A modeling system for simulation of dial-a-ride services

Carl Henrik Häll, Magdalena Högberg and Jan T. Lundgren

Linköping University Post Print

N.B.: When citing this work, cite the original article.

The original publication is available at www.springerlink.com:

Carl Henrik Häll, Magdalena Högberg and Jan T. Lundgren, A modeling system for simulation of dial-a-ride services, 2012, Public transport, (4), 1, 17-37.

<http://dx.doi.org/10.1007/s12469-012-0052-6>

Copyright: Springer Verlag (Germany)

<http://www.springerlink.com/?MUD=MP>

Postprint available at: Linköping University Electronic Press

<http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-68057>

A Modeling System for Simulation of Dial-a-Ride Services

Carl H. Häll^{*†} Magdalena Högborg[‡] Jan T. Lundgren^{*}

Abstract

We present a modeling system for simulation of dial-a-ride services. It can be used as a tool for understanding and study how different designs, and different ways to operate a dial-a-ride service, affect the performance and efficiency of the service. The system simulates the operation of a dynamic dial-a-ride service that operates with multiple fleets of vehicles with different capacities, schedules and depots. It can be used to investigate how the setting of service and cost parameters and the design of the service affect the total cost for the operator and level of service for the customer. We describe the different modules in the system and the possible uses of the system. A short simulation study is performed to exemplify how it can be used. In this study the effects of including costs for customer discomfort are evaluated.

1 Introduction

Local public transport systems are usually based on different combinations of fixed route services, such as trains, subway, trams and buses. These services provide regular and often frequent services between fixed locations in the network. The vehicles travel along predefined routes at scheduled times. However, the fixed route services do not satisfy the needs of all the customers wanting to use the public transport system. For many people, especially the elderly and disabled persons, it can be difficult to use the services (due to the person's reduced mobility). It is therefore necessary to introduce some form of transportation service that is specially designed for these customers. This service is often called paratransit. It is normally demand responsive and is used as a complement to the regular, fixed route service, to build a local public transportation system that satisfies the mobility needs of the whole population.

One common form of paratransit is to operate it as a dial-a-ride service, where customers call in requests to a call center and the transportation is carried out from door to door. A service of this kind can be designed in many different ways, depending on the

^{*}Linköping University, Dept of Science and Technology (ITN), SE-60174 Norrköping, SWEDEN

[†]E-MAIL: carha@itn.liu.se

[‡]Astando AB, 11221 Stockholm, Sweden

level of service the operator wishes to provide the customers with. It is a question of balancing the service level against the cost of operating the service. Dial-a-ride service is a quite expensive form of transportation service, and in order to develop a cost-efficient service, it is important to know how different system design affects both the level of service for the customers and the cost for the operator. It is also important that the procedures for the operational planning of the service, in terms of routing and scheduling, work efficiently, and that the operator knows how changes in the design and use of procedures affect the solution; both in terms of customer service and operating costs.

When operating a dial-a-ride service, the planning problem is to design routes and schedules for a fleet of vehicles, given a set of requests for transportation and also possibly taking future requests into account. The set of travel requests can include individual passengers or groups of passengers, with specified origins and destinations. The schedules have to satisfy passengers' requirements in terms of user inconvenience, e.g. defined by waiting times, maximal travel times or deviations from desired departure and arrival times. The overall goal is to minimize a combination of user inconvenience and operating costs. This problem is known as the dial-a-ride problem (DARP) and is a special case of the vehicle routing problem with pick-ups and deliveries. What makes DARP special is the necessity of balancing user inconvenience (service level) against the operating costs for transporting passengers.

In this paper, we present a modeling system for simulating dynamic dial-a-ride services. The system is called DARS (Dial-a-Ride Simulator). DARS uses a micro simulation approach. Given a set of requests with specified call-in times for a planning period, the system simulates the operative planning and generates vehicle itineraries based on customer requirements and operational costs. The aim of this paper is to describe the system, and the need for, and use of, it. To exemplify the use of DARS, we perform a simulation study evaluating the impact of customer related costs.

DARS can be used as a tool to understand and study how different system designs and different ways to operate a dial-a-ride service affect the performance and efficiency of the service. Also, it can be used to investigate how the settings of service and cost parameters affect the total cost for the operator and level of service for the customers. Further, DARS can also be used to study changes in planning methodology and investigate the effects of using different heuristics to compute travel schedules.

The outline of the paper is as follows. In Section 2 we present a literature review of modeling and simulation of dial-a-ride services, and in Section 3 we present an overview of the modeling system DARS. Section 4, further describes the system and specifies what input is necessary, in order to run simulations in DARS. Section 5 describes how the construction of vehicle schedules is made in DARS. In Section 6, the operational use of DARS is described, and in Section 7 simulations are performed to evaluate the impact of customer related costs. Finally, Section 8 presents the conclusions of this study and future research projects to be studied within this area.

2 Modeling and simulation of DARP

In practice there are many different types of dial-a-ride services, which mean there are also many different mathematical models to describe the problem of planning such services. In a practical application the choice of model and solution procedure is therefore of great importance for an efficient generation of schedules. In this section we will describe the most important models found in the literature for dial-a-ride services.

Dial-a-ride services are operated according to one of two modes, static (off-line) or dynamic (on-line). The static mode is defined by the fact that all requests are known in advance, which also allows vehicle schedules to be planned in advance. Static versions of the dial-a-ride problem (DARP) are described for instance in Psaraftis (1983), Jaw et al. (1986), Ioachim et al. (1995), Borndörfer et al. (1997), Toth and Vigo (1997), Fu (2002a), Cordeau and Laporte (2003b) and Cordeau (2006). In the dynamic mode, described for example in Madsen et al. (1995), Teodorovic and Radivojevic (2000), Colorni and Righini (2001) and Attanasio et al. (2004), the number of requests gradually increases as the customers call in requests, and the planning starts before all requests are known. However, the modeling of the dynamic case is in many ways similar to the modeling of the static case. Some requests are usually known before the scheduling starts, and a starting solution to the dynamic case can be obtained by solving the initial static case, given the requests at hand.

Most papers describing solution methods for DARP are based on heuristic methods, since the problem type is known to be NP-hard and optimal solutions are therefore very difficult to find. However, Cordeau (2006) and Ropke et al. (2007) describe exact formulations of the DARP as well as giving branch-and-cut algorithms for finding optimal solutions. Among the papers presenting metaheuristics for solving the DARP, tabu search is most commonly used, see for example Cordeau and Laporte (2003b); Attanasio et al. (2004) and Cordeau and Laporte (2003b). Other techniques such as simulated annealing and genetic algorithms have also been used, see for example Baugh et al. (1998) and Uchimura et al. (2002). More detailed descriptions of the DARP can be found in Cordeau and Laporte (2003a) and in Cordeau et al. (2004). Good surveys of the topic are presented in Cordeau and Laporte (2007) and Parragh et al. (2008), and a survey specifically on dynamic problems is presented in Berbeglia et al. (2010).

Some more specific types of dial-a-ride services, and therefore more specific cases of the DARP, have also been studied in the literature. Ho and Haugland (2011) describes a dial-a-ride service where each request will be made with a certain probability, and in Häll et al. (2009), a dial-a-ride service is combined with a fixed-route bus service.

Many real life systems, especially those found in Sweden, are dynamic, and a customer may request service with very short call ahead time. This dynamic feature puts hard constraints on the computation time available for creating a feasible schedule when a new customer makes a request. Since the customer must be given a pick-up time while waiting on the phone, the time available for constructing a feasible route is usually only a couple of seconds. Therefore, insertion heuristics are most often used in the commercial systems used for planning paratransit.

In a practical application of the DARP, the planning objective can be seen from two different perspectives. From the operator's perspective, the goal is most often to minimize the total number of vehicles needed, or to minimize the total travel time of the vehicles. From the customer's perspective, the goal is often to minimize ride time or service time deviations, (Fu and Teply, 1999). For both the static and dynamic cases, most models assume a homogeneous vehicle fleet located at a single depot. This is however not always the case in practice. There might be several depots for the vehicles and also different vehicle types, for example vehicles equipped to handle wheelchairs.

When designing a dial-a-ride service, whether dynamic or static, it is important to know how different service characteristics and routing policies affect the service. Simulation is a tool that for many years has been used to study the effects of many-to-many dial-a-ride systems (see e.g. Heathington et al., 1968; Wilson et al., 1969; Gerrard, 1974). How the usability of the systems changes due to the number of passengers have been studied in some papers. Bailey and Clark (1987) uses simulation to study the interaction between demand, service rate and policy alternatives for a taxi service. The usability of a dial-a-ride service in comparison to fixed route systems is studied through simulation in the work of Noda et al. (2003).

The use of simulations to study the effects of different changes in a dial-a-ride service is described in Quadrifoglio et al. (2008) who also studies how time window settings and zoning vs. no-zoning strategies affect the total trip time, deadhead miles and fleet size. The impact of fleet size is studied in Diana et al. (2006). A continuous approximation model is used instead of simulation to determine the number of vehicles needed to give a predefined quality of the service.

Deflorio et al. (2002) propose a simulation system that is able to evaluate quality and efficiency parameters of a dial-a-ride service. The system can simulate a number of uncertainties caused by both passengers and drivers. Another simulation system is described in Fu (2002b). The purpose of this system is to evaluate what effects new technologies, such as automatic vehicle location, can have on a dial-a-ride service.

The specific conditions in dial-a-ride services when we have very short call ahead times makes it impossible to use the previously described models and systems. The high level of service given to the customers puts special requirements on how time windows are constructed. In addition, DARS also handles a wider framework and can be used to study any general type of dial-a-ride services. For example, the instance of the dynamic DARP we consider in this paper can have multiple user groups, multiple vehicles and (possibly) multiple operators using different types of vehicles and different depots.

3 General description of DARS

This section describes the components of the modeling system DARS, which is a standalone system developed in C#.NET. It also describes how the components interact with each other. The main components are the *Input manager*, the *Simulation control*, the *Request manager* and the *Output generator*. A schematic description of the DARS archi-

tecture is given in Figure 1.

Figure 1: A schematic description of the DARS architecture

The Input manager handles all information that defines the scenario which we want to study and evaluate, and this information has to be clearly specified before a simulation can be run in DARS. The information includes vehicles available, travel demand structure, transportation network, design of the service and cost structure. These categories of input are further described in Section 4.

The Simulation control is the component that simulates the events occurring in the real world. At the start of a simulation, requests are generated with a given call-in time and inserted into a time ordered event queue. Normally, historical data is used to generate requests. The main purpose of the Simulation control is to keep track of time, and to know the specific times when requests (events) are to occur. When the simulation time equals the time a request is to be called in, the Simulation control notifies the Request manager that the request has to be scheduled. Today, DARS does not consider any stochastic events such as accidents, no-shows and cancellations. If these kinds of events are to be included in DARS, they should be included in the Simulation control.

The aim of the Request manager is to simulate the operational work in a real planning situation, in other words, to create the travel schedules for the vehicles given the incoming requests from the Simulation control. A travel schedule specifies the itinerary for a vehicle, including the planned pick-up times and delivery times. For each point in time during the simulation, the current planning situation is given by the available schedules, which are based on all requests that have arrived so far.

The Request manager uses two different procedures to create travel schedules. Each time a new request arrives, an insertion heuristic is used to insert the new request into the current schedule. If it is impossible to schedule the request, it is rejected. In order to simulate a real planning situation, this insertion procedure is not allowed to take more than a few seconds. When there are no new requests to insert, the computational time is used by a reoptimization heuristic in order to improve the current schedules. The two heuristics

are described in Section 5.1 – 5.2. The program will be executed until all requests have been processed. While executing the two types of heuristics, the schedule control in the Request manager makes sure that there is always an up-to-date travel schedule for all vehicles.

When the simulation has ended, the component Output generator provides three different types of statistics for analyzing the results. These are General statistics, Distance related statistics and Time related statistics. In the current version of DARS, the following information is generated

- General statistics
 - the number of scheduled requests
 - the number of rejected requests
 - the number of vehicles used
 - the value of the objective function
 - the maximum and average number of requests onboard a vehicle
 - the maximum and average number of passengers onboard a vehicle
- Distance related statistics
 - the total distance traveled by all vehicles
 - the total distance traveled without passengers
 - the total direct travel distance, which is the sum of the direct travel distances of all requests
- Time related statistics
 - the total time traveled by all vehicles and all passengers
 - the total idle time and service time for all vehicles
 - the total (and worst) excess ride time
 - the total (and worst) deviation from the planned pick-up times

There are of course a number of other statistics that can be calculated as a result of the scheduling made. Any user of DARS can specify their own statistics based on the given output. In addition to the output statistics, DARS also provides a logging function that logs the final schedule, and two graphical tools to visualize the results. These are explained in Section 6.

4 Input manager

The Input manager handles all the information about the dial-a-ride service and about the specific scenarios to be simulated. In this section, we describe the input needed to make a simulation and what type of scenarios can be simulated using the current version of DARS. This current version is based on a real-world dial-a-ride service operating in Sweden. However, all the components should be implemented as accurately as possible for the specific scenario one is studying.

4.1 Vehicle specifications

DARS offers the simulation of two different types of vehicles, ordinary taxi vehicles and specially equipped vehicles in which it is possible to transport passengers in wheelchairs.

For each vehicle, the start depot and the end depot are specified as well as the time of the day the vehicle can be used. Three different capacities are also specified for each vehicle; the number of ordinary seats, the number of seats with extra legroom and the number of wheelchair places (zero for ordinary taxi vehicles). A vehicle may only be reassigned when arriving to, or waiting at, a given stop, i.e., it is not possible to reassign a vehicle while the vehicle is moving. This assumption is not as unreasonable as it may sound, since it is natural to assume that vehicles on route should be allowed to finish the current leg of the route before receiving any new assignments, for instance due to convenience of the drivers. Vehicle break-downs and other stochastic vehicle-related events are not considered.

4.2 Travel demand

The planning takes place in a dynamic setting where all of the demand is not known at the start of the day and at the start of the schedule planning. Customers may call in and demand transportation at any time. Each customer specifies a pick-up time, a pick-up location and a drop-off location (all such locations are henceforth referred to as nodes) and optionally, any special requirements on legroom or wheelchair space.

When simulating the planning of schedules in DARS, we generate all requests in advance. The requests are specified based on historical data and for each request it is specified at what time the request shall be placed, i.e. the call-in time.

One simplification made to the requests concerns the baggage capacity of the vehicles. In a real-world application, some customers may use walkers or wheelchairs that are foldable, or customers may have other types of baggage. This may limit which vehicle can be used to serve the customer. However, in the current version of DARS, customers using foldable or aids that in any other way are storable, are treated as ordinary customers. No consideration is taken of any form of baggage.

When transporting elderly and/or handicapped individuals, the time required to enter or exit the vehicle might be substantial due to a passenger's reduced mobility, and therefore these service times have to be taken into account. If not, the schedule created might

be unrealistic to be implemented in a real world setting. Different service times can be used for different categories of customers. In the current version of DARS, two such categories are used; for customers with a wheelchair and customers without a wheelchair. If more detailed information about customer requirements is available, we can include additional categories.

4.3 Network

It is assumed that travel times between all nodes (including depot locations) are given by a matrix. In the current version, this matrix is based on travel times between different zones. The geographical area is divided into such zones, and the travel times are calculated in the road network based on given distances and average speeds on the links. In this way, the road network structure is not needed when the simulation is running. For each journey-leg, only the zones to which the start-node and end-node belongs are checked, and the travel time follows. Travelling between two locations in the same zone is handled in the same way, i.e. such a journey-leg is assumed to take a certain amount of time (individual for each zone).

The model can be extended to account for different travel times at different times of the day, e.g. to handle dynamic changes due to different traffic situations (such as congestion in the rush hour). All stochastic events, such as traffic jams caused by accidents, are disregarded in the model.

4.4 Service levels

Since we wish to use DARS to evaluate different designs of a dial-a-ride service, for each design it must be possible to define what level of service is guaranteed to the customers. When designing a dial-a-ride service, there are two main things that affect the level of service which have to be taken into account. The first is how specific (exact) in time the passengers can demand to be picked up, and the second is how long a requested journey is allowed to take. The matters of time window width and maximum ride time are both considered in DARS.

The pick-up time window ensures that the transport service of a request is started within an acceptable time from when the customer requested the service. The maximal ride time ensures that no customer gets a too long excess ride time. Before running a simulation in DARS, the parameters defining these constraints must be given.

4.4.1 Maximal ride time constraints

A maximal ride time constraint restricts the time a customer can be transported by a vehicle. In the current version, the maximal ride time of a journey is proportional to the direct ride time. For short journeys, a constant time is also added. Let

DRT_i	=	the direct ride time between pick-up node and drop-off node of request i
A	=	the added ride time for shorter journeys
θ	=	the proportionality parameter used for longer journeys
M	=	the threshold ride time

Then the maximal allowed ride time for request i can be computed as

$$MRT_i = \begin{cases} \theta DRT_i & \text{if } DRT_i \geq M \\ \theta DRT_i + A & \text{if } DRT_i < M \end{cases}$$

The ride time of a customer is defined as the time that has elapsed from the departure from the pick-up node until arrival at the delivery node. Consequently the service times of the user, both at the pick-up node and at the delivery node, are not included in the ride time.

4.4.2 Time window constraints

In a real life situation, a customer i that requests to be picked up at time RPT_i is given a planned pick-up time PPT_i while still on the phone. The planned pick-up time must be within a certain time window around the requested pick-up time. After the customer has been given this planned pick-up time, the planner is allowed to change the actual pick-up time so that it occurs within a small time window around the planned pick-up time. In the following we describe the calculation of these two time windows. The time windows are used when calculating feasible insertions of new requests in the current schedule.

As noted earlier, service times at pick-up and delivery nodes should be considered when calculating a schedule. Let STP_i denote the service time required at the pick-up node of request i , and let STD_i be the service time at the delivery node of request i . It is assumed that the time the customer gives when making a reservation, is the time the customer wishes the service at the pick-up node to begin.

Let EPT_i be the earliest possible time to start the service at the pick-up node of request i , and let LPT_i be the latest possible time to start the service of request i . The pick-up node must then be visited by a vehicle within the time window $[EPT_i, LPT_i]$, which is of length TW_1 . Assuming that the time window is symmetric around RPT_i , we have that $EPT_i = RPT_i - 0.5TW_1$ and $LPT_i = RPT_i + 0.5TW_1$. A special rule should be applied if the customer requests service within $0.5TW_1$ time units from the time that the request is called in. In such a case, the start of the time window is the same as the call-in time. If not using this rule, a customer making a late call may be serviced closer to its requested time compared to a customer making a request far in advance.

To simplify the calculation of schedule feasibility, it is customary to calculate time windows for the delivery nodes too. The earliest time a customer can arrive at the delivery node is given by the earliest possible pick-up time plus the direct travel time between the pick-up and delivery nodes, plus service times. The latest possible end of the service at

the delivery node is given as the latest possible start of service at the pick-up node, plus the maximum ride time for this customer, plus the service times at the nodes. Therefore, the time window $[EDT_i, LDT_i]$ at the delivery node is calculated as:

$$EDT_i = EPT_i + DRT_i + STP_i + STD_i$$

and

$$LDT_i = LPT_i + MRT_i + STP_i + STD_i$$

In Figure 2, the principles used for constructing time windows around the pick-up node and the delivery node are shown.

Figure 2: Illustration of how time windows for the pick-up node and the delivery node are constructed when a pick-up time has been requested by a customer

The relationship between the maximal ride time MRT_i and the direct ride time DRT_i in the figure is dependent on how the values of the constants (θ and M) that form the maximal ride time constraints (as described above) are chosen. The construction of time windows is similar to that presented in Jaw et al. (1986), with the difference that we explicitly include the service times (STP_i and STD_i), which makes it possible to study the effects of different service times.

DARS also allows the use of a second type of time window with width TW_2 . This type is used for dial-a-ride services where customers are guaranteed to be picked up within a certain time from the planned pick-up time. The longer the time window is, the easier it is for the planner to create good vehicle schedules, but at the expense of decreased customer service.

Given the requested time from the customer, and given the time window length TW_1 , the planning procedure tries to find the best feasible insertion of the request, and the time the vehicle is planned to visit the customers pick-up node is given as the planned pick-up time PPT_i . This time may then be delayed TW_2 time units, so the final time window within which the vehicle must visit the pick-up node becomes $[PPT_i, PPT_i + TW_2]$. An associated time window at the delivery node is calculated as $[EDT_i, LDT_i]$, where

$$EDT_i = PPT_i + DRT_i + STP_i + STD_i$$

and

$$LDT_i = PPT_i + TW_2 + MRT_i + STP_i + STD_i$$

The calculations of time windows for pick-up and delivery nodes are illustrated in Figure 3.

Figure 3: Illustration of how time windows for the pick-up node and the delivery node are constructed when a planned pick-up time has been given to the customer

There are some practical aspects about the specification of the value of the time windows widths that need to be considered. Wider time windows, TW_1 and TW_2 , give more and better planning possibilities but give a lower level of service for the customers. A larger value of TW_1 means that a customer is not guaranteed to be served as close to the requested time as he or she would otherwise be, and a large value of TW_2 creates a larger insecurity for the customers as to when they really will be picked up. In summary, large time windows are efficient for the operator but undesired by the customers. It is therefore very important to find good values for the time window widths in order to balance the conflicting objectives of the operator and the customer.

4.5 Cost structure

The objective function in a dynamic DARP is usually defined as to minimize the operational costs of the service, to minimize some measure of disutility for the customers, or as a combination of these two different perspectives. In the current version of DARS, a weighted combination of operational costs and costs perceived by the customers is used. The objective is to minimize a total cost consisting of the sum of three types of costs defined as

$$\text{Vehicle cost} = \sum_{k=1}^K (\alpha_k DT_k + \beta_k IT_k + \gamma_k)$$

$$\text{Waiting time cost} = \sum_{i=1}^N (v_1^i WT_i + v_2^i WT_i^2 + v_4^i WT_i^4)$$

$$\text{Excess ride time cost} = \sum_{i=1}^N (\vartheta_1^i ET_i + \vartheta_2^i ET_i^2 + \vartheta_4^i ET_i^4)$$

where

- K = the set of scheduled vehicles
- N = the set of customer requests
- DT_k = the total drive time including waiting time with customers for vehicle k
- IT_k = the total idle time defined as waiting time without customers for vehicle k
- WT_i = the waiting time of customer i
- ET_i = the excess ride time of customer i , defined as the difference between the ride time provided by the system and the direct ride time

and the cost parameters are

- α_k = the drive time cost per time unit associated with the use of vehicle k
- β_k = the idle time cost per time unit associated with the idle time of vehicle k
- γ_k = the fixed cost associated with the use of vehicle k
- v_1^i = the cost associated with the linear part of the waiting time for customer i
- v_2^i = the cost associated with the quadratic part of the waiting time for customer i
- v_4^i = the cost associated with the quadruple part of the waiting time for customer i
- ϑ_1^i = the cost associated with the linear part of the excess ride time of customer i
- ϑ_2^i = the cost associated with the quadratic part of the excess ride time of customer i
- ϑ_4^i = the cost associated with the quadruple part of the excess ride time of customer i

In this way, the vehicle cost is a combination of drive time cost, idle time cost and a fixed cost for taking a vehicle into service. The expressions for the waiting time cost and the excess ride time cost include three components related to how the waiting time and excess ride time are assumed to influence the cost. These expressions will give the user some flexibility of how to define the impact of these service indicators on the total cost.

5 Construction of travel schedules

The travel schedules are constructed in the Request manager component. DARS uses a time based simulation approach and a new request arrives to the Request manager when the simulation time equals the call-in time of this request in the event queue. The new request is inserted into the current schedule using an insertion heuristic at the place which is the most favorable, based on the defined cost in the objective function. The procedure is described in more detail in Section 5.1.

The time available until the simulation time is equal to the call-in time of the next request in the queue is used by a reoptimization algorithm to improve the current schedule. The reoptimization algorithm is run continuously until a new request is generated. If a new request appears before the reoptimization has found a better schedule, the reoptimization is interrupted and the old schedule is used when the new request is inserted. The reoptimization algorithm is described in Section 5.2.

This full procedure is illustrated in Figure 4, and describes the link between the “simulation control” and the “request manager” (see Figure 1). Since the insertion and reoptimization procedures are run as separate threads, it is easy to change the specific algorithms used in these procedures. Thus, the effects of using different solution methods may be evaluated.

5.1 Insertion of requests

The insertion of a new request i in the current schedule is performed according to the following three steps:

1. For each vehicle, generate all feasible positions to insert request i and calculate the change in objective function value.
2. Insert the request into the position with the lowest incremental cost.
3. If no feasible insertion exists, add the request to a list of requests not served.

This is a quite general procedure which is also used in two of the most cited descriptions of insertion heuristics for the DARP, the ADARTW heuristic developed by Jaw et al. (1986) and the REBUS heuristic developed by Madsen et al. (1995).

As described in Madsen et al. (1995), a time consuming part of insertion heuristics such as the one described above, is to check if an insertion is feasible. For a solution to

Figure 4: Description of the different steps in the creation of a new schedule

be feasible, the customers of the new request must of course be picked up and delivered by the same vehicle and the precedence constraints must be fulfilled. Furthermore, all the other constraints must also be checked. These are the maximum ride time constraints, time window constraints, waiting time constraints and vehicle capacity constraints. The constraints must be checked for all scheduled requests that are affected by the newly inserted request.

To examine whether a given route is feasible with respect to maximal ride times and time windows, the algorithm presented in Hunsaker and Savelsbergh (2002), originally developed for the static dial-a-ride, is used. The algorithm calculates the earliest feasible departure and arrival times, ensuring that the constraints are fulfilled. Some minor modifications have been made to the original algorithm, adapting it to the dynamic case and to the use of a heterogeneous fleet of vehicles. Since these modifications are only minor, the details of the feasibility tests are not presented here. Any reader requiring these details is referred to the paper by Hunsaker and Savelsbergh (2002).

There is however one known problem with this heuristic. In some specific cases it fails to find a feasible solution although such a solution exists. This problem has been described in Högberg (2008), Tang et al. (2010) and Haugland and Ho (2010). If the algorithm finds a feasible solution, then indeed the solution is feasible, but for some cases it can state that no feasible solution exists in situations when this is not true. In spite of this problem we have chosen to use this algorithm, since it is a fast algorithm. One should also remember, that when working with multiple vehicles, most of the times when the problem with the algorithm occurs, it only means that the best place to insert the new request have not been found, but quite often some feasible insertion can still be found in the itinerary of another vehicle.

A problem with any insertion heuristic is that the quality of solutions produced tends to be rather poor. The reason is that a new request is inserted in a local optimal position, and the heuristic does not attempt to move the solution towards a global optimum. This was recognized by Psaraftis (1980) for these kinds of heuristics. To create better solutions for this kind of problem, a reoptimization of some kind has to be used.

5.2 Reoptimization

The aim of a reoptimization procedure is to search for improvements to the current schedule. In our problem, the computational time available for the procedure depends on the number and frequency of new requests, and how much time remains after the insertion heuristic has been applied. There are many options for how the reoptimization algorithm can be designed. One of the aims with the development of DARS is to provide an environment in which it is easy to change the reoptimization algorithm. This gives the possibility to easily implement a reoptimization procedure that is suitable for the specific transportation service one is studying, or to implement different algorithms to compare them and study what effects they have on the solutions obtained.

Therefore, in the current version, we have not tried to find the very best reoptimization algorithm to implement, but to create a module in which various algorithms can be tested. This module is implemented in such a way that it suits any ruin-and-recreate heuristic. In this way, any such heuristic can be tested, no matter if it is a very simple one just removing one request at a time and reinserting it, or if all requests are considered in some type of metaheuristic such as e.g. in Beaudry et al. (2010).

All planned requests, for which the pick-up node is not yet visited, can be extracted from the current schedule. Before the reoptimization heuristic starts, a copy of the schedule is made. The reoptimization (extractions and reinsertions of requests) is then made in the copied version. In this way, there is always a feasible solution to return to if the reoptimization does not find a better solution before the time of the next event (a new request to insert or a planned request that begins to be served) occurs.

Requests that are removed from the schedule may not be inserted earlier than the PPT_i given to the customer, and may not be delayed more than TW_2 minutes, as described in Section 4.4.

6 The graphical user interface

In order to be able to use a simulation platform such as DARS in an efficient and practical way, it should be easy to define scenarios and to study the effects of different parameter settings, different demand structures and different planning heuristics. We therefore need a graphical user interface (GUI) from where we can provide and change the input parameters and read the output results. Some changes in the design of the service and regarding the heuristics used for generating schedules have still to be made from inside the code. This section describes the GUI of DARS and gives an example on how the effects of changes to a certain parameter can be studied.

The *Main form* of the GUI is shown in Figure 5. Here, the input and output directories, as well as the simulation start time, are set. The time factor is also set in the Main form. DARS simulates the scheduling of requests arriving in real time, or real time scaled by some factor χ . This time factor is introduced to reduce the time needed to complete a simulation. Using a high value of χ will speed up the simulation time but also influence the results negatively, since an increased simulation speed decreases the available computation time for finding good schedules. From the Main form, it is also possible to choose to run either a single simulation or batch simulations. A batch simulation uses predefined projects, with saved cost settings, constraint settings and simulation settings. One project is run at a time and output is created for each project file.

Figure 5: Description of the Main form in the graphical user interface

In the *Cost settings form*, all cost parameters are set. This form is shown in Figure

6. The existing fields in the form correspond to the costs used to describe the objective function (as described in Section 4.5). In the *Simulation settings form* all constraint parameters as well as general simulation parameters are set, as shown in Figure 7. It is possible to save and load files which include all cost parameters and constraint parameters. This makes the preparation of new simulations and new project files easier.

The screenshot shows a window titled "CostSettingsDialog" with a blue title bar and standard Windows window controls. The window is divided into two main columns of input fields. The left column is titled "Request Related Costs" and contains three sections: "Excess ride time costs" with fields for "SEK/Hour (ERT_LIN)" (0), "SEK/Hour^2 (ERT_QUAD)" (200), and "SEK/Hour^4 (ERT_QUADRO)" (0); "Waiting time costs" with fields for "SEK/Hour (WAIT_LIN)" (0), "SEK/Hour^2 (WAIT_QUAD)" (2500), and "SEK/Hour^4 (WAIT_QUADRO)" (0). The right column is titled "Vehicle Related Costs" and contains two sections: "Small vehicle costs" with fields for "Fixed cost (S_FIX)" (25), "Drive time cost (S_DRIVE)" (250), and "Idle time cost (S_IDLE)" (250); and "Van vehicle costs" with fields for "Fixed cost (V_FIX)" (30), "Drive time cost (V_DRIVE)" (300), and "Idle time cost (V_IDLE)" (300). At the bottom of the window, there are four buttons: "Save to file...", "Load from file...", "OK", and "Cancel".

Figure 6: Description of the Cost settings form in the graphical user interface. Each field corresponds to a part of the total cost function

The Cost settings form and the Simulation settings form give the user the possibility to easily study the effects of different parameter settings. To study the effects of using different types of heuristics to generate schedules, the user must choose a method (or implement a new method) from within the code.

In addition to the possibility of analyzing the results using the statistics described in Section 3, the GUI also provides the possibility to analyze the results by visualizing the routes both in time and space. When visualizing a route in time, the tool plots time windows, maximum ride times, service times, waiting times and the ride times of each request on the route. These plots show what the vehicle is doing at every moment along the route. The tool that plots the routes in space, draws the routes on a map and gives the user the possibility to view one route at a time. This possibility is especially demanded by practitioners, as a complement to the output statistics.

7 Effects of customer related costs

This section will give an example that illustrates how DARS can be used, and the kind of studies it can be used for. In this study, we will evaluate the impact of the customer

Simulation Settings

Maximum Ride Time Constraints

Max ride time if short ride (MRT_SR): 20

Short ride time (SRT): 20

Max ride time factor (MRT_F): 2.5

Time Window Constraints

Time window length 1 (TW1): 30

Time window length 2 (TW2): 15

Time window stretch factor when delivery time specified: 1

Stretch factor for 'right-now' requests: 1

'Right-now' time definition: 0

Waiting Times

Maximum waiting times

Maximum waiting time at pickup (WMAX_P): 10

Maximum waiting time at delivery (WMAX_D): 10

Customers in wheel chairs

Minimum waiting time at pickup (WMIN_WC_P): 4

Minimum waiting time at delivery (WMIN_WC_D): 4

Ambulatory customers

Minimum waiting time at pickup (WMIN_AMB_P): 2

Minimum waiting time at delivery (WMIN_AMB_D): 2

Background Optimization Options

☐ Switch stops

☒ Reassign requests

Nr of Requests to reassign (BO_NR): 1

Number of reassignments (BO_RNR): 11

Save to file...

Load from file...

OK

Cancel

Figure 7: Description of the Simulation settings form in the graphical user interface

related costs. How the customer related costs affect the solution is very important to know when evaluating operational costs against the level of service provided to the customers. The input to the simulations is based on a real-world service operating in the city of Gothenburg, Sweden. In this example, 3072 requests are to be scheduled, and all the requests are to be performed during the same day. Out of these requests, 62% are known the day before (not necessarily 24 hours in advance, but at least before the day of service begins).

There are two different types of vehicles used to serve the requests, conventional taxi vehicles and vans equipped to handle customers using wheelchair. The taxi vehicles have the capacity to serve three riders in ordinary seats, one rider in a seat with extra legroom (front seat) and no riders using wheelchair. The vans can serve three riders in seats with extra legroom and three riders using wheelchair. A rider that does not require extra legroom can of course still be served in such a seat, but not vice versa. The number of vehicles is regarded as unlimited. In practice, there are always as many vehicles available as needed, since the operator requests vehicles from several different taxi companies.

Most paratransit services operating in Sweden consider both hard and soft constraints regarding the waiting time and maximum ride time. The hard constraints of course say that a request must be picked-up within a specific time-window (TW_2) and that the total

ride time of the request must be less than a specific time (MRT_i). However, the customer related costs (the waiting time costs and excess ride time costs) are acting as soft constraints, limiting the deviations from planned pick-up times and from direct ride times, since such deviations impose additional costs to the total cost function.

So, in order to evaluate the effects of using the customer related costs, compared to only using the hard constraints, two different scenarios of the service will be evaluated. In the first scenario, the costs of waiting time and excess ride time will be set in a way that is realistic for Swedish paratransit services. In the second scenario, all customer related costs will be neglected, in this way creating a scenario where only the hard constraints are considered. In this way, no consideration is taken to any customer discomfort.

The parameters that determine the maximum ride time and the time windows are in both scenarios set as

$$\begin{aligned}\theta &= 2.5 \\ M &= 20 \text{ minutes} \\ A &= 20 \text{ minutes} \\ TW_1 &= 30 \text{ minutes} \\ TW_2 &= 15 \text{ minutes}\end{aligned}$$

and the service times are set as

$$STP_i = STD_i = \begin{cases} 4 \text{ minutes,} & \text{if customer } i \text{ uses wheelchair} \\ 2 \text{ minutes,} & \text{otherwise} \end{cases}$$

Vehicle related cost parameters are set as

$$\alpha_k = \beta_k = \begin{cases} 250, & \text{if vehicle } k \text{ is a taxi vehicle} \\ 300, & \text{if vehicle } k \text{ is a van} \end{cases}$$

and

$$\gamma_k = \begin{cases} 25, & \text{if vehicle } k \text{ is a taxi vehicle} \\ 30, & \text{if vehicle } k \text{ is a van} \end{cases}$$

The customer related costs are in the first scenario set to $v_2^i = 4000$ and $\vartheta_2^i = 1000$) (and all other customer related cost parameters are set to 0) for all requests. In the second scenario, also v_2 and ϑ_2 are set to 0.

The results of the simulations are presented in Table 1. From these results, it can be seen very clearly that from the operator's point of view, the efficiency of the service increases. The distance travelled with passengers has only decreased a little, but the distance travelled without passengers has decreased a lot (57%), resulting in that the total distance has decreased by 26%. These shorter distances also mean that ride-sharing has

increased. The maximum number of requests simultaneously onboard a vehicle has increased from only 2 (meaning that very little ride-sharing occurs) to 6. As a result of this, also the average number of requests onboard a vehicle has improved substantially.

Table 1: Simulation results

	Scenario 1	Scenario 2	Impact
Maximum No. of requests onboard a vehicle	2	6	200%
Average No. of requests onboard a vehicle	0.64	1.64	156.5%
Total distance of vehicles (km)	34088465	25178987	-26.1%
Total distance with passengers (km)	19730196	18953817	-3.9%
Total distance without passengers (km)	14358269	6225170	-56.6%
Total request distance (km)	22235909	33283116	49.7%
Total request excess ride time (s)	165892	1817440	995.6%
Total deviation from <i>PPT</i> (s)	35152	886327	2421.4%

As a result of the increased ride-sharing, the discomfort in form of excess ride times and deviations from planned pick-up times has also increased. The total excess ride time has increased very much, giving an increase in average per request from just less than 1 minute to almost 10 minutes. The deviation from planned pick-up times has increased even more, from on average only 11 seconds (meaning that very few requests were re-planned) to an average of almost 5 minutes.

The results show that it makes a huge difference to the efficiency of the service if one values the discomfort of passengers as in the first scenario, or if only the hard constraints are considered, as in the second scenario. This small example shows the possibilities of making an analysis of the simulation results and it also shows the complex relations between the design and parameter setting and the simulation output.

In many real-world applications it would not be realistic to completely neglect customer related costs. However, this provides an estimation of which improvements from the operator’s perspective could be reached. More realistic is to reduce the customer related costs, without fully neglecting them.

8 Conclusions and future research

This paper has presented DARS, a modeling system for simulation of dial-a-ride services. DARS contributes to the development of more efficient transportation of the elderly and disabled, since it makes it possible to study the effects of using alternative designs, cost structures and planning methodologies for such services. DARS was developed on the basis of the planning conditions of dial-a-ride services occurring in most larger Swedish cities. One important characteristic of the planning conditions in these cities is that customers have the right to demand service with very short call ahead time, resulting in a need to use fast heuristics for generating schedules. Further, the conditions can be very

different from city to city with regard to the type of design that is implemented and with regard to chosen cost structures. We have thus made it easy to change the design and cost parameters, and to use alternative heuristics to generate schedules, using DARS. The specifications in the current version of DARS presented in this paper can easily be modified to represent most other planning conditions and rules of operation, and we therefore believe that DARS can be of great use in the development of any future dial-a-ride service.

The short simulation study performed to illustrate the usability of DARS, has shown which efficiency improvements can be reached if less consideration is taken to customer discomfort. They also show what sort of customer discomfort would be the price to pay for these improvements.

DARS can be used both by practitioners and researchers. Practitioners can use DARS directly to study how different designs and cost parameter settings affect the efficiency of their service. Results from such studies may very easily be adapted and thereby lead to direct changes to how the service is operated. Researchers can use the simulation framework to study how different regulations, design and planning methods affect the service in a more general context. Examples of research questions that are being studied currently under Swedish conditions include the effects of using exact (address-based) distances instead of precalculated distance matrices, the effects of using different reoptimization algorithms and the effects of various parameter settings.

Acknowledgements

This research is part of a project financed by the Swedish Governmental Agency for Innovation Systems (VINNOVA) and the Swedish Transport Administration. The work has been done in collaboration with Planit Sweden AB and Malmator.

References

- Attanasio, A., J. Cordeau, G. Ghiani, and G. Laporte (2004). Parallel tabu search heuristics for the dynamic multi-vehicle dial-a-ride problem. *Parallel Computing* 30, 377–387.
- Bailey, W. and T. Clark (1987). A simulation analysis of demand and fleet size effects on taxicab service rates. In *WSC '87: Proceedings of the 19th conference on Winter simulation*, pp. 838–844.
- Baugh, J., G. Kakivaya, and J. Stone (1998). Intractability of the dial-a-ride problem and a multiobjective solution using simulated annealing. *Engineering Optimization* 30, 91–123.
- Beaudry, A., G. Laporte, T. Melo, and S. Nickel (2010). Dynamic transportation of patients in hospitals. *OR Spectrum* 32, 77–107.

- Berbeglia, G., J. Cordeau, and G. Laporte (2010). Dynamic pickup and delivery problems. *European Journal of Operational Research* 202, 8–15.
- Borndörfer, R., M. Grötschel, F. Klostermeier, and C. Küttner (1997). Telebus Berlin: Vehicle scheduling in a dial-a-ride system. Technical Report SC 97-23, Konrad-Zuse-Zentrum für Informationstechnik, Berlin.
- Colomi, A. and G. Righini (2001). Modeling and optimizing dynamic dial-a-ride problems. *International transactions in operational research* 8, 155–166.
- Cordeau, J. (2006). A branch-and-cut algorithm for the dial-a-ride problem. *Operations Research* 54, 573–586.
- Cordeau, J. and G. Laporte (2003a). The dial-a-ride problem (DARP): Variants, modeling issues and algorithms. *4OR - A Quarterly Journal of the Belgian, French and Italian Operations Research Societies* 1, 89–101.
- Cordeau, J. and G. Laporte (2003b). A tabu search heuristic for the static multi-vehicle dial-a-ride problem. *Transportation Research Part B* 37, 579–594.
- Cordeau, J. and G. Laporte (2007). The dial-a-ride problem: Models and algorithms. *Annals of Operations Research* 153, 29–46.
- Cordeau, J., G. Laporte, J. Potvin, and M. Savelsbergh (2004). Transportation on demand. Technical Report CRT-2004-25, Centre For Research on Transportation.
- Deflorio, F., B. D. Chiara, and A. Murro (2002). Simulation and performance of DRTS in a realistic environment. In *Proceedings of the 13th Mini-Euro Conference Handling uncertainty in the analysis of Traffic and Transportation systems and the 9th Meeting of the Euro Working Group on Transportation Intermodality, Sustainability and Intelligent transport systems*, pp. 622–628.
- Diana, M., M. Dessouky, and N. Xia (2006). A model for the fleet sizing of demand responsive transportation services with time windows. *Transportation Research Part B* 40, 651–666.
- Fu, L. (2002a). Scheduling dial-a-ride paratransit under time-varying, stochastic congestion. *Transportation Research Part B* 36, 485–506.
- Fu, L. (2002b). A simulation model for evaluating advanced dial-a-ride paratransit systems. *Transportation Research Part A* 36, 291–307.
- Fu, L. and S. Tepley (1999). On-line and off-line routing and scheduling of dial-a-ride paratransit vehicles. *Computer-Aided Civil and Infrastructure Engineering* 14, 309–319.
- Gerrard, M. (1974). Comparison of taxi and dial-a-bus services. *Transportation Science* 8, 85–101.

- Häll, C., H. Andersson, J. Lundgren, and P. Värbrand (2009). The integrated dial-a-ride problem. *Public Transport* 1, 39–54.
- Haugland, D. and S. Ho (2010). Feasibility testing for dial-a-ride problems. In B. Chen (Ed.), *Lecture notes in computer science*, 6124, pp. 170–179. Springer-Verlag Berlin Heidelberg.
- Heathington, K., J. Miller, R. Knox, G. Hoff, and J. Bruggeman (1968). Computer simulation of a demand scheduled bus system offering door-to-door service. *Highway Research Record* 251, 26–40.
- Ho, S. and D. Haugland (2011). Local search heuristics for the probabilistic dial-a-ride problem. *OR Spectrum* 33, 961–988.
- Högberg, M. (2008). On improving paratransit scheduling by using more accurate distance matrices, local search and demand estimation. Master Thesis E332, Optimization and Systems Theory, KTH.
- Hunsaker, B. and M. Savelsbergh (2002). Efficient feasibility testing for dial-a-ride problems. *Operations Research Letters* 30, 169–173.
- Ioachim, I., J. Desrosiers, Y. Dumas, M. Solomon, and D. Villeneuve (1995). A request clustering algorithm for door-to-door handicapped transportation. *Transportation Science* 29, 63–78.
- Jaw, J., A. Odoni, H. Psaraftis, and N. Wilson (1986). A heuristic algorithm for the multi-vehicle advance request dial-a-ride problem with time windows. *Transportation Research Part B* 20, 243–257.
- Madsen, O., H. Ravn, and J. Moberg Rygaard (1995). A heuristic algorithm for a dial-a-ride problem with time windows, multiple capacities, and multiple objectives. *Annals of Operations Research* 60, 193–208.
- Noda, I., M. Ohta, K. Shinoda, Y. Kumada, and H. Nakashima (2003). Evaluation of usability of dial-a-ride systems by social simulation. In *Proc. of Fourth International Workshop on Multi-Agent-Based Simulation*, pp. 139–152.
- Parragh, S., K. Doerner, and R. Hartl (2008). A survey on pickup and delivery problems: Part II: Transportation between pickup and delivery locations. *Journal für Betriebswirtschaft* 58, 81–117.
- Psaraftis, H. (1980). A dynamic programming solution to the single vehicle many-to-many immediate request dial-a-ride problem. *Transportation Science* 14, 130–154.
- Psaraftis, H. (1983). An exact algorithm for the single vehicle many-to-many dial-a-ride problem with time windows. *Transportation Science* 17, 351–357.

- Quadrifoglio, L., M. Dessouky, and F. Ordóñez (2008). A simulation study of demand responsive transit system design. *Transportation Research Part A* 42, 718–737.
- Ropke, S., J. Cordeau, and G. Laporte (2007). Models and branch-and-cut algorithms for pickup and delivery problems with time windows. *Networks* 49, 258–272.
- Tang, J., Y. Kong, H. Lau, and A. Ip (2010). A note on "Efficient feasibility testing for dial-a-ride problems". *Operations Research Letters* 38, 405 – 407.
- Teodorovic, D. and G. Radivojevic (2000). A fuzzy logic approach to dynamic dial-a-ride problem. *Fuzzy sets and systems* 116, 23–33.
- Toth, P. and D. Vigo (1997). Heuristic algorithms for the handicapped persons transportation problem. *Transportation Science* 31, 60–71.
- Uchimura, K., H. Takahashi, and T. Saitoh (2002). Demand responsive services in hierarchical public transportation system. *IEEE Transactions on Vehicular Technology* 51, 760–766.
- Wilson, N., J. Sussman, L. Goodman, and B. Higonnet (1969). Simulation of a computer aided routing system (CARS). In *Proceedings of the third conference on Applications of simulation*, pp. 171–183.