bar{x} - \bar{\pi}b$. Hadjar et al. (2006) computed a first feasible solution $\bar{x}$ by performing a depth-first search without backtracking in the branch-and-bound tree and imposing multiple decisions at each branching node. At each node of the branching tree, part of the dual solution $\bar{\pi}$ is provided by the dual solution produced by the column generation method at this node and the remainder is found by solving shortest path problems. This variable fixing strategy, which is performed at each node of the tree, can fix over 90% of the X_{ij}^k variables in most instances treated by Hadjar et al. (2006).

Hadjar et al. (2006) proposed to add at each node of the search tree cutting planes that are related to the odd cycles in the MDVSP underlying network. These valid inequalities are lifted through a heuristic procedure. The authors showed that, under certain conditions, the lifted inequalities define facets of the convex hull of the feasible solution set. With this approach, Hadjar et al. (2006) succeeded in solving randomly generated MDVSP instances that involve up to 6 depots and 750 trips. It should, however, be mentioned that these results are difficult to compare with the results obtained by Löbel (1998) or Kliewer et al. (2006) (see below) because the characteristics of the test problems differ significantly from one paper to the other.

Instead of using model (26)–(30), Kliewer et al. (2006) developed a multicommodity network flow model based on a time-space network. In fact, when several trips start and end at the same terminus, a substantial reduction in the number of variables can be obtained by using a sequence of waiting arcs at each terminus instead of representing explicitly all possible connections. Kliewer

et al. (2006) also applied an aggregation procedure for reducing the number of arcs representing potential deadhead trips. The resulting model is solved to optimality with the CPLEX MIP solver. With this approach, the authors report solving large real-world instances. In particular, they solved one instance that involves 7068 trips, 5 depots, and 124 termini in approximately 3 hours of computational time.

An interesting extension to the MDVSP is the possibility of changing the scheduled departure times of the trips within certain time intervals, called *time windows*. Such flexibility on departure times, which can be considered when the frequency on a route is not too high, can often yield significant savings by providing additional possible deadhead trips that would be infeasible otherwise. To our knowledge, this extension has been tackled first by Mingozzi et al. (1995) who adapted the methodology they have developed for the MDVSP in Bianco et al. (1994). Their approach consists of solving by branch-and-bound a set partitioning model that contains a reduced set of columns. The reduction is obtained by variable fixing as described above. More recently, Desaulniers et al. (1998b) have proposed a branch-and-price approach for this extension that generalizes the work of Ribeiro and Soumis (1994). They also showed that, with a slight modification, this approach is capable of handling an exact cost on the waiting time occurring between two consecutive trips. Given the time windows, such waiting times and their ensuing waiting costs cannot be computed a priori; they must be computed during the solution process.

4.2 Duty scheduling

Duty scheduling, also known as driver scheduling, is the second step in the process of planning the operations for a public transit agency. As with the vehicle scheduling problem, the duty scheduling problem (DSP) is important from an economic point of view since it determines most of the wages paid to the drivers. The DSP is separable by depot and consists of determining the work days (also called *duties*) of the drivers based at a depot in order to cover all the vehicle blocks assigned to this depot. Since a driver exchange can occur at various points along a vehicle block, all blocks are divided into a sequence of *segments* according to these *relief points*. The consecutive segments along a block assigned to the same driver are collectively called a *piece of work*. Duties are therefore composed of pieces of work that are usually separated by breaks. Different duty types, differing, for instance, by the number of pieces of work they can contain and their possible starting times and durations, can be considered. As examples, there may exist straight duties that contain a single piece of work, and split duties containing two pieces of work. Duties are subject to a wide variety of safety regulations and collective agreement rules such as a maximum duty spread, a maximum duration of a piece of work, and a predefined time interval in which a break must be awarded. These rules vary according to the duty type.

In general, the objective of the DSP is twofold and consists of minimizing first the total number of duties and second the total number of worked hours. Using duty fixed costs and an hourly rate for the worked hours, this objective is usually transformed into one that minimizes total cost. In summary, the DSP can be stated as follows. Given the segments of a set of vehicle blocks, find a set of valid duties that covers all these segments and minimizes total cost. Additional constraints such as a limit on the number of duties of a certain type can also be taken into account.

The DSP can be formulated as a set partitioning problem that relies on the following notation. Let S be the set of block segments to cover and D the set of valid duties. Denote by c_d the cost of duty d (including fixed costs and wages) and by a_d^s a binary parameter that takes value 1 if duty d covers segment s, and 0 otherwise. Finally, define a binary variable Y_d for each duty $d \in D$ that indicates if duty d is retained in the solution. Using this notation the DSP is formulated as

$$\text{minimize} \quad \sum_{d \in D} c_d Y_d \tag{32}$$

subject to:

$$\sum_{d \in D} a_d^s Y_d = 1, \quad \forall s \in S, \tag{33}$$

$$Y_d \in \{0, 1\}, \quad \forall d \in D. \tag{34}$$

The objective function (32) consists of minimizing total duty costs. Constraint set (33) ensures that a driver is assigned to each block segment. Finally, binary requirements on the Y_d variables are expressed by (34). It should be noted that all work rules defining the validity of the duties are taking into account in the definition of set D. In some cases, this set or part of it can be enumerated a priori using an enumeration algorithm that considers these rules (for instance, see Smith and Wren, 1988). Otherwise, it can be defined implicitly as a set of constrained paths in one or several networks, where the constraints are used to model the complex work rules (for instance, see Desrochers and Soumis, 1989).

Several authors formulate the DSP as a set covering model that allows the over-covering of each segment (the equalities in (33) are replaced by greater-than-or-equal-to inequalities). This over-covering is usually not acceptable, but solving this model often produces a solution that contains very little or no over-covering at all, especially when assigning one driver to a segment is cheaper than assigning several drivers to it. In this case, over-covering can usually be easily eliminated a posteriori using a heuristic procedure. The main advantage of a set covering model over a set partitioning model is its flexibility which allows more rapid computation of a feasible continuous solution and good heuristic integer solutions.

The DSP, modeled as a set partitioning or a set covering problem, is much more difficult to solve than the MDVSP due to the complexity of the work

rules and the huge number of duties that are valid in real-world DSP instances. Indeed, since some of these rules can only be modeled using nonlinear relationships, one has to rely on a formulation similar to (32)–(34) to avoid explicit nonlinear constraints. This formulation however contains a huge number of variables in practice, making it very difficult to solve to optimality.

As surveyed in Wren and Rousseau (1995), several heuristic approaches were proposed before the 1990s. One of the most successful consists essentially of generating a priori a subset of the valid duties and solving a set partitioning/covering model (32)–(34) restricted to this subset (for instance, see Smith and Wren, 1988). The subset of valid duties is composed of promising duties that offer various possibilities for covering each block segment. The restricted model is generally solved by a heuristic integer linear programming method.

Research on this type of approach is still ongoing. In 1999, an attempt was made by Curtis et al. (1999) to solve the restricted set partitioning model using a hybrid constraint programming/linear programming heuristic method where the linear programming solutions are used to guide variable and value ordering in the constraint programming algorithm. One nice feature of this methodology is that its constraint programming component is capable of handling nonlinear constraints which can arise from certain work rules. However, the authors could not solve instances involving more than 203 segments and 26 duties. More recently, Fores et al. (2002) have incorporated a column generation strategy to the solution process of the restricted set covering model. Column generation, which is applied only at the root node of the branch-and-bound search tree, consists of generating as needed negative reduced cost columns from a superset of a priori enumerated valid duties. This novelty improves the quality of the solutions while slightly increasing solution times. For instance, they can solve medium-size instances involving close to 90 duties in one hour of computational time.

Column generation for the DSP was introduced by Desrochers and Soumis (1989). In their approach the master problem corresponds to the linear relaxation of the set covering model. The subproblems are constrained shortest path problems which are solved by a generalized version of the dynamic programming algorithm of Desrochers and Soumis (1988) (see Desaulniers et al., 1998a; Irnich and Desaulniers, 2005). Integer solutions are found by a branch-and-bound scheme, where column generation is used at each node of the search tree to compute a lower bound. The overall approach obtains optimal solutions for small instances and near-optimal solutions for larger instances. In 1995, Desrosiers and Rousseau (1995) reported solving DSP instances involving 156 duties using a commercial version of this branch-and-price approach.

In the last fifteen years, branch-and-price approaches have been applied with success to a wide variety of vehicle routing and crew scheduling problems. When the size of these problems are huge as in real-world DSPs, accelerating strategies, various heuristics, and stabilization techniques, such as the ones reported in Barnhart et al. (1998), du Merle et al. (1999), and Desaulniers et al. (2002), must be included in these approaches to produce good quality solutions

in acceptable computational times. In a recent paper on the DSP, Borndörfer et al. (2003) pursued this direction by presenting a heuristic branch-and-price approach that calls upon several speed-up strategies. Firstly, instead of solving the master problem by the simplex or barrier algorithm, they proposed to solve its dual using a heuristic coordinate ascent method combined with a boxstep stabilization method. Secondly, the constrained shortest path subproblems are solved by an enumerative algorithm that relies on lower bounds computed by Lagrangian relaxation. Finally, they suggested a heuristic branching scheme without backtracking that fixes at each branching node a duty variable Y_d to one. This variable is selected from a set of twenty candidate variables using a probing strategy. For each candidate variable, this strategy fixes it to one and evaluates the impact of this decision on the master problem solution value without generating additional columns. The selected variable is the one yielding the smallest deterioration of the master problem solution value. After branching, when this deterioration is less than a predetermined threshold value, no columns are generated and variable fixing is performed again. Using this customized column generation approach, Borndörfer et al. (2003) solved large real-world DSP instances involving close to 2000 segments and over 110 duties.

Freling et al. (1999, 2003) proposed to solve the linear relaxation of the DSP by a column generation approach where the master problem is approximately solved by Lagrangian relaxation. At each column generation iteration, a constrained shortest path subproblem is solved in two main steps: first, pieces of work are generated by solving an all-pairs shortest path problem; and second, duties are generated from these pieces of work by solving a constrained shortest path problem. To avoid generating the same column twice, the Lagrange multipliers are modified before generating new columns in such a way that all columns already generated get a nonnegative reduced cost. Once the linear relaxation is solved, a set covering problem involving all columns generated along the way is heuristically solved to find a feasible integer solution. As reported in Freling et al. (2003), this approach can easily solve small-size DSP instances with 238 segments and 24 duties. No results on larger instances are reported, however, since the DSP was solved only for comparison with an integrated vehicle and crew scheduling approach that is discussed in the next section.

One drawback of the column generation approach that relies on constrained shortest path subproblems is the inability to model all work rules that may define the validity of a duty. One way of overcoming this drawback is to ignore these rules at the subproblem level and simply to reject all generated columns that violate them. Another way that was suggested by Borndörfer et al. (2003) consists of defining an infeasible path constraint for each identified illegal column and including these constraints in the subproblems. A third alternative has been proposed by de Silva (2001), who suggested formulating the subproblems as flexible constraint programming models and solving them using constraint programming tools. He succeeded in solving real-world DSP instances with complex work rules involving up to 495 segments.

Recently, Gintner et al. (2004) proposed a DSP approach that benefits from the fact that several vehicle schedules may be optimal. Indeed, given a feasible vehicle schedule, the segments derived from the blocks of this schedule can be rearranged into different blocks yielding the same vehicle costs. Their DSP model allows for this possibility by enumerating all feasible pieces of work and duties that can be obtained from the segments. Their approach solves the linear relaxation of this model using column generation and computes, using the CPLEX MIP solver, an integer solution for the restricted set covering model containing only the columns generated. This approach was tested on random datasets involving up to 400 trips with one segment per trip. These tests showed that substantial savings in the number of duties can be achieved from this additional flexibility.

Research investigating the use of metaheuristics for solving the DSP has also been carried out recently. Kwan et al. (1999, 2001) proposed a genetic algorithm that relies on the linear relaxation solution of a restricted set covering model to identify important traits that should appear in the optimal integer solution. Shen and Kwan (2001) developed a tabu search approach that involves multiple neighborhoods and an appropriate memory scheme. Lourenço et al. (2001) introduced a genetic and a tabu search algorithm to solve DSPs involving multiple objectives such as minimizing the number of duties, minimizing the number of duties with a single piece of work, minimizing the number of vehicle changes, and minimizing the over-covering when allowed. Both of these algorithms use for large instances a greedy randomized adaptive search procedure (GRASP) as an intensification tool. All these metaheuristics are fast and produce solutions that are comparable (in terms of quality) to the solutions produced by an approach based on a restricted set partitioning/covering model, similar to that of Smith and Wren (1988). For instance, Shen and Kwan (2001) reported solving a DSP involving 859 segments and 106 duties in less than 18 minutes with their tabu search algorithm.

4.3 Integrated vehicle and duty scheduling

In general, vehicle scheduling is performed before duty scheduling in the operational planning process of a public transit agency. Since driver relief opportunities are numerous in most contexts, an efficient duty schedule can often be obtained from a near-optimal bus schedule to yield an overall high-quality solution. On the other hand, when these relief opportunities are rare, as is the case in extra-urban mass transit systems or for a line-by-line scheduling process, a very efficient vehicle schedule may lead to a poor duty schedule or even to an infeasible DSP. Integrating vehicle scheduling and duty scheduling is therefore essential in these situations, and research on this topic has been conducted recently.

The integrated vehicle and duty scheduling problem (IVDSP) can be stated as follows. Given a set of timetabled trips and a fleet of vehicles assigned to several depots, find minimum-cost vehicle blocks and valid driver duties such

that each active trip is covered by one block, each active trip segment is covered by one duty, and each deadhead, pull-in, and pull-out trip (hereafter called an *inactive trip*) used in the vehicle schedule is also covered by one duty. As in the MDVSP, each block must start and end at the same depot and, as in the DSP, driver duties must comply with a set of work rules and each duty must be composed of trips that are covered by buses originating from the same depot. This last requirement is often mandatory since drivers are usually assigned to a depot. Additional constraints such as vehicle availability can also be imposed.

Next, we propose a formulation for the IVDSP that combines the models presented above for the MDVSP and the DSP. Besides the notation introduced in the previous two sections, the following notation is required. Let D^k be the set of valid duties for a driver assigned to depot k, and b_{dij} be a binary parameter equal to 1 if duty $d \in D^k$ covers the trip associated with arc $(i, j) \in A^k$ and to 0 otherwise. For each depot $k \in K$ and each duty $d \in D^k$, we define a binary variable Y_d^k that takes the value 1 if duty d is selected and the value 0 otherwise.

The proposed formulation for the IVDSP is as follows:

$$\text{minimize} \quad \sum_{k\in K} \sum_{(i,j)\in A^k} c_{ij} X_{ij}^k + \sum_{k\in K} \sum_{d\in D^k} c_d Y_d^k \tag{35}$$

subject to:

$$\sum_{k\in K} \sum_{i:(i,j)\in A^k} X_{ij}^k = 1, \quad \forall j \in \mathcal{T}, \tag{36}$$

$$\sum_{j\in\mathcal{T}} X_{k,j}^k \leqslant v^k, \quad \forall k \in K, \tag{37}$$

$$\sum_{i:(i,j)\in A^k} X_{ij}^k - \sum_{i:(j,i)\in A^k} X_{ji}^k = 0, \quad \forall k \in K, j \in \mathcal{T} \cup \{k\}, \tag{38}$$

$$X_{ij}^k \in \{0, 1\}, \quad \forall k \in K, (i, j) \in A^k, \tag{39}$$

$$\sum_{k\in K} \sum_{d\in D^k} a_d^s Y_d^k = 1, \quad \forall s \in S, \tag{40}$$

$$\sum_{d\in D^k} b_{dij} Y_d^k - X_{ij}^k = 0, \quad \forall k \in K, (i, j) \in A^k, \tag{41}$$

$$Y_d^k \in \{0, 1\}, \quad \forall k \in K, d \in D^k. \tag{42}$$

The objective (35) minimizes the sum of the vehicle and duty costs. Constraints (36)–(39) define the vehicle scheduling problem. They are identical to (27)–(30). Constraint sets (40) and (42) are the counterparts of (33) and (34) for the multidepot case. Finally, constraints (41) establish the link between the vehicle schedule and the duty schedule; that is, each inactive trip covered by a bus must also be covered by a duty assigned to the depot from which this bus originates.

In comparison with the DSP, the IVDSP is highly combinatorial since the inactive trips are unknown a priori; they have to be determined by the optimization process. In consequence, the number of possible valid duties is very large especially when multiple depots are considered. Given its complexity and its lesser importance, the IVDSP has not been addressed in the literature as much as the MDVSP and the DSP. Indeed, as surveyed in Freling et al. (1999), only a few heuristic approaches have been proposed in the 1980s and the early 1990s. However, it seems that this problem has lately attracted the attention of several researchers who developed solution approaches based on mathematical programming decomposition techniques.

Freling et al. (1999, 2003) addressed the single-depot IVDSP where no bus availability constraints are considered but the main objective consists of minimizing the overall number of buses and duties required to cover all active trips. Similar to the approach they proposed for the DSP, they developed a column generation approach where the master problem is solved by Lagrangian relaxation. In this case, all constraints involving duty variables are relaxed in the Lagrangian function, yielding a Lagrangian subproblem that corresponds to pricing out the duty variables and solving a single-depot vehicle scheduling problem. Thus, a feasible vehicle schedule is computed each time that the Lagrangian subproblem is solved. When the linear relaxation of the IVDSP is satisfactorily solved using this process, the last computed vehicle schedule is kept and used to define a DSP that is solved by their DSP column generation approach (see Section 4.2) to derive a feasible duty schedule. In Freling et al. (2003), the authors report solving real-world IDVSP instances involving up to 148 segments and 23 duties in reasonable computation times. Their results also show that small gains in the total number of buses and duties can be attained by solving the IVDSP instead of solving the vehicle scheduling problem and the duty scheduling problem sequentially. These gains are more substantial when drivers are not allowed to change buses after a break (see also Freling et al., 2001a).

In 2001, Haase et al. (2001) introduced a formulation that only involves duty variables and one bus counter variable which is used to apply a fixed cost per bus. This model can be partially derived from model (35)–(42), adapted to the single-depot case and without availability constraints, by substituting the X variables according to their definition in constraint set (41). Bus-count constraints, similar to the plane-count constraints of Klabjan et al. (2002), are added to complete the model. These constraints provide lower bounds on the number of buses required at specific times of the horizon, namely each time that a bus can leave the depot to reach just in time the beginning location of an active trip. Solving this model provides optimal duties and ensures that an optimal vehicle schedule can be obtained a posteriori using a simple polynomial-time procedure. To do so, Haase et al. (2001) proposed a branch-and-price approach that relies on several accelerating strategies such as dynamically generating the bus-count constraints and reducing the average number of nonzero elements in the constraint coefficient matrix by an appropriate constraint sub-

stitution. Two versions of this approach are presented: an exact version where branching is performed at the subproblem level, and a heuristic version where multiple branching decisions on the duty variables are made at every branching node. With the exact version of the algorithm, randomly generated IVDSP instances involving up to 400 segments and 60 duties were solved in less than 3 hours, while the heuristic version succeeded in solving instances with 700 segments and 121 duties within the same time frame.

Recently, Elhallaoui et al. (2005) developed a dynamic constraint aggregation algorithm for speeding up the solution process of set partitioning type problems solved by a column generation approach. This exact algorithm aggregates and disaggregates, as needed, the set partitioning constraints in order to reduce the size of the master problem and degeneracy. They tested this new approach on the single-depot IVDSP instances of Haase et al. (2001). They report reducing the time needed for solving the linear relaxation by up to 80% on instances involving up to 1280 segments. Furthermore, they observed that the number of fractional-valued variables in a linear relaxation solution decreases considerably with this methodology, yielding high expectations to compute rapidly optimal integer solutions.

The multidepot version of the IVDSP has been investigated in Huisman et al. (2005), where the authors presented two formulations for this problem that are generalizations of the single-depot models developed in Freling et al. (2003) and in Haase et al. (2001). Hereafter, we refer to these formulations as the MD-FHW model and the MD-HDD model, respectively. Two similar solution approaches, that are adaptations of the approach proposed for the single-depot case in Freling et al. (2003), are also proposed. Both approaches contain two phases: the first phase computes a lower bound on the optimal value, while the second one finds a feasible solution. The lower bound is computed by approximately solving a linear relaxation using a combined column generation/Lagrangian relaxation method. The first approach relies on the linear relaxation of the MD-FHW model while the second one uses that of the MD-HDD model. The second approach also includes a special treatment of the bus-count constraints which are added one at a time. The second phase is identical for both approaches. A heuristic feasible vehicle schedule is found by applying Lagrangian relaxation on the MD-FHW model, where only the duty variables generated during the first phase are considered. Once this schedule is established, a duty schedule is computed for each depot using the DSP approach proposed in Freling et al. (1999, 2003). A series of comparative tests on real-life and randomly generated datasets involving up to 653 segments showed that both integrated approaches can solve these instances to yield substantial savings when compared to the traditional bus-first, duty-second sequential approach. Furthermore, neither of the integrated approaches could clearly outperform the other one, even though the second one regularly provided weaker lower bounds than those produced by the first approach. To reduce solution times and solve larger instances, de Groot and Huisman (2004) devised and

compared different heuristic strategies for splitting an instance into smaller ones which are thereafter solved individually by an integrated approach.

For the multidepot IVDSP, Borndörfer et al. (2004) used an integer programming formulation that essentially combines together model (26)–(30) for the MDVSP and model (32)–(34) for the DSP and adds synchronization constraints between the buses and the drivers on the deadhead, pull-in and pull-out trips. They proposed a heuristic solution approach based on a Lagrangian relaxation of these synchronization constraints. The Lagrangian dual is solved by a proximal bundle method and integer solutions are obtained through a heuristic branch-and-bound procedure. With this approach, they report solving large real-world instances.

4.4 Crew rostering

Given a set of anonymous duties defined over a certain time horizon (typically, a week or a month) for the drivers assigned to a particular depot, crew rostering consists of assigning these duties to the available drivers to form their work schedules (called *rosters*). As with the duties, the validity of the rosters is restricted by safety regulations and collective agreement rules. For instance, a driver cannot work more than a certain number of consecutive days. In most North American public transit agencies, drivers build their own rosters in order of seniority, leaving no place for optimization. On the other hand, in many European agencies, the main objective of the crew rostering problem is to distribute the work load evenly among the drivers, yielding an interesting optimization problem.

As surveyed in Odoni et al. (1994), the common practice for solving transit rostering problems consists of first solving a sequence of assignment problems to build an initial solution and then using a local improvement procedure to better this solution. In the first phase, for each day of the horizon, an assignment problem is defined to assign the duties of the corresponding day to the partial rosters that were built by the previous assignment problems. The cost structure aims at balancing the workload among the drivers. It can also incorporate bonuses to account for the preferences of the drivers for certain duties. An iteration of the second-phase heuristic procedure can be, for example, to select a day, divide all the rosters into two parts according to that day, and solve an assignment problem to match the first parts of the current rosters with possibly different second parts. Such heuristic approaches were developed in the mid-1980s and are still in use due to their computational speed and their flexibility with regards to the work rules.

From a mathematical programming point of view, the crew rostering problem can be formulated as a set partitioning or a set covering problem where a row is defined for each duty and a column is associated with each valid roster. Solving such a model is however not popular for crew rostering problems encountered in public transit systems. This is in contrast to the air and rail contexts where various mathematical programming approaches based on a

set partitioning/covering formulation of the crew rostering problem have been proposed lately in the literature. One major difference between transit and air/rail rostering problems appears to be in the size of the real-world instances, which is larger for transit problems. Indeed, the transit problems are not separable per vehicle type since the drivers are usually allowed to drive all the buses. Furthermore, in these problems, the tasks to cover correspond to individual duties, while they correspond to tours of duties (also known as pairings) that may span up to six days in air/rail rostering problems. Nevertheless, we think that most methodological advances in air/rail crew rostering can be adapted for public transit rostering (at least for small- and medium-size problems). We thus refer the interested reader to Chapters 1 and 2 of this book for a review of the latest advances in air and rail crew rostering.

4.5 Parking and dispatching

An operational planning problem that has attracted little attention in the literature is the management of the parking area in vehicle depots. In congested cities, depots are often restrained in space and quite crowded from late evening to early morning. They also contain different types of buses that are needed by particular bus routes. Therefore, when a bus of a particular type has to leave the depot in the morning, several other buses might need to be moved to clear the way out, resulting in a delay. Two alternatives can be considered to avoid delays. The first one consists of always assigning a directly accessible bus to every morning pull-out even if the bus type is not the one requested for the corresponding bus schedule. In this case, we say that a *mismatch* occurs. The second alternative is to reorder the buses during the night so that they all are properly positioned for the morning pull-outs. An exchange of parking slots between two buses is called a *crossing* or a *maneuver*.

Given a sequence of bus arrivals during the evening, a set of timetabled pull-outs in the morning, and a required bus type for each pull-out, the vehicle parking and dispatching problem consists of parking the buses in the depot upon their arrival and dispatching them to the pull-outs such that the number of mismatches or crossings is minimized while satisfying the following constraints. For safety reasons, buses are not allowed to go backwards in the depot. Therefore, assuming that the depot is made up of lanes that operate as queues, buses enter the lanes at one end and exit them at the other. Obviously, lane capacity must not be exceeded. Finally, given the limited space available to perform crossings, they are only permitted between vehicles of the same lane.

The vehicle parking and dispatching problem is an operational planning problem that usually needs to be solved on a daily basis due to the high variability of the bus availability per type. This variability arises from regular maintenance requirements and unexpected breakdowns. To our knowledge, Winter and Zimmermann (2000) were the first to introduce this problem,

which was defined for tram operations. They showed that it is an NP-hard problem and formulated it as a quadratic assignment model with side constraints. By linearizing this model and adding valid inequalities, they could only solve small-size instances to optimality using the CPLEX MIP solver. Consequently, they proposed heuristics for solving larger instances.

In 2001, Gallo and Di Miele (2001) addressed the vehicle parking and dispatching problem in the context of mass transit buses. They proposed an integer programming model, suitable for both objectives stated above, that relies on three variable types: a first type to assign arriving buses to lanes, a second type to assign morning pull-outs to lanes, and a third type to identify the matchings between the buses and the pull-outs. To solve this model, they developed a three-step heuristic approach. In the first step, Lagrangian decomposition is applied to fix the values of the first two types of variables. In this decomposition, these two types of variables are duplicated to yield two generalized assignment subproblems (one for the arrivals and the other for the pull-outs) that are solved by the CPLEX MIP solver, and a set of *design noncrossing matching* subproblems (one for each lane) which can be solved in polynomial time. In this context, a design noncrossing matching problem consists of matching arriving buses with morning pull-outs with no crossings while selecting, as design decisions, the subsets of buses and pull-outs to consider in the associated lane. A bundle method is used for solving the Lagrangian dual problem. In the second step, after assigning the buses and pull-outs to the lanes according to the last computed solutions in the first step, a simplified design noncrossing matching problem (the design decisions are fixed) is solved for each lane to obtain a complete solution that may contain undesirable crossings or mismatches. A heuristic procedure is then invoked in the third step in an attempt to improve this solution. Using this three-step approach, Gallo and Di Miele (2001) reported solving real-life instances involving up to 4 bus types, 12 lanes, and 77 buses in a few minutes.

Very recently, Hamdouni et al. (2006) argued that an optimal solution to the vehicle parking and dispatching problem, as stated in Winter and Zimmermann (2000) and in Gallo and Di Miele (2001), may be difficult to use in practice due to the randomness of the bus arrival times. Indeed, such a solution may contain a large number of pairs of consecutive slots in a lane to which buses of different types are assigned. Each such pair of slots is likely to lead to a mismatch during the operations if the buses planned for these slots arrive in reverse order. In order to increase the solution robustness, Hamdouni et al. (2006) proposed a restricted definition for the vehicle parking and dispatching problem in which a lane can contain a maximum of two bus types, each bus type being confined to a single block of consecutive parking slots. In this case, a maximum of one pair of consecutive slots per lane is susceptible to mismatches. They also suggest replacing the objective of minimizing the number of crossings by the objective of minimizing the number of lanes that need to be reordered. This suggestion is motivated by the fact that all the buses in a lane must be moved out of the depot when a crossing must be performed

in that lane. Therefore, performing crossings in a lane costs approximately the same independently of the number of crossings to perform. For this version of the problem, Hamdouni et al. (2006) have presented an integer programming model that is based on an enumeration of the possible patterns that can be used to divide a lane into a maximum of two blocks. This model is solved by the CPLEX MIP solver after adding a series of cuts that reduce the feasible region without hindering the search for an optimal solution. Real-world instances involving up to 4 bus types, 16 lanes, and 144 buses were solved to optimality in less than twenty seconds of computation time.

4.6 Maintenance scheduling

Another area where operations research can be helpful in planning public transit operations is bus maintenance scheduling. Since maintenance costs are one of the largest expense categories in a typical transit system, some transit agencies are now investing in maintenance scheduling systems to help them reduce maintenance costs while maintaining a reliable, safe, and attractive transit system. Unfortunately, maintenance systems are not applicable everywhere. When parking depots have limited space, bus assignment is performed on a daily basis according to the day's parking pattern and it is impossible to predict how many miles each bus will travel on the following days. In this case, buses are simply withdrawn from the fleet when they are due for maintenance and maintenance resources are available. These last-minute withdrawals can often reduce service quality. On the other hand, for spacious parking depots where almost all parking slots are directly accessible at any time, it might be desirable to assign buses to specific vehicle schedules that are valid for a long period of time. The rationale behind this strategy is that drivers can then be assigned to the same bus every day in hope that they will be more sensitive to mechanical anomalies of their bus and report minor problems before they become major ones. In this context where vehicle schedules are fixed for a long time horizon, a maintenance system can be devised for scheduling maintenance activities when buses are not supposed to be in service, and in such a way to maximize maintenance resource utilization.

As stated in Haghani and Shafahi (2002), the bus maintenance scheduling problem can be defined as follows. Given the buses' operating schedules, their maintenance requirements, and maintenance resource and crew availability, schedule buses for maintenance and assign them to existing facilities such that each bus is maintained in time, while the amount of time that the buses are out of service is minimized. Maintenance requirements, involving various types of maintenance, are expressed in terms of a maximum mileage or number of days in between two consecutive maintenance activities. Note that maintenance facilities cannot all be used for all types of maintenance.

To our knowledge, the bus maintenance scheduling problem has only been addressed by Haghani and Shafahi (2002). These authors presented three integer programming models for this problem. The first model is very general but

unsolvable in practice. The second one relies on the assumption that the buses requiring maintenance for each type are identified and sorted in the order they should be maintained. This assumption often can be held in practice. The third model is a network model with side constraints that also relies on assumptions that are usually valid in small to medium-size agencies, namely: regular inspections can be performed in all maintenance bays; and, when a bus is due for more than one inspection in a planning period, it is possible to compute for these inspections nonoverlapping maintenance intervals in which these activities will be scheduled. Haghani and Shafahi (2002) proposed three heuristic approaches for solving the second model and a fourth heuristic for the third model. The first two heuristics are simple branch-and-bound methods, while the third heuristic fixes a large number of variables to zero based on the linear relaxation solution, before solving the resulting problem by an exact branch-and-bound scheme. Finally, the fourth heuristic is a network-based algorithm. Using each of these four approaches, Haghani and Shafahi (2002) solved a series of instances arising from simulated operations involving 181 buses and five maintenance types. The results show that all heuristics can solve these instances in less than five minutes on average to yield acceptable solutions. As mentioned by these authors, future research on this problem can focus on developing better heuristic procedures as well as exact solution approaches based on decomposition methods.

5 Real-time control

In actual operations, a wide variety of exogenous and endogenous factors can affect service delivery, such as weather, incidents, variations in traffic conditions, vehicle breakdowns, etc. These factors may degrade the level of service experienced by transit passengers. For this discussion, we differentiate *minor* and *major* service disruptions: minor disruptions are those that create small perturbations from the schedule (e.g., 5–10 minutes), and major disruptions cause longer breaks in the schedule. The distinction is made in order to differentiate the typical responses to these service problems.

To address these challenges, a transit operator may employ a variety of operations control techniques (Turnquist, 1981; Levinson, 1991). These will generally vary depending on the magnitude of the perturbation to service. In normal service with only minor perturbations from the schedule and small service disruptions, vehicle holding and transit signal priority are the most common techniques that are applied. In holding, a vehicle may be held at a stop to improve passenger service. The first part of this section (Section 5.1) addresses the vehicle holding problem. For transit signal priority, transit vehicles may be given preferential treatment in signal timing in order to move more rapidly through a signalized intersection. For reasons of scope, signal priority is not discussed in this chapter.

When major service disruptions occur, more serious control measures may be considered. These can include skipping stops on a route, including *expressing* over parts of the route or skipping particular stops, in order to catch up on the schedule. It may also be useful to re-position a vehicle on the route. In *short-turning*, a vehicle on a route is emptied and placed in service traveling in the other direction on the route, in order to accommodate passenger demand in the other direction. Real-time *deadheading* may also be employed to re-locate a vehicle to another part of the route (or to another route entirely) where it may be of greater service. In addition, extra vehicles and drivers can be made available, to be inserted into service as the need arises. Models and methods for these types of control measures are outlined in Section 5.2.

5.1 Vehicle holding

The transit vehicle holding problem has been explored by many researchers over the past 30 years. Early approaches to this problem have generally focused on either *threshold-based* holding or *schedule-based* holding. The threshold technique involves holding a vehicle only if the preceding headway is below a certain amount of time (e.g., the desired headway or some other threshold value). In this case, the vehicle is held only until the threshold time and then dispatched. If the vehicle arrives after the threshold value, it is dispatched immediately. On the other hand, schedule-based holding involves holding a vehicle only until its scheduled departure time; if it arrives later than the scheduled time, it is dispatched immediately. More recently, in contrast to such holding policies, models have been developed to determine optimal holding times for each vehicle individually. In all of the modeling approaches, the objective is to minimize the total passenger delay (or waiting time), as measured by the delay or waiting time for passengers waiting to board, passengers already on board, and passengers at downstream locations. Generally, for cases where passengers may arrive at stops according to a printed schedule, the delay is measured in terms of the deviation from the schedule. In cases where passengers arrive randomly at the stops, the average passenger waiting time was given by Welding (1957) in the following expression:

$$E[WT] = \frac{E[H]}{2}\left(1 + \frac{\mathrm{Var}[H]}{E[H]^2}\right), \tag{43}$$

where $E[WT]$ is the expected waiting time per person, $E[H]$ is the expected headway, and $\mathrm{Var}[H]$ is the variance of the headway.

Analytic approaches, considering idealized routes with stochastic service characteristics, were studied in the 1970s. The analytic work at this time focused on optimal threshold policies for simplistic networks, as analytic models for optimal threshold-based holding policies were not easily found for more realistic problems. Partly as a consequence, most subsequent analysis schedule-based and threshold-based holding from the 1970s through the 1990s has been conducted using simulation, rather than analytic techniques.

In contrast to methods to find an optimal threshold value, Barnett (1974) introduced a model to solve directly for the optimal holding time of each vehicle at a control stop. Barnett (1974) derived approximations for optimal holding times at a single control point along a transit route, using simplified two-point discrete distributions of the vehicle lateness to that point. The optimal holding time for each vehicle is found by minimizing the expected waiting time, given as a quadratic function of the holding time. The optimal holding times are a function of the mean and variance of the headway distribution, the ratio of the passenger load at the control stop to the load downstream, and the covariance of vehicle arrivals at the control stop.

Turnquist and Blume (1980) extended the analysis by Barnett (1974) to consider the effectiveness of holding decisions. They show that holding will only serve to reduce the passenger waiting time if

$$\mathrm{COV}[H] > \frac{0.5\gamma}{1-\gamma}, \tag{44}$$

where COV[H] is the coefficient of variation of headways at the control stop and γ is the ratio of passengers on board at the control stop to those at downstream stops. The obvious implications of this formula are that control is best implemented when the coefficient of variation of headways is large, and/or when the ratio of on-board passengers to downstream passengers is small. This can be used to select whether and at what locations holding is implemented.

More rigorous analytic methods to solve for individual vehicle holding times have only emerged more recently. Most of the recent holding models include detailed models of transit operations, such as dwell times, passenger boarding and alighting processes, and minimum headway and capacity constraints. For holding decisions, most models assume that vehicle dwell times at stops are modeled explicitly as a linear function of the number of boarding and/or alighting passengers, plus any holding time added at the control stop. More critically, this level of detail allows a certain realism in the modeling of the effect of holding decisions: once a hold is effected, the dwell time of subsequent vehicles at the same stop, and their trajectories downstream, will be changed.

The model formulation of Adamski and Turnau (1998) addressed the problem of minimizing schedule deviations on route. The problem is formulated as an optimal control problem, and the operating dynamics are explained through a set of linear difference equations as a vehicle moves across stops on a route. Using these linear difference equations and a quadratic objective function in the vehicle departure times, the determination of a control (a holding time) can be solved through traditional control methods. This approach appears to be most applicable for maintaining schedule adherence on lower-frequency transit lines, where schedule adherence may be more important than maintaining regular headways.

For higher-frequency service, headway regularity becomes the dominant factor in minimizing passenger waiting time. Holding in this case becomes a problem of adjusting vehicle headways to minimize this variability. Here, we

present the vehicle holding problem formulation based on the work of Eberlein (1995) and Eberlein et al. (2001). A transit route has stations $K = \{1, \ldots, k_t\}$, where k_t is a terminus with sufficient layover time to recover from a minor service disruption. A control is exerted only at stop k, and affects only downstream stops from k to k_t. Let $K' = \{k+1, \ldots, k_t\}$ be the set of downstream stops. Vehicle trips affected by the hold are in the set $I_m = \{i, i+1, \ldots, i+m-1\}$ where trip m is not controlled. The decision variables are the departure times for each vehicle at the control stop ($d_{j,k}$ for vehicle j at the control stop k). The headways, then, are measured as the time between consecutive departures from each stop.

In the model of operations, passengers arrive at the stop k' at a rate of $\lambda_{k'}$. The load on board vehicle j after leaving stop k' is $L_{j,k'}$, and the percentage alighting at any stop k' is given by $q_{k'}$. Also, $a_{j,k'}$ is the arrival time of vehicle j at stop k', $s_{j,k'}$ is the dwell time of vehicle j at stop k' to allow passengers to alight and board, $R_{k'}$ is the running time from stop $k'-1$ to k'. The dwell time is based on a linear function of the alighting and boarding passengers, where c_0 is a constant term, c_1 is the incremental time necessary for one passenger to board, and c_2 is the incremental time for one passenger to alight. Also, delay may propagate at a downstream timepoint k_c if the departure time d_{j,k_c} after passenger alighting and boarding is greater than the scheduled departure time t_{j,k_c}. The departure time d^0_{j,k_c} is the departure time at this timepoint if no control action is taken. Finally, no vehicle may enter a stop for a period of time $h_{\min}$ after a vehicle has departed, and the minimum headway upon entering the stop must be at least h_0.

The formulation is as follows, where θ is a weight on the delay to on-board passengers compared to waiting passengers:

$$\text{minimize} \quad \sum_{j \in I_m} \sum_{k'=k}^{k_t} \frac{\lambda_{k'}}{2} h^2_{j,k'} + \theta L_{i,k}(d_{i,k} - a_{i,k} - s_{i,k}) \tag{45}$$

subject to:

$$d_{j,k} - a_{j,k} - s_{j,k} \geqslant 0, \quad \forall j \in I_m, \tag{46}$$

$$d_{j,k'} - a_{j,k'} - s_{j,k'} = 0, \quad \forall j \in I_m, \forall k' \in K', \tag{47}$$

$$d_{i+m,k} - a_{i+m,k} - s_{i+m,k} = 0, \tag{48}$$

$$a_{j,k'} - d_{j-1,k'} \geqslant h_{\min}, \quad \forall j \in I_m, \forall k' \in K', \tag{49}$$

$$d_{j,k_c} \geqslant \max\{t_{j,k_c}, d^0_{j,k_c}\}, \quad \forall j \in I_m, \tag{50}$$

$$a_{j,k'} = \max\{d_{j,k'-1} + R_{k'}, d_{j-1,k'} + h_0\}, \quad \forall j \in I_m, k' \in K' \cup \{k\}, \tag{51}$$

$$s_{i,k'} = c_0 + c_1 \lambda_{k'} h_{i,k'} + c_2 q_{k'} L_{i,k'-1}, \quad \forall j \in I_m, k' \in K' \cup \{k\}, \tag{52}$$

$$h_{j,k'} = d_{j,k'} - d_{j-1,k'}, \quad \forall j \in I_m, k' \in K' \cup \{k\}, \tag{53}$$

$$L_{j,k'} = \lambda_{k'} h_{j,k'} + (1 - q_{k'}) L_{j,k'-1},$$
$$\forall j \in I_m, k' \in K' \cup \{k\}. \tag{54}$$

The first term in (45) gives the total waiting time experienced by passengers at the current stop and all downstream stops, and the second term is the weighted objective value of delay to passengers on board during the hold. The constraint sets (46)–(54) account for the most commonly used dynamics of operations across the different vehicles and stops. In this formulation, (46)–(48) account for arrival and dwell times at stops; (49)–(50) account for maintaining minimum time separation between vehicles at all stops and at the next timepoint, respectively; (51) accounts for run time between stops ensuring a minimum headway; (52) gives a linear accounting of the dwell time for boarding and alighting passengers at a stop; (53) defines the headways in terms of consecutive departure times; and (54) gives the load on board the vehicle upon leaving a stop.

The formulation in Eberlein et al. (2001) excluded the second term in the objective function (the delay imposed to those on board). The formulation includes as decision variables the departure times $d_{j,k}$ for all vehicles $j \in I_m$ at the control stop k. Note that once these departure times are determined, the system evolution is automatic using (47)–(54). Also, this formulation of the vehicle holding problem is a quadratic program but generally not convex. To solve the problem, a heuristic to optimize each departure time, sequentially across vehicles from i to $i + m$ and iterating until convergence, is proposed.

A similar mathematical formulation is presented by Zhao et al. (2001), but using a more general formulation of the cost function. However, the formulation includes only one decision variable, the holding time of the current vehicle i. The proposed solution technique uses a multiagent system approach, in which the vehicle and the set of impacted stations engage in negotiation using the marginal costs of the proposed holding time. This method is proved to be optimal for convex cost functions.

An extension of these models is described by Sun and Hickman (2004). This work considers the use of multiple holding stations along a route, in order to minimize a weighted sum of passenger waiting costs and on-board delay. A heuristic is proposed which solves for the optimal holding times of each vehicle at its next holding station. The models are solved sequentially, beginning at the holding station furthest downstream and moving upstream along the route. The solution at each holding station is solved using a steepest descent method.

The vehicle holding problem formulation by Hickman (2001) differs from these previous formulations in that the analysis explicitly includes stochastic elements, in contrast to these strictly deterministic models. In Hickman (2001), the passenger arrival and alighting processes and the vehicle running times are considered stochastic. The objective function and the operational dynamics incorporate these processes through the expected values and variances of the vehicle headways and loads. The total passenger waiting time, based on (43),

yields the following objective function, minimizing over the holding time $t \geqslant 0$:

$$\text{minimize} \quad \sum_{k'=k}^{k_t} \frac{\lambda_{k'}}{2} \sum_{j \in I_m} \left(\text{Var}[h_{j,k'}|t] + E[h_{j,k'}|t]^2 \right) + \theta E[L_{i,k}]t, \quad (55)$$

where the variables are defined as before. This objective is minimized, subject to the operational dynamics including both expectations and variances of headway and load. To this end, the operations model of Marguier (1985) was used, in which the expectations and variances of vehicle headways and loads are formulated as linear difference equations. As a result, the model becomes a (convex) quadratic optimization problem with linear constraints, in the single decision variable of the holding time t. A simple line search is proposed to find the optimal holding time, while accounting for these operational dynamics.

Another version of the holding problem considers holding strategies for vehicles at a timed transfer terminal. In this condition, passengers arriving late on one route may not be able to make a transfer to a connecting route. The purpose of terminal holding is to hold a vehicle so that transfer passengers can make a connection. In this situation, the objective includes the delay to passengers on board or downstream on the held route, and the delay to passengers wishing to transfer. A hold reduces the delay to transfer passengers while increasing the delay to passengers on board or downstream. The critical variables include the lateness of vehicles at the terminal and the volume of transfer, boarding, and downstream passengers on each route.

In a deterministic operating environment where the passenger boarding and transfer passenger loads are known, and the lateness of any vehicle is known with certainty, the problem reduces to the situation where either the vehicle is not held, or it is held until the moment when a vehicle on another route arrives. In Hall et al. (2001), this model was extended to stochastic vehicle arrivals, giving a distribution of vehicle lateness on each route. In this case, analytic methods may be used to determine the optimal holding time for each route. The objective function reaches a global minimum either with no holding or at one of the local minima in the neighborhood of each expected vehicle arrival time. An extension of this model to the case of technology that allows one to forecast vehicle arrivals at a transfer terminal is described in Dessouky et al. (2003).

5.2 *Other strategies*

There have been a variety of other control strategies that have been analyzed using mathematical programming techniques. In all the cases cited here, deterministic models of transit operations are used. This means that the objective function uses the waiting time as a function of the square of the headway.

Li et al. (1991, 1992) presented a model in which stop-skipping and holding are considered simultaneously in order to bring a route back on schedule after a service disruption. In this case, skipping stops will reduce dwell times for a

vehicle, reducing the preceding headway. This may reduce the average passenger waiting time, but this effect is weighed against the extra waiting time for passengers whose stops are skipped. With this in mind, the objective function is to minimize the total passenger waiting time along the route, by selecting for each vehicle which stops to include in its trip. The decision variables include binary variables indicating if vehicle j is to stop at stop k, and the continuous variable of the departure time of each vehicle at each stop. Constraints include the vehicle operating dynamics and the vehicle capacity. Three solution heuristics are proposed that iterate among local improvement techniques, with each iteration considering changes in only a subset of decision variables.

Fu et al. (2003) proposed a variation on stop-skipping in which every second vehicle is considered for stop-skipping. In this way, at most one vehicle will pass a stop before it is served. The problem is formulated as a nonlinear 0–1 programming problem, and is solved separately for each dispatched vehicle. The objective function includes passenger waiting time, passenger in-vehicle time, and the bus travel time. Constraints include the typical operations dynamics of passenger boarding and alighting processes and bus running times. The problem is solved using explicit enumeration for each bus. Results suggest that this can be solved in real-time for routes with a small number of potential stops; there were 14 stops in their case study.

These previous studies assume that a stop-skipping decision is made before the vehicle is dispatched from a terminal. Sun and Hickman (2005) extended this concept to consider a real-time stop-skipping policy, made while the vehicle is traveling on the route. Based on the latest vehicle location and disruption information, the problem is formulated as solving for a skipping segment along the route, considering passenger waiting time and in-vehicle time. The model solves for the start and end point of the skipped segment using explicit enumeration. A route with 41 stops was analyzed in their case study, with the model being solved in real time (i.e., in seconds of CPU time).

Eberlein (1995) examined the real-time control actions of expressing and deadheading. The expressing problem is defined as determining if a vehicle should skip over a route segment. In this definition, both the starting stop and ending stop of the *express segment* are determined. However, only one express segment is considered per vehicle. In the deadheading problem, a vehicle is to be dispatched from a terminal, and the decision faced is whether to begin revenue service at the terminal or to begin revenue service further down the route. If deadheading is preferred, a *deadhead segment* is created over which the vehicle runs empty. The deadheading work was later published in a separate paper (Eberlein et al., 1998). Because of similarities in the formulation of these problems, only the deadheading work is presented here.

In Eberlein (1995) and Eberlein et al. (1998), this problem was formulated as a nonlinear integer program in order to identify at what downstream stop to resume revenue service. The complication of vehicle dynamics make an analytic solution impossible and significantly complicates the solution using mathematical programming techniques. However, under simplifying assumptions

about the passenger boarding and alighting processes and the vehicle running times in the deadhead segment, the problem becomes analytically tractable. With this simplification, the objective function (minimizing total passenger waiting time) is convex in the number of stops to skip. Using a continuous relaxation, the end of the deadhead segment is calculated as a real value, which is then rounded up or down to the nearest integer to find the stop that minimizes the objective function.

In Eberlein (1995) and Eberlein et al. (1999), a set of models are proposed to simultaneously examine holding, deadheading, and expressing. It is observed that the strategy of holding has the opposite effect of deadheading and expressing: holding a vehicle will lengthen the preceding headway, while deadheading and expressing a vehicle will shorten the preceding headway. As a result, at most one control strategy will be applied to a single vehicle i. This means the control problems are separable, and an efficient heuristic is applied. In the first step, a holding time is determined, following the heuristic from Eberlein et al. (2001). If station skipping is feasible, deadheading and expressing are considered for vehicle i, and holding is considered for subsequent vehicles $i+1$ to $i+m$ to minimize the total passenger waiting time.

Two other recent studies have examined additional operations control strategies under service disruptions, particularly for rail lines. O'Dell and Wilson (1999) performed a comparison of holding and short-turning. The holding model formulation has as its objective minimizing the passenger waiting time along the route, but uses a piecewise linear function as an approximation of the traditional quadratic objective function. Hard capacity constraints are included, resulting in integer constraints and, as a result, a mixed integer linear program. The short-turning model extends the holding model to accommodate a vehicle that may be turned around on the route at a control stop; it is also formulated as a mixed-integer linear program. Commercial software is used to solve both the holding and short-turning models for a set of disruption scenarios.

Shen and Wilson (2001) extended the model of O'Dell and Wilson (1999) to include expressing, in addition to short-turning and holding. The objective function again includes a piecewise linear approximation to the quadratic function of passenger waiting time, but includes a large number binary variables for whether or not to skip a stop (during expressing) and whether or not to short-turn a vehicle. Additional linear approximations of several nonlinear constraints are used to create a mixed-integer linear program. Commercial mixed-integer programming software is used to solve these models.

Finally, a model has been presented by Li et al. (2004) which considers bus re-routing to accommodate vehicle breakdowns. Such a model can be used to insert a replacement vehicle or re-assign existing service vehicles when a vehicle must unexpectedly be removed from service. The model uses an auction heuristic for the multidepot vehicle scheduling problem (MDVSP) to solve practical instances in real time.

6 Conclusion

This survey chapter has reviewed the operations research literature applied to the domain of public transit, with a focus on recent contributions. It has highlighted a fruitful cooperation between the public transit agencies and the operations research community. Indeed, public transit has provided interesting and challenging problems to operations research, while operations research has been successful at solving efficiently several important public transit problems (for instance, network design, timetabling, vehicle scheduling, and crew scheduling). Research on these problems is still going on with the aim of developing new solution approaches or improving existing ones that will allow to solve larger instances and to address additional complexities such as stochasticity and complicated operational rules that were previously ignored.

This survey has also shown that new problems (integration of vehicle and crew scheduling, bus parking and dispatching, as well as a wide variety of real-time control problems), presenting new challenges to the operations research community, have also been studied recently. Research on these problems has already suggested innovative models and solution methodologies which might be applicable in practice in a near future.

This fruitful collaboration between transit agencies and operations research will certainly continue for a long time as transit agencies continue to strive to provide a good quality service at minimum cost, with continual pressure from budgetary restrictions. Operations research should therefore remain an essential tool for helping the agencies plan and run their operations efficiently.

References

Books from CASPT conferences in chronological order

Wren, A. (Ed.) (1981). *Computer Scheduling of Public Transport*. North-Holland, Amsterdam.

Rousseau, M. (Ed.) (1985). *Computer Scheduling of Public Transport 2*. North-Holland, Amsterdam.

Daduna, J.R., Wren, A. (Eds.) (1988). *Computer-Aided Transit Scheduling*. *Lecture Notes in Economics and Mathematical Systems*, vol. 308. Springer-Verlag, Heidelberg.

Desrochers, M., Rousseau, M. (Eds.) (1992). *Computer-Aided Transit Scheduling*. *Lecture Notes in Economics and Mathematical Systems*, vol. 386. Springer-Verlag, Heidelberg.

Daduna, J.R., Branco, I., Paixão, J.M.P. (Eds.) (1995). *Computer-Aided Transit Scheduling*. *Lectures Notes in Economics and Mathematical Systems*, vol. 430. Springer-Verlag, Heidelberg.

Wilson, N.H.M. (Ed.) (1999). *Computer-Aided Transit Scheduling*. *Lecture Notes in Economics and Mathematical Systems*, vol. 471. Springer-Verlag, Heidelberg.

Voss, S., Daduna, J.R. (Eds.) (2001). *Computer-Aided Scheduling of Public Transport*. *Lecture Notes in Economics and Mathematical Systems*, vol. 505. Springer-Verlag, Heidelberg.

Hickman, M., Mirchandani, P., Voss, S. (Eds.) (in press). *Computer-Aided Scheduling of Public Transport*. *Lecture Notes in Economics and Mathematical Systems*, Springer-Verlag, Heidelberg, in press.

Cited references

Adamski, A., Turnau, A. (1998). Simulation support tool for real-time dispatching control in public transport. *Transportation Research – Part A* 32 (2), 73–87.

Andreasson, I. (1977). A method for the analysis of transit networks. In: Roubens, M. (Ed.), *Advances in Operations Research*. North-Holland, pp. 1–8.

Baaj, M.H., Mahmassani, H.S. (1990). TRUST: A LISP program for the analysis of transit route configurations. *Transportation Research Record* 1283, 125–135.

Baaj, M.H., Mahmassani, H.S. (1992). Artificial intelligence-based system representation and search procedures for transit route network design. *Transportation Research Record* 1358, 67–70.

Baaj, M.H., Mahmassani, H.S. (1995). Hybrid route generation heuristic algorithm for the design of transit networks. *Transportation Research – Part C* 3 (1), 31–50.

Barnett, A. (1974). On controlling randomness in transit operations. *Transportation Science* 8, 102–116.

Barnhart, C., Johnson, E.L., Nemhauser, G.L., Savelsbergh, M.W.P., Vance, P.H. (1998). Branch-and-price: Column generation for solving huge integer programs. *Operations Research* 46, 316–329.

Bertossi, A.A., Carraresi, P., Gallo, G. (1987). On some matching problems arising in vehicle scheduling models. *Networks* 17, 271–281.

Bianco, L., Mingozzi, A., Ricciardelli, S. (1994). A set partitioning approach to the multiple depot vehicle scheduling problem. *Optimization Methods and Software* 3, 163–194.

Bielli, M., Caramia, M., Carotenuto, P. (2002). Genetic algorithms in bus network optimization. *Transportation Research – Part C* 10 (1), 19–34.

Bookbinder, J.H., Désilets, A. (1992). Transfer optimization in a transit network. *Transportation Science* 26 (2), 106–118.

Borndörfer, R., Grötschel, M., Löbel, A. (2003). Duty scheduling in public transit. In: Jäger, W., Krebs, H.-J. (Eds.), *Mathematics – Key Technology for the Future*. Springer-Verlag, New York, pp. 653–674.

Borndörfer, R., Löbel, A., Weider, S. (2004). A bundle method for integrated multi-depot vehicle and duty scheduling in public transit. ZIB-Report 04-14, Konrad-Zuse-Zentrum für Informationstchnik, Berlin, Germany. In: Hickman, M., Mirchandani, P., Voss, S. (Eds.), *Computer-Aided Transit Scheduling. Lecture Notes in Economics and Mathematical Systems*. Springer-Verlag, Heidelberg, in press.

Bouzaïene-Ayari, B., Gendreau, M., Nguyen, S. (2001). Modeling bus stops in transit networks: A survey and new formulations. *Transportation Science* 35 (3), 304–321.

Bowman, L.A., Turnquist, M.A. (1981). Service frequency, schedule reliability and passenger wait times at transit stops. *Transportation Research – Part A* 15 (6), 465–471.

Carpaneto, G., Dell'Amico, M., Fischetti, M., Toth, P. (1989). A branch and bound algorithm for the multiple vehicle scheduling problem. *Networks* 19, 531–548.

Ceder, A. (1984). Bus frequency determination using passenger count data. *Transportation Research – Part A* 18 (5–6), 439–453.

Ceder, A. (1986). Methods for creating bus timetables. *Transportation Research – Part A* 21 (1), 59–83.

Ceder, A. (1989). Optimal design of transit short-turn trips. *Transportation Research Record* 1221, 8–22.

Ceder, A., Israeli, Y. (1998). User and operator perspectives in transit network design. *Transportation Research Record* 1623, 3–7.

Ceder, A., Tal, O. (1999). Timetable synchronization for buses. In: Wilson, N.H.M. (Ed.), *Computer-Aided Transport Scheduling*. Springer-Verlag, Heidelberg, pp. 245–258.

Ceder, A., Wilson, N.H.M. (1986). Bus network design. *Transportation Research – Part B* 20 (4), 331–344.

Ceder, A., Golany, B., Tal, O. (2001). Creating bus timetables with maximal synchronization. *Transportation Research – Part A* 35, 913–928.

Chien, S., Schonfeld, P. (1998). Joint optimization of a rail transit line and its feeder bus system. *Journal of Advanced Transportation* 32 (3), 253–284.

Chowdhury, S., Chien, S. (2002). Intermodal transit system coordination. *Transportation Planning and Technology* 25, 257–287.

Chriqui, C., Robillard, P. (1975). Common bus lines. *Transportation Science* 9 (1), 115–121.

Cominetti, R., Correa, J. (2001). Common-lines and passenger assignment in congested transit networks. *Transportation Science* 35 (3), 250–267.

Costa, A., Branco, I., Paixão, J.M.P. (1995). Vehicle scheduling problem with multiple type of vehicles and a single depot. In: Daduna, J.R., Branco, I., Paixão, J.M.P. (Eds.), *Computer-Aided Transit Scheduling. Lecture Notes in Economics and Mathematical Systems*, vol. 430. Springer-Verlag, Heidelberg, pp. 115–129.

Curtis, S.D., Smith, B.M., Wren, A. (1999). Forming bus driver schedules using constraint programming. In: *Proceedings of the 1st International Conference on the Practical Application of Constraint Technologies and Logic Programming (PACLP99)*. The Practical Application Company, London, pp. 239–254.

Daduna, J.R., Paixão, J.M.P. (1995). Vehicle scheduling for public mass transit – an overview. In: Daduna, J.R., Branco, I., Paixão, J.M.P. (Eds.), *Computer-Aided Transit Scheduling*. *Lectures Notes in Economics and Mathematical Systems*, vol. 430. Springer-Verlag, Heidelberg, pp. 76–90.

Dantzig, G.B., Wolfe, P. (1960). Decomposition principle for linear programs. *Operations Research* 8, 101–111.

de Cea, J., Fernández, E. (1989). Transit assignment to minimal routes: An efficient new algorithm. *Traffic Engineering and Control* 30 (10), 491–494.

de Cea, J., Fernández, E. (1993). Transit assignment for congested public transport systems: An equilibrium model. *Transportation Science* 27 (2), 133–147.

de Cea, J., Fernández, E. (1996). An empirical comparison of equilibrium and non-equilibrium transit assignment models. *Traffic Engineering and Control* 37 (7), 441–445.

de Cea, J., Fernández, E. (2000). Transit-assignment models. In: Hensher, D.A., Button, K.J. (Eds.), *Handbook of Transport Modelling*. *Handbooks in Transport*. Elsevier, Amsterdam, pp. 497–508.

de Groot, S.W., Huisman, D. (2004). Vehicle and crew scheduling: Solving large real-world instances with an integrated approach. Report EI2004-13, Econometric Institute, Erasmus University of Rotterdam, The Netherlands. In: Hickman, M., Mirchandani, P., Voss, S. (Eds.), *Computer-Aided Transit Scheduling. Lecture Notes in Economics and Mathematical Systems*. Springer-Verlag, Heidelberg, in press.

de Silva, A. (2001). Combining constraint programming and linear programming on an example of bus driver scheduling. *Annals of Operations Research* 108, 277–291.

Dell'Amico, M., Fischetti, M., Toth, P. (1993). Heuristic algorithms for the multiple depot vehicle scheduling problem. *Management Science* 39 (1), 115–125.

Desaulniers, G., Desrosiers, J., Ioachim, I., Solomon, M.M., Soumis, F., Villeneuve, D. (1998a). A unified framework for deterministic time constrained vehicle routing and crew scheduling problems. In: Crainic, T.G., Laporte, G. (Eds.), *Fleet Management and Logistics*. Kluwer Academic, Norwell, MA, pp. 57–93.

Desaulniers, G., Lavigne, J., Soumis, F. (1998b). Multi-depot vehicle scheduling with time windows and waiting costs. *European Journal of Operational Research* 111, 479–494.

Desaulniers, G., Desrosiers, J., Solomon, M.M. (2002). Accelerating strategies for column generation methods in vehicle routing and crew scheduling problems. In: Ribeiro, C.C., Hansen, P. (Eds.), *Essays and Surveys in Metaheuristics*. Kluwer Academic, Norwell, MA, pp. 309–324.

Desrochers, M., Soumis, F. (1988). A generalized permanent labeling algorithm for the shortest path problem with time windows. *INFOR* 26, 191–212.

Desrochers, M., Soumis, F. (1989). A column generation approach to the urban transit crew scheduling problem. *Transportation Science* 23, 1–13.

Desrosiers, J., Rousseau, J.-M. (1995). Results obtained with crew-opt: A column generation method for transit crew scheduling. In: Daduna, J.R., Branco, I., Paixão, J.M.P. (Eds.), *Computer-Aided Transit Scheduling*. *Lectures Notes in Economics and Mathematical Systems*, vol. 430. Springer-Verlag, Heidelberg, pp. 349–358.

Desrosiers, J., Dumas, Y., Solomon, M.M., Soumis, F. (1995). Time constrained routing and scheduling. In: Ball, M.O., Magnanti, T.L., Monma, C.L., Nemhauser, G.L. (Eds.), *Network Routing*. *Handbooks in Operations Research and Management Science*, vol. 8. Elsevier, Amsterdam, pp. 35–139.

Dessouky, M., Hall, R., Zhang, L., Singh, A. (2003). Real-time control of buses for schedule coordination at a terminal. *Transportation Research – Part A* 37, 145–164.

Dial, R.B. (1967). Transit pathfinder algorithm. *Highway Research Record* 205, 67–85.

du Merle, O., Villeneuve, D., Desrosiers, J., Hansen, P. (1999). Stabilized column generation. *Discrete Mathematics* 194, 229–237.

Dubois, D., Bel, G., Llibre, M. (1979). A set of methods in transportation network synthesis and analysis. *Journal of the Operational Research Society* 30 (9), 797–808.

Eberlein, X.-J. (1995). Real-time control strategies in transit operations: Models and analysis. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA.

Eberlein, X.-J., Wilson, N.H.M., Barnhart, C., Bernstein, D. (1998). The real-time deadheading problem in transit operations control. *Transportation Research – Part B* 32 (2), 77–100.

Eberlein, X.-J., Wilson, N.H.M., Bernstein, D. (1999). Modeling real-time control strategies in public transit operations. In: Wilson, N.H.M. (Ed.), *Computer-Aided Transit Scheduling. Lecture Notes in Economics and Mathematical Systems*, vol. 471. Springer-Verlag, Heidelberg, pp. 325–346.

Eberlein, X.-J., Wilson, N.H.M., Bernstein, D. (2001). The holding problem with real-time information available. *Transportation Science* 35 (1), 1–18.

Elhallaoui, I., Villeneuve, D., Soumis, F., Desaulniers, G. (2005). Dynamic aggregation of set partitioning constraints in column generation. *Operations Research* 53 (4), 632–645.

Fan, W., Machemehl, R.B. (2004). Optimal transit route network design problem: Algorithms, implementations, and numerical results. Report SWUTC/04/167244-1, Center for Transportation Research, University of Texas at Austin.

Forbes, M.A., Holt, J.N., Watts, A.M. (1994). An exact algorithm for multiple depot bus scheduling. *European Journal of Operational Research* 72 (1), 115–124.

Fores, S., Proll, L., Wren, A. (2002). TRACS II: A hybrid IP/heuristic driver scheduling system for public transport. *Journal of the Operational Research Society* 53, 1093–1100.

Freling, R., Paixão, J.M.P. (1995). Vehicle scheduling with time constraint. In: Daduna, J.R., Branco, I., Paixão, J.M.P. (Eds.), *Computer-Aided Transit Scheduling. Lecture Notes in Economics and Mathematical Systems*, vol. 430. Springer-Verlag, Heidelberg, pp. 130–144.

Freling, R., Wagelmans, A.P.M., Paixão, J.M.P. (1999). An overview of models and techniques for integrating vehicle and crew scheduling. In: Wilson, N.H.M. (Ed.), *Computer-Aided Transit Scheduling. Lecture Notes in Economics and Mathematical Systems*, vol. 471. Springer-Verlag, Heidelberg, pp. 441–460.

Freling, R., Huisman, D., Wagelmans, A.P.M. (2001a). Applying an integrated approach to vehicle and crew scheduling in practice. In: Voss, S., Daduna, J.R. (Eds.), *Computer-Aided Scheduling of Public Transport. Lecture Notes in Economics and Mathematical Systems*, vol. 505. Springer-Verlag, Heidelberg, pp. 73–90.

Freling, R., Wagelmans, A.P.M., Paixão, J.M.P. (2001b). Models and algorithms for single-depot vehicle scheduling. *Transportation Science* 35 (2), 165–180.

Freling, R., Huisman, D., Wagelmans, A.P.M. (2003). Models and algorithms for integration of vehicle and crew scheduling. *Journal of Scheduling* 6, 63–85.

Fu, L., Liu, Q., Calamai, P. (2003). A real-time optimization model for dynamic scheduling of transit operations. In: *The 82nd Annual Meeting of the Transportation Research Board*, Washington, DC, January.

Furth, P.G. (1985). Alternating deadheading in bus route operations. *Transportation Science* 19 (1), 13–28.

Furth, P.G. (1986). Zonal route design for transit corridors. *Transportation Science* 20 (1), 1–12.

Furth, P.G. (1987). Short turning on transit routes. *Transportation Research Record* 1108, 42–52.

Furth, P.G., Wilson, N.H.M. (1982). Setting frequencies on bus routes: Theory and practice. *Transportation Research Record* 818, 1–7.

Gallo, G., Di Miele, F. (2001). Dispatching buses in parking depots. *Transportation Science* 35 (3), 322–330.

Gao, Z., Sun, H., Shan, L.L. (2004). A continuous equilibrium network design model and algorithm for transit systems. *Transportation Research – Part B* 38 (3), 235–250.

Gentile, G., Nguyen, S., Pallottino, S. (2005). Route choice on transit networks with online information at stops. *Transportation Science* 39 (3), 289–297.

Gintner, V., Kliewer, N., Suhl, L. (2004). A crew scheduling approach for public transit enhanced with aspects from vehicle scheduling. DSOR Working Paper WP0407, University of Paderborn, Germany. In: Hickman, M., Mirchandani, P., Voss, S. (Eds.) (2004). *Computer-Aided Transit Scheduling. Lecture Notes in Economics and Mathematical Systems*. Springer-Verlag, Heidelberg, in press.

Haase, K., Desaulniers, G., Desrosiers, J. (2001). Simultaneous vehicle and crew scheduling in urban mass transit systems. *Transportation Science* 35 (3), 286–303.

Hadjar, A., Marcotte, O., Soumis, F. (2006). A branch-and-cut algorithm for the multiple depot vehicle scheduling problem. *Operations Research* 54 (1), 130–149.

Haghani, A., Shafahi, Y. (2002). Bus maintenance systems and maintenance scheduling: Model formulations and solutions. *Transportation Research – Part A* 36, 453–482.

Hall, R.W. (1985). Vehicle scheduling at a transportation terminal with random delay en route. *Transportation Science* 19 (3), 308–320.

Hall, R.W. (1986). The fastest path through a network with random time-dependent travel times. *Transportation Science* 20 (3), 182–188.

Hall, R., Dessouky, M., Lu, Q. (2001). Optimal holding times at transfer stations. *Computers & Industrial Engineering* 40, 379–397.

Hamdouni, M., Desaulniers, G., Soumis, F., Marcotte, O., Van Putten, M. (2006). Parking and dispatching buses in depots using block patterns. *Transportation Science* 40 (3), 364–377.

Han, A.F., Wilson, N.H.M. (1982). The allocation of buses in heavily utilized networks with overlapping routes. *Transportation Research – Part B* 16 (3), 221–232.

Hasselström, D. (1981). Public transportation planning – a mathematical programming approach. Doctoral dissertation, University of Göteborg, Sweden.

Hickman, M. (2001). An analytic stochastic model for the transit vehicle holding problem. *Transportation Science* 35 (3), 215–237.

Hickman, M.D., Bernstein, D.H. (1997). Transit service and path choice models in stochastic and time-dependent networks. *Transportation Science* 31 (2), 129–146.

Hickman, M.D., Wilson, N.H.M. (1995). Passenger travel time and path choice implications of real-time transit information. *Transportation Research – Part C* 3 (4), 211–226.

Huisman, D., Freling, R., Wagelmans, A.P.M. (2005). Multiple-depot integrated vehicle and crew scheduling. *Transportation Science* 39 (4), 491–502.

Hurdle, V.F. (1973a). Minimum cost schedules for a public transportation route – I. Theory. *Transportation Science* 7 (2), 109–137.

Hurdle, V.F. (1973b). Minimum cost schedules for a public transportation route - II. Examples. *Transportation Science* 7 (2), 138–157.

Irnich, S., Desaulniers, G. (2005). Shortest path problems with resource constraints. In: Desaulniers, G., Desrosiers, J., Solomon, M.M. (Eds.), *Column Generation*. Springer-Verlag, New York, pp. 33–65.

Israeli, Y. (1992). Transit route and scheduling design at the network level. Doctoral dissertation, Technion Israel Institute of Technology, Haifa, Israel.

Israeli, Y., Ceder, A. (1989). Designing transit routes at the network level. In: *Proceedings of the First Vehicle Navigation and Information Systems Conference*. IEEE Vehicular Technology Society, pp. 310–316.

Israeli, Y., Ceder, A. (1995). Transit route design using scheduling and multiobjective programming techniques. In: Daduna, J.R., Branco, I., Piaxão, J. (Eds.), *Computer-Aided Transit Scheduling. Lecture Notes in Economics and Mathematical Systems*, vol. 430. Springer-Verlag, Heidelberg, pp. 56–75.

Israeli, Y., Ceder, A. (1996). Public transportation assignment with passenger strategies for overlapping route choice. In: Lesort, B. (Ed.), *Transportation and Traffic Theory: Proceedings of the 13th International Symposium on Transportation and Traffic Theory*. Pergamon, pp. 561–588.

Jansson, K., Ridderstolpe, B. (1992). A method for the route choice problem in public transport systems. *Transportation Science* 26 (3), 246–251.

Jordan, W.C., Turnquist, M.A. (1979). Zone scheduling of bus routes to improve service reliability. *Transportation Science* 13 (3), 242–268.

Klabjan, D., Johnson, E.L., Nemhauser, G.L., Gelman, E., Ramaswamy, S. (2002). Airline crew scheduling with time windows and plane-count constraints. *Transportation Science* 36, 337–348.

Klemt, W.D., Stemme, W. (1988). Schedule synchronization for public transit networks. In: Daduna, J.R., Wren, A. (Eds.), *Computer-Aided Transit Scheduling*. Springer-Verlag, New York, pp. 327–335.

Kliewer, N., Mellouli, T., Suhl, L. (2006). A time-space network based exact optimization model for multi-depot bus scheduling. *European Journal of Operational Research* 175 (3), 1616–1627.

Knoppers, P., Muller, T. (1995). Optimized transfer opportunities in public transport. *Transportation Science* 29 (1), 101–105.

Koutsopoulos, H.N., Odoni, A., Wilson, N.H.M. (1985). Determination of headways as a function of time varying characteristics on a transit network. In: Rousseau, J.M. (Ed.), *Computer Scheduling of Public Transport 2*. North-Holland, Amsterdam, pp. 391–414.

Kwan, A.S.K., Kwan, R.S.K., Wren, A. (1999). Driver scheduling using genetic algorithms with embedded combinatorial traits. In: Wilson, N.H.M. (Ed.), *Computer-Aided Transit Scheduling*. *Lecture Notes in Economics and Mathematical Systems*, vol. 471. Springer-Verlag, Heidelberg, pp. 81–102.

Kwan, A.S.K., Kwan, R.S.K., Wren, A. (2001). Evolutionary driver scheduling with relief chains. *Evolutionary Computing* 9, 445–460.

Lam, W.H.K., Gao, Z.Y., Chan, K.S., Yang, H. (1999). A stochastic user equilibrium assignment model for congested transit networks. *Transportation Research – Part B* 33, 351–368.

Lam, W.H.K., Zhou, J., Sheng, Z.-H. (2002). A capacity restraint transit assignment with elastic line frequency. *Transportation Research – Part B* 36, 919–938.

Lampkin, W., Saalmans, P.D. (1967). The design of routes, service frequencies, and schedules for a municipal bus undertaking: A case study. *Operational Research Quarterly* 18 (4), 375–397.

Last, A., Leak, S.E. (1976). Transept: A bus model. *Traffic Engineering and Control* 18 (1), 14–20.

le Clercq, F. (1972). A public transport assignment method. *Traffic Engineering and Control* 14 (2), 91–96.

Lee, K.T., Schonfeld, P. (1991). Optimal slack time for timed transfers at a transit terminal. *Journal of Advanced Transportation* 25 (3), 281–308.

Levinson, H. (1991). Supervision strategies for improved reliability of bus routes. In: *Synthesis of Transit Practice 15*, National Cooperative Transit Research and Development Program.

Li, Y., Rousseau, J.-M., Gendreau, M. (1991). Real-time scheduling on a transit bus route: A 0–1 stochastic programming model. Publication 772, Centre de Recherche sur les Transports, Université de Montréal.

Li, Y., Rousseau, J.-M., Wu, F. (1992). Real-time scheduling on a transit bus route. In: Desrochers, M., Rousseau, M. (Eds.), *Computer-Aided Transit Scheduling*. *Lecture Notes in Economics and Mathematical Systems*, vol. 386. Springer-Verlag, Heidelberg, pp. 213–235.

Li, J., Mirchandani, P., Borenstein, D. (2004). Parallel auction algorithm for bus rescheduling. In: Hickman, M., Mirchandani, P., Voss, S. (Eds.), *Computer-Aided Scheduling of Public Transport*. *Lecture Notes in Economics and Mathematical Systems*. Springer-Verlag, Heidelberg, in press.

Lo, H., Yip, C.W., Wan, K.H. (2003). Modeling transfers and nonlinear fare structure in multi-modal network. *Transportation Research – Part B* 37 (2), 149–170.

Lo, H., Yip, C.W., Wan, Q.K. (2004). Modeling competitive multi-modal transit services: A nested logit approach. *Transportation Research – Part C* 12, 251–272.

Löbel, A. (1998). Vehicle scheduling in public transit and Lagrangean pricing. *Operations Research* 44 (12), 1637–1649.

Lourenço, H.R., Paixão, J.P., Portugal, R. (2001). Multiobjective metaheuristics for the bus-driver scheduling problem. *Transportation Science* 35 (3), 331–343.

Magnanti, T.L., Wong, R.T. (1984). Network design and transportation planning: Models and algorithms. *Transportation Science* 18 (1), 1–55.

Marguier, P.H.J. (1985). Bus route performance evaluation under stochastic conditions. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA.

Marguier, P.H.J., Ceder, A. (1984). Passenger waiting strategies for overlapping bus routes. *Transportation Science* 18 (3), 207–230.

Mesquita, M., Paixão, J. (1999). Exact algorithms for the multi-depot vehicle scheduling problem based on multicommodity network flow type formulations. In: Wilson, N.H.M. (Ed.), *Computer-Aided Transit Scheduling*. *Lecture Notes in Economics and Mathematical Systems*, vol. 471. Springer-Verlag, Heidelberg, pp. 221–243.

Mingozzi, A., Bianco, L., Ricciardelli, S. (1995). An exact algorithm for combining vehicle trips. In: Daduna, J.R., Branco, I., Paixão, J.M.P. (Eds.), *Computer-Aided Transit Scheduling*. *Lecture Notes in Economics and Mathematical Systems*, vol. 430. Springer-Verlag, Heidelberg, pp. 145–172.

Newell, G.F. (1971). Dispatching policies for a transportation route. *Transportation Science* 5 (1), 91–105.

Nguyen, S., Pallottino, S. (1988). Equilibrium traffic assignment for large scale transit networks. *European Journal of Operational Research* 37, 176–186.
Nguyen, S., Pallottino, S., Malucelli, F. (2001). A modeling framework for passenger assignment on a transport network with timetables. *Transportation Science* 35 (3), 238–249.
Nielsen, O.A. (2000). A stochastic transit assignment model considering differences in passengers utility functions. *Transportation Research – Part B* 34, 377–402.
Nielsen, O.A. (2004). A large scale stochastic multi-class schedule-based transit model with random coefficients. In: Wilson, N.H.M., Nuzzolo, A. (Eds.), *Schedule-Based Dynamic Transit Modeling: Theory and Applications*. Kluwer Academic, Boston, pp. 53–77.
Nuzzulo, A. (2003). Transit path choice and assignment model approaches. In: Lam, W.H.K., Bell, M.G.H. (Eds.), *Advanced Modeling for Transit Operations and Service Planning*. Pergamon, Amsterdam, pp. 93–124.
Nuzzulo, A., Russo, F., Crisalli, U. (2001). A doubly dynamic schedule-based assignment model for transit networks. *Transportation Science* 35 (3), 268–285.
O'Dell, S., Wilson, N.H.M. (1999). Optimal real-time control strategies for rail transit operations during disruptions. In: Wilson, N.H.M. (Ed.), *Computer-Aided Transit Scheduling*. *Lecture Notes in Economics and Mathematical Systems*, vol. 471. Springer-Verlag, Heidelberg, pp. 299–323.
Odoni, A.R., Rousseau, J.-M., Wilson, N.H.M. (1994). Models in urban and air transportation. In: Pollock, S.M., Rothkopf, M.H., Barnett, A. (Eds.), *Operations Research and the Public Sector*. *Handbooks in Operations Research and Management Science*, vol. 6. North-Holland, Amsterdam, pp. 107–150.
Pattnaik, S.B., Mohan, S., Tom, V.M. (1998). Urban bus transit network design using genetic algorithm. *Journal of Transportation Engineering* 124 (4), 368–375.
Poon, M.H., Wong, S.C., Tong, C.O. (2004). A dynamic schedule-based model for congested transit networks. *Transportation Research – Part B* 38, 343–368.
Rapp, M.H., Gehner, C.D. (1976). Transfer optimization in an interactive graphic system for transit planning. *Transportation Research Record* 619, 27–33.
Ribeiro, C.C., Soumis, F. (1994). A column generation approach to the multiple depot vehicle scheduling problem. *Operations Research* 42 (1), 41–52.
Salzborn, F.J.M. (1972). Optimum bus scheduling. *Transportation Science* 6 (2), 137–148.
Salzborn, F.J.M. (1980). Scheduling bus systems with interchanges. *Transportation Science* 14 (3), 211–220.
Scheele, S. (1980). A supply model for public transit services. *Transportation Research – Part B* 14, 133–146.
Sheffi, Y., Sugiyama, M. (1982). Optimal bus scheduling on a single route. *Transportation Research Record* 895, 46–52.
Shen, S., Wilson, N.H.M. (2001). An optimal integrated real-time disruption control model for rail transit systems. In: Voss, S., Daduna, J. (Eds.), *Computer-Aided Scheduling of Public Transport*. *Lecture Notes in Economics and Mathematical Systems*, vol. 503. Springer-Verlag, Heidelberg, pp. 335–363.
Shen, Y., Kwan, R.S.K. (2001). Tabu search for driver scheduling. In: Voss, S., Daduna, J.R. (Eds.), *Computer-Aided Scheduling of Public Transport*. *Lecture Notes in Economics and Mathematical Systems*, vol. 505. Springer-Verlag, Heidelberg, pp. 121–135.
Silman, L.A., Barzily, Z., Passy, U. (1974). Planning the route system for urban buses. *Computers & Operations Research* 1, 210–211.
Site, P.D., Filippi, F. (1998). Service optimization for bus corridors with short-turn strategies and variable vehicle size. *Transportation Research – Part A* 32 (1), 19–38.
Smith, B.M., Wren, A. (1988). A bus crew scheduling system using a set covering formulation. *Transportation Research – Part A* 22, 97–108.
Spiess, H. (1983). On optimal route choice strategies in transit networks. Publication 285, Centre de Recherche sur les Transports, Université de Montréal.
Spiess, H., Florian, M. (1989). Optimal strategies: A new assignment model for transit networks. *Transportation Research – Part B* 23 (2), 83–102.
Stern, H.I., Ceder, A. (1983). An improved lower bound to the minimum fleet size problem. *Transportation Science* 17 (4), 471–477.

Sun, A., Hickman, M. (2004). The holding problem at multiple holding stations. In: Hickman, M., Mirchandani, P., Voss, S. (Eds.), *Computer-Aided Scheduling of Public Transport*. *Lecture Notes in Economics and Mathematical Systems*. Springer-Verlag, Heidelberg.

Sun, A., Hickman, M. (2005). The real-time stop-skipping problem. *Journal of Intelligent Transportation Systems* 9 (2), 91–109.

Ting, C.-J., Schonfeld, P. (2005). Schedule coordination in a multiple hub transit network. *Journal of Urban Planning and Development* 131 (2), 112–124.

Tom, V.M., Mohan, S. (2003). Transit route network design using frequency coded genetic algorithm. *Journal of Transportation Engineering* 129 (2), 186–195.

Tong, C.O., Richardson, A.J. (1984). A computer model for finding the time-dependent minimum path in a transit system with fixed schedules. *Journal of Advanced Transportation* 18 (2), 145–161.

Tong, C.O., Wong, S.C. (1999). A stochastic transit assignment model using a dynamic schedule-based network. *Transportation Research – Part B* 33, 107–121.

Turnquist, M.A. (1978). A model for investigating the effects of service frequency and reliability on bus passenger waiting times. *Transportation Research Record* 663, 70–73.

Turnquist, M.A. (1981). Strategies for improving reliability of bus service. *Transportation Research Record* 818, 7–13.

Turnquist, M.A., Blume, S.W. (1980). Evaluating potential effectiveness of headway control strategies for transit systems. *Transportation Research Record* 746, 25–29.

van Nes, R., Hamerslag, R., Immers, B.H. (1988). Design of public transport networks. *Transportation Research Record* 1202, 74–83.

Verma, A., Dinghra, S.L. (2005). Feeder bus routes generation within integrated mass transit planning framework. *Journal of Transportation Engineering* 131 (11), 822–834.

Wahba, M., Shalaby, A. (2005). A multi-agent learning-based approach to the transit assignment problem: A prototype. *Transportation Research Record* 1926, 96–105. Paper taken from CD-ROM of Conference Proceedings.

Welding, P.I. (1957). The instability of a close-interval service. *Operational Research Quarterly* 8 (3), 133–148.

Wilson, N.H.M., Nuzzolo, A. (Eds.) (2004). *Schedule-Based Dynamic Transit Modeling: Theory and Applications*. Kluwer Academic, Boston.

Winter, T., Zimmermann, U.T. (2000). Real-time dispatch of trams in storage yards. *Annals of Operations Research* 96, 287–315.

Wirasinghe, S.C., Liu, G. (1995). Optimal schedule design for a transit route with one intermediate time point. *Transportation Planning and Technology* 19, 121–145.

Wren, A., Rousseau, J.-M. (1995). Bus driver scheduling – an overview. In: Daduna, J.R., Branco, I., Paixão, J.M.P. (Eds.), *Computer-Aided Transit Scheduling*. *Lecture Notes in Economics and Mathematical Systems*, vol. 430. Springer-Verlag, Heidelberg, pp. 173–187.

Wu, J.H., Florian, M. (1993). A simplicial decomposition method for the transit equilibrium assignment problem. *Annals of Operations Research* 44, 245–260.

Wu, J.H., Florian, M., Marcotte, P. (1994). Transit equilibrium assignment: A model and solution algorithms. *Transportation Science* 28 (3), 193–203.

Zhao, J., Dessouky, M., Bukkapatnam, S. (2001). Distributed holding control of bus transit operations. In: *Proceedings of the IEEE Intelligent Transportation Systems Council (ITSC) Conference*, Oakland, CA, August.